uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Amir Homayoun Kamkar-Parsi
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Electrical Engineering)
GRADE / DEGREE

School of Information Technology and Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Signal Estimators and Noise Reduction Schemes for High-End Binaural Hearing Aids
TITRE DE LA THÈSE / TITLE OF THESIS

Dr. Martin Bouchard
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Miodrag Bolic

Benoît Champagne (McGill University)

Christian Giguère

Rafik Goubran

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Signal Estimators and Noise Reduction Schemes for High-End Binaural Hearing Aids

A. Homayoun Kamkar-Parsi

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-61398-6
Our file Notre référence
ISBN: 978-0-494-61398-6

UNIVERSITY OF OTTAWA

OTTAWA-CARLETON INSTITUTE FOR ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies and Postdoctoral Studies for acceptance a thesis entitled "**Signal Estimators and Noise Reduction Schemes for High-End Binaural Hearing Aids**" by **A. Homayoun Kamkar-Parsi** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Computer Engineering.**

Dated: August 2009

Supervisor:

Dr. Martin Bouchard

Readers:

Dr. Benoît Champagne

Dr. Rafik A. Goubran

Dr. Christian Giguère

Dr. Miodrag Bolic

Abstract

There is a variety of hearing aid models available in the marketplace, which vary in terms of size, shape and effectiveness. Due to size constraints, for certain types of hearing aids only a single microphone per hearing aid can be fitted, and only single-channel monaural noise reduction schemes can be integrated in them. In the near future, binaural hearing aids will become available, making use of a wireless link to share information between the two ears. In this thesis, a binaural diffuse noise power spectral density estimator is first developed. The estimator does not require a voice activity detection algorithm, it provides an estimate at any time i.e. during speech activity or not, it has no noise tracking latency and it does not assume that the target speaker is in front of the binaural hearing aid user. An instantaneous binaural target speech power spectral density estimator is then developed, for conditions of speech corrupted by lateral interfering noise. The estimator does not require the knowledge of the direction of the noise source, and the noise source can be highly non-stationary or transient. Finally, a complete binaural noise reduction scheme is designed, by incorporating the two binaural estimators mentioned above. The proposed reduction scheme is designed for complex real-life environments composed of time-varying diffuse noise, multiple non-stationary directional noises and reverberant conditions. The proposed scheme also preserves the original interaural cues of both the target speech and the background noises. Simulations using recorded signals provided by a hearing aid manufacturer are performed in the thesis to validate the performance of the proposed estimators and the proposed noise reduction algorithm. The results indicate that the proposed binaural noise reduction scheme proves to be a good candidate for the noise reduction stage of upcoming binaural hearing aids.

Acknowledgements

First and foremost, I would like to express my profound respect and gratitude to my supervisor, Dr. Martin Bouchard. Throughout my graduate studies, Dr. Bouchard offered me continuous support and encouragement as well as invaluable help and insight into my research. Furthermore, he had always been the person whom I could turn to generate ideas and make informative technical discussions.

I also would like to thank very much my parents and my sister for their ongoing support, care, love and encouragement throughout this long journey. A special thank you to my father who always pushed me to achieve excellence from the very beginning. As well, I would like to thank very much a very special person in my life, my close friend Manos. Manos has always provided me with unfailing support and encouragement, and a unique friendship and kindness all those years.

Finally, I am very grateful for the financial support I have received for my Ph.D. research. I would to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship, Siemens Audiologische Technik GmbH Group and the University of Ottawa Excellence Postgraduate Scholarship.

Contents

LIST OF FIGURES ..... **ix**

LIST OF TABLES ..... **xiii**

LIST OF ACRONYMS ..... **xiv**

CHAPTER 1. INTRODUCTION **1**

1.1 MOTIVATION AND PREVIOUS WORK 1

1.2 OBJECTIVES 5

1.3 CONTRIBUTIONS 8

CHAPTER 2. GENERAL OVERVIEW OF THE HUMAN EAR AND BINAURAL HEARING CONCEPTS **10**

2.1 BRIEF DESCRIPTION OF THE HUMAN EAR 10

2.2 BINAURAL HEARING 14

2.2.1 Lateral Sound Source Perception (Azimuth) *15*

2.2.2 Sound Elevation Perception *18*

2.2.3 Distance Perception *19*

2.2.4 Auralization *20*

2.2.5 Binaural Speech Intelligibility *22*

CHAPTER 3. BACKGROUND ON NOISE PSD ESTIMATION AND MONAURAL NOISE REDUCTION TECHNIQUES **24**

3.1 OVERVIEW OF ADVANCED NOISE PSD ESTIMATION TECHNIQUES 25

3.1.1 Minimum Statistics Method *25*

3.1.2 Cross-Power Spectral Density Method *28*

3.1.2.1 Two-Microphone System Description 29

3.1.2.2 Noise PSD Estimation via Cross-Power Spectral Density Method 30

3.2 OVERVIEW OF MONAURAL NOISE REDUCTION TECHNIQUES 31
3.2.1 Spectral Subtraction *33*
3.2.2 Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator *35*
3.2.3 Geometric Approach to Spectral Subtraction *37*
3.2.4 Kalman Filtering-Based Speech Enhancement *39*

CHAPTER 4. OVERVIEW OF PREVIOUS BINAURAL NOISE REDUCTION SCHEMES 45

4.1 BINAURAL WIENER FILTERING 45
4.1.1 Binaural Wiener Filtering Design *46*
4.1.2 Insights on Binaural Wiener Solution *51*
4.1.3 Binaural Wiener Implementation Proposed in [BOG'07] *56*
4.2 BINAURAL MVDR-BASED BEAMFORMING WITH POST-PROCESSING 61
4.3 RELATION BETWEEN THE BINAURAL WIENER FILTER AND MVDR-CONSTRAINED BEAMFORMING 65

CHAPTER 5. OVERVIEW OF OBJECTIVE MEASURES FOR NOISE REDUCTION PERFORMANCE EVALUATION 69

CHAPTER 6. PROPOSED BINAURAL DIFFUSE NOISE POWER SPECTRUM DENSITY ESTIMATOR 73

6.1 ACOUSTICAL ENVIRONMENT: DIFFUSE NOISE FIELD 73
6.2 BINAURAL SYSTEM DESCRIPTION 75
6.3 PROPOSED BINAURAL DIFFUSE NOISE POWER SPECTRUM ESTIMATION 77
6.3.1 Diffuse Noise PSD Estimation *78*
6.3.2 Direct Computation of the Error Auto-Power Spectrum *82*
6.3.3 The Direct Computation $\Gamma_{EE}(\omega)$ Versus the Indirect Computation $\Gamma_{EE\_1}(\omega)$ *85*

6.3.4 Simulation Results Comparing the Error Components: $\Gamma_{EE\_err}(\omega)$ Versus $\Gamma_{EE\_1\_err}(\omega)$ *87*

6.3.5 Low frequency Compensation *91*

6.4 CASE OF ADDITIONAL DIRECTIONAL INTERFERENCES 97

CHAPTER 7. DIFFUSE NOISE ESTIMATION TECHNIQUES EVALUATION 99

7.1 SIMULATION SETUP AND HEARING SITUATIONS 99

7.2 NOISE ESTIMATION TECHNIQUES EVALUATION 102

7.2.1 PBNE versus CPSM *105*

7.2.2 PBNE versus MSA *117*

7.2.3 Summary of Proposed Binaural Noise Estimation Technique Performance *124*

CHAPTER 8. BINAURAL TARGET SPEECH PSD ESTIMATOR FOR TARGET SPEECH CORRUPTED BY DIRECTIONAL NOISE 125

8.1 ACOUSTICAL ENVIRONMENT: LATERAL (DIRECTIONAL) NOISE 127

8.2 BINAURAL SYSTEM DESCRIPTION FOR PROPOSED PROBLEM 129

8.3 PROPOSED BINAURAL INSTANTANEOUS TARGET SPEECH SPECTRUM ESTIMATOR 130

8.3.1 Target Speech PSD Estimation *131*

8.3.2 Direct Computation of the Error Auto-Power Spectrum *134*

8.3.3 Finalizing the Target Speech PSD Estimator *138*

8.3.4 Case of Non-Frontal Target Sources *140*

CHAPTER 9. AN EXAMPLE INTEGRATING THE BINAURAL TARGET SPEECH PSD ESTIMATOR INTO A NOISE REDUCTION SCHEME 143

9.1 BINAURAL WIENER FILTERING NOISE REDUCTION SCHEME 145

9.2 INTEGRATION OF THE TARGET SPEECH PSD ESTIMATOR 146

9.3 MODIFICATION TO PRESERVE INTERAURAL CUES 148
9.4 SIMULATION SETUP AND HEARING SITUATIONS 149
9.5 RESULTS AND DISCUSSIONS 153

CHAPTER 10. PROPOSED COMPLETE BINAURAL NOISE REDUCTION SCHEME OPERATING IN COMPLEX ACOUSTIC NOISY ENVIRONMENTS 158

10.1 BINAURAL SYSTEM DESCRIPTION 160
10.2 PROPOSED BINAURAL NOISE REDUCTION SCHEME 161
10.3 DESCRIPTION OF EACH STAGE OF THE PROPOSED SCHEME 165
10.3.1 Stage 1 165
10.3.2 Stage 2 168
10.3.3 Stage 3 170
10.3.4 Stage 4 171
10.3.5 Stage 5 175
10.3.6 Case of Non-Frontal Target Source 176
10.4 SIMULATION SETUP AND SELECTED COMPLEX HEARING SITUATIONS 181
10.5 SIMULATION RESULTS AND DISCUSSION 182

CHAPTER 11. CONCLUSION AND FUTURE WORK 189

11.1 CONCLUSION 189
11.2 FUTURE WORK 190

CHAPTER 12. REFERENCES 193

List of Figures

Figure 2-1: Cross-sectional view of the human ear 11
Figure 2-2: Simplified hearing concept 13
Figure 2-3: Diagram of the spherical coordinate system 14
Figure 2-4: Cone of confusion, different spatial directions producing identical ITDs and ILDs 16
Figure 2-5: Illustration of two signal waves from a frontal sound source reaching the ear and a signal reflection caused by the pinna 19
Figure 2-6: Simple binaural synthesis 21
Figure 3-1: Partial block diagram of the noise power spectral density estimation via Minimum Statistics Approach 27
Figure 4-1: Binaural multichannel Wiener system 46
Figure 4-2: Binaural superdirective beamformer with post-filtering proposed in [LOT'06] 61
Figure 6-1: Binaural hearing aids user in a noisy diffuse field environment with a single target source speaker 75
Figure 6-2: Overall diagram for proposed binaural noise estimation method 78
Figure 6-3: True Wiener filter versus noisy Wiener filter magnitude responses 88
Figure 6-4: a) Magnitude graph of analytical $\tilde{\Gamma}_{EE}(\omega)$ versus the corresponding experimental graph; b) Magnitude graph of the analytical error, $\left|\varepsilon_{EE\_err}(\omega)\right|$, versus the corresponding experimental error graph. All obtained using the direct computation method 89
Figure 6-5: a) Magnitude graph of analytical $\tilde{\Gamma}_{EE\_1}(\omega)$ versus the corresponding experimental magnitude graph; b) Magnitude graph of the analytical error, $\left|\varepsilon_{EE\_1\_err}(\omega)\right|$, versus the corresponding experimental error graph. All obtained using the indirect computation method 90
Figure 6-6: Magnitude of analytical error for the direct and indirect computation methods 91

Figure 7-1: Scenario A, the target speaker is in front of the binaural hearing aid user in a diffuse noise field environment 100
Figure 7-2: Scenario B, the target speaker is at 90° on the right of the binaural hearing aid user in a diffuse noise field environment 101
Figure 7-3: Experimental coherence of real binaural cafeteria noise (Blue) versus the corresponding modified analytical diffuse noise model (Red) and the unmodified analytical diffuse noise model assuming 2 omni-directional microphones in free space (Green) 105
Figure 7-4: Left (top - red) and right (bottom - black) noise-free speech signals for scenario A 108
Figure 7-5: Left (top - red) and right (bottom - black) noisy speech signals and the frame that has been processed (blue) for scenario A 109
Figure 7-6: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs for scenario A 109
Figure 7-7: Noise estimation results for scenario A. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black).. 110
Figure 7-8: Noise estimation results shown for frequencies from 0 to 500 Hz only. PBNE (top - blue) versus CPSM (bottom - red), original left and right noise PSDs (top or bottom - black) 111
Figure 7-9: Noise estimation results from an average over 20 realizations. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black) 111
Figure 7-10: Noise estimation results for PBNE with the low frequency compensation (top – blue) and without (bottom – blue), original left and right noise PSDs (top or bottom – black) 112
Figure 7-11: PBNE noise PSD estimation results (blue) from an average over 20 realizations with various head diameters and gain factors. Original left and right noise PSDs (black). 112
Figure 7-12: Error offsets of the PBNE noise PSD estimation for various head diameters and factors 113

Figure 7-13: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs used for scenario B 114
Figure 7-14: Noise estimation results for scenario B. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black) 115
Figure 7-15: Noise estimation results for an average of 20 realizations for scenario B. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black) 115
Figure 7-16: Scenario C, average over 20 realizations for the Left (blue) and right (red) white noise target source PSDs and corresponding left (black) and right (green) original noise PSDs 116
Figure 7-17: Noise estimation results for an average of 20 realizations for scenario C. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black) 117
Figure 7-18: Left (top - black) and right (bottom - red) noisy speech signals 119
Figure 7-19: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs used for scenario A at SNR=11 dB 119
Figure 7-20: Noise estimation results over an average of 585 frames and at SNR = 11dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSDs (top or bottom - black) 120
Figure 7-21: Noise estimation results over an average of 585 frames and at SNR = 5dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSD (top or bottom - black) 120
Figure 7-22: Noise estimation results over an average of 585 frames and at SNR = 0dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSD (top or bottom - black) 121
Figure 7-23: Noisy right speech signal power (blue) at each frame and corresponding noise power (back) increased at frame index 200 and then decreased at index 400 (black), used in Scenario D 122

Figure 7-24: Noise power estimation result at each frame for MSA in scenario D with highly non-stationary noise. MSA (red), original right noise power (black) 123
Figure 7-25: Noise power estimation result for PBNE at each frame in scenario D with highly non-stationary noise. PBNE (red), original right noise power (black) 123
Figure 8-1: Binaural hearing aid user in front of the target speaker and a lateral interfering talker in background 128
Figure 8-2: Overall diagram for the proposed binaural target speech PSD estimation method 131
Figure 10-1: Overall structure of the Proposed Binaural Noise Reduction (PBNR) scheme 162
Figure 10-2: Enhancement results using the BSBp, GeoSP and PBNR methods. Only the results for the left channel are shown, for a short segment. 188

List of Tables

Table 9-1: Objective Performance Results for Scenario a). 157
Table 9-2: Objective Performance Results for Scenario b). 157
Table 10-1: Diffuse Noise PSD Estimator 178
Table 10-2: Classifier and Noise PSD Adjuster 178
Table 10-3: MMSE-STSA 179
Table 10-4: Target Speech PSD Estimator 179
Table 10-5: Kalman Filtering 180
Table 10-6: Objective Performance Results for Left and Right Input SNRs at 2.1 dB and 4.6 dB Respectively 187
Table 10-7: Objective Performance Results for Left and Right Input SNRs at -3.9 dB and -1.4 dB Respectively 188
Table 10-8: Objective Performance Results for Left and Right Input SNRs at –13.5 dB and -11.0 dB Respectively 188

List of Acronyms

AR	Auto Regressive
BSP	Binaural Superdirective Beamformer
BSPp	Binaural Superdirective Beamformer with Post-filtering
BTE	Behind-The-Ear
CPSM	Cross-Power Spectral Density Method
CSII	Coherence Speech Intelligibility Index
DFT	Discrete Fourier Transform
EBMW	Extended Binaural Multichannel Wiener
FFT	Fast Fourier Transform
GeoSP	Geometric approach Spectral subtraction
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
ILD	Interaural Level Difference
ITC	In-The-Canal
ITD	Interaural Time Difference
ITE	In-The-Ear
KEMAR	Knowles Electronic Manikin for Acoustic Research
LPC	Linear Predictive Coding
MIT	Massachusetts Institute of Technology
MMSE-STSA	Minimum Mean-Square Error Short-time Spectral Amplitude Estimator
MSA	Minimum Statistics based Approach
MSC	Magnitude-Squared Coherence
MVDR	Minimum Variance Distortionless Response
PBNE	Proposed Binaural Noise Estimator
PBNR	Proposed Binaural Noise Reduction scheme
PBTE_NR	Proposed Binaural Target Estimator - Noise Reduction
PESQ	Perceptual Evaluation of Speech
PSD	Power Spectral Density

PSM	Perceptual Similarity Measure
SegSNR	Segmental SNR
SNR	Signal-To-Noise Ratio
TMR	Target-To-Masker energy Ratio
VAD	Voice Activity Detection
WB-PESQ	Wideband PESQ

Chapter 1. INTRODUCTION

1.1 Motivation and Previous Work

One of the major problems for sensorineural hearing-impaired individuals is the increased difficulty in understanding speech in noisy environments caused mainly by the loss of temporal and spectral resolution in the auditory processing of the impaired ear. The primary basis of hearing loss is due to aging, damage caused by severe noise exposure, or other trauma. Studies have shown that this loss in terms of Signal-to-Noise Ratio (SNR) can be about 4 to 10 dB [HAM'05]. However, with the constant development in semi-conductor technology, current advanced digital hearing aids incorporate powerful speech processing techniques to combat this issue. Speech processing in hearing aids can be essentially divided into three domains. The first domain focuses on using advanced speech signal processing techniques to identify and compensate for diverse hearing impairments. For example, the compensation for the loss of loudness (i.e. inability to hear soft sounds) performed by the use of perceptual-models based on multi-band compression and amplification techniques, and also the use of spectral contrast enhancement to adjust for the loss of frequency selectivity (i.e. inability to separate sounds or sound components of different frequencies). As a result, they almost provide complete restoration of speech intelligibility in quiet to the degree of normal hearing. However, they are not able to restore speech intelligibility in noisy environments. This is due to the fact that only applying an amplification stage to compensate for the loss of loudness or for the reduction of dynamic range in one or several frequency bands will also amplify speech as well as noise. Hence, the second domain is devoted to speech enhancement or noise reduction techniques prior to the amplification stage in advanced hearing aids. The latter will be the entire focus of this thesis. Finally, the third domain explores the daily use of hearing aids, addressing issues such as flexibility, convenience, cosmetic, acoustic feedback problems and other artefacts [LUO'03].

Even though noise reduction techniques are implemented in current high-end hearing aids to help against noisy environment, most current noise reduction algorithms are still monaural, that is each hearing aid is using only the information available from its own channel (i.e. ear) to perform a noise reduction scheme. As a result, considering the psychoacoustics of how humans hear and perceive desired speech in noisy environments, and in particular the way that humans are able to "focus" on a given speech even in very noisy environments with background noise or competing speech i.e. the so-called "cocktail party effect", the use of monaural signal processing methods is sub-optimal due for instance to the lack of spatial information. Monaural noise reductions strategies may improve the signal-to-noise ratio but they could not so far establish an increase of speech intelligibility [HAM'05][MAR'03].

Nevertheless, in the near future, a new generation of binaural hearing aids will be available. Those intelligent hearing aids will use and combine the simultaneous information available from hearing aid microphones in each ear (i.e. left and right channels). Such a system is called a binaural system, as in the binaural hearing of humans, taking advantage of the two ears and the relative differences found in the signals received by the two ears. Binaural hearing plays a significant role for understanding speech when speech and noise are spatially separated. Those new binaural hearing aids will allow the sharing and exchange of information (or signals) received via a wireless link from both left and right hearing aid microphones, and will also generate an output for the left and right ear. Unfortunately, the few previous attempts to include binaural processing in hearing aids noise reduction algorithms have not yet been able to successfully demonstrate the potential improvement to be granted by such processing.

For instance, in [WIT'97], a potential technique that would use binaural information to suppress reverberation and lateral noise source was introduced. Even though in a laboratory environment hearing-impaired subjects claimed an improved speech perception in certain noise conditions, however, in real-life circumstances, the results turned out to be quite poor due to unstable and unreliable directional filter gain factors in

the presence of multiple interferers or diffuse noise [WIT'03]. In [HAM'02], the performance evaluation of monaural and binaural spectral subtraction algorithms as well as a conventional directional microphone was investigated. In the monaural case, the noise power spectrum was estimated as in [MAR'94] by tracking the spectral minima of a smoothed power spectrum estimate of the noisy speech signal, followed by a spectral subtraction scheme. For the binaural case, the work in [DOE'96] was implemented where the noise power spectrum is estimated binaurally but with again a monaural spectral subtraction enhancement scheme applied to each channel respectively. The evaluation was conducted under a diffuse noise scenario (i.e. cafeteria babble noise), with a loudspeaker as a target source in front of a dummy head wearing two standard Behind-The-Ear hearing aids. As a result, the best performance in terms of noise reduction was obtained with the directional microphone, as opposed to both spectral subtraction schemes. However, much greater noise reduction was achieved with the combination of a directional microphone with a monaural or a binaural spectral subtraction technique applied to each hearing channel. The binaural spectral subtraction technique only provided minor improvements compared to the monaural one. This is partly because the methods have been focusing on performing binaural noise estimation only, and not on performing truly binaural de-noising or speech enhancement. Also, despite the fact that directional microphones have some important benefit in terms of noise reduction and very low speech distortion, they cannot be implemented in all types of hearing aids (i.e. Inside-The-Ear) due to size constraints (i.e. space limitations on the device [LOT'06] [HAM'05]) and because the assumption of free sound field is not met inside the ear canal [HAM'05].

In [DES'97] and [WEL'97], the authors present binaural noise reduction schemes with a tradeoff between noise reduction and the preservation of interaural cues. For instance, in [WEL'97] the algorithm first filters the binaural inputs into a high-pass and a low-pass segment using the same cut-off frequency. For binaural localization, studies have shown that ITD cues are relevant mostly at low frequencies [GEL'04]. The noise reduction scheme then only processes the high frequency segment and the result is added to the delayed low frequency segment. However the major drawback to this approach is that the

low frequency region contains target speech with undistorted ITD but with unprocessed noise. Therefore, there is a trade-off between noise reduction and speech ITD cues preservation. As the cut-off frequency increases, the preservation of the ITD cues improves at the cost of noise reduction.

In [WIT'03], the authors proposed a selective noise reduction strategy depending on environment classification such as diffuse noise field or lateral noise only. The classification is performed by general diffusiveness or coherence of the sound field. For a diffuse noise environment classification, spectral gains are computed based on the coherence function and applied to input noisy frames reducing diffuse noise. For a directional noise environment classification: Interaural Level and Time Differences (ILD, ITD) are compared with reference values for a frontal target direction. Signal segments with interaural differences deviating from the reference are filtered (attenuated), otherwise they remain unchanged (directional filtering). However, this directional filtering works only with low levels of diffuse background noise, otherwise the interaural measures are unreliable causing artefacts and should be switched off. Then again, this binaural algorithm presents performance degradation in the presence of both diffuse noise and lateral noises.

In [YAN'03], a binaural noise reduction system intended for potential binaural hearing aids was explored based on the initial framework in [ALL'77] to reduce lateral noise. Their approach is to divide at first the binaural noisy signal into several frequency bands, in which the interaural difference in phase and gain are determined by the short-term auto-correlation and cross-correlation. Then a second gain is evaluated based on a single frequency band. The combination of the two gains is examined and the outcome is applied to each frequency band. The resulting gain allows highly correlated signals to pass, while uncorrelated signals are attenuated. The author claims an improvement of SNR greater than the Delay-and-Sum approach. But under reverberant conditions, the target signal cannot be retrieved. Also, one of the major drawbacks is that only a monaural signal is produced. Consequently, the spatial cues of any lateral residual noise sources will also be lost, which is not acceptable for binaural hearing aids applications.

Furthermore, research has shown that hearing impaired individuals often localize sounds better without their bilateral hearing aids than with them, due to current noise reduction schemes in hearing aids which were not designed to preserve localization cues [DES'97][BOG'06]. In [KLA'06], a new binaural Wiener filter based noise reduction scheme was introduced, which contains a modified cost function that includes terms for ITD and ILD cues. They affirm that they can achieve some noise reduction while limiting the distortion of the speech and noise localization cues. However, their approach relies again on a perfect VAD for the speech and noise statistics to adapt their Wiener filter parameters. The method essentially works for a single lateral noise source under a low reverberation environment and without the presence of diffuse noise. The various statistics have also been estimated off-line in their work. Furthermore, the lateral noise source considered is not another interfering speech signal, which is highly non-stationary, but instead a continuous directional background babble-noise source signal with short-time stationary characteristics. Therefore, in this scenario only, the use of a perfect VAD becomes possible. As a result, complex auditory environments involving non-stationary and/or transient lateral noises cannot be managed using this algorithm.

From this summary of the previous work achieved, it can be concluded that there is a lot of room for improvement in the design of a complete binaural speech de-noising or enhancement scheme that could be implemented on the upcoming binaural hearing aid technology.

1.2 Objectives

In most speech de-noising techniques also referred to as speech enhancement techniques, it is necessary to estimate a priori the characteristics of the noise corrupting the desired signal. Usually, most speech enhancement techniques require the need of Voice Activity Detection (VAD), to estimate the noise power spectrum during speech pauses. However, the use of a VAD to estimate the noise power will mostly be efficient for highly stationary noise environments, which is not found in many daily activities, and the VADs

often fail under situations with low Signal-To Noise Ratios (SNRs) [LOT'06]. Also, for speech enhancement applications, an accurate noise estimate is required at all times (i.e. even during speech activity) in order to perform effectively under non-stationary noise conditions.

Some advanced techniques, which do not require a VAD, were published, for example in [MAR'01]. But these techniques are mostly based on a monaural microphone system, where only a single noisy signal is available for processing. Consequently, only the temporal and spectral characteristics of the received noisy signal can be analysed. In contrast, multiple microphone systems can additionally take into account the spatial distribution of noise and speech sources, using techniques such as beamforming to enhance the noisy speech signal. In the near future, as introduced earlier, the development of binaural hearing aids will lead to new types of noise reduction techniques, taking advantage of the left and right reference noisy signals.

In this thesis, the first objective is to develop a new approach for binaural noise power spectrum estimation in a binaural noise reduction system, which would be implemented in up-coming binaural hearing aids. This proposed binaural noise power spectrum estimator should be able to provide an accurate estimate of the noise during speech pauses as well as during speech activities without the use of a VAD. The performance of the proposed binaural noise estimation algorithm under a diffuse noise field environment is to be investigated. In simple terms, a diffuse noise field occurs when the resulting noise at the two ears comes from all directions, with no particular dominant direction. Such noise is typical of several practical situations (e.g. background babble noise in cafeteria, car noise etc.), and even in non-diffuse noise conditions there is often a significant diffuse noise component due to room reverberation. In a diffuse noise field, the noise components received at both ears are not correlated (i.e. one noise cannot be predicted from the other noise), and they also have roughly the same frequency content (spectral magnitude shape). On the other hand, the speech signal coming from a speaker produces highly correlated components at the left and right ear, especially under low reverberation environments. Consequently, using these conditions, it is possible to derive an exact

formula to identify the spectral magnitude shape of the noise components at the left and right ear. This estimation of the spectral magnitude shape of the noise components is often the key factor affecting the performance of most existing noise reduction or speech enhancement algorithms. Therefore, having a new method that can instantaneously provide a good estimate of this spectral shape, without any assumption about speaker location (i.e. no specific direction of arrival required for the target speech signal) and without any assumption on speech activity, is a useful result. The proposed method is also suitable for highly non-stationary colored noise under the diffuse noise constraint. Some theoretical analysis for the performance of the different versions of the proposed method is achieved in this thesis. The proposed method will be compared with the work of two current advanced noise power estimation techniques from [MAR'01] and [DOE'96]. In [MAR'01], the noise power spectrum is estimated by tracking the spectral minima of a smoothed power spectrum estimate of the noisy speech signal. In [DOE'96], their approach uses the left and right signals of a binaural hearing aid where a combination of auto- and cross-power spectral densities of the noisy binaural signals are used to extract the power spectral density. Moreover, both of these advanced noise power estimation techniques do not rely on voice activity detectors.

Other types of noise such as directional noise sources are also frequently encountered in real-life listening situations and can reduce greatly the understanding of the target speech. For instance, directional noise sources can emerge from strong multi-talkers and transient noises such as hammering, dishes clattering etc. occurring in the background. Therefore the second objective in this thesis is to develop another binaural estimator to retrieve the desired target speech PSD from the noisy binaural input signals by exploiting spatial information available in binaural systems. By incorporating this estimated desired target speech PSD into a noise reduction scheme, it is possible to effectively reduce non-stationary directional noise as well. The proposed scheme will be compared in terms of noise reduction performance to a recent binaural Wiener filtering-based noise reduction scheme designed to reduce directional noise developed in [BOG'07].

Finally, the third objective of this thesis is to develop a complete binaural noise reduction scheme, which should be robust against complex noisy environments such as time-varying diffuse noise, multiple directional non-stationary noises in the background and under reverberating conditions. The proposed scheme will take advantage of the two estimators developed above. Furthermore, the proposed binaural noise reduction scheme should also maintain all the interaural cues of the target speech and the background noises. The spatial impression of the environments should thus mostly remain unchanged after processing. The proposed scheme will be compared to various noise reduction schemes including another binaural beamforming-based noise reduction scheme proposed in [LOT'06]. Various objective measures (providing insights on the speech distortion level, the intelligibility improvement, SNR, etc.) will be used to assess the noise reduction performance results of each scheme under complex noisy environments.

1.3 Contributions

a) In Chapter 6, an improved binaural diffuse noise PSD estimator was developed where the binaural hearing aid user is located in a diffuse noise field environment. Comprehensive performance comparisons with other noise PSD estimation schemes were carried out in Chapter 7 in terms of accuracy and noise tracking latency. The work in Chapters 6 and 7 was structured into a paper published in the following journal:

- A. H. Kamkar-Parsi, M. Bouchard, "Improved Noise Power Spectrum Density Estimation For Binaural Hearing Aids Operating in a Diffuse Noise Field Environment", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, n.4, pp. 521-533, May 2009.

b) In Chapter 8, an instantaneous binaural target speech PSD estimator was developed to extract the desired target PSD corrupted by non-stationary directional noise. In Chapter 9, this estimator was incorporated into a binaural noise reduction scheme referred to as Instantaneous Binaural Wiener Filtering to reduce directional noise. The noise reduction

performance resulting from this Instantaneous Binaural Wiener Filtering was compared with another recent binaural noise reduction scheme proposed in [BOG'07]. The work in Chapters 8 and 9 was structured into a paper and submitted to the following journal:

- A.H. Kamkar-Parsi, and M. Bouchard, "Instantaneous Target Speech Power Spectrum Estimation for Binaural Hearing Aids and Reduction of Directional Nonstationary Noise with Preservation of Interaural Cues", submitted (in Feb. 2009) for publication in *IEEE Trans. on Audio, Speech and Language Processing.*

c) In Chapter 10, the proposed diffuse noise PSD estimator and the proposed target speech PSD estimator were both incorporated into a new binaural noise reduction scheme structured into five stages. The proposed scheme guarantees interaural cues preservation for both the target speech and the background noise. It was designed to be robust for complex acoustic noisy environments consisting of time-varying diffuse noise and multiple non-stationary directional noise sources, under reverberating conditions. The noise reduction performance was also compared with several other noise reduction schemes including another binaural noise reduction scheme proposed in [LOT'06]. This work was submitted to the following journal:

- A.H. Kamkar-Parsi, and M. Bouchard, "Advanced Binaural Noise Reduction Scheme For Binaural Hearing Aids Operating In Complex Noisy Environments", submitted (in Feb. 2009) for publication in *IEEE Trans. on Audio, Speech and Language Processing.*

d) The contributions in a, b and c were gathered into *US provisional patent # 61/149,363 Method and System for a Multi-Microphone Noise Reduction,* filed by Siemens Audiologische Technik Group.

Chapter 2. GENERAL OVERVIEW OF THE HUMAN EAR AND BINAURAL HEARING CONCEPTS

2.1 Brief Description of the Human Ear

The auditory system is defined as the sensory system for hearing. It consists of the ears and their connections to and within the central nervous system [GEL'04]. The function of the human ear in the area of hearing is to translate sound waves into nerve impulses. These impulses are then perceived and deduced by the brain as sound. The human ear can perceive sounds in the range of 20 to 20 kHz [WEB'1].

The peripheral ear consists of three distinctive areas: the outer, the middle and the inner ear. Figure 2-1 illustrates the cross-sectional view of the human ear. The outer ear is made up of the pinna and the ear canal. The pinna is the visible part of the outer ear and more precisely the folds of cartilage surrounding the ear canal. It protects the eardrum and funnels sound waves into the ear canal, but it also acts as a direction-dependent filter that attenuates and reflects sound wave. One of the main roles of the pinna is that it influences localization due to its asymmetric shape and its filtering varies depending on the direction of arrival of the sound source. These resulting spectral variations provide important directional cues, which help the brain determining elevation and front/back distinctions of a sound source. This phenomenon will be more elaborated in the next section.

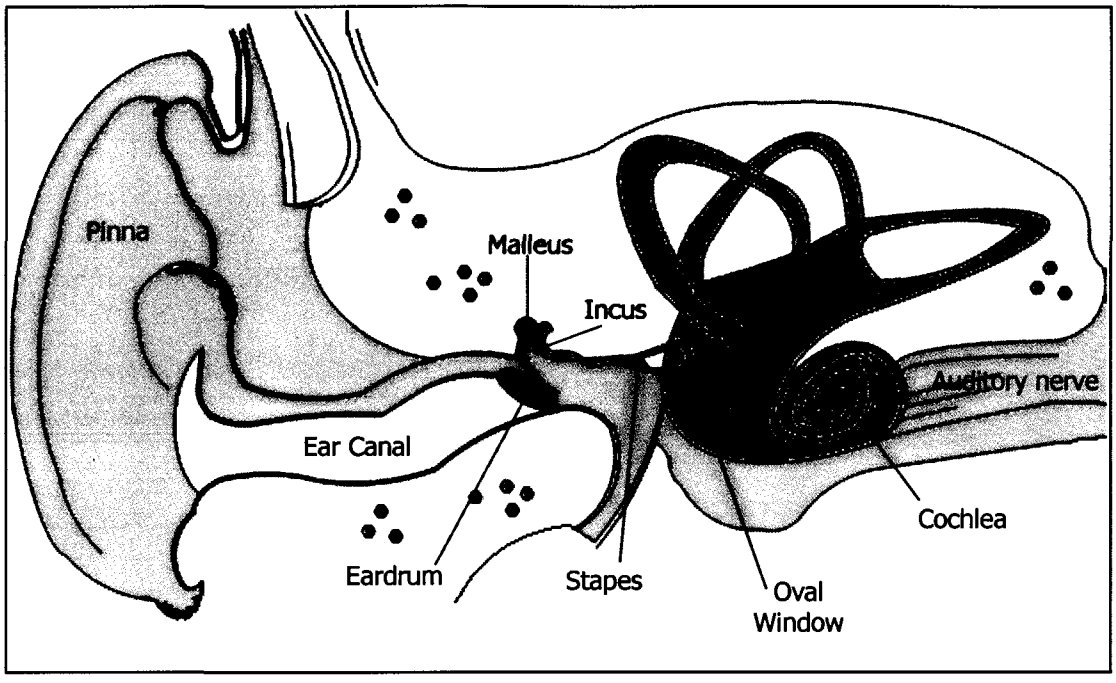

Figure 2-1: Cross-sectional view of the human ear

The ear canal may be visualized as a resonating tube open at one end and closed at the other end by the eardrum. Sounds passing through the ear canal are therefore affected by the acoustic characteristics of this resonator. The human ear canal is about 2.3 cm long and its resonance occurs at the frequency around 3.8 kHz, amplifying sounds in the range of 2-5 kHz [GEL'04][WEB'2].

The eardrum, also referred to as the tympanic membrane, divides the outer and the middle ears. It receives vibrations crossing the ear canal and transmits them across the air-filled middle ear cavity by a series of tiny bones (i.e. ossicles) to the oval window of the inner ear. The oval window is a smaller membrane at the entrance of the inner ear. The ossicles consist of three bones named as malleus (hammer), incus (anvil) and stapes (stirrup). The middle ear's function is to perform an impedance transformation between the air of the outer ear and the liquid of the inner ear by the help of these ossicles. One of the mechanisms of the ossicles is then to achieve an amplification by a so-called "mechanical lever action", which converts lower-pressure eardrum sound vibrations into

higher-pressure at the oval window. Otherwise, if a wave is transferred from a low-impedance medium to one of high impedance such as a liquid without prior impedance matching, a significant amount of its energy is reflected and fails to enter the liquid.

Therefore, for hearing to occur a higher pressure is necessary since beyond the oval window the inner ear contains fluids, and the acoustic impedance of the inner ear fluid is about 4000 times that of air [SHA'00]. At this stage, it should be pointed out that the middle ear still maintains the sound information in waveform and it is only converted to nerve impulses in the cochlea (in the inner ear). Also, another function of the middle ear is to protect the inner ear from loud sounds. When excess pressure is felt, the muscles attached to the ossicles tighten the eardrum and reduces the transfer of force to the oval window, keeping the inner ear from receiving the full impact of sounds that are too loud.

Finally, the cochlea, which is located in the inner ear and is a "snail looking" tube filled with a fluid, converts the vibrations at the oval window input into electrical nerve impulses. The cochlea can be imagined as a microphone converting sound pressure impulses from the outer ear into electrical impulses [GEL'04][WEB'3]. Those electrical impulses are passed on to the brain via the auditory nerves. Figure 2-2 shows a diagram summarizing the concept of hearing and its various stages. This diagram is a modified version of the one found in [WEB'1].

Further processing still remains on the nerve impulses, leading to a multitude of auditory reactions and sensations which are then perceived and interpreted by the brain as sound. This remaining processing is beyond the scope of this thesis, but the reader may refer to [SHA'00][GEL'04] for more explanations.

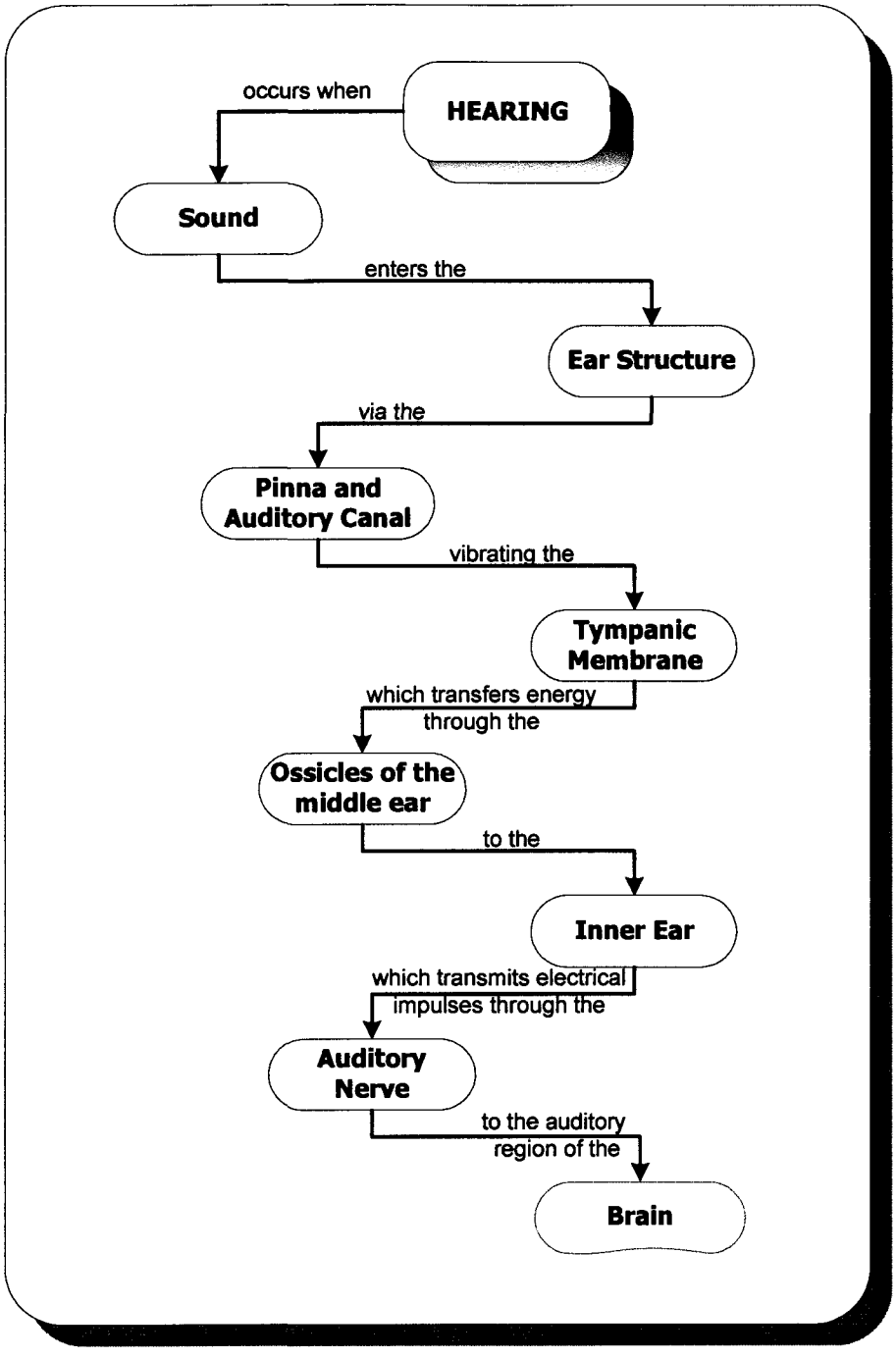

Figure 2-2: Simplified hearing concept

2.2 Binaural Hearing

The term binaural hearing is defined as hearing with both ears as opposed to only one, where the latter is referred to as monaural hearing. Binaural hearing provides the ability to extract auditory information that would not be feasible using one ear only. One major benefit that can be attributed to binaural hearing is the sense of localization of a sound source. Although some degree of localization is achievable in monaural listening (i.e. in the vertical plane), binaural hearing significantly improves our ability to sense the direction of arrival of a sound source in the horizontal plane. Sound source locations are expressed in spherical rather than in cartesian coordinates. The spherical coordinate system is governed by three parameters: azimuth, elevation and distance as illustrated in Figure 2-3. All locations are referenced to the centre of the head. The azimuthal angle is defined as the position from the centre (0°) of the head covering the range from -180° to 180° in the horizontal plane. Positive angle values start from the right-hand side. The elevation angle is defined as the position above and below the head, covering the range from –90° to 90°. Positive angle values start from above the head. The distance is defined in meters originating from the centre of the head.

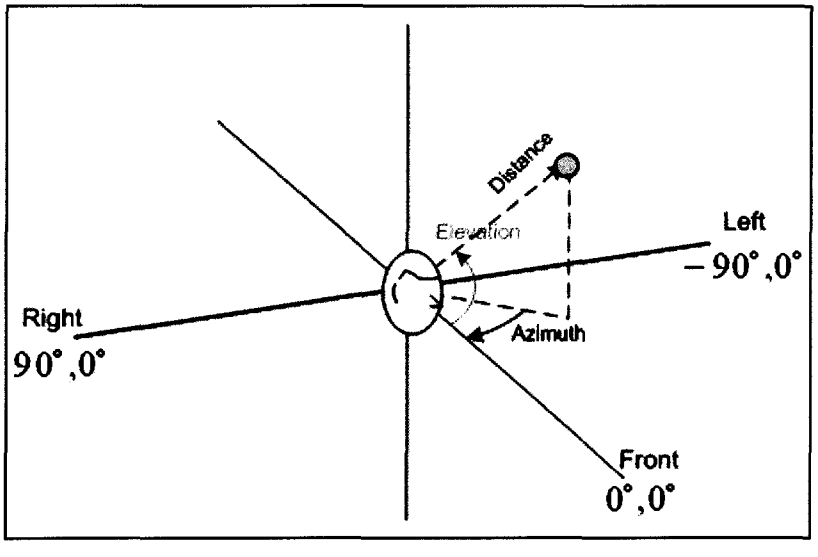

Figure 2-3: Diagram of the spherical coordinate system

2.2.1 Lateral Sound Source Perception (Azimuth)

About a century ago, Lord Rayleigh developed a so-called theory referred to as "duplex theory". The duplex theory is one of the first models for explaining sound source localization. It assigns our localization abilities in the horizontal plane to two primary cues: Interaural Time Differences (ITD) and Interaural Level Differences (ILD). For instance, when a sound is in front of the listener, the distance to travel to each ear is equal. Therefore, the sound reaches both ears simultaneously. However, when the source is coming from the right side as an example, the sound will obviously reach the right ear before the left one since it has to travel a longer path. ITD is defined as the difference between the times that the sound reaches the left and right ear, whereas ILD is the difference in intensity (or level) of the sound received at the two ears. Using the same right side source example, ILD occurs because some of the incident sound waves are diffracted or obstructed by the head in the path before reaching the left ear, causing significant intensity reduction at the left ear. This situation is known as the head-shadow effect. Moreover, Lord Rayleigh recognized that ITD and ILD are in fact frequency dependent and complementary. His duplex theory explains that localization can be found on the basis of time differences between the ears for the lower frequencies and level differences between the ears for the higher frequencies [GEL'04]. For the lower frequencies, the wavelength of the sound is greater than the diameter of the head, therefore not much head shadowing or diffraction will occur (i.e. hardly any difference in signal attenuation). Consequently, at low frequencies, ILD will not provide much information regarding the sound source location. However, it turns out that the ITD will be much more significant at lower frequencies, since with a wavelength greater than the diameter of the head, the phase difference between the received signals at both ears can provide a unique ITD. ITD corresponding to only a fraction of a cycle (i.e. signal period) can be easily detectable. For the higher frequencies, the wavelength of the sound wave is smaller than the diameter of the head, leading to a much stronger attenuation (i.e. higher frequencies are blocked by the head in the path traveling to the far left ear). Therefore, ILD will tend to be quite high and it will be much more relevant as a localization cue. On the other hand, there will be an ambiguity using the ITD, since ITD will result in several

delay cycles due to the short wavelength at high frequencies. Studies have shown that ITD becomes relevant only below 1500Hz, providing an unambiguous localization cue [GEL'04][BLA'97]. Generally, above that frequency the acoustic shadow produced by the head becomes effective and thus ILD becomes far more reliable.

Now, the obvious question that can be asked is the following: how do we differentiate between a sound source coming directly from the front versus a sound source directly from behind since both cases will theoretically yield identical ITDs and ILDs? In fact, ITD and ILD from the duplex theory can only partially explain our ability to distinguish among different spatial directions, since different spatial locations can produce identical ITDs and ILDs.

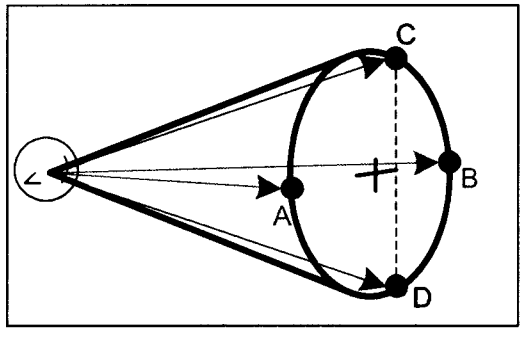

Figure 2-4: Cone of confusion, different spatial directions producing identical ITDs and ILDs

According to the duplex theory, for a spherical model of the head, identical values of ITD and ILD can be evaluated at two opposite points anywhere on the surface of a cone formed by a circle with the center of the head as depicted in Fig. 2-4. This dilemma is referred to as the "cone of confusion" [BEG'94]. It can be seen in Fig. 2-4 that the positions A and B illustrate the front-back ambiguity and positions C and D illustrate the elevation ambiguity. Fortunately, the cone of confusion is resolved by the following:
Research has shown that the shape of the head, torso and primarily the outer ear (pinna) affect the spectral contents of a sound originating from the free space to the inner ear. As introduced in the previous section, the pinna acts as a filter on the sound source spectrum

from a given direction. Most importantly, due to its irregular shape, this filtering is spatially dependent i.e. varies with the orientation of the sound source relative to the ear. This filtering effect is referred to as "Head-Related Transfer Function (HRTF)". The HRTF varies according to the elevation and azimuth of the sound source, and consequently, there is a unique HRTF (one for each ear) for each sound source orientation. For given sound source, the left and right HRTFs would contain the ILD and ITD cues, and the incorporation of spectral cues caused mostly by the irregular shape of the pinna. Since the magnitude and phase responses of this filter are direction-dependent, it is hence possible for the listener to distinguish between different source locations producing identical ITDs and ILDs and therefore resolving the "cone of confusion" dilemma.

In practice, a common technique to empirically measure the left and right ear HRTFs at different spatial locations is to insert a microphone into a subject's ears deep inside the ear canal [CHE'99]. A known stimulus such as a burst of wideband noise is then played through a loudspeaker oriented at a specific azimuth and elevation from the subject's head. For each location, from the known characteristics of the signal sent and the received signal recorded by the microphone inside each ear, the left and right transfer functions (or impulse responses) can be computed. Basically, the measurements will provide the direction-dependent acoustic filtering (i.e. HRTF) that a sound source undergoes due to the head, torso and pinna [CHE'99]. It should be noted that the effect of the microphone and the loudspeaker transfer functions should be removed from the raw measurements. Also, all HRTFs computations are performed in an anechoic chamber in order to avoid any type of reverberation.

In 1994, an extensive set of HRTF measurements have been done by a research group at the Massachusetts Institute of Technology (MIT) using a KEMAR (Knowles Electronic Manikin for Acoustic Research) placed in an anechoic chamber. The KEMAR is an anthropomorphic manikin whose dimensions match those of a median human. The set of measurements for 710 different positions in terms of azimuth and elevation can be found in [WEB'4] along with the complete environment setup description.

HRTF measurements are difficult and time consuming, and also the measurements might differ between different individuals due to various shapes of the head, ears, pinnaes etc. Often a general set of HRTFs not tailored to a specific individual is used in applications such as in virtual reality games and 3-D sound modeling, which will be briefly addressed in Section 2.2.4. However, it has been reported that the use of non-individualized HRTFs causes an increased perception error in elevation and front-back localization, but only a minor increase of perception error for lateral localization (i.e. azimuthal angle) [WEN'93].

2.2.2 Sound Elevation Perception

From the previous section, the primary cues for lateral localization of a sound source (i.e. azimuthal angle of the originating sound source) are binaural. In contrast, the cues for the elevation perception of a sound source are for the most part monaural [MID'91]. Relating to auditory localization, monaural hearing is limited to pinna cues since binaural cues such as ILDs and ITDs are obviously not feasible in this case. Therefore, sound source elevation detection relies above all on spectral cues (i.e. also referred to as pinna cues or monaural cues). Sound waves that reach the irregular-shaped pinna are reflected and diffracted in a complex geometrical way, which varies according to the angle of incidence of the sound source and its original frequency content.

Those reflections can result in destructive interference, which appear as a notch in the HRTF. A simple illustration of two waves originating from a frontal sound source where one wave undergoes a short direct path towards the ear and the other one a longer path due to a reflection from the pinna is shown in Figure 2-5. The delayed signal, that is the wave traveling a longer path due to reflection, can interfere in a destructive manner if it is out of phase with the direct signal. The notch is at its deepest level (i.e. largest attenuation) for sources located in the front rather than for sources above the head.

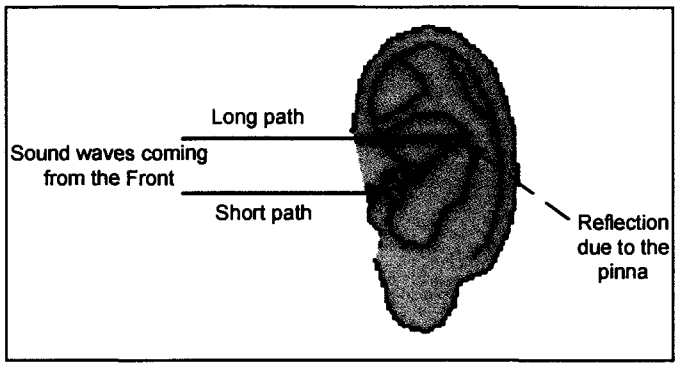

Figure 2-5: Illustration of two signal waves from a frontal sound source reaching the ear and a signal reflection caused by the pinna

In general, variations in elevation cause changes in the high-frequency spectrum of the HRTF. It can be seen that the notch moves between 5 and 11 kHz as the sound source elevates above and around the head [GEL'04]. Starting from below the head, there is a notch around 6 kHz and the notch shifts in higher frequency as the sound source moves upward toward the front. When the sound source is above the head (on top), the notch is essentially not present. As the source moves downward towards the back of the head and then below again, the notch shifts lower in frequency [GEL'04]. Thus, these acoustical modifications of the original sounds spectrum caused by the pinna provide the necessary monaural information to determine where the sound is coming from.

2.2.3 Distance Perception

The cues regarding the azimuth estimation are easily explainable, the cues for elevation are comparatively less understood, but the cues leading to our perception of distance from the originating sound source are still under investigation. The distance cues used by the auditory system include: the sound intensity at the listener's location relative to the originating source (since the intensity loss under ideal conditions is a function of the distance obeying an inverse square law), and the ratio of direct to reverberant sound

energy, which decreases as the source distance increases [ZAH'02]. Also, the resulting spectral shape can be considered as a potential cue. With relatively long distances, the frequency content of a sound changes with distance from the sound source due to the air absorption. As a result, the higher frequencies are greater attenuated compared to the lower ones. It has been noticed from past research that ILDs for frequencies below 3 kHz contribute to the distance perception when the sound source is within one meter from the head. Moreover, familiarity or experience with the original sound source and the environment has some influence on our distance perception. For instance, prior knowledge (or familiarity) of a sound such as speech would allow the listener to make distance judgements based on the modified sound level caused by the change of distance [GEL'04].

2.2.4 Auralization

The research on understanding how the sound is localized in various environments and the benefit of binaural hearing has received an increased attention over the years with the development of highly performing computers and the major interest in the creation of virtual environments. Therefore, virtual environments make use of auralization systems where sophisticated signal processing techniques and binaural hearing concepts are strongly applied. Those types of systems, commonly referred to as virtual reality systems, virtual room synthesis, or 3-D sound modeling, are defined as:

" *the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space*" [KLE'93].

In other words, an auralization system takes the original sound source and alters its frequency content according to:
a) the modeled space/environment where the sound source is propagating. For instance, the system should replicate and combine acoustical interactions such as reverberation,

distance cues, air absorption and object occlusion in order to create a complete simulation of the acoustic scene such as inside a theatre, an airplane, the natural world, etc.

b) the chosen position/location of the sound source and of the listener

c) the shape and characteristic of head/ears/torso of the listener

The outcome of such a system should provide the listener a binaural signal (i.e. one for each ear), played for example through stereo headphones, which attempts to create an aural illusion of the sound originating from the specified location in the chosen environment. A simple example of binaural synthesis is illustrated in Figure 2-6, where a sound signal is processed by HRTFs (i.e. the left and right input sound signals are convolved with the impulse responses corresponding to the left and right HRTF respectively) and the binaural output is listened to over headphones, leading to the sound localization cues for each ear to be reproduced. Therefore, the listener should perceive the sound at the location specified by the HRTFs.

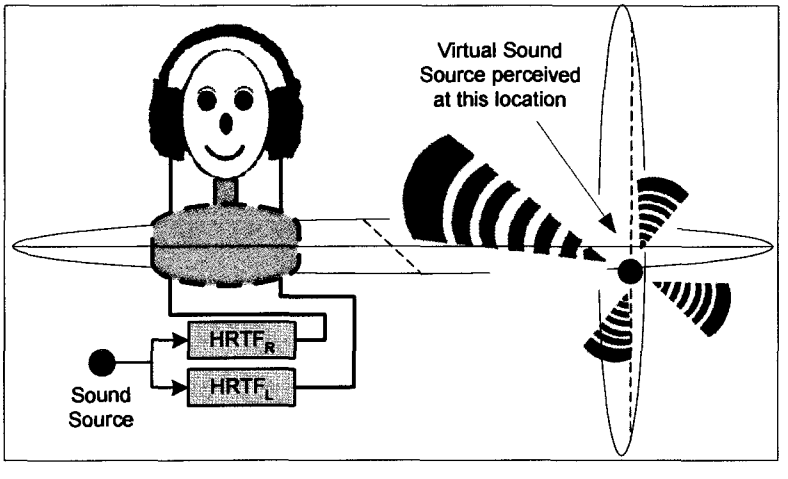

Figure 2-6: Simple binaural synthesis

2.2.5 Binaural Speech Intelligibility

From the previous section, one of the benefits of binaural hearing is to provide the listener with information about sound source location using spatial auditory cues. In addition, research has shown that those cues also help to improve speech intelligibility. For instance, in many social environments, an individual hears simultaneous sounds from many sources. The listener has the ability to focus on a single target voice among a mixture of competing talkers and background noises, disregarding other conversions such as in crowded restaurants. This phenomenon has been referred to as "cocktail party effect". Extensive studies have been conducted to explain this phenomenon, which involves both monaural and binaural processes [DEV'03]. One of the key factors has been referred to as spatial unmasking. Spatial unmasking is defined as the improvement in detection or identification of the masked source (i.e. target source) that arises when the target source and the interfering maskers (i.e. competing sources) are at different spatial locations, relative to when all sources are spatially coincident [DEV'03]. Spatial unmasking is caused by both monaural energetic effects and binaural processing. The monaural energetic effects arise when for instance the target and the masker sources are spatially collocated, the Target-To-Masker energy ratio (TMR) will then be approximately equal at the two ears. If the target or masker is relocated, the TMR will then increase at one ear referred to as the "better ear" due to head-shadowing and decreases at the other ear, leading to an improvement in performance due to monaural spatial unmasking on the "better ear". Moreover, further spatial unmasking due to binaural cues (i.e. in addition to the changes in the "better ear" TMR from head-shadow) occurs when the target and masker exhibit different ITDs and ILDs [DEV'03][SHI'02]. An improvement of detection especially arises for the low frequency part of the target signal, mostly helped by the differences in ITDs between the target sound and competing sources [HAW'04]. In [HAW'94], the authors investigated the benefit of binaural hearing in a cocktail party for various locations of the maskers/interferers, the number of maskers and also the type of interferers such as speech-shape noise, regular speech and speech with linguistic content similar with the content of the target voice, which creates the so-called informational masking. Their results for those various scenarios were evaluated in

terms of binaural advantage. The binaural advantage was obtained by subtracting the speech reception thresholds (SRTs) when listening with the "better-ear" from those of binaural hearing (i.e. corresponding to the overall advantage). Their work suggested that the listener's ability to perceive and understand the target speech in the complex noisy situations not only relies on the type, number and position of interfering sounds, but also on interactions between these factors. Detailed results can be found in their studies, however they are beyond the scope of this thesis.

Chapter 3. BACKGROUND ON NOISE PSD ESTIMATION AND MONAURAL NOISE REDUCTION TECHNIQUES

The enhancement of speech corrupted by noise is a challenging research area and is still considered as an emerging field. Over the years, there are a large number of noise reduction techniques that have been proposed such as spectral subtraction schemes, Wiener filtering, Kalman filter-based methods, etc.. They have been intended for instance to be used in numerous speech communication applications such as mobile telephony, speech recognition tools, vehicle hands-free voice systems, restoration of degraded audio recordings, and plenty more.

A very significant requirement of several noise reduction schemes is the estimation of the noise power spectrum density prior to any type of enhancement procedure. For some techniques, this estimation is simply performed at the beginning of the speech signal, where it is assumed that no speech is present and that the noise is relatively stationary. The noise statistics can then be easily computed over an average of a few initial frames. However, most realistic noisy environments are characterized by non-stationary noise and it is very crucial to often update the noise power spectrum density to maintain the effectiveness of the noise reduction scheme. More sophisticated algorithms employ a voice activity detection (VAD) scheme to update the noise power spectrum density during speech pauses. However, the detection mechanism of VADs degrades in highly noisy environments (i.e. false speech or noise detection at low SNRs), leading to false estimation (or poor estimation) of the noise characteristics. Also, the noise estimation is even more degraded under non-stationary noise conditions, since VAD-based noise estimation techniques cannot update the noise estimate during speech activity, resulting in poor enhancement or even additional target speech distortions. Section 3.1 will review two advanced noise power spectrum estimation schemes where a VAD is not required. Section 3.2 will then provide an overview of a few well-known monaural noise reduction techniques.

3.1 Overview of Advanced Noise PSD Estimation Techniques

Two advanced noise power spectrum estimation techniques will be described, which do not require the use of VADs. Since VADs usually fail at low SNR [MAR'01], both techniques do not employ any explicit thresholds to differentiate between speech activity and speech pause. Also, any shape of colored noise can be considered, and noise stationarity is not essentially required. The first technique is based on minimum statistics and uses a single channel system. On the other hand, the second technique based on a cross-power spectrum density method requires a two-channel system.

3.1.1 Minimum Statistics Method

The author in [MAR'01] proposed a new approach to estimate the noise power density from a noisy speech signal based on minimum statistics. The technique relies on two main observations: at first, the speech and the corrupting noise are usually considered statistically uncorrelated, and secondly, the power of the noisy speech signal often decays to the power spectrum level of the corrupting noise. It has been suggested that based on those two observations, it is possible to derive an accurate noise power spectral density estimate by tracking the minimum of the noisy speech signal PSD over the past several frames for each frequency bin, and then by applying a bias compensation to adjust/regulate the power level. The main components of the algorithm can be summarized as follows:

Let $y(i)$ be the received noisy speech signal from a single microphone (a single channel environment) where $s(i)$ is the clean speech and $n(i)$ is the additive uncorrelated noise signal as in (3-1). i is the sampling time index.

$$y(i) = s(i) + n(i) \qquad \text{(EQ 3-1)}$$

The short-time DFT of the noisy signal is defined as [MAR'94]:

$$Y(\lambda,k)=\sum_{i=0}^{W_{DFT}-1} y(\lambda R+i)\cdot w(i)\cdot e^{(-j\frac{2\pi ik}{W_{DFT}})} \qquad \text{(EQ 3-2)}$$

where the window shift is of *R samples,* $w(i)$ is a weighting window for a frame size of L samples, k is the discrete sub-band frequency with W_{DFT} subbands (also $L= W_{\mathrm{DFT}}$) and λ is the frame index.

The short time noisy signal power density, $P_Y(\lambda,k)$, using recursively-smoothed periodograms is then computed as:

$$P_Y(\lambda,k)=\alpha(\lambda,k)\cdot P_Y(\lambda-1,k)+\left(1-\alpha(\lambda,k)\right)\cdot\left|Y(\lambda,k)\right|^2 \qquad \text{(EQ 3-3)}$$

where $\alpha(\lambda,k)$ is a variable smoothing parameter.

As briefly explained earlier, this noise power spectrum estimate is based on the observation that even during speech activity, a short-term power spectral density estimate of the noisy signal often decays to the noise power level values. The reasoning is that during speech pauses or within brief periods in between words and syllables, the speech energy is close or equal to zero [MAR'01]. Therefore, the noise power estimate, $P_N(\lambda,k)$ in Equation (3-4), is then obtained by taking a weighted minimum of subsequent noisy signal short-time power spectrum density estimates of P_Y, within a finite window of D frames. The latter window should be large enough to envelop high-energy speech segments.

$$P_N(\lambda,k)=Bias(\lambda,k)\cdot P_{MIN}(\lambda,k) \qquad \text{(EQ 3-4)}$$

where $P_{MIN}(\lambda,k)=\min\left\{P_Y(\lambda-D,k),P_Y(\lambda-D+1,k),P_Y(\lambda-D+2,k),\cdots,P_Y(\lambda,k)\right\}$ is the estimated minimum noise PSD and $Bias(\lambda,k)$ is a bias compensation factor regulating (or weighting) the minimum noise power estimate.

The major aspects for the accuracy of the noise PSD estimation in [MAR'01] rely on the computations of the time and frequency varying smoothing factor $\alpha(\lambda,k)$ in (3-3), and of

the bias compensation factor $Bias(\lambda,k)$ in (3-4). For instance, a fixed smoothing factor could lead to inaccurate noise estimates as the search window for the minimum could fall into broad speech peaks or a high smoothing factor value (i.e. close to one) could lead to a noise estimate with fairly large fluctuations. Also, the bias compensation is present to regulate the noise tracking estimation so that it is not pushed towards low values and, in case of increasing noise power, it does not lag behind. Finding the minimum of D subsequent PSD estimates of P_Y will cause a computational complexity and intrinsic delay, depending on how frequently the estimated minimum is being updated.

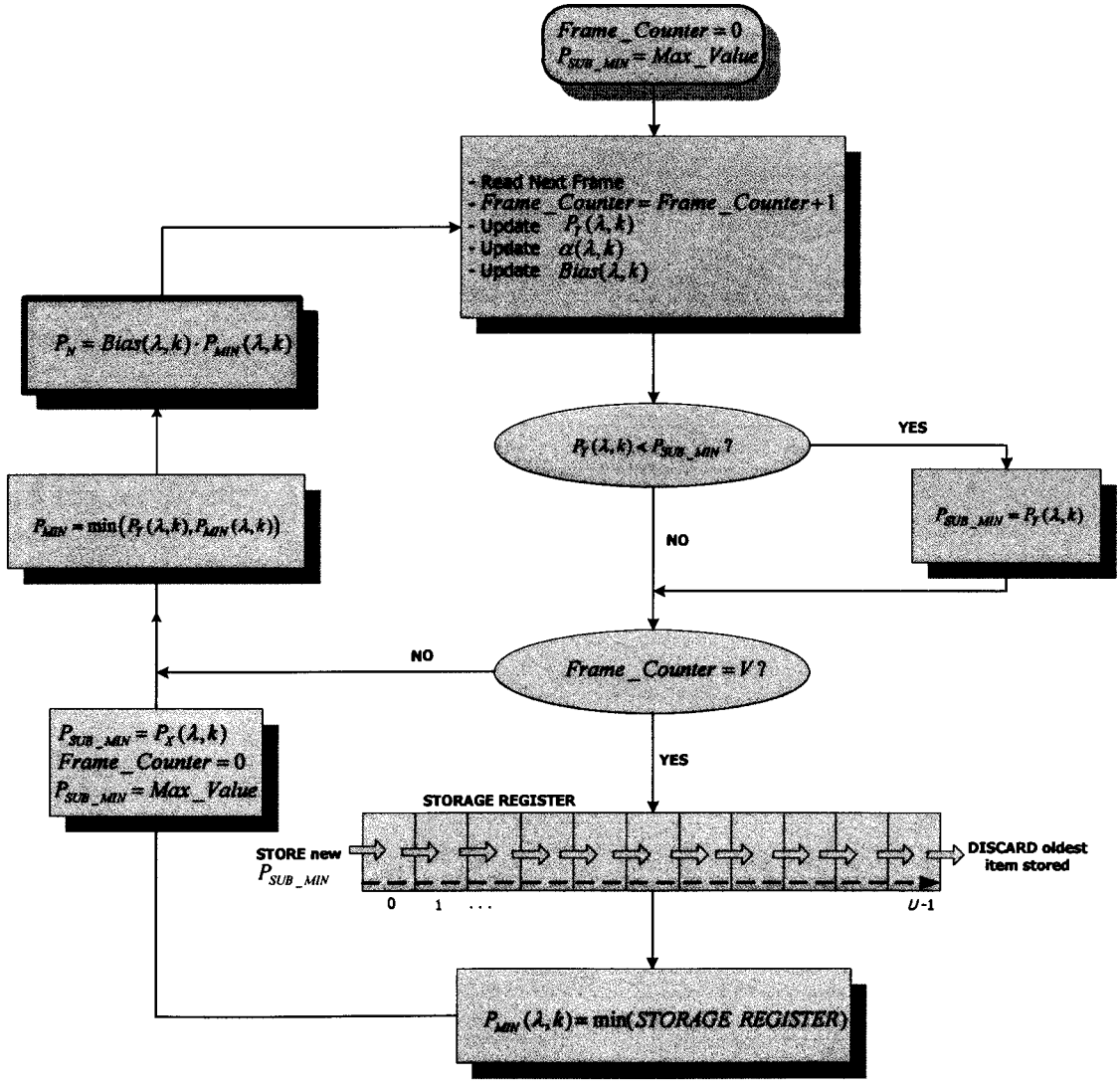

Figure 3-1: Partial block diagram of the noise power spectral density estimation via Minimum Statistics Approach

In [MAR'94][MAR'01], the implementation is performed via a tree search to balance the complexity and the update rate, where a global window of D frames is decomposed into U sub-windows of V frames (i.e. $D = U \cdot V$). Figure 3-1 illustrates a partial block diagram featuring the main components of the noise PSD estimation based on minimum statistics. At first, the minimum power of the last V frames is found by comparing P_{SUB_MIN} (initially set to a maximum value i.e. *Max_Value*) with the current $P_Y(\lambda,k)$ on a frame-by-frame basis. Then, whenever V frames have been processed, this minimum power of the last V frames i.e. P_{SUB_MIN} is stored, and P_{SUB_MIN} is again set to *Max_value*. Finally, the minimum power of the global window of D frames , $P_{MIN}(\lambda,k)$, is then obtained by just taking the minimum of the last U stored P_{SUB_MIN} minima values. Also, if the current sub-band power, $P_Y(\lambda,k)$, is lower than the minimum estimated noise power, $P_{MIN}(\lambda,k)$, then $P_{MIN}(\lambda,k)$ immediately takes the value of $P_Y(\lambda,k)$. It should be noted (in Fig. 3-1) that all the computations are embedded into loops over all frequencies and all time indices. Further details can be found in [MAR'01] regarding the explicit computations of the time and frequency varying smoothing and bias compensation factors. Also, an additional bias compensation is present in the algorithm for the shorter sub-windows, improving the tracking performance in case of non-stationary noise.

3.1.2 Cross-Power Spectral Density Method

In contrast to the single channel system explained in the previous section, where only one microphone is used for the noise PSD estimation, the method suggested in [DOE'96] makes use of a secondary microphone i.e. a dual-microphone system, which can be aimed for a potential use in binaural hearing aids. Their noise power spectral density estimation is based on the combination of the auto- and the cross-power spectral densities of the received noisy binaural signals, to extract the power spectral density of the noise under a diffuse noise field. However, the primary assumption taken is that the target speaker should be in front of the hearing aid user, and thus, it assumes that the speech

components in the left and right signals received from each microphone have followed equal attenuation and delay paths.

3.1.2.1 Two-Microphone System Description

For the derivation of the dual-channel noise power spectrum density estimator, the authors in [DOE'96] suppose at first that the speaker is close enough to the two microphones. Consequently, the short distance between the speaker and the microphones is such that the microphones receive essentially speech through a direct path from the nearby speaker. Therefore, the speech signals received by the two microphones are highly correlated, corresponding to a Magnitude-Squared Coherence (MSC) between the microphones close to one [DOE'96]. They model this acoustic situation by $H_L(\omega)$, $H_R(\omega)$, the left and right frequency responses between the target speaker and the left and right microphones. Secondly, they specify that the speaker is in a diffuse noise field environment. In a diffuse noise field environment, the left and right received noise signals are poorly correlated over the entire frequency spectrum (except at low frequencies). Section 6.1 of this thesis discusses the diffuse noise field characteristics in more details.

Neglecting the fact that at low frequencies a diffuse noise field has a high coherence by definition [MCC'03], the authors assume a very low magnitude-squared coherence (i.e. close to zero) for the noise over most of the entire frequency band. As a result, the noise picked up by the microphones can still be modeled by two additive, uncorrelated noise sources with $N_L(\omega)$ and $N_R(\omega)$ corresponding to the left and right short-time complex Fourier spectra of the noise, with a cross-power spectral density of: $\Gamma_{N_L N_R}(\omega) \approx 0$ (except at low frequencies). Their system model, in the spectral domain, is hence given by:

$$Y_L(\omega) = S(\omega)H_L(\omega) + N_L(\omega) \qquad \text{(EQ 3-5)}$$

$$Y_R(\omega) = S(\omega)H_R(\omega) + N_R(\omega) \qquad \text{(EQ 3-6)}$$

where $Y_L(\omega)$, $Y_R(\omega)$ are the short-time complex Fourier spectrum of the noisy signals received at the left and right microphones, and $S(\omega)$ denotes the short-time complex Fourier spectrum of the target source speech signal.

Furthermore, the authors have made two additional assumptions:

- The speech components in the left and right signals have followed almost equal attenuation and delay paths, which is equivalent to have a target speaker at equal distance to both microphones, and located in front (or behind) the listener. Hence:

$$H_L(\omega) \approx H_R(\omega) = H(\omega) \qquad \text{(EQ 3-7)}$$

- Also, due to the diffuse sound field characteristic, an assumption holds for the left and right noise power spectral densities:

$$\Gamma_{N_L N_L}(\omega) \approx \Gamma_{N_R N_R}(\omega) = \Gamma_{NN} \qquad \text{(EQ 3-8)}$$

3.1.2.2 Noise PSD Estimation via Cross-Power Spectral Density Method

The following is the reproduction of the work from [DOE'96] to obtain their noise power spectral density estimator. From the acoustical environment described in Section 3.1.2.1, the magnitude cross-power spectral density using the left and right received signals is equal to:

$$\begin{aligned}
|\Gamma_{LR}(\omega)| &= \left| E\left\{ Y_L(\omega) \cdot Y_R^*(\omega) \right\} \right| \\
&= \left| E\left\{ \left(S(\omega) \cdot H_L(\omega) + N_L(\omega) \right) \cdot \left(S^*(\omega) \cdot H_R^*(\omega) + N_R^*(\omega) \right) \right\} \right| \\
&= \left| \begin{array}{l} E\left\{ |S(\omega)|^2 \right\} \cdot H_L(\omega) H_R^*(\omega) + E\left\{ S(\omega) \cdot N_R^*(\omega) \right\} \cdot H_L(\omega) \\ + E\left\{ S^*(\omega) \cdot N_L(\omega) \right\} \cdot H_R^*(\omega) + E\left\{ N_L(\omega) \cdot N_R^*(\omega) \right\} \end{array} \right|
\end{aligned} \qquad \text{(EQ 3-9)}$$

Since the left and right noise signals are uncorrelated (neglecting the low frequency correlation as mentioned in the previous section) and the target speech signal and the noise signals are mutually uncorrelated as well, Equation (3-9) can be simplified to the following:

$$|\Gamma_{LR}(\omega)| = \left|E\left\{|S(\omega)|^2\right\} \cdot H_L(\omega)H_R^*(\omega)\right| \\ = \Gamma_{SS}(\omega) \cdot |H(\omega)|^2 \qquad \text{(EQ 3-10)}$$

where $\Gamma_{SS}(\omega) = E\left\{|S(\omega)|^2\right\}$

Similarly, the product of the left and right power spectral densities of the microphone signals is given by:

$$\Gamma_{LL}(\omega) \cdot \Gamma_{RR}(\omega) = E\left\{|Y_L(\omega)|^2\right\} \cdot E\left\{|Y_R(\omega)|^2\right\}$$

$$= E\left\{|N_L(\omega)|^2\right\} \cdot E\left\{|N_R(\omega)|^2\right\} + \left(E\left\{|S(\omega)|^2\right\}\right)^2 \cdot |H_L(\omega)|^2 \cdot |H_R(\omega)|^2$$

$$+E\left\{|S(\omega)|^2\right\} \cdot \left(|H_L(\omega)|^2 \cdot E\left\{|N_R(\omega)|^2\right\} + |H_R(\omega)|^2 \cdot E\left\{|N_L(\omega)|^2\right\}\right)$$

$$= \left(E\left\{|N(\omega)|^2\right\} + E\left\{|S(\omega)|^2\right\} \cdot |H(\omega)|^2\right)^2$$

$$= \left(\Gamma_{NN}(\omega) + \Gamma_{SS}(\omega) \cdot |H(\omega)|^2\right)^2 \qquad \text{(EQ 3-11)}$$

where $\Gamma_{NN}(\omega) \triangleq E\left\{|N(\omega)|^2\right\}$

Finally, substituting Equation (3-10) into (3-11) and by taking the square root of the result, the noise power spectral density can be extracted and expressed as:

$$\Gamma_{NN}(\omega) = \sqrt{\Gamma_{LL}(\omega) \cdot \Gamma_{RR}(\omega)} - |\Gamma_{LR}(\omega)| \qquad \text{(EQ 3-12)}$$

3.2 Overview of Monaural Noise Reduction Techniques

As introduced earlier, a variety of speech noise reduction or speech enhancement techniques have been proposed in the literature over the years. Typical speech enhancement algorithms are based on short-time spectral magnitude estimation, such as spectral subtraction, Wiener filtering, the Ephraim-Malah algorithm, etc.. One of the reasons for this is that the noise found in the phase component of the degraded speech signal is perceptually less important in contrast to the noise affecting the speech magnitude [SHA'06]. Therefore, numerous techniques have tried to find the optimal

estimate of the magnitude of the input speech signal and pass the unaltered noisy phase component directly to the output.

One of the main drawbacks of speech enhancement is the trade-off between noise reduction and speech distortion. A higher noise suppression tends to produce a higher speech distortion and vice-versa. The most common artefact of short-time spectral-based noise reduction schemes is musical noise. Musical noise is often due to poor estimation of the noise power spectrum used in the enhancement scheme. It can be defined as residual sinusoidal components with random frequencies that vary in each frame of the enhanced speech spectrum. The term "musical" is related to the presence of pure tones in the residual noise after enhancement [CAP'94]. To reduce this undesirable phenomenon, some techniques may apply an overestimation of the noise spectrum in their noise suppression rules, to attenuate much more than necessary the short-time noisy spectrum. But then again a higher distortion of the speech signal will occur, consequently deteriorating the speech intelligibility. Some advanced techniques may use a perceptual approach in their noise suppression rule to reduce the prominence of the musical noise. This concept is based on the fact that the human auditory system is able to tolerate additive noise only if it is lower than a masking threshold [VIR'99][MA'06]. Therefore, the noise reduction gain can be less aggressive, but sufficient enough to bring the noise level below this threshold. As a result, the enhanced speech signal can be less distorted.

Furthermore, speech enhancement techniques can be performed by employing one or multiple microphones. Usually, the greater the number, the better is the enhancement, since more information is gathered and more features can be extracted and analysed. Of course multiple microphone systems tend to be costly, or even sometimes the implementation is not feasible on some devices due to size constraints.

The following sections will overview a few monaural (i.e. single channel) noise reduction techniques based on spectral estimation. However, the last one is a model-based approach belonging to another category of speech enhancement scheme. The latter does not suffer

from musical noise but the enhanced speech can be somewhat muffled. Also, all the techniques chosen are suitable for speech contaminated with colored noise.

3.2.1 Spectral Subtraction

Speech enhancement using spectral subtraction is one of the most traditional and popular techniques due to its simplicity. As a first step, similar to most noise reduction techniques, it is assumed that the noisy speech signal, $y(i)$, is composed of the target speech signal, $s(i)$, corrupted by an additive uncorrelated noise signal, $n(i)$, as the following:

$$y(i) = s(i) + n(i) \qquad \text{(EQ 3-13)}$$

Secondly, as the name of the technique is quite self-explanatory, the enhancement is performed (i.e. recovering an estimate of the target signal, $s(i)$) by subtracting the estimated power spectrum of the noise from the noisy input power spectrum signal. Often, this operation is performed with short-time analysis. Prior to the subtraction, an estimate of the power spectrum of the noise has to be carried out, usually by averaging multiple frames at the beginning of the sequence where no speech is present, or over known noise-only segments, using possibly a VAD to perform frequent updates of the noise power spectrum estimation.

In mathematical terms, using the same notation and signal definitions as in Section 3.1.1, this technique can be represented as follows:

Let $Y(\lambda,k)$ be the corresponding short-time spectrum of the input noisy speech signal and $|N(\lambda,k)|^2$ be the short-time squared estimated spectral magnitude of the noise. Then, the estimated short-time squared spectral magnitude of the desired speech can be found as the following:

$$|\tilde{S}(\lambda,k)|^2=\begin{cases}|Y(\lambda,k)|^2-|N(\lambda,k)|^2, \text{if } |Y(\lambda,k)|^2\geq|N(\lambda,k)|^2\\ 0, \quad \text{otherwise}\end{cases} \qquad \text{(EQ 3-14)}$$

After finding the magnitude estimate, the last step is to combine the magnitude estimate with the actual unprocessed input phase (i.e. directly from the noisy input signal). It is assumed in this method that the noise does not affect significantly the phase of the target speech signal. The resulting enhanced short-time spectrum estimate of the target speech signal can then be evaluated as:

$$\tilde{S}(\lambda,k)=\left|\tilde{S}(\lambda,k)\right|\cdot e^{j\angle Y(\lambda,k)} \qquad \text{(EQ 3-15)}$$

Although spectral subtraction is a very simple technique and can effectively suppress unwanted additive noise, as previously mentioned it tends to introduce a specific disturbance commonly known as musical noise that can be very annoying or unacceptable depending on the application. Nevertheless, in the literature, many methods have been investigated regarding how the subtraction should be applied. For instance, a more general form of Equation (3-14) can be expressed as follows:

$$|\tilde{S}(\lambda,k)|=\begin{cases}\left(|Y(\lambda,k)|^\alpha-\beta|N(\lambda,k)|^\alpha\right)^{\frac{1}{\alpha}}, \text{if } |Y(\lambda,k)|^\alpha\geq\beta|N(\lambda,k)|^\alpha\\ 0, \quad \text{otherwise}\end{cases} \qquad \text{(EQ 3-16)}$$

where the parameter α has some impact on speech intelligibility, whereas β controls the degree of noise reduction [GOH'98]. Unfortunately, single channel enhancement systems are restrained by a tradeoff between noise reduction and speech distortion. For example, a large β value would substantially attenuate the corrupting noise and suppress the musical noise generated in the spectral subtraction procedure. However, the major drawback is that it can also attenuate the target speech spectral contents (i.e. causing a larger amount of speech distortion), therefore reducing speech intelligibility. Further research has been done for example by incorporating perceptual properties of the human hearing in the technique, in order to find the parameter values that will provide the best

tradeoff between noise suppression and speech distortion in a perceptual sense [VIR'99]. The bottom line is that the choice of the parameter values is very subjective and varies depending on the environment or the application. Also, one of the key factors closely associated to the performance of this technique is to find an accurate estimation of the noise power spectrum before any subtractive rule.

3.2.2 Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator

The following noise reduction technique proposed by Ephraim and Malah is derived from a Minimum Mean-Square Error Short-Time Spectral Amplitude estimator (MMSE-STSA), where the speech and noise spectral components are modelled as statistically independent Gaussian random variables [EPH'84]. Their algorithm has raised the interest of many researchers in the fields of mobile hands-free communications, restoration of degraded audio recordings [SCH'02] [SCA'96] [CAP'94], etc.. This is due to the fact that, in contrast to spectral subtraction techniques, one of the main advantages of the noise suppression proposed by Ephraim and Malah is that it exhibits minor residual musical noise artefacts after enhancement [MAR'03][CAP'94]. This is quite important since an enhanced signal (i.e. improved SNR in the context here) containing musical noise artefacts can be very tiresome and disturbing, especially for hearing aid users.

The main equations of their technique are summarized below with some explanations.
At first, the enhanced signal, $Z(\lambda,k)$, in the frequency domain is given by:

$$Z(\lambda,k) = G(\lambda,k) \cdot Y(\lambda,k) \qquad \text{(EQ 3-17)}$$

where the MMSE STSA estimator is expressed as a spectral gain, $G(\lambda,k)$, applied to each short-time spectrum noisy input component, $Y(\lambda,k)$.

This MMSE-STSA spectral gain is computed over each time frame index and each discrete frequency sub-band as [CAP'94]:

$$G(\lambda,k)=\frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+\xi(\lambda,k)}\right)\cdot\left(\frac{\gamma(\lambda,k)}{1+\gamma(\lambda,k)}\right)}\cdot M\left[\left(1+\xi(\lambda,k)\right)\cdot\left(\frac{\gamma(\lambda,k)}{1+\gamma(\lambda,k)}\right)\right] \quad \text{(EQ 3-18)}$$

where $M[\vartheta]$ is function evaluated as:

$$M[\vartheta]=e^{\left(\frac{-\vartheta}{2}\right)}\cdot\left[(1+\vartheta)\cdot I_0\cdot\left(\frac{\vartheta}{2}\right)+\vartheta\cdot I_1\cdot\left(\frac{\vartheta}{2}\right)\right] \quad \text{(EQ 3-19)}$$

where $I_0(.)$ and $I_1(.)$ denote the modified Bessel functions of zero and first order respectively.

As it can be noticed in Equation (3-18), the spectral gain is governed by two parameters, $\xi(\lambda,k)$ and $\gamma(\lambda,k)$. They can be interpreted as the *a posterior* SNR and *a priori* SNR, respectively.

The *a posterior* SNR is defined as:

$$\xi(\lambda,k)=\frac{|Y(\lambda,k)|^2}{|N(\lambda,k)|^2}-1 \quad \text{(EQ 3-20)}$$

and the *a priori* SNR is evaluated as:

$$\gamma(\lambda,k)=(1-\sigma)\cdot P[\xi(\lambda,k)]+\sigma\frac{|G(\lambda-1,k)\cdot Y(\lambda-1,k)|^2}{|N(\lambda,k)|^2} \quad \text{(EQ 3-21)}$$

where $P[x]=x$ *if* $x\geq 0$, and $P[x]=0$ otherwise.

The *a posterior* SNR can be interpreted as the instantaneous SNR. Meanwhile, looking at Equation (3-21), it can be seen that the *a priori* SNR is in fact a weighted sum (with weighting parameter σ) between the instantaneous *a posterior* SNR and the SNR, computed using the previous enhanced frame power at time index λ-1. It is commonly argued that this estimation approach contributes to a large extent to the reduction of musical noise in the Ephraim and Malah technique [MAR'03][CAP'94]. This approach is referred to as "the decision direct method" in [EPH'84].

In addition to the above scheme, the authors in [EPH'84] derived a modified MMSE STSA amplitude estimator that takes into account "the uncertainty of signal presence" in the noisy signal observations. This modification was pushed by the fact that the target speech signal is absent in the noisy signal quite frequently (i.e. during speech pauses) or appears with only low energy level in some noisy spectral components, for instance in unvoiced speech. The authors found that taking into account the uncertainty of signal presence in their MMSE-STSA estimator yields a further reduction of the colourless residual noise, with negligible additional distortions in the enhanced speech signal [EPH'84].

The gain of the new STSA estimator under uncertainty of signal presence, $G_{sp}(\lambda,k)$, is computed as:

$$G_{sp}(\lambda,k)=\frac{\Lambda}{1+\Lambda}\cdot G(\lambda,k)\Big|_{\gamma_{sp}(\lambda,k)} \qquad \text{(EQ 3-22)}$$

where the modified *a priori* SNR is defined as:

$$\gamma_{sp}(\lambda,k)=(1-q_k)\cdot\gamma(\lambda,k) \qquad \text{(EQ 3-23)}$$

The parameter, q_k, represents the probability of signal absence in the k^{th} spectral component and Λ is given by:

$$\Lambda=\frac{1-q_k}{q_k}\cdot\frac{1}{1+\gamma_{sp}(\lambda,k)}e^{\left[\frac{\gamma_{sp}(\lambda,k)}{1+\gamma_{sp}(\lambda,k)}\right]\cdot(1+\xi(\lambda,k))} \qquad \text{(EQ 3-24)}$$

3.2.3 Geometric Approach to Spectral Subtraction

In [HU'08a], a new geometric approach to spectral subtraction was proposed, which addresses the shortcomings of the spectral subtraction algorithm such as musical noise distortions (see Section 3.2.1). It also possesses properties similar to the traditional STSA-MMSE algorithm briefly described in Section 3.2.2.

Recalling the spectral subtraction technique as in Section 3.2.1, by taking the short-time Fourier transform of the input noisy signal represented by Equation (3-13), as follows:

$$Y(\lambda,k) = S(\lambda,k) + N(\lambda,k) \quad \text{(EQ 3-25)}$$

The power spectrum of the noisy input signal is then:

$$|Y(\lambda,k)|^2 = Y^*(\lambda,k)\cdot Y(\lambda,k)$$

$$= |S(\lambda,k)|^2 + |N(\lambda,k)|^2 + S(\lambda,k)\cdot N^*(\lambda,k) + S^*(\lambda,k)\cdot N^*(\lambda,k) \quad \text{(EQ 3-26)}$$

Then if we assume that the noise is zero mean and uncorrelated with the clean speech signal, the statistical expectation of the cross terms $E\{S(\lambda,k)\cdot N^*(\lambda,k)\}$ and $E\{S^*(\lambda,k)\cdot N^*(\lambda,k)\}$ in Equation (3-26) can reduce to zero. As a result, the estimated clean speech power spectrum can be estimated as (similar to Equation (3-14) for the general spectral subtraction scheme):

$$|\tilde{S}(\lambda,k)|^2 = |Y(\lambda,k)|^2 - |N(\lambda,k)|^2 \quad \text{(EQ 3-27)}$$

Equation (3-27) can be written in another form as follows:

$$|\tilde{S}(\lambda,k)|^2 = H^2(\lambda,k)\cdot |Y(\lambda,k)|^2 \quad \text{(EQ 3-28)}$$

where $H(\lambda,k)$ is a gain function expressed as:

$$H(\lambda,k) = \sqrt{1 - \frac{|N(\lambda,k)|^2}{|Y(\lambda,k)|^2}} \quad \text{(EQ 3-29)}$$

However, the work in [HU'08a] demonstrated that the assumption that cross terms are negligible (i.e. equal to zero) in Equation (3-26) is not any more valid for frequency regions where the input spectral SNR values (i.e. SNR(k)) are near 0 dB.

As a result, the algorithm proposed in [HU'08a] computes a modified gain function of Equation (3-29), which makes no assumptions about the cross terms being zero. Without going into the derivation details, the new gain function based on a geometric is expressed as follows:

$$H_{GA}(\lambda,k)=\sqrt{1-\frac{\left(\gamma_{GA}(\lambda,k)+1-\xi_{GA}(\lambda,k)\right)^2}{4\cdot\gamma_{GA}(\lambda,k)}\cdot\frac{\left(\gamma_{GA}(\lambda,k)-1-\xi_{GA}(\lambda,k)\right)^2}{4\cdot\xi_{GA}(\lambda,k)}} \qquad \text{(EQ 3-30)}$$

where $\gamma_{GA}(\lambda,k)$ is expressed as:

$$\gamma_{GA}(\lambda,k)=\beta\cdot\gamma_{GA}(\lambda-1,k)+(1-\beta)\cdot\min\left(\frac{|Y(\lambda,k)|^2}{|N(\lambda,k)|^2},20\right) \qquad \text{(EQ 3-31)}$$

and $\xi_{GA}(\lambda,k)$ is expressed as:

$$\xi_{GA}(\lambda,k)=\alpha\cdot\frac{|H_{GA}(\lambda-1,k)\cdot Y(\lambda-1,k)|^2}{|N(\lambda-1,k)|^2}+(1-\alpha)\cdot\left(\sqrt{\frac{|Y(\lambda,k)|^2}{|N(\lambda,k)|^2}}-1\right)^2 \qquad \text{(EQ 3-32)}$$

with α and β being weighting factors.

It can be seen that $\gamma_{GA}(\lambda,k)$ and $\xi_{GA}(\lambda,k)$ are closely related to the *a posterior* and *a priori* SNRs found in the Ephraim and Malah "decision direct method" described in Section 3.2.2 (i.e. STSA-MMSE algorithm). This also explains why the new geometrical gain function $H_{GA}(\lambda,k)$ exhibits negligible musical noise artefacts since it inherits some of the properties of the STSA-MMSE technique.

3.2.4 Kalman Filtering-Based Speech Enhancement

Kalman filtering based methods belong to a different category of speech enhancement algorithms known as model-based. They start from the state-space formulation of a linear dynamical system and offer a recursive solution to linear optimal filtering problems [HAY'01].

Various approaches based on Kalman filtering have been presented in the literature [GAB'05][MA'04][POP'98]. Kalman filtering based methods operate usually in two stages: first, the driving process statistics (i.e. the noise and the speech model parameters)

are estimated, then secondly, the speech estimation is performed by using Kalman filtering. These approaches vary essentially by the choice of the method used to estimate and to update the different model parameters for the speech and the additive noise [GAB'04]. In the literature, most Kalman filtering methods have been proposed for speech contaminated with additive white noise, but only a few are aimed for additive colored noise such as [GAB'05][MA'04]. In this section, the main steps of the Kalman filtering method for colored noise from [GAB'05] will be examined.

Again, we use a noisy speech signal, $y(i)$, composed of the clean speech signal, $s(i)$, corrupted by additive colored noise, $n(i)$. The i^{th} sample of the observation (i.e. noisy speech signal sample) can be expressed as:

$$y(i) = s(i) + n(i) \qquad \text{(EQ 3-33)}$$

For the Kalman procedure, as a first step, both the speech signal and the colored additive noise of Equation (3-33) are individually modeled as two Auto-Regressive (AR) processes with orders p and q respectively:

$$s(i) = \sum_{l=1}^{p} a_l \cdot s(i-l) + u(i) \qquad \text{(EQ 3-34)}$$

$$n(i) = \sum_{k=1}^{q} b_k \cdot n(i-k) + w(i) \qquad \text{(EQ 3-35)}$$

where a_l is the l^{th} AR speech model parameter and b_k is the k^{th} AR noise model parameter. $u(i)$ and $w(i)$ are uncorrelated Gaussian white noise sequences with zeros means and variances σ_u^2 and σ_w^2 respectively. More specifically, $u(i)$ and $w(i)$ are referred to as the model driving noise processes (not to be confused with the colored additive acoustic noise $n(i)$).

Secondly, the state-space model of the AR speech model is obtained by concatenating p consecutive samples of the speech samples denoted by $\mathbf{z}_S(i) = [s(i-p+1), \cdots, s(i)]^T$ and translating the corresponding equations into matrix form, yielding:

$$\mathbf{z}_S(i) = \mathbf{A}_S \cdot \mathbf{z}_S(i-1) + \mathbf{B}_S \cdot u(i) \quad \text{(EQ 3-36)}$$

$$s(i) = \mathbf{C}_S \cdot \mathbf{z}_S(i) \quad \text{(EQ 3-37)}$$

where $\mathbf{A}_S$ is the clean speech transition matrix $\mathbf{A}_S = \begin{bmatrix} 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & 0 \cdots 1 \\ a_p & a_{p-1} & a_{p-2} \cdots a_1 \end{bmatrix}$ (EQ 3-38)

with input vector $\mathbf{B}_S = [0,\cdots,0,1]^T$ and observation vector $\mathbf{C}_S = [0,\cdots,0,1]$.

Similarly, the space-space model for the colored noise is obtained by concatenating q consecutive noise samples denoted by $\mathbf{z}_n(i) = [n(i-q+1),\cdots,n(i)]^T$. The corresponding matrix form is then:

$$\mathbf{z}_n(i) = \mathbf{A}_n \cdot \mathbf{z}_n(i-1) + \mathbf{B}_n \cdot w(i) \quad \text{(EQ 3-39)}$$

$$n(i) = \mathbf{C}_n \cdot \mathbf{z}_n(i) \quad \text{(EQ 3-40)}$$

where $\mathbf{A}_n$ is the noise transition matrix of $\mathbf{A}_n = \begin{bmatrix} 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & 0 \cdots 1 \\ b_q & b_{q-1} & b_{q-2} \cdots b_1 \end{bmatrix}$ (EQ 3-41)

with input vector $\mathbf{B}_n = [0,\cdots,0,1]^T$ and observation vector $\mathbf{C}_n = [0,\cdots,0,1]$.

Therefore, the entire process can be modeled by the following state-space equations:

$$\mathbf{z}_S(i) = \mathbf{A}_S \cdot \mathbf{z}_S(i-1) + \mathbf{B}_S \cdot u(i) \quad \text{(EQ 3-42)}$$

$$\mathbf{z}_n(i) = \mathbf{A}_n \cdot \mathbf{z}_n(i-1) + \mathbf{B}_n \cdot w(i) \quad \text{(EQ 3-43)}$$

$$y(i) = s(i) + n(i) = C_S \cdot \mathbf{z}_S(i) + C_n \cdot \mathbf{z}_n(i) \quad \text{(EQ 3-44)}$$

By combining the states in Equations (3-42) and (3-43), the augmented system will have the following form:

$$\mathbf{z}(i) = \mathbf{A} \cdot \mathbf{z}(i-1) + \mathbf{d}(i) \quad \text{(EQ 3-45)}$$

$$y(i) = \mathbf{C} \cdot \mathbf{z}(i) \quad \text{(EQ 3-46)}$$

where the augmented state vector $\mathbf{z}(i)$ is as follows:

$$\mathbf{z}(i) = \begin{bmatrix} \mathbf{z}_S(i) \\ \mathbf{z}_n(i) \end{bmatrix} \quad \text{(EQ 3-47),}$$

The augmented transition matrix of size $(p+q)\times(p+q)$ is represented as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_S & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_n \end{bmatrix} \quad \text{(EQ 3-48)}$$

The $(p+q)\times 1$ augmented driving process vector is:

$$\mathrm{d}(i) = \begin{bmatrix} \mathbf{B}_S \cdot u(i) \\ \mathbf{B}_n \cdot w(i) \end{bmatrix} = [0,\cdots,0,u(i),0,\cdots,0,w(i)]^T \quad \text{(EQ 3-49)}$$

and the $1\times(p+q)$ augmented observation vector is:

$$\mathbf{C} = [\mathbf{C}_S \ \mathbf{C}_n] = [0,\cdots,0,1,0,\cdots,0,1] \quad \text{(EQ 3-50)}$$

It can be noticed that the system represented by Equations (3-45) and (3-46) is driven by the white noise driving process $\mathbf{d}(i)$ with corresponding covariance matrix $\mathbf{Q}(i) = E\{\mathbf{d}(i)\cdot\mathbf{d}^{\mathrm{T}}(i)\}$, but with no measurement noise as seen in Equation (3-46). This is due to the fact that by including the colored noise model as an additional state in the global state-space model, the noise that directly affects the observation $y(i)$ has been incorporated into the state $\mathbf{z}(i)$ [POP'98].

Having defined all the parameters of the global state-space model and without going into all the derivations which can be found in the literature, the standard Kalman filter estimation and updating equations are then expressed as follows:

$$e(i) = y(i) - \mathbf{C}\cdot\hat{\mathbf{z}}(i/i-1) \quad \text{(EQ 3-51)}$$

$$\mathbf{K}(i) = \mathbf{P}(i/i-1)\cdot\mathbf{C}\cdot\left[\mathbf{C}\cdot\mathbf{P}(i/i-1)\cdot\mathbf{C}^{\mathrm{T}}\right]^{-1} \ \mathbf{K}(i) = \mathbf{P}(i/i-1)\cdot\mathbf{C}\times\left[\mathbf{C}\cdot\mathbf{P}(i/i-1)\cdot\mathbf{C}^{\mathrm{T}}\right]^{-1} \quad \text{(EQ 3-52)}$$

$$\hat{\mathbf{z}}(i/i) = \hat{\mathbf{z}}(i/i-1) + \mathbf{K}(i)\cdot e(i) \quad \text{(EQ 3-53)}$$

$$\mathbf{P}(i/i) = [\mathbf{I} - \mathbf{K}(i) \cdot \mathbf{C}] \cdot \mathbf{P}(i/i-1) \quad \text{(EQ 3-54)}$$

$$\hat{\mathbf{z}}(i+1/i) = \mathbf{A} \cdot \hat{\mathbf{z}}(i/i) \quad \text{(EQ 3-55)}$$

$$\mathbf{P}(i+1/i) = \mathbf{A} \cdot \mathbf{P}(i/i) \cdot \mathbf{A}^{\mathbf{T}} + \mathbf{Q}(i) \quad \text{(EQ 3-56)}$$

where:

- $\hat{\mathbf{z}}(i/i-1)$ is the minimum mean-square estimate of the state vector $\mathbf{z}(i)$ given the past observations $y(1), \cdots, y(i-1)$
- $\mathbf{P}(i/i-1)$ is the predicted (*a priori*) state-error covariance matrix
- $\hat{\mathbf{z}}(i/i)$ is the filtered estimate of the state vector $\mathbf{z}(i)$
- $\mathbf{P}(i/i)$ is the filtered state-error covariance matrix
- $\mathbf{Q}(i)$ is the driving processes covariance matrix
- $e(i)$ is the innovation sequence
- $\mathbf{K}(i)$ is the Kalman gain.

Finally, the enhanced speech signal $\hat{s}(i)$ can be obtained from the p^{th} component of the state-vector estimator, $\hat{\mathbf{z}}(i/i)$, which is considered as the output of the Kalman filter. However, selecting the p^{th} component is not the only choice. This Kalman filter algorithm will be used in Chapter 10, where it will be explained that selecting instead the first component of the state-vector provides a better enhancement result (but it introduces a delay). This technique is then referred to as the delayed Kalman filter.

In practice, the performance of the Kalman filtering technique strongly relies on the estimation of several unknown parameters. For instance, in practice the AR coefficients of the clean speech and noise are not available since only the noisy signal is accessible. Some Kalman filtering techniques run a pre-enhancement scheme to get a pseudo-enhanced signal prior to estimate the AR coefficients, and then the Kalman filtering is applied. Often, the AR coefficients can be found by Linear Predictive Coding (LPC) or via the Levinson-Durbin (LD) algorithm, commonly used in speech coding, filter design and spectral estimation. In simple terms, they find an all-pole filter based on the

autocorrelation sequence of the target signal. The estimated filter coefficients are also the required AR coefficients. Moreover, the colored noise covariance matrix and the variance of the noise driving processes are unknown parameters as well, and they have to be estimated. Some approaches are presented in [GAB'04][GAB'05][MA'04] to estimate those parameters. Furthermore, the transition matrices obtained from the AR coefficients are in fact time-varying, since the noise and especially the speech are not stationary. Typically, they should be updated within 10-30 ms, since under this time constraint the speech signal can be assumed more or less stationary [QUA'02]. Furthermore, Kalman filtering is known not to produce musical noise since it is not a spectral subtraction based technique, but it can have other artefacts such as the enhanced speech being muffled a bit.

Chapter 4. Overview of Previous Binaural Noise Reduction Schemes

In this chapter, we provide the description of two fairly recent binaural noise reduction schemes in the literature proposed in [BOG'07] and in [LOT'06]. In [BOG'07] (which complements the work in [KLA'06] and in several related publications such as [KLA'07],[DOC'05b]), a binaural Wiener filtering technique with a modified cost function was developed to specifically reduce directional noise, and also to have some control over the distortion level of the binaural interaural cues for both the speech and noise components. In [LOT'06], a binaural noise reduction scheme partially based on a Minimum Variance Distortionless Response (MVDR) beamforming concept was developed, more explicitly referred to as a superdirective beamformer with dual-channel input and output, followed by an adaptive post-filter. This scheme was designed to maintain all the interaural cues.

Section 4.1 will describe the Binaural Wiener filtering scheme and its practical implementation proposed in [BOG'07]. Section 4.2 will describe the binaural noise reduction scheme proposed in [LOT'07] based on MVDR-constrained beamforming with post-processing. A more elaborate derivation of the Binaural Wiener filtering scheme as opposed to the MVDR-constrained beamformer will be provided in this chapter, since in Chapter 9 we present a modified version of Binaural Wiener Filtering scheme to investigate the performance of the target power spectrum density estimator developed in Chapter 8. Finally, section 4.3 will show the relation/link between a Binaural Wiener Scheme and a MVDR-constrained beamformer.

4.1 Binaural Wiener filtering

The first subsection will provide the analytical derivation of binaural multichannel Wiener filtering along with MSE computations and the theoretical noise reduction result

that can be achieved when the target signal is corrupted by one directional noise. In the second subsection, the implementation of the binaural Wiener scheme proposed in [BOG'07] will be discussed along with the assumptions used.

4.1.1 Binaural Wiener Filtering Design

Figure 4-1 illustrates the layout of a typical binaural multichannel Wiener system under the assumption of the wireless link between the multiple left and right microphone signals (implying that the processing is performed with all the available microphone signals). M_L and M_R correspond to the number of left and right hearing aid microphones on each side of the head respectively. Therefore, there are $M = M_L + M_R$ microphones in total. Also, d_{LR} is the head diameter and d is the inter-microphone distance from the same hearing aid.

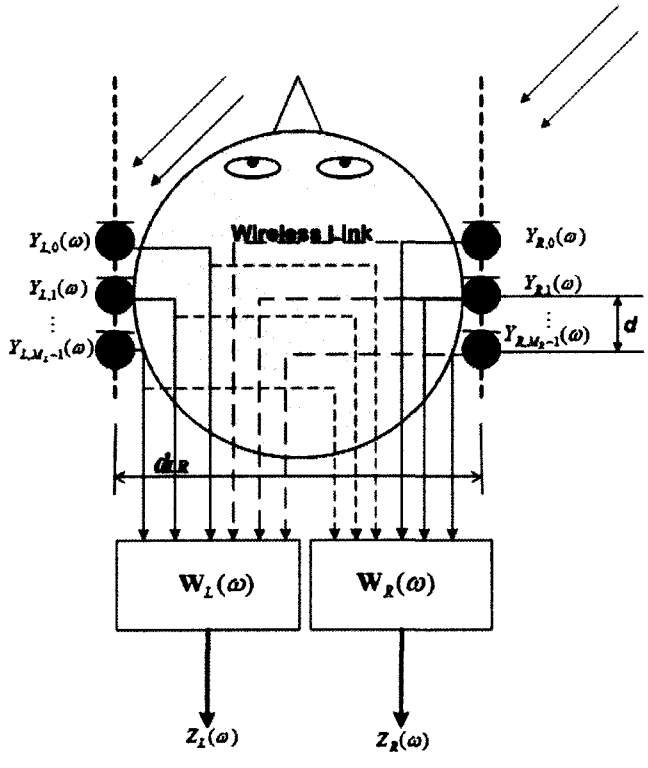

Figure 4-1: Binaural multichannel Wiener system

The binaural noisy input vector is formed by gathering all available left and right microphones signals in the frequency domain as follows:

$$\mathbf{Y}(\omega) = \begin{bmatrix} \mathbf{Y}_L(\omega) \\ \mathbf{Y}_R(\omega) \end{bmatrix} = \mathbf{S}(\omega) + \mathbf{V}(\omega) \qquad \text{(EQ 4-1)}$$

where $\mathbf{S}(\omega)$ is the binaural input target speech vector expressed as:

$$\mathbf{S}(\omega) = \begin{bmatrix} S_{L,0}(\omega) \\ S_{L,1}(\omega) \\ \vdots \\ S_{L,M_L-1}(\omega) \\ S_{R,0}(\omega) \\ S_{R,1}(\omega) \\ \vdots \\ S_{R,M_R-1}(\omega) \end{bmatrix} = \begin{bmatrix} H_{L,0}(\omega) \\ H_{L,1}(\omega) \\ \vdots \\ H_{L,M_L-1}(\omega) \\ H_{R,0}(\omega) \\ H_{R,1}(\omega) \\ \vdots \\ H_{R,M_R-1}(\omega) \end{bmatrix} \cdot S_s(\omega) \qquad \text{(EQ 4-2)}$$

$\mathbf{V}(\omega)$ is the binaural input noise vector expressed as:

$$\mathbf{V}(\omega) = \begin{bmatrix} V_{L,0}(\omega) \\ V_{L,1}(\omega) \\ \vdots \\ V_{L,M_L-1}(\omega) \\ V_{R,0}(\omega) \\ V_{R,1}(\omega) \\ \vdots \\ V_{R,M_R-1}(\omega) \end{bmatrix} = \begin{bmatrix} K_{L,0}(\omega) \\ K_{L,1}(\omega) \\ \vdots \\ K_{L,M_L-1}(\omega) \\ K_{R,0}(\omega) \\ K_{R,1}(\omega) \\ \vdots \\ K_{R,M_R-1}(\omega) \end{bmatrix} \cdot V_s(\omega) \qquad \text{(EQ 4-3)}$$

S_s and V_s are the speech and noise sources respectively. $H_{L,i}$ and $H_{R,i}$ are the HRTFs between the speech source and the i^{th} reference microphone on the left and right hearing aids, respectively. Similarly, $K_{L,i}$ and $K_{R,i}$ are the HRTFs between the noise source and the r^{th} reference microphone on the left and right hearing aids, respectively. It should be noted that in this representation we are assuming here that we are dealing with a single dominant noise (i.e. a single directional noise) present in the environment.

From the system above, the left and right hearing aid output signals are obtained as the following:

$$\begin{aligned} Z_L(\omega) &= \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega) \\ &= \mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega)+\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) \end{aligned} \quad \text{(EQ 4-4)}$$

$$\begin{aligned} Z_R(\omega) &= \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega) \\ &= \mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)+\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega) \end{aligned} \quad \text{(EQ 4-5)}$$

where $\mathbf{W}_L(\omega)$ and $\mathbf{W}_R(\omega)$ are left and right Wiener filters defined as:

$$\mathbf{W}_L(\omega)=\begin{bmatrix} W_{L,0}(\omega) \\ W_{L,1}(\omega) \\ \vdots \\ W_{L,M-1}(\omega) \end{bmatrix} \quad \text{(EQ 4-6)}$$

$$\mathbf{W}_R(\omega)=\begin{bmatrix} W_{R,0}(\omega) \\ W_{R,1}(\omega) \\ \vdots \\ W_{R,M-1}(\omega) \end{bmatrix} \quad \text{(EQ 4-7)}$$

Also, left and right Wiener filters can be concatenated into a single Mx2 matrix as in Equation (4-8) and applied to the noisy input vector $\mathbf{Y}(\omega)$ as in Equation (4-9):

$$\mathbf{W}(\omega)=\begin{bmatrix}\mathbf{W}_L(\omega) & \mathbf{W}_R(\omega)\end{bmatrix} \quad \text{(EQ 4-8)}$$

$$\mathbf{Z}(\omega)=\begin{bmatrix} Z_L(\omega) \\ Z_R(\omega) \end{bmatrix}=\mathbf{W}^H(\omega)\cdot\mathbf{Y}(\omega) \quad \text{(EQ 4-9)}$$

The total number of rows in $\mathbf{W}(\omega)$ corresponds to the total number of available microphone signals and the total number of columns in $\mathbf{W}(\omega)$ corresponds to the number of desired outputs. For instance, for the binaural system shown in Fig. 4-1, $\mathbf{W}(\omega)$ is Mx2.

In the hearing aid application, the goal is to find an estimate of the target speech signal from the noisy input signals received. For instance, in the context of binaural hearing, the left and right output signals from the system could then be an estimate of the speech

component in the first left microphone signal and an estimate of the speech component in the first right microphone signal i.e. $Z_L(\omega) \approx S_{L,0}(\omega)$ and $Z_R(\omega) \approx S_{R,0}(\omega)$

Let's define the vector of desired (target speech) signals to estimate as:

$$\mathbf{D}(\omega) = \begin{bmatrix} S_{L,0}(\omega) \\ S_{R,0}(\omega) \end{bmatrix} = \begin{bmatrix} S_L(\omega) \\ S_R(\omega) \end{bmatrix} \qquad \text{(EQ 4-10)}$$

Using a mean square error (MSE) cost function:

$$J_{MSE}\left(\mathbf{W}(\omega)\right) = E\left\{\left\|\left[\mathbf{D}(\omega) - \mathbf{W}^H(\omega)\cdot\mathbf{Y}(\omega)\right]\right\|^2\right\} = E\left\{\left\|\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\end{bmatrix}\right\|^2\right\}$$

(EQ 4-11)

the Wiener filters $\mathbf{W}(\omega) = \left[\mathbf{W}_L(\omega) \quad \mathbf{W}_R(\omega)\right]$ are then found by minimizing the cost function of Equation (4-11), which corresponds to the optimum solution in a Minimum MSE (MMSE) sense.

At first, let's further expand the cost function in (4-11) as follows:

$$J_{MSE}\left(\mathbf{W}(\omega)\right) = E\left\{\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\end{bmatrix}^H \cdot \begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\end{bmatrix}\right\}$$

$$= E\left\{\left[\left(S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\right)^H \quad \left(S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\right)^H\right]\cdot\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\end{bmatrix}\right\}$$

$$= E\left\{\left\|S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\right\|^2\right\} + E\left\{\left\|S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\right\|^2\right\}$$

$$= \quad J_{MSE_L}\left(\mathbf{W}_L(\omega)\right) \quad + \quad J_{MSE_R}\left(\mathbf{W}_R(\omega)\right)$$

(EQ 4-12)

As it can seen in (4-12), the overall cost function $J_{MSE}\left(\mathbf{W}(\omega)\right)$ can be decomposed into the sum of two independent systems (or the sum of two cost functions i.e. $J_{MSE_L}\left(\mathbf{W}_L(\omega)\right)$ and $J_{MSE_R}\left(\mathbf{W}_R(\omega)\right)$). Both systems use the same inputs $\mathbf{Y}(\omega)$ but generate an output using its own filter i.e. $\mathbf{W}_L(\omega)$ for the left output signal and $\mathbf{W}_R(\omega)$ for the right output signal. Therefore, the filter coefficients $\mathbf{W}_L(\omega)$ and $\mathbf{W}_R(\omega)$ can be found separately by minimizing their respective cost functions i.e. $\frac{\delta J_{MSE_L}}{\partial \mathbf{W}_L} = 0$, $\frac{\delta J_{MSE_R}}{\partial \mathbf{W}_R} = 0$.

Below, we will show the steps to find $\mathbf{W}_L(\omega)$:

$$\begin{aligned} J_{MSE_L}\left(\mathbf{W}_L(\omega)\right) &= E\left\{\left\|S_L(\omega)-\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\right\|^2\right\} \\ &= E\left\{\left(S_L(\omega)-\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\right)\cdot\left(S_L(\omega)-\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\right)^H\right\} \\ &= E\left\{\begin{matrix} S_L(\omega)\cdot S_L^H(\omega)-S_L(\omega)\cdot\mathbf{Y}^H(\omega)\cdot\mathbf{W}_L-\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\cdot S_L^H(\omega) \\ +\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\cdot\mathbf{Y}^H(\omega)\cdot\mathbf{W}_L \end{matrix}\right\} \\ &= E\left\{S_L(\omega)\cdot S_L^H(\omega)\right\}-E\left\{S_L(\omega)\cdot\mathbf{Y}^H(\omega)\right\}\cdot\mathbf{W}_L-\mathbf{W}_L^H(\omega)\cdot E\left\{\mathbf{Y}(\omega)\cdot S_L^H(\omega)\right\} \\ &\quad +\mathbf{W}_L^H(\omega)\cdot E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^H(\omega)\right\}\cdot\mathbf{W}_L \end{aligned}$$

(EQ 4-13)

By taking the gradient of (4-13) and setting it to zero, $\mathbf{W}_L(\omega)$ is found:

$$\frac{\partial J_{MSE_L}}{\partial \mathbf{W}_L}=-2\cdot E\left\{\mathbf{Y}(\omega)\cdot S_L^H(\omega)\right\}+2\cdot E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^H(\omega)\right\}\cdot\mathbf{W}_L(\omega)=0 \qquad \text{(EQ 4-14)}$$

$$\mathbf{W}_L(\omega)=E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^H(\omega)\right\}^{-1}\cdot E\left\{\mathbf{Y}(\omega)\cdot S_L^H(\omega)\right\} \qquad \text{(EQ 4-15)}$$

We can rewrite Equation (4-15) in a more compact form:

$$\mathbf{W}_L(\omega)=\mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega) \qquad \text{(EQ 4-16)}$$

with $\mathbf{\Gamma}_{\mathbf{YY}}(\omega)=E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^H(\omega)\right\}$, (EQ 4-17)

$$\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega)=E\left\{\mathbf{Y}(\omega)\cdot S_L^*(\omega)\right\} \qquad \text{(EQ 4-18)}$$

and $\mathbf{\Gamma}$ is defined here as a PSD function

Similarly, for the right channel estimation of $S_R(\omega)$, the binaural Wiener solution is then:

$$\mathbf{W}_R(\omega)=\mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{YS}_R}(\omega) \qquad \text{(EQ 4-19)}$$

with $\mathbf{\Gamma}_{\mathbf{YS}_R}(\omega)=E\left\{\mathbf{Y}(\omega)\cdot S_R^*(\omega)\right\}$ (EQ 4-20)

Regrouping Equations (4-19) and (4-20), the overall Wiener solution of the system can be expressed as:

$$\mathbf{W}(\omega)=\left[\mathbf{W}_L(\omega)\quad\mathbf{W}_R(\omega)\right]=\mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{YD}}(\omega) \qquad \text{(EQ 4-21)}$$

with:

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{YD}}(\omega) &= E\left\{\mathbf{Y}(\omega)\cdot\mathbf{D}^{H}(\omega)\right\} \\ &= \left[\boldsymbol{\Gamma}_{\mathbf{YS}_L}(\omega) \quad \boldsymbol{\Gamma}_{\mathbf{YS}_R}(\omega)\right]\end{aligned} \qquad \text{(EQ 4-22)}$$

From the filter solutions obtained, we can also express $J_{MSE}\left(\mathbf{W}(\omega)\right)$ in another form. At first using (4-16), the MSE cost function for the left output channel is rewritten as:

$$\begin{aligned}J_{MSE_L}\left(\mathbf{W}_L(\omega)\right) &= E\left\{S_L(\omega)\cdot S_L^{H}(\omega)\right\} - E\left\{S_L(\omega)\cdot\mathbf{Y}^{H}(\omega)\right\}\cdot\mathbf{W}_L - \mathbf{W}_L^{H}(\omega)\cdot E\left\{\mathbf{Y}(\omega)\cdot S_L^{H}(\omega)\right\} \\ &\quad + \mathbf{W}_L^{H}(\omega)\cdot E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^{H}(\omega)\right\}\cdot\mathbf{W}_L \\ &= \Gamma_{S_LS_L}(\omega) - \Gamma_{S_L\mathbf{Y}}(\omega)\cdot\mathbf{W}_L - \mathbf{W}_L^{H}(\omega)\cdot\Gamma_{\mathbf{YS}_L}(\omega) + \mathbf{W}_L^{H}(\omega)\cdot\Gamma_{\mathbf{YY}}(\omega)\cdot\mathbf{W}_L \\ &= \Gamma_{S_LS_L}(\omega) - \Gamma_{\mathbf{YS}_L}^{H}\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_L}(\omega)\end{aligned} \qquad \text{(EQ 4-23)}$$

Similarly, using (4-19), the MSE cost function for the right output channel is then:

$$\Gamma_{S_RS_R}(\omega) - \Gamma_{\mathbf{YS}_R}^{H}\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_R}(\omega) \qquad \text{(EQ 4-24)}$$

The overall cost function of the entire system is re-written as follows:

$$\begin{aligned}J_{MSE}\left(\mathbf{W}(\omega)\right) &= J_{MSE_L}\left(\mathbf{W}_L(\omega)\right) + J_{MSE_R}\left(\mathbf{W}_R(\omega)\right) \\ &= \Gamma_{S_LS_L}(\omega) - \Gamma_{\mathbf{YS}_L}^{H}\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_L}(\omega) + \Gamma_{S_RS_R}(\omega) - \Gamma_{\mathbf{YS}_R}^{H}\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_R}(\omega)\end{aligned} \qquad \text{(EQ 4-25)}$$

The next section will provide further insight into the binaural Wiener solution and the resulting theoretical left and right MSEs, when the binaural target speech signal is corrupted by a single dominant directional noise source.

4.1.2 Insights on Binaural Wiener Solution

Let's assume a binaural system with one microphone on each side (i.e. M_L=1 and M_R=1 then M=2). The system is then reduced to:

$$\mathbf{Y}(\omega) = \begin{bmatrix} Y_L \\ Y_R \end{bmatrix} = \mathbf{S}(\omega) + \mathbf{V}(\omega) \qquad \text{(EQ 4-26)}$$

$$\text{with } \mathbf{S}(\omega)=\begin{bmatrix}S_L(\omega)\\S_R(\omega)\end{bmatrix}=\begin{bmatrix}H_L(\omega)\\H_R(\omega)\end{bmatrix}\cdot S_s(\omega) \qquad \text{(EQ 4-27)}$$

We assume that there is only a single dominant noise source (directional noise) emerging from an arbitrary lateral location corrupting the binaural target speech signal. The binaural noise signal can be then represented as follows:

$$\mathbf{V}(\omega)=\begin{bmatrix}V_L(\omega)\\V_R(\omega)\end{bmatrix}=\begin{bmatrix}K_L(\omega)\\K_R(\omega)\end{bmatrix}\cdot V_s(\omega) \qquad \text{(EQ 4-28)}$$

The desired target vector is then:

$$\mathbf{D}(\omega)=\begin{bmatrix}S_L(\omega)\\S_R(\omega)\end{bmatrix} \qquad \text{(EQ 4-29)}$$

The binaural outputs are then:

$$Z_L(\omega)=\mathbf{W}_L^H(\omega)\cdot\mathbf{Y}(\omega)\approx S_L(\omega) \qquad \text{(EQ 4-30)}$$

$$\text{and } Z_R(\omega)=\mathbf{W}_R^H(\omega)\cdot\mathbf{Y}(\omega)\approx S_R(\omega) \qquad \text{(EQ 4-31)}$$

Let's evaluate the MSE of the estimated output signals. For convenience, the MSE computations are shown for the left output channel only, i.e. using Equation (4-23):

$$MSE_L\left(\mathbf{W}_L(\omega)\right)=\Gamma_{S_LS_L}(\omega)-\Gamma_{\mathbf{YS}_L}^H\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_L}(\omega) \qquad \text{(EQ 4-32)}$$

We shall now explicitly evaluate each parameter of MSE_L. Assuming that the target speech and the noise components are uncorrelated i.e.

$\Gamma_{S_LV_L}(\omega)=\Gamma_{S_LV_R}(\omega)=\Gamma_{S_RV_R}(\omega)=\Gamma_{S_RV_L}(\omega)=0$):

$$\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega)=E\left\{\begin{bmatrix}S_L(\omega)+V_L(\omega)\\S_R(\omega)+V_R(\omega)\end{bmatrix}\cdot S_L^*(\omega)\right\}=\begin{bmatrix}\Gamma_{S_LS_L}(\omega)\\\Gamma_{S_RS_L}(\omega)\end{bmatrix} \qquad \text{(EQ 4-33)}$$

and

$$\Gamma_{\mathbf{YY}}(\omega)=E\left\{\mathbf{Y}(\omega)\cdot\mathbf{Y}^{H}(\omega)\right\}=E\left\{\begin{bmatrix}S_L(\omega)+V_L(\omega)\\S_R(\omega)+V_R(\omega)\end{bmatrix}\cdot\begin{bmatrix}S_L(\omega)+V_L(\omega)\\S_R(\omega)+V_R(\omega)\end{bmatrix}^{H}\right\}$$
$$=\begin{bmatrix}\Gamma_{S_LS_L}(\omega)+\Gamma_{V_LV_L}(\omega) & \Gamma_{S_LS_R}(\omega)+\Gamma_{V_LV_R}(\omega)\\ \Gamma_{S_RS_L}(\omega)+\Gamma_{V_RV_L}(\omega) & \Gamma_{S_RS_R}(\omega)+\Gamma_{V_RV_R}(\omega)\end{bmatrix} \quad \text{(EQ 4-34)}$$

$$\text{then } \Gamma_{\mathbf{YY}}^{-1}(\omega)=\begin{pmatrix}a & b\\ c & d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}$$

$$=\frac{1}{\begin{pmatrix}\Gamma_{S_LS_L}\cdot\Gamma_{S_RS_R}+\Gamma_{S_LS_L}\cdot\Gamma_{V_RV_R}+\Gamma_{V_LV_L}\cdot\Gamma_{S_RS_R}+\Gamma_{V_LV_L}\cdot\Gamma_{V_RV_R}\\ -\Gamma_{S_LS_R}\cdot\Gamma_{S_LS_R}-\Gamma_{S_LS_R}\cdot\Gamma_{V_RV_L}-\Gamma_{V_LV_R}\cdot\Gamma_{V_RV_L}-\Gamma_{V_LV_R}\cdot\Gamma_{S_RS_L}\end{pmatrix}}\cdot\begin{bmatrix}\Gamma_{S_RS_R}+\Gamma_{V_RV_R} & -\Gamma_{S_LS_R}-\Gamma_{V_LV_R}\\ -\Gamma_{S_RS_L}-\Gamma_{V_RV_L} & \Gamma_{S_LS_L}+\Gamma_{V_LV_L}\end{bmatrix}$$

(EQ 4-35)

Equation (4-35) can be further simplified by using the following equality principle from the coherence function of two correlated signals (refer to Chapter 6, Section 6.3.5 for more explanation). For instance, the coherence function between the left and right speech signals is defined as:

$$\psi_{S_LS_R}(\omega)=\frac{\Gamma_{S_LS_R}(\omega)}{\sqrt{\Gamma_{S_LS_L}(\omega)\cdot\Gamma_{S_RS_R}(\omega)}} \quad \text{(EQ 4-36)}$$

and inversely between the right and left speech signal it is defined as:

$$\psi_{S_RS_L}(\omega)=\psi^{*}_{S_LS_R}(\omega)=\frac{\Gamma_{S_RS_L}(\omega)}{\sqrt{\Gamma_{S_LS_L}(\omega)\cdot\Gamma_{S_RS_R}(\omega)}} \quad \text{(EQ 4-37)}$$

Using Equation (4-36) and (4-37), we deduce the following equality:

$$\Gamma_{S_LS_R}(\omega)\cdot\Gamma_{S_RS_L}(\omega)=\left|\psi_{S_LS_R}(\omega)\right|\cdot\Gamma_{S_LS_L}(\omega)\cdot\Gamma_{S_RS_R}(\omega)$$
$$=1\cdot\Gamma_{S_LS_L}(\omega)\cdot\Gamma_{S_RS_R}(\omega) \quad \text{(EQ 4-38)}$$

Note that a coherence magnitude of 1 (i.e. $\left|\psi_{S_LS_R}(\omega)\right|=1$) in Equation (4-38) is only valid if there exists a full correlation between the left and right target speech signals.

Similarly, the same equality can be derived for the left and right noise signals (since we assume that there is a single directional noise source present) as follows:

$$\begin{aligned}\Gamma_{V_LV_R}(\omega)\cdot\Gamma_{V_RV_L}(\omega) &= \left|\psi_{V_LV_R}(\omega)\right|\cdot\Gamma_{V_LV_L}(\omega)\cdot\Gamma_{V_RV_R}(\omega)\\ &= 1\cdot\Gamma_{V_LV_L}(\omega)\cdot\Gamma_{V_RV_R}(\omega)\end{aligned} \quad \text{(EQ 4-39)}$$

Using (4-37 and 4-38), four terms in Equation (4-35) can then be cancelled:

$$\mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega) = \frac{1}{\begin{pmatrix}\Gamma_{S_LS_L}\cdot\Gamma_{V_RV_R}+\Gamma_{V_LV_L}\cdot\Gamma_{S_RS_R}\\ -\Gamma_{S_LS_R}\cdot\Gamma_{V_RV_L}-\Gamma_{V_LV_R}\cdot\Gamma_{S_RS_L}\end{pmatrix}}\cdot\begin{bmatrix}\Gamma_{S_RS_R}+\Gamma_{V_RV_R} & -\Gamma_{S_LS_R}-\Gamma_{V_LV_R}\\ -\Gamma_{S_RS_L}-\Gamma_{V_RV_L} & \Gamma_{S_LS_L}+\Gamma_{V_LV_L}\end{bmatrix} \quad \text{(EQ 4-40)}$$

Finally, by grouping all the parameters (i.e. 4-33,4-34 and 4-40) in (4-32), the result of the MSE for the left output signal is then found:

$$MSE_L\left(\mathbf{W}_L(\omega)\right) = \Gamma_{S_LS_L}(\omega) - \Gamma_{\mathbf{YS}_L}^{H}\cdot\Gamma_{\mathbf{YY}}^{-1}(\omega)\cdot\Gamma_{\mathbf{YS}_L}(\omega) = 0 \quad \text{(EQ 4-41)}$$

Similarly, $MSE_R\left(\mathbf{W}_R(\omega)\right) = 0$ (EQ 4-42)

Since the MSE is zero, it can be concluded that for the case where the target source is corrupted by a dominant directional noise source emerging from an arbitrary location, by applying the theoretical optimum binaural Wiener filter solution $\mathbf{W}(\omega) = \left[\mathbf{W}_L(\omega)\ \ \mathbf{W}_R(\omega)\right]$, the left and right target speech signals can be perfectly retrieved i.e. $Z_L(\omega) = S_L(\omega)$ and $Z_R(\omega) = S_R(\omega)$.

If we take a closer look at the actual form of the Wiener solution, by working on the left channel and recalling Equation (4-16):

$$\begin{aligned}\mathbf{W}_L(\omega) &= \mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega)\\ &= \frac{1}{\begin{pmatrix}\Gamma_{S_LS_L}\cdot\Gamma_{V_RV_R}+\Gamma_{V_LV_L}\cdot\Gamma_{S_RS_R}\\ -\Gamma_{S_LS_R}\cdot\Gamma_{V_RV_L}-\Gamma_{V_LV_R}\cdot\Gamma_{S_RS_L}\end{pmatrix}}\cdot\begin{bmatrix}\Gamma_{V_RV_R}\cdot\Gamma_{S_LS_L}-\Gamma_{V_LV_R}\cdot\Gamma_{S_RS_L}\\ -\Gamma_{V_RV_L}\cdot\Gamma_{S_LS_L}+\Gamma_{V_LV_L}\cdot\Gamma_{S_RS_L}\end{bmatrix}\end{aligned} \quad \text{(EQ 4-43)}$$

Using (4-27) and (4-28), we can express all the parameters of the Wiener filter (4-43) in terms of HRTFs belonging to the target source and the noise source, with their corresponding source PSDs.

$$\Gamma_{S_L S_L}(\omega) = E\{S_L(\omega)\cdot S_L^*(\omega)\} = E\{S_s(\omega)\cdot H_L(\omega)\cdot S_s^*(\omega)H_L^*(\omega)\} = E\{S_s(\omega)\cdot S_s^*(\omega)\}\cdot|H_L(\omega)|^2$$
$$= \Gamma_{S_s S_s}(\omega)\cdot|H_L(\omega)|^2 \qquad \text{(EQ 4-44)}$$

$$\Gamma_{S_L S_R}(\omega) = E\{S_L(\omega)\cdot S_R^*(\omega)\} = \Gamma_{S_s S_s}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega) \qquad \text{(EQ 4-45)}$$

$$\Gamma_{S_R S_R}(\omega) = \Gamma_{S_s S_s}\cdot|H_R(\omega)|^2 \qquad \text{(EQ 4-46)}$$

$$\Gamma_{V_L V_L}(\omega) = \Gamma_{V_s V_s}(\omega)\cdot|K_L(\omega)|^2 \qquad \text{(EQ 4-47)}$$

$$\Gamma_{V_L V_R}(\omega) = \Gamma_{V_s V_s}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega) \qquad \text{(EQ 4-48)}$$

Substituting (4-44) to (4-48) into (4-43), yields:

$$\mathbf{W}_L(\omega) = \begin{bmatrix} \Gamma_{S_s S_s}(\omega)\cdot|H_R(\omega)|^2\cdot\Gamma_{V_s V_s}(\omega)\cdot|K_R(\omega)|^2 - \Gamma_{V_s V_s}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)\cdot\Gamma_{S_s S_s}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) \\ -\Gamma_{V_s V_s}(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega)\cdot\Gamma_{S_s S_s}(\omega)\cdot|H_L(\omega)|^2 + \Gamma_{V_s V_s}(\omega)\cdot|K_L(\omega)|^2\cdot\Gamma_{S_s S_s}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) \end{bmatrix}$$
$$\cdot\frac{1}{\begin{pmatrix} \Gamma_{S_s S_s}(\omega)\cdot|H_L(\omega)|^2\cdot\Gamma_{V_s V_s}(\omega)\cdot|K_R(\omega)|^2 + \Gamma_{V_s V_s}(\omega)\cdot|K_L(\omega)|^2\cdot\Gamma_{S_s S_s}(\omega)\cdot|H_R(\omega)|^2 \\ -\Gamma_{S_s S_s}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)\cdot\Gamma_{V_s V_s}(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega) - \Gamma_{V_s V_s}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)\cdot \\ \Gamma_{S_s S_s}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) \end{pmatrix}}$$
$$= \begin{bmatrix} |H_R(\omega)|^2\cdot|K_R(\omega)|^2 - K_L(\omega)\cdot K_R^*(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) \\ -K_L^*(\omega)\cdot K_R(\omega)\cdot|H_L(\omega)|^2 + |K_L(\omega)|^2\cdot H_L^*(\omega)\cdot H_R(\omega) \end{bmatrix}$$
$$\cdot\frac{1}{\begin{pmatrix} |H_L(\omega)|^2\cdot|K_R(\omega)|^2 + |K_L(\omega)|^2\cdot|H_R(\omega)|^2 \\ -H_L(\omega)\cdot H_R^*(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega) - K_L(\omega)\cdot K_R^*(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) \end{pmatrix}}\cdot\frac{\Gamma_{S_s S_s}(\omega)\cdot\Gamma_{V_s V_s}(\omega)}{\Gamma_{S_s S_s}(\omega)\cdot\Gamma_{V_s V_s}(\omega)}$$
(EQ 4-49)

As it can be noticed in (4-49), the binaural multichannel Wiener solution for the case with a target source and one directional noise source is independent of the actual target or noise PSDs. The solution is only a function of the HRTFs from the target and the directional noise source, which implies that the Wiener solution becomes only a function of the spatial position of each source. Applying this Wiener solution will theoretically allow a perfect noise reduction (i.e. perfect extraction of the left and right noise-free target speech signal) and if we do an analogy with beamforming, it is equivalent to

inserting a notch in the beampattern for the direction of arrival of the single directional interference.

In the next section, a practical implementation of the binaural Wiener filter proposed in [BOG'07] will be shown for the same case where the target speech is corrupted by one directional noise, along with typical assumptions.

4.1.3 Binaural Wiener Implementation Proposed in [BOG'07]

In Section 4.1.1 the standard binaural multichannel Wiener filter was described and its performance against a single directional noise was described in Section 4.1.2. The optimum ideal binaural Wiener filter coefficients were found using Equations (4-21) and (4-22). However, to compute those coefficients, the statistical cross-correlation vectors (i.e. Equations (4-18) and (4-20)) between the binaural noisy input signals and the binaural target speech signals are required. Unfortunately, in practice, those cross-correlation vectors are not directly accessible. A practical approach is proposed in [BOG'07] where the binaural noise reduction scheme is first based on the standard binaural Wiener filters described in Section 4.1.1, with the parameters of the Wiener filters (such as the unknown statistical cross-correlation vectors) estimated by using a robust VAD and relying on two assumptions:

i) the target speech and noise are statistically independent, therefore Equation (4-17) can be rewritten as:

$$\mathbf{\Gamma}_{YY}(\omega) = \mathbf{\Gamma}_{SS}(\omega) + \mathbf{\Gamma}_{VV}(\omega) \qquad \text{(EQ 4-50)}$$

where $\mathbf{\Gamma}_{SS}(\omega)$ is the statistical cross-correlation matrix of the binaural target speech input signals defined as:

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{SS}}(\omega) &= E\left\{\mathbf{S}(\omega)\cdot\mathbf{S}^{H}(\omega)\right\} = E\left\{\begin{bmatrix}S_L(\omega)\\S_R(\omega)\end{bmatrix}\cdot\begin{bmatrix}S_L(\omega)\\S_R(\omega)\end{bmatrix}^{H}\right\} \\ &= \begin{bmatrix}\boldsymbol{\Gamma}_{\mathbf{SS}_L}(\omega) & \boldsymbol{\Gamma}_{\mathbf{SS}_R}(\omega)\end{bmatrix}\end{aligned} \quad \text{(EQ 4-51)}$$

and $\boldsymbol{\Gamma}_{\mathbf{VV}}(\omega)$ is the statistical correlation matrix of the binaural noise signals defined as:

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{VV}}(\omega) &= E\left\{\mathbf{V}(\omega)\cdot\mathbf{V}^{H}(\omega)\right\} \\ &= E\left\{\begin{bmatrix}V_L(\omega)\\V_R(\omega)\end{bmatrix}\cdot\begin{bmatrix}V_L(\omega)\\V_R(\omega)\end{bmatrix}^{H}\right\}\end{aligned} \quad \text{(EQ 4-52)}$$

Using the assumption i), the statistical cross-correlation vectors in (4-18) and (4-20) can be then simplified to:

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{YS}_L}(\omega) &= E\left\{\mathbf{Y}(\omega)\cdot S_L^{*}(\omega)\right\} = E\left\{\mathbf{S}(\omega)\cdot S_L^{*}(\omega)\right\} \\ &= \boldsymbol{\Gamma}_{\mathbf{SS}_L}(\omega)\end{aligned} \quad \text{(EQ 4-53)}$$

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{YS}_R}(\omega) &= E\left\{\mathbf{Y}(\omega)\cdot S_R^{*}(\omega)\right\} = E\left\{\mathbf{S}(\omega)\cdot S_R^{*}(\omega)\right\} \\ &= \boldsymbol{\Gamma}_{\mathbf{SS}_R}(\omega)\end{aligned} \quad \text{(EQ 4-54)}$$

And using (4-53) and (4-54), Equation (4-22) reduces to:

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{YD}}(\omega) &= \begin{bmatrix}\boldsymbol{\Gamma}_{\mathbf{YS}_L}(\omega) & \boldsymbol{\Gamma}_{\mathbf{YS}_R}(\omega)\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{\Gamma}_{\mathbf{SS}_L}(\omega) & \boldsymbol{\Gamma}_{\mathbf{SS}_R}(\omega)\end{bmatrix}\end{aligned} \quad \text{(EQ 4-55)}$$

ii) The noise signal is considered short-term stationary implying that $\boldsymbol{\Gamma}_{\mathbf{VV}}(\omega)$ is the same whether it is calculated during noise-only periods or during target speech + noise periods.

In [BOG'07][KLA'07][DOC'05b], from assumption ii) and having access to an ideal VAD, $\boldsymbol{\Gamma}_{\mathbf{VV}}(\omega)$ could then be estimated using an average over "noise-only" periods resulting in estimate $\tilde{\boldsymbol{\Gamma}}_{\mathbf{VV}}(\omega)$, and $\boldsymbol{\Gamma}_{\mathbf{YY}}(\omega)$ could be estimated using "speech+noise"

periods producing estimate $\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)$. Consequently, an estimate of $\mathbf{\Gamma}_{\mathbf{SS}}(\omega)$ could be found by using (4-50) as follows:

$$\begin{aligned}\tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega) &= \tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega) - \tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega) \\ &= \left[\tilde{\mathbf{\Gamma}}_{\mathbf{SS}_L}(\omega) \quad \tilde{\mathbf{\Gamma}}_{\mathbf{SS}_R}(\omega)\right]\end{aligned} \qquad \text{(EQ 4-56)}$$

An estimate of $\mathbf{\Gamma}_{\mathbf{YD}}(\omega)$ in (4-55) can be then replaced by Equation (4-56) as follows:

$$\mathbf{\Gamma}_{\mathbf{YD}}(\omega) = \left[\mathbf{\Gamma}_{\mathbf{SS}_L}(\omega) \quad \mathbf{\Gamma}_{\mathbf{SS}_R}(\omega)\right] \approx \tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega) \qquad \text{(EQ 4-57)}$$

As a result, the binaural Wiener coefficients are then estimated using (4-57) as the following:

$$\begin{aligned}\mathbf{W}(\omega) = \mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega) \cdot \mathbf{\Gamma}_{\mathbf{YD}}(\omega) &\approx \left(\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)\right)^{-1} \cdot \tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega) \\ &\approx \left(\tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega) + \tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega)\right)^{-1} \cdot \tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega)\end{aligned} \qquad \text{(EQ 4-58)}$$

Unfortunately, applying the Wiener filters obtained using Equation (4-58) will reduce the noise but it will not maintain the interaural cues of the original (i.e. unprocessed) noise in the binaural outputs i.e. ($\left[Z_L(\omega) \quad Z_R(\omega)\right]^T = \mathbf{W}^H(\omega) \cdot \mathbf{Y}(\omega)$). The filtering tends to shift the original interaural cues of the noise components towards the cues of the speech component i.e. losing their spatial separation due to interaural cues distortion. However, the second part of the work in [BOG'07] was to find an approach to control the level of interaural cues distortion for both the target speech and noise, while reducing the noise. It was found that by extending the cost function defined in Equation (4-11) to include two extra expressions involving the interaural transfer functions of the target speech and the noise (referred to as ITF_S and ITF_V respectively), it is possible to control the interaural cues distortion level and the noise reduction strength. The first expression is the ITF cost function of noise defined as:

$$J_{ITF}^{V}\left(\mathbf{W}(\omega)\right) = E\left\{\left|\frac{\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega)}{\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)} - ITF_{in}^{V}\right|^2\right\} = E\left\{\frac{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) - ITF_{in}^{V}\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\right|^2}{\left|\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\right|^2}\right\} \qquad \text{(EQ 4-59)}$$

$$\text{with } ITF_{in}^{V} = \frac{V_L(\omega)}{V_R(\omega)} \approx \frac{E\{V_L(\omega)\cdot V_R^*(\omega)\}}{E\{V_R(\omega)\cdot V_R^*(\omega)\}} = \frac{\mathbf{e}_L^H\cdot\boldsymbol{\Gamma}_{\mathbf{vv}}\cdot\mathbf{e}_R^H}{\mathbf{e}_R^H\cdot\boldsymbol{\Gamma}_{\mathbf{vv}}\cdot\mathbf{e}_R^H} \quad \text{(EQ 4-60)}$$

In Equation (4-60), $\mathbf{e}_L$ and $\mathbf{e}_R$ are M-dimensional vectors with one non-zero element equal to 1.0, corresponding to the selected reference microphone signal on each side, and the other elements are 0.

Similarly, the second expression is the ITF cost function of the speech defined as:

$$J_{ITF}^{S}\left(\mathbf{W}(\omega)\right) = E\left\{\frac{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) - ITF_{in}^{V}\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\right|^2}{\left|\mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\right|^2}\right\} \quad \text{(EQ 4-61)}$$

$$\text{with } ITF_{in}^{S} = \frac{S_L(\omega)}{S_R(\omega)} \approx \frac{E\{S_L(\omega)\cdot S_R^*(\omega)\}}{E\{S_R(\omega)\cdot S_R^*(\omega)\}} = \frac{\mathbf{e}_L^H\cdot\boldsymbol{\Gamma}_{\mathbf{ss}}\cdot\mathbf{e}_R^H}{\mathbf{e}_R^H\cdot\boldsymbol{\Gamma}_{\mathbf{ss}}\cdot\mathbf{e}_R^H} \quad \text{(EQ 4-62)}$$

It should be noted that to estimate Equations (4-60) and (4-62), another assumption made in [BOG'07] is that the speech and noise are stationary (i.e. they do not relocate or move) and they can be computed using the received binaural noisy signals.

Using (4-59) and (4-61), the extended cost function of (4-11) becomes then:

$$\begin{aligned} J^{Total}\left(\mathbf{W}(\omega)\right) &= E\left\{\left\|\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) \\ \mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\end{bmatrix}\right\|^2\right\} \\ &\quad + \alpha\cdot J_{ITF}^{S}\left(\mathbf{W}(\omega)\right) + \beta\cdot J_{ITF}^{V}\left(\mathbf{W}(\omega)\right) \\ &= E\left\{\left\|\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) \\ \mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\end{bmatrix}\right\|^2\right\} \\ &\quad + \alpha\cdot E\left\{\frac{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) - ITF_S\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\right|^2}{\left|\mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\right|^2}\right\} + \beta\cdot E\left\{\frac{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) - ITF_V\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\right|^2}{\left|\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\right|^2}\right\} \end{aligned} \quad \text{(EQ 4-63)}$$

It is stated that no closed expression is available to minimize $J^{Total}(\mathbf{W}(\omega))$ without requiring iterative techniques. However, simplified quadratic ITF cost functions of the speech and noise are used instead and $J^{Total}(\mathbf{W}(\omega))$ is simplified to the following:

$$J^{Total}(\mathbf{W}(\omega)) = E\left\{\left\|\begin{bmatrix} S_L(\omega) - \mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) \\ S_R(\omega) - \mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\end{bmatrix}\right\|^2 + \mu\left\|\begin{bmatrix}\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) \\ \mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\end{bmatrix}\right\|^2\right\}$$
$$+\alpha\cdot E\left\{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{S}(\omega) - ITF_S\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{S}(\omega)\right|^2\right\}\beta\cdot E\left\{\left|\mathbf{W}_L^H(\omega)\cdot\mathbf{V}(\omega) - ITF_V\cdot\mathbf{W}_R^H(\omega)\cdot\mathbf{V}(\omega)\right|^2\right\}$$
(EQ 4-64)

Without going to the details such as in Section 4.1.1, solving this extended cost function yields the extended binaural Wiener filter with cues distortions control (or cues preservation tradeoff) as follows:

$$\mathbf{W}^{BWF\_ITF}(\omega) = \begin{bmatrix}\mathbf{W}_L^{BWF\_ITF}(\omega) \\ \mathbf{W}_R^{BWF\_ITF}(\omega)\end{bmatrix} = \left(\mathbf{R}_{\mathbf{Rs}}(\omega) + \mu\cdot\mathbf{R}_{\mathbf{Rv}}(\omega) + \alpha\cdot\mathbf{R}_{\mathbf{Rsc}}(\omega) + \beta\cdot\mathbf{R}_{\mathbf{Rvc}}(\omega)\right)^{-1}\cdot\mathbf{r}_{\mathbf{X}}(\omega)$$
(EQ 4-65)

where $\mathbf{r}_{\mathbf{X}} = \mathbf{\Gamma}_{\mathbf{YD}}^T(\omega) = \begin{bmatrix}\mathbf{\Gamma}_{SS_L}(\omega) \\ \mathbf{\Gamma}_{SS_R}(\omega)\end{bmatrix}$ (EQ 4-66),

$$\mathbf{R}_{\mathbf{Rs}}(\omega) = \begin{bmatrix}\mathbf{\Gamma}_{\mathbf{SS}}(\omega) & \mathbf{0}_{\mathbf{M\times M}} \\ \mathbf{0}_{\mathbf{M\times M}} & \mathbf{\Gamma}_{\mathbf{SS}}(\omega)\end{bmatrix}$$ (EQ 4-67),

$$\mathbf{R}_{\mathbf{Rv}}(\omega) = \begin{bmatrix}\mathbf{\Gamma}_{\mathbf{VV}}(\omega) & \mathbf{0}_{\mathbf{M\times M}} \\ \mathbf{0}_{\mathbf{M\times M}} & \mathbf{\Gamma}_{\mathbf{VV}}(\omega)\end{bmatrix}$$ (EQ 4-68)

$$\mathbf{R}_{\mathbf{Rsc}}(\omega) = \begin{bmatrix}\mathbf{\Gamma}_{\mathbf{SS}}(\omega) & -ITF_S^*\cdot\mathbf{\Gamma}_{\mathbf{SS}}(\omega) \\ -ITF_S\cdot\mathbf{\Gamma}_{\mathbf{SS}}(\omega) & |ITF_S|^2\cdot\mathbf{\Gamma}_{\mathbf{SS}}(\omega)\end{bmatrix}$$ (EQ 4-69),

$$\mathbf{R}_{\mathbf{Rvc}}(\omega) = \begin{bmatrix}\mathbf{\Gamma}_{\mathbf{VV}}(\omega) & -ITF_V^*\cdot\mathbf{\Gamma}_{\mathbf{VV}}(\omega) \\ -ITF_V\cdot\mathbf{\Gamma}_{\mathbf{VV}}(\omega) & |ITF_V|^2\cdot\mathbf{\Gamma}_{\mathbf{VV}}(\omega)\end{bmatrix}$$ (EQ 4-70)

In the binaural Wiener filters expressed as in Equation (4-65), the variable μ provides a tradeoff between noise reduction and speech content distortion. α controls the speech cues distortion and β controls the noise cues distortion. For instance, placing more emphasis on cues preservation (i.e. increasing α and β) will decrease the noise reduction performance, which is again another tradeoff. A more detailed analysis on the interaction of those variables can be found in [BOG'07].

4.2 Binaural MVDR-based Beamforming with Post-Processing

In [LOT'06], a binaural noise reduction scheme was proposed, which consists of a superdirective beamformer and a postfilter. However, their proposed scheme outputs enhanced binaural signals with full preservation of the interaural cues of the original signal, unlike the multichannel Wiener solution of the previous section. Figure 4-2 illustrates the structure of their algorithm in the frequency domain.

In Fig. 4-2, a binaural system with M=2 microphones is shown (i.e. one microphone per hearing aid and under the assumption of a perfect binaural link between the left and right microphones of each ear), and $Y_L(\omega)$ and $Y_R(\omega)$ are the left and right noisy input speech signals expressed in the frequency domain.

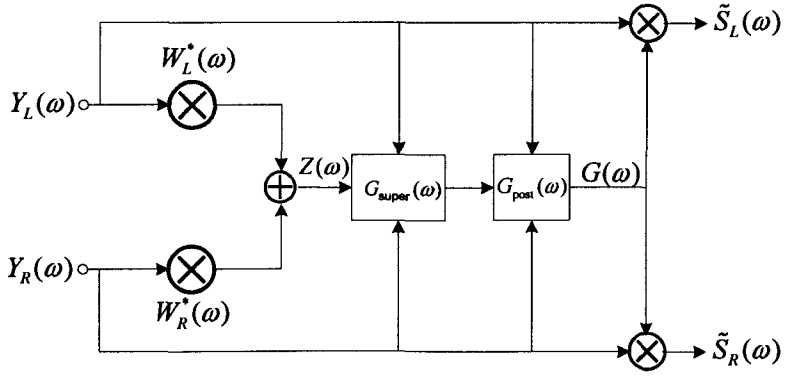

Figure 4-2: Binaural superdirective beamformer with post-filtering proposed in [LOT'06]

At first, the algorithm applies a beamforming scheme (i.e. $\mathbf{W}(\omega)=\left[W_L(\omega) \quad W_R(\omega)\right]^T$), and produces a monaural output $Z(\omega)$ as follows:

$$Z(\omega)=\mathbf{W^H}(\omega)\cdot\mathbf{Y}(\omega)=W_L^*(\omega)\cdot Y_L(\omega)+W_R^*(\omega)\cdot Y_R(\omega) \qquad \text{(EQ 4-71)}$$

However, a monaural output signal implies a complete loss of spatial hearing. Instead, here the output $Z(\omega)$ will serve as a reference for the calculation of common real-valued spectral gains $G_{\text{super}}(\omega)$ defined as:

$$G_{\text{super}}(\omega)=\frac{|Z(\omega)|}{|Y_L(\omega)|+|Y_R(\omega)|} \qquad \text{(EQ 4-72)}$$

Therefore, by having for each frequency a common real-valued gain applied to both left and right noisy input signals, this will guaranty the preservation of the interaural time and level differences in the binaural enhanced output signals. Consequently, spatial distortions will be minimized in the output signal. Furthermore, the algorithm imposes the constraint of having a distortionless response for the desired target source direction when applying $G_{\text{super}}(\omega)$ that is:

$$G_{\text{super}}(\theta_S,k)=1 \qquad \text{(EQ 4-73)}$$

where θ_S is the angle of arrival (i.e. azimuth) of the target speech source signal $S(\omega)$.

To achieve this, a modified Minimum Variance Distortionless Response (MVDR) constraint with a correction factor $corr_{\text{super}}(\theta_S,\omega)=|H_L(\theta_S,\omega)|+|H_R(\theta_S,\omega)|$ (based on Equation (4-72)) is used to initially compute the MVDR beamformer coefficients of $\mathbf{W}(\omega)$:

$$\min_{\mathbf{W}}\mathbf{W}^H(\omega)\cdot\mathbf{\Gamma}_{\text{NN}}(\omega)\cdot\mathbf{W}(\omega) \qquad \text{(EQ 4-74)}$$

with the constraint that

$$\mathbf{W}^H(\omega)\cdot\mathbf{H}(\theta_S,\omega)=corr_{\text{super}}(\theta_S,\omega) \qquad \text{(EQ 4-75)}$$

In Equation (4-75), $\mathbf{H}(\theta_{S,}\omega)=\left[H_L(\theta_S,\omega)\quad H_R(\theta_S,\omega)\right]^T$ is the propagation vector of the desired target source (i.e. consisting of the HRTFs between the target speech and the left and right hearing aid microphones respectively), that is:

$$\begin{bmatrix}S_L(\omega)\\S_R(\omega)\end{bmatrix}=\mathbf{H}(\theta_{S,}\omega)\cdot S(\omega)=\begin{bmatrix}H_L(\theta_S,\omega)\\H_R(\theta_S,\omega)\end{bmatrix}\cdot S(\omega) \qquad \text{(EQ 4-76)}$$

Solving (4-74) with the constraint in (4-75), the design of the superdirective vector $\mathbf{W}(\omega)$ with desired target source angle θ_S is the following:

$$\mathbf{W}(\omega)=\left(corr_{super}(\theta_S,\omega)\right)\cdot\mathbf{W}_{MVDR}(\omega) \qquad \text{(EQ 4-77)}$$

$$\text{where } \mathbf{W}_{MVDR}(\omega)=\frac{\left(\mathbf{\Gamma}_{\mathbf{NN}}^{-1}(\omega)+\mu_s\mathbf{I}\right)\mathbf{H}(\theta_{S,}\omega)}{\mathbf{H}^H(\theta_S,\omega)\left(\mathbf{\Gamma}_{\mathbf{NN}}^{-1}(\omega)+\mu_s\mathbf{I}\right)\mathbf{H}(\theta_S,\omega)} \qquad \text{(EQ 4-78)}$$

$\mathbf{W}_{MVDR}(\omega)$ is the "original" MVDR solution without correction factor in the constraint (i.e. $\mathbf{\Gamma}_{SS}(\omega)\mathbf{H}(\theta_{S,}\omega)=1$), where μ_s is a regularization factor that can also be seen as a tradeoff between directivity and robustness [LOT'06]. $\mathbf{\Gamma}_{\mathbf{NN}}(\omega)$ corresponds to a noise cross power-spectral density (cross-PSD) matrix defined as:

$$\mathbf{\Gamma}_{\mathbf{NN}}(\omega)=\begin{bmatrix}1 & \Gamma_{N_LN_R}(\omega)\\ \Gamma_{N_LN_R}(\omega) & 1\end{bmatrix} \qquad \text{(EQ 4-79)}$$

$\mathbf{\Gamma}_{\mathbf{NN}}(\omega)$ remains fixed (non-adaptive) and is optimized a for diffuse noise environment. Having $\mathbf{\Gamma}_{\mathbf{NN}}(\omega)$ equal to the cross-PSD matrix of a diffuse noise field, the filter solution maximizes the directivity index (or equivalently, the array gain). Therefore, the resulting system is called "superdirective" [BRA'01]. $\mathbf{\Gamma}_{\mathbf{N_LN_R}}(\omega)$ is computed in practice by averaging a number B equidistant HRTFs across the horizontal plane, $0\le\theta<2\pi$ as follows:

$$\Gamma_{N_LN_R}(\omega)=\frac{\sum_{n=1}^{B}H_L(\theta_n,\omega)\cdot H_R(\theta_n,\omega)}{\sqrt{\left(\sum_{n=1}^{B}\left|H_L(\theta_n,\omega)\right|^2\right)\cdot\left(\sum_{n=1}^{B}\left|H_R(\theta_n,\omega)\right|^2\right)}} \qquad \text{(EQ 4-80)}$$

In [LOT'06], an angular resolution of 5 degrees in the horizontal plane is used to compute equation (4-80), corresponding to $B = 72$.

Secondly, to further reduce the noise after initially enhancing with $G_{\text{super}}(\omega)$, another gain called the Post-filtering gain is applied, defined as :

$$G_{\text{post}}(\omega) = \frac{2 \cdot |Z(\omega)|^2}{|Y_L(\omega)|^2 + |Y_R(\omega)|^2} \cdot corr_{\text{post}}(\theta_n, \omega) \qquad \text{(EQ 4-81)}$$

where $corr_{\text{post}}(\theta_s, \omega)$ is again a correction factor to ensure that a distorsionless response in the target source direction is still maintained (i.e. $G_{\text{post}}(\theta_S, \omega) = 1$). $corr_{\text{post}}(\theta_s, \omega)$ is expressed as the following:

$$corr_{\text{post}}(\theta_n, \omega) = \frac{|H_L(\theta_S, \omega)|^2 + |H_R(\theta_S, \omega)|^2}{2 \cdot \left(|H_L(\theta_S, \omega)| + |H_R(\theta_S, \omega)|\right)^2} \qquad \text{(EQ 4-82)}$$

As a result, the final spectral gain per frequency $G(\omega)$ is obtained by taking the product of $G_{\text{super}}(\omega)$ and $G_{\text{post}}(\omega)$ as follows:

$$G(\omega) = G_{\text{super}}(\omega) \cdot G_{\text{post}}(\omega) \qquad \text{(EQ 4-83)}$$

The enhancement results obtained with this proposed binaural noise reduction scheme (with and without Post-filtering) will be evaluated with the results of other noise reduction schemes in Chapter 10. This scheme will be referred to as the Binaural Superdirective Beamformer with Post-filtering (BSBp) i.e. using $\text{BSBp} \triangleq G(\omega)$, and as the Binaural Superdirective Beamformer without Post-filtering (BSB) i.e. using $\text{BSB} \triangleq G_{\text{super}}(\omega)$.

4.3 Relation Between the Binaural Wiener Filter and MVDR-Constrained Beamforming

In this section, we will present how the Binaural Wiener Filter solution as derived in Section 4.1 can be decomposed into the product of a MVDR-constrained beamformer and a single channel post-filter.

If we take again a closer look at the actual form of the Binaural Wiener Filter solution, recalling Equation (4-21) (for clarity the Binaural Wiener Filter is relabeled as: $\mathbf{W}_{BWF}(\omega) \triangleq \mathbf{W}(\omega)$):

$$\mathbf{W}_{BWF}(\omega) = \mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega) \cdot \mathbf{\Gamma}_{\mathbf{YD}}(\omega) \qquad \text{(EQ 4-84)},$$

then assuming that the target speech and the noise are uncorrelated, we showed in Section 4.1 that Equation (4-84) can be simplified to:

$$\mathbf{W}_{BWF}(\omega) = \left(\mathbf{\Gamma}_{\mathbf{SS}}(\omega) + \mathbf{\Gamma}_{\mathbf{VV}}(\omega)\right)^{-1} \cdot \mathbf{\Gamma}_{\mathbf{SS}}(\omega) \qquad \text{(EQ 4-85)}$$

$\mathbf{\Gamma}_{\mathbf{SS}}(\omega)$ can be expanded using the HRTFs of the target source signal as follows:

$$\mathbf{\Gamma}_{\mathbf{SS}}(\omega) = \mathbf{\Gamma}_{\mathbf{SS}}(\omega) = \Gamma_{S,S_i}(\omega) \cdot \mathbf{H}(\omega) \cdot \mathbf{H}^{H}(\omega) \qquad \text{(EQ 4-86)}$$

where $\mathbf{H}(\omega)$ is equivalent to the propagation vector of the desired target source $\mathbf{H}(\theta_S, k) = \left[H_L(\theta_S, k) \quad H_R(\theta_S, k)\right]^T$ in (4-78) of Section 4.2.

Substitution Equation (4-86) into (4-85), $\mathbf{W}_{BWF}(\omega)$ can be written in another form as follows:

$$\mathbf{W}_{BWF}(\omega) = \left(\mathbf{\Gamma}_{\mathbf{VV}}(\omega) + \Gamma_{S,S_i}(\omega) \cdot \mathbf{H}(\omega) \cdot \mathbf{H}^{H}(\omega)\right)^{-1} \cdot \Gamma_{S,S_i}(\omega) \cdot \mathbf{H}(\omega) \cdot \mathbf{H}^{H}(\omega) \qquad \text{(EQ 4-87)}$$

We then apply a decomposition using the Matrix Inversion Lemma (or Sherman-Morrison-Woodbury Formual) [HEN'81] to $\left(\mathbf{\Gamma}_{\mathbf{VV}}(\omega) + \Gamma_{S,S_i}(\omega) \cdot \mathbf{H}(\omega) \cdot \mathbf{H}^{H}(\omega)\right)^{-1}$ in (4-87) as follows:

$$\left(\mathbf{A}+b\cdot\mathbf{u}\cdot\mathbf{v}^{H}\right)^{-1}=\mathbf{A}^{-1}-\frac{b}{1+b\cdot\mathbf{v}^{H}\cdot\mathbf{A}^{-1}\cdot\mathbf{u}}\mathbf{A}^{-1}\cdot\mathbf{u}\cdot\mathbf{v}^{H}\cdot\mathbf{A}^{-1} \quad \text{(EQ 4-88)}$$

with $\mathbf{A}=\mathbf{\Gamma}_{\mathbf{VV}}(\omega)$; $b=\Gamma_{S_1S_1}(\omega)$; $\mathbf{u}=\mathbf{H}(\omega)$; $\mathbf{v}^{H}=\mathbf{H}^{H}(\omega)$

$$\left(\mathbf{\Gamma}_{\mathbf{VV}}(\omega)+\Gamma_{S_1S_1}(\omega)\cdot\mathbf{H}(\omega)\cdot\mathbf{H}^{H}(\omega)\right)^{-1}=$$
$$\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)-\frac{\Gamma_{S_1S_1}(\omega)}{1+\Gamma_{S_1S_1}(\omega)\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)\cdot\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega) \quad \text{(EQ 4-89)}$$

Substituting Equation (4-89) into (4-87) and after simplification, $\mathbf{W}_{BWF}(\omega)$ is reduced to the following:

$$\mathbf{W}_{BWF}(\omega)=\frac{\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\Gamma_{S_1S_1}(\omega)\cdot\mathbf{H}(\omega)\cdot\mathbf{H}^{H}(\omega)}{1+\Gamma_{S_1S_1}(\omega)\cdot\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)} \quad \text{(EQ 4-91)}$$

Multiplying Equation (4-91) by $\frac{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}$ yields:

$$\mathbf{W}_{BWF}(\omega)=\frac{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\cdot\frac{\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\Gamma_{S_1S_1}(\omega)\cdot\mathbf{H}(\omega)\cdot\mathbf{H}^{H}(\omega)}{1+\Gamma_{S_1S_1}(\omega)\cdot\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)} \quad \text{(EQ 4-92)}$$

Re-arranging Equation (4-92), the following the new expression for $\mathbf{W}_{BWF}(\omega)$ is obtained:

$$\mathbf{W}_{BWF}(\omega)=\mathbf{W}_{MVDR}(\omega)\cdot\mathbf{H}^{H}(\omega)\cdot Post(\omega) \quad \text{(EQ 4-93)}$$

where $\mathbf{W}_{MVDR}(\omega)=\frac{\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}$ (EQ 4-94)

and $Post(\omega)=\frac{\Gamma_{S_1S_1}(\omega)}{\Gamma_{S_1S_1}(\omega)+\frac{1}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{\mathbf{VV}}^{-1}(\omega)\cdot\mathbf{H}(\omega)}}$ (EQ 4-95)

Therefore, it can be concluded that the Binaural Wiener Filter solution in Equation (4-93) is then the product of the "original" MVDR-constrained beamformer and a scalar post-filter $Post(\omega)$. The extra vector $\mathbf{H}^{H}(\omega)$ in Equation (4-93) is simply a factor that scales

the "original" MVDR beamformer output (i.e. $\mathbf{W}_{MVDR}^{H}(\omega)\cdot\mathbf{Y}(\omega)$) so that it would keep the same power as one of the input source signal.

Futhermore, the following will demonstrate that $Post(\omega)$ is in fact a single channel Wiener filter as well.

We first evaluate the output PSD of the desired target signal at the output of the MVDR beamformer as follows:

$$\Gamma_{S,S_1}^{MVDR}(\omega)=\mathbf{W}_{MVDR}^{H}(\omega)\cdot\mathbf{\Gamma}_{SS}(\omega)\cdot\mathbf{W}_{MVDR}(\omega)$$
$$=\left(\frac{\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\right)^{H}\cdot\Gamma_{S,S_1}(\omega)\cdot H(\omega)\cdot H^{H}(\omega)\cdot\left(\frac{\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\right)$$
$$=\Gamma_{S,S_1}(\omega)$$

(EQ 4-96)

The result in Equation (4-96) verifies the distorstionless power response in the desired target signal direction as imposed by the MVDR-constrained beamformer solution.

Secondly, we evaluate the output PSD of the noise signal at the output of the MVDR beamformer as follows:

$$\Gamma_{V,V_1}^{MVDR}(\omega)=\mathbf{W}_{MVDR}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}(\omega)\cdot\mathbf{W}_{MVDR}(\omega)$$
$$=\left(\frac{\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\right)^{H}\cdot\mathbf{\Gamma}_{VV}(\omega)\cdot\left(\frac{\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}\right)=\frac{1}{\mathbf{H}^{H}(\omega)\cdot\mathbf{\Gamma}_{VV}^{-1}(\omega)\cdot\mathbf{H}(\omega)}$$

(EQ 4-97)

Substituting Equations (4-96) and (4-97) into (4-95), $Post(\omega)$ is then expressed as follows:

$$Post(\omega)=\frac{\Gamma_{S,S_1}^{MVDR}(\omega)}{\Gamma_{S,S_1}^{MVDR}(\omega)+\Gamma_{V,V_1}^{MVDR}(\omega)}$$

(EQ 4-98)

The post-filter in equation (4-98) is equivalent to a single channel Wiener filter solution [BRA'01], that depends on the SNR output of the beamformer as follows:

$$Post(\omega) = \frac{SNR_{out}(\omega)}{1 + SNR_{out}(\omega)} \qquad \text{(EQ 4-99)}$$

where $SNR_{out}(\omega) = \frac{\Gamma_{S_s S_s}^{MVDR}(\omega)}{\Gamma_{V_s V_s}^{MVDR}(\omega)}$ (EQ 4-100).

To further interpret the decomposition result of the Binaural Wiener Filter in Equation (4-93), it can be said that the MVDR-beamformer component (i.e. $\mathbf{W}_{MVDR}(\omega)$) contributing in $\mathbf{W}_{BWF}(\omega)$ is data independent. In general, $\mathbf{W}_{MVDR}(\omega)$ is defined completely by the spatial configuration of the target signal and the noise signals. For instance, in [LOT'06] and also mentioned in Section 4.2, the noise cross-PSD matrix $\mathbf{\Gamma}_{\mathbf{VV}}(\omega)$ is replaced by a fixed noise cross-PSD matrix of a diffuse noise field. Hence, $\mathbf{W}_{MVDR}(\omega)$ provides then a spatial filtering. However, the post-Wiener filter in Equation (4-99) is data dependent since $\mathbf{\Gamma}_{\mathbf{SS}}(\omega)$ must be estimated in order to compute $SNR_{out}(\omega)$. In practice, $Post(\omega)$ provides further noise reduction.

Chapter 5. OVERVIEW OF OBJECTIVE MEASURES FOR NOISE REDUCTION PERFORMANCE EVALUATION

Various types of objective measures such as the Signal-to-Noise Ratio (SNR), the Segmental SNR (segSNR), the WideBand Perceptual Evaluation of Speech Quality (WB-PESQ), the Perceptual Similarity Measure (PSM) and the Coherence Speech Intelligibility Index (CSII) were used to evaluate the noise reduction performance of a variety of noise reduction schemes used in this thesis. In addition, three objective measures referred to as *composite objective measures* were also used to evaluate and compare the noise reduction schemes. They are referred to as the predicted rating of speech distortion (*Csig*), the predicted rating of background noise intrusiveness (*Cbak*) and the predicted rating of overall quality (*Covl*) as proposed in [HU'06]. A brief description of the different objective measures is presented below.

SNR: the SNR calculation is performed by taking the power of the noise-free target speech signal in the time-domain and dividing it by the power of the overall noise signal. The result is then converted into decibels.

SegSNR: the Segmental SNR measure proposed in [HAN'98] is the time-domain frame-based SNR. All the target speech and noise signals are first decomposed into frames.
In this work, each frame is windowed with a Hann window and a frame size of 30 ms is used with 75% overlap. The SegSNR is then computed by averaging the SNR estimates of each frame. Also as suggested in [HAN'98], the frames with SNRs above 35 dB are maintained to 35 dB (i.e. upper bound) since they do not reflect major perceptual differences. On the other hand, when the frames with SNR values become very negative due to small speech signal energies, the lower bound for those frames is set to -10 dB since they do not truly reflect the perceptual contributions of the signal.

WB-PESQ: PESQ was originally recommended by ITU-T under standard P.862.1 for speech quality assessment. It is designed to predict the subjective Mean Opinion Score

(MOS) of narrowband (3.1 kHz) handset telephony and narrowband speech coders [ITU'01]. In [HU'08b], a study was conducted to evaluate several quality measures for speech enhancement (i.e. PESQ, segmental SNR, frequency weighted SNR, Log-likelihood ratio, Itakura-Saito distance etc.). PESQ provided the highest correlation with subjective evaluations in terms of overall quality and signal distortion. PESQ scores are based on the MOS scale which is defined as follows: 5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 - Bad. Recently, ITU-T standardized the WB-PESQ (Wideband PESQ) under P.862.2, which is the extension of the model used in PESQ for wideband speech signals and operates at a sampling rate of 16 kHz [ITU'07].

PSM was proposed in [HUB'06] to estimate the perceptual similarity between the processed signal and the clean speech signal, in a way similar to the Perceptual Evaluation of Speech Quality (PESQ) [ITU'01]. PESQ was optimized for speech quality however, while PSM is also applicable to processed music and transients, thus also providing a prediction of perceived quality degradation for wideband audio signals [HUB'06], [ROH'05]. PSM has demonstrated high correlations between objective and subjective data and it has been used for quality assessment of noise reductions algorithms in [ROH'07],[ROH'05]. In terms of noise reduction evaluation, PSM is first obtained by using the unprocessed noisy signal and the target speech signal, and then by using the processed "enhanced" signal with the target speech signal. The difference between the two PSM results (referred to as ΔPSM) provides a noise reduction performance measure. A positive ΔPSM value indicates a higher quality obtained from the processed signal compared to the unprocessed one, whereas a negative value implies signal deterioration.

CSII was proposed in [KAT'05] as the extension of the speech intelligibility index (SII), which estimates speech intelligibility under conditions of additive stationary noise or bandwidth reduction. CSII further extends the SII concept to also estimate intelligibility in the occurrence of non-linear distortions such as broadband peak-clipping and center-clipping. To relate to our work, the non-linear distortion can also be caused by the result of de-noising or speech enhancement algorithms. The method first partitions the speech input signal into three amplitude regions (low-, mid- and high-level regions). The CSII

calculation is performed on each region (referred to as the three-level CSII) as follows: Each region is divided into short overlapping time segments of 16 ms to better consider fluctuating noise conditions. Then the signal-to-distortion ratio (SDR) of each segment is estimated, as opposed to the standard SNR estimate in the SII computation. The SDR is obtained using the mean-squared coherence function. The CSII result for each region is based on the weighed sum of the SDRs across the frequencies, similar to the frequency weighted SNR in the SII computation. Finally, the intelligibility is estimated from a linear weighted combination of the CSII results gathered from each region. It is stated in [KAT'05] that applying the three-level CSII approach and the fact that the SNR is replaced by the SDR provide much more information about the effects of the distortion on the speech signal. CSII provides a score between 0 and 1. A score of "1" represents a perfect intelligibility and a score of "0" represents a completely unintelligible signal.

The composite measures C*sig*, C*bak* and C*ovl* proposed in [HU'06] were obtained by combining numerous existing objective measures using nonlinear and nonparametric regression models, which provided much higher correlations with subjective judgments of speech quality and speech/noise distortions than conventional objective measures. For instance, the composite measure C*sig* is obtained by weighting and combining the Weighted-Slope Spectral (WSS) distance, the Log Likelihood Ratio (LLR) and the PESQ, and is represented by a five-point scale as follows: 5 – very natural, no degradation, 4 – fairly natural, little degradation, 3 – somewhat natural, somewhat degraded, 2 – fairly unnatural, fairly degraded, 1- very unnatural, very degraded. C*bak* combines segmental SNR, PESQ and WSS, and is represented by a five-point scale of background intrusiveness as follows: 5 – Not noticeable, 4 – Somewhat noticeable, 3 – Noticeable but not intrusive, 2 – Fairly conspicuous, somewhat intrusive, 1 – Very conspicuous, very intrusive. Finally, C*ovl* combines PESQ, LLR and WSS, and it uses the scale of the mean opinion score (MOS) as follows: 5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 - Bad.

It should be noted that more recent updated composite measures were proposed in [HU'08b], further extending the results in [HU'06] in terms of objective measure

selections and weighting rules. However, they were not used in the work of this thesis since the updated composite measures were selected and optimized in environments with higher SNR/PESQ levels than the SNR/PESQ levels in this work. Therefore, the original composite measures from [HU'06] were used. Moreover, the correlation of the original composite measures with subjective results were optimized for signals sampled at 8kHz. Therefore, in our work, to properly get the assessments from the *Csig*, *Cbak* and *Covl* composite measures the simulation signals (after processing) were downsampled from 20kHz to 8kHz. However, the CSII and WB-PESQ objective measures were applied for wideband speech signals downsampled to 16kHz.

To sum up, the SNR, SegSNR, *Covl* , PSM and WB-PESQ measures will provide feedback regarding the overall quality of the signal after processing, *Cbak* will provide feedback about the distortions that affect the background noise (i.e. noise distortion/noise intrusiveness), *Csig* will give information about the distortions that impinges on the target speech signal itself (i.e. signal distortion), whereas the CSII measure will indicate the potential speech intelligibility improvement of the processed speech versus the noisy unprocessed speech signal. It should be noted that all our objective measures compare the output after processing with a target noise-free input speech signal. However, the input signal can be reverberant if the hearing scenario is a reverberant condition. We thus define the "noise-free" input signal as the input signal without the additions of diffuse babble noise and directional interfering noises, but the input speech signal is still in a reverberant condition if the hearing scenario tested contains reverberation.

Chapter 6. PROPOSED BINAURAL DIFFUSE NOISE POWER SPECTRUM DENSITY ESTIMATOR

In this chapter, an improved diffuse noise power spectrum density estimator is developed, targeting future binaural hearing aids. It will be shown in the next chapter that the proposed technique can outperform the advanced noise estimators previously described in Chapter 3. The proposed technique requires a binaural system, which is equivalent to a two-channel system where the processing is performed on two received noisy speech signals, picked up from the left and right hearing aid microphones. The noise power spectrum density estimation is designed for a diffuse noise field environment such as babble talk in a crowded cafeteria. Therefore, the characteristics of noise can be quite non-stationary and colored. This proposed estimator does not rely on voice activity detectors and will be shown to provide a more accurate noise estimation at low SNRs and also during speech activity. Furthermore, the target speaker can be located anywhere around the binaural hearing aid user, that is without any assumption about speaker location or direction of target speech arrival.

Section 6.1 will explain the selected acoustical environment where the noise power spectrum density is estimated. Section 6.2 will describe the binaural system adapted to this environment, with signal definitions. Finally, in Section 6.3 the proposed binaural noise spectrum estimator will be presented in detail.

6.1 Acoustical Environment: Diffuse Noise Field

For a hearing aid user, listening to a nearby target speaker in a diffuse noise field as in Fig. 6-1 is a common environment encountered in many typical noisy situations i.e. the babble noise in an office or a cafeteria, the engine noise and the wind blowing in a car, etc. [MCC'03],[GUE'03],[HAM'02],[DOE'96]. Considering the situation of a person

being in a diffuse noise field environment, the two ears would receive noise signals that have propagated from all directions, originally with equal amplitude and a random phase [ABU'04]. In the literature, a diffuse noise field has also been defined as uncorrelated noise signals of equal power propagating in all directions simultaneously [MCC'03]. A diffuse noise field assumption has been proven to be a suitable model for a number of practical reverberant noise environments often encountered in speech enhancement applications [MEY'97] [BIT'99],[HAM'02],[MCC'03],[CAM'03] and it has often been applied in array processing such as in superdirective beamformers [ELK'00]. It has been observed through empirical results that a diffuse noise field exhibits a high coherence (i.e. high correlation) at low frequencies and a low coherence over the remaining frequency spectrum. However, it is different from a localized noise source, where a dominant noise source is coming from a specific direction. Most importantly, with the occurrence of a localized noise source or directional noise, the noise signals received by the left and right microphones become highly correlated over most of the frequency content of the noise signals. The proposed binaural noise spectrum estimator has been designed for binaural hearing aids operating in a diffuse noise field environment.

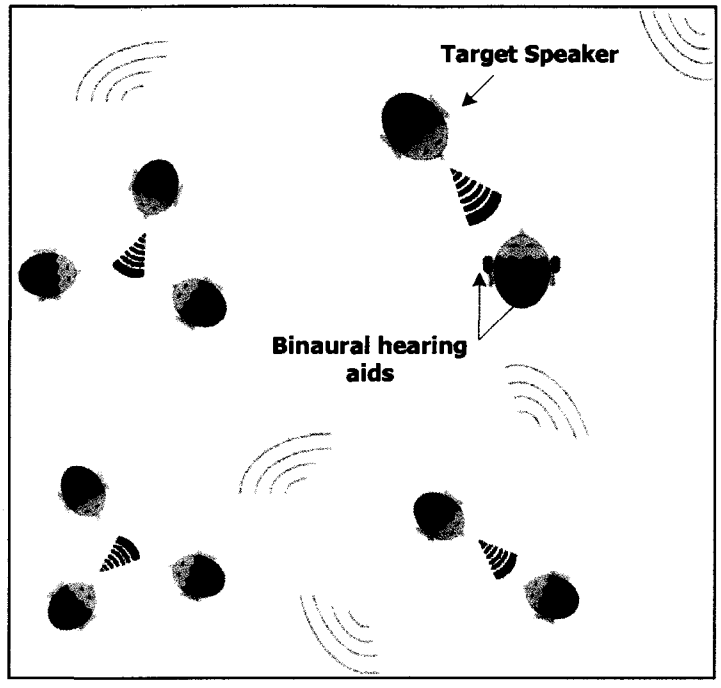

Figure 6-1: Binaural hearing aids user in a noisy diffuse field environment with a single target source speaker

6.2 Binaural System Description

Let $l(i)$, $r(i)$ be the noisy signals received at the left and right hearing aid microphones, defined here in the temporal domain as:

$$l(i) = s(i) \otimes h_l(i) + n_l(i) \quad \text{(EQ 6-1)}$$

$$r(i) = s(i) \otimes h_r(i) + n_r(i) \quad \text{(EQ 6-2)}$$

where $s(i)$ is the target source speech signal and $\otimes$ represents a linear convolution sum operation. $n_l(i)$ and $n_r(i)$ are respectively the left and right received additive noise signals.

It is assumed for now that the distance between the speaker and the two microphones (one placed on each ear) is such that they receive essentially speech through a direct path from the nearby speaker, implying that the received left and right signals are highly correlated (i.e. the direct component dominates its reverberation components). Hence, the left and right received signals can be modeled by left and right impulse responses, $h_l(i)$ and $h_r(i)$, convolved with the target source speech signal. In the context of binaural hearing, those impulse responses are often referred to as the left and right head-related impulse responses (HRIRs) between the target speaker and the left and right hearing aids microphones.

Prior to estimating the noise power spectrum, the following assumptions are made (comparable to [DOE'96] as in Section 3.1.2):

a) the target speech and noise signals are uncorrelated, and the hearing aid user is in a diffuse noise field environment as described in Section 6.1.

b) $n_l(i)$ and $n_r(i)$ are mutually uncorrelated, which is a known characteristic of a diffuse noise field, except at very low frequencies [DOE'96],[CAM'03]. In fact, neglecting this high correlation at low frequencies will lead to an underestimation of the noise power spectrum density at low frequencies. The noise power estimator in [DOE'96] previously explained in Section 3.1.2 suffers from this effect [HAM'02]. However, this low correlation will be taken into consideration in Section 6.3.5, by adjusting the proposed noise estimator with a compensation method for the low frequencies. But in this section, uncorrelated left and right noise are assumed for now, over the entire frequency spectrum.

c) the left and right noise power spectral densities are considered approximately equal, that is: $\Gamma_{N_L N_L}(\omega) \approx \Gamma_{N_R N_R}(\omega) \approx \Gamma_{NN}$. This approximation is again a realistic characteristic of diffuse noise fields, and it has been verified from experimental recordings.

Additionally, as opposed to [DOE'96], the target speaker can be anywhere around the hearing user, that is the direction of arrival of the target speech signal does not need to be frontal (azimuthal angle $\neq 0°$).

Using the assumptions above and along with Equations (6-1) and (6-2), the left and right auto power spectral densities, $\Gamma_{LL}(\omega)$ and $\Gamma_{RR}(\omega)$, can be expressed as the following:

$$\Gamma_{LL}(\omega) = F.T.\{\gamma_{ll}(\tau)\} = \Gamma_{SS}(\omega)\,|\,H_L(\omega)\,|^2 + \Gamma_{NN}(\omega) \qquad \text{(EQ 6-3)}$$

$$\Gamma_{RR}(\omega) = F.T.\{\gamma_{rr}(\tau)\} = \Gamma_{SS}(\omega)\,|\,H_R(\omega)\,|^2 + \Gamma_{NN}(\omega) \qquad \text{(EQ 6-4)}$$

where $\gamma_{yx}(\tau) = E[y(i+\tau)\cdot x(i)]$ represents a statistical correlation function.

6.3 Proposed Binaural Diffuse Noise Power Spectrum Estimation

In this section, the proposed binaural diffuse noise power spectrum density estimation will be developed. Section 6.3.1 will present the overall diagram of the proposed noise power spectrum estimation. It will be shown that the noise spectrum estimation is found by applying first a Wiener filter to perform a prediction of the left noisy speech signal from the right noisy speech signal, followed by taking the auto-power spectral density of the difference between the left noisy signal and its prediction. As a second step, a quadratic equation is formed by combining 1) the auto-power spectral density of the previous difference signal, with 2) the auto-power spectral densities of the left and right noisy speech signals. As a result, the solution of the quadratic equation provides the power spectral density of the noise. In practice, the estimation error on one of the variables used in the quadratic system causes the noise power spectrum estimation to be less accurate. This is because the estimated value of this variable is computed indirectly i.e. it is obtained from a combination of several other variables. However, Section 6.3.2 will show that there is also an alternative and direct way to compute the value of this variable, which is less intuitive but provides a better accuracy in practice. Therefore,

solving the quadratic equation by using the direct computation of this variable will give a better noise spectrum estimation. The direct and indirect computations of this variable will be compared analytically in Section 6.3.3 and experimentally in Section 6.3.4, explaining the differences in practice. Finally, Section 6.3.5 will show how to adjust the noise spectrum estimator at low frequencies to account for the low-frequency coherence found in diffuse noise field environments.

6.3.1 Diffuse Noise PSD Estimation

Figure 6-2 shows a diagram of the overall proposed estimation method including a Wiener prediction filter and the final quadratic equation estimating the diffuse noise power spectral density.

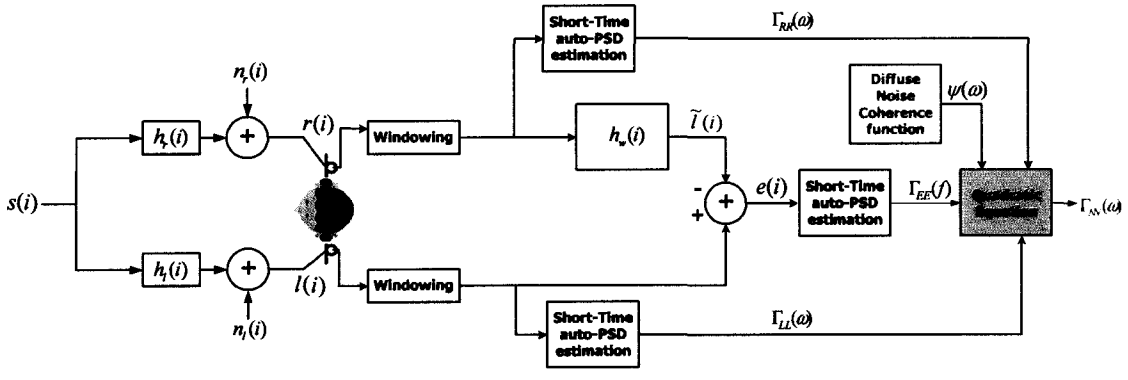

Figure 6-2: Overall diagram for proposed binaural noise estimation method

In a first step, a filter $h_w(i)$ is used to perform a linear prediction of the left noisy speech signal from the right noisy speech signal.

Using a minimum mean square error criterion (MMSE), the optimum solution is the Wiener solution, defined here in the frequency domain as:

$$H_W(\omega) = \Gamma_{LR}(\omega)/\Gamma_{RR}(\omega) \qquad \text{(EQ 6-5)}$$

where $\Gamma_{LR}(\omega)$ is cross-power spectral density between the left and the right noisy signals.

$\Gamma_{LR}(\omega)$ is obtained as follows:

$$\Gamma_{LR}(\omega) = F.T.\{\gamma_{lr}(\tau)\} = F.T.\{E[l(i+\tau)\cdot r(i)]\} \quad \text{(EQ 6-6)}$$

with $\gamma_{lr}(\tau) = E\left(\left[s(i+\tau)\otimes h_l(i) + n_l(i+\tau)\right]\cdot\left[s(i)\otimes h_r(i) + n_r(i)\right]\right)$

$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau) + \gamma_{sn_r}(\tau)\otimes h_l(\tau) + \gamma_{n_l s}(\tau)\otimes h_r(-\tau) + \gamma_{n_l n_r}(\tau) \quad \text{(EQ 6-7)}$$

Using the previously defined assumptions in Section 6.2, Equation (6-7) can then be simplified to:

$$\gamma_{lr}(\tau) = \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau) \quad \text{(EQ 6-8)}$$

The cross-power spectral density expression then becomes:

$$\Gamma_{LR}(\omega) = \Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega) \quad \text{(EQ 6-9)}$$

Therefore, substituting (6-9) into (6-5) yields:

$$H_W(\omega) = \Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)/\Gamma_{RR}(\omega) \quad \text{(EQ 6-10)}$$

Furthermore, using (6-3) and (6-4), the squared magnitude response of the Wiener filter in (6-10) can also be expressed as:

$$|H_W(\omega)|^2 = \frac{\left(\Gamma_{LL}(\omega) - \Gamma_{NN}(\omega)\right)\cdot\left(\Gamma_{RR}(\omega) - \Gamma_{NN}(\omega)\right)}{\Gamma_{RR}^2(\omega)} \quad \text{(EQ 6-11)}$$

For the second step of the noise estimation algorithm, Equation (6-11) is rearranged into a quadratic equation as the following:

$$\Gamma_{NN}{}^2(\omega) - \Gamma_{NN}(\omega)\cdot\left(\Gamma_{LL}(\omega) + \Gamma_{RR}(\omega)\right) + \Gamma_{EE\_1}(\omega)\cdot\Gamma_{RR}(\omega) = 0 \quad \text{(EQ 6-12)}$$

where $\Gamma_{EE\_1}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{RR}(\omega)\cdot|H_W(\omega)|^2$ (EQ 6-13)

Consequently, the noise power spectral density $\Gamma_{NN}(\omega)$ can be estimated by solving the quadratic equation in (6-12), which will produce two solutions:

$$\Gamma_{NN}(\omega)=\frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)\pm\Gamma_{LRavg}(\omega) \qquad \text{(EQ 6-14)}$$

where $\Gamma_{LRavg}(\omega)=\frac{1}{2}\sqrt{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2-4\cdot\Gamma_{EE\_1}(\omega)\cdot\Gamma_{RR}(\omega)}$ (EQ 6-15)

Below we demonstrate that $\Gamma_{LRavg}(\omega)$ in Equation (6-15) is equivalent to the average of the left and right noise-free speech power spectral densities. Consequently, the "negative root" in (6-14) will be the one leading to the correct estimation for $\Gamma_{NN}(\omega)$.

Substituting (6-13) into (6-15) yields:

$$\begin{aligned}\Gamma_{LRavg}(\omega)&=\frac{1}{2}\sqrt{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2-4\cdot\left(\Gamma_{LL}(\omega)-\Gamma_{RR}(\omega)\cdot|H_W(\omega)|^2\right)\cdot\Gamma_{RR}(\omega)}\\&=\frac{1}{2}\sqrt{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2-4\cdot\left(\Gamma_{LL}(\omega)\cdot\Gamma_{RR}(\omega)-\Gamma_{RR}{}^2(\omega)\cdot|H_W(\omega)|^2\right)}\end{aligned} \qquad \text{(EQ 6-16)}$$

Substituting (6-11) into (6-16) yields:

$$\Gamma_{LRavg}(\omega)=\frac{1}{2}\sqrt{\begin{array}{l}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2\\-4\cdot\left(\Gamma_{LL}(\omega)\cdot\Gamma_{RR}(\omega)-\left(\left(\Gamma_{LL}(\omega)-\Gamma_{NN}(\omega)\right)\cdot\left(\Gamma_{RR}(\omega)-\Gamma_{NN}(\omega)\right)\right)\right)\end{array}} \qquad \text{(EQ 6-17)}$$

After a few simplifications, the following is obtained:

$$\begin{aligned}\Gamma_{LRavg}(\omega)&=\frac{1}{2}\sqrt{\left(\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-2\cdot\Gamma_{NN}(\omega)\right)^2}\\&=\frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)-2\cdot\Gamma_{NN}(\omega)\right)\end{aligned} \qquad \text{(EQ 6-18)}$$

As expected, looking at Equation (6-18), $\Gamma_{LRavg}(\omega)$ is equal to the average of the left and right noise-free speech power spectral densities. Consequently, substituting (6-18) into (6-14), it can easily be noticed that only the "negative root" will lead to the correct solution for $\Gamma_{NN}(\omega)$ as the following:

$$\begin{aligned}\Gamma_{NN}(\omega) &= \frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\Gamma_{LRavg}(\omega) \\ &= \frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)-2\cdot\Gamma_{NN}(\omega)\right) \\ &= \Gamma_{NN}(\omega)\end{aligned} \qquad \text{(EQ 6-19)}$$

Consequently, the noise power spectral density can be determined at this moment using the following expressions:

$$\Gamma_{NN}(\omega) = \frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\Gamma_{LRavg}(\omega) \qquad \text{(EQ 6-20)}$$

$$\text{where } \Gamma_{LRavg}(\omega) = \frac{1}{2}\sqrt{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2-4\cdot\Gamma_{EE\_1}(\omega)\cdot\Gamma_{RR}(\omega)} \qquad \text{(EQ 6-21)}$$

$$\text{and} \quad \Gamma_{EE\_1}(\omega) = \Gamma_{LL}(\omega)-\Gamma_{RR}(\omega)\cdot|H_W(\omega)|^2 \qquad \text{(EQ 6-22)}$$

However, using $\Gamma_{EE\_1}(\omega)$, or more explicitly Equation (6-22) in that current form, does not yield an accurate estimate of $\Gamma_{NN}(\omega)$ in practice, as briefly explained at the beginning of Section 6.3. It will be shown in Section 6.3.2 that $\Gamma_{EE\_1}(\omega)$ is in fact the auto-power spectral density of the prediction residual (or error) $e(i)$ shown in Fig. 6-2. The direct computation of this auto-power spectral density from the samples of $e(i)$ will be referred to as $\Gamma_{EE}(\omega)$, while the indirect computation using (6-22) is referred to as $\Gamma_{EE\_1}(\omega)$. $\Gamma_{EE\_1}(\omega)$ and $\Gamma_{EE}(\omega)$ are theoretically equivalent, however only estimates of the different power spectral densities are available in practice to compute (6-5), (6-20), (6-21) and (6-22), and the resulting estimation of $\Gamma_{NN}(\omega)$ in (6-20) is not as accurate if $\Gamma_{EE\_1}(\omega)$ is used. This is because the difference between the true and the estimated Wiener solutions for (6-5) can lead to large fluctuations in $\Gamma_{EE\_1}(\omega)$, when evaluated using (6-22). As opposed to $\Gamma_{EE\_1}(\omega)$, the direct estimation of $\Gamma_{EE}(\omega)$ is not subject to those large fluctuations. The direct and indirect computations of this variable have been compared analytically and experimentally in Sections 6.3.3 and 6.3.4, by taking into

consideration a non-ideal (i.e. estimated) Wiener solution. It is will shown that using the direct computation yields a much greater accuracy in terms of the noise PSD estimation.

6.3.2 Direct Computation of the Error Auto-Power Spectrum

This section will demonstrate that $\Gamma_{EE\_1}(\omega)$ is also the auto-power spectral density of the prediction residual (or error) $e(i)$ represented in Fig. 6-2. It will also finalize the proposed algorithm for estimating the noise PSD in a diffuse noise field environment.

The prediction residual error is defined as:

$$e(i) = l(i) - \tilde{l}(i) \quad \text{(EQ 6-23)}$$

$$= l(i) - r(i) \otimes h_w(i) \quad \text{(EQ 6-24)}$$

As previously mentioned in Section 6.3.1, the direct computation of this auto-power spectral density from the samples of $e(i)$ is referred to as $\Gamma_{EE}(\omega)$ and the indirect computation using (6-22) is referred to as $\Gamma_{EE\_1}(\omega)$.

From Fig. 6-2 and the definition of $e(i)$, we have:

$$\Gamma_{EE}(\omega) = F.T.\left(\gamma_{ee}(\tau)\right) \quad \text{(EQ 6-25)}$$

where $\gamma_{ee}(\tau) = E\left(e(i+\tau)\cdot e(i)\right)$

$$= E\left(\left[l(i+\tau) - \tilde{l}(i+\tau)\right]\cdot\left[l(i) - \tilde{l}(i)\right]\right)$$

$$= E\left[l(i+\tau)l(i)\right] - E\left[l(i+\tau)\tilde{l}(i)\right] - \left[E\tilde{l}(i+\tau)l(i)\right] + E\left[\tilde{l}(i+\tau)\tilde{l}(i)\right]$$

$$= \gamma_{ll}(\tau) - \gamma_{l\tilde{l}}(\tau) - \gamma_{\tilde{l}l}(\tau) + \gamma_{\tilde{l}\tilde{l}}(\tau) \quad \text{(EQ 6-26)}$$

As seen in (6-26), $\gamma_{ee}(\tau)$ is thus the sum of 4 terms, where the following temporal and frequency domain definitions for each term are:

$$\gamma_{ll}(\tau) = E\left(\left[s(i+\tau)\otimes h_l(i) + n_l(i+\tau)\right]\cdot\left[s(i)\otimes h_l(i) + n_l(i)\right]\right)$$

$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_l(-\tau) + \gamma_{nn}(\tau) \quad \text{(EQ 6-27)}$$

$$\Gamma_{LL}(\omega) = \Gamma_{SS}(\omega)\,|\,H_L(\omega)\,|^2 + \Gamma_{NN}(\omega) \qquad \text{(EQ 6-28)}$$

$$\gamma_{l\tilde{l}}(\tau) = E\Big(\big[s(i+\tau)\otimes h_l(i) + n_l(i+\tau)\big]\cdot\big[\big[s(i)\otimes h_r(i) + n_r(i)\big]\otimes h_w(i)\big]\Big)$$

$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau)\otimes h_W(-\tau) \qquad \text{(EQ 6-29)}$$

$$\Gamma_{L\tilde{L}}(\omega) = \Gamma_{SS}(\omega)H_L(\omega)H_R^*(\omega)H_W^*(\omega) \qquad \text{(EQ 6-30)}$$

$$\gamma_{\tilde{l}l}(\tau) = E\Big(\big[\big[s(i+\tau)\otimes h_r(i) + n_r(i+\tau)\big]\otimes h_w(i)\big]\cdot\big[s(i)\otimes h_l(i) + n_l(i)\big]\Big)$$

$$= \gamma_{ss}(\tau)\otimes h_l(-\tau)\otimes h_r(\tau)\otimes h_W(\tau) \qquad \text{(EQ 6-31)}$$

$$\Gamma_{\tilde{L}L}(\omega) = \Gamma_{SS}(\omega)H_L^*(\omega)H_R(\omega)H_W(\omega) \qquad \text{(EQ 6-32)}$$

$$\gamma_{\tilde{l}\tilde{l}}(\tau) = E\left(\begin{matrix}\big[\big[s(i+\tau)\otimes h_r(i) + n_r(i+\tau)\big]\otimes h_w(i)\big] \\ \cdot\big[\big[s(i)\otimes h_r(i) + n_r(i)\big]\otimes h_w(i)\big]\end{matrix}\right)$$

$$= \gamma_{ss}(\tau)\otimes h_r(\tau)\otimes h_r(-\tau)\otimes h_W(\tau)\otimes h_W(-\tau) + \gamma_{nn}(\tau)\otimes h_W(\tau)\otimes h_W(-\tau) \qquad \text{(EQ 6-33)}$$

$$\Gamma_{\tilde{L}\tilde{L}}(\omega) = \Gamma_{SS}(\omega)\,|\,H_R(\omega)\,|^2\,|\,H_W(\omega)\,|^2 + \Gamma_{NN}(\omega)\,|\,H_W(\omega)\,|^2 \qquad \text{(EQ 6-34)}$$

From Equation (6-26), we can write:

$$\Gamma_{EE}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{L\tilde{L}}(\omega) - \Gamma_{\tilde{L}L}(\omega) + \Gamma_{\tilde{L}\tilde{L}}(\omega) \qquad \text{(EQ 6-35)}$$

and substituting all the terms in their respective frequency domain forms (i.e. (6-28),(6-30), (6-32) and (6-34) into (6-35)), yields:

$$\Gamma_{EE}(\omega) = \Gamma_{NN}(\omega)\cdot(1+|\,H_W(\omega)\,|^2) + \Gamma_{SS}(\omega)\cdot|\,H_L(\omega)\,|^2 - 2\cdot\Gamma_{SS}(\omega)\cdot Re\Big(H_L^*(\omega)\cdot H_R(\omega)\cdot H_W(\omega)\Big) + \Gamma_{SS}(\omega)\cdot|\,H_R(\omega)\,|^2\cdot|\,H_W(\omega)\,|^2 \qquad \text{(EQ 6-36)}$$

Multiplying both sides of (6-10) by $H_L^*(\omega)\cdot H_R(\omega)$, it can be observed that $H_L^*(\omega)H_R(\omega)\,H_W(\omega)$ in (6-36) is in fact real-valued:

$$H_L^*(\omega)H_R(\omega)\,H_W(\omega) = \frac{\Gamma_{SS}(\omega)\big|H_L(\omega)\big|^2\big|H_R(\omega)\big|^2}{\Gamma_{RR}(\omega)} = \mathrm{Re}\Big\{H_L^*(\omega)H_R(\omega)\,H_W(\omega)\Big\} \qquad \text{(EQ 6-37)}$$

Using (6-3) and (6-4), the latter becomes:

$$\Gamma_{SS}(\omega)\,\mathrm{Re}\left\{H_L^*(\omega)H_R(\omega)\,H_W(\omega)\right\}=\frac{\left(\Gamma_{LL}(\omega)-\Gamma_{NN}(\omega)\right)\left(\Gamma_{RR}(\omega)-\Gamma_{NN}(\omega)\right)}{\Gamma_{RR}(\omega)} \quad \text{(EQ 6-38)}$$

and using (6-11), we then obtain:

$$\Gamma_{SS}(\omega)\,\mathrm{Re}\left\{H_L^*(\omega)H_R(\omega)\,H_W(\omega)\right\}=\Gamma_{RR}(\omega)\,|H_W(\omega)|^2 \quad \text{(EQ 6-39)}$$

Using (6-39) into (6-36), and applying (6-3) and (6-4) again yields after simplification:

$$\begin{aligned}\Gamma_{EE}(\omega)&=\Gamma_{NN}(\omega)\cdot(1+|H_W(\omega)|^2)+\Gamma_{LL}(\omega)-\Gamma_{NN}(\omega)-2\cdot\Gamma_{RR}(\omega)\cdot|H_W(\omega)|^2\\&\quad+\left(\Gamma_{RR}(\omega)-\Gamma_{NN}(\omega)\right)\cdot|H_W(\omega)|^2\\&=\Gamma_{LL}(\omega)-\Gamma_{RR}(\omega)\cdot|H_W(\omega)|^2 \quad \text{(EQ 6-40)}\end{aligned}$$

We find that (6-40) is identical to (6-22), and thus $\Gamma_{EE\_1}(\omega)$ in (6-22) represents the auto-PSD of $e(i)$.

To sum up, an estimate for $\Gamma_{EE}(\omega)$ computed directly from the signal $e(i)$ as depicted in Fig. 6-2 is to be used in practice instead of estimating $\Gamma_{EE\_1}(\omega)$ indirectly through (6-22). Consequently, replacing $\Gamma_{EE\_1}(\omega)$ by $\Gamma_{EE}(\omega)$ in (6-21), the proposed noise estimation algorithm is obtained.

As a result, the noise power spectral density can be found using the following equations:

$$\Gamma_{NN}(\omega)=\frac{1}{2}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\Gamma_{LRavg}(\omega) \quad \text{(EQ 6-41)}$$

$$\text{where } \Gamma_{LRavg}(\omega)=\frac{1}{2}\sqrt{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)^2-4\cdot\Gamma_{EE}(\omega)\cdot\Gamma_{RR}(\omega)} \quad \text{(EQ 6-42)}$$

$$\text{and} \quad \Gamma_{EE}(\omega)=F.T.\left(\gamma_{ee}(\tau)\right)=F.T.\left\{E\left(e(i+\tau)\cdot e(i)\right)\right\} \quad \text{(EQ 6-43)}$$

6.3.3 The Direct Computation $\Gamma_{EE}(\omega)$ Versus the Indirect Computation $\Gamma_{EE\_1}(\omega)$

As briefly introduced in Section 6.3, the difference between the true and the estimated Wiener solutions for (6-5) can lead to large fluctuations in $\Gamma_{EE\_1}(\omega)$ when evaluated using (6-22). However, the direct estimate of $\Gamma_{EE}(\omega)$ (i.e. by taking the auto-power spectral density of the prediction $e(i)$) is not subject to those large fluctuations, as it will be demonstrated analytically here and experimentally in the next section.

For instance, instead of having the true Wiener filter solution in (6-5), in practice we can have the following model:

$$\tilde{h}_W(i) = h_W(i) + h_e(i) \qquad \text{(EQ 6-44)}$$

where $\tilde{h}_W(i)$ is the result of using a practical adaptive filter or a least-squares estimation (i.e. the approximation of the true Wiener filter solution), and $h_e(i)$ can be seen as a residual random error signal associated to each Wiener filter coefficient.

The Fourier transform of (6-44) is then:

$$\tilde{H}_W(\omega) = H_W(\omega) + H_E(\omega) \qquad \text{(EQ 6-45)}$$

The corresponding magnitude squared response can be expressed as follows:

$$|\tilde{H}_W(\omega)|^2 = \tilde{H}_W(\omega)\cdot\tilde{H}_W^*(\omega)$$

$$= |H_W(\omega)|^2 + H_W(\omega)\cdot H_E^*(\omega) + H_W^*(\omega)\cdot H_E(\omega) + |H_E(\omega)|^2 \qquad \text{(EQ 6-46)}$$

In the first step, the error component that will appear in the indirect computation of $\Gamma_{EE\_1}(\omega)$ evaluated using (6-22) will be found. Replacing $|H_W(\omega)|^2$ by $|\tilde{H}_W(\omega)|^2$ into (6-22) yields:

$$\tilde{\Gamma}_{EE\_1}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{RR}(\omega)\cdot|\tilde{H}_W(\omega)|^2 \qquad \text{(EQ 6-47)}$$

$$= \Gamma_{LL}(\omega) - \Gamma_{RR}(\omega) \cdot \begin{pmatrix} |H_W(\omega)|^2 + H_W(\omega) \cdot H_E^*(\omega) \\ + H_W^*(\omega) \cdot H_E(\omega) + |H_E(\omega)|^2 \end{pmatrix} \quad \text{(EQ 6-48)}$$

$$= \Gamma_{EE\_1}(\omega) + \varepsilon_{EE\_1\_err}(\omega) \quad \text{(EQ 6-49)}$$

where the analytical error component in the indirect computation in the frequency domain is:

$$\varepsilon_{EE\_1\_err}(\omega) = -\left(H_W(\omega) \cdot H_E^*(\omega) + H_W^*(\omega) \cdot H_E(\omega) + |H_E(\omega)|^2 \right) \cdot \Gamma_{RR}(\omega) \quad \text{(EQ 6-50)}$$

In the second step, the error component appearing in the direct computation of $\Gamma_{EE}(\omega)$ evaluated using (6-43) is found. Replacing $H_W(\omega)$ by $\tilde{H}_W(\omega)$ and $|H_W(\omega)|^2$ by $|\tilde{H}_W(\omega)|^2$ into (6-36) yields:

$$\tilde{\Gamma}_{EE}(\omega) = \Gamma_{NN}(\omega) \cdot \left(1 + |\tilde{H}_W(\omega)|^2\right) + \begin{pmatrix} \Gamma_{SS}(\omega) \cdot |H_L(\omega)|^2 \\ -2 \cdot \Gamma_{SS}(\omega) \cdot Re\left(H_L^*(\omega) \cdot H_R(\omega) \cdot \tilde{H}_W(\omega)\right) \\ + \Gamma_{SS}(\omega) \cdot |H_R(\omega)|^2 \cdot |\tilde{H}_W(\omega)|^2 \end{pmatrix} \quad \text{(EQ 6-51)}$$

$$= \Gamma_{NN}(\omega) \cdot \begin{pmatrix} 1 + |H_W(\omega)|^2 + H_W(\omega) \cdot H_E^*(\omega) \\ + H_W^*(\omega) \cdot H_E(\omega) + |H_E(\omega)|^2 \end{pmatrix} + \begin{pmatrix} \Gamma_{SS}(\omega) \cdot |H_L(\omega)|^2 \\ -2 \cdot \Gamma_{SS}(\omega) \cdot Re\left(H_L^*(\omega) \cdot H_R(\omega) \cdot \left(H_W(\omega) + H_E(\omega)\right)\right) \\ + \Gamma_{SS}(\omega) \cdot |H_R(\omega)|^2 \\ \cdot \begin{pmatrix} |H_W(\omega)|^2 + H_W(\omega) \cdot H_E^*(\omega) \\ + H_W^*(\omega) \cdot H_E(\omega) + |H_E(\omega)|^2 \end{pmatrix} \end{pmatrix} \quad \text{(EQ 6-52)}$$

After simplification, and grouping the terms belonging to $\Gamma_{EE}(\omega)$ from (6-36), we obtain the following:

$$\tilde{\Gamma}_{EE}(\omega) = \Gamma_{EE}(\omega) + \varepsilon_{EE\_err}(\omega) \quad \text{(EQ 6-53)}$$

where the analytical error component in the direct computation is obtained as:

$$\varepsilon_{EE\_err}(\omega) = \left(H_W(\omega)\cdot H_E^*(\omega) + H_W^*(\omega)\cdot H_E(\omega) + |H_E(\omega)|^2\right)\cdot\left(\Gamma_{SS}(\omega)|H_R(\omega)|^2 + \Gamma_{NN}(\omega)\right) - 2\cdot\Gamma_{SS}(\omega)\cdot \mathrm{Re}\left(H_L^*(\omega)\cdot H_R(\omega)\cdot H_E(\omega)\right)$$

(EQ 6-54)

and using (6-4) again yields:

$$\varepsilon_{EE\_err}(\omega) = \left(H_W(\omega)\cdot H_E^*(\omega) + H_W^*(\omega)\cdot H_E(\omega) + |H_E(\omega)|^2\right)\cdot\Gamma_{RR}(\omega) - 2\cdot\Gamma_{SS}(\omega)\cdot \mathrm{Re}\left(H_L^*(\omega)\cdot H_R(\omega)\cdot H_E(\omega)\right)$$

(EQ 6-55)

In the next section, we will make use of the results from (6-50) and (6-55) to further illustrate that $\Gamma_{EE}(\omega)$ is a better estimate than $\Gamma_{EE\_1}(\omega)$.

6.3.4 Simulation Results Comparing the Error Components: $\Gamma_{EE\_err}(\omega)$ Versus $\Gamma_{EE\_1\_err}(\omega)$

This section will illustrate that since in practice we can only obtain the estimated Wiener solution as opposed to the true solution, evaluating $\Gamma_{EE}(\omega)$ in (6-43) using the direct computation instead of the indirect computation will produce an error component of lower magnitude.

For the simulation setup of the system described in Section 6.2, more specifically referring to the diagram in Fig. 6-2, the target source signal has been generated at 0° azimuth (i.e. in front of the hearing aid user), at *fs*=20 kHz. The target source was purposely chosen to be a white Gaussian noise source (with variance σ_S^2=150) in order to have a frequency content covering the entire frequency spectrum. For the system defined by (6-1) and (6-2), additional left and right white Gaussian noise signals were added to the received signals, with variances equal to 2. The true Wiener filter frequency response of this system has been obtained using (6-10) with known HRTFs $H_L(\omega)$ and $H_R(\omega)$ for a target source in front of the hearing aid user. From the true Wiener filter solution, an estimated Wiener filter, $\tilde{h}_W(i)$, was modeled by adding to the time domain coefficients of

the true Wiener filter white Gaussian noise components with a variance $\sigma_n^2 = 0.00005$. Figure 6-3 shows the true (i.e. "noise-free") as well as the resulting "noisy" Wiener filter magnitude frequency responses.

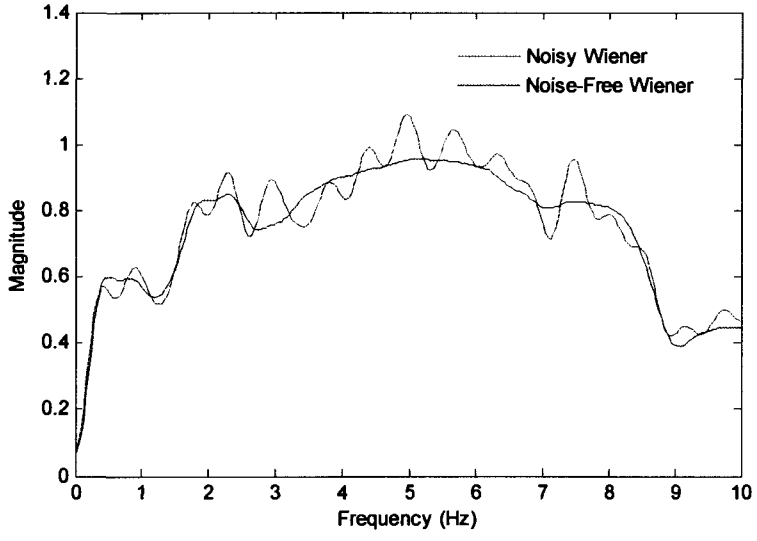

Figure 6-3: True Wiener filter versus noisy Wiener filter magnitude responses

Having the estimated Wiener solution, $\Gamma_{EE}(\omega)$ has been evaluated in both ways by: a) the direct computation (i.e. $\tilde{\Gamma}_{EE}(\omega)$) and b) the indirect computation (i.e. $\tilde{\Gamma}_{EE\_1}(\omega)$), as discussed in Section 6.3.3.

a) Applying the **direct** computation:

Figure 6-4a shows the analytical graph of $\tilde{\Gamma}_{EE}(\omega)$ using (6-52) (computed with all the theoretical PSDs), versus the experimental graph obtained by computing $\tilde{\Gamma}_{EE}(\omega)$ (averaged over several realizations) as in (6-25) with the estimated Wiener filter $\tilde{h}_W(i)$. It can be seen that there is a good match between the two curves.

Figure 6-4b then shows the corresponding analytical error magnitude component $\left|\varepsilon_{EE\_err}(\omega)\right|$ using (6-55), versus the experimental error component obtained by taking the difference between 1) $\tilde{\Gamma}_{EE}(\omega)$ computed (and averaged over several realizations) from (6-25) with the estimated Wiener filter $\tilde{h}_W(i)$, and 2) $\Gamma_{EE}(\omega)$computed from the true Wiener solution. Again a good match can be observed between the curves.

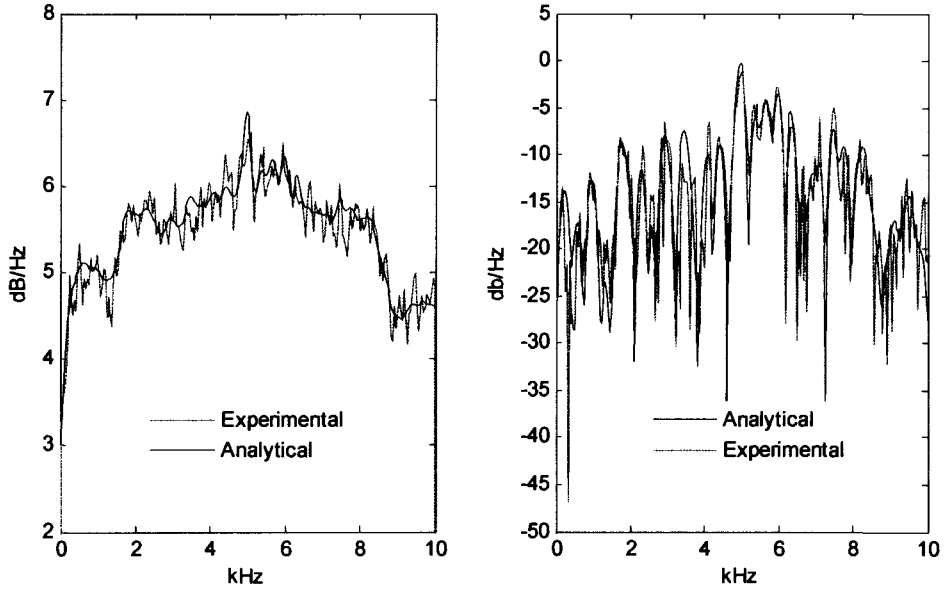

Figure 6-4: a) Magnitude graph of analytical $\tilde{\Gamma}_{EE}(\omega)$ versus the corresponding experimental graph; b) Magnitude graph of the analytical error, $\left|\varepsilon_{EE\_err}(\omega)\right|$, versus the corresponding experimental error graph. All obtained using the direct computation method.

b) Similarly, applying the **indirect** computation:

Figure 6-5a shows the analytical graph of $\tilde{\Gamma}_{EE\_I}(\omega)$ using (6-48) (computed with the theoretical PSDs), versus the experimental graph obtained by computing $\tilde{\Gamma}_{EE\_I}(\omega)$

(averaged over several realizations) as in (6-22) with the estimated Wiener filter $\tilde{h}_W(i)$. It can be seen that there is a good match between the two curves.

Figure 6-5b then shows the corresponding analytical error magnitude component $\left|\varepsilon_{EE\_1\_err}(\omega)\right|$ obtained using (6-50), versus the experimental error component obtained by taking the difference between 1) $\tilde{\Gamma}_{EE\_1}(\omega)$ computed from (6-22) (and averaged over several realizations) with the estimated Wiener filter $\tilde{h}_W(i)$, and 2) $\Gamma_{EE\_1}(\omega)$ computed from the true Wiener solution (also identical to $\Gamma_{EE}(\omega)$). Again a good match can be observed between the curves.

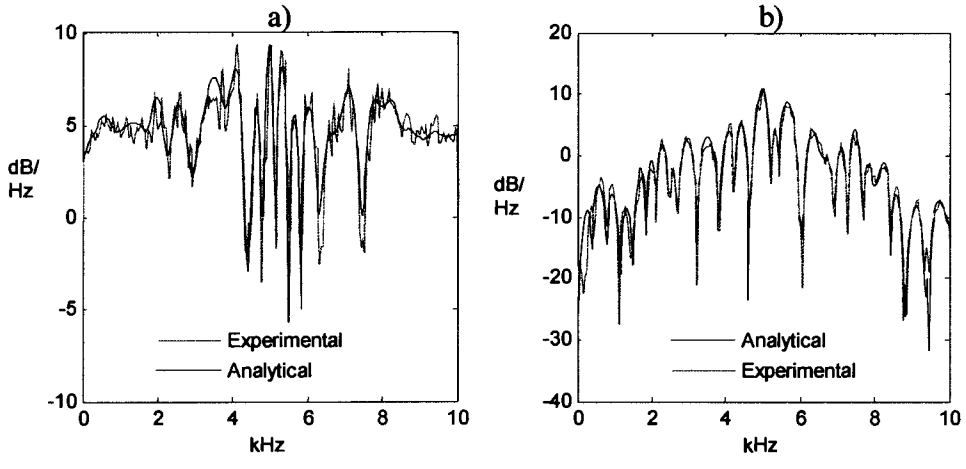

Figure 6-5: a) Magnitude graph of analytical $\tilde{\Gamma}_{EE\_1}(\omega)$ versus the corresponding experimental magnitude graph; b) Magnitude graph of the analytical error, $\left|\varepsilon_{EE\_1\_err}(\omega)\right|$, versus the corresponding experimental error graph. All obtained using the indirect computation method.

It can be seen that all the experimental graphs closely verify their corresponding analytical results. For all the experimental graphs, the PSD estimations have been achieved using Welch's method with 50% overlap, and each segment was windowed with a Hann window. Also, all the experimental graphs result from an average over 25 realizations.

Finally, Fig. 6-6 compares the analytical error magnitude graphs obtained from the direct and indirect computation methods.

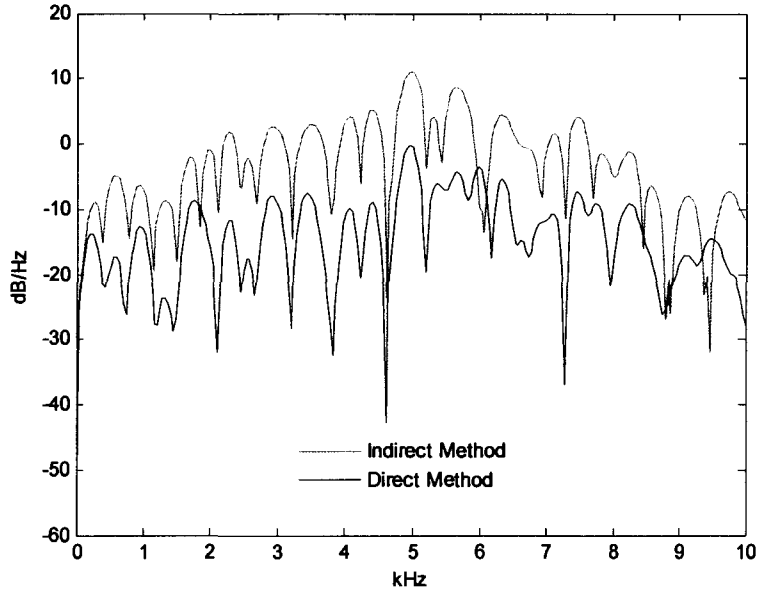

Figure 6-6: Magnitude of analytical error for the direct and indirect computation methods

From Fig. 6-6, it can easily be observed that the error magnitude $\left|\varepsilon_{EE\_err}(\omega)\right|$ from the direct computation method is lower than the error magnitude $\left|\varepsilon_{EE\_1\_err}(\omega)\right|$ obtained from the indirect computation method. Therefore, $\tilde{\Gamma}_{EE}(\omega)$ should be evaluated using the direct computation as in Equation (6-43).

6.3.5 Low frequency Compensation

Analogous to the noise estimation approach in [DOE'96], the technique proposed in the previous sub-sections will produce an underestimation of the noise PSD at low frequencies. This is due to the fact that a diffuse noise field exhibits a high coherence between the left and right channels at low frequencies, which is a known characteristic as

explained in Section 6.1. The left and right noise channels are then uncorrelated over most of the frequency spectrum except at low frequencies. The technique proposed in the previous sub-sections assumes uncorrelated noise components, thus it considers the correlated noise components to belong to the target speech signal, and consequently, an underestimation of the noise PSD occurs at low frequencies. The following will show how to circumvent this underestimation:

For a speech enhancement platform where the noise signals are picked up by two or more microphones such as in beamforming systems or any type of multi-channel noise reduction schemes, a common measure to characterize noise fields is the complex coherence function [MCC'03][ABU'04]. The latter can be seen as a tool that provides the correlation of two received noise signals based on the cross- and auto-power spectral densities. This coherence function can also be referred to as the spatial coherence function and is evaluated as follows:

$$\psi_{LR}(\omega) = \frac{\Gamma_{LR}(\omega)}{\sqrt{\Gamma_{LL}(\omega) \cdot \Gamma_{RR}(\omega)}} \qquad \text{(EQ 6-57)}$$

We assume here to have a 2-channel system with the microphones labelled as the left and right microphones and that the distance between them is d_{LR}. Then, $\Gamma_{LR}(\omega)$ is the cross-power spectral density between the left and right received noise signals, and $\Gamma_{LL}(\omega)$ and $\Gamma_{RR}(\omega)$ are the auto-power spectral densities of left and right signals respectively. The coherence has a range of $|\psi_{LR}(\omega)| \leq 1$ and is primarily a normalized measure of correlation between the signals at two points (i.e. positions) in a noise field at a given frequency. Moreover, it was found that the coherence function of a diffuse noise field is in fact real-valued and an analytical model has been developed. The approximation provided by the model is given by [MCC'03],[COO'55]:

$$\psi_{LR}(f) = \operatorname{sinc}\left(\frac{2 \cdot \pi \cdot f \cdot d_{LR}}{c}\right) \qquad \text{(EQ 6-58)}$$

where d_{LR} is the distance between the left and right microphones and c is the speed of sound.

However, this model was derived for two omni-directional microphones in free space. For binaural hearing, the directionality and diffraction/reflection due to the pinna and the head will have some influence, and the analytical model assuming microphones in free space represented in (6-58) should be re-adjusted. The adjustment has been undertaken as follows: the coherence function of (6-57) was evaluated using real diffuse cafeteria noise signals. The left and right noise signals used in the simulation were provided by a hearing aids manufacturer and were collected from hearing aids microphone recordings mounted on a KEMAR. The distance parameter was then equal to the distance between the dummy head ears. The KEMAR was placed in a crowded university cafeteria. It has been found that the effect brought by having the microphones placed on human ears as opposed to free space reduces the bandwidth of the low frequency range where the high correlation part of a diffuse noise field is present, and it also slightly decreases the correlation magnitudes. Consequently, it was established by simulation that by simply increasing the distance parameter of the analytical diffuse noise model of (6-58) (i.e. with microphones in free space) and applying a factor less than one to the latter, it was possible to have a modified analytical model matching the coherence function evaluated using the real binaural cafeteria noise, as it will shown in the simulation results.

Now, in order to use the above notions and modify the noise PSD estimation equations found for uncorrelated noise signals, some of the key equations previously derived need to be re-written by taking into account the noise correlation at low frequencies. The cross-power spectral density between the left and right noisy channels in Equation (6-9) becomes at low frequencies:

$$\Gamma_{LR}^{C}(\omega) = \Gamma_{SS}(\omega) \cdot H_L(\omega) \cdot H_R^*(\omega) + \Gamma_{N_L N_R}(\omega) \quad \text{(EQ 6-59)}$$

where $\Gamma_{N_L N_R}(\omega)$ is the noise cross-power spectral density between the left and right channel. The upper script "*C*" is to differentiate between the previous Equation (6-9) and the new compensated one taking into account the low frequency noise correlation.

Therefore, the Wiener solution becomes:

$$H_W^C(\omega) = \frac{\Gamma_{LR}(\omega)}{\Gamma_{RR}(\omega)} = \frac{\Gamma_{SS}(\omega) \cdot H_L(\omega) \cdot H_R^*(\omega) + \Gamma_{N_L N_R}(\omega)}{\Gamma_{RR}(\omega)} \quad \text{(EQ 6-60)}$$

Using the definition in (6-57), the coherence function of a diffuse noise field can be expressed as:

$$\psi(\omega) = \frac{\Gamma_{N_L N_R}(\omega)}{\sqrt{\Gamma_{N_L N_L}(\omega)\Gamma_{N_R N_R}(\omega)}} = \frac{\Gamma_{N_L N_R}(\omega)}{\Gamma_{NN}(\omega)} \quad \text{(EQ 6-61)}$$

Consequently, the noise cross-power spectral density, $\Gamma_{N_L N_R}(\omega)$, can replaced by:

$$\Gamma_{N_L N_R}(\omega) = \psi(\omega) \cdot \Gamma_{NN}(\omega) \quad \text{(EQ 6-62)}$$

For the remaining of this section, the noise cross-power spectral density $\Gamma_{N_L N_R}(\omega)$, in any equation, will be replaced by $\psi(\omega) \cdot \Gamma_{NN}(\omega)$.

Following the procedure employed to find the noise PSD estimator derived in Section 6.3.1, starting again from the squared magnitude response of the Wiener filter, we get:

$$|H_W(\omega)|^2 = \frac{\left(\Gamma_{LL}(\omega) - \Gamma_{NN}(\omega)\right) \cdot \left(\Gamma_{RR}(\omega) - \Gamma_{NN}(\omega)\right) + \psi^2(\omega) \cdot \Gamma_{NN}^2(\omega) + \Gamma_A(\omega)}{\Gamma_{RR}^2(\omega)} \quad \text{(EQ 6-63)}$$

where $\Gamma_A(\omega) = 2 \cdot \psi(\omega) \cdot \Gamma_{NN}(\omega) \cdot \Gamma_{SS}(\omega) \cdot \mathrm{Re}\left\{H_L(\omega) \cdot H_R^*(\omega)\right\}$ (EQ 6-64)

and using Equation (6-60), $\Gamma_A(\omega)$ can be rewritten as:

$$\begin{aligned}\Gamma_A(\omega) &= 2\cdot\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\mathrm{Re}\left\{H_W^C(\omega)\Gamma_{RR}(\omega)-\psi(\omega)\cdot\Gamma_{NN}(\omega)\right\}\\ &= 2\cdot\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\Gamma_{RR}(\omega)\cdot\mathrm{Re}\left\{H_W^C(\omega)\right\}-2\cdot\psi^2(\omega)\cdot\Gamma_{NN}^2(\omega)\end{aligned} \quad \text{(EQ 6-65)}$$

Substituting (6-65) into (6-63) and after few simplifications, the noise PSD estimation is found by solving the following quadratic equation:

$$\left(1-\psi^2(\omega)\right)\cdot\Gamma_{NN}^2(\omega)+\Gamma_{NN}(\omega)\cdot\begin{pmatrix}-\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)+\\ 2\cdot\psi(\omega)\cdot\mathrm{Re}\left\{H_W^C(\omega)\right\}\end{pmatrix}+\Gamma_{EE\_1}(\omega)\cdot\Gamma_{RR}(\omega))=0 \quad \text{(EQ 6-66)}$$

where again $\Gamma_{EE\_1}^C(\omega)=\Gamma_{LL}(\omega)-\Gamma_{RR}(\omega)\cdot|H_W^C(\omega)|^2$, which was referred to as the indirect computation approach explained in Section 6.3.1.

Similar to Section 6.3.2, it will be demonstrated here again that $\Gamma_{EE\_1}^C(\omega)$ is still equal to the auto-power spectral density of the prediction error $e(i)$ (i.e. $\Gamma_{EE}^C(\omega)=F.T.(\gamma_{ee}(\tau))$), and therefore $\Gamma_{EE}^C(\omega)$ is referred to as the direct computation approach as explained in Section 6.3.2. We had established in Section 6.3.2 that the auto-power spectral density of the residual error was the sum of four terms as shown in Equation (6-35). By taking into account the low frequency noise correlation, two of the terms in (6-35), namely $\Gamma_{L\tilde{L}}(\omega)$ and $\Gamma_{\tilde{L}L}(\omega)$, will be modified as follows:

$$\Gamma_{EE}^C(\omega)=\Gamma_{LL}(\omega)-\Gamma_{L\tilde{L}}^C(\omega)-\Gamma_{\tilde{L}L}^C(\omega)+\Gamma_{\tilde{L}\tilde{L}}(\omega) \quad \text{(EQ 6-67)}$$

where:

$$\begin{aligned}\Gamma_{L\tilde{L}}^C(\omega) &= \Gamma_{L\tilde{L}}(\omega)+\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\left(H_W^C(\omega)\right)^*\\ &= \Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)\cdot\left(H_W^C(\omega)\right)^*+\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\left(H_W^C(\omega)\right)^*\end{aligned} \quad \text{(EQ 6-68)}$$

and

$$\begin{aligned}\Gamma_{\tilde{L}L}^C(\omega) &= \Gamma_{\tilde{L}L}(\omega)+\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot H_W^C(\omega)\\ &= \Gamma_{SS}(\omega)\cdot H_R^*(\omega)\cdot H_L(\omega)\cdot H_W^C(\omega)+\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot H_W^C(\omega)\end{aligned} \quad \text{(EQ 6-69)}$$

Adding all the terms in (6-67), we get:

$$\Gamma_{EE}^{C}(\omega)=\Gamma_{EE}(\omega)+2\cdot\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\mathrm{Re}\left\{H_{W}^{C}\right\} \qquad \text{(EQ 6-70)}$$

$$=\Gamma_{NN}(\omega)\cdot(1+|H_{W}^{C}(\omega)|^{2})+\Gamma_{SS}(\omega)\cdot|H_{L}(\omega)|^{2}+\Gamma_{SS}(\omega)\cdot|H_{R}(\omega)|^{2}\cdot|H_{W}^{C}(\omega)|^{2}+\Gamma_{B}(\omega)$$

(EQ 6-71)

where $\Gamma_{B}(\omega)=-2\cdot\Gamma_{SS}(\omega)\cdot Re\left(H_{L}^{*}(\omega)\cdot H_{R}(\omega)\cdot H_{W}^{C}(\omega)\right)+2\cdot\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\mathrm{Re}\left\{H_{W}^{C}(\omega)\right\}$

(EQ 6-72)

Using the complex conjugate of (6-60) (i.e.$\left(H_{W}^{C}(\omega)\right)^{*}$) in (6-62), Equation (6-72) simplifies to:

$$\Gamma_{B}(\omega)=-2\cdot Re\left\{\left(\left(H_{W}^{C}(\omega)\right)^{*}\cdot\Gamma_{RR}(\omega)-\psi(\omega)\cdot\Gamma_{NN}(\omega)\right)\cdot H_{W}^{C}(\omega)\right\}+2\cdot\psi(\omega)\cdot\Gamma_{NN}(\omega)\cdot\mathrm{Re}\left\{H_{W}^{C}(\omega)\right\}$$

$$=2\cdot\left|H_{W}^{C}(\omega)\right|^{2}\cdot\Gamma_{RR}(\omega)$$

(EQ 6-73)

Replacing (6-73) in (6-71) and using (6-3) and (6-4), $\Gamma_{EE}^{C}(\omega)$ becomes:

$$\Gamma_{EE}^{C}(\omega)=\Gamma_{LL}(\omega)-\Gamma_{RR}(\omega)\cdot|H_{W}^{C}(\omega)|^{2} \qquad \text{(EQ 6-74)}$$

We can see that the equality still holds, that is: $\Gamma_{EE}^{C}(\omega)=\Gamma_{EE\_1}^{C}(\omega)$.

To finalize, solving the quadratic equation in (6-66) and using $\Gamma_{EE}^{C}(\omega)$ instead of $\Gamma_{EE\_1}^{C}(\omega)$, the noise PSD estimation for a diffuse noise field environment without neglecting the low frequency correlation is given by (6-75)- (6-77):

$$\Gamma_{NN}(\omega)=\frac{1}{2\cdot\left(1-\psi^{2}(\omega)\right)}\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)-2\cdot\psi(\omega)\cdot\Gamma_{RR}(\omega)\cdot\mathrm{Re}\left\{H_{W}^{C}(\omega)\right\}-\Gamma_{root}(\omega)\right)$$

(EQ 6-75)

where $$\Gamma_{root}(\omega)=\sqrt{\begin{pmatrix}-\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)\\+2\cdot\psi(\omega)\cdot\Gamma_{RR}(\omega)\cdot\mathrm{Re}\left\{H_{W}^{C}(\omega)\right\}\end{pmatrix}^{2}-4\cdot\left(1-\psi^{2}(\omega)\right)\cdot\Gamma_{EE}^{C}(\omega)\Gamma_{RR}(\omega)}$$

(EQ 6-76)

and $\Gamma_{EE}^{C}(\omega) = F.T.\left(\gamma_{ee}(\tau)\right) = F.T.\left\{E\left(e(i+\tau)\cdot e(i)\right)\right\}$ (EQ 6-77)

From (6-60), the product $\Gamma_{RR}(\omega)\cdot \mathrm{Re}\left\{H_W^C(\omega)\right\}$ in (6-75) and (6-76) is equivalent to:

$$\Gamma_{RR}(\omega)\cdot \mathrm{Re}\left\{H_W^C(\omega)\right\} = \Gamma_{RR}(\omega)\cdot\frac{H_W^C(\omega)+\left(H_W^C(\omega)\right)^*}{2} = \mathrm{Re}\left\{\Gamma_{LR}(\omega)\right\}.$$

The complete diffuse noise PSD estimator algorithm is summarized in Table 10-1 (in Chapter 10).

6.4 Case of Additional Directional Interferences

Chapter 6 focused on noise PSD estimation for the case of a single directional target source combined with background diffuse noise. For more general cases where there would also be directional interferences (i.e. directional noise sources), the behavior of the proposed diffuse noise PSD estimator is briefly summarized below. The components on the left and right channels that remain fully or strongly cross-correlated are called here the "equivalent" left and right directional source signals, while the components on the left and right channel that have poor or zero cross-correlation are called here the "equivalent" left and right noise signals. Note that with this definition some of the equivalent noise signal components include directional target and directional interference signal components that can no longer be predicted from the other channel, because predicting a sum of directional signals from another sum of directional signals no longer allows a perfect prediction (i.e. the cross-correlation between the two sums of signals is reduced). With these equivalent source and noise signals, if the proposed noise PSD estimator remains the same as described in this chapter, some of the assumptions made in the development of the estimator may no longer be fully met: 1) the PSD of the left and right equivalent noise components may no longer be the same, and 2) the equivalent source and noise signals on each channel may no longer be fully uncorrelated. The PSD noise estimator may thus become biased in such cases. Nevertheless, it will be shown in Chapter 10 with several speech enhancement experiments under complex acoustic

environments (including reverberation, diffuse noise, and several non-stationary directional interferences) that the proposed diffuse noise PSD estimator can still provide a useful estimate as part of an overall noise reduction system.

It should also be noted that for the case of a single directional speech source combined with diffuse noise but under a reverberant environment, the speech components received at the two ears would become partly diffuse (from the tail of the reverberation), and the proposed PSD noise estimator would then detect the reverberant (or diffuse) part of the speech as noise (which is a good thing). This estimator could thus potentially be used by a speech enhancement algorithm to also reduce the reverberation found in the received speech signal.

The next chapter will perform experimental evaluations of the binaural noise PSD estimation method proposed in this chapter, compared with the previously described advanced noise estimation methods discussed in Chapter 3.

Chapter 7. DIFFUSE NOISE ESTIMATION TECHNIQUES EVALUATION

In the first section of this chapter, various simulated hearing scenarios will be described where the target speaker is located anywhere around a binaural hearing aid user in a noisy environment. In the second section, the accuracy of the proposed binaural diffuse noise estimation technique elaborated in Chapter 6 will be compared with the two advanced noise estimation techniques described in Section 3.1, namely the noise estimation approach based on minimum statistics (discussed in Section 3.1.1) [MAR'01] and the cross-power spectral density method (derived in Section 3.1.2) [DOE'96]. The noise estimation will be performed on the scenarios presented in the first section. The performance under highly non-stationary noise conditions will also be analyzed. Finally, the fourth section will summarize the quality of the proposed binaural noise PSD estimation technique.

7.1 Simulation Setup and Hearing Situations

The following is the description of various simulated hearing scenarios where the noise PSD will be estimated. It should be noted that all data used in the simulations such as the binaural speech signals and the binaural noise signals were provided by a hearing aid manufacturer and obtained from "Behind The Ear" (BTE) hearing aids microphone recordings, with microphones installed at the left and the right ears of a KEMAR dummy head, with a 17.5 cm distance between the ears. For instance, the dummy head was rotated at different positions to receive the target source speech signal at diverse azimuth angles and the source speech signal was produced by a loudspeaker at 1.5 meters from the KEMAR. Also, the KEMAR had been installed in different noisy environments such as a university cafeteria, to collect real life noise-only data. Speech and noise sources were recorded separately. It should be noted that the target speech source used in the simulation was purposely recorded in a reverberant free environment, to avoid an

overestimation of the diffuse noise PSD due to the tail of reverberation (then considered as noise). As briefly introduced at the end of Section 6.4, this overestimation can actually be beneficial since the proposed binaural estimator can also be used by a speech enhancement algorithm to reduce reverberation.

Scenario A):

The target speaker is in front of the binaural hearing aid user (i.e. azimuth = 0°) and the additive corrupting binaural noise used in the simulation has been obtained from the binaural recordings in a university cafeteria (i.e. cafeteria babble-noise). The noise has the characteristics of a diffuse noise field as discussed in Chapter 6. Figure 7-1 illustrates this scenario.

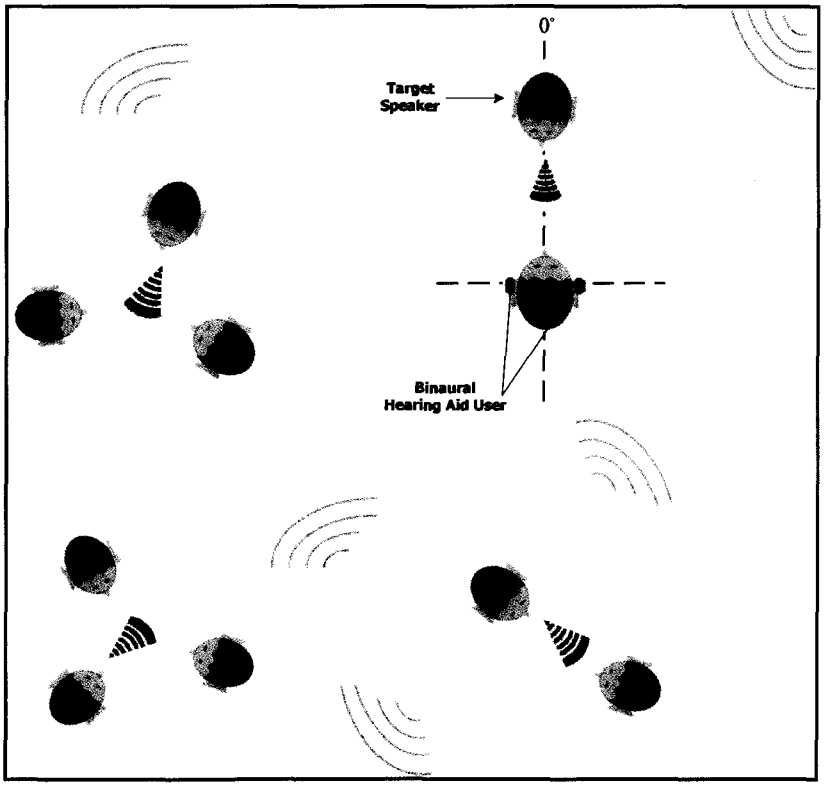

Figure 7-1: Scenario A, the target speaker is in front of the binaural hearing aid user in a diffuse noise field environment

Scenario B):

The target speaker is at 90° to the right of the binaural hearing aid user (i.e. azimuth = 90°) and located again in a diffuse noise field environment (i.e. cafeteria babble-noise) as depicted in Fig. 7-2.

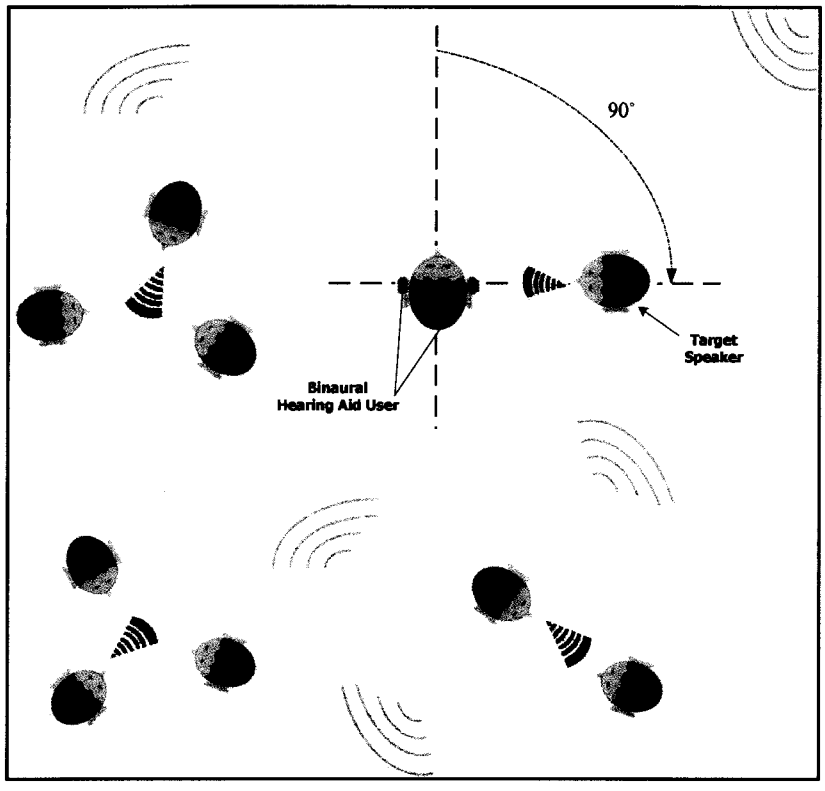

Figure 7-2: Scenario B, the target speaker is at 90° on the right of the binaural hearing aid user in a diffuse noise field environment

Scenario C):

Similar to scenario B, the target speaker is at 90° to the right of the binaural hearing aid user (i.e. azimuth = 90°) in the noisy cafeteria environment. However, instead of using real target speech signals, the target source was purposely chosen to be white noise in order to have a frequency spectrum covering the entire wideband frequency spectrum as opposed to a speech signal spectrum. This scenario is mostly for comparing the proposed binaural noise estimation with the cross-power spectral density method.

Scenario D):

The target speaker is in front of the binaural hearing aid user (i.e. azimuth = 0°) similar to scenario A. However, even though the original noise coming from a cafeteria is quite non-stationary, its diffuse power level will be purposely increased and decreased during a selected time period to simulate highly non-stationary noise conditions. This scenario could be encountered for example if the user is entering or exiting a noisy cafeteria, etc..

7.2 Noise Estimation Techniques Evaluation

For simplicity, the proposed binaural noise estimation technique of Chapter 6 will be given the acronym: **PBNE**. The cross-power spectral density method and the minimum statistics based approach of Chapter 3 will be given the acronyms: **CPSM** and **MSA** respectively. For our proposed technique, in this chapter (as opposed to Chapter 10 where a more efficient implementation will be used), a least-squares algorithm with 80 coefficients has been used to estimate the Wiener solution of (6-5), which performs a prediction of the left noisy speech signal from the right noisy speech signal updated on a frame-by-frame basis as illustrated in Fig. 6-2.

It should be noted that the least-squares solution of the Wiener filter should also include a causality delay (i.e. Δ):

$$e(i) = l(i-\Delta) - r(i) \otimes h_w(i) \qquad \text{(EQ 7-1)}$$

The causality delay should be large enough to handle two criteria:

a) The causality delay is mandatory to allow the Wiener filter to be on either channel since the position of the target speech signal is arbitrary:

- The largest lag between the left and right received noisy signal is when the target speaker is at $\theta = 90^{\circ}$:
- It yields the highest interaural time delay (ITD) [HAR'05]:

$$ITD \approx \frac{3 \cdot r}{c} \sin(\theta) = 763\mu s \; ITD \approx \frac{3 \cdot r}{c} \sin(\theta) = 741\mu s \qquad \text{(EQ 7-2)}$$

- At f_s = 20kHz, the maximum ITD is about 15 samples with head radius r =8.75 cm and a sound speed of c=344 m/s. Note that the factor 3 in equation (7-2) is an adjustment for the ITD computation for the low frequencies in a reverberant condition [HAR'05]. From above, the causality delay should be at least a minimum of $\Delta = 15$ samples. Therefore, a causality delay of M/2=40 samples is sufficient enough to consider the largest ITD possible, but:

b) The causality delay should be also sufficient enough to take into account non-causal samples resulting from the prediction of the inverse of an HRIR. For instance: assuming that a Wiener filter is placed on the right channel and performs a prediction of the left noisy speech signal from the right noisy speech signal, the Wiener solution is found to be:

$$H_W(\omega) = \Gamma_{SS}(\omega) \cdot H_L(\omega) \cdot H_R^*(\omega) / \Gamma_{RR}(\omega) \qquad \text{(EQ 7-3)}$$

For simplicity here, let's assume that there is no diffuse noise present i.e. $\Gamma_{NN}(\omega) = 0$, the Wiener filter is then simplified to:

$$H_W(\omega) = \frac{\Gamma_{SS}(\omega) \cdot H_L(\omega) \cdot H_R^*(\omega)}{\Gamma_{RR}(\omega)} = \frac{\Gamma_{SS}(\omega) \cdot H_L(\omega) \cdot H_R^*(\omega)}{\Gamma_{SS}(\omega) \cdot |H_R(\omega)|^2 + \Gamma_{NN}(\omega)} = H_L(\omega) \cdot \frac{1}{H_R(\omega)} \qquad \text{(EQ 7-4)}$$

In the time domain, the Wiener filter is then:

$$h(n) = h_l(n) \otimes inv(h_r(n)) \qquad \text{(EQ 7-5)}$$

From the above solution, it can be observed that the Wiener filter solution is then the convolution between the left HRIR and the inverse of the right HRIR. The ideal inverse of the right HRIR will typically have some non-causal samples (i.e. non minimal phase

HRIR) and therefore the least-squares estimate of the Wiener solution should include a causality delay.

A modified distance parameter of 35 cm (i.e. double of the actual distance between the ears of the KEMAR (i.e. d= d_{LR} x 2=17.5 cm x 2) has been selected for the analytical diffuse noise model of (6-58). The model of (6-58) has also been multiplied by a factor of α =0.8. This factor of 0.8 is actually a conservative value because from our empirical results, the practical coherence obtained from the binaural cafeteria recordings would vary between 1.0 and 0.85 at the very low frequencies (below 500 Hz). The lower bound factor of 0.8 was selected to prevent a potential overestimation of the noise PSD at the very low frequencies, while still providing good low frequency compensation. Figure 7-3 illustrates the experimental coherence obtained from the binaural cafeteria babble-noise recordings for the setup explained in Section 7.1 using Equation (6-57) and the corresponding modified analytical diffuse noise model of (6-58) used in our technique expressed as:

$$\psi(\omega) = \psi_{LR}^{\text{modified}}(\omega) = \alpha \cdot sinc\left(\frac{\omega \cdot d_{LR} \cdot 2}{c}\right) \qquad \text{(EQ 7-6)}$$

For comparison, Fig. 7-3 also shows the unmodified analytical diffuse noise model represented in Equation (6-58), which assumes two omni-directional microphones in free space with d_{LR}=17.5 cm. The authors in [CAM'03] claim that because of the directionality and diffraction/reflection due to the pinna and the head (i.e. head shadowing effects), the first zero of the practical coherence of a diffuse noise field will be pushed back to around 500Hz instead, and frequencies above about 350Hz will have coherence less than 0.5. As it can be seen in Fig. 7-3, those characteristic results were confirmed from our measurements.

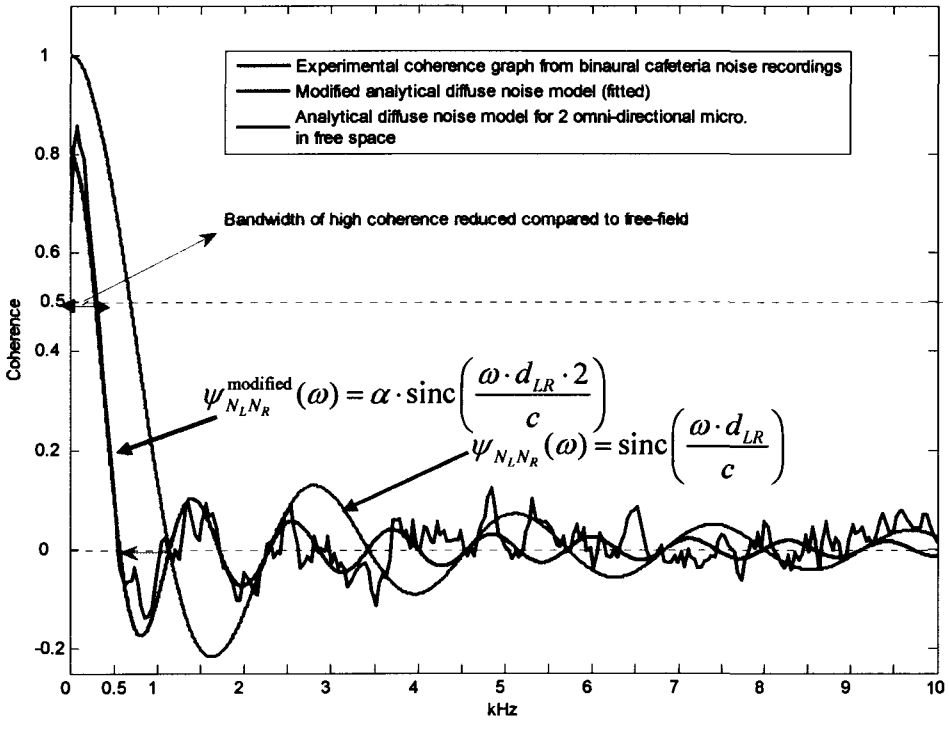

Figure 7-3: Experimental coherence of real binaural cafeteria noise (Blue) versus the corresponding modified analytical diffuse noise model (Red) and the unmodified analytical diffuse noise model assuming 2 omni-directional microphones in free space (Green)

Moreover, all the PSD calculations have been made using Welch's method with 50% overlap, and a Hann window has been applied to each segment.

7.2.1 PBNE versus CPSM

Results for Scenario A):

The left and right noise-free speech signals are shown in Fig. 7-4 and the corresponding noisy speech signals received by the binaural hearing user are illustrated in Fig. 7-5. The left and right SNRs are both equal to 5dB since the speaker is in front of the hearing aid

user. PBNE and CPSM have the advantage to estimate the noise on a frame-by-frame basis, that is both techniques do not require the knowledge of previous frames to perform their noise estimation. Fig. 7-5 shows the frame where the noise PSD has been estimated. A frame length of 25.6ms has been used at a sampling frequency of 20kHz. Also, the selected frame purposely contained the presence of both speech and noise. The left and right received noise-free speech PSDs and the left and right measured noise PSDs are depicted in Fig. 7-6. It can be noticed that the measured noise obtained from the cafeteria has approximately the same left and right PSDs, which verifies one of the characteristics of a diffuse noise field. Therefore, for convenience, the original left and right noise PSDs will be represented with the same color in all figures related to noise estimation results. The noise estimation results are given in Fig. 7-7.

For clarity, the results obtained with PBNE have been shifted vertically above the results from CPSM. It can be seen that both techniques provide a good noise PSD estimate, which closely tracks the original colored noise PSDs (i.e. cafeteria babble-noise). However, it can be noticed that CPSM suffers from an under estimation of the noise at low frequencies (here below about 500Hz) as indicated in [DOE'96] [HAM'02]. Fig. 7-8 shows a close-up of the noise estimation results at low frequencies up to 500Hz to illustrate the under estimation from CPSM. The underestimation is about 7dB for this case. To better compare the results, instead of showing the results from only a single realization of the noise sequences, the results over an average of 20 realizations but still maintaining the same speech signal (i.e. by processing the same speech frame index with different noise sequences) are shown in Fig. 7-9. The under estimation at low frequencies is clearly noticeable from the CPSM result. On the other hand, PBNE provides a good estimation even at low frequencies due to the compensation method developed earlier. Finally, Fig. 7-10 shows the noise estimation results with and without the compensation method using PBNE. It can be observed that the noise PSD estimation at low frequencies is well compensated.

Even though the diameter of the head could be provided during the fitting stage for future high-end binaural hearing aids, the effect of the low frequency compensation by the PBNE approach was evaluated with different head diameters (d_{LR}) and gain factors, to

evaluate the robustness of the approach in the case where the parameters selected for the modified diffuse noise model are not optimum. From the binaural cafeteria recordings provided by a hearing aids manufacturer, the experimental coherence obtained is as illustrated in Fig. 7-3. The optimum model parameters are d_{LR} =17.5 cm (which is multiplied by 2 in our modified analytical diffuse noise model for microphones not in free-field) and a factor α =0.8. Fig. 7-11 shows the PBNE noise estimation results with various non-optimized head diameters and gain factors used with our approach, followed by the corresponding error "offset" of the PBNE noise PSD estimate for the various parameter settings as depicted in Fig. 7-12. Each error offset was computed by taking the difference between the noise PSD estimate (in decibels) of Fig. 7-11 and the linear average of the original left and right noise PSDs converted in decibels. All the noise estimation results were obtained using Equations (6-75) to (6-77), which incorporate the low frequency compensator. It can be seen that even with d_{LR} =14 cm (2 cm below the actual head diameter of the KEMAR) and a factor of α =1.0, only a slight overestimation is noticeable at around 500 Hz. On the other hand, even with d_{LR} = 20cm (4 cm higher than the actual head diameter) where an underestimation result is expected at the low frequencies, the proposed method still provides a better noise PSD estimation than having no low frequency compensation for the lower frequencies (i.e. the result with d_{LR} = 17.5 cm with factor α =0.0).

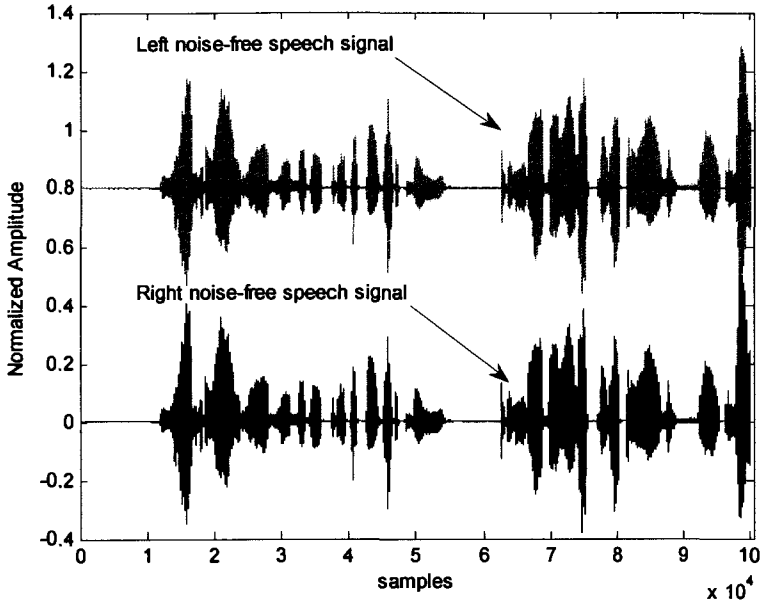

Figure 7-4: Left (top - red) and right (bottom - black) noise-free speech signals for scenario A

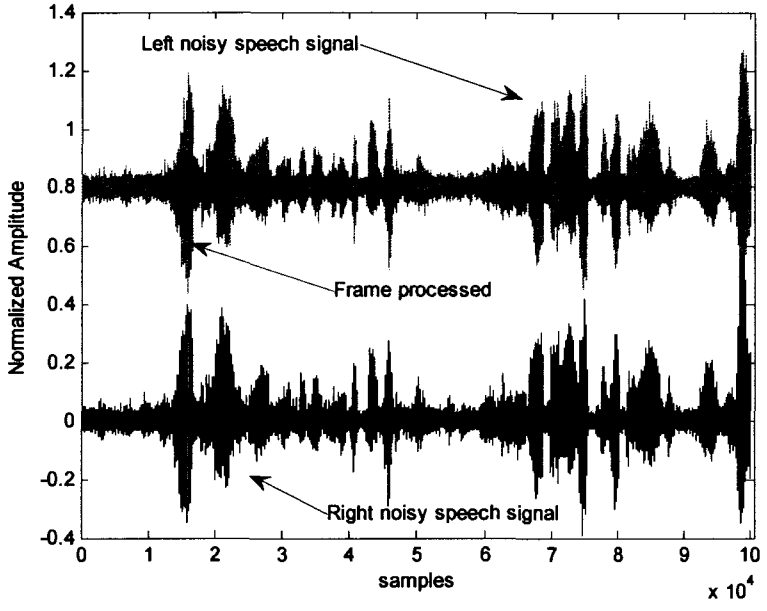

Figure 7-5: Left (top - red) and right (bottom - black) noisy speech signals and the frame that has been processed (blue) for scenario A

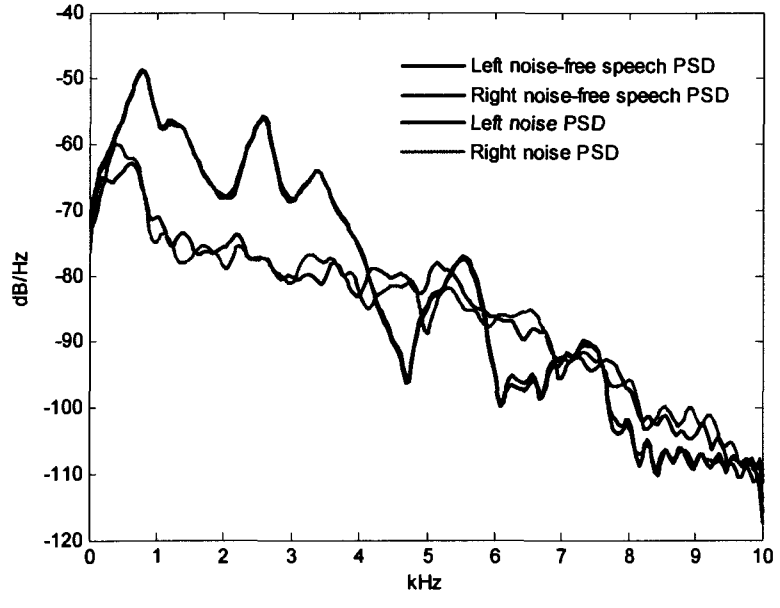

Figure 7-6: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs for scenario A

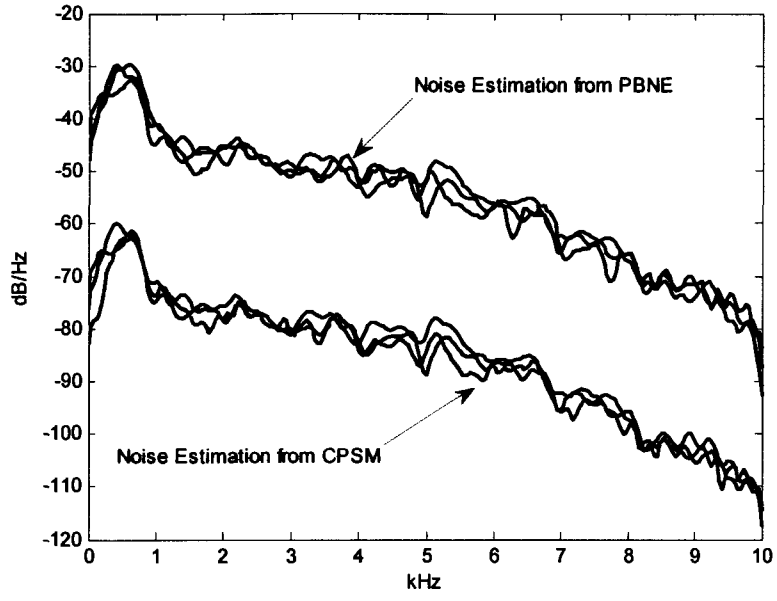

Figure 7-7: Noise estimation results for scenario A. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black)

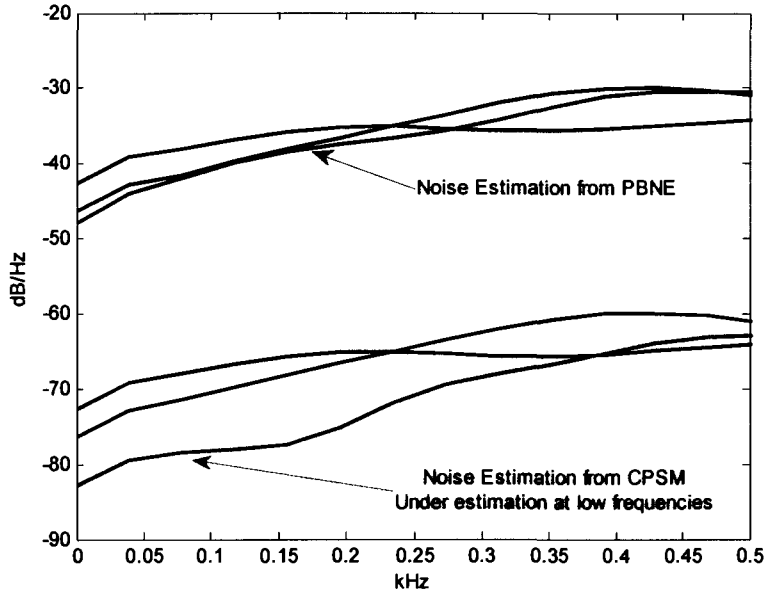

Figure 7-8: Noise estimation results shown for frequencies from 0 to 500 Hz only. PBNE (top - blue) versus CPSM (bottom - red), original left and right noise PSDs (top or bottom - black)

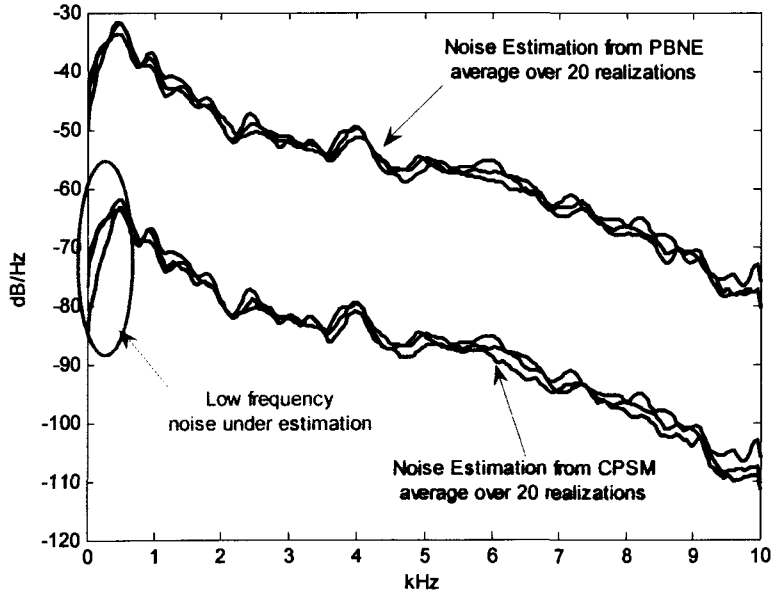

Figure 7-9: Noise estimation results from an average over 20 realizations. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black)

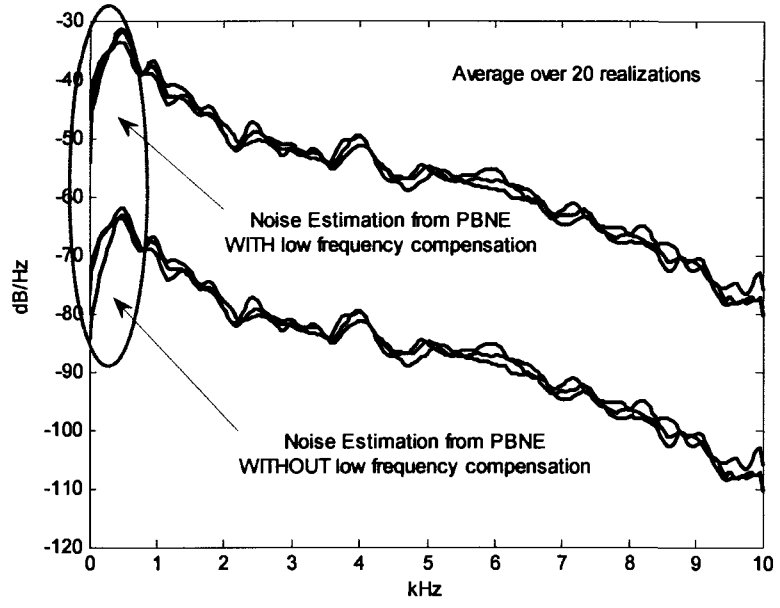

Figure 7-10: Noise estimation results for PBNE with the low frequency compensation (top – blue) and without (bottom – blue), original left and right noise PSDs (top or bottom – black)

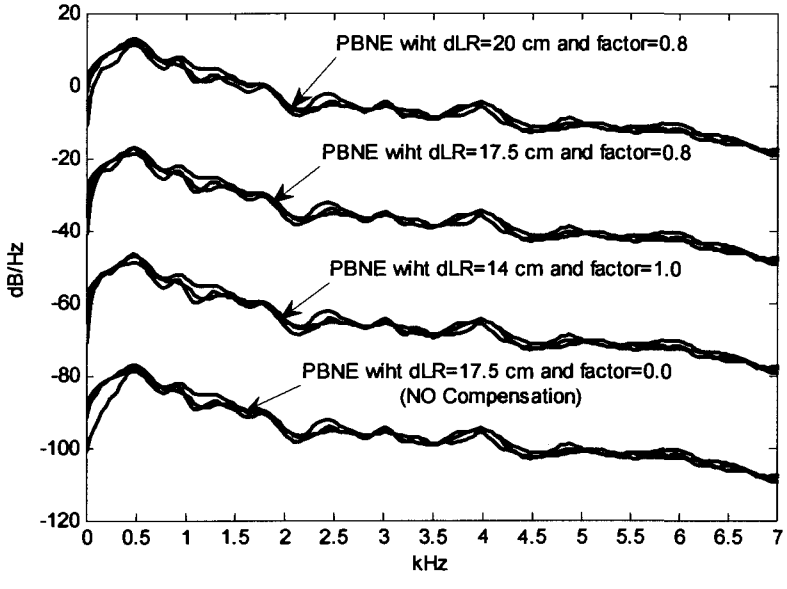

Figure 7-11: PBNE noise PSD estimation results (blue) from an average over 20 realizations with various head diameters and gain factors. Original left and right noise PSDs (black).

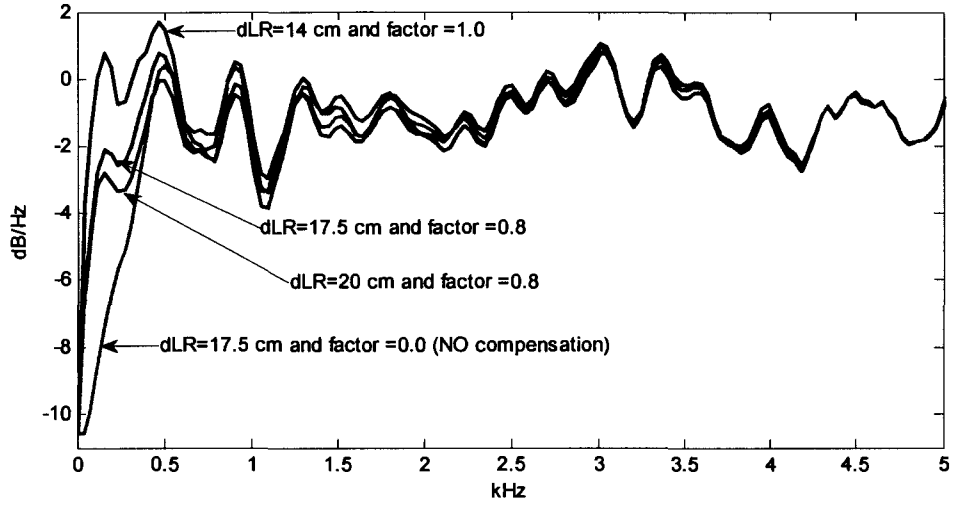

Figure 7-12: Error offsets of the PBNE noise PSD estimation for various head diameters and factors

Results for Scenario B):

In contrast to scenario A, the location of the speaker has been changed from the front position to 90° on the right of the binaural hearing aid user. The left and right SNRs are 2.2 dB and 8.6 dB repsectively. Figure 7-13 illustrates the received signal PSDs for this configuration corresponding to the same frame time index as selected in Fig. 7-5. From Fig. 7-14, it can be seen that for this scenario, the noise estimation from PBNE clearly outperforms the one from CPSM. CPSM produces an overestimation of the noise PSD. This is due to the fact that the technique in [DOE'96] assumes that the left and right source speech signals follows the same attenuation path before reaching the hearing aid microphones i.e. assuming equivalent left and right HRTFs. This situation only appends if the speaker is frontal (or at the back), implying that the received speech PSD levels in each frequency band should be comparable, which is definitely not the case as shown in Fig. 7-13 for a speaker at 90° azimuth. CPSM was not designed to provide an exact solution when the target source is not in front of the user. The noise results over an

average of 20 realizations are also shown in Fig. 7-15. We can easily see the bias occurring in the estimated noise from CPSM, producing an overestimation.

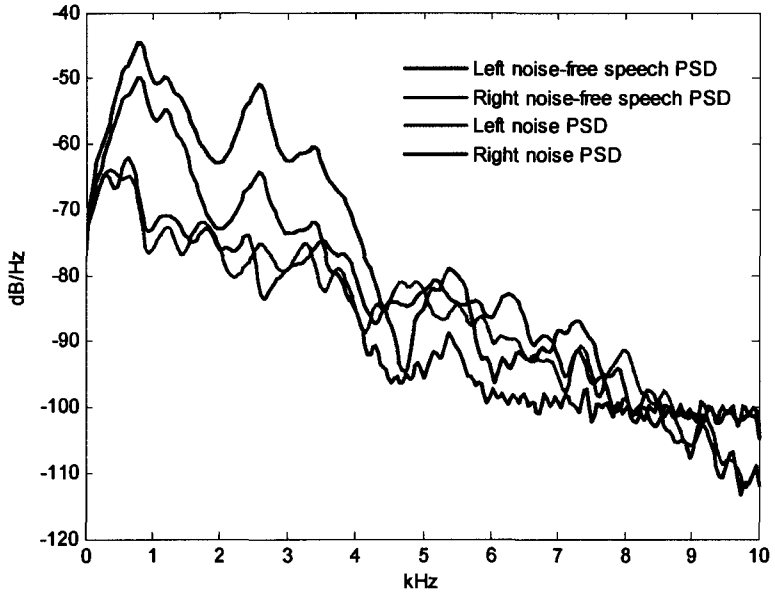

Figure 7-13: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs used for scenario B

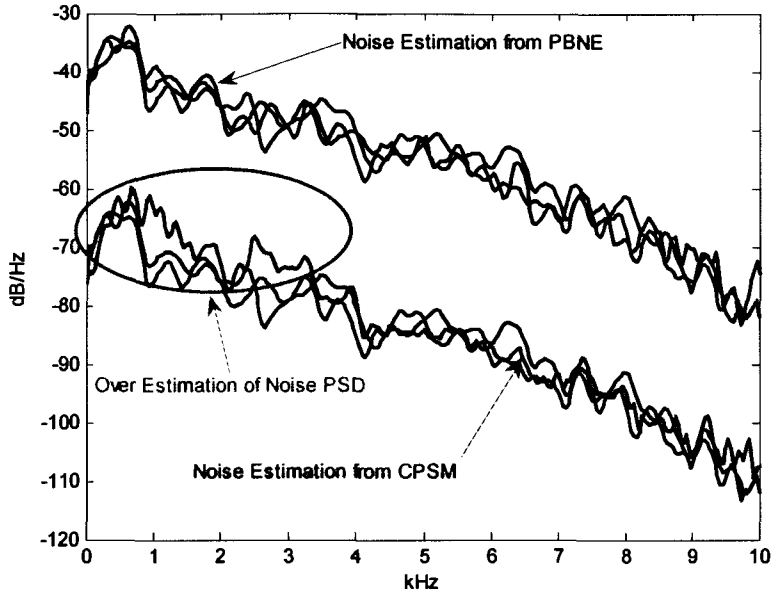

Figure 7-14: Noise estimation results for scenario B. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black)

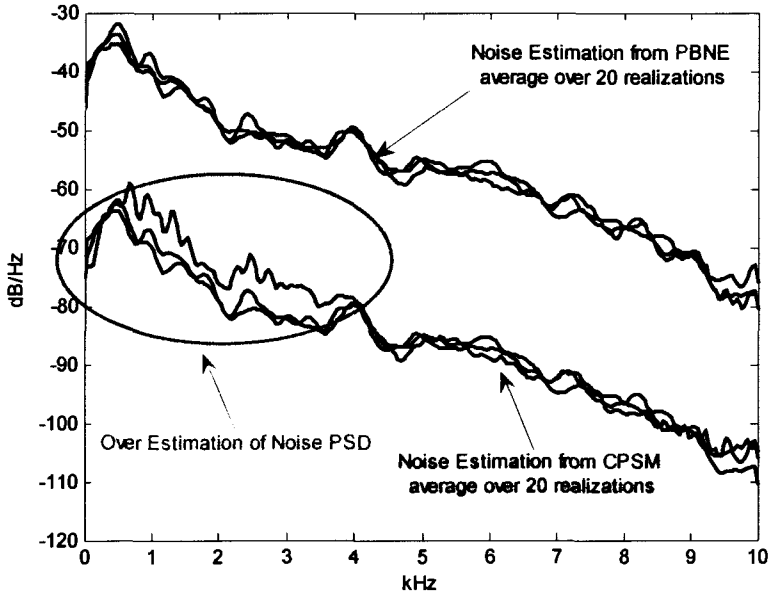

Figure 7-15: Noise estimation results for an average of 20 realizations for scenario B. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black)

Results for Scenario C):

In this scenario, the input target signal is white noise instead of a speech source signal and the remaining setup is the same as the scenario in B. Figure 7-16 shows the received PSD levels. It can be seen that the received left and right target source signals cover the entire wideband frequency spectrum and the corresponding target source PSD levels are at significantly different magnitudes due to the HRTFs corresponding to 90° azimuth. The difference of PSD magnitudes is larger at high frequencies since the effect of head shadowing is greater at higher frequencies. The results shown in Fig. 7-17 clearly illustrate the overestimation of noise from CPSM. In broad terms, it can be shown by analysis that for each frequency band, the larger the difference between the left and right target source PSD levels, the larger the noise overestimation is on that frequency band. More specifically, the larger the difference between the left and right SNRs at that particular frequency, the greater will be the overestimation for that frequency. Finally, it can easily be observed that PBNE closely tracks the original noise PSDs, leading to a better estimation, independently of the direction of arrival of the target source signal.

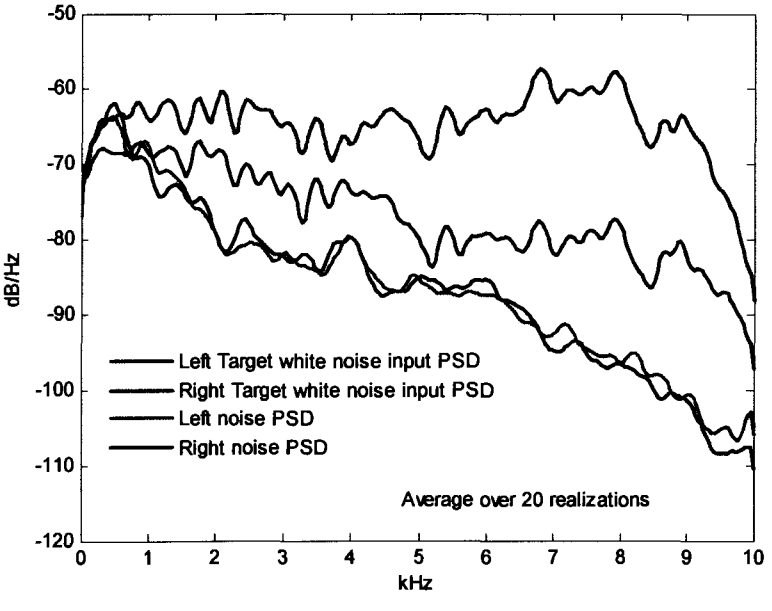

Figure 7-16: Scenario C, average over 20 realizations for the Left (blue) and right (red) white noise target source PSDs and corresponding left (black) and right (green) original noise PSDs

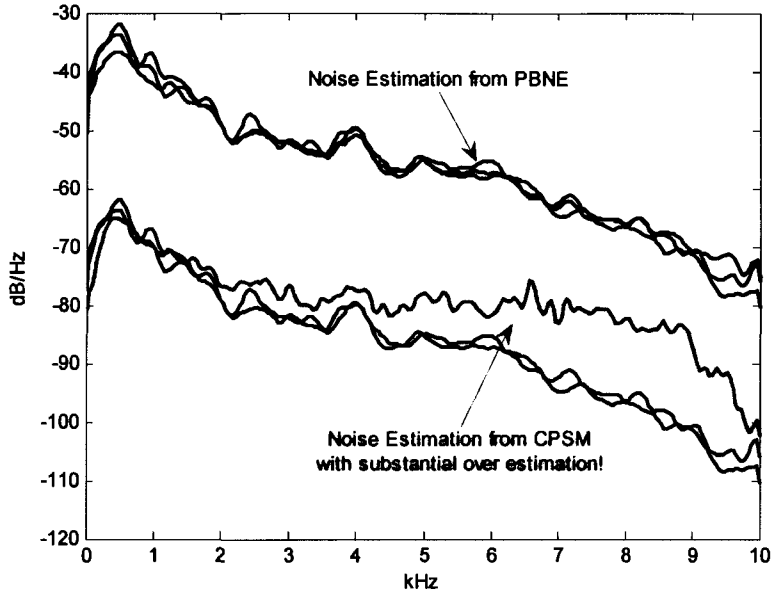

Figure 7-17: Noise estimation results for an average of 20 realizations for scenario C. PBNE (top - blue) versus CPSM (bottom – red), original left and right noise PSDs (top or bottom – black)

7.2.2 PBNE versus MSA

One of the drawbacks of MSA with respect to PBNE is that the technique requires knowledge of previous frames (i.e. previous noisy speech signal segments) in order to estimate the noise PSD on the current frame. Therefore, it also requires an initialization period before the noise estimation can be considered reliable. Also, a greater number of parameters (such as various smoothing parameters and search window sizes etc.) belonging to the technique must be chosen prior to run time. These parameters have a direct effect on the noise estimation accuracy and tracking latency in case of highly non-stationary noise. Secondly, as described in Section 3.1.1, the target source must be only a speech signal, since the algorithm estimates the noise within syllables, speech pauses, etc., with the assumption that the power of the speech signal often decays to the noise

power level. On the other hand, PBNE could be applied to any type of target source, as long as there is a degree of correlation between the received left and right signals.

Results for scenario A):

Figure 7-18 illustrates the left and right noisy speech signals. Since the MSA requires the knowledge of previous frames as opposed to PBNE or CPSM, the noise PSD estimation will not be compared on a frame-by-frame basis. MSA does not have an exact mathematical representation to estimate the noise PSD for a given frame only since it relies on the noise search over a range of past noisy speech signal frames. Unlike the preceding section where the noise estimation was obtained by averaging the results over multiple realizations (i.e. by processing the same speech frame index with different noise sequences), in this case it is not realistic to perform the same procedure because MSA can only find or update its noise estimation within a window of noisy speech frames as opposed to a single frame. Instead, to make an adequate comparison with PBNE, it is more suitable to make an average over the noise PSD estimates of consecutive frames. The left and right noisy speech signals represented in Fig. 7-18 have been decomposed into a total of 585 frames of 25.6ms with 50% for overlap at 20 kHz sampling frequency. The left and right noise-free speech PSDs and the left and right measured noise PSDs averaged over 585 frames are depicted in Fig. 7-19. It should be noted that all the PSD averaging has been done in linear scale. The left and right SNRs are approximately equal to 11 dB. Figures 7-20 to 7-22 illustrate the noise PSD estimation results from MSA versus PBNE, averaged over 585 frames at 11, 5 and 0 dB SNRs respectively. Only the noise estimation results on the right noisy speech signal are shown, since similar results were obtained for the left noisy signal. It can be observed that the accuracy of PBNE noise estimation is higher than the one from MSA. It can also be seen that the PBNE performance is maintained at the various input SNRs in contrast to MSA, where the accuracy is reduced at lower SNRs.

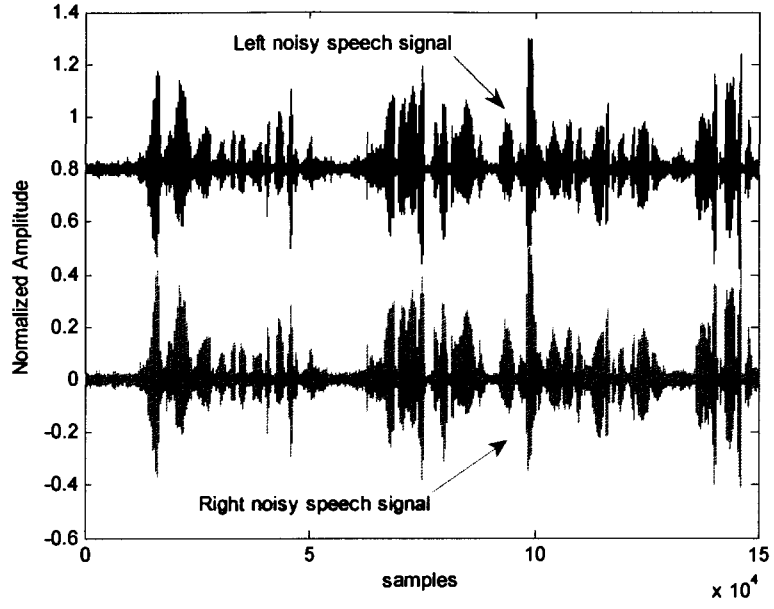

Figure 7-18: Left (top - black) and right (bottom - red) noisy speech signals

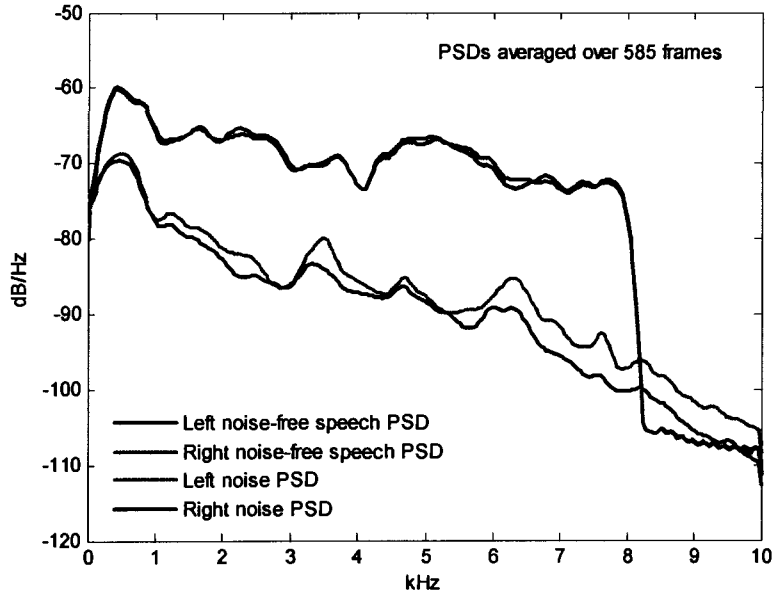

Figure 7-19: Left (blue) and right (red) noise-free speech PSDs and corresponding left (black) and right (green) original noise PSDs used for scenario A at SNR=11 dB

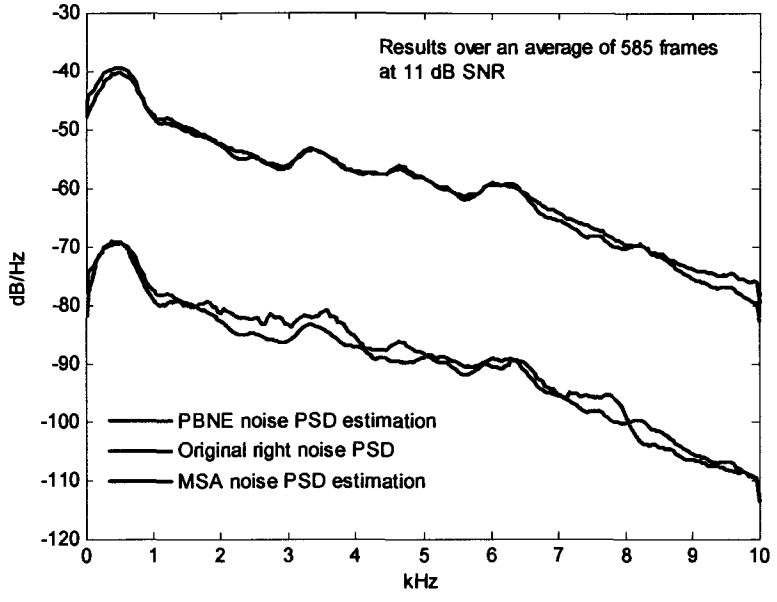

Figure 7-20: Noise estimation results over an average of 585 frames and at SNR = 11dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSDs (top or bottom - black)

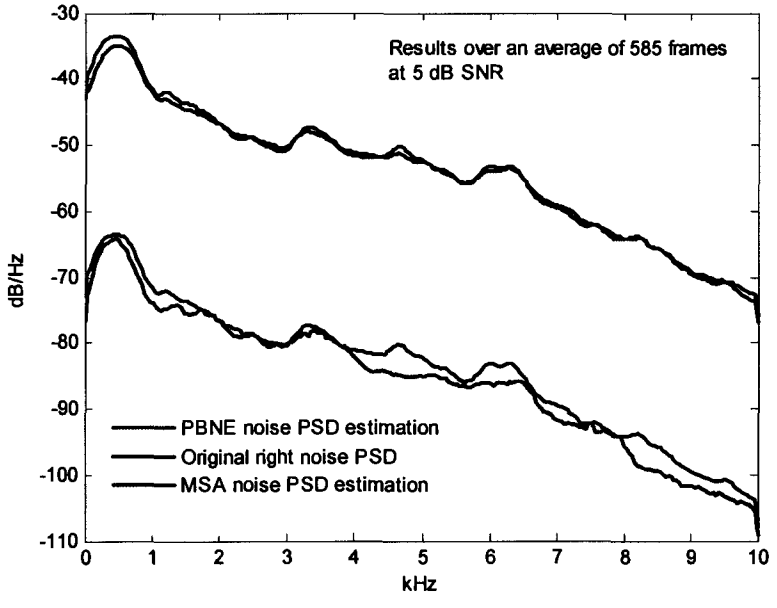

Figure 7-21: Noise estimation results over an average of 585 frames and at SNR = 5dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSD (top or bottom - black)

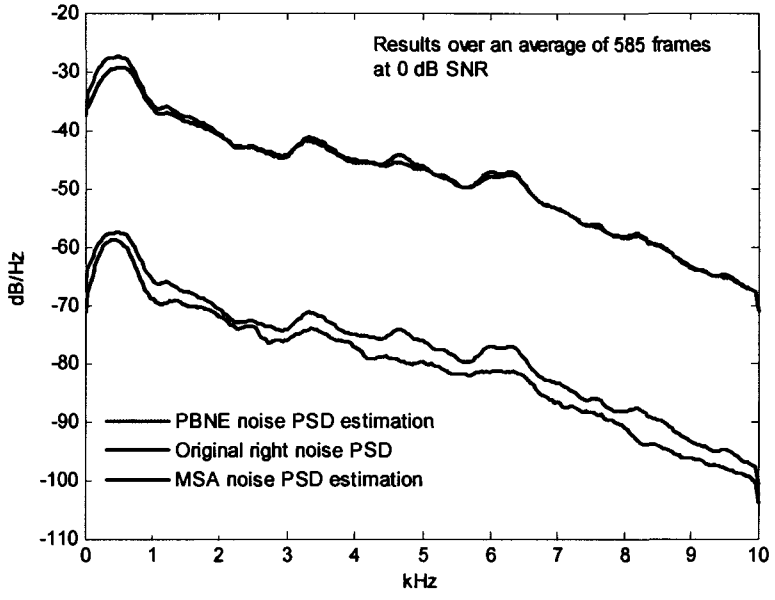

Figure 7-22: Noise estimation results over an average of 585 frames and at SNR = 0dB for scenario A. PBNE (top - blue) versus MSA (bottom – red), right noise PSD (top or bottom - black)

Results for scenario D:

In this scenario, the noise tracking capability of both techniques is evaluated in the event of a jump or a drop of the noise power level for various reasons. For instance, if the hearing aid user is leaving or entering a crowded cafeteria, or just relocating to a less noisy area. To simulate those conditions, the original noise power has been increased by 12 dB at frame index 200 and then reduced again by 12 dB from frame index 400. Fig. 7-23 shows the power of the noisy right input speech signal for each frame and the corresponding noise power with the sudden fluctuations. To perform the comparison, the total noise power calculated for each frame has been compared with the corresponding total noise power estimates (evaluated by integrating the noise PSD estimates) at each frame. The results from MSA and PBNE are shown in Figs. 7-24 and 7-25 respectively. Again, only the noise estimation results on the right noisy speech signal are shown. As it can be noticed, MSA experiences some latency tracking the noise jump. This latency is

related to the tree search implementation in the MSA technique [MAR'01], discussed in Section 3.1.1. It is essentially governed by the selected number of sub-windows, U, and the number of frames, V, in each sub-window. In [MAR'01], the latency for a substantial noise jump is given as follows: $Latency = U \cdot V + V$. For this scenario, U was assigned a value of 8 and V a value of 6, giving a latency of 56 frames, as demonstrated in Fig. 7-24.. For a sudden noise drop, the latency is equal to a maximum of V frames [MAR'01]. Fortunately, the latency is much lower for a sudden noise decrease as it can be seen in Fig. 7-24. Otherwise, having a long period of noise overestimation when used by a noise reduction scheme would greatly attenuate the target speech signal, therefore affecting its intelligibility. Of course, it is possible to reduce the latency by shrinking the search window length but the drawback is that the accuracy will be lowered as well. The search window length (i.e. $U \cdot V$) must be large enough to bridge any speech activity, but short enough to track non-stationary noise fluctuations. It is a trade-off of the MSA method. On the other hand, as expected PBNE tracks easily the increase or the decrease of the noise power level, since the algorithm relies only on the current frame being processed.

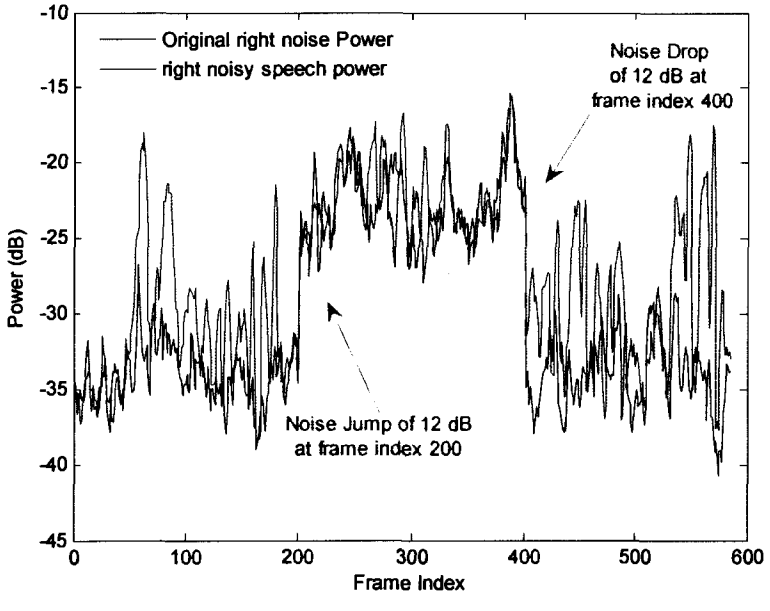

Figure 7-23: Noisy right speech signal power (blue) at each frame and corresponding noise power (back) increased at frame index 200 and then decreased at index 400 (black), used in Scenario D

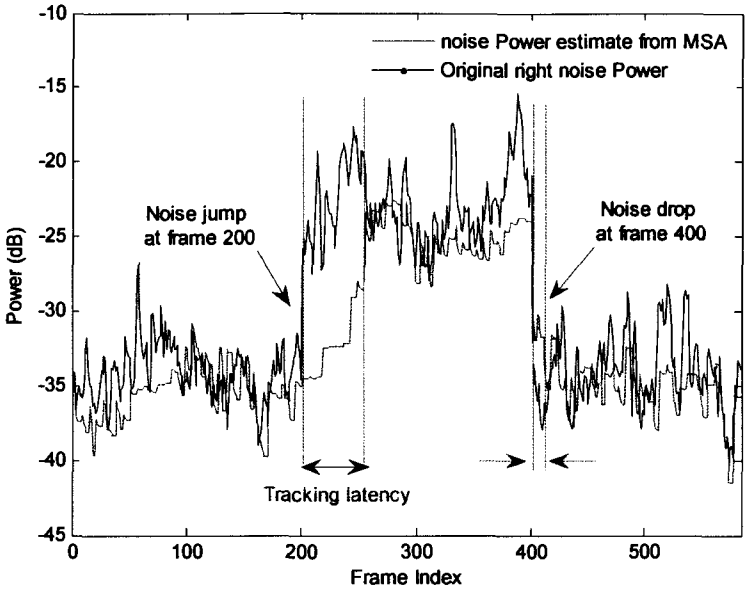

Figure 7-24: Noise power estimation result at each frame for MSA in scenario D with highly non-stationary noise. MSA (red), original right noise power (black)

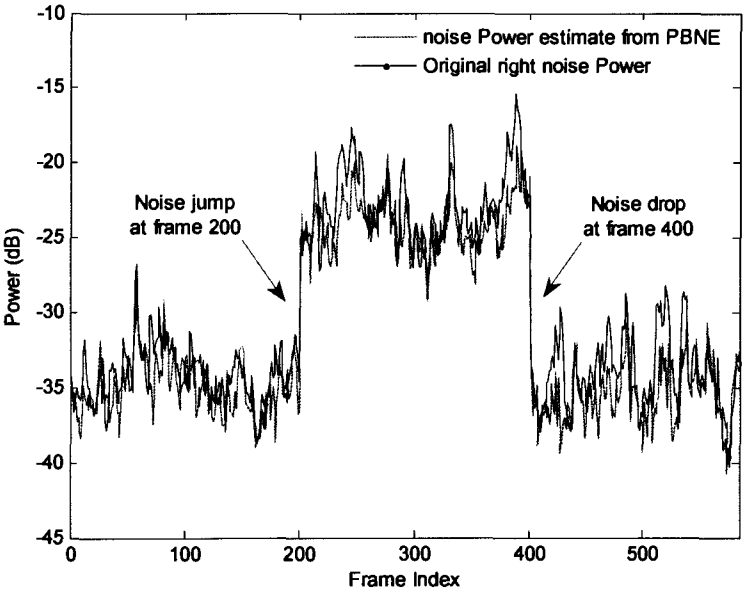

Figure 7-25: Noise power estimation result for PBNE at each frame in scenario D with highly non-stationary noise. PBNE (red), original right noise power (black)

7.2.3 Summary of Proposed Binaural Noise Estimation Technique Performance

To summarize, the proposed binaural noise estimation technique developed in Chapter 6 has been designed for a diffuse noise field environment. The target speaker can be at any location around the binaural hearing aid user, as long as the speaker is at proximity of the hearing aid user in the noisy environment. Therefore, the direction of arrival of the source speech signal can be arbitrary. Also, the target source signal can be other than a speech signal, as long as there is a high degree of correlation between the left and right noisy signals. The noise estimation is accurate even at high or low SNRs, and it is performed on a frame-by-frame basis. It does not employ any voice activity detection algorithm, and the noise can be estimated during speech activity or not. It can track highly non-stationary noise conditions and any type of colored noise, provided that the noise has diffuse field characteristics. Also, in practice, if the noise is considered stationary over several frames, to further increase the accuracy the noise estimation can be achieved by averaging estimates obtained over consecutives frames.

Chapter 8. BINAURAL TARGET SPEECH PSD ESTIMATOR FOR TARGET SPEECH CORRUPTED BY DIRECTIONAL NOISE

In Chapter 6, a proposed binaural noise PSD estimator was designed for hearings aids operating in a diffuse noise field environment. However, other types of noise such as directional noise sources are frequently encountered in real-life listening situations and can reduce greatly the understanding of the target speech. For instance, directional noise sources can emerge from strong multi-talkers in addition to permanent diffuse noise in the background. This situation really degrades speech intelligibility since some other issues may arise such as informational masking (defined as the interfering speech carrying linguistic content, which can be confused with the content of the target speaker [HAW'04]), which has an even greater negative impact for a hearing impaired individual. Also, transient lateral noise may occur in the background such as hammering, dishes clattering etc. Those intermittent noises can create unpleasant auditory sensations even in a quiet environment i.e. without diffuse background noise.

In a monaural system where only a single channel is available for processing, the use of spatial information is not feasible. Consequently it is very difficult for instance to distinguish between the speech coming from a target speaker or from interferers, unless the characteristics of the lateral noise/interferers are known in advance, which is not realistic in real life situations. Also, most monaural noise estimation schemes such as the noise PSD estimation using minimum statistics in [MAR'01] assume that the noise characteristics vary at a much slower rate than the target speech signal. Therefore, noise estimation schemes such as in [MAR'01] will not detect for instance lateral transient noise like dishes clattering, hammering sounds, etc..

As a solution to mitigate the impact of one or two dominant directional noise sources, high-end multi-microphone (but still monaural i.e. on one ear only) hearing aids incorporate advanced directional microphones where directivity is achieved for example by differential processing of two omni-directional microphones placed on the hearing aid

[HAM'05]. The directivity can also be adaptive so that it can constantly estimate the direction of the noise arrival and then steer a notch in a "beampattern" to match the main direction of the noise arrival. Two or three microphone array systems (i.e. beamformers) provide great benefits in today's high end hearing aids, however due to size constrains only certain models such as Behind-The-Ear (BTE) can accommodate two or even three microphones. Smaller models such as In-The-Canal (ITC) or In-The-Ear (ITE) only permits the fitting of a single microphone. Consequently beamforming cannot be applied for such cases.

Furthermore, it has been reported that hearing impaired individuals localize sounds better without their bilateral hearing aids (or by having the noise reduction program switched off) than with them [BOG'06]. This is due to the fact that current noise reduction schemes implemented in bilateral (i.e. combination of two monaural, one from each side) hearing aids are not designed to preserve localization cues. As a result, it creates an inconvenience for the hearing aid user and it should be pointed out that in some cases such as in street traffic, incorrect sound localization may be endangering.

Thus, all the reasons above provide a further motivation to place more importance towards a binaural system and to investigate the potential improvement of current noise reduction schemes against noise coming from lateral directions, such as an interfering background talker or transient noise, and most importantly without altering the interaural cues of both the speech and the noise.

In this chapter, an instantaneous binaural target speech PSD estimator is developed, where the target speech PSD is retrieved from the received binaural noisy signals corrupted by lateral interfering noise. The proposed estimator does not require the knowledge of the direction of the noise source (i.e. computations of Interaural Transfer Functions (ITFs) are not required). The noise can be highly non-stationary (i.e. fluctuating noise statistics) such as an interfering speech signal from a background talker or just transient noise (i.e. dishes clattering or door opening/closing in the background). Moreover, the estimator does not require a voice activity detector (VAD) or any

classification, and it is performed on a frame-by-frame basis with no memory (which is the rationale for calling the proposed estimator "instantaneous"). Consequently, the background noise source can also be moving (or equivalently, switching from one main interfering noise source to another at a different direction). This section will focus on the scenario where the target speaker is assumed to remain in front of the binaural hearing aid user, although it will be shown in the last subsection that the proposed target source PSD estimator can also be extended to non-frontal target source directions. In practice, a signal coming from the front is often considered to be the desired target signal direction, especially in the design of standard directional microphones implemented in hearing aids [HAM'05],[PUD'06].

Section 8.1 will explain the selected acoustical environment where the target power spectrum density is estimated for binaural hearing aids. Section 8.2 will describe the binaural system adapted to this environment, with signal definitions, and finally in Section 8.3, the proposed binaural target spectrum density estimator will be demonstrated in detail.

It should be noted that the next chapter will show an example on how to integrate the proposed target speech PSD estimator developed in this chapter into a selected noise reduction scheme, to remove lateral noises with preservation of all interaural cues.

8.1 Acoustical Environment: Lateral (Directional) Noise

In the considered environment, the binaural hearing aids user is in front of the target speaker with a strong lateral interfering noise in the background as illustrated in Fig. 8-1. The interfering noise can be a background talker (i.e. speech-like characteristic), which often occurs when chatting in a crowded cafeteria, or it can be dishes clattering, hammering sounds in the background etc., which are referred to as transient noise. Those types of noise are characterized as being highly non-stationary and may occur at random instants around the target speaker in real-life environments. Moreover, those noise signals

are referred to as localized noise sources or directional noise. In the presence of a localized noise source as opposed to a diffuse noise field environment, the noise signals received by the left and right microphones are highly correlated. In the considered environment, the noise can originate anywhere around the binaural hearing aids user, implying that the direction of arrival of the noise is arbitrary, however it should differ from 0° (or from the assumed target direction) to provide a spatial separation between the target speech and the noise.

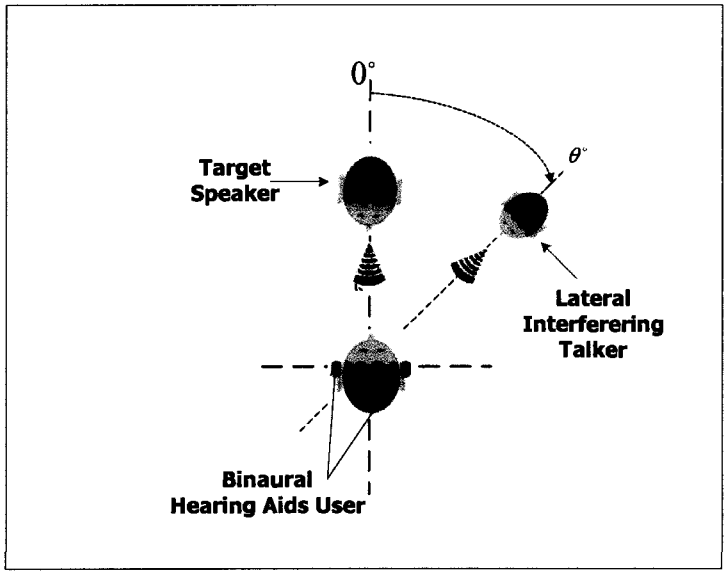

Figure 8-1: Binaural hearing aid user in front of the target speaker and a lateral interfering talker in background

8.2 Binaural System Description for Proposed Problem

Let $l(i)$, $r(i)$ be the noisy signals received at the left and right hearing aid microphones, defined here in the temporal domain as:

$$\begin{aligned} l(i) &= s(i) \otimes h_l(i) + v(i) \otimes k_l(i) \\ &= s_l(i) + v_l(i) \end{aligned} \quad \text{(EQ 8-1)}$$

$$\begin{aligned} r(i) &= s(i) \otimes h_r(i) + v(i) \otimes k_r(i) \\ &= s_r(i) + v_r(i) \end{aligned} \quad \text{(EQ 8-2)}$$

where $s(i)$ and $v(i)$ are the target and interfering directional noise sources respectively. It is assumed that the distance between the speaker and the two microphones (one placed on each ear) is such that they receive essentially speech through a direct path from the speaker. This implies that the received target speech left and right signals are highly correlated (i.e. the direct component dominates its reverberation components). The same reasoning applies for the interfering directional noise. The left and right received noise signals are then also highly correlated as opposed to diffuse noise, where left and right received signals would be poorly correlated over most of the frequency spectrum. Hence, in the context of binaural hearing, $h_l(i)$ and $h_r(i)$ are the left and right head-related impulse responses (HRIRs) between the target speaker and the left and right hearing aids microphones. $k_l(i)$ and $k_r(i)$ are the left and right head-related impulse responses between the interferer and the left and right hearing aids microphones. As a result, $s_l(i)$ is the received left target speech signal and $v_l(i)$ corresponds to the lateral interfering noise on the left channel. Similarly, $s_r(i)$ is the received right target speech signal and $v_r(i)$ corresponds to the lateral interfering noise received on the right channel.

Prior to estimating the target speech PSD, the following assumptions are made:

i) The target speech and the interfering noise are not correlated

ii) The direction of arrival of the target source speech signal is approximately frontal that is: $h_l(i) \approx h_r(i) = h(i)$ (EQ 8-3)

(note: the case of a non-frontal target source is discussed in Section 8.3.4)

iii) the noise source can be anywhere around the hearing aids user, that is the direction of arrival of the noise signal is arbitrary but it can not be the same as the target source, otherwise it will be considered as a target source.

Using the assumptions above along with Equations (8-1) and (8-2), the left and right auto power spectral densities, $\Gamma_{LL}(\omega)$ and $\Gamma_{RR}(\omega)$, can be expressed as the following:

$$\Gamma_{LL}(\omega) = \Gamma_{SS}(\omega)|H(\omega)|^2 + \Gamma_{VV}(\omega)|K_L(\omega)|^2 \quad \text{(EQ 8-4)}$$

$$\Gamma_{RR}(\omega) = \Gamma_{SS}(\omega)|H(\omega)|^2 + \Gamma_{VV}(\omega)|K_R(\omega)|^2 \quad \text{(EQ 8-5)}$$

8.3 Proposed Binaural Instantaneous Target Speech Spectrum Estimator

In this section, a new binaural target speech spectrum estimation method is developed. Subsection 8.3.1 presents the overall diagram of the proposed target speech spectrum estimation. It is shown that the target speech spectrum estimate is found by initially applying a Wiener filter to perform a prediction of the left noisy speech signal from the right noisy speech signal, followed by taking the difference between the auto-power spectral density of the left noisy signal and the auto-power spectral density of its prediction. As a second step, an equation is formed by combining the PSD of this difference signal, the auto-power spectral densities of the left and right noisy speech signals and the cross-power spectral density between the left and right noisy signals. The solution of the equation provides the target speech PSD. In practice, similar to the implementation of the binaural diffuse noise power spectrum density estimator in Chapter

6, the estimation of one of the variables used in the equation causes the target speech power spectrum estimation to be less accurate in some cases. However, there are two ways of computing this variable: an indirect form, which is obtained from a combination of several other variables, and a direct form, which is less intuitive. It was observed through empirical results that combining the two estimates (obtained using the direct and indirect computations) provides a better target speech power spectrum estimation. Therefore, Subsection 8.3.2 will present the alternate way (i.e. the direct form) of computing the estimate, and finally Subsection 8.3.3 will show the effective combination of those two estimates (i.e. direct and indirect forms), finalizing the proposed target speech power spectrum estimation technique. Subsection 8.3.4 will show how to extend the proposed target source PSD estimator to non-frontal target source directions.

8.3.1 Target Speech PSD Estimation

Figure 8-2 shows a diagram of the overall proposed estimation method. It includes a Wiener prediction filter and the final equation estimating the target speech power spectral density.

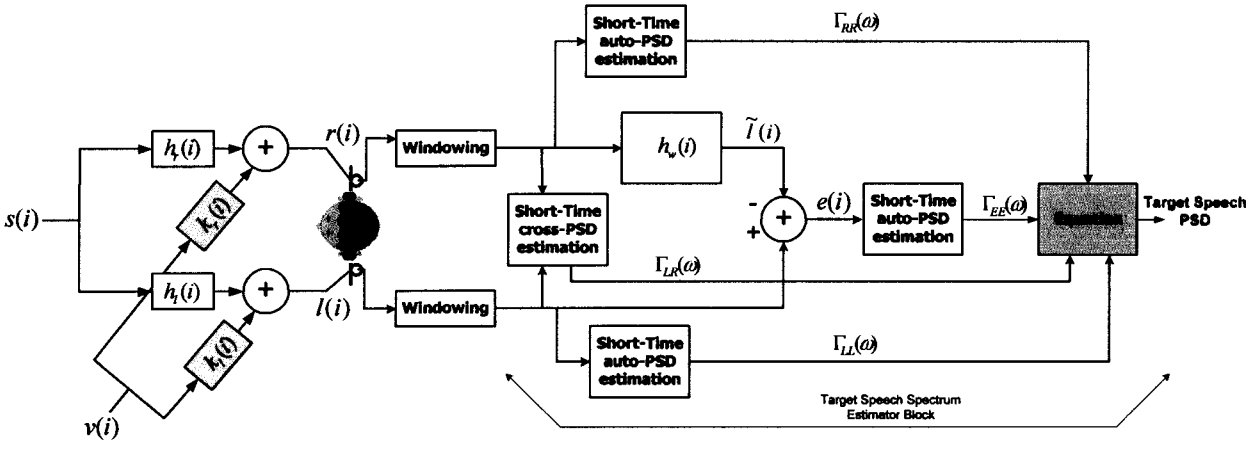

Figure 8-2: Overall diagram for the proposed binaural target speech PSD estimation method

In a first step, a filter $h_w^r(i)$ is used to perform a linear prediction of the left noisy speech signal from the right noisy speech signal. Using a minimum mean square error criterion (MMSE), the optimum solution is the Wiener solution, defined here in the frequency domain as:

$$H_W^R(\omega) = \Gamma_{LR}(\omega)/\Gamma_{RR}(\omega) \qquad \text{(EQ 8-6)}$$

where $\Gamma_{LR}(\omega)$ is the cross-power spectral density between the left and the right noisy signals.

$\Gamma_{LR}(\omega)$ is obtained as follows:

$$\Gamma_{LR}(\omega) = F.T.\{\gamma_{lr}(\tau)\} = F.T.\{E[l(i+\tau)\cdot r(i)]\} \qquad \text{(EQ 8-7)}$$

with:

$$\gamma_{lr}(\tau) = E\left(\left[s(i+\tau)\otimes h_l(i) + v(i+\tau)\otimes k_l(i)\right]\cdot\left[s(i)\otimes h_r(i) + v(i)\otimes k_r(i)\right]\right)$$
$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau) + \gamma_{vv}(\tau)\otimes k_l(\tau)\otimes k_r(-\tau) + \gamma_{sv}(\tau)\otimes h_l(\tau)\otimes k_r(-\tau) + \gamma_{vs}(\tau)\otimes k_l(\tau)\otimes h_r(-\tau) \qquad \text{(EQ 8-8)}$$

Using the previously defined assumptions in Section 8.2, Equation (8-8) can then be simplified to:

$$\gamma_{lr}(\tau) = \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau) + \gamma_{vv}(\tau)\otimes k_l(\tau)\otimes k_r(-\tau) \qquad \text{(EQ 8-9)}$$

The cross-power spectral density expression then becomes:

$$\Gamma_{LR}(\omega) = \Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega) + \Gamma_{VV}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega) \qquad \text{(EQ 8-10)}$$

$$= \Gamma_{SS}(\omega)\cdot|H(\omega)|^2 + \Gamma_{VV}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega) \qquad \text{(EQ 8-11)}$$

Using (8-6), the squared magnitude response of the Wiener filter is computed as follows:

$$\left|H_W^R(\omega)\right|^2 = \frac{\left|\Gamma_{LR}(\omega)\right|^2}{\Gamma_{RR}^2(\omega)} = \frac{\Gamma_{LR}(\omega)\cdot\Gamma_{LR}^*(\omega)}{\Gamma_{RR}^2(\omega)} \qquad \text{(EQ 8-12)}$$

Furthermore, substituting (8-10) into (8-11) the squared magnitude response of the Wiener filter in (8-12) can also be expressed as:

$$\left|H_W^R(\omega)\right|^2 = \frac{1}{\Gamma_{RR}^2(\omega)}\left\{\begin{pmatrix}\Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)\\ +\Gamma_{VV}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)\end{pmatrix}\cdot\begin{pmatrix}\Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)\\ +\Gamma_{VV}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)\end{pmatrix}^*\right\} \quad \text{(EQ 8-13)}$$

$$= \frac{1}{\Gamma_{RR}^2(\omega)}\left\{\begin{matrix}\Gamma_{SS}^2(\omega)\cdot\left|H_L(\omega)\right|^2\cdot\left|H_R(\omega)\right|^2 + \Gamma_{SS}(\omega)\cdot\Gamma_{VV}(\omega)\cdot\begin{pmatrix}H_L^*(\omega)\cdot H_R(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)\\ +H_L(\omega)\cdot H_R^*(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega)\end{pmatrix}\\ +\Gamma_{VV}^2(\omega)\cdot\left|K_L(\omega)\right|^2\cdot\left|K_R(\omega)\right|^2\end{matrix}\right\}$$

(EQ 8-14)

$$= \frac{1}{\Gamma_{RR}^2(\omega)}\left\{\begin{matrix}\left(\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)^2 + \Gamma_{SS}(\omega)\cdot\Gamma_{VV}(\omega)\cdot\left|H(\omega)\right|^2\cdot\begin{pmatrix}K_L(\omega)\cdot K_R^*(\omega)+\\ K_L^*(\omega)\cdot K_R(\omega)\end{pmatrix}+\\ \left(\Gamma_{VV}(\omega)\cdot\left|K_L(\omega)\right|\cdot\left|K_R(\omega)\right|\right)^2\end{matrix}\right\}$$

(EQ 8-15)

In the previous equation, the left and right directional noise interferer HRTFs are still unknown parameters, however they can be substituted into (8-15) using (8-11) as well, as its complex conjugate form, as follows:

$$\left|H_W^R(\omega)\right|^2 = \frac{1}{\Gamma_{RR}^2(\omega)}\left\{\begin{matrix}\left(\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)^2 + \Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\cdot\begin{pmatrix}\left(\Gamma_{LR}(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)\\ +\left(\Gamma_{LR}^*(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)\end{pmatrix}\\ +\Gamma_{VV}^2(\omega)\cdot\left|K_L(\omega)\right|^2\cdot\left|K_R(\omega)\right|^2\end{matrix}\right\}$$

(EQ 8-16)

From (8-16), the remaining unknown parameters (such as in the left and right directional noise HRTFs magnitudes) can be substituted using (8-4) and (8-5) as follows:

$$\left|H_W^R(\omega)\right|^2 = \frac{1}{\Gamma_{RR}^2(\omega)}\left\{\begin{matrix}\left(\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)^2 + \Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\cdot\begin{pmatrix}\left(\Gamma_{LR}(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)+\\ \left(\Gamma_{LR}^*(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)\end{pmatrix}\\ +\left(\Gamma_{LL}(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)\cdot\left(\Gamma_{RR}(\omega)-\Gamma_{SS}(\omega)\cdot\left|H(\omega)\right|^2\right)\end{matrix}\right\}$$

(EQ 8-17)

After simplification and rearranging the terms in (8-17), the target speech PSD is found by solving the following equation:

$$\Gamma_{SS}(\omega)\cdot|H(\omega)|^2 = \frac{\Gamma_{RR}(\omega)\cdot\Gamma^R_{EE\_1}(\omega)}{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\left(\Gamma_{LR}(\omega)+\Gamma^*_{LR}(\omega)\right)} \qquad \text{(EQ 8-18)}$$

$$= \Gamma^R_{SS}(\omega)$$

where $\Gamma^R_{EE\_1}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{RR}(\omega)\cdot|H^R_W(\omega)|^2$ (EQ 8-19)

It should be noted that the Wiener filter coefficients used in (8-19) were computed using the right noisy speech signal as a reference input to predict the left channel, as illustrated in Fig. 8-1. However, to diminish the distortion on the interfering noise spatial cues, when audible residual interfering noise still remains in the estimated target speech spectrum, the target speech PSD should also be estimated by using the dual procedure, that is: using the left noisy speech signal input as a reference for the Wiener filter instead of the right. This configuration for the setup of the Wiener filter is referred to as $H^L_W(\omega)$ or as $h^l_w(\omega)$ in the time domain.

To sum up, the target speech PSD retrieved from the right channel is referred to as $\Gamma^R_{SS}(\omega)$ and is found using (8-18) and (8-19). Similarly, the target speech PSD retrieved from the left channel is referred to as $\Gamma^L_{SS}(\omega)$ and is found using the following equations:

$$\Gamma^L_{SS}(\omega) = \frac{\Gamma_{LL}(\omega)\cdot\Gamma^L_{EE\_1}(\omega)}{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right)-\left(\Gamma_{LR}(\omega)+\Gamma^*_{LR}(\omega)\right)} \qquad \text{(EQ 8-20)}$$

where $\Gamma^L_{EE\_1}(\omega) = \Gamma_{RR}(\omega) - \Gamma_{LL}(\omega)\cdot|H^L_W(\omega)|^2$ (EQ 8-21)

and the Wiener filter coefficients in (8-21) are computed using the left noisy channel as a reference input to predict the right channel.

8.3.2 Direct Computation of the Error Auto-Power Spectrum

As briefly introduced in Section 8.3, the accuracy of the retrieved target speech PSD can be improved by adjusting the estimate of the variables $\Gamma^R_{EE\_1}(\omega)$ and $\Gamma^L_{EE\_1}(\omega)$ used in

(8-18) and (8-20). For the remaining part of this subsection, we will focus on $\Gamma^R_{EE\_1}(\omega)$, but the same development applies to $\Gamma^L_{EE\_1}(\omega)$. As shown in Equation (8-19), $\Gamma^R_{EE\_1}(\omega)$ is obtained by taking the difference between the auto-power spectral density of the left noisy signal and the auto-power spectral density of its prediction. However, it will be shown in this subsection that $\Gamma^R_{EE\_1}(\omega)$ is in fact the auto-power spectral density of the prediction residual (or error) $e(i)$ shown in Fig. 8-2, which is somewhat less intuitive. The direct computation of this auto-power spectral density from the samples of $e(i)$ is referred to as $\Gamma^R_{EE}(\omega)$ here, while the indirect computation using (8-19) is referred to as $\Gamma^R_{EE\_1}(\omega)$. $\Gamma^R_{EE\_1}(\omega)$ and $\Gamma^R_{EE}(\omega)$ are theoretically equivalent, however only estimates of those power spectral densities are available in practice to compute (8-5), (8-18) and (8-19). It was found through empirical results that the estimation of $\Gamma^R_{SS}(\omega)$ in (8-18) yields a more accurate result by using $\Gamma^R_{EE\_1}(\omega)$ or $\Gamma^R_{EE}(\omega)$ depending on different cases, and that using a combination of both performs better. The next subsection will show the appropriate use of $\Gamma^R_{EE\_1}(\omega)$ and $\Gamma^R_{EE}(\omega)$ for the estimation of $\Gamma^R_{SS}(\omega)$.

In Chapter 6, using a similar binaural system, the analytical equivalence between $\Gamma^R_{EE}(\omega)$ and $\Gamma^R_{EE\_1}(\omega)$ was derived in details for the hearing scenario where the binaural hearing aids user is located in a diffuse background noise. This section deals with directional background noise instead. Using similar derivation steps as in Chapter 6, it is possible to prove again that $\Gamma^R_{EE}(\omega)$ and $\Gamma^R_{EE\_1}(\omega)$ are analytically equivalent.

Starting from the prediction residual error as shown in Fig. 8-2, which can be defined as:

$$e(i) = l(i) - \tilde{l}(i) = l(i) - r(i) \otimes h^r_w(i) \qquad \text{(EQ 8-22)}$$

we have:

$$\Gamma^R_{EE}(\omega) = F.T.\left(\gamma_{ee}(\tau)\right) \qquad \text{(EQ 8-23)}$$

where

$$\gamma_{ee}(\tau) = E\left(e(i+\tau)\cdot e(i)\right)$$

$$= E\left(\left[l(i+\tau)-\tilde{l}(i+\tau)\right]\cdot\left[l(i)-\tilde{l}(i)\right]\right)$$
$$= E\left[l(i+\tau)l(i)\right]-E\left[l(i+\tau)\tilde{l}(i)\right]-E\left[\tilde{l}(i+\tau)l(i)\right]+E\left[\tilde{l}(i+\tau)\tilde{l}(i)\right]$$
$$= \gamma_{ll}(\tau)-\gamma_{l\tilde{l}}(\tau)-\gamma_{\tilde{l}l}(\tau)+\gamma_{\tilde{l}\tilde{l}}(\tau) \quad \text{(EQ 8-24)}$$

As derived in (8-24), $\gamma_{ee}(\tau)$ is thus the sum of 4 terms, where the following temporal and frequency domain definitions for each term are:

$$\gamma_{ll}(\tau) = E\left(\left[s(i+\tau)\otimes h_l(i)+v(i+\tau)\otimes k_l(i)\right]\cdot\left[s(i)\otimes h_l(i)+v(i)\otimes k_l(i)\right]\right) \quad \text{(EQ 8-25)}$$

$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_l(-\tau)+\gamma_{vv}(\tau)\otimes k_l(\tau)\otimes k_l(-\tau) \quad \text{(EQ 8-26)}$$

$$\Gamma_{LL}(\omega) = \Gamma_{SS}(\omega)\,|\,H_L(\omega)\,|^2 + \Gamma_{YY}(\omega)\,|\,K_L(\omega)\,|^2 \quad \text{(EQ 8-27)}$$

$$\gamma_{l\tilde{l}}(\tau) = E\left(\left[s(i+\tau)\otimes h_l(i)+v(i+\tau)\otimes k_l(i)\right]\cdot\left[\left[s(i)\otimes h_r(i)+v(i)\otimes k_r(i)\right]\otimes h_w^r(i)\right]\right)$$
$$= \gamma_{ss}(\tau)\otimes h_l(\tau)\otimes h_r(-\tau)\otimes h_w^r(-\tau)+\gamma_{vv}(\tau)\otimes k_l(\tau)\otimes k_r(-\tau)\otimes h_w^r(-\tau)$$
(EQ 8-28)

$$\Gamma_{L\tilde{L}}(\omega) = \Gamma_{SS}(\omega)H_L(\omega)H_R^*(\omega)\left(H_W^R(\omega)\right)^* + \Gamma_{VV}(\omega)K_L(\omega)K_R^*(\omega)\left(H_W^R(\omega)\right)^*$$
(EQ 8-29)

$$\gamma_{\tilde{l}l}(\tau) = E\left(\left(\left[s(i+\tau)\otimes h_r(i)+v(i+\tau)\otimes k_r(i)\right]\otimes h_w^r(i)\right)\cdot\left[s(i)\otimes h_l(i)+v(i)\otimes k_l(i)\right]\right)$$
$$= \gamma_{ss}(\tau)\otimes h_l(-\tau)\otimes h_r(\tau)\otimes h_w^r(\tau)+\gamma_{vv}(\tau)\otimes k_l(-\tau)\otimes k_r(\tau)\otimes h_w^r(\tau)$$
(EQ 8-30)

$$\Gamma_{\tilde{L}L}(\omega) = \Gamma_{SS}(\omega)H_L^*(\omega)H_R(\omega)H_W^R(\omega) + \Gamma_{VV}(\omega)K_L^*(\omega)K_R(\omega)H_W^R(\omega)$$
(EQ 8-31)

$$\gamma_{\tilde{l}\tilde{l}}(\tau) = E\left(\left(\left[\begin{matrix}s(i+\tau)\otimes h_r(i)+\\ v(i+\tau)\otimes k_r(i)\end{matrix}\right]\otimes h_w^r(i)\right)\cdot\left(\left[\begin{matrix}s(i)\otimes h_r(i)+v(i+\tau)\\ \otimes k_r(i)\end{matrix}\right]\otimes h_w^r(i)\right)\right)$$

$$= \gamma_{ss}(\tau)\otimes h_r(\tau)\otimes h_r(-\tau)\otimes h_w^r(\tau)\otimes h_w^r(-\tau)+\gamma_{vv}(\tau)\otimes k_r(\tau)\otimes k_r(-\tau)\otimes h_w^r(\tau)\otimes h_w^r(-\tau)$$
(EQ 8-32)

$$\Gamma_{\tilde{L}\tilde{L}}(\omega) = \Gamma_{SS}(\omega)\,|H_R(\omega)|^2\,|H_W^R(\omega)|^2 + \Gamma_{VV}(\omega)\,|K_R(\omega)|^2\,|H_W^R(\omega)|^2 \qquad \text{(EQ 8-33)}$$

From (8-24), we can write:

$$\Gamma_{EE}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{L\tilde{L}}(\omega) - \Gamma_{\tilde{L}L}(\omega) + \Gamma_{\tilde{L}\tilde{L}}(\omega) \qquad \text{(EQ 8-34)}$$

and substituting all the terms in their respective frequency domain forms (i.e. (8-27), (8-29), (8-31) and (8-33)) into (8-34) yields:

$$\begin{aligned}\Gamma_{EE}(\omega) &= \Gamma_{SS}(\omega)\cdot|H_L(\omega)|^2 + \Gamma_{SS}(\omega)\cdot|H_L(\omega)|^2\cdot|H_W^R(\omega)|^2 + \Gamma_{SS}(\omega)\cdot|H_R(\omega)|^2\cdot|H_W^R(\omega)|^2 - 2\cdot\Gamma_{AA}(\omega)\\ &= \Gamma_{LL}(\omega) + \Gamma_{RR}(\omega)\cdot|H_R(\omega)|^2 - \Gamma_{AA}(\omega)\end{aligned} \qquad \text{(EQ 8-35)}$$

where:

$$\begin{aligned}\Gamma_{AA}(\omega) &= \Gamma_{SS}(\omega)\cdot\left(H_L(\omega)\cdot H_R^*(\omega)\cdot\left(H_W^R(\omega)\right)^* + H_L^*(\omega)\cdot H_R(\omega)\cdot H_W^R(\omega)\right)\\ &\quad + \Gamma_{YY}(\omega)\cdot\begin{pmatrix}K_L(\omega)\cdot K_R^*(\omega)\cdot\left(H_W^R(\omega)\right)^*\\ +K_L^*(\omega)\cdot K_R(\omega)\cdot H_W^R(\omega)\end{pmatrix}\\ &= 2\cdot\mathrm{Re}\left\{\left(\Gamma_{SS}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) + \Gamma_{VV}(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega)\right)\cdot H_W^R(\omega)\right\}\end{aligned} \qquad \text{(EQ 8-36)}$$

Substituting Equations (8-6) and (8-10) into (8-36), $\Gamma_{AA}(\omega)$ is equal to:

$$\begin{aligned}&= 2\cdot\mathrm{Re}\left\{\frac{\left(\Gamma_{SS}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega) + \Gamma_{YY}(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega)\right)\cdot}{\Gamma_{SS}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega) + \Gamma_{VV}(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega)}{\Gamma_{RR}(\omega)}\right\}\\ &= \frac{2}{\Gamma_{RR}(\omega)}\cdot\left\{\begin{matrix}\left(\Gamma_{SS}(\omega)\cdot|H_L(\omega)|\cdot|H_R(\omega)|\right)^2 + \Gamma_{SS}(\omega)\cdot\Gamma_{VV}(\omega)\cdot H_L^*(\omega)\cdot H_R(\omega)\cdot K_L(\omega)\cdot K_R^*(\omega) +\\ \Gamma_{SS}(\omega)\cdot\Gamma_{VV}(\omega)\cdot H_L(\omega)\cdot H_R^*(\omega)\cdot K_L^*(\omega)\cdot K_R(\omega) + \left(\Gamma_{VV}(\omega)\cdot|K_L(\omega)|\cdot|K_R(\omega)|\right)^2\end{matrix}\right\}\end{aligned} \qquad \text{(EQ 8-37)}$$

Looking at Equation (8-37) and matching the terms belonging to the squared magnitude response of the Wiener filter i.e. $|H_W^R(\omega)|^2$ in Equation (8-14), Equation (8-37) can be simplified to the following:

$$\Gamma_{AA}(\omega) = 2\cdot\Gamma_{RR}(\omega)\cdot|H_W^R(\omega)|^2 \qquad \text{(EQ 8-38)}$$

Replacing (8-38) into (8-35), we get:

$$\Gamma_{EE}(\omega) = \Gamma_{LL}(\omega) - \Gamma_{RR}(\omega) \cdot |H_W^R(\omega)|^2 \quad \text{(EQ 8-39)}$$

Equation (8-39) is identical to (8-19), and thus $\Gamma^R_{EE\_1}(\omega)$ in (8-19) represents the auto-PSD of $e(i)$. Consequently, $\Gamma^R_{EE}(\omega)$ and $\Gamma^R_{EE\_1}(\omega)$ are then analytically equivalent. Similarly, $\Gamma^L_{EE\_1}(\omega)$ in (8-21) is then also equivalent to $\Gamma^L_{EE}(\omega)$ found by directly taking the auto power spectral density of the prediction error defined as:

$$e(i) = r(i) - l(i) \otimes h^l_w(i) \quad \text{(EQ 8-40)}$$

8.3.3 Finalizing the Target Speech PSD Estimator

This subsection will propose an effective combination of $\Gamma^R_{EE}(\omega)$ and $\Gamma^R_{EE\_1}(\omega)$ to estimate $\Gamma^R_{SS}(\omega)$ (and thus also an effective combination of $\Gamma^L_{EE}(\omega)$ and $\Gamma^L_{EE\_1}(\omega)$ to estimate $\Gamma^L_{SS}(\omega)$), and it will finalize the target speech PSD estimator. Throughout the remaining of the section, the effective combination of $\Gamma^j_{EE}(\omega)$ and $\Gamma^j_{EE\_1}(\omega)$ will be referred to as $\Gamma^i_{EE\_FF}(\omega)$ with j corresponding to either the left channel (i.e. j=L) or the right channel (i.e. j=R).

First, the magnitude of interaural offset in a dB scale between the left and right received noisy PSDs is computed as follows:

$$Offset\_dB(\omega) = \left|10 \cdot \log\left(\Gamma_{LL}(\omega)\right) - 10 \cdot \log\left(\Gamma_{RR}(\omega)\right)\right| \quad \text{(EQ 8-41).}$$

Secondly, the interval of frequencies (i.e. ω_int) where *Offset_dB* is greater than a selected threshold *th_offset* are found as follows:

$$\omega\_\text{int} \quad \text{subject to}: Offset\_dB(\omega\_\text{int}) > th\_offset \quad \text{(EQ 8-42)}$$

Considering for instance the target speech estimation on the right channel: if the offset is greater than *th_offset*, it implies that there is a strong presence of directional noise

interference at those particular frequencies (i.e. ω_int), under the assumption that the target speech is approximately frontal. Consequently, in the context of speech de-noising or enhancement, it is reasonable that the received input noisy speech PSD should be more attenuated at that frequency. Through empirical results, it was observed that for large offsets the estimate of $\Gamma^R_{EE}(\omega)$ estimated via Equation (8-23) yields a lower magnitude than the magnitude of $\Gamma^R_{EE\_1}(\omega)$ estimated via Equation (8-19). As a result, for large offsets it is then more suitable to use $\Gamma^R_{EE}(\omega)$ instead of $\Gamma^R_{EE\_1}(\omega)$ to compute the target speech PSD $\Gamma^R_{SS}(\omega)$ in (8-18). This will yield a greater attenuation of the original noisy speech PSD at that particular frequency i.e. ω_int , therefore more interference will be removed. Inversely, if the offset is not large enough (below *th_offset*) implying that the interference is not as strong, it was noticed empirically that $\Gamma^R_{EE\_1}(\omega)$ should be used instead. Thus, from the above observations, in our work the effective combination of the two estimates was taken as follows:

$$\Gamma^j_{EE\_FF}(\omega) = \begin{cases} \Gamma^j_{EE\_1}(\omega), & \text{for } \omega \notin \omega\_\text{int} \\ \alpha \cdot \Gamma^j_{EE}(\omega) + (1-\alpha)\cdot \Gamma^j_{EE\_1}(\omega), & \text{for } \omega \in \omega\_\text{int} \end{cases} \qquad \text{(EQ 8-43)}$$

where ω_int is found using (8-42) and j corresponds again to either the left channel (i.e. $j=L$) or the right channel (i.e. $j=R$). The weighting coefficient α in (8-43) and *th_offset* in (8-42) were set to 0.85 and 3dB respectively.

Finally, using (8-43), the proposed binaural target speech PSD estimator is defined as the following:

$$\Gamma^j_{SS}(\omega) = \min\left(\left| \frac{\Gamma_{jj}(\omega)\cdot \Gamma^j_{EE\_FF}(\omega)}{\left(\Gamma_{LL}(\omega)+\Gamma_{RR}(\omega)\right) - \left(\Gamma_{LR}(\omega)+\Gamma^*_{LR}(\omega)\right)} \right|, \Gamma_{jj}(\omega) \right) \qquad \text{(EQ 8-44)}$$

It should be noted that the target speech PSD estimator in (8-44) is upper-limited to the input PSD of its corresponding channel to prevent amplification due to the division operator. Also, if $Offset\_dB(\omega)$ is very small at a particular frequency (e.g. $Offset\_dB(\omega) < 0.5$ dB), implying that the frame contains mostly a target speech signal, the result in (8-44) would be close to

$\frac{\Gamma_{jj}(\omega)\cdot 0}{0}$. Therefore, to avoid numerical errors, (8-44) was set to $\Gamma_{SS}^{j}(\omega)=\Gamma_{jj}(\omega)$ for frequencies where $Offset\_dB(\omega)$ is below 0.5 dB.

In the case of frontal targets, $\Gamma_{SS}^{j}(\omega)$ computed for the left and right sides could then be averaged to get a better estimate, but in the case of non-frontal targets (next section), the estimates $\Gamma_{SS}^{j}(\omega)$ for the left and right sides will be different.

8.3.4 Case of Non-Frontal Target Sources

In the previous subsections, the target source PSD estimator was designed under the assumption that the target source was frontal and that a directional interference source was at any arbitrary (unknown) direction in the background. This is the focus and the scope of this chapter. However, it is possible to slightly modify the solution found in (8-29) for a frontal target source to take into account a non-frontal target source as follows: First, if the direction of the non-frontal target source is known, or more specifically if the ratio between the left and right HRTFs for the target is known (from measurements or from a model based on the direction of arrival), then this ratio can be defined as:

$$\Delta_{LR}(\omega)=\frac{H_R(\omega)}{H_L(\omega)} \qquad \text{(EQ 8-45)}$$

Secondly, to find for instance the right target speech PSD (i.e. $\Gamma_{SS}^{R}(\omega)$), the approach is to compensate or pre-adjust the left noisy signal to the direction of the right noisy signal, by using the HRTFs ratio of the target speech defined in (8-45). In the frequency domain, the left noisy input signal "pre-adjusted" can be then computed as follows:

$$Y_L^{AD}(\omega)=Y_L(\omega)\cdot\Delta_{LR}(\omega) \qquad \text{(EQ 8-46)}$$

where $Y_L(\omega)$ is the short-time Fourier transform of the original left noisy input signal as defined in (8-1) (i.e. $Y_L(\omega) = F.T\big(l(i)\big)$). For simplicity, the corresponding time domain "pre-adjusted" representation of $Y_L^{AD}(\omega)$ is referred to as: $l^{ad}(i)$.

Finally, by performing this pre-adjustment, the solution developed in (8-44) for a frontal target can be applied again (i.e. the solution remains valid) but all the required parameters should then be computed using $l^{ad}(i)$ instead of $l(i)$. The final result of (8-44) will yield the estimation of the right target speech PSD i.e. $\Gamma_{SS}^{R}(\omega)$.

Reciprocally, to find the left target PSD i.e. $\Gamma_{SS}^{L}(\omega)$, the original left noisy input signal i.e. $l(i)$ remains unchanged but the right noisy input signal i.e. $r(i)$ in (8-2) should be at first pre-adjusted by using the inverse of (8-45). Consequently, $\Gamma_{SS}^{L}(\omega)$ is found by using $l(i)$ and the pre-adjusted right noisy input signal referred to as $r^{ad}(i)$ instead of $r(i)$ in (8-44).

It should be noted that by pre-adjusting the left or right input noisy signals to compute the left or right target PSDs, the residual directional noise remaining in the left and right target PSD estimations will also be shifted. Consequently, the interaural cues of the noise would not be preserved. However, regardless of the direction of the target source, it will be shown in Chapter 9 how to fully preserve all the interaural cues for both the target speech and noise. Chapter 9 will also show an example on how to integrate the proposed target speech PSD estimator from this chapter into a selected recent binaural noise reduction scheme, to remove lateral noise with preservation of all interaural cues.

It should noted that the target speech estimator equation using (8-44) (with a binaural system composed of one microphone on each side of the head) yields an exact solution when the target speech is corrupted by one directional non-statonary noise in the background. However, when multiple directional noises occur simultaneously in the background, the target speech estimator equation will no longer be exact. However, it

will be shown in Chapters 9 and 10 that the target estimator proposed in Equation (8-44) can still be a useful component of a noise reduction system in such environments, leading to good noise reduction performance even with multiple lateral noises in the background.

Chapter 9. AN EXAMPLE INTEGRATING THE BINAURAL TARGET SPEECH PSD ESTIMATOR INTO A NOISE REDUCTION SCHEME

In Chapter 4, we provided an overview of previous binaural noise reduction schemes such as the work in [BOG'07] (which complements the work in [KLA'06] and in several related publications such as [KLA'07],[DOC'05b]). As we recall from Chapter 4 in Section 4.1.3, in [BOG'07] a binaural Wiener filtering technique with a modified/extended cost function was developed to reduce directional noise but also to have some control over the distortion level of the binaural cues for both the speech and directional noise components. The results showed that the binaural cues can be maintained after processing but there is a tradeoff between the noise reduction and the preservation of the binaural cues. Another major drawback of the technique in [BOG'07] is that all the statistics for the design of the Wiener filter parameters were estimated off-line in their work and their estimations relied strongly on an ideal VAD. As a result, the directional background noise is restrained to be stationary or slowly fluctuating and the noise source should not relocate during speech activity since its characteristics are only computed during speech pauses. Furthermore, the case where the noise is a lateral interfering speech causes additional problems, because an ideal spatial classification is also needed to distinguish between lateral interfering speech and target speech segments. Regarding the preservation of the interaural cues, the technique in [BOG'07] requires the knowledge of the original interaural transfer functions (ITFs) for both the target speech and the directional noise, under the assumption that they are constant and that they could be directly measured with the microphone signals [BOG'07]. Unfortunately, in practice, the Wiener filter coefficients and the ITFs are not always easily computable, especially when the binaural hearing aids user is in an environment with non-stationary and moving background noise or with the additional presence of stationary diffuse noise in the background. The occurrence of those complex but realistic environments in real-life hearing situations will decrease the performance of the technique in [BOG'07].

In this chapter, the objective is to demonstrate that working with a binaural system, it is possible to significantly reduce non-stationary directional noise and still preserve interaural cues. By incorporating the proposed instantaneous binaural target speech PSD estimator developed in Chapter 8 into a simple binaural noise reduction scheme, it will be shown that non-stationary interfering noise can be efficiently attenuated, while perfectly preserving the cues. In contrast to the work in [BOG'07], the proposed binaural noise reduction scheme using the target PSD estimator will not require the knowledge of the direction of the noise source (i.e. computations of Interaural Transfer Functions (ITFs) are not required). The noise can be highly non-stationary such as an interfering speech signal from a background talker or just transient noise (i.e. dishes clattering or door opening/closing in the background). The proposed binaural noise reduction scheme will not require a VAD or any spatial classification. The noise reduction will be performed on a frame-by-frame basis since the target PSD is retrieved on a frame-by-frame basis (i.e. instantaneous target estimator). Consequently, the background noise source can also be moving (or equivalently, switching from one main interfering noise source to another at a different direction). It will be shown how to adjust the proposed binaural noise reduction scheme to preserve the interaural cues of the target speech and the residual noise after processing. As a result, the spatial impression of the environment will remain unchanged.

As introduced earlier, the work in [BOG'07] is based on a general binaural multichannel Wiener filter to reduce directional noise but with a modified/extended cost function in order to have some control over the distortion level of the binaural cues for both the speech and noise components. In Section 9.1, we will first briefly recall the general binaural Wiener filtering main equations as previously described in Chapter 4 (without the extended cost function for interaural cues preservation, and for a binaural system with only a single microphone per hearing aid on each side). In Section 9.2, we will then demonstrate the integration of our proposed target speech PSD estimator into the binaural Wiener scheme, and in Section 9.3 we will show how to fully preserve the interaural cues. With this integration, our proposed binaural noise reduction scheme will be then referred to as "instantaneous binaural Wiener filter". Finally, Section 9.4 will present

simulation results comparing the work in [BOG'07] under ideal conditions with our proposed instantaneous binaural Wiener filter, in terms of noise reduction performance.

9.1 Binaural Wiener Filtering Noise Reduction Scheme

From the binaural system and signal definitions defined in Section 8.2, the left and right received noisy signal can be represented in the frequency domain as the following:

$$Y_L(\omega) = S_L(\omega) + V_L(\omega) \quad \text{(EQ 9-1)}$$

$$Y_R(\omega) = S_R(\omega) + V_R(\omega) \quad \text{(EQ 9-2)}$$

Each of these signals can be seen as the result of a Fourier transform obtained from a single measured frame of the respective time signals. Combining (9-1) and (9-2) into a vector form referred to as the "binaural noisy input vector" yields:

$$\mathbf{Y}(\omega) = \begin{bmatrix} Y_L(\omega) \\ Y_R(\omega) \end{bmatrix} = \mathbf{S}(\omega) + \mathbf{V}(\omega) \quad \text{(EQ 9-3)}$$

where $\mathbf{S}(\omega) = \begin{bmatrix} S_L(\omega) \\ S_R(\omega) \end{bmatrix}$ is the binaural speech input vector and $\mathbf{V}(\omega) = \begin{bmatrix} V_L(\omega) \\ V_R(\omega) \end{bmatrix}$ is the binaural noise input vector.

The output signals for the left and right hearing aids referred to as $Z_L(\omega)$ and $Z_R(\omega)$ are expressed as:

$$Z_L(\omega) = \mathbf{W}_L^H(\omega) \cdot \mathbf{Y}(\omega) = \mathbf{W}_L^H(\omega) \cdot \mathbf{S}(\omega) + \mathbf{W}_L^H(\omega) \cdot \mathbf{V}(\omega) \quad \text{(EQ 9-4)}$$

$$Z_R(\omega) = \mathbf{W}_R^H(\omega) \cdot \mathbf{Y}(\omega) = \mathbf{W}_R^H(\omega) \cdot \mathbf{S}(\omega) + \mathbf{W}_R^H(\omega) \cdot \mathbf{V}(\omega) \quad \text{(EQ 9-5)}$$

where $\mathbf{W}_L(\omega)$ and $\mathbf{W}_R(\omega)$ are M-dimensional complex weighting vectors for the left and right channels. As mentioned earlier, the binaural system is composed of only a single microphone per hearing aid (i.e. one for each ear). Therefore, the total number of available channels for processing is M=2.

$\mathbf{W}_L(\omega)$ and $\mathbf{W}_R(\omega)$ are also regrouped into a M x 2 complex matrix as the following:

$$\mathbf{W}(\omega) = \left[\mathbf{W}_L(\omega) \quad \mathbf{W}_R(\omega)\right] \qquad \text{(EQ 9-6)}$$

The objective is to find the filter coefficients $\mathbf{W}_L(\omega)$ and $\mathbf{W}_R(\omega)$ used in (9-6), which would produce an estimate of the target speech $S_L(\omega)$ for the left ear and $S_R(\omega)$ for the right ear. As we demonstrated in Section 4.1, the optimum solution in the MMSE sense is the multichannel Wiener solution defined as:

$$\mathbf{W}(\omega) = \left[\mathbf{W}_L(\omega) \quad \mathbf{W}_R(\omega)\right] = \mathbf{\Gamma}_{\mathbf{YY}}^{-1}(\omega) \cdot \mathbf{\Gamma}_{\mathbf{YD}}(\omega) \qquad \text{(EQ 9-7)}$$

with:

$$\begin{aligned} \mathbf{\Gamma}_{\mathbf{YD}}(\omega) &= E\left\{\mathbf{Y}(\omega) \cdot \mathbf{D}^H(\omega)\right\} \\ &= \left[\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega) \quad \mathbf{\Gamma}_{\mathbf{YS}_R}(\omega)\right] \end{aligned} \qquad \text{(EQ 9-8)}$$

In Equation (9-8), $\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega)$ is the $M \times 1$ statistical cross-correlation vector between the binaural noisy inputs and the left target speech signal, and similarly $\mathbf{\Gamma}_{\mathbf{YS}_R}(\omega)$ is the statistical cross-correlation vector between the binaural noisy input and the right target speech signal defined respectively as:

$$\mathbf{\Gamma}_{\mathbf{YS}_L}(\omega) = E\left\{\mathbf{Y}(\omega) \cdot S_L^*(\omega)\right\} \qquad \text{(EQ 9-9)}$$

$$\mathbf{\Gamma}_{\mathbf{YS}_R}(\omega) = E\left\{\mathbf{Y}(\omega) \cdot S_R^*(\omega)\right\} \qquad \text{(EQ 9-10)}$$

9.2 Integration of the Target Speech PSD Estimator

From the binaural Wiener filtering solution described in Section 9.1, it can seen that the optimum solution expressed in (9-7)-(9-10) requires the knowledge of the statistics of the actual left and right target speech signals i.e. $S_L(\omega)$ and $S_R(\omega)$ respectively, required more specifically in Equations (9-9) and (9-10) in order to compute the statistical cross-correlation vectors between the binaural noisy input signals and the binaural target speech signals. Obviously, those cross-correlation vectors are not directly accessible since

$S_L(\omega)$ and $S_R(\omega)$ are not directly available in practice. The work in [BOG'07] also summarized in Chapter 4 (Section 4.1.3) showed possible practical estimations of those Wiener filter parameters but relying strongly on a robust VAD and stationary noise signals. However, using the target speech PSD estimator developed in Chapter 8 and without the need of a VAD, it is possible instead to find an instantaneous estimate of those cross-correlation vectors from the estimated target speech magnitude spectrum, under the assumption that the target speaker is approximately frontal.

First, using the proposed target speech estimator expressed in (8-44), the left and right target speech magnitude spectrum estimates can be computed as:

$$\left|\hat{S}_j(k)\right| = \left|\hat{S}_j(\omega)\right|_{\omega=\frac{2\pi\cdot k}{N}} = \sqrt{\Gamma_{SS}^{j}(k)\cdot N} \qquad \text{(EQ 9-11)}$$

where j corresponds again to either the left channel (i.e. j=L) or the right channel (i.e. j=R) channel, N is the number of frequency bins in the DFT and k is the discrete frequency bin.

Secondly, it is known that the noise found in the phase component of the degraded speech signal is perceptually less important in contrast to the noise affecting the speech magnitude [SHA'06]. Consequently, the unaltered noisy left and right input phases will be used in the computations of cross-correlations vectors in (9-9) and (9-10). However, as mentioned in Chapter 8, one of the key elements of the target speech PSD estimator is that the target speech magnitude can be estimated on a frame-by-frame basis without the need of a voice activity detector. Hence, we can compute the instantaneous estimates (i.e. estimation on a frame-by-frame basis) of the cross-correlation vectors defined in (9-9) and (9-10) as the following:

$$\hat{\mathbf{\Gamma}}_{\mathbf{YS}_i}(k) = \hat{\mathbf{\Gamma}}_{\mathbf{YS}_i}(\omega)\Big|_{\omega=\frac{2\pi\cdot k}{N}} = \mathbf{Y}(k)\cdot\left|\hat{S}_i(k)\right|\cdot e^{j\angle Y_i^*(k)} \qquad \text{(EQ 9-12)}$$

Similarly, the instantaneous correlation matrix of the binaural input signals can be computed as:

$$\hat{\mathbf{\Gamma}}_{\mathbf{YY}}(k) = \hat{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)\Big|_{\omega=\frac{2\pi\cdot k}{N}} = \mathbf{Y}(k)\cdot\mathbf{Y}^H(k) \qquad \text{(EQ 9-13)}$$

As a result, the proposed instantaneous (or adaptive) binaural Wiener filter incorporating the target speech PSD estimator is then found as follows:

$$\hat{\mathbf{W}}_{inst}(k)=\left[\hat{\mathbf{W}}_L^{inst}(k)\quad \hat{\mathbf{W}}_R^{inst}(k)\right]=\left(\hat{\mathbf{\Gamma}}_{\mathbf{YY}\_avg}(k)+\delta\cdot\mathbf{I}_{2\times 2}\right)^{-1}\cdot\hat{\mathbf{\Gamma}}_{\mathbf{YD}}(k) \qquad \text{(EQ 9-14)}$$

where $\hat{\mathbf{\Gamma}}_{\mathbf{YY}\_avg}(k)=0.1\cdot\hat{\mathbf{\Gamma}}_{\mathbf{YY}\_avg}(k)+0.9\cdot\hat{\mathbf{\Gamma}}_{\mathbf{YY}}(k)$ (EQ 9-15)

and $\hat{\mathbf{\Gamma}}_{\mathbf{YD}}(k)=\left[\hat{\mathbf{\Gamma}}_{\mathbf{YS}_L}(k)\quad \hat{\mathbf{\Gamma}}_{\mathbf{YS}_R}(k)\right]$ (EQ 9-16)

It should be noted that $\hat{\mathbf{\Gamma}}_{\mathbf{YY}}(k)$ is not invertible (i.e. matrix of rank =1), therefore, in equation (9-14), the inverse of $\hat{\mathbf{\Gamma}}_{\mathbf{YY}\_avg}(k)$ is computed instead. Also, δ provides additional numerical conditioning (i.e. matrix regularization, typically δ = 1e-16).

It will be shown in the simulation results that the effect of having an instantaneous estimate for the binaural Wiener filter becomes very advantageous when the background noise is transient and/or moving, without relying on a VAD or any signal content classifier.

9.3 Modification to Preserve Interaural Cues

Using the proposed instantaneous binaural Wiener filters computed using (9-14)-(9-16), the enhanced left and right output signals are then found by multiplying the noisy binaural input vector with its corresponding Wiener filter as follows:

$$Z_i^{inst}(k)=\left(\hat{\mathbf{W}}_i^{inst}(k)\right)^H\cdot\mathbf{Y}(k) \qquad \text{(EQ 9-17)}$$

However, similar to the work in [LOT'06], to preserve the original interaural cues for both the target speech and the directional noise after enhancement, it is beneficial to determine a common real-valued enhancement gain per frequency to be applied to both left and right noisy input spectral coefficients. This will guaranty that the interaural time and level differences (ILDs and ITDs) of the enhanced binaural output signals will match the ITDs and ILDs of the original unprocessed binaural input signals.

First, using (9-16), the left and right real-valued spectral enhancement gains are computed as the following:

$$G_L(k) = \min\left(\left|Z_L^{inst}(k)\right| / \left|Y_L(k)\right|, 1\right) = \min\left(\left|\left(\hat{\mathbf{W}}_L^{inst}(k)\right)^H \cdot \mathbf{Y}(k)\right| / \left|Y_L(k)\right|, 1\right) \quad \text{(EQ 9-18)}$$

$$G_R(k) = \min\left(\left|Z_R^{inst}(k)\right| / \left|Y_R(k)\right|, 1\right) = \min\left(\left|\left(\hat{\mathbf{W}}_R^{inst}(k)\right)^H \cdot \mathbf{Y}(k)\right| / \left|Y_R(k)\right|, 1\right) \quad \text{(EQ 9-19)}$$

It should be noted that the spectral gains in (9-18) and (9-19) are upper-limited to one to prevent amplification due to the division operator.

Secondly, (9-18) and (9-19) are then combined into a single common real-valued spectral enhancement gain as follows:

$$G_{ENH}(k) = \sqrt{G_L(k) \cdot G_R(k)} \quad \text{(EQ 9-20)}$$

Finally, using (9-20), the left and right output enhanced signals with interaural cues preservation are then estimated as the following:

$$\hat{S}_L(k) = G_{ENH}(k) \cdot Y_L(k) \quad \text{(EQ 9-21)}$$

$$\hat{S}_R(k) = G_{ENH}(k) \cdot Y_R(k) \quad \text{(EQ 9-22)}$$

The common gain ensures that the ITD and ILD between the two sides remain intact, while the real valued gain (i.e. with zero phase response) will ensure that no frequency dependent phase distortion is introduced.

9.4 Simulation Setup and Hearing Situations

The following is the description of various simulated hearing scenarios. As already specified in Chapter 6, all data used in the simulations such as the binaural speech signals and the binaural noise signals were provided by a hearing aid manufacturer and obtained from "Behind The Ear" (BTE) hearing aids microphone recordings, with hearing aids installed at the left and the right ears of a KEMAR dummy head. The target speech

source and directional interfering noise recordings used in the simulations for this chapter were purposely taken in a reverberant free environment to avoid the addition of diffuse noise on top of the directional noise. In a reverberant environment, the noise and target speech signals received are the sum of several components such as components emerging from the direct sound path, from the early reflections and from the tail of the reverberation [MEE'02]. However, the components emerging from the tail of the reverberation have diffuse characteristics and consequently are no longer considered directional. By integrating in a noise reduction scheme both the proposed binaural target speech PSD estimator from Chapter 8 and the binaural diffuse noise PSD estimator developed in Chapter 6, speech enhancement experiments in complex acoustic scenes composed of time-varying diffuse noise, multiple directional noises and reverberant environments have shown that it becomes possible to effectively diminish those combined diverse noise sources. However, this will be the subject of Chapter 10. The scope of this Chapter 9 is therefore to demonstrate the efficiency of the proposed target source PSD estimator in the presence of an interfering directional noise, using as an example a basic binaural algorithm for such a scenario (i.e. the binaural Wiener filter).

Scenario a): The target speaker is in front of the binaural hearing aid user (i.e. azimuth = 0°) and two background lateral interfering talkers are at azimuth = 90° and 240° in the background.

Scenario b): The target speaker is in front of the binaural hearing aid user with a lateral interfering talker (at 90° azimuth) and transient noises (at 210° azimuth) both occurring in the background.

For simplicity, the proposed binaural noise reduction incorporating the target speech spectrum estimator technique (i.e. Sections 9.2 and 9.3) will be given the acronym: **PBTE_NR** (Proposed Binaural Target Estimator - Noise Reduction). The extended binaural noise reduction scheme in [BOG'07] will be given the acronym: **EBMW** (Extended Binaural Multichannel Wiener). The complete algorithm in [BOG'07] is summarized in Chapter 4 Section 4.1.3.

For the simulations, the results were obtained on a frame-by-frame basis with 25.6ms of frame length and 50% overlap. A Hann window was applied to each binaural input frames with a FFT-size of N=512 at a sampling frequency of fs=20kHz. After processing each frame, the enhanced signals were reconstructed using the Overlap-and-Add method.

The PBTE_NR defined in Equations (9-21),(9-22) was configured as follows: for each binaural frame received, the proposed target speech PSD estimator is evaluated using (8-44). Also, for each binaural frame received, the Wiener solution of (8-6) was estimated directly in the frequency domain as in (9-23) to reduce complexity. It performs a prediction of the left noisy speech signal from the right noisy speech signal as illustrated in Fig. 8-2.

$$H_W^R(\omega) = \Gamma_{LR}(\omega) \cdot e^{-j\omega\Delta} / \Gamma_{RR}(\omega) \qquad \text{(EQ 9-23)}$$

$H_W^R(\omega)$ was then converted back in the time domain, and the first 150 coefficients were kept (discarding the rest) to obtain $h_W^R(i)$. It should be noted that the Wiener filter also included a causality delay of Δ = 60 samples. It can easily be shown that for instance when only directional noise is present without frontal target speech activity, the time domain Wiener solution of (8-6) is then the convolution between the left HRIR and the inverse of the right HRIR. The ideal inverse of the right-side HRIR will typically have some non-causal samples (i.e. non minimal phase HRIR) and therefore the estimate of the Wiener solution should include a causality delay. Furthermore, this causality delay allows the Wiener filter to be on either side of the binaural system to consider the largest possible ITD. The complete target speech PSD estimator algorithm is summarized in Table 9-3.

Once the target speech spectrum is estimated, the result is incorporated in (9-14), to get our so-called instantaneous (i.e. adapted on frame-by-frame basis) binaural Wiener filter, $\hat{\mathbf{W}}_{inst}(\omega)$ as described in Sections 9.2 and 9.3. Moreover, the results obtained with PBTE_NR neither require the use of a VAD (or any classifier) nor a training period.

The EBMW algorithm defined in (4-65) was configured as follows: First, the estimates of the noise and noisy input speech correlation matrices (i.e. $\tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega)$ and $\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)$ respectively) are obtained to compute $\tilde{\mathbf{\Gamma}}_{\mathbf{SS}}(\omega)$ in (4-56). In [BOG'07] the enhancement results were obtained for an environment with stationary directional background noise and all the estimates were calculated off-line using an ideal VAD. However, in this chapter, the scenarios described earlier involve interfering speech and/or transient directional noise in the background, which makes it more complex to obtain those estimates. For instance, each binaural frame received can be classified into one of those four following categories: i) "speech-only" frame (i.e. target speech activity only), ii) "noisy" frame (i.e. target speech activity + noise activity), iii) "noise-only" frame (i.e. noise activity only) and iv) "silent" frame (i.e. without any activities). Consequently, an ideal frame classifier combined with the ideal VAD is also required since $\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)$ has to be estimated using frames belonging to category ii) only and $\tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega)$ has to be estimated using frames belonging to category iii) only. This ideal classifier required for the method from [BOG'07] is also assumed capable of perfectly distinguishing between target speech and interfering speech. Moreover, to obtain all the required estimates, the EBMW also requires a training period. In the simulations, the estimates were obtained offline using three different training periods: a) estimations resulting from 3 seconds of category ii) and 3 seconds of category iii); b) estimations resulting from 6 seconds of category ii) and 6 seconds of category iii); and finally c) estimations resulting from 9 seconds of category ii) and 9 seconds of category iii). The noise reduction results for each training period will be presented in section 9.5. Furthermore, for the EBMW μ was set to 1 and α and β were set to 0, to purposely get the maximum noise reduction possible (please refer to Section 4.1.3 for details about the EBMW algorithm). Thus interaural cues distortion will not be considered by the EBMW algorithm. This setup was chosen so that it becomes possible to demonstrate that even under the ideal conditions for the EBMW from a noise reduction and speech distortion perspective (with a perfect VAD and perfect classifier, and with the algorithm focusing only on noise reduction and speech distortion), the proposed PBTE_NR which does not rely on any VAD or classifier and which guarantees

that the interaural cues are preserved can still outperform the EBMW in most practical cases. It should be mentioned again that unlike the proposed PBTE_NR, the EBMW could only minimize the interaural cues distortion (i.e. not fully preserving the cues) by using non-zero values for α and β, at the cost of achieving less noise reduction.

9.5 Results and Discussions

In this chapter, the objectives measures selected to evaluate the noise reduction performance of PBTE_NR and EBMW were: SNR, WB-PESQ, *Csig*, *Cbak*, *Covl* and CSII. Those objective measures are described in Chapter 5. It should be noted here that the objective measures specific for the evaluation of interaural cues distortion such as in [BOG'07] were not used, since the proposed PBTE_NR algorithm guarantees cues preservation.

The noise reduction performance results for scenarios a) and b) are gathered in Table 9-1 and Table 9-2 respectively. The performance measures for the PBTE_NR and EBMW algorithms were obtained over 8.5 seconds of data (i.e. 8.5 seconds of enhanced binaural signal corresponding to each scenario). However, as mentioned in Section 4.1.3, the reference EBMW algorithm requires a training period to estimate the noise and the noisy input speech correlation matrices (i.e. $\tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega)$ and $\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)$ respectively) before processing. In all the tables, the notation 'x sc + x sc' represents the number of seconds of category ii) and iii) signals that were used off-line (in addition to the 8.5 seconds of data used to evaluate the de-noising performance) to obtain those estimates. As defined in the previous section, category ii) represents the "noisy" frames required for the computation of $\tilde{\mathbf{\Gamma}}_{\mathbf{YY}}(\omega)$ and category iii) represents the "noise-only" frames required for the computation of $\tilde{\mathbf{\Gamma}}_{\mathbf{VV}}(\omega)$. Similar to [BOG'07], all the parameters estimation for the reference EBMW algorithm were performed offline assuming a perfect VAD but also assuming a perfect classifier as well, to distinguish between an interfering speech and the target speech. For the training period of the reference EBMW algorithm, it should be

noted that in order to attain the longest training period represented by "9sc + 9sc", the actual off-line training data required was well over 18 seconds, since the degraded speech data is additionally composed of the two other remaining categories, such as the "speech-only" frames (i.e. category i) and "silent" frames (i.e. category iv) respectively. For instance, the longest training period took close to 40 seconds of data to obtain the appropriate 9 seconds periods of data belonging to categories ii) and iii). The 8.5 seconds of data used for the evaluation of the de-noising performance was also included in the data used for the off-line estimation of the parameters in the EBMW algorithm, which could also be considered as a favourable case. In contrast, the proposed PBTE_NR algorithm did not make use of any prior training period.

Looking at the performance results for scenario a) shown in Table 9-1 for the case where two lateral interfering talkers are in the background at two fixed locations, it can be seen that our proposed PBTE_NR algorithm strongly reduces the overall noise, with left and right SNR gains of about 7.5dB. Most importantly, while the noise is greatly reduced, the overall quality of the binaural signals after processing was also improved, as represented by a gain in the *Covl* measure and the increase of WB-PESQ. The target speech distortion is reduced as represented by the increase of the *Csig* measure on both channels. The overall residual noise in the binaural enhanced signals is less intrusive as denoted by the increase of the *Cbak* measure on both channels again. Finally, since there is a good gain in the CSII measure (on both channels), the binaural enhanced signals from our proposed PBTE_NR scheme have a potential speech intelligibility improvement. Overall it can be seen in Table 9-1 that the PBTE_NR algorithm outperformed the reference EBMW algorithm even under an ideal setup for this algorithm (i.e. long training period i.e. "9sc+9sc", perfect VAD and classifier, and without taking into account any preservation of interaural cues).

In [BOG'07],[KLA'06],[KLA'07],[DOC'05a],[DOC'05b], the EBMW algorithm strongly relied on the assumption that the noise signal is considered short-term stationary, that is, $\tilde{\mathbf{\Gamma}}_{\mathbf{vv}}(\omega)$ is equivalent whether it is calculated during noise-only periods (i.e. category iii) or during target speech + noise periods (i.e. category ii). This implies that

$\tilde{\Gamma}_{VV}(\omega)$ should be equivalent to the averaged noise correlation matrix found in $\tilde{\Gamma}_{YY}(\omega)$, since as shown in (4-50) $\tilde{\Gamma}_{YY}(\omega)$ can be decomposed into the sum of the noise and the binaural target speech correlation matrices. However, when the background noise is an interfering speech signal and due to the non-stationary nature of speech, it was found that this equivalence is only achievable on average over a long training period (i.e. long term average). Moreover, to maintain the same performance once a selected adequate training period is completed, the background noise/interferers should not move or relocate, otherwise the estimated statistics required for the computation of the Wiener filter coefficients will become again suboptimal. In practice, those estimates should be frequently updated in order to follow the environment changes, but this implies a shorter training period. However, as shown in the performance results for scenario a), even under ideal conditions (i.e. perfect VAD and classifier, with the interferers remaining at fixed locations and no emphasis on the preservation of the interaural cues), a non-negligible training period of 12 seconds (i.e. 6sc+6sc) still yields a much lower performance result than the one obtained with the proposed PBTE_NR algorithm. The reason is that the PBTE_NR algorithm provides binaural enhancement gains that are continuously updated using the proposed instantaneous target speech PSD estimator. More specifically, since a new target speech PSD estimate is available on frame-by-frame basis (in this simulation, every 25ms corresponding to the frame length), the coefficients of the binaural Wiener filter are also updated at the same rate. The binaural Wiener filter is then better suited for the reduction of transient non-stationary noise. Furthermore, it should be reminded that another important advantage of the PBTE_NR algorithm is that the interaural cues of both the speech and noise will not be distorted at all since in the PBTE_NR algorithm, the left and right (i.e. binaural) instantaneous Wiener filters are combined into a common real-valued spectral enhancement gain as developed in Section 9.3. This gain is then applied to both the left and right noisy input signals, to produce the left and right enhanced hearing aid signals as shown in (9-20)-(9-21). As a result, this enhancement approach guarantees interaural cues preservation and will therefore maintain original spatial awareness.

In scenario b), the interference is coming from a talker and from some dishes clattering in the background. Since those two noise sources are originating at different directions (90° and 210° azimuths respectively) and the noise coming from the dishes clattering is transient, scenario b) can also be described as a single moving noise source, which quickly alternates between those two different directions. It is clear that this type of scenario will decrease the performance of the reference EBMW algorithm, since the overall background noise is even more fluctuating. However, to make the reference EBMW algorithm work even under this scenario, the background transient noise i.e. the dishes clattering was designed to occur periodically in the background over the entire noisy data. Consequently, this helped acquiring better estimates for $\tilde{\Gamma}_{YY}(\omega)$ and $\tilde{\Gamma}_{VV}(\omega)$ during the offline training period. Otherwise, if the transient noise was occurring at random times, $\tilde{\Gamma}_{YY}(\omega)$ and $\tilde{\Gamma}_{VV}(\omega)$ should be estimated online to be able to adapt to this sudden apparition of noise. However, this is also a tradeoff since a short adaptation period will also lead to a lower accuracy of $\tilde{\Gamma}_{YY}(\omega)$ and $\tilde{\Gamma}_{VV}(\omega)$, especially if the interfering noise is another competing speech.

Comparatively, the proposed PBTE_NR algorithm still produced a good performance for the second scenario, which can be verified by the increase of all the objective measures. This is due again to the fact that the adaptation is on a frame-by-frame basis, which allows to quickly adapt to the sudden change of noise direction even when the noise is just a burst (i.e. transient) such as dishes clattering. Moreover, using the proposed PBTE_NR algorithm, the interaural cues for the two background noises and the target speaker are not affected due to its common real-valued spectral gain.

Informal listening tests showed that using the reference EBMW algorithm without the compensation for interaural cues tends to shift the interaural cues of the noise components towards the cues of the speech component i.e. losing their spatial separation due to interaural cues distortion. The original binaural files and the resulting enhanced speech files for scenarios a) and b) using the PBTE_NR and EBMW algorithms are available for download at the address:

http://www.site.uottawa.ca/~bouchard/papers/files-targetPSD-estimator.zip

Table 9-1: Objective Performance Results for Scenario a).

	SNR		WB-PESQ		*Csig*		*Cbak*		*Covl*		CSII	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
EBMW (6sc+6sc)	4.55	5.36	1.24	1.34	3.60	3.45	2.58	2.49	2.90	2.75	0.77	0.84
PBTE_NR	**9.51**	**9.44**	**1.80**	**1.90**	**4.05**	**4.08**	**2.93**	**2.96**	**3.32**	**3.35**	**0.95**	**0.96**

Table 9-2: Objective Performance Results for Scenario b).

	SNR		WB-PESQ		*Csig*		*Cbak*		*Covl*		CSII	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
EBMW (6sc+6sc)	3.80	5.77	1.27	1.31	3.49	3.49	2.63	2.65	2.77	2.78	0.86	0.87
PBTE_NR	**9.15**	**9.47**	**1.60**	**1.68**	**4.07**	**3.97**	**3.03**	**2.94**	**3.34**	**3.27**	**0.95**	**0.94**

While this chapter has been focusing on the case of one or two directional interferences, the next chapter will present a binaural noise reduction scheme that can operate in complex acoustic environments, making use of both the diffuse noise PSD estimator and the target PSD estimator that have been developed in this thesis.

Chapter 10. PROPOSED COMPLETE BINAURAL NOISE REDUCTION SCHEME OPERATING IN COMPLEX ACOUSTIC NOISY ENVIRONMENTS

Most multi-microphone noise reduction systems are designed to reduce only a specific type of noise, or they have proved to be efficient against only certain types of noise encountered in an environment. As a result, under difficult practical situations their noise reduction performance will substantially decrease. For instance, in [BOG'07] or as we recall from Chapter 9, a binaural Wiener filtering technique with a modified cost function was developed to specifically reduce directional noise, and also to have some control over the distortion level of the binaural interaural cues for both the speech and noise components. However, the noise reduction performance results reported in [BOG'07] were performed in an environment with a single directional noise in the background.

In this chapter, a new advanced binaural noise reduction scheme is proposed, which finalizes this thesis, where the binaural hearing aid user is situated in complex noisy environments. The binaural system is composed of one microphone per hearing aid on each side of the head and under the assumption again of having a binaural link between the hearing aids. However, the proposed scheme could also be extended to hearing aids having multiple microphones on each side. The proposed scheme can overcome a wider range of noises with highly fluctuating statistics encountered in real-life environments such as a combination of time varying diffuse noise (i.e. babble-noise in a crowded cafeteria), multiple non-stationary directional noises (i.e. interfering speeches, dishes clattering etc.) and all under reverberant conditions.

The proposed binaural noise reduction scheme first relies on the integration of the two binaural PSD estimators that were developed in Chapter 6 and in Chapter 8. Chapter 6 introduced an instantaneous binaural diffuse noise PSD estimator designed for binaural hearing aids operating in a diffuse noise field environment such as babble-talk in a

crowded cafeteria, with an arbitrary target source direction. This binaural noise Power Spectral Density (PSD) estimator was proven to provide a greater accuracy (and without noise tracking latency) compared to advanced noise spectrum estimation schemes such as in [MAR'01] and in [DOE'96]. Detailed comparisons under various scenarios were performed in Chapter 7. The second binaural estimator integrated in our proposed binaural noise reduction scheme is the instantaneous target speech PSD estimator presented in Chapter 8. This binaural estimator is able to recover a target speech PSD (with a known direction) from received binaural noisy signals corrupted by non-stationary directional interfering noise such as an interfering speech or transient noise (i.e. dishes clattering) coming from arbitrary directions.

The overall proposed binaural noise reduction scheme is structured into five stages, where two of those stages directly involve the computation of the two binaural PSD estimators previously mentioned. Our proposed scheme does not rely on any voice activity detection, and it does not require the knowledge of the direction of the noise sources. Moreover, our proposed scheme fully preserves the interaural cues of the target speech and any directional background noise. Therefore the original spatial impression of the environment is mostly maintained.

Our proposed binaural noise reduction scheme will be compared to another recent advanced binaural noise reduction scheme proposed in [LOT'06] (described in Chapter 4) and also to an advanced monaural scheme in [HU'08a] (described in Chapter 3), in terms of noise reduction and speech intelligibility improvement, evaluated by various objective measures. In [LOT'06], a binaural noise reduction scheme partially based on a Minimum Variance Distortionless Response (MVDR) beamforming concept was developed, more explicitly referred to as a superdirective beamformer with dual-channel input and output, followed by an adaptive post-filter. This scheme can maintain all the interaural cues. In [HU'08b], a monaural noise reduction scheme based on geometric spectral subtraction approach was designed. It produces no audible musical noise and possesses similar properties to the traditional Minimum Mean Square Error (MMSE) algorithm such as in [EPH'84].

Section 10.1 will provide the binaural system description and the description of the complex acoustical environment where the binaural hearing aid user is found. Section 10.2 will summarize the five stages constituting the proposed binaural noise reduction scheme. Section 10.3 will detail each stage with their respective algorithm. Finally, Section 10.4 will present simulation results comparing the work in [LOT'06] and in [HU'08a] with our proposed binaural noise reduction scheme, in terms of noise reduction performance and speech intelligibility improvement in a complex noisy environment.

10.1 Binaural System Description

By taking the same signal definitions as in Chapter 8, let $l(i)$, $r(i)$ be the noisy signals received at the left and right hearing aid microphones, defined here in the time domain as:

$$l(i) = s(i) \otimes h_l(i) + n_l(i) = s_l(i) + n_l(i) \quad \text{(EQ 10-1)}$$

$$r(i) = s(i) \otimes h_r(i) + n_r(i) = s_r(i) + n_r(i) \quad \text{(EQ 10-2)}$$

$s(i)$ is the target source. It is again assumed that the distance between the target speaker and the two microphones (one placed on each ear) is such that they receive essentially speech through a direct path from the target speaker. This implies that the received target speech left and right signals are highly correlated (i.e. the direct component dominates its reverberation components). Note that although the basic model above assumes the dominance of the direct path from the target source over its reverberant components, the overall system introduced later in this chapter is applicable to reverberant environments, as it will be demonstrated. $s_l(i)$ is the received left target speech signal and $s_r(i)$ is the received right target speech signal with corresponding left and right HRIRs (i.e. $h_l(i)$ and $h_r(i)$). In this chapter, the acoustical environment includes a combination of diverse interfering noises in the background where $n_l(i)$ and $n_r(i)$ are the received left and right overall interfering noises signals, respectively. The left and right noise signals received can be seen as the sum of the left and right noise signals received from several directional noise sources located at different azimuths, implying specific HRIRs for each directional noise source location, with the addition of diffuse background noise. The interfering

noises can aggregate several background directional talkers, with also the additional presence of transient noises such as dishes clattering, hammering sounds in the background, etc..

It is assumed for now that the direction of arrival of the target source speech signal is approximately frontal (i.e. the binaural hearing aid user is facing the target speaker), and we have:

$$h_l(i) \approx h_r(i) = h(i) \qquad \text{(EQ 10-3).}$$

The case of non-frontal target sources will be discussed in Section 10.3.6.

From the above binaural system and signal definitions, the left and right received noisy signals can be represented in the frequency domain as follows:

$$Y_L(\lambda,\omega) = S_L(\lambda,\omega) + N_L(\lambda,\omega) \qquad \text{(EQ 10-4)}$$

$$Y_R(\lambda,\omega) = S_R(\lambda,\omega) + N_R(\lambda,\omega) \qquad \text{(EQ 10-5)}$$

It should be noted that each of these signals can be seen as the result of a Fourier transform (i.e. FFT) obtained from a single measured frame of the respective time signals, with λ as the frame index and ω as the angular frequency.

The left and right auto power spectral densities, $\Gamma_{LL}(\lambda,\omega)$ and $\Gamma_{RR}(\lambda,\omega)$, can be expressed as follows:

$$\Gamma_{LL}(\lambda,\omega) = \Gamma_{SS}(\lambda,\omega)\,|H(\omega)|^2 + \Gamma_{N_LN_L}(\lambda,\omega) = \Gamma_{S_LS_L}(\lambda,\omega) + \Gamma_{N_LN_L}(\lambda,\omega) \qquad \text{(EQ 10-6)}$$

$$\Gamma_{RR}(\lambda,\omega) = \Gamma_{SS}(\lambda,\omega)\,|H(\omega)|^2 + \Gamma_{N_RN_R}(\lambda,\omega) = \Gamma_{S_RS_R}(\lambda,\omega) + \Gamma_{N_RN_R}(\lambda,\omega) \qquad \text{(EQ 10-7)}$$

10.2 Proposed Binaural Noise Reduction Scheme

Figure 10-1 illustrates the entire structure of the proposed binaural noise reduction scheme. The entire scheme is composed of five stages, briefly described as follows.

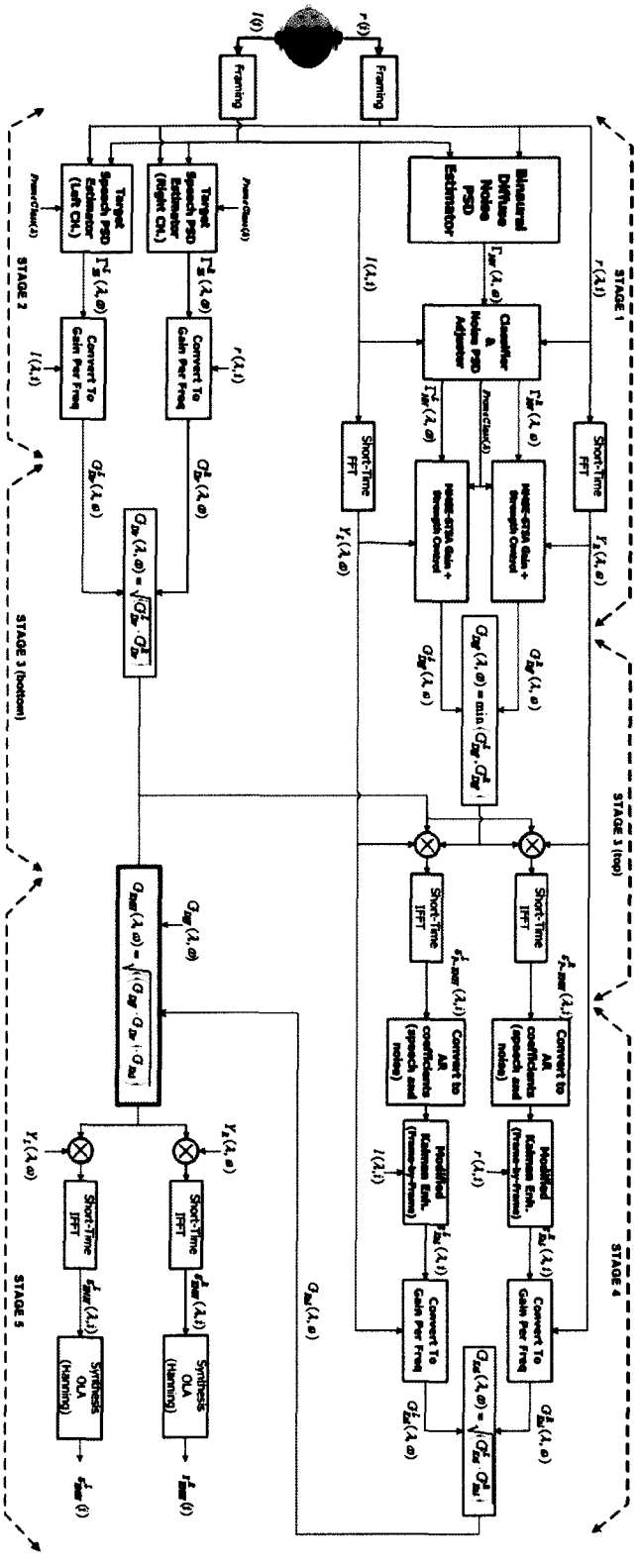

Figure 10-1: Overall structure of the Proposed Binaural Noise Reduction (PBNR) scheme

In the first stage, the Binaural Diffuse Noise PSD Estimator developed in Chapter 6, a classification module and a noise PSD adjuster are used to estimate the left and right noise PSDs for each incoming left and right noisy frames. The noise PSD estimates are then incorporated into a pre-enhancement scheme such as the Minimum Mean Square Short-Time Spectral Amplitude Estimator (MMSE-STSA) developed in [EPH'84] [CAP'94] to produce spectral gains for each respective channel. Those gains are aimed to reduce the presence of diffuse noise and they are referred to as "diffuse noise gains". The reason for using the MMSE-STSA algorithm is because this algorithm produces relatively low musical noise [CAP'96] compared to other spectral gains-based algorithms.

In the second stage, the target speech PSD estimator developed in Chapter 8 is used to extract the target speech PSD (assumed to be frontal for now). Next, the ratio between the target speech PSD estimate and the corresponding noisy input PSD is taken to generate corresponding spectral gains for each respective channel (i.e. left and right) aimed to reduce the directional noises. The resulting spectral gains are referred to as "directional noise gains".

In the third stage, the diffuse noise gains and the directional noise gains are combined (with a weighting rule) and applied to the FFTs of the current left and right noisy input frames. The latter products are then transformed back into the time-domain, resulting into pre-enhanced left and right side frames with interaural cues preservation (i.e. preserving the ILDs and ITDs for both the target speech and directional noises), which will be used in the fourth stage.

In the fourth stage, the binaural noisy input frames are passed through a modified version of Kalman filtering for colored noise from [GAB'05] and also summarized in Chapter 3. The pre-enhanced binaural frames obtained from the third stage are used to calculate the Auto-Regressive (AR) coefficients for the speech and noise models, which are required parameters in the selected Kalman filtering method. Then, similarly to the previous stage, by taking the ratio between the PSDs of the resulting left and right Kalman filtered

frames and the original noisy signal PSDs, a new set of spectral gains referred to as "Kalman-based gains" are obtained.

In the fifth and final stage, the diffuse noise gains, the directional noise gains and the Kalman-based gains are combined with a weighting rule to produce the final set of spectral enhancement gains in the proposed binaural noise reduction scheme. Those gains are then applied to the FFTs of the original noisy left and right frames. The latter products are then transformed back into the time-domain, yielding the final enhanced left and right frames. Most importantly, the same set of spectral gains (which are also real-valued i.e. they do not introduce varying group delays between frequencies) are applied to both the left and right noisy input FFTs, to ensure the preservation of Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) in the enhanced signals, similarly to the approach taken in [LOT'06]. This will avoid spatial distortion (i.e. guarantees preservation of all interaural cues).

It should be noted that the binaural outputs of the third stage could be the final outputs of the binaural noise reduction scheme in some simpler implementations. However, even though the spectral gains obtained by the MMSE-STSA enhancement scheme (in the first stage) provide a relatively low level of musical noise compared to other spectral gain-based algorithms, in our experiments they still produced some robotic-sounding artifacts in the processed speech. Of course the level of distortion can be adjusted by removing less noise, which is the classic tradeoff in speech enhancement. Because of the presence of robotic-sounding speech for the level of noise reduction that we had targeted, a fourth stage was added to the system, where a Kalman filtering enhancement scheme is used (i.e. a model-based approach). As opposed to generating musical noise or robotic sounding speech, in our experiments these types of model-based enhancement algorithms tended to introduce distortion appearing as muffled-sounding speech after processing. It was heuristically observed that combining the gains from the two types of enhancement schemes (in our case spectral-gain based MMSE-STSA and model-based Kalman filtering) provides a more "natural-sounding" speech after processing, with a much reduced level of audible musical noise. This combination is performed in the fifth stage.

10.3 Description of Each Stage of the Proposed Scheme

In this section, the five stages constituting the proposed binaural noise reduction scheme will be explained in details. The left and right signals are decomposed into frames of size D (referred to as binaural noisy input frames) with 50% overlap. The left noisy frames are denoted by $l(\lambda,i)$ and the right noisy frames are denoted by $r(\lambda,i)$. $l(\lambda,i)$ and $r(\lambda,i)$ are the inputs of each stage. The PSD estimates of $l(\lambda,i)$ and $r(\lambda,i)$ were calculated using Welch's method with a Hann data window. However, except for these computations of the PSD estimates, no segmentation or windowing was performed elsewhere on the input data.

10.3.1 Stage 1

First, the Binaural Diffuse Noise PSD Estimator proposed in Chapter 6 is applied using the binaural noisy input frames (i.e. $l(\lambda,i)$ and $r(\lambda,i)$) to estimate the diffuse background noise PSD, $\Gamma_{NN}(\lambda,\omega)$, present in the environment. The Binaural Diffuse Noise PSD Estimator algorithm is summarized in Table 10-1. It should be noted that in Table 10-1, the algorithm proposed in Chapter 6 requires to first estimate $h_W^R(\lambda,i)$. $h_W^R(\lambda,i)$ is a Wiener filter that predicts the current left noisy input frame $l(\lambda,i)$ using the right current input noisy frame $r(\lambda,i)$ as a reference. In this chapter, and as opposed to Chapter 7 where a more costly time domain least-squares implementation had been used, the Wiener filter was estimated directly in the frequency domain to reduce complexity (as shown in Equation 10-8) and a causality delay of Δ = 60 samples was used (since the target speaker can be anywhere around the hearing aid user). $H_W^R(\lambda,\omega)$ was then converted back in the time domain, and the first 150 coefficients were kept (discarding the rest) to obtain $h_W^R(\lambda,i)$.

$$H_W^R(\lambda,\omega)=\Gamma_{LR}(\lambda,\omega)\cdot e^{-j\omega\Delta}/\Gamma_{RR}(\lambda,\omega) \qquad \text{(EQ 10-8)}$$

Secondly, $l(\lambda,i)$, $r(\lambda,i)$ and $\Gamma_{NN}(\lambda,\omega)$ are fed to a block entitled "*Classifier & Noise PSD Adjuster*" as shown in Fig. 10-1. The function of this block is to further alter/update the previous diffuse noise PSD estimate $\Gamma_{NN}(\lambda,\omega)$, and to produce distinct left and right noise PSD estimates $\Gamma_{NN}^{L}(\lambda,\omega)$ and $\Gamma_{NN}^{R}(\lambda,\omega)$ respectively, as illustrated in Fig. 10-1. The Classifier & noise PSD adjuster block is described as follows:

It first computes the interaural coherence magnitude, $0 \le C_{LR}(\omega) \le 1$ between the left and right input noisy signals defined as:

$$C_{LR}(\omega) = \left|\Gamma_{LR}(\omega)\right|^2 / \left(\Gamma_{LL}(\omega)\cdot\Gamma_{RR}(\omega)\right) \qquad \text{(EQ 10-9)}$$

Then, the mean coherence over a selected bandwidth is computed and it is expressed as:

$$\overline{C_{LR}} = \frac{1}{BW}\int_{BW} C_{LR}(\omega)d\omega \qquad \text{(EQ 10-10)}$$

where BW is the selected bandwidth. The bandwidth selected should at least cover a speech signal spectrum (e.g. 300 Hz to 6 kHz) since it is applied for a hearing aid application.

Furthermore, the noise PSD estimation of the current frame is initialized to the estimate returned by the binaural diffuse noise PSD estimator, that is $\Gamma_{NN}^{R}(\lambda,\omega) = \Gamma_{NN}(\lambda,\omega)$ for the right channel and $\Gamma_{NN}^{L}(\lambda,\omega) = \Gamma_{NN}(\lambda,\omega)$ for the left channel. The result obtained using (10-9) will be used to find the frequencies where the coherence magnitude is below a very low coherence threshold referred to as *Th_Coh_vl*. The noise PSD adjuster will increase the initial noise PSD estimate to the level of the noisy input PSD at those frequencies. This implies that only incoherent noise is present at those frequencies. Next, the Classifier will use the result of Equation (10-10) to help classify the binaural noisy input frames received as diffuse noise-only frames or frames also carrying target speech content and/or directional noise. The two possible outcomes for the Classifier are evaluated as follows:

a) A frame is classified as carrying <u>only diffuse noise</u> if there is a low correlation between the left and right received signals over most of the frequency spectrum. In a

speech application, only frequencies relevant to speech content are considered important. Therefore, only a low average correlation over those frequencies will classify the frame as diffuse noise. Analytically, the frame containing only diffuse noise is found by taking the average coherence over typical speech bandwidth using (10-10) and the result should be below a selected low threshold *Th_Coh*. If it is the case, then the value of the variable *FrameClass* is set to 0. In this case, the *Noise PSD Adjuster* takes the initial noise PSD estimate and increases it close to the input noisy PSD of the corresponding frame being processed. More precisely, the adjusted noise PSD estimation is set equal to the geometric mean between the initial noise PSD estimate and the input noisy PSD. The input noisy PSD could also be weighted.

b) A frame is classified as not-diffuse noise if there is a significant correlation between the left and right received signals. This implies that the frame may also contain (on top of some diffuse noise) some target speech content and/or directional background noise such as interfering talker/transient noise. *FrameClass* is then set to 1 if the average coherence over the speech bandwidth using (10-10) is above *Th_Coh*. In this case, the Noise PSD Adjuster will not make any further adjustments in order to be on the conservative side, even though this frame might only contain directional interfering noise. But this will be taken into account in Stage 2.

It is often beneficial to extend a classification period over several frames. For instance, if a frame has been classified as not-diffuse noise, it might then contain target speech content. Therefore, in that case it is safer to force the forthcoming frames to be also classified as not-diffuse noise frames, overruling the actual instantaneous classification result. Table 10-2 summarizes the "*Classifier & Noise PSD Adjuster*" block.

Finally, the last step of stage 1 is to integrate the left and right noise PSDs (i.e. outputs of the "*Classifier & Noise PSD Adjuster*" block) into a Minimum Mean Square Short-Time Spectral Amplitude Estimator (MMSE-STSA). Table 10-3 summarizes the MMSE-STSA algorithm proposed in [EPH'84]. The latter is a SNR-type amplitude estimator speech enhancement scheme (monaural), which is known to produce low musical noise

distortion [CAP'94]. It was introduced in Chapter 3. Applying the MMSE-STSA scheme to each channel with its corresponding noise PSD estimate obtained from the output of the *Noise PSD Adjuster* (i.e. $\Gamma_{NN}^{L}(\lambda,\omega)$ for left channel and $\Gamma_{NN}^{R}(\lambda,\omega)$ for the right channel), real-valued spectral enhancement gains are then obtained. They are denoted by $G_{Diff}^{L}(\lambda,\omega)$ for the left channel and by $G_{Diff}^{R}(\lambda,\omega)$ for the right channel. Those gains are aimed to reduce diffuse noise if it is present (and for reverberant environments they also help reducing the tail of reverberation causing diffuseness). $G_{Diff}^{L}(\lambda,\omega)$ and $G_{Diff}^{R}(\lambda,\omega)$ are referred to as "diffuse noise gains". A strength control is also applied to control the level of noise reduction by not letting the spectral gains drop below a minimum gain, $g_{MIN\_ST1}(\lambda)$. This noise reduction strength control is incorporated as follows:

$$G_{Diff}^{j}(\lambda,\omega) = \max\left(G_{Diff}^{i}(\lambda,\omega), g_{MIN\_ST1}(\lambda)\right), \quad j = L \text{ or } R \qquad \text{(EQ 10-11)}$$

where j corresponds to either the left channel (i.e. $j = L$) or the right channel (i.e. $j = R$).

10.3.2 Stage 2

The goal of Stage 2 is to find spectral enhancement gains which will remove lateral noises. Similar to the first stage, the Instantaneous Target Speech PSD Estimator proposed in Chapter 8 is applied according to the frame classification output $FrameClass(\lambda)$. The Instantaneous Target Speech PSD Estimator algorithm is summarized in Table 10-4. This estimator is designed to extract on a frame-by-frame basis the target speech PSD corrupted by lateral interfering noise with possibly highly non-stationary characteristics. The Instantaneous Target Speech PSD Estimator is applied to each channel (i.e. to the left and right noisy input frames). The target speech PSD estimate obtained from the left noisy input frame is referred to as $\Gamma_{SS}^{L}(\lambda,\omega)$ and the estimate from the right noisy input frame is referred to as $\Gamma_{SS}^{R}(\lambda,\omega)$.

It should be noted that in Table 10-4, the algorithm proposed in Chapter 8 requires to estimate $H_{W}^{L}(\lambda,\omega)$, $h_{W}^{L}(\lambda,i)$, $H_{W}^{R}(\lambda,\omega)$ and $h_{W}^{R}(\lambda,i)$. $H_{W}^{R}(\lambda,\omega)$ and $h_{W}^{R}(\lambda,i)$ were

already computed in Stage 1. Reciprocally, $h_W^L(\lambda,i)$ is a Wiener filter that predicts the current right noisy input frame $r(\lambda,i)$ using the left current input noisy frame $l(\lambda,i)$ as a reference. This Wiener filter was again estimated directly in the frequency domain for lower complexity, as follows:

$$H_W^L(\lambda,\omega)=\Gamma_{RL}(\lambda,\omega)\cdot e^{-j\omega\Delta}/\Gamma_{LL}(\lambda,\omega) \qquad \text{(EQ 10-12)}$$

$H_W^L(\lambda,\omega)$ was then converted back in the time domain and again the first 150 coefficients were kept (discarding the rest) to obtain $h_W^L(\lambda,i)$. It should be noted that the causality delay is required here since the directional noise can emerge from either side of the binaural hearing aid user.

The next step is to convert the target speech PSD estimates computed above into real-valued spectral gains aimed for directional noise reduction, illustrated by the block entitled "*Convert To Gain Per Freq*" depicted in Fig. 10-1. The conversion into spectral gains is performed in order to ease the control of the noise reduction strength by allowing spectral flooring, as done in stage 1 for the diffuse noise gains. In addition, it will permit to easily combine all the gains from the different stages, which will be done in stage 5. In this stage, the corresponding left and right spectral gains referred to as "directional noise gains" are defined as follows:

$$G_{Dir}^L(\lambda,\omega)=\min\left(\sqrt{\Gamma_{SS}^L(\lambda,\omega)/\Gamma_{LL}(\lambda,\omega)},1\right) \qquad \text{(EQ 10-13)}$$

$$G_{Dir}^R(\lambda,\omega)=\min\left(\sqrt{\Gamma_{SS}^R(\lambda,\omega)/\Gamma_{RR}(\lambda,\omega)},1\right) \qquad \text{(EQ 10-14)}$$

It should be noted that the spectral gains in (10-13) and (10-14) are upper-limited to one to prevent amplification due to the division operator.

10.3.3 Stage 3

The objective of the third stage is to provide pre-enhanced binaural output frames with interaural cues preservation to Stage 4 (i.e. preserving the ILDs and ITDs for the both the target speech and directional noises). First, the left and right spectral gains $G_{Diff}^{L}(\lambda,\omega)$ and $G_{Diff}^{R}(\lambda,\omega)$ obtained from the output of Stage 1 are combined into a single real-valued gain per frequency as follows:

$$G_{Diffuse}(\lambda,\omega) = \min\left(G_{Diff}^{L}(\lambda,\omega), G_{Diff}^{R}(\lambda,\omega)\right) \quad \text{(EQ 10-15)}$$

Secondly, the left and right directional gains obtained from the Stage 2 are also combined into a single real-valued gain per frequency as follows:

$$G_{Dir}(\lambda,\omega) = \sqrt{G_{Dir}^{L}(\lambda,\omega)\cdot G_{Dir}^{R}(\lambda,\omega)} \quad \text{(EQ 10-16)}$$

Finally, the gains from Stages 1 and 2 are then combined as follows:

$$G_{Diffuse\_Dir}(\lambda,\omega) = \max\left(G_{Diffuse}(\lambda,\omega)\cdot G_{Dir}(\lambda,\omega), g_{MIN\_ST3}(\lambda)\right) \quad \text{(EQ 10-17)}$$

where a strength control is applied again to control the level of noise reduction, by not allowing the spectral gains to drop below a minimum selected gain referred to as $g_{MIN\_ST3}(\lambda)$.

This real-valued spectral gain above will be applied to both the left and right noisy input frames to produce the corresponding pre-enhanced binaural output frames as follows:

$$s_{P-ENH}^{j}(\lambda,i) = IFFT\left(G_{Diffuse\_Dir}(\lambda,\omega)\cdot Y_{j}(\lambda,\omega)\right),\ j = R \text{ or } L \quad \text{(EQ 10-18)}$$

where $j = L$ corresponds to the left frame and $j = R$ corresponds to the right frame. As previously mentioned, applying a unique real-valued gain to both channels will ensure the preservation of ITDs and ILDs for both the target speech and the remaining directional noises in the enhanced signals (i.e. no spatial cues distortion).

10.3.4 Stage 4

In Stage 4, another category of monaural speech enhancement algorithm known as Kalman filtering is performed. In contrast to the MMSE-STSA algorithm performed in Stage 1, Kalman filtering based methods are model-based, starting from the state-space formulation of a linear dynamical system, and they offer a recursive solution to linear optimal filtering problems [HAY'01]. In this thesis, the Kalman filtering algorithm examined is a modified version of the Kalman Filtering for colored noise proposed in [GAB'05], which was also summarized in Chapter 3. We will present our approach as to how to estimate the various parameters required such as the AR coefficients of speech and noise, the Gaussian noise sequences and the driving process correlation matrix.

As we recall from Chapter 3, in [GAB'05], the Kalman filter uses an Auto-Regressive (AR) model for the target speech signal but also for the noise signal. The speech signal and the colored additive noise (for each channel) are individually modeled as two Auto-Regressive (AR) processes with orders p and q respectively:

$$s_j(i) = \sum_{k=1}^{p} a_k^j \cdot s_j(i-k) + u_j(i) \qquad \text{(EQ 10-19)}$$

$$n_j(i) = \sum_{k=1}^{q} b_k^j \cdot n_j(i-k) + w_j(i) \qquad \text{(EQ 10-20)}$$

where a_k^j is the k^{th} AR speech model coefficient and b_k^j is the k^{th} AR noise model coefficient, and j corresponds to either the left frame (i.e. $j = L$) or the right frame (i.e. $j = R$) . $u_j(i)$ and $w_j(i)$ are uncorrelated Gaussian white noise sequences with zeros means and variances $\left(\sigma_u^j\right)^2$ and $\left(\sigma_w^j\right)^2$ respectively. More specifically, $u_j(i)$ and $w_j(i)$ are referred to as the model driving noise processes (not to be confused with the colored additive acoustic noise i.e. $n_j(i)$ as in Equations (10-1) and (10-2)).

In this work, the Kalman filtering scheme in [GAB'05] was modified to operate on a frame-by-frame basis. All the parameters are frame index dependent (i.e. λ) and the AR models and driving noise processes are updated on a frame-by-frame basis as well (i.e.

$a_k^j(\lambda)$ and $b_k^j(\lambda)$). Since in practice the clean speech and noise signals of each channel are not separately available (i.e. only the sum of those two signals are available for the left and right frames i.e. $l(\lambda,i)$ and $r(\lambda,i)$), the AR coefficients for the left and right target clean speech models in Equation (10-19) are found by applying Linear Predictive Coding (LPC) to the left and right pre-enhanced frames obtained from the outputs of the Stage 3 referred to as s_{P-ENH}^L and s_{P-ENH}^R respectively. The AR coefficients for the noise models in Equation (20) are evaluated by applying LPC on the estimated noise signals extracted from the left and right input noisy frames. The noise signals for each channel are extracted using the pre-enhanced frames as follows:

$$n_{P-ENH}^L(\lambda,i) = l(\lambda,i) - s_{P-ENH}^L(\lambda,i) \quad \text{(EQ 10-21)}$$

$$n_{P-ENH}^R(\lambda,i) = r(\lambda,i) - s_{P-ENH}^R(\lambda,i) \quad \text{(EQ 10-22)}$$

The AR coefficients are then used to find the driving noise processes in (10-19) and (10-20) by computing the LPC residuals (also known as the prediction errors) defined as follows:

$$\hat{u}_j(\lambda,i) = s_{P-ENH}^j(i) - \sum_{k=1}^{p} a_k^j(\lambda) \cdot s_{P-ENH}^j(\lambda,i-k),\ i = 0,1,...,D-1 \quad \text{(EQ 10-23)}$$

$$\hat{w}_j(\lambda,i) = n_{P-ENH}^j(i) - \sum_{k=1}^{q} b_k^j(\lambda) \cdot n_{P-ENH}^j(\lambda,i-k),\ i = 0,1,...,D-1 \quad \text{(EQ 10-24)}$$

After having obtained the required AR coefficients and correlation statistics from the corresponding driving noise sequences for the speech and noise models for each channel, Kalman filtering is then applied to the left and right noisy input frames, producing the left and right enhanced output frames (i.e. Kalman filtered frames) referred to as $s_{Kal}^L(\lambda,i)$ and $s_{Kal}^R(\lambda,i)$ respectively. Table 10-5 summarizes the modified Kalman filtering algorithm for colored noise proposed in [GAB'05], where $\mathbf{A}^j$ represents the augmented state matrix structured as:

$$\mathbf{A}^j(\lambda) = \begin{bmatrix} \mathbf{A}_s^j(\lambda) & \mathbf{0}_{p\times p} \\ \mathbf{0}_{q\times q} & \mathbf{A}_n^j(\lambda) \end{bmatrix} \quad \text{(EQ 10-25),}$$

$\mathbf{A}_s^j$ and $\mathbf{A}_n^j$ correspond to the clean speech and noise transition matrices expressed as:

$$\mathbf{A}_s^j(\lambda)=\begin{bmatrix} 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & 0 \cdots 1 \\ a_p^j & a_{p-1}^j & a_{p-2}^j \cdots a_1^j \end{bmatrix} \quad \text{(EQ 10-26)},$$

$$\mathbf{A}_n^j(\lambda)=\begin{bmatrix} 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & 0 \cdots 1 \\ b_q^j & b_{q-1}^j & b_{q-2}^j \cdots b_1^j \end{bmatrix} \quad \text{(EQ 10-27)},$$

$\mathbf{Q}_j(\lambda)$ corresponds to the driving process correlation matrix computed as:

$$\mathbf{Q}_j(\lambda)=\begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots 0 & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & \vdots \\ 0_{p,1} & \cdots & 0_{p,p-1} & E\left(u_j(i)\cdot u_j(i)\right) & 0_{p,p+1} & \cdots & 0_{p,p+q-1} & E\left(u_j(i)\cdot w_j(i)\right) \\ \vdots & & \vdots & \vdots & & & \vdots & \vdots \\ 0_{p+q,1} & \cdots & 0_{p+q,p-1} & E\left(w_j(i)\cdot u_j(i)\right) & 0_{p+q,p+1} & \cdots & 0_{p+q,p+q-1} & E\left(w_j(i)\cdot w_j(i)\right) \end{bmatrix}$$

(EQ 10-28)

Theoretically, since the target speech signal and the interfering noise signal are statically uncorrelated, the driving noise processes from the speech and noise models in (10-19) and (10-20) should be uncorrelated. This implies that the cross terms in (10-28) (i.e. $E\left(u_j(i)\cdot w_j(i)\right)$ and $E\left(w_j(i)\cdot u_j(i)\right)$) could be assumed to be zero. However, those assumptions do not generally hold true. In a speech application, only short-time estimations are used due to the non-stationary nature of a speech signal. Also, to compute the AR coefficients of the target speech and noise, only estimates of target speech and noise signals are accessible in practice (i.e. herein the estimates were obtained using (10-18) and (10-23)-(10-24)). Therefore, $s_{P-ENH}^j(\lambda,i)$ still contains some residual noise and reciprocally, $n_{P-ENH}^j(\lambda,i)$ still contains some residual target speech signal. Consequently,

those residuals will be also reflected in the computation of the driving noise processes (i.e. obtained from prediction errors using (10-23) and (10-24)), causing non-negligible cross terms due to their correlation. In this work, the cross terms were estimated using (10-23) and (10-24) (assuming short-time stationary and ergotic processes) as follows:

$$E\left(u_j(i)\cdot w_j(i)\right)\approx\frac{1}{D}\sum_{i=0}^{D-1}\hat{u}_j(\lambda,i)\cdot\hat{w}_j(\lambda,i) \qquad \text{(EQ 10-29)}$$

$E\left(u_j(i)\cdot u_j(i)\right)$ and $E\left(w_j(i)\cdot w_j(i)\right)$ are also approximated in a similar way as above.

Still in Table 10-5, $\hat{\mathbf{z}}_j(\lambda,i/i)$ is the filtered estimate of $\mathbf{z}_j(\lambda,i)$, and they are (p+q) by 1 augmented state vectors formulated as:

$$\mathbf{z}_j(\lambda,i)=\left[s_j(\lambda,i-p+1),\cdots,s_j(\lambda,i),n_j(\lambda,i-q+1),\cdots,n_j(\lambda,i)\right]^T \qquad \text{(EQ 10-30)}$$

$$\hat{\mathbf{z}}_j(\lambda,i)=\left[\hat{s}_j(\lambda,i-p+1),\cdots,\hat{s}_j(\lambda,i),\hat{n}_j(\lambda,i-q+1),\cdots,\hat{n}_j(\lambda,i)\right]^T \qquad \text{(EQ 10-31),}$$

$\hat{\mathbf{z}}_j(\lambda,i/i-1)$ is the minimum mean-square estimate of the state vector $\mathbf{z}_j(\lambda,i)$ given the past observations $y(1),\cdots,y(i-1)$. $\mathbf{P}(\lambda,i/i-1)$ is the predicted (*a priori*) state-error covariance matrix, $\mathbf{P}(\lambda,i/i)$ is the filtered state-error covariance matrix, $e(\lambda,i)$ is the innovation sequence and finally, $\mathbf{K}(\lambda,i)$ is the Kalman gain.

The enhanced speech signal at frame index λ and at time index i (i.e. $s_{Kal}^{j}(\lambda,i)=\hat{s}_j(\lambda,i)$) can be obtained from the p^{th} component of the state-vector estimator, i.e. $\hat{\mathbf{z}}(\lambda,i/i)$, which can be considered as the output of the Kalman filter. However, in [PAL'87] it was observed that at time instant i, the first component of $\hat{\mathbf{z}}(i/i)$ (i.e. $\hat{s}(i-p+1)$) yields a better estimate of the speech signal for a previous time index $i-p+1$, since this estimate is based on p-1 additional observations (i.e. $y(i-p+2),\cdots,y(i)$). Consequently, the best estimate of $s_j(i)$ is obtained at time index $i+p-1$. This approach delays the retrieval of $\hat{s}_j(i)$ until the time index $i+p-1$ is reached (i.e. a lag of p – 1 samples). In [PAL'87], this approach is referred to as the delayed Kalman filter, which was also used in our work.

Furthermore, as previously mentioned, we designed our Kalman filter to operate on a frame-by-frame basis with 50% overlap, and by also having the AR coefficients updated on a frame-by-frame basis. Therefore, for each noisy input frame received, the state space vector $\mathbf{z}_j(\lambda,i)$ and the predicted state-error covariance matrix $\mathbf{P}(\lambda,i/i-1)$ were initialized (i.e. at sample index i = 0) with their respective values obtained at sample index $i = D/2-1$ from frame index λ-1.

Similar to Stage 2, the next step is to convert the Kalman filtering results into corresponding real-valued spectral gains. The spectral gains in this stage are referred to as Kalman-based gains and are obtained by taking the ratio between the Kalman filtered frames PSDs and the corresponding input noisy PSDs. The left and right Kalman-based gains are defined as follows:

$$G_{Kal}^{L}(\lambda,\omega) = \min\left(\sqrt{\Gamma_{S_{Kal}S_{Kal}}^{L}(\lambda,\omega)/\Gamma_{LL}(\lambda,\omega)},1\right) \quad \text{(EQ 10-32)}$$

$$G_{Kal}^{R}(\lambda,\omega) = \min\left(\sqrt{\Gamma_{S_{Kal}S_{Kal}}^{R}(\lambda,\omega)/\Gamma_{RR}(\lambda,\omega)},1\right) \quad \text{(EQ 10-33)}$$

where $\Gamma_{S_{Kal}S_{Kal}}^{L}(\lambda,\omega)$ and $\Gamma_{S_{Kal}S_{Kal}}^{R}(\lambda,\omega)$ are the PSDs of the left and right Kalman filtered frames $s_{Kal}^{L}(\lambda,i)$ and $s_{Kal}^{R}(\lambda,i)$ respectively.

10.3.5 Stage 5

In the fifth and final stage, the spectral gains designed in all the stages (i.e. the diffuse noise gains, the directional noise gains and the Kalman-based gains) are weighted and combined to produce the final set of spectral enhancement gains for the proposed binaural enhancement structure. The final enhancement real-valued spectral gains are computed as follows:

$$G_{ENH}(\lambda,\omega) = \max\left(\sqrt{\left(\left(G_{Diff}(\lambda,\omega)\cdot G_{Dir}(\lambda,\omega)\right)\cdot G_{Kal}(\lambda,\omega)\right)}, g_{MIN\_ST5}(\lambda)\right) \quad \text{(EQ 10-34)}$$

where $G_{Kal}(\lambda,\omega)$ is obtained from the left and right Kalman-based gains at the output of Stage 4 combined into a single real-valued gain per frequency as follows:

$$G_{Kal}(\lambda,\omega)=\sqrt{G^{L}_{Kal}(\lambda,\omega)\cdot G^{R}_{Kal}(\lambda,\omega)} \qquad \text{(EQ 10-35)}$$

and $g_{MIN\_ST5}(\lambda)$ is a minimum spectral gain floor.

Finally, the enhancement gains are then applied to the short-time FFTs of the original noisy left and right frames. The latter products are then transformed back into the time-domain (i.e. inverse FFT) yielding the left and right enhanced output frames of the proposed binaural noise reduction scheme as follows:

$$s^{j}_{ENH}(\lambda,i)=IFFT\left(G_{ENH}(\lambda,\omega)\cdot Y_{j}(\lambda,\omega)\right),\ j=R\ or\ L \qquad \text{(EQ 10-36)}$$

In this final stage, having a common real-valued enhancement spectral gain as computed in (10-34) and applied to both channels will ensure both that no frequency dependent phase shift (group delay) is introduced, and that the interaural cues of all directional sources are preserved (i.e. preservation of lateral localization cues). It should be noted that even though all the interaural cues are preserved by appling a common real-valued gain, the preservation of elevation cues is not guaranteed. As briefly explained in Section 2.2.2, source elevation detection relies above all on spectral cues (i.e. monaural cues as opposed to binaural cues). Consequently, the input spectral cues may get distorted due to the spectral coloring caused by applying the enhancement spectral gain to the noisy input spectrum.

10.3.6 Case of Non-Frontal Target Source

So far in this chapter a frontal target source has been assumed in the developments of the proposed method, which as previously mentioned is a realistic and commonly used assumption for hearing aids. In the case of a non-frontal target source, the only step in our proposed scheme that would require a modification is at Stage 2. Stage 2 is designed to remove lateral interfering noises using the target speech PSD estimator proposed in Chapter 8 under the assumption of a frontal target. In Chapter 8 Section 8.3.4, it was explained that it is possible to slightly modify the algorithm in Table 10-4 to take into

account a non-frontal target source. Essentially, the algorithm in Table 10-4 would remain the same except that the left and right input frames (i.e. $l(\lambda,i)$ and $r(\lambda,i)$) would be pre-adjusted before applying the algorithm. The algorithm would then essentially require to know the direction of arrival (DOA) of the non-frontal target source, or more specifically it would need to know the ratio between the left and right HRTFs for the non-frontal target (perhaps from a model and based on the known DOA). The tracking of the target source DOA to be used in stage 2 of our proposed noise reduction scheme for the case of non-frontal targets is one of the possible future work that could be performed from this thesis.

Table 10-1: Diffuse Noise PSD Estimator

Initialization:

- d_{LR} = 0.175 m; c = 344 m/s; $\alpha = 0.99999$;

Δ=60

- $\psi_{LR}(\omega) = \alpha \cdot sinc\left(\frac{\omega \cdot d_{LR} \cdot 2}{c}\right)$

(Note: ω is in radians/sec)

- $\lambda = 0$

START : for each binaural input frames received compute:

1- $H_W^R(\lambda,\omega) = \Gamma_{LR}(\lambda,\omega) \cdot e^{-j\omega\Delta} / \Gamma_{RR}(\lambda,\omega)$

Note: for $H_W^R(\lambda,\omega)$ estimation only, a segmentation of 4 with 50% overlap was used for the computation of $\Gamma_{LR}(\lambda,\omega)$, $\Gamma_{RR}(\lambda,\omega)$.

2- $h_W^R(\lambda,i)$ (refer to section 10.3.1))

3- $e(i) = l(\lambda, i-\Delta) - r(\lambda,i) \otimes h_W^R(\lambda,i)$

4- $\Gamma_{EE}(\lambda,\omega) = PSD\ of\ e(i)$

5-

$$\Gamma_{root}(\lambda,\omega) = \sqrt{\begin{pmatrix}-\left(\Gamma_{LL}(\lambda,\omega)+\Gamma_{RR}(\lambda,\omega)\right)\end{pmatrix}^2 \begin{pmatrix}+2\cdot\psi(\omega)\cdot \mathrm{Re}\{\Gamma_{LR}(\lambda,\omega)\}\end{pmatrix} -4\cdot\left(1-\psi^2(\omega)\right)\cdot\Gamma_{EE}(\omega)\Gamma_{RR}(\lambda,\omega)}$$

6-

$$\Gamma_{NN}(\lambda,\omega) = \frac{1}{2\cdot\left(1-\psi^2(\omega)\right)}\begin{pmatrix}\Gamma_{LL}(\lambda,\omega)+\Gamma_{RR}(\lambda,\omega)-2\cdot\psi(\omega)\\ \cdot \mathrm{Re}\{\Gamma_{LR}(\lambda,\omega)\}-\Gamma_{root}(\lambda,\omega)\end{pmatrix}$$

7- $\lambda = \lambda + 1$

END

Note: for $\Gamma_{EE}(\lambda,\omega)$ and $\Gamma_{LR}(\lambda,\omega)$ computations in steps 5 and 6, a segmentation of 4 with 50% overlap was used.

Table 10-2: Classifier and Noise PSD Adjuster

Initialization:

α =0.5;

Th_Coh_vl =0.1; Th_Coh =0.2;

$ForcedClassFlag = 0$;

$NumberOfForcedFrames$ =5;

$\lambda = 0$

START: for each incoming frame received compute:

1- $C_{LR}(\lambda,\omega)$; $\overline{C_{LR}}(\lambda)$; $\theta_N(\lambda)$;

Note: for the PSD computations in $\overline{C_{LR}}(\lambda)$, a segmentation of 8 with 50% overlap was used.

2- $\Gamma_{NN}^j(\lambda,\omega) = \Gamma_{NN}(\lambda,\omega), \forall\omega$

3- *Find* ω_N *subject to* $C_{LR}(\lambda,\omega_N) < Th\_Coh\_vl$

4- $\Gamma_{NN}^j(\lambda,\omega_N) = \Gamma_{jj}(\lambda,\omega_N)$

5-

if $\overline{C_{LR}}(\lambda) < Th\_Coh$ & $ForcedClassFlag = 0$

$\Rightarrow FrameClass(\lambda) = 0$

$$\Rightarrow \Gamma_{NN}^j(\lambda,\omega) = \sqrt{\max\left(\alpha\cdot\Gamma_{jj}(\lambda,\omega), \Gamma_{NN}^j(\lambda,\omega)\right)\cdot\Gamma_{NN}^j(\lambda,\omega)}$$

else

$\Rightarrow FrameClass(\lambda) = 1$

$\Rightarrow \Gamma_{NN}^j(\lambda,\omega) = \Gamma_{NN}^j(\lambda,\omega), \forall\omega$

$\Rightarrow ForcedClassFlag = 1$

$\Rightarrow ForcedFrameCount = 0$

end

$\Rightarrow ForcedFrameCount = ForcedFrameCount + 1$

if $ForcedFrameCount > NumberOfForcedFrames$

$\Rightarrow ForcedClassFlag = 0$

end

6- $\lambda = \lambda + 1$

END

Note: Steps 1 to 6 is performed with: $j = L$ and $j = R$

Table 10-3: MMSE-STSA

Initialization:

$\beta = 0.8; q = 0.2; \sigma = 0.98; W_{DFT} = 512;$

$\lambda = 0 ; N_j(-1,\omega) = N_j(0,\omega) ; Y_j(-1,\omega) = Y_j(0,\omega);$

START with $j = L$, for each incoming frame received compute:

1- $N_j(\lambda,\omega) = \sqrt{\Gamma^j_{NN}(\lambda,\omega) \cdot W_{DFT}}$

2- $N_j(\lambda,\omega) = \beta \cdot N_j(\lambda,\omega) + (1-\beta) \cdot N_j(\lambda-1,\omega)$

3- $\xi_j(\lambda,\omega) = \dfrac{\left|Y_j(\lambda,\omega)\right|^2}{\left|N_j(\lambda,\omega)\right|^2} - 1$

4-

$$\gamma_j(\lambda,\omega) = (1-\sigma) \cdot \max\left(\xi_j(\lambda,\omega), 0\right) + \sigma \frac{\left|G^j(\lambda-1,\omega) \cdot Y_j(\lambda-1,\omega)\right|^2}{\left|N_j(\lambda,\omega)\right|^2}$$

5- $\hat{\gamma}_j(\lambda,\omega) = (1-q) \cdot \gamma_j(\lambda,\omega)$

6- $\vartheta = \left(1 + \xi_j(\lambda,\omega)\right) \cdot \left(\dfrac{\hat{\gamma}_j(\lambda,\omega)}{1+\hat{\gamma}_j(\lambda,\omega)}\right)$

7- $M[\vartheta] = e^{\left(-\frac{\vartheta}{2}\right)} \cdot \left[(1+\vartheta) \cdot I_0 \cdot \left(\dfrac{\vartheta}{2}\right) + \vartheta \cdot I_1 \cdot \left(\dfrac{\vartheta}{2}\right)\right]$

8-

$$G^j(\lambda,\omega) = \frac{\sqrt{\pi}}{2} \sqrt{\left(\frac{1}{1+\xi_j(\lambda,\omega)}\right) \cdot \left(\frac{\hat{\gamma}_j(\lambda,\omega)}{1+\hat{\gamma}_j(\lambda,\omega)}\right)} \cdot M[\vartheta]$$

9- $\Lambda = \dfrac{1-q}{q} \cdot \dfrac{1}{1+\hat{\gamma}_j(\lambda,\omega)} e^{\left[\frac{\hat{\gamma}_j(\lambda,\omega)}{1+\hat{\gamma}_j(\lambda,\omega)}\right] \cdot \left(1+\xi_j(\lambda,\omega)\right)}$

10- $G^j_{Diff}(\lambda,\omega) = \dfrac{\Lambda}{1+\Lambda} \cdot G^j(\lambda,\omega)$

11- $\lambda = \lambda + 1$

END

Note: $I_0(.)$ and $I_1(.)$ denote the modified Bessel functions of zero and first order respectively.

Repeat steps 1 to 11 with $j = R$

Table 10-4: Target Speech PSD Estimator

Initialization:

α =0.85; th_offset =3; Δ=60;

$\lambda = 0$;

START: with $j = L$, for each incoming frame received compute:

1- *if* $(j == R)$,

(refer to section 10.3.2)

$H^R_W(\lambda,\omega) = \Gamma_{LR}(\lambda,\omega) \cdot e^{-j\omega\Delta} / \Gamma_{RR}(\lambda,\omega)$

$\Gamma^R_{EE\_1}(\lambda,\omega) = \Gamma_{LL}(\lambda,\omega) - \Gamma_{RR}(\lambda,\omega) \cdot |H^R_W(\lambda,\omega)|^2$

$e(i) = l(\lambda, i-\Delta) - r(\lambda,i) \otimes h^R_w(\lambda,i)$

else

$H^L_W(\lambda,\omega) = \Gamma_{RL}(\lambda,\omega) \cdot e^{-j\omega\Delta} / \Gamma_{LL}(\lambda,\omega)$

$\Gamma^L_{EE\_1}(\lambda,\omega) = \Gamma_{RR}(\lambda,\omega) - \Gamma_{LL}(\lambda,\omega) \cdot |H^L_W(\lambda,\omega)|^2$

$e(i) = r(\lambda, i-\Delta) - l(\lambda,i) \otimes h^L_w(\lambda,i)$

end

Note: for $H^R_W(\lambda,\omega)$ and $H^L_W(\lambda,\omega)$ estimation only, a segmentation of 4 with 50% overlap was used for the computation of $\Gamma_{LR}(\lambda,\omega)$, $\Gamma_{RL}(\lambda,\omega)$, $\Gamma_{RR}(\lambda,\omega)$ and $\Gamma_{LL}(\lambda,\omega)$.

2- $\Gamma^j_{EE}(\lambda,\omega) = PSD\ of\ e(i)$

3-

$Offset\_dB(\omega) = \left|10 \cdot \log\left(\Gamma_{LL}(\lambda,\omega)\right) - 10 \cdot \log\left(\Gamma_{RR}(\lambda,\omega)\right)\right|$

4-

Find ω_int subject to:

$Offset\_dB(\omega\_\text{int}) > th\_offset$

5- *if* $(FrameClass(\lambda) == 0)$,

$\Gamma^j_{EE\_FF}(\lambda,\omega) = \sqrt{\Gamma^j_{EE\_1}(\lambda,\omega) \cdot \Gamma_{JJ}(\lambda,\omega)}$

else

$$\Gamma^j_{EE\_FF}(\lambda,\omega) = \begin{cases} \Gamma^j_{EE\_1}(\lambda,\omega), & for\ \omega \notin \omega\_\text{int} \\ \alpha \cdot \Gamma^j_{EE}(\lambda,\omega) + (1-\alpha) \cdot & \\ \Gamma^j_{EE\_1}(\lambda,\omega), & for\ \omega \in \omega\_\text{int} \end{cases}$$

end

6-

$$\Gamma^j_{SS}(\lambda,\omega) = \min\left(\left|\frac{\Gamma_{jj}(\lambda,\omega) \cdot \Gamma^j_{EE\_FF}(\lambda,\omega)}{\left(\Gamma_{LL}(\lambda,\omega) + \Gamma_{RR}(\lambda,\omega)\right) - \left(\Gamma_{LR}(\lambda,\omega) + \Gamma^*_{LR}(\lambda,\omega)\right)}\right|, \Gamma_{jj}(\lambda,\omega)\right)$$

7- $\lambda = \lambda + 1$

END

Repeat steps 1 to 7 with $j=R$

Table 10-5: Kalman Filtering

Initialization:

p=20; q=20;

$\mathbf{C} = [0_1, \cdots, 0_{p-1}, 1, 0_1, \cdots, 0_{q-1}, 1_q]_{1\times(p+q)}$

$\lambda = 0$:

$\hat{\mathbf{z}}_j(\lambda, 0/-1)$=draw vector of $(p+q)$ random numbers ~ $N(0,1)$

$\mathbf{P}_j(\lambda, 0/-1) = \boldsymbol{I}_{(p+q)\times(p+q)}$;

START with $j = L$, for each incoming frame received compute:

1- *if* $(j == L)$, $y(i) = l(\lambda, i)$

else $y(i) = r(\lambda, i)$ *end*

2- Update $\mathbf{A}_s^j$ and $\mathbf{A}_n^j$ into $\mathbf{A}^j(\lambda)$

And $\mathbf{Q}_j(\lambda)$

3-START iteration from $i = 0$ to D - 1,

$e(\lambda, i) = y(\lambda, i) - \mathbf{C} \cdot \hat{\mathbf{z}}_j(\lambda, i/i-1)$

$\mathbf{K}(\lambda, i) = \mathbf{P}_j(\lambda, i/i-1) \cdot \mathbf{C} \times \left[\mathbf{C} \cdot \mathbf{P}_j(\lambda, i/i-1) \cdot \mathbf{C}^{\mathbf{T}}\right]^{-1}$

$\hat{\mathbf{z}}_j(\lambda, i/i) = \hat{\mathbf{z}}_j(\lambda, i/i-1) + \mathbf{K}(\lambda, i) \cdot e(\lambda, i)$

$\mathbf{P}_j(\lambda, i/i) = [I - \mathbf{K}(\lambda, i) \cdot \mathbf{C}] \cdot \mathbf{P}_j(\lambda, i/i-1)$

$\hat{\mathbf{z}}_j(\lambda, i+1/i) = \mathbf{A}_j(\lambda) \cdot \hat{\mathbf{z}}_j(\lambda, i/i)$

$\mathbf{P}_j(\lambda, i+1/i) = \mathbf{A}_j(\lambda) \cdot \mathbf{P}_j(\lambda, i/i) \cdot \mathbf{A}_j^{\mathbf{T}}(\lambda) + \mathbf{Q}_j(\lambda)$

if $(i \geq p-1)$

$s_{Kal}^j(\lambda, i-p+1)$ =1st component of $\hat{\mathbf{z}}_j(\lambda, i/i)$

end

if $(i == D/2-1)$,

$\hat{\mathbf{z}}_j^{temp} = \hat{\mathbf{z}}_j(\lambda, i/i-1)$

$\mathbf{P}_j^{temp} = \mathbf{P}_j(\lambda, i/i-1)$

end

END

4- $\lambda = \lambda + 1$

5- $\hat{\mathbf{z}}_j(\lambda, 0/-1) = \hat{\mathbf{z}}_j^{temp}$

6- $\mathbf{P}_j(\lambda, 0/-1) = \mathbf{P}_j^{temp}$

END

Repeat steps 1 to 6 with $j = R$

10.4 Simulation Setup and Selected Complex Hearing Situations

The following is the description of the simulated complex hearing scenario. As in previous experimental chapters (Chapter 7 and 9), all the data used in the simulations such as the binaural speech signals and the binaural noise signals were provided by a hearing aid manufacturer and obtained from "Behind The Ear" (BTE) hearing aids microphone recordings. All the signals used for this section were recorded in a reverberant environment with an average reverberation time of 1.76 sec. Speech and noise sources were recorded separately. The signals fed to the noise reduction schemes were 8.5 seconds in length.

Scenario: a female target speaker is in front of the binaural hearing aid user (at 0.75m from the hearing aid user), with two male lateral interfering talkers at 270° and 120° azimuths respectively (both at 1.5m from the hearing aid user), with transient noises (i.e. dishes clattering) at 330° azimuth and with time-varying diffuse-like babble noise from crowded cafeteria recordings added in the background. It should be noted that all the speech signals are occurring simultaneously and the dishes are clattering several times in the background during the speech conversation. Moreover, the power level of the original babble-noise coming from a cafeteria recording was purposely abruptly increased by 12 dB at 4.25 secs to simulate even more non-stationary noise conditions, which could be encountered for example if the hearing aid user is entering a noisy cafeteria.

The performance of each considered enhancement or de-noising scheme will be evaluated using this acoustic scenario at three different overall input SNRs varying from about -13.5 dB to 4.6 dB. For simplicity, the Proposed Binaural Noise Reduction scheme will be given the acronym PBNR. The Binaural Superdirective Beamformer with and without Post-filtering noise reduction scheme in [LOT'06] will be given the acronyms BSBp and BSB respectively. The monaural noise reduction scheme proposed in [HU'08a] based on geometric approach spectral subtraction will be given the acronym

GeoSP. It should be noted that the GeoSP scheme was described in Chapter 3 and the BSPp and BSP schemes were described in Chapter 4.

For all the simulations, the signals were processed on a frame-by-frame basis with D = 25.6 ms of frame length and 50% overlap. A FFT-size of N = 512 and a sampling frequency of fs = 20 kHz were used. For the BSBp, BSB and GeoSP schemes, a Hann window was applied to each binaural input frame. After processing each frame, the left and right enhanced signals were reconstructed using the Overlap-and-Add (OLA) method. For the PBNR scheme, the left and right enhancement frames obtained from the output of Stage 5 were windowed using Hann coefficients and then synthesized using the OLA method. The reason for not applying windowing to the binaural input frames for the PBNR scheme is because the implementation of Welch's method that the PBNR scheme uses for PSD computations already involves a windowing operation. The spectral gain floors were set to 0.35 (i.e. $g_{MIN\_ST1}(\lambda)$ = 0.35) for Stage 1 and to 0.1 for Stages 2 to 5. Moreover, the GeoSP scheme requires a noise PSD estimation prior to enhancement, and the monaural noise PSD estimation based on minimum statistics in [HU'08a] was used to update the noise spectrum estimate. The GeoSP algorithm was also slightly modified by applying to the enhancement spectral gain a spectral floor gain set to 0.35, to reduce the noise reduction strength. Both results (i.e. with and without this additional spectral flooring) will be presented. The result with spectral flooring will be referred to as GeoSPo.35.

10.5 Simulation Results and Discussion

In this chapter, the objectives measures selected to evaluate the noise reduction performance of all the schemes were: SNR, Segmental SNR (seg-SNR), ΔPSM, C*sig*, C*bak*, C*ovl* and CSII. Those objective measures are described in Chapter 5.

Table 10-6 shows the noise reduction performance results for the complex hearing scenario described in Section 10.4. Table 10-6 corresponds to the scenario with left and

right input SNR levels of 2.1 dB and 4.6 dB respectively. The performance results are tabulated with processed signals of 8.5 seconds. Figure 10-2 illustrates the corresponding enhanced signals (i.e. processed signals) resulting from the BSPp, GeoSP and PBNR algorithms. Only the results for the left channels are shown, and only for a short segment to facilitate the visual comparison between the schemes. The unprocessed noisy speech segment shown in Fig. 10-2 contains contamination from transient noise (dishes clattering), interfering speeches and background babble noise. The original noise-free speech segment is also depicted in Fig. 10-2 for comparison.

Looking at the objective performance results shown in Table 10-6, it can be seen that our proposed PBNR scheme strongly reduces the overall noise, with left and right SNR gains of about 7.7dB and 5.5dB respectively. Most importantly, while the noise is greatly reduced, the overall quality of the binaural signals after processing was also improved, as represented by a gain in the *Covl* measure and a positive ΔPSM. The target speech distortion is reduced as represented by the increase of the *Csig* measure on both channels. The overall residual noise in the binaural enhanced signals is less intrusive as denoted by the increase of the *Cbak* measure on both channels again. Finally, since there is a gain in the CSII measure (on both channels), the binaural enhanced signals from our proposed PBNR scheme have a potential speech intelligibility improvement. Overall it can be seen in Table 10-6 that the PBNR scheme clearly outperforms the results obtained by the BSPp, BSP, GeoSP and GeoSP0.35 schemes in all the various objective measures. To further analyze the results, it is noticed from Fig. 10-2 that our proposed binaural PBNR scheme visibly attenuated all the combinations of noises around the hearing aid user (transient noise from the dishes clattering, interfering speech and babble noise). The BSPp scheme also reduced those various noises (i.e. directional or diffuse) but the overall noise remaining in the enhanced signal is still significantly higher than for the PBNR. It should also be noted that the enhancement signals obtained by the BSP and BSPp algorithms contain musical noise as can easily be perceived through listening. As for the GeoSP scheme, it can be visualized that it greatly reduced the background babble-noise, but the transient noise and the interfering speech were not attenuated.

The following two paragraphs will provide some further analysis regarding the BSP/BSPp and GeoSP approaches, to explain the results obtained in Fig. 10-2 and the musical noise perceived in the BSP/BSPp enhanced signals. In [LOT'06], the binaural noise scheme BSBp uses a pre-beamforming stage based on the MVDR approach. One of the parameters implemented for the design of the MVDR-type beamformer is a predetermined matrix of cross-power spectral densities (cross-PSD) of the noise under the assumption of a diffuse field. In [LOT'06], this matrix is always maintained fixed (i.e. non-adaptive). Consequently, the BSBp scheme is not optimized to reduce directional interfering noise originating from a specific location. To be more precise, since the noise cross-PSD is designed for a diffuse field, the BSBp scheme will approximately aim to attenuate simultaneously noise originating from all spatial locations except in the desired target direction. The main advantage of this scheme is that it does not require the estimation of the interfering directional noise sources locations. On the other hand, the level of noise attenuation achievable is then reduced since a beamforming notch is not adaptively steered towards the main direction of arrival for the noise. Nevertheless, all the objective measures were improved in our setup with the BSPp and BSP schemes. As briefly mentioned in Section 10.4, the BSP corresponds to the approach without post-processing. The post-processing consists of a Wiener post-filter to further increase the performance, which was the case as shown in Table 10-6 by looking at the results obtained using the BSBp. However, it was noticed that the BSP or BSPp approach causes the appearance of musical noise in the enhanced signals. This is not easily intuitive since in general beamforming approaches should not suffer from musical noise. But as mentioned earlier, the scheme in [LOT'06] uses a beamforming stage which initially produces a single output. By definition, beamforming operates by combining and weighting an array of spatially separated sensor signals (here using the left and right hearing aid microphone signals) and it typically produces a single (i.e. monaural) enhanced output signal. This output is free of musical noise. Unfortunately, in binaural hearing, having a monaural output represents a complete loss of interaural cues of all the sources. In [LOT'06], to circumvent this problem, the output of the beamformer was converted into a common real-valued spectral gain, which was then applied to both binaural input channels. This produces binaural enhanced signals with cues preservation

as mentioned earlier, but it also introduces musical noise in the enhanced signals produced from complex acoustic environments. The conversion to a single gain can no longer be considered as a "true" beamforming operation, since the left or the right enhanced output is obtained by altering/modifying its own respective single channel input, and not by combining input signals from different array sensors. The BSP or BSPp approach thus become closer to other classic speech enhancement methods with Wiener-type enhancement gains, which are often prone to musical noise.

In contrast, the GeoSP scheme in [HU'08a] does not introduce much musical noise. The approach possesses properties similar to the traditional MMSE-STSA algorithm in [EPH'84], in terms of enhancement gains composed of *a priory* and *a posteriori* SNR smoothing, helping in the elimination of musical noise [CAP'94]. However, the GeoSP scheme is based on a monaural system where only a single channel is available for processing. Therefore, the use of spatial information is not feasible, and only spectral and temporal characteristics of the noisy input signal can be examined. Consequently, it is very difficult for instance for the scheme to distinguish between the speech coming from a target speaker or from interferers, unless the characteristics of the lateral noise/interferers are fixed and known in advance, which is not realistic in real life situations. Also, most monaural noise estimation schemes such as the noise PSD estimation using minimum statistics in [MAR'01] assume that the noise characteristics vary at a much slower pace than the target speech signal, and therefore these noise estimation schemes will not detect for instance lateral transient noise such as dishes clattering, hammering sounds, etc. As a result, the monaural noise reduction scheme GeoSP from [HU'08a], which implements the noise estimation scheme in [MAR'01] to update its noise power spectrum, will only be able to attenuate diffuse babble noise as depicted in Fig. 10-2. Also, it was noticed that reducing the noise reduction strength of the original version of the monaural noise reduction scheme proposed in [HU'08a] helped improving its performance (the scheme referred to as GeoSPo.35). The spectral gain floor was set to 0.35, which is the same level that was used in Stage 1 of the PBNR scheme. This modification caused more residual babble noise to be left in the binaural output signals (i.e. decrease of SNR and segSNR gains), however the output signals were less

distorted, which is very important in a hearing aid application. As shown in Table 10-6, all the objective measures (except SNR and SegSNR) were improved using GeoSPo.35, compared to the results obtained with the original scheme GeoSP. It should be mentioned that the results obtained with GeoSPo.35 still produced a slight increase of speech distortion (i.e. a lower C*sig* value) with respect to the original unprocessed noisy signals. Therefore it seems that perhaps the spectral gain floor could be further raised.

The performance of all the noise reduction schemes were also evaluated under lower SNR levels. For the same hearing scenario, Table 10-7 shows the results for input left and right SNR levels of about -3.9 dB and -1.5 dB, representing an overall noise of 6dB higher than the settings used in Table 10-6. Table 10-8 shows the results with a noise level further increased by 9 dB, corresponding to left and right SNRs of -13.5 dB and -11 dB respectively (simulating a very noisy environment).

It can be assessed that the PBNR scheme is efficient even under very low SNR levels as shown in Tables 10-7 and 10-8. All the objective measures were improved on both channels with respect to the unprocessed results and to the other noise reduction schemes. This performance is due to the fact the PBNR approach is divided into different stages addressing various problems and using minimal assumptions. The first two stages are designed to resolve the contamination from various types of noises without the use of a voice activity detector. For instance, Stage 1 designs enhancement gains to reduce diffuse noise only, while the purpose of Stage 2 is to reduce directional noise only. Stage 3 and 4 produce new sets of spectral gains using a Kalman filtering approach from the pre-enhanced binaural signals obtained by combining and applying the gains from stages 1 and 2. In the design phase of the PBNR algorithm, it was found through informal listening tests that combining the gains from the two types of enhancement schemes (MMSE-STSA and Kalman filtering, combined in Stage 5) provides a more "natural-sounding" speech after processing, with negligible musical noise. As previously mentioned, the proposed PBNR also guarantees the preservation of the interaural cues of the directional background noises and of the target speaker, just like the BSPp and BSP schemes. As a result, the spatial impression of the environment will remain mostly

unchanged. Informal listening can easily show the improved performance of the proposed scheme, and the resulting binaural original and enhanced speech files corresponding to the results in Tables 10-6, 10-7 and 10-8 for the different schemes are available for download at the address:

http://www.site.uottawa.ca/~bouchard/papers/TASLP_Complete_Binaural_Enhancement_Syst.zip

It should be noted that the parameters of the proposed scheme (i.e. spectral gain floors, the combination of gains at different stages etc.) were not specifically tuned to this specific noisy environment. Similar performance results were obtained by varying the positions of the various directional noise sources. However, if the noise sources get closer to the target source (i.e. the angle between the target and the noise sources is reduced), the noise reduction performance will decrease since the HRTFs of the noise sources will be similar to the HRTFs of the target source. Furthermore, it should be reminded that it is not possible to reduce directional noise coming directly from behind the binaural hearing aid user, since the binaural input signals are obtained from two microphones in a broadside array configuration.

Table 10-6: Objective Performance Results for Left and Right Input SNRs at 2.1 dB and 4.6 dB Respectively

	SNR		SegSNR		*Csig*		*Cbak*		*Covl*		ΔPSM		CSII	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
BSB	4.07	6.83	0.63	0.46	3.44	3.63	2.27	2.40	2.75	2.94	0.031	0.026	0.73	0.84
GeoSP	3.79	6.64	-0.23	0.85	2.65	2.93	2.02	2.19	2.17	2.44	0.021	0.012	0.59	0.71
PBNR	**9.80**	**10.15**	**2.88**	**3.15**	**3.73**	**3.81**	**2.63**	**2.68**	**3.06**	**3.14**	**0.128**	**0.088**	**0.95**	**0.96**

Table 10-7: Objective Performance Results for Left and Right Input SNRs at -3.9 dB and -1.4 dB Respectively

	SNR		SegSNR		*Csig*		*Cbak*		*Covl*		ΔPSM		CSII	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
BSB	-1.83	1.01	-4.25	-3.41	2.82	3.03	1.69	1.83	2.18	2.38	0.029	0.027	0.34	0.48
GeoSP	-1.56	2.04	-3.20	-2.26	1.94	2.32	1.44	1.65	1.51	1.86	0.021	0.007	0.30	0.36
PBNR	**5.84**	**6.10**	**-0.55**	**-0.21**	**3.14**	**3.23**	**2.08**	**2.14**	**2.51**	**2.59**	**0.118**	**0.086**	**0.63**	**0.75**

Table 10-8: Objective Performance Results for Left and Right Input SNRs at –13.5 dB and -11.0 dB Respectively

	SNR		SegSNR		*Csig*		*Cbak*		*Covl*		ΔPSM		CSII	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
BSB	-11.28	-8.37	-8.17	-7.72	1.98	2.17	1.01	1.11	1.42	1.59	0.022	0.021	0.12	0.14
GeoSP	-10.90	-6.90	-6.76	-6.01	1.64	1.50	1.23	1.01	1.53	1.14	0.016	0.003	0.07	0.13
PBNR	**-1.26**	**-1.07**	**-5.11**	**-4.83**	**2.20**	**2.31**	**1.28**	**1.35**	**1.61**	**1.71**	**0.082**	**0.062**	**0.17**	**0.25**

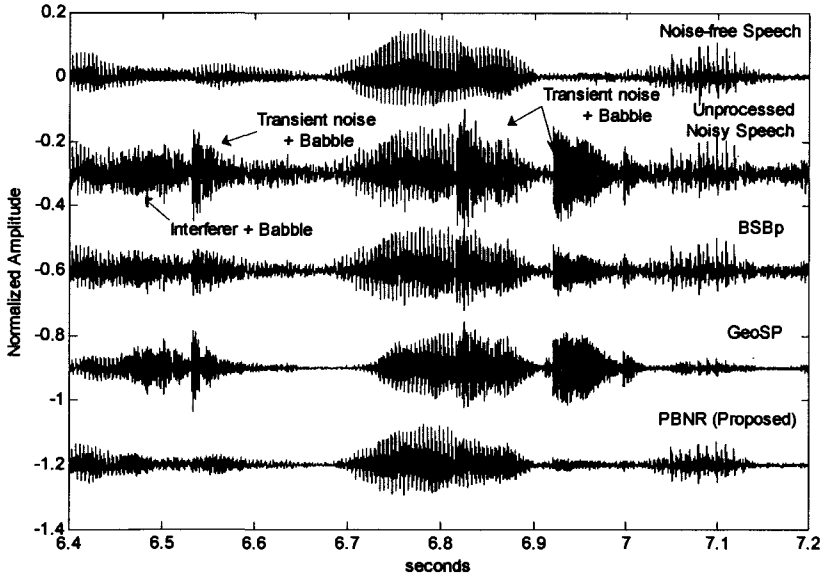

Figure 10-2: Enhancement results using the BSBp, GeoSP and PBNR methods. Only the results for the left channel are shown, for a short segment.

Chapter 11. CONCLUSION AND FUTURE WORK

11.1 Conclusion

The primary objective of this thesis was to develop a complete binaural noise reduction scheme to be implemented in near future high-end binaural hearing aids, making use of the signals received from the left and right hearing aid microphones. Due to the size constraints of some hearing aid models, only a single microphone per hearing aid was considered to develop the binaural noise reduction scheme in this thesis.

In Chapter 6, an improved binaural diffuse noise power spectrum density estimator was developed, where the binaural hearing aid user is located in a diffuse noise field environment. The target speaker can be at any location around the binaural hearing aid user, implying that the direction of arrival of the source speech signal can be arbitrary. This estimator does not employ any voice activity detection algorithm, and the noise can be estimated during speech activity or not. It can track highly non-stationary noise conditions and any type of colored noise, provided that the noise has diffuse field characteristics. In Chapter 7, this proposed estimator was compared with two other advanced noise power spectrum estimation techniques: [MAR'01] and [DOE'96], and the proposed estimator demonstrated a better accuracy and a lower tracking latency under various hearing scenarios. The proposed diffuse noise PSD estimator can then be beneficial to any noise reduction scheme that requires an accurate noise PSD estimate to achieve a satisfactory de-noising performance.

In Chapter 8, an instantaneous binaural target speech power spectrum estimator was developed. It allows the instantaneous target speech power spectrum retrieval in a noisy environment composed of a background interfering talker or transient noise (i.e. highly non-stationary directional noises). It was demonstrated that incorporating the proposed estimator in a binaural Wiener filter algorithm, referred to as the instantaneous binaural

Wiener filter, can efficiently reduce non-stationary as well moving directional background noise. It was shown in Chapter 9 that the proposed instantaneous binaural Wiener filter outperformed the binaural Wiener filtering scheme in [BOG'07] in terms of highly non-stationary directional noise reduction. Most importantly, the proposed instantaneous binaural Wiener filter does not use any voice activity detection or frame content classifier, it does not require any training period (it is "instantaneous" on a frame by frame basis), and it fully preserves both the target speech and directional noise interaural cues.

Finally, in Chapter 10 a new binaural noise reduction scheme was proposed, based on the two developed binaural PSD estimators and a combination of speech enhancement techniques. It was designed for complex acoustic environments and was the main goal of this work. From the simulation results and an evaluation using several objective measures, the proposed scheme confirmed to be effective for complex real-life acoustic environments composed of multiple time-varying directional noises sources, time-varying diffuse noise, and reverberant conditions. Also, the proposed scheme produces enhanced binaural output signals for the left and right ears with full preservation of the original interaural cues of the target speech and the directional background noises. Consequently, the spatial impression of the environment remains mostly unchanged after processing. The proposed binaural noise reduction scheme is thus a good candidate for the noise reduction stage of upcoming binaural hearing aids.

11.2 Future Work

As introduced in the description of this thesis, new high-end binaural hearing aids will be soon available in the marketplace. They will allow the use of information or signals received from both left and right hearing aid microphones to generate an output for the left and right ear. The cross-communication between the left and right microphone hearing aid signals will be ideally achieved via a wireless link.

From this work, a new binaural noise reduction scheme was developed for binaural hearing aids. From simulation results, the proposed scheme confirmed to be efficient for complex environments such as when the binaural hearing aids is located in environments composed of diffuse noise and directional noises in reverberant conditions.

Some future work would be the investigation of the proposed binaural noise reduction in real-time scenarios, which include practical effects related to complexity, delay, and non-ideal binaural wireless link bandwidth as well as jitter. More precisely, the work should consider tradeoffs such as where the processing should be performed and how much data should be exchanged between binaural wireless units. Therefore, efficient implementations of digital signal processing algorithms should be explored. Also, the proposed binaural noise reduction scheme operates on noisy speech segments of 25.6 ms at a time, but in practice the segment length should be ideally reduced to 10 ms for hearing aids. Consequently, methods of decreasing the current 25.6 ms frame size to a more realistic 10 ms frame size should be examined under the restriction that an acceptable frequency resolution should still be maintained. In fact, the entire processing delay of the hearing aid should not exceed 10 ms [STO'02]. Furthermore, the option of including a delay (i.e. jitter buffer) to circumvent a binaural wireless link jitter should be investigated. A real-time implementation would certainly allow to further tune the performance of the noise reduction scheme, based on the real-time behavior of the algorithm under real-life scenarios.

Another future work would be to develop an advanced Direction of Arrival (DOA) estimator to locate the position of the target speaker, since the current main limitation of the developed binaural noise reduction scheme is that the target speaker is either assumed to be approximatively in front of the binaural hearing aids user, or the direction of the speaker is assumed to be known (with corresponding HRTF ratios). This tracking feature or DOA estimator should be tailored for the specific case of binaural hearing aids. For instance, in the context of binaural hearing, the tracking for target speaker could be performed mostly between +20 degrees to -20 degrees (where 0 degree represents a frontal speaker). This tracking range is considered as being a typical head movement of a

hearing user during a conversation. Consequently, having this additional feature would provide clear benefits for the binaural noise reduction scheme since any signal carrying speech content within that search range will not be considered as directional noise (i.e. directional interfering background talker). This is mainly advantageous for stage 2 of the proposed noise reduction scheme, where directional noise reduction gains are obtained and can therefore be easily adjusted for non-frontal target sources as explained in Section 10.3.6. Otherwise, due to head movement the noise reduction scheme could falsely attenuate the target speech instead. It should be noted that in the recent work undertaken in [ROH'08], it was reported that the performance of state of the art DOA algorithms when used for binaural hearing aids was found not to provide a very satisfactory performance yet.

Chapter 12. REFERENCES

[ABU'04] H. Abutalebi, H. Sheikhzadeh, L. Brennan, "A Hybrid Subband Adaptive System for Speech Enhancement in Diffuse Noise Fields", *IEEE Signal Processing Letters*, vol. 11, no. 1, pp. 44-47, Jan. 2004.

[ALL'77] J. B. Allen, D. A. Berkley and J. Blauert, "Multi-microphone Signal-Processing Technique to Remove Room Reverberation from Speech Signals", *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912-915, 1977.

[BEG'94] D. Begault, "3-D Sound for Virtual Reality and Multimedia", Academic Press Professional Inc., San Diego, CA, 1994

[BIT'99] J. Bitzer, K. U. Simmer, and K. Kammeyer, "Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceller (GSC) for Speech Enhancement", *Proc. of ICASSP 1999*, vol. 5, pp. 2965-2968, March 1999.

[BLA'96] J. Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization", MIT Press, 1997.

[BLA'99] J. Blauert, J., "Spatial Hearing: The Psychophysics of Human Sound Localization", MIT Press, 1999.

[BOG'06] T. Van den Bogaert, T. Klasen, L. Van Deun, J. Wouters, and M. Moonen, "Horizontal Localization with Bilateral Hearing Aids: Without Is Better Than With", *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 515-526, Jan. 2006.

[BOG'07] T. Bogaert, S. Doclo, M. Moonen, "Binaural Cue Preservation for Hearing Aids Using an Interaural Transfer Function Multichannel Wiener Filter," in *Proc. IEEE ICASSP*, vol. 4, pp. 565 - 568, Apr. 2007.

[BRA'01] M. Brandstein, D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001

[CAM'03] D.R. Campbell, P.W. Shield, "Speech Enhancement Using Subband Adaptive Griffiths-Jim Signal Processing", *Speech Communication*, vol. 39, pp. 97-110, Jan. 2003.

[CAP'94] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 2, pp. 345-349, 1994.

[CHE'99] C.I. Cheng, G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTF's): Representations of HRTF's in Time, Frequency, and Space (Invited Paper)", *Proceedings of the Audio Engineering Society 107th Convention*, New York, 1999.

[COO'55] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr., "Measurement of Correlation Coefficients in Reverberant Sound Fields", *Journal of the Acoustical Society of America*, vol. 27, pp. 1072-1077, Nov. 1955.

[DES'97] J. Desloge, W. Rabinowitz, and P. Zurek, "Microphone-Array Hearing Aids with Binaural Output-Part I: Fixed-processing Systems", *IEEE Transaction on Speech and Audio Processing*, vol. 5, no. 6, pp. 529-542, Nov. 1997.

[DEV'03] S. Devore, B. Shinn-Cunningham, "Perceptual Consequences of Including Reverberation in Spatial Auditory Displays", *Proceedings of the 2003 International Conference on Auditory Display*, Boston, July 2003.

[DOC'05a] S. Doclo, M. Moonen, "Multi-microphone Noise Reduction Using Recursive GSVD-Based Optimal Filtering with ANC Postprocessing Stage", *IEEE Trans. on Audio, Speech and Audio Processing,* vol. 13, no.1 pp. 53-69, Jan. 2005.

[DOC'05b] S. Doclo, T. Klasen, J. Wouters, S. Haykin, M. Moonen, "Extension of the Multi-Channel Wiener Filter with ITD cues for Noise Reduction in Binaural Hearing Aids," in *Proc. IEEE WASPAA*, pp. 70 - 73, Oct. 2005.

[DOE'96] M. Doerbecker, and S. Ernst, "Combination of Two-Channel Spectral Subtraction and Adaptive Wiener Post-filtering for Noise Reduction and Dereverberation", *Proc. of 8th European Signal Processing Conference (EUSIPCO '96)*, Trieste, Italy, pp. 995–998, Sept. 1996.

[ELK'00] G. W. Elko, "Superdirectional Microphone Arrays", *Acoustical Signal Processing for Telecommunication, Kluwer Academic Publisher*, vol. 10, pp. 181-237, March 2000.

[EPH'84] Y. Ephraim, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, and signal Processing*, Vol. ASSP-32, No.6, pp. 1109-1121, Dec 1984.

[GAB'04] M. Gabrea, "Robust Adaptive Kalman Filtering-Based Speech Enhancement Algorithm", *IEEE Transactions of Acoustics, Speech and Signal Processing*, vol. 1, pp. I-301-4, 2004.

[GAB'05] M. Gabrea, "An Adaptive Kalman Filter for the Enhancement of Speech Signals in Colored Noise", *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Palz, NY, pp. 45-48, October, 2005.

[GEL'04] S. A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, Marcel Dekker, Fourth Edition, 2004.

[GOH'98] Z. Goh, K. Tan, B. Tan, "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction", *IEEE Transactions of Acoustics, Speech and Signal Processing* , vol. 6, no. 3, pp. 287-292, 1998.

[GUE'03] A. Guerin, R. Le Bouquin-Jeannes, G. Faucon, "A Two-Sensor Noise Reduction System: Applications for Hands-Free Car Kit", *EURASIP Journal on Applied Signal Processing*, pp. 1125-1134, 2003.

[HAM'02] V. Hamacher, "Comparison of Advanced Monaural and Binaural Noise Reduction Algorithms for Hearing Aids", *Proc. of ICASSP'02*, vol.4, pp. IV-4008-4011, Orlando, Florida, May 2002.

[HAM'05] V. Hamacher, J. Chalupper, J. Eggers, E. Fisher, U. Kornagel, H. Puder, and U. Rass, " Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends", *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 18, pp. 2915-2929, 2005.

[HAN'98] J.H.L. Hansen, B.L. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms", *International Conference on Spoken Language Processing*, vol. 7, pp. 2819–2822, Sydney, Australia, Dec. 1998.

[HAR'05] W. Hartmann, B. Rakerd, K. Aaron, "Binaural Coherence in Rooms", *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 451-462, May/June 2005.

[HAW'04] M. Hawley, R. Litovsky, "The Benefit of Binaural Hearing in a Cocktail Party: Effect of Location and Type of Interferer", *Journal of the Acoustical Society of America*, vol. 114, pp. 833-843, 2004.

[HAY'01] S. Haykin, *Kalman Filtering and Neural Networks*, John Wiley and Sons, Inc., 2001.

[HEN'81] H.V. Henderson, S. R. Searle, "On Deriving the Inverse of a Sum of Matrices", *Society for Industrial and Applied Mathematics*, vol. 23, no. 1, pp. 53-60, Jan 1981.

[HU'06] Y. Hu and P. Loizou, "Subjective Comparison of Speech Enhancement Algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 153–156, 2006.

[HU'08a]Y. Hu and P. C. Loizou, "A Geometric Approach to Spectral Subtraction", *Speech Communication*, vol. 50, No. 6, pp. 453-466, Jan. 2008.

[HU'08b] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE Trans. Audio Speech Language Processing*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[HUB'06] R. Huber and B. Kollmeier, "PEMO-Q - A New Method for Objective Audioquality Assessment using a Model of Auditory Perception." *IEEE Trans. on Audio, Speech and Language Processing*, vol.14, no. 6, pp. 1902-1911, Nov. 2006.

[ITU'01] ITU-T, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", *Series P: Telephone Transmission Quality Recommendation P.862, International Telecommunications Union*, Feb 2001.

[ITU'07] ITU-T. "Wideband Extension to Recommendation p.862 for the Assessment of Wideband Telephone Networks and Speech Codecs", *Recommendation P.862.2, International Telecommunication Union*, Nov. 2007.

[KAT'05] J.M. Kates and K.H. Arehart, "Coherence and the Speech Intelligibility Index", *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224 - 2237, April 2005.

[KLA'06] T. Klasen, S. Doclo, T. V. Bogaert, M. Moonen, and J. Wouters, "Binaural Multi-Channel Wiener Filtering for Hearing Aids: Preserving Interaural Time and Level Differences", *ICASSP 2006*, Toulouse, France, vol. 5, pp. 145.148, May 2006.

[KLA'07] T. J. Klasen, T. Bogaert, M. Moonen, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Processing*, vol. 55, no. 4, pp. 1579 - 1585, Apr. 2007.

[KLE'93] M. Kleiner, B.I. Dalenbäck, and P. Svensson, "Auralization - an Overview", *Journal of the Audio Engineering Society*, vol. 41, pp. 861-875, 1993.

[LOT'06] T. Lotter and P. Vary, "Dual-Channel Speech Enhancement by Superdirective Beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1-14, 2006.

[LUC'93] R. D. Luce, *Sound and Hearing: A Conceptual Introduction*, Laurence Erlbaum Associates Inc., 1993.

[LUO'03] F. Luo, B. Widow, C. Pavlovic, "Special Issue on Speech Processing for Hearing Aids", *Speech Communications*, vol. 39, pp. 1-3, 2003.

[MA'04] N. Ma, M. Bouchard, R. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise", *ICASSP* 2004, vol. 1, pp. 717-720, May 2004.

[MA'06] N. Ma, M. Bouchard, "Speech Enhancement Using a Masking Threshold Contrained Kalman Filter and its Heuristic Implementations", *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 19-32, Jan 2006

[MAR'01] 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, July 2001.

[MAR'03] R. Martin, "Statistical Method for The Enhancement of Noisy Speech", *International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, pp. 1-6, Sept 2003.

[MAR'94] R. Martin, "Spectral Subtraction based on Minimum Statistics", *Proceedings of EUSIPCO-94*, Lausanne, Switzerland, 1182-85, 1994.

[MCC'03] I. McCowan, and H. Bourland, "Microphone Array Post-Filter Based on Diffuse Noise Field", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709-716, Nov. 2003.

[MEE'02] K. Meesawat, D. Hammershoi, "An Investigation of the Transition from Early Reflections to a Reverberation Tail in a BRIR", *Proc. of the 2002 International Conference on Auditory Display*, Kyoto, Japan, July 2002.

[MEY'97] J. Meyer and K. U. Simmer, "Multi-channel Speech Enhancement in a Car Environment Using Wiener Filtering and Spectral Subtraction", *Proc. of ICASSP 1997*, Munich, Germany, vol. 2, pp. 1167-1170, April 1997.

[MID'89] J.C. Middlebrooks, J.C. Makous and D.M. Green, "Directional Sensitivity of Sound-Pressure Levels in the Human Ear Canal", *Journal of the Acoustical Society of America*, vol. 86, pp. 89-108, 1989.

[MID'91] J. C. Middlebroks, D. M. Green, "Sound Localization by Human Listeners", *Annual Review of Psychology*, vol. 42, pp. 135-159, 1991.

[PAL'87] K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 12, pp. 297-300, April 1987.

[POP'98] D. C. Popescu, I. Zeljkovic, "Kalman Filtering of Colored Noise for Speech Enhancement", *IEEE Transactions of Acoustics, Speech and Signal Processing*, vol. 2, pp. 997-1000, 1998.

[PUD'06] H. Puder, "Adaptive Signal Processing for Interference Cancellation in Hearing Aids", *Signal Processing*, vol. 86, no. 6, pp. 1239-1253, June 2006.

[QUA'02] T. E. Quateri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2002.

[ROH'05] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Perceptual Quality Measures for the Evaluation of Noise Reduction Schemes", *in 9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, pp. 169-172, 2005.

[ROH'07] T. Rohenburg, V. Hohmann, B. Kollmeir, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms By Means of Objective Perceptual Quality Measures", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 315-318, NY, Oct. 21, 2007.

[ROH'07] T. Rohenburg, V. Hohmann, B. Kollmeir, "Robustness Analysis of Binaural Hearing Aid Beamformer Algorithms by Means of Objective Perceptual Quality Measures", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 315-318, NY, Oct. 21,2007.

[SCH'02] S. Schmitt, M. Sandrock, J. Cronemeyer, "Single Channel Noise Reduction for Hands Free Operation in Automotive Environments", *Audio Engineering Society 112th Convention*, Munich, Germany, May 2002.

[SHA'00] D. O'Shaughnessy, *Speech Communications: Human and Machines*, IEEE Press, 2000.

[SHI'02] B. Shinn-Cunningham, "Speech Intelligibility, Spatial Unmasking, and Realism in Reverberant Spatial Auditory Displays", *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, July 2002.

[STO'02] M. A. Stone, B. C. J. Moore, "Tolerable Hearing Aid Delays-II: Estimation of Limits Imposed During Speech Production", *Ear and Hearing*, vol. 23, no. 4, pp. 325-338, 2002

[VIR'99] 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, 126-137, 1999.

[WEL'97] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-Array Hearing Aids with Binaural output. ii. a Two-Microphone Adaptive System," *IEEE Transaction on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[WEB'1] http://hyperphysics.phy-astr.gsu.edu/hbase/sound/hearcon.html
[WEB'2] http://hyperphysics.phy-astr.gsu.edu/hbase/sound/ear.html#c3
[WEB'3] http://hyperphysics.phy-astr.gsu.edu/hbase/sound/oss.html#c1
[WEB'4] http://sound.media.mit.edu/KEMAR.html

[WEN'93] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman, "Localization Using Non-Individualized Head-Related Transfer Functions", *Journal of the Acoustical Society of America*, vol. 94, pp. 111-123, 1993

[WIT'03] T. Wittkop, V. Hohmann, "Strategy-selective Noise Reduction for Binaural Digital Hearing Aids", *Speech Communication*, vol. 39, pp. 111-138, 2003.

[WIT'97] T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. Woods, B. Kollmeier, "Speech Processing for Hearing Aids: Noise Reduction Motivated by Models of Binaural Interaction", *Acustica, Acta Acustica*, vol. 83, pp. 684-699, 1997.

[YAN'03] Z. Yan, L. Du, Jian, J. Wei, H. Zeng, "Two-Channel Microphone Array Processing for Speech Enhancement", *Circuits and Systems*, vol. 2, pp. 548-551, May 2003.