uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Zhengwei Luo

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical Engineering)

GRADE / DEGREE

Department of Electrical Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Beamforming for Binaural Hearing Aids

TITRE DE LA THÈSE / TITLE OF THESIS

M. 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

H. Dajani

R. Goubran

Gary W. Slater

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

Beamforming for Binaural Hearing Aids

by

Zhengwei Luo

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

Master of Applied Science

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
May 2009

Your file Votre référence
ISBN: 978-0-494-58199-5
Our file Notre référence
ISBN: 978-0-494-58199-5

Abstract

Binaural hearing aids making use of a wireless link are becoming a trend in hearing-aids design. However, it is still not clear how much gain can be obtained in complex real-life acoustic environments when using binaural hearing aids compared to monaural ones, and whether binaural hearing aids are worth the additional effort and complexity. This thesis aims to provide some answers to this question. In particular, it will compare the performance of different microphone array configurations, study the effects of using different head models for fixed beamforming design, assess the effect of head model mismatch and direction of arrival information mismatch, investigate methods to preserve the binaural cues, evaluate combinations of fixed binaural beamforming followed by other noise reduction algorithms, and assess the performance of the different algorithms using both classical beamforming metrics and objective measures related to speech quality and intelligibility.

Acknowledgements

I would like to extend my sincere thanks to my supervisor Dr. Martin Bouchard. In a graduate course which he offered I obtained the fundamental knowledge of speech processing. He guided me step by step during my whole research process, and has always been very generous with his time and efforts. In addition, he provided me with considerable insights into the topic of this thesis. His way of doing research, his passion for his work and his professional dedication to his field have contributed to my achievements in the last two years, are having a great influence on my current work, and will play an important role in my future development as well. I respect him in every aspect and will always feel grateful to have had the chance of working with him.

Thanks also to Dr. Tyseer Aboulnasr, who gave me two of my six graduate level courses, and who is always willing to offer her comments on other aspects of my overseas life. I truly enjoyed every moment with her. Dr. Eric Dubois also gave me a good lesson on how to write a thesis appropriately when evaluating my course project. I owe thanks to him.

Moreover, Homayoun Kamkar-Parsi and Frédéric Mustière kindly referred me to information on objective speech quality measurements. Homayoun Kamkar-Parsi also tested the combination of our respective algorithms and provided me with results. I thank both of them.

Finally, special thanks to my boyfriend Shikuan Li and my family, no matter what happens to me, they are always there for me.

Table of Content

List of Figures

List of Tables

List of Acronyms

AG	Array Gain
AGJ	Adaptive Griffiths-Jim
ANC	Adaptive Noise Canceller
BTE	Behind-The-Ear
CIC	Completely In the Canal
CSII	Coherence Speech Intelligibility Index
DI	Directivity Index
DOA	Direction Of Arrival
FF	Free Field
GSC	Generalized Sidelobe Canceller
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
HSE	Head Shadow Effects
ILD	Interaural Level Difference
ITC	In The Canal
ITD	Interaural Time Difference
ITE	In-The-Ear
ITF	Interaural Transfer Function
LLR	Log Likelihood Ratio
MHRTF	Measured HRTF
MMSE	Minimum Mean Square-Error
MMSE-STSA	Minimum Mean Square-Error Short-Time Spectral Amplitude
MOS	Mean Opinion Score
MVDR	Minimum Variance Distortionless Response
MWF	Multi-Channel Wiener Filter
NG	Noise gain
OLA	Overlap-And-Add
PBTE-NR	Proposed Binaural Target Estimator – Noise Reduction

PESQ	Perceptual Evaluation of Speech Quality
PHAT-GCC	PHAse Transform Generalized Cross Correlation
PSD	Power Spectral Density
PZS	Pole-Zero Spherical
RDS	Range Dependent Spherical
SDR	Signal to Distortion Ratio
segSNR	Segmental SNR
SII	Speech Intelligibility Index
SNR	Signal to Noise Ratio
VAD	Voice Activity Detection
WPF	Wiener Post-Filter
WSS	Weighted Spectral Slope

Chapter 1 Introduction

1.1 Motivation and Previous Work

The human auditory system is composed of three parts. First, the *outer ear* includes the pinna, which assists in directing the sound into the ear and provides directional cues. It also includes the auditory canal with a resonator effect. The *middle ear* is located behind the eardrum for impedance conversion (i.e. from air to liquid). Finally, the *inner ear* is where hair cells in the cochlea convert the sound into electric nerve stimulation [PUD'06]. One major cause of hearing loss is found in the inner ear where hair cells can be damaged by, for example, a too strong sound exposure, and a higher hearing threshold is one of the consequences of this. Another characteristic of this type of hearing loss is a stronger masking effect with respect to both frequency and time, which results in reduced speech intelligibility [PUD'06]. To help in mitigating the effects of hearing loss, hearing aids have been designed and used for several decades. A hearing aid is an electro-acoustic body-worn apparatus designed to amplify and enhance sounds for hearing-impaired users [GOR'08], and modern digital hearing aids include noise reduction algorithms. Noise reduction algorithms aim to eventually mimic some of the unique noise reduction features of the human auditory system, such as the ability to deal with the "cocktail-party effect", i.e. the ability to focus our listening attention on a single talker in a complex noisy environment, which can include background noises and interfering speeches [DIL'01].

Because of the good noise reduction and the fairly low amount of distortion that multi-microphone processing algorithms can typically provide, multi-microphone hearing aids have been investigated for a while. Beamforming, which is defined as a general signal processing technique used to control the directionality of the reception or transmission of a signal on a transducer array, has already proved to be able to compensate for some of the effects of a target speech signal masking, which causes a reduction of intelligibility [GAN'09],[PUD'06]. More recently, fast developments in electronics such as reduction in

die size and power consumption have provided more room for hearing-aid designers to apply more sophisticated noise reduction algorithms, and many beamforming algorithms existing in the scientific literature have been investigated to be integrated in digital hearing-aids [DIL'03]. Beamforming systems exploit spatial information in addition to the temporal and spectral information of the desired speech and noise signals, in order to achieve one or more selected "look" directions [ELK'07]. At the same time, sounds coming from other directions are attenuated.

The simplest multi-microphones system are directional microphones, which place two or three microphones in an endfire configuration [TEU'07] (i.e. microphones are then co-linear in the direction of a target sound source). Designers have applied directional microphones to hearing aids for several years, and using directional microphones has proven to increase the speech intelligibility in various noisy environments [HAM'02]. The Adaptive Griffiths-Jim (AGJ) beamformer is a classic structure which is adaptive to signal statistics and to interference locations [GRI'82]. It is mostly suitable for the reduction of directional interfering noise. However, it requires a voice activity detection (VAD) for filter adaptation during the noise-only periods, and VAD failure can cause cancellation of desired speech, for example during low SNR periods. Moreover, the AGJ is sensitive to the target direction of arrival (DOA) estimates. In [CAM'03], the AGJ was combined with a sub-band scheme to further improve the speech intelligibility. Using the concepts of constrained optimization, more sophisticated beamforming algorithms such as the Generalized Sidelobe Canceller (GSC) and the Minimum Variance Distortionless Response (MVDR) (to be described in more details in a later section) have also been developed. The GSC consists of a fixed, spatial pre-processor and an adaptive noise canceller (ANC) [SPR'05]. A beamformer such as the MVDR is used as the pre-processor.

Another category of beamformer is the Multi-Channel Wiener Filter (MWF), which is an optimal noise reduction algorithm in the mean square-error (MMSE) sense [BEN'05]. In [SIM'01], it was proven that the MWF can be decomposed into a cascade of a MVDR "front-end" plus a single-channel Wiener filter as a post-filter, to obtain a MMSE

estimate of the source speech. In any case, the adaptive ANC and MWF algorithms still rely on the availability of a robust VAD algorithm and/or robust classification algorithm to distinguish a target source from interfering sources. This can be very challenging in practice for complex non-stationary acoustics environments [KAM'09C].

When applying beamforming into hearing-aid applications, the array length is limited because of aesthetical and practical reasons [SPR'05]. With the development of wireless transmission techniques, a wireless link which allows to share some signal information between hearing aids located on the left and right sides of the head is becoming available, and the resulting *binaural hearing aids* are recently becoming a trend in hearing-aids design. With binaural hearing aids it becomes possible for designers to implement noise reduction systems which make use of the binaural characteristics of the human auditory system, or which simply make use of the extra spatial information that becomes available from the binaural signals.

[WIT'97] proposed a potential technique which used binaural information to suppress reverberation and lateral noise source. However, the results proved to be quite poor in real-life circumstances because of the presence of multiple interferers or diffuse noise. In [WIT'03], a selective coherence-based binaural noise reduction strategy depending on environment classification was proposed. Nevertheless, this approach was found to work only with low levels of diffuse background noise, and it also presents performance degradation in the presence of both diffuse noise and lateral noises. Based on the framework in [ALL'77], [YAN'03] established a binaural noise reduction system intended for reducing lateral noise. However, the target signal cannot be retrieved under reverberant conditions. Moreover, this method only produced one monaural output signal, which indicates that the spatial cues of any lateral residual noise and sources will not be preserved. [KLA'06] introduced a multichannel binaural Wiener filter based noise reduction algorithm which contains a modified cost function that includes terms for ITD and ILD cues. But this approach essentially works for a single lateral noise source under an environment with low reverberation and without the presence of diffuse noise. It also relies on a perfect VAD for the speech and noise statistics.

In [GOR'08], a comparison between binaural and monaural fixed MVDR beamforming tuned for diffuse noise reduction was made, using classic beamforming performance measures (i.e. not using speech quality or intelligibility objectives measures). The environment was assumed to be free-field (e.g. no head shadow effects considered, etc.)., and the preservation of binaural cues was not addressed. In binaural hearing aids, this preservation of the binaural cues on both sides is desirable. By binaural cues, we mainly mean the interaural time differences (ITD) between the right and left ears, the interaural level differences (ILD) between the right and left ears, the speech onset of each ear (a.k.a. the precedence effect), and the direction dependent spectral shaping by the pinnae (a.k.a. the anatomical transfer function) [HAR'99]. Not being able to localize sounds correctly could put a hearing-aid user into a dangerous situation, for example on a noisy road where the hearing-aid user might not be able to locate the direction of incoming cars in time. It has been observed that hearing impaired persons localize sounds better without using the currently available (and monaural) hearing aids [KLA'06]. In fact, almost all of the single-channel hearing-aid users are suffering from this problem. In [KLA'06] and [BOG'07], in order to partially preserve the ITD and ILD of the speech and noise components, an extension of the binaural MWF was proposed with a cost function including interaural transfer function (ITF) terms (which basically combines the information from the ITD and ILD). However, with this approach only a tradeoff between noise reduction and cues preservation is possible, and preserving the cues significantly affects the resulting noise reduction [KLA'06]. Moreover, the MWF requires a VAD or a sophisticated classifier, and it is mostly suitable for stationary noise environments with only a few localized sources [KAM'09C].

Another method for binaural hearing aids aiming to preserve binaural cues is to apply a real-valued common gain on the binaural signals from both sides, which guarantees to preserve the binaural cues. In [LOT'06], a binaural beamforming system with one microphone on each side was presented using this technique. Under the assumption of a known Direction of Arrival (DOA) for a target sound source, a binaural MVDR beamformer designed for diffuse noise was converted to generate a real-valued common gain and combined with a Wiener post-filter. Results from real-life recordings with

directional speech and reverberation showed some potential for the proposed scheme, although the target direction of 60 degrees (allowing a larger beamforming gain) and the absence of diffuse noise in the experiments may be considered as a favorable and non-typical setup.

Along the same lines, a GSC binaural beamformer combining a fixed MVDR beamformer designed for diffuse noise and an Adaptive Griffiths-Jim part, which can be turned on and off, was introduced and tested in [ROH'07]. "Fixed" MVDR design here means that it is tuned for fixed noise statistics (i.e. the correlation matrices of diffuse noise are known and fixed), however based on the DOA of the target source and knowing the corresponding directivity vectors, the MVDR design can actually be modified dynamically. The authors in [ROH'07] compared four different head models used for the MVDR beamformer design and for the delay compensation in the Adaptive Griffith-Jim beamformer. In addition, three different ways to generate the binaural outputs from the beamformer output were investigated: the common gain approach described earlier, independent bilateral processing which does not preserve the binaural cues, and using head models to re-synthesize the target cues only (i.e. not the directional noise or interference cues). Among those approaches, the common gain method was found to perform best. Results in [ROH'07] also showed that for a fixed MVDR beamformer tuned for diffuse noise, the robustness to DOA steering errors was much more satisfactory than for the Adaptive Griffiths-Jim beamformer, and this limits the applicability of the Adaptive Griffiths-Jim component in practice. In another recent paper from the same authors [ROH'08], DOA estimation for the fixed MVDR beamformer tuned to diffuse noise and producing a common gain was done by estimating the signal delay between microphone pairs. It was found that a good DOA accuracy is required in order to achieve a satisfactory beamforming performance, and that practical DOA algorithms can not always achieve such a good accuracy.

The previous work from the literature indicates that under some conditions binaural beamforming with cues preservation has a good potential to improve the quality and intelligibility of speech in binaural hearing aids. However, it is still not clear from the

literature how much gain (in terms of overall speech quality or intelligibility) can be obtained in complex real-life acoustic environments when using binaural hearing aids compared to monaural ones, and whether binaural hearing aids are worth the additional effort and complexity. Using real-life data provided by a hearing aid manufacturer, this thesis aims to provide some answers in terms of the performance gains which can be provided by using binaural beamforming algorithms.

1.2 Thesis Objectives and Organization

The objectives of this thesis are as follows:

- to provide a basic review of fixed MVDR-based beamforming design in the context of binaural hearing aids
- to compare the performance of different microphone array configurations for hearing aids (monaural, binaural, with different number of microphones)
- to investigate the performance of the considered algorithms in complex noisy acoustic environments including time-varying diffuse-like noise, multiple directional interfering sources (speeches and transient sounds), and reverberation.
- to investigate the performance of the considered algorithms using objective measures related to speech quality and speech intelligibility
- to further investigate the performance of using different head models for the fixed MVDR beamformer design
- to further explore the effects of head model mismatch and DOA mismatch for fixed MVDR beamformer design
- to further develop and evaluate methods of converting the binaural beamforming output to a common gain in order to preserve the binaural cues
- to investigate the potential of using fixed binaural MVDR beamformers as a pre-processor followed by other noise reduction or speech enhancement algorithms.

To achieve these objectives, the thesis is organized in the following way:

Chapter 2 provides a review of MVDR beamforming and it explains why the MVDR is chosen in this thesis. It also introduces classical beamformer performance criteria.

Chapter 3 describes the concept of binaural hearing aids and provides a review of the binaural beamforming problem, given the unique physical limitation in hearing-aid applications. In addition, it provides a detailed description of the different microphone configurations and head models used in this thesis.

Chapter 4 explains the rationale for binaural algorithms using a common gain and it introduces different methods to convert the binaural beamforming output to a common gain.

Chapter 5 describes some structures where a fixed binaural MVDR beamformer is used as a pre-processor and followed by different noise reduction or speech enhancement algorithms. A new structure based on using the output of independent monaural microphone arrays is also proposed.

Chapter 6 presents simulation results of fixed MVDR beamformers tuned for diffuse noise, using classic beamforming performance measures. It compares different array configurations, different head models, DOA mismatch and head model mismatch, for frontal and non-frontal target sources.

Chapter 7 first introduces some speech quality and intelligibility related objective measures. Using these objective measures, it then evaluates the performances of fixed MVDR beamformers tuned for diffuse noise and operating in complex acoustic environments. Moreover, simulation results are also presented for the different structures presented in Chapter 5, combining noise reduction algorithms with the fixed MVDR beamformer as a pre-processor, and for the different methods to generate a common gain as presented in Chapter 4.

Chapter 8 concludes the thesis and provides some suggestions for future research work related to this topic.

1.3 List of Contributions

The following list provides the main new contributions from the thesis:

- the development in Chapter 4 of new and more general ways to produce a common gain in order to preserve binaural cues when using binaural beamforming algorithms
- the combination in Chapter 5 of a fixed binaural MVDR beamformer with other algorithms such as the monaural Minimum Mean Square Short-Time Spectral Amplitude Estimator (MMSE-STSA) [EPH'84], in order to reduce the musical noise
- the introduction in Chapter 5 of a structure using the output of independent monaural microphone arrays and followed by a binaural processing algorithm
- the comparison in Chapter 6 of the binaural beamforming performance for different array configurations, with frontal and non-frontal targets, using different head models, with head model mismatch and DOA mismatch, and using classical beamforming measures
- the performance assessment in Chapter 7 of the considered binaural beamforming algorithms under complex acoustic environments which are composed of time varying diffuse noise, multiple directional non-stationary interferences, and reverberation, using speech quality and intelligibility related objective measures
- the evaluation in Chapter 7 of head model mismatch and DOA mismatch using speech quality and intelligibility related objective measures.

Chapter 2 Review of Beamforming and MVDR Design

2.1 Basics of MVDR Design

2.1.1 Single linear microphone array

A general formulation of a linear 1-D uniform array beamforming problem in the frequency domain is shown in Figure 2.1, and it will be referred to as a monaural array in this thesis.

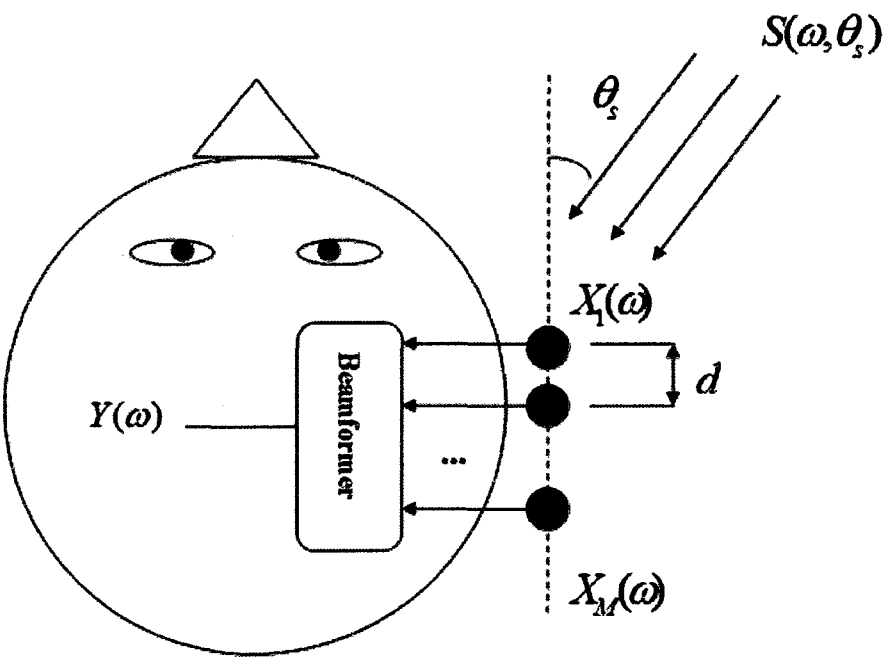


Figure 2.1 Block diagram of a linear 1-D beamformer

$S(\omega)$, $X_i(\omega)$ $i=1...M$, and $Y(\omega)$ are all frequency domain signals which can be considered as the output of a short-time Fourier transform applied to an input frame of time samples. θ_s is the direction of the source signal in the azimuth plane, $S(\omega,\theta_s)$ is the source signal coming from angle θ_s , $X_i(\omega)$ $i=1...M$ are the microphone array received

signals, $Y(\omega)$ is the monaural output of the beamformer, M is the number of microphones belonging to the array, and the M sensors are placed in a linear array of sensors equally spaced by a distance of d. The relationship between the target source signal and the source signal component measured by the microphone array is:

$$\underline{X}_s(\omega) = \underline{D}(\omega,\theta_s)S(\omega,\theta_s) \qquad (2.1)$$

where the directivity vector $\underline{D}(\omega,\theta_s)$ is defined here as the frequency response from a far field source (plane wave propagation, point source) at an angle θ_s, and at a constant distance (or radius) for any angle. The environment considered for $\underline{D}(\omega,\theta_s)$ is anechoic, it can be a free field or it can include head shadow effects (then the $\underline{D}(\omega,\theta_s)$ become a set of Head-Related Transfer Functions (HRTFs)). The HRTFs describe the complex filtering effects of the sound diffraction and reflection caused by the pinnae, head, shoulders and torso before the sound reaches the ear drums [BRO'98]. They also include the interaural effects such as the ITD and ILD. The HRTFs vary with azimuth, elevation, range, and frequency in a complex way. They also vary considerably from person to person. The corresponding representation of the HRTFs in time are called the head-related impulse responses (HRIRs), which capture the same information as the HRTFs. The HRTFs include the head shadow effect, which is due to the fact that a sound from one side has to travel through and around the head in order to reach the ear on the other side. This obstruction includes both amplitude attenuation and complex filtering effects, which are called head shadow effects (HSE). At the opposite, free field (FF) propagation means an environment with no physical object affecting propagation, so there is no head shadow effect existing in free field, and motion in a free field can be written by simple equations.

By linear filtering and summing the measured signals using a set of weights, a basic "filter-and-sum" beamformer can be obtained [GOR'08], and if the weights of the beamformers are noted as $\underline{W}(\omega)$ we then have:

$$Y(\omega) = \underline{W}^H(\omega)\underline{X}(\omega) = \underline{W}^H(\omega)\underline{D}(\omega,\theta_s)S(\omega,\theta_s)$$
$$\underline{W}(\omega) = [W_1(\omega)\ \ W_2(\omega)\ \ \cdots\ \ W_M(\omega)]^T \tag{2.2}$$
$$\underline{X}(\omega) = [X_1(\omega)\ \ X_2(\omega)\ \ \cdots\ \ X_M(\omega)]^T$$

2.1.2 MVDR beamforming

In this thesis, the minimum variance distortionless response (MVDR) beamforming design algorithm is chosen mainly for two reasons: its formulation can be directly applied for arbitrary geometries and environments (arbitrary target direction, free field or with head shadow, 1-Dimentional array or 2-Dimentional array, uniform or non-uniform array, etc); and it allows for fixed design, which is more robust [COX'87], [ROH'07]. The MVDR solution belongs to the family of constrained optimization, and it is obtained by minimizing the Mean Square Error (MSE) of the beamformer output with the constraint that the power gain in the target direction is unity [COX'87]. The problem can be written as follows:

$$\min_{\underline{W}(\omega)} E\{|Y(\omega)|^2\} \text{ subject to } \underline{W}^H(\omega)\underline{D}(\omega,\theta_s) = 1$$

Under far field conditions (plane wave propagation, point sources), this leads to the following solution:

$$\underline{W}(\omega) = \frac{\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right] + \mu I\right)^{-1}\underline{D}(\omega,\theta_s)}{\underline{D}^H(\omega,\theta_s)\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right] + \mu I\right)^{-1}\underline{D}(\omega,\theta_s)} \tag{2.3}$$

$\underline{X}_v(\omega)$ is the noisy portion of the signals measured by the sensors, so if we assume that the noise source is $V(\omega,\theta_v)$, where θ_v is the direction of the noise, and the directivity vector (or HRTFs) of the noise source is $\underline{D}(\omega,\theta_v)$, then similar to equation (2.1) we can write:

$$\underline{X}_v(\omega) = \underline{D}(\omega,\theta_v)V(\omega,\theta_v) \tag{2.4}$$

It should be noted that μ has a very small value and the term μI is introduced either to improve numerical conditioning, to avoid the loss of directivity by microphone mismatch, or to reduce the noise gain, thus it is a tradeoff factor between directivity and robustness.

Keeping in mind that for the case of a target source from a steering direction, the output of the beamformer is obtained by equation (2.2), and also that the MVDR has the constraint of equation (2.3), we also readily see that without any noise from the target source direction θ_s , we then have $Y_s(\omega) = S(\omega, \theta_s)$, and therefore $E\left[|Y_s(\omega)|^2\right] = E\left[|S(\omega, \theta_s)|^2\right]$.

For a narrowband input, by minimizing the noise power subject to the constraint of a distortionless response for a desired direction, the fixed MVDR beamformer (also called superdirective beamformer) produces the best possible signal-to-noise ratio [MOZ'80],[LOT'06].

2.1.3 Diffuse Noise Field

A diffuse noise field is defined as equally distributed and uncorrelated noise source components coming from all directions [LEU'06]. Diffuse noise is representative of many noise environments, for example the babble noise in the cafeteria of a university is quite diffuse-like. The distributed sources $V(\omega, \theta_v)$ have the same average power $E[|V(\omega, \theta_v)|^2] = \sigma_v^2(\omega)$ from any direction, and there is no coherence between the different sources $V(\omega, \theta_v)$ at different angles. In the equations of this thesis, the diffuse noise at the input array sensors is described as resulting from a *discrete* average of the contribution from uniformly distributed sources $V(\omega, \theta_v)$ on the *azimuth plane only* (i.e. the elevation angle is assumed to be zero). That is to say, the diffuse noise field corresponds to what is sometimes referred to in the literature as a "cylindrically isotropic field" [ELK'01], for the particular case where uncorrelated noise components are coming

equally from different directions on the azimuth plane only (from all azimuth angles but for a single elevation angle corresponding to the azimuth plane, which is 0 degree or 90 degrees depending on the notation system being used for the elevation). Moreover, wave propagation in this thesis is always assumed to be in the azimuth plane only, whether for a diffuse noise field or a directional point source. In the literature, a diffuse noise field is normally defined for a "spherically isotropic noise field", where uncorrelated noise components are coming equally from all possible spherical directions. For a microphone array whose components are located on the azimuth plane, these two definitions lead to two different results for the coherence found between two sensor signals in free field, as will be illustrated shortly.

For the case of diffuse noise, equation (2.3) becomes the following solution:

$$\begin{aligned}
\underline{W}(\omega) &= \frac{\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]+\mu I\right)^{-1}\underline{D}(\omega,\theta_s)}{\underline{D}^H(\omega,\theta_s)\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]+\mu I\right)^{-1}\underline{D}(\omega,\theta_s)} \\
&= \frac{\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]\sigma_v^2(\omega)+\mu I\right)^{-1}\underline{D}(\omega,\theta_s)}{\underline{D}^H(\omega,\theta_s)\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]\sigma_v^2(\omega)+\mu I\right)^{-1}\underline{D}(\omega,\theta_s)} \\
&= \frac{\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]+\mu' I\right)^{-1}\underline{D}(\omega,\theta_s)}{\underline{D}^H(\omega,\theta_s)\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]+\mu' I\right)^{-1}\underline{D}(\omega,\theta_s)}
\end{aligned} \tag{2.5}$$

In equation (2.5), N_θ is the number of directions from which the diffuse noise components originate, and the function of the term $\mu' I$ is the same as μI (essentially regularization). It should also be noted that in the case of a directional noise interference the term $1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]$ becomes simply $\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)$.

We see that an overall scale factor on $E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]$ (or on $1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]$) has no effect in the resulting design of $\underline{W}(\omega)$. However, an overall scale factor on $\underline{D}(\omega,\theta_s)$ has a direct inverse relation on the gain of the solution $\underline{W}(\omega)$ produced. We will come back to this in a section to follow.

Moreover, several comments regarding the noise covariance matrix $E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]$ or $1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]$ can be made as follows:

1) It does not always necessarily have a Toeplitz structure, that is to say, the components on each diagonal of the matrix are not necessarily identical, for example in the case of directional (non-diffuse) interference sources in the presence of head shadow effects. In free field the correlation matrix would be Toeplitz, and in head shadow cases under diffuse noise conditions, the matrix would also be approximately Toeplitz.

2) To obtain a correlation matrix with a diagonal having 1.0 values on average, a normalization by the average diagonal component is possible:

$$\frac{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]}{trace\left\{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]\right\}/M}=\frac{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]}{E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)\right]/M}=\frac{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]}{\sum_{i=1}^{M}E[\left|X_{v_i}(\omega)\right|^2/M} \qquad (2.6)$$

The same normalization can also be done for the matrix $1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]$.

This normalization of the matrix $E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]$ by $trace\left\{E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]\right\}/M$ or $E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)\right]/M$ or $\sum_{i=1}^{M}E[\left|X_{v_i}(\omega)\right|^2/M$ can be seen as a global normalization by the global arithmetic average of the sensors power. This can be different from the following alternative normalization, which is to use the pair-wise geometric average $\sqrt{E[\left|X_{v_i}(\omega)\right|^2]}\sqrt{E[\left|X_{v_j}(\omega)\right|^2]}$ $(i,j=1,2...M)$ for a normalization that varies for each element in the correlation matrix. This normalization by $\sqrt{E[\left|X_{v_i}(\omega)\right|^2]}\sqrt{E[\left|X_{v_j}(\omega)\right|^2]}$ $(i,j=1,2...M)$

is the same as the definition of coherence. If $E\left[\left|X_{v_i}(\omega)\right|^2\right]$ is the same for each sensor (for example in free field, or approximately in head shadow cases with diffuse noise), then the two types of normalization become equivalent.

3) For diffuse noise, the coherence function between two input sensors in free field conditions (no head shadow, far field plane wave propagation, anechoic conditions) is often expressed in the literature by the function $\frac{\sin(\omega\, d_{ij}/c)}{\omega\, d_{ij}/c}$, where d_{ij} is the distance between two sensors. It should be noted that this classical form is corresponding to the "spherically isotropic noise field", while the corresponding result for a "cylindrically isotropic field" with the scenario described earlier produces a coherence function with higher sidelobes [ELK'01]. This can be evaluated by a discrete average over azimuth angles of the correlation between theoretical directivity vector components, or it can be obtained from measured directivity vectors obtained in free field. The theoretical coherence function for an isotropic field and for a cylindrically isotropic field are compared in Figure 2.2:

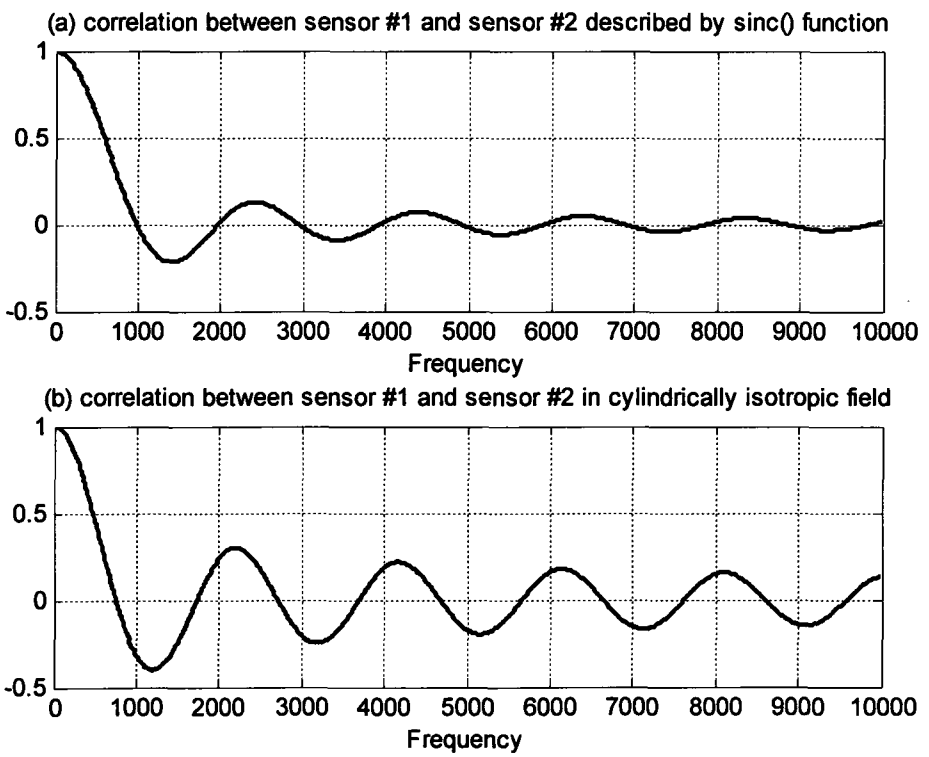


Figure 2.2 Coherences in spherically and cylindrically isotropic diffuse noise fields

The parameters for these figures were chosen as $d_{12} = 0.175$ m, and $c = 344$ m/s which is the speed of sound. It should be noted that the $\frac{\sin(\omega\, d_{ij}/c)}{\omega\, d_{ij}/c}$ formula does illustrate clearly the inversely proportional relationship of the coherence with sensor distance and with frequency.

2.1.4 Normalization of MVDR Beamforming

In order to make sure that the level and phase of the target source component in the beamformer output is the same as the level and phase of the target source component found in one specific input sensor signal (e.g. to avoid signal amplification and group delay distortion in hearing aids), we can consider the problem of scaling the weights of a "regular" MVDR beamformer design from section 2.1.2. In the case of the target source, equation (2.2) becomes:

$$Y_s(\omega) = \underline{W}^H(\omega)\underline{X}_s(\omega) = \underline{W}^H(\omega)\underline{D}(\omega,\theta_s)S(\omega,\theta_s) \qquad (2.7)$$

If the objective is for $Y_s(\omega)$ to keep the same level and phase as the k^{th} target component $X_{s,k}(\omega)$ in the target input sensor vector $\underline{X}_s(\omega)$, then we have that:

$$Y_s(\omega) = \underline{W}^H(\omega)\underline{D}(\omega,\theta_s)S(\omega,\theta_s) = X_{s,k}(\omega) = D_k(\omega,\theta_s)S(\omega,\theta_s) \qquad (2.8)$$

and thus

$$\underline{W}^H(\omega)\underline{D}(\omega,\theta_s) = D_k(\omega,\theta_s) \qquad (2.9)$$

Recognizing that the MVDR design normally has the constraint in equation (2.3):

$$\underline{W}^H(\omega)\underline{D}(\omega,\theta_s) = 1 \qquad (2.10)$$

this simply means that the coefficients $\underline{W}^H(\omega)$ from the "normal" MVDR beamformer design must be scaled by a factor $D_k(\omega,\theta_s)$, or equivalently that the coefficients $\underline{W}(\omega)$ from the "normal" MVDR beamformer design must be scaled by a factor $D_k^*(\omega,\theta_s)$.

The MVDR design above with the above normalization or scaling factor $D_k^*(\omega,\theta_s)$ has some other advantages, on top of preserving the target signal level:

1) Without the above normalization, and assuming here for simplicity that the different components (sensor signals $D_k(\omega,\theta_s)$) in $\underline{D}(\omega,\theta_s)$ have similar magnitude shapes, the filters in the "regular" MVDR design $\underline{W}(\omega)$ end up having a frequency magnitude shaped by the inverse of $\left|D_k(\omega,\theta_s)\right|$. This can be observed directly from the MVDR design equations as in (2.5). On the other hand, with the above normalization and under the same assumption, the filters in the above normalized MVDR design $\underline{W}(\omega)$ end up having a frequency magnitude shape which is more white (i.e. flat).

Similarly, in terms of the beamformer output $Y(\omega)$ signal magnitude, the output of the "regular" MVDR design ends up having a magnitude spectrum with an approximate additional shaping factor $\frac{1}{\left|D_k(\omega,\theta_s)\right|}$ (compared to the signals in $\underline{X}(\omega)$), while on the other hand the output of the above normalized MVDR design does not have this additional shaping of the magnitude spectrum (i.e. it follows more closely the spectral shape of the components in the input vector $\underline{X}(\omega)$, including its original level).

2) Since using the normalized MVDR design of this section produces beamforming filters whose magnitude frequency response is flatter, it also leads to solutions which behave better numerically and require less regularization, for example when converting to time domain beamforming filters. This also means that shorter time domain responses are obtained, thus less coefficients and less complexity.

3) Using the normalized MVDR design of this section will produce beamforming weights $\underline{W}(\omega)$ whose amplitude will no longer vary with the level of the directivity vectors $\underline{D}(\omega,\theta_s)$. This will facilitate the comparison of the Noise Gain (to be defined in a later section) for the beamformers obtained using different sets of directivity vectors $\underline{D}(\omega,\theta_s)$ (or head models), each possibly with different overall scaling factors.

2.2 Classical Beamformer Performance Measures

2.2.1 Beampattern

The beampattern, also called directivity pattern $DP(\omega,\theta)$, is defined as the squared magnitude response (or power gain) for a signal from a single arbitrary direction. From equation (2.7), we can have

$$\begin{aligned} E\left[\left|Y_s(\omega)\right|^2\right] &= E\left[\left(\underline{W}^H(\omega)\underline{X}_s(\omega)\right)\left(\underline{W}^H(\omega)\underline{X}_s(\omega)\right)^H\right] \\ &= \underline{W}^H(\omega)E\left[\underline{X}_s(\omega)\underline{X}_s{}^H(\omega)\right]\underline{W}(\omega) \\ &= \underline{W}^H(\omega)\left(\underline{D}(\omega,\theta_s)\underline{D}^H(\omega,\theta_s)E\left[\left|S(\omega,\theta_s)\right|^2\right]\right)\underline{W}(\omega) \\ &= \left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|^2 E\left[\left|S(\omega,\theta_s)\right|^2\right] \end{aligned} \tag{2.11}$$

and thus we obtain:

$$DP(\omega,\theta_s) = \left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|^2 \tag{2.12}$$

The beampattern is defined here for plane waves (far field propagation, point sources). If those conditions are not met, then there are other factors that would need to be included: distance, source radiation model and orientation, etc.

The beampattern is a very effective and direct way to show the directivity of the beamformers. As an illustration, in free field for a frontal target with $M = 2$, the beampattern in the following figure shows that for diffuse noise a fixed first-order differential microphone combining two back-to-back cardioids performs similarly to a fixed MVDR beamformer tuned with an endfire configuration. The first-order differential microphone system is designed as follows [GOR'08]:

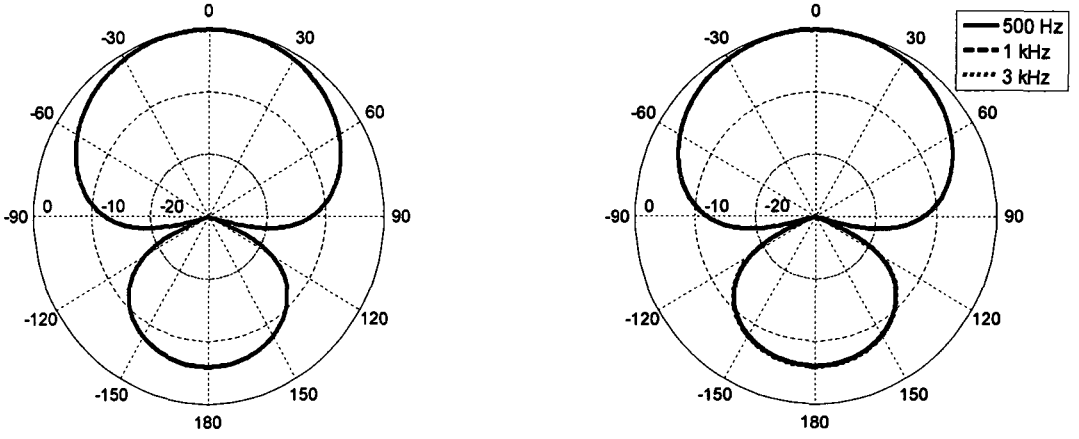
$$\underline{W}(\omega)=\begin{bmatrix}\dfrac{1+\alpha e^{-j\omega T}}{1-e^{-j\omega 2T}}\\ \dfrac{-\alpha-e^{-j\omega T}}{1-e^{-j\omega 2T}}\end{bmatrix} \tag{2.13}$$

$$\alpha=\frac{1+\cos\theta_v}{1-\cos\theta_v}, T=\frac{d}{c}$$

$$90°\le\theta_V\le 180°$$

$\theta_V = 109°$ for diffuse noise

and the respective beampatterns for the fixed 1st order directional microphone and fixed MVDR beamformer are shown in Figure 2.3 [GOR'08]. Please note that in Figure 2.3, the beampatterns for frequencies 500 Hz, 1 kHz and 3 kHz are all the same.



(a) 1st order directional microphone (b) endfire MVDR beamformer ($M = 2$)

Figure 2.3 Comparison between directional microphone and endfire MVDR beamformer tuned for diffuse noise ($M = 2$)

2.2.2 Array Gain

The Array Gain $AG(\omega)$ is defined as:

$$AG(\omega) = \frac{\text{signal to noise ratio (SNR) with beamforming}}{\text{signal to noise ratio (SNR) without beamforming}} \tag{2.14}$$

where each "signal to noise ratio (SNR)" is defined as the average power from a directional source (for which the beamformer is steered) over the average power from the all the noise components. Below we have considered the case of diffuse noise. In the derivations, the simpler directional noise case could be obtained by simply removing the angle-dependant averages on the noise statistics.

Similar to equation (2.11), the beamformer noise output for the diffuse noise field case is:

$$\begin{aligned}
&E\left[\left|Y_V(\omega)\right|^2\right] = E\left[\left(\underline{W}^H(\omega)\underline{X}_v(\omega)\right)\left(\underline{W}^H(\omega)\underline{X}_v(\omega)\right)^H\right] \\
&= \underline{W}^H(\omega)E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]\underline{W}(\omega) \\
&= \underline{W}^H(\omega)\left(1/N_\theta \sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]\sigma_v^2(\omega)\right)\underline{W}(\omega) \\
&= 1/N_\theta \sum_{\theta_v}\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_v)\right|^2\sigma_v^2(\omega)
\end{aligned} \tag{2.15}$$

For the target signals measured by the sensors we have:

$$\begin{aligned}
&E\left[\left\|\underline{X}_s(\omega)\right\|^2\right] = E\left[\underline{X}_s{}^H(\omega)\underline{X}_s(\omega)\right] \\
&= \underline{D}^H(\omega,\theta_s)\underline{D}(\omega,\theta_s)E\left[\left|S(\omega,\theta_s)\right|^2\right] \\
&= \left\|\underline{D}(\omega,\theta_s)\right\|^2 E\left[\left|S(\omega,\theta_s)\right|^2\right]
\end{aligned} \tag{2.16}$$

where $\underline{D}(\omega,\theta_v)$ follows the same definition as $\underline{D}(\omega,\theta_s)$ earlier.

Correspondingly, for the noise signals measured by the sensors we have:

$$\begin{aligned}
&E\left[\left\|\underline{X}_v(\omega)\right\|^2\right] = E\left[\underline{X}_v{}^H(\omega)\underline{X}_v(\omega)\right] \\
&= 1/N_\theta \sum_{\theta_v}\left[\underline{D}^H(\omega,\theta_v)\underline{D}(\omega,\theta_v)\right]\sigma_v^2(\omega) \\
&= 1/N_\theta \sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_v)\right\|^2\right]\sigma_v^2(\omega)
\end{aligned} \tag{2.17}$$

Thus the signal to noise ratio (SNR) with beamforming is the following:

$$\frac{E\left[\left|Y_s(\omega)\right|^2\right]}{E\left[\left|Y_v(\omega)\right|^2\right]} = \frac{\underline{W}^H(\omega)E\left[\underline{X}_s(\omega)\underline{X}_s{}^H(\omega)\right]\underline{W}(\omega)}{\underline{W}^H(\omega)E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right]\underline{W}(\omega)}$$
$$= \frac{\underline{W}^H(\omega)\left(\underline{D}(\omega,\theta_s)\underline{D}^H(\omega,\theta_s)E\left[\left|S(\omega,\theta_s)\right|^2\right]\right)\underline{W}(\omega)}{\underline{W}^H(\omega)\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_v)\underline{D}^H(\omega,\theta_v)\right]\sigma_v^2(\omega)\right)\underline{W}(\omega)} \qquad (2.18)$$
$$= \frac{\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|^2}{1/N_\theta\sum_{\theta_v}\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_v)\right|^2}\frac{E\left[\left|S(\omega,\theta_s)\right|^2\right]}{\sigma_v^2(\omega)}$$

and the signal to noise ratio without beamforming is the following:

$$\frac{E\left[\left\|\underline{X}_s(\omega)\right\|^2\right]}{E\left[\left\|\underline{X}_v(\omega)\right\|^2\right]} = \frac{E\left[\underline{X}_s{}^H(\omega)\underline{X}_s(\omega)\right]}{E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)\right]}$$
$$= \frac{\left\|\underline{D}(\omega,\theta_s)\right\|^2 E\left[\left|S(\omega,\theta_s)\right|^2\right]}{1/N_\theta\sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_v)\right\|^2\right]\sigma_v^2(\omega)} \qquad (2.19)$$

The ratio between the two SNRs is the Array Gain $AG(\omega)$ and it can have the following forms:

$$AG(\omega)=\frac{\underline{W}^H(\omega)E\left[\underline{X}_s(\omega)\underline{X}_s{}^H(\omega)\right]\underline{W}(\omega)}{\underline{W}^H(\omega)E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]\underline{W}(\omega)}\times\frac{E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)\right]}{E\left[\underline{X}_s{}^H(\omega)\underline{X}_s(\omega)\right]}$$

$$=\frac{\underline{W}^H(\omega)\dfrac{E\left[\underline{X}_s(\omega)\underline{X}_s{}^H(\omega)\right]}{E\left[\underline{X}_s{}^H(\omega)\underline{X}_s(\omega)\right]}\underline{W}(\omega)}{\underline{W}^H(\omega)\dfrac{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]}{E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)\right]}\underline{W}(\omega)}$$

$$=\frac{\underline{W}^H(\omega)\dfrac{E\left[\underline{X}_s(\omega)\underline{X}_s{}^H(\omega)\right]}{E\left[\underline{X}_s{}^H(\omega)\underline{X}_s(\omega)/M\right]}\underline{W}(\omega)}{\underline{W}^H(\omega)\dfrac{E\left[\underline{X}_V(\omega)\underline{X}_V{}^H(\omega)\right]}{E\left[\underline{X}_V{}^H(\omega)\underline{X}_V(\omega)/M\right]}\underline{W}(\omega)}$$

$$=\frac{\underline{W}^H(\omega)R_{ss}(\omega)\underline{W}(\omega)}{\underline{W}^H(\omega)R_{VV}(\omega)\underline{W}(\omega)}$$

$$AG(\omega)=\left(\frac{\underline{W}^H(\omega)\left(\underline{D}(\omega,\theta_s)\underline{D}^H(\omega,\theta_s)E\left[|S(\omega,\theta_s)|^2\right]\right)\underline{W}(\omega)}{\underline{W}^H(\omega)\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_V)\underline{D}^H(\omega,\theta_V)\right]\sigma_v^2(\omega)\right)\underline{W}(\omega)}\right)\times\left(\frac{1/N_\theta\sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_V)\right\|^2\right]\sigma_v^2(\omega)}{\left\|\underline{D}(\omega,\theta_s)\right\|^2E\left[|S(\omega,\theta_s)|^2\right]}\right)$$

$$=\left(\frac{\underline{W}^H(\omega)\left(\underline{D}(\omega,\theta_s)\underline{D}^H(\omega,\theta_s)\right)\underline{W}(\omega)}{\underline{W}^H(\omega)\left(1/N_\theta\sum_{\theta_v}\left[\underline{D}(\omega,\theta_V)\underline{D}^H(\omega,\theta_V)\right]\right)\underline{W}(\omega)}\right)\times\left(\frac{1/N_\theta\sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_V)\right\|^2\right]}{\left\|\underline{D}(\omega,\theta_s)\right\|^2}\right)$$

$$=\frac{\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|^2}{1/N_\theta\sum_{\theta_v}\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_V)\right|^2}\times\left(\frac{1/N_\theta\sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_V)\right\|^2\right]}{\left\|\underline{D}(\omega,\theta_s)\right\|^2}\right)=\frac{DP(\omega,\theta_s)}{1/N_\theta\sum_{\theta_v}DP(\omega,\theta_v)}\times\frac{1/N_\theta\sum_{\theta_v}\left[\left\|\underline{D}(\omega,\theta_V)\right\|^2\right]}{\left\|\underline{D}(\omega,\theta_s)\right\|^2}$$

(2.20)

It should be noted that for a MVDR beamformer the design is done subjected to the condition in equation (2.10) , and therefore we have:

$$\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|^2=DP(\omega,\theta_s)=1 \qquad (2.21)$$

Moreover, in free field we also have:

$$1/N_\theta \sum_{\theta_v} \left[\left\| \underline{D}(\omega,\theta_v) \right\|^2 \right] = \left\| \underline{D}(\omega,\theta_s) \right\|^2 \qquad (2.22)$$

Therefore, in free field the Array Gain $AG(\omega)$ becomes equivalent to the Directivity Index $DI(\omega)$, which is defined as the ratio of the power gain from the target direction over the average power gain from all the other directions, shown here on the azimuth plane only and for a discrete sum of angles:

$$DI(\omega) = \frac{DP(\omega,\theta_s)}{1/N_\theta \sum_{\theta_v} DP(\omega,\theta_v)} \qquad (2.23)$$

Note that the Directivity Index is specific to diffuse noise (and moreover limited to far field propagation, plane wave, point source model), while the Array Gain can be defined for other types of noise as well (e.g. directional, etc.) and for other types of propagation models.

2.2.3 Noise Gain

The definition used in this thesis for the noise gain $G(\omega)$ is the squared magnitude response (or power gain) for self-generated (e.g. internal, thermal, spatially uncorrelated, spatially white) microphone noise. Alternative definitions also exist in the literature.

From equation (2.15), for a randomly distributed source we have:

$$E\left[|Y_v(\omega)|^2 \right] = \underline{W}^H(\omega) E\left[\underline{X}_v(\omega) \underline{X}_v{}^H(\omega) \right] \underline{W}(\omega) \qquad (2.24)$$

where $E\left[\underline{X}_v(\omega) \underline{X}_v{}^H(\omega) \right]$ for spatially white noise with the same power σ_w^2 on each sensor has the form $\sigma_w^2 \underline{I}$, where $\underline{I}$ is an identity matrix.

The output component power can then be written as:

$$E\left[|Y_v(\omega)|^2 \right] = \underline{W}^H(\omega) \underline{W}(\omega) \sigma_w^2 \qquad (2.25)$$

Thus the noise power gain between the input noise source and the output of the beamformer becomes:

$$G(\omega) = \underline{W}^{H}(\omega)\underline{W}(\omega) \tag{2.26}$$

It can be observed easily that the Noise Gain is directly influenced by the values of the beamforming weights. By normalizing the MVDR beamformer weights as in section 2.1.4, we can make sure that the comparisons of Noise Gains between different Head Models (section 3.5 from the next chapter) will all be on the same scale.

Chapter 3 MVDR Beamformer Design for Binaural Hearing Aids

3.1 Concept of Binaural Hearing Aids

The hearing aids currently existing in the market are either monaural hearing aids or bilateral hearing aids. A monaural hearing aid is a hearing aid which combines only its own microphone inputs to generate an output. Figure 2.1 could serve as an example of a monaural hearing aid. To avoid confusion, it is important to emphasize the differences between bilateral and binaural hearing aids. A user wearing a monaural hearing aid on each ear is said to be using *bilateral hearing aids* [KLA'07]. That is to say, a pair of bilateral hearing aids is composed of two independent monaural hearing aids on each side, and typically little information is shared between the hearing aids from the two ears. In contrast, in binaural hearing aids, it is assumed that a wireless link allowing the exchange of information between the two sides is available, so that the outputs of both ears are obtained by processing the sensor inputs from both the left and right sides.

As we can see, developing and using binaural hearing aids means more complexity and cost, so this brings the question: what are the advantages of binaural hearing aids compared with monaural and bilateral hearing aids? Some potential advantages are:

- Binaural hearing aids perform best in terms of the beamforming noise reduction, because it is known that doubling the number of microphones in a microphone array can normally increase the SNR of the output. Due to the size limitations, three microphones is the maximum number of sensors that can be found in practice in a hearing aid, and normally only two microphones are found, as users prefer small or even unnoticeable hearing aids.
- Let's consider a scenario where a hearing impaired person is wearing a monaural hearing aid on his or her left ear, but a target source signal is coming from the right ear direction. The SNR at the output of the left ear hearing aid will still be

low due to the head shadow effect. Wearing bilateral hearing aids could possibly help in this case, but for people who are having asymmetric hearing loss, more confusion and sound localization problems can be brought when using bilateral hearing aids.

- If they can preserve the spatial cues, binaural hearing aids can further help to improve the intelligibility of the speech, by preserving the spatial masking property of the human binaural processing, e.g. for the "cocktail party effect".

3.2 Design Constraints for Beamforming in Hearing Aids

For a practical realization, some the design constraints for beamforming processing in hearing aids are:

3.2.1 Physical Constraints

Comparing with other microphone array applications, one limitation that hearing aids have is their physical constraints such as size and shape. Generally speaking, commercial hearing aids are available in four forms including completely-in-the-canal (CIC), in the canal (ITC), in the ear (ITE) and behind the ear (BTE) models [DIL'01]. Among them, CIC and ITC are designed to be very small, which can accommodate only one microphone per unit, so that they can fit within the ear canal. In contrast, ITE and BTE models are bigger, which can hold two or even occasionally three microphones, but the distances between each microphone is limited to 8-10 mm [PUD'06].

3.2.2 Microphone Noise

Microphone noise can be amplified during the processing, as indicated by the classical Noise Gain measure defined in Section 2.2.3. It should be noted that there is a tradeoff between the noise gain and the directivity index of beamformers [COX'87].

3.2.3 Head Shadow Effects and Model Mismatch

Depending on the beamforming design algorithm used, head shadow effects and model mismatch can also produce some performance degradation. For example, a fixed MVDR design for diffuse noise assumes some model for the sound propagation from different directions. If a free field propagation is assumed and if in practice there are some head shadow effects, then this can lead to a loss of performance. Alternatively, if the head shadow effects are modeled using some known HRTFs and in practice the HRTFs are not the same as the ones used for the fixed design, then this can also lead to a loss of performance. The effect of model mismatch will be evaluated in some of the experiments of this thesis.

3.2.4 Microphone Mismatch

Theoretically speaking, all of the microphones in the microphone array should be identical to each other. However, it is not the case in real life, and the amplitude or phase mismatch among microphones can produce some performance degradations [PUD'06]. Most of the experiments in this thesis are based on recordings provided by a hearing aid manufacturer and which already include some typical microphone mismatch conditions, therefore microphone mismatch will not be further considered in this thesis.

3.2.5 Position Mismatch

Movements of the hearing aid user or touching the hearing aid can cause changes in the positions of the microphones. This means that the head model used for fixed MVDR design which was assuming that the hearing aid was in a specific orientation (e.g. on the horizontal plane) is not exactly valid, and there will be an additional mismatch there. This problem of position mismatch is common to both monaural and binaural fixed MVDR beamformers, and it will not be further considered in this thesis. Therefore, the results presented in this thesis can be considered as "favourable" in the sense that they do not consider the effect of position mismatch.

3.3 Binaural MVDR Beamforming

The structure of a binaural hearing aid using a binaural MVDR beamformer is shown below:

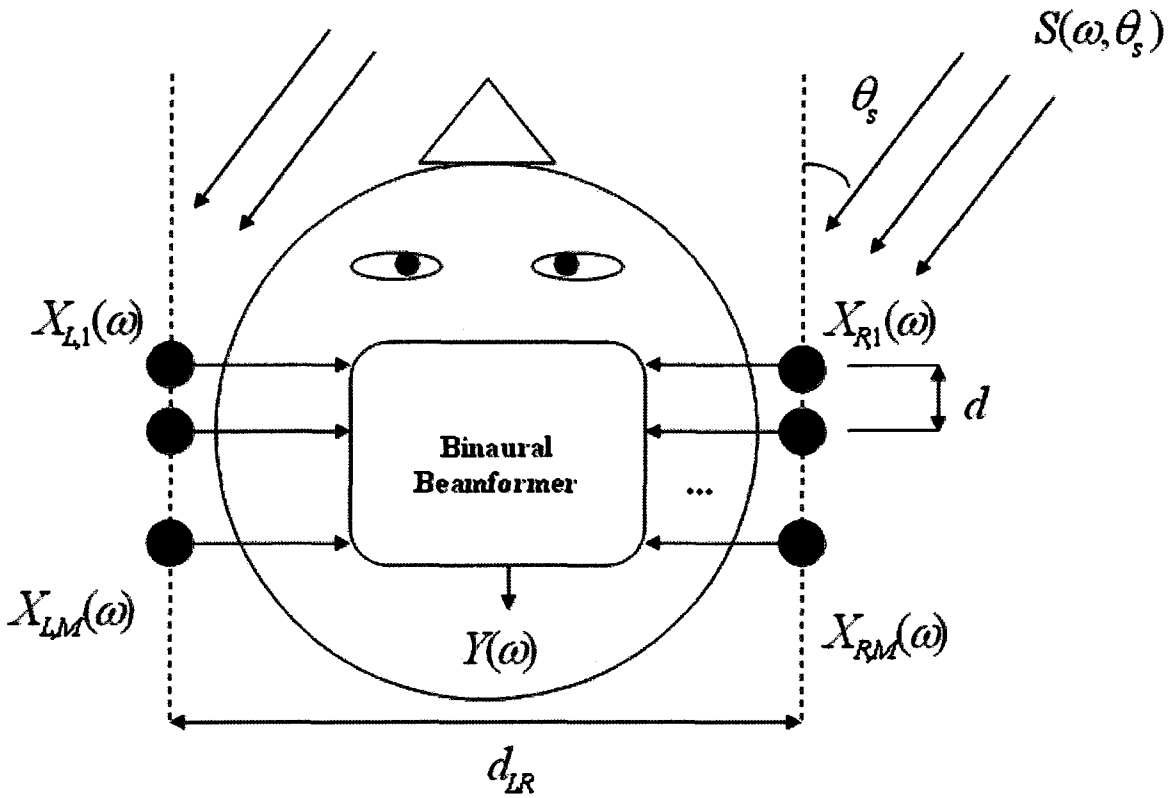


Figure 3.1 Block diagram of a binaural hearing aid beamformer

From Figure 3.1, it can be seen that a binaural hearing aid consists of two linear microphone arrays mounted on the right and left ear separately. Each side is composed of M microphones equally spaced by distance d, and θ_s is the angle of the target signal. The distance between the arrays on the two sides is d_{LR}, and $X_{R,i}$ and $X_{L,i}$ $(i=1,2...M)$ are the frequency domain signals measured by the sensor on the right and left side respectively. Analogous to Section 2.1.1, for the target source $S(\omega,\theta_s)$, we have:

$$\underline{X}_{s,R}(\omega)=\underline{D}_R(\omega,\theta_s)S(\omega,\theta_s) \qquad (3.1)$$

$$\underline{X}_{s,L}(\omega) = \underline{D}_L(\omega,\theta_s)S(\omega,\theta_s) \tag{3.2}$$

where $\underline{D}_R(\omega,\theta_s)$ and $\underline{D}_L(\omega,\theta_s)$ are the HRTFs (or directivity vectors) for the target source direction of the right and left ears respectively. Similarly for the noise $V(\omega,\theta_v)$ we have

$$\underline{X}_{v,R}(\omega) = \underline{D}_R(\omega,\theta_v)V(\omega,\theta_v) \tag{3.3}$$

$$\underline{X}_{v,L}(\omega) = \underline{D}_L(\omega,\theta_v)V(\omega,\theta_v) \tag{3.4}$$

Thus the measured signals are:

$$\underline{X}_R(\omega) = \underline{X}_{s,R} + \underline{X}_{v,R} \tag{3.5}$$

$$\underline{X}_L(\omega) = \underline{X}_{s,L} + \underline{X}_{v,L} \tag{3.6}$$

Writing the input signal as $\underline{X}(\omega) = [\underline{X}_R^T(\omega) \;\; \underline{X}_L^T(\omega)]^T$, the output of a binaural beamformer is generated by linearly filtering and combining the $2M$ microphone signals using a length-$2M$ linear filter $\underline{W}(\omega)$:

$$Y(\omega) = \underline{W}^H(\omega)\underline{X}(\omega) \tag{3.7}$$

where $\underline{X}(\omega)$ and $\underline{W}(\omega)$ are $2M \times 1$ vectors. Considering the right and left augmented source direction vector $\underline{D}\,(\omega,\theta_s) = [\underline{D}_R^T(\omega,\theta_s) \;\; \underline{D}_L^T(\omega,\theta_s)]^T$ and the augmented noise vector $\underline{X}_v(\omega) = [\underline{X}_{v,R}^T(\omega) \;\; \underline{X}_{v,L}^T(\omega)]^T$, the design criteria of MVDR beamforming becomes:

$$\min_{\underline{W}\,(\omega)} E\{|Y(\omega)|^2\} \text{ subject to } \underline{W}^H(\omega)\underline{D}\,(\omega,\theta_s) = 1 \tag{3.8}$$

and the binaural MVDR beamformer coefficients can be obtained as:

$$\underline{W}(\omega) = \frac{\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right] + \mu I\right)^{-1}\underline{D}(\omega,\theta_s)}{\underline{D}^H(\omega,\theta_s)\left(E\left[\underline{X}_v(\omega)\underline{X}_v{}^H(\omega)\right] + \mu I\right)^{-1}\underline{D}(\omega,\theta_s)} \tag{3.9}$$

It is of great importance to highlight the point that as in monaural beamformers, a binaural beamformer only produces a single monaural output $Y(\omega)$. We will come back to this limitation in a future section.

3.4 Array Configurations for Hearing Aids

3.4.1 Monaural 1

The Monaural 1 configuration is a configuration in which only one microphone is mounted on either the right ear or left ear. Monaural 1 can be seen as the case of Section 2.1.1 when $M = 1$, which is the extreme case of a microphone array with only a single microphone, i.e. there is no beamforming performed at all. It should be noted that in Figure 3.2, only the case for the right ear is shown.

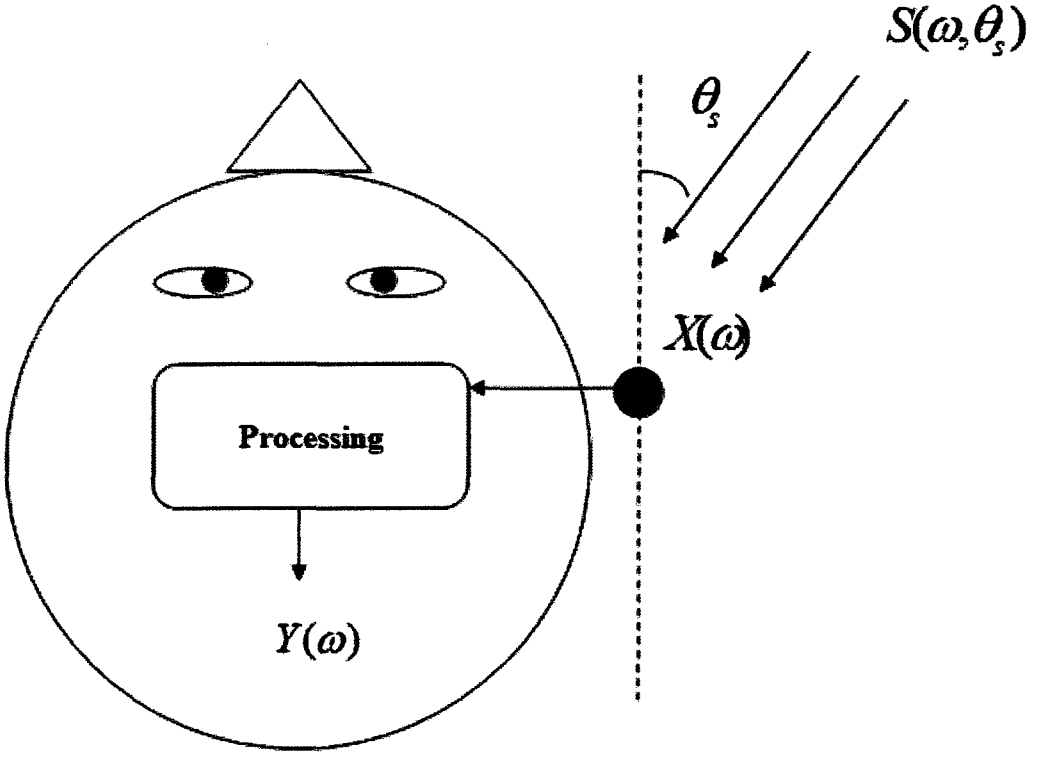


Figure 3.2 Block diagram of configuration Monaural 1

3.4.2 Monaural 2

Monaural 2 is the case of Section 2.1.1 when $M = 2$. For a frontal target case, it represents an endfire microphone array with two microphones and spacing d. Again only the case for the right ear is shown here:

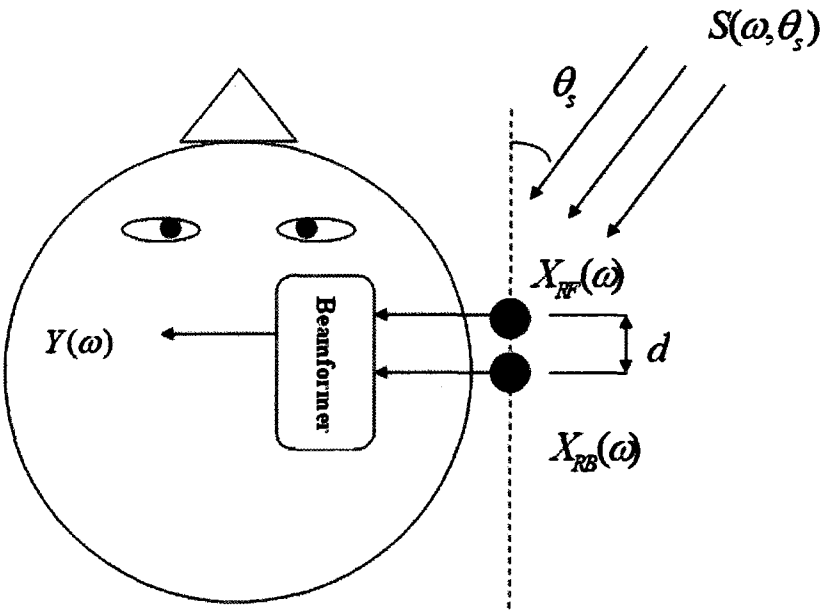


Figure 3.3 Block diagram of configuration Monaural 2

In Figure 3.3 and subsequent figures, we use $X_{RF}(\omega)$, $X_{RB}(\omega)$, $X_{LF}(\omega)$ and $X_{LB}(\omega)$ to represent signals measured by right-frontal, right-back, left-frontal and left-back microphones respectively.

3.4.3 Binaural 1 + 1

Binaural 1 + 1 is the case of Figure 3.1 with $M = 1$, which means one microphone for each ear. In contrast to the Monaural 2 configuration, for the frontal target case this is a broadside array configuration. It is important to emphasize that the two microphones are not independent in this case, and that there is link available to share the information between the two sides.

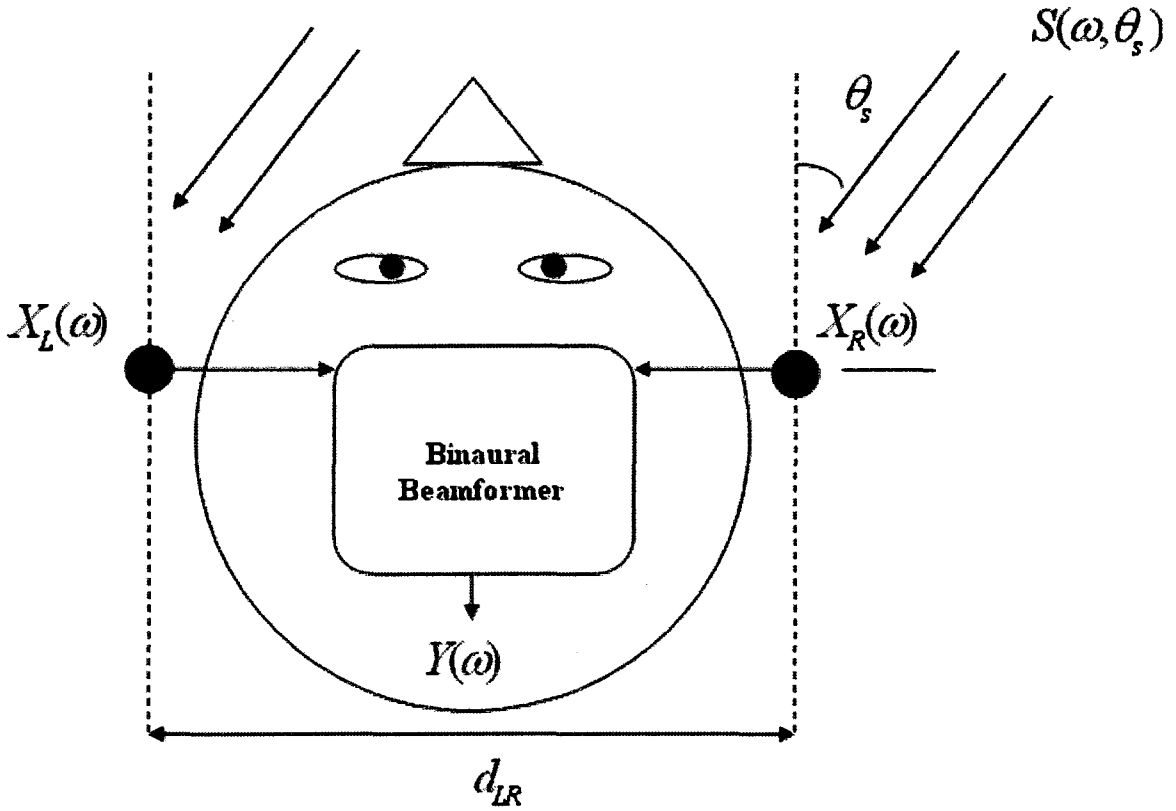


Figure 3.4 Block diagram of configuration Binaural 1 + 1

3.4.4 Binaural 2 + 2

Binaural 2 + 2 is the case of Figure 3.1 with $M = 2$, which means there are two-microphones for each ear, as shown in Figure 3.5. For the frontal target case, this is a 2-D array that has both endfire and broadside components. As for the Binaural 1+1 case, there is a link between the two sides allowing the sharing of signal information.

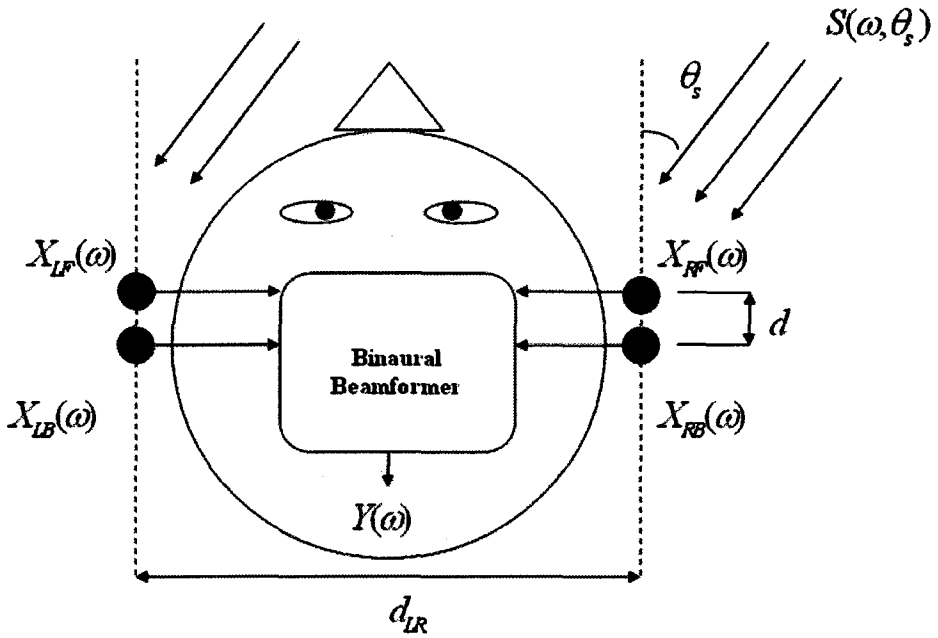


Figure 3.5 Block diagram of configuration Binaural 2 + 2

3.5 Head Models

3.5.1 Pole-Zero Spherical Head Model

In [BRO'98], an approximation of a spherical head model was presented to synthesize binaural sounds from a monaural sound, producing well-controlled horizontal and vertical effects. ITDs are well approximated by Woodworth and Schlosberg's frequency-independent formula [MOO'89], while ILDs are modeled by a single-pole and a single-zero head shadow filter, cascaded with a delay element for the ITDs. This provides a good approximation of the ideal frequency response of a rigid sphere.

The model is implemented by the following equations:

$$\tau_{\text{mod}} = \begin{cases} \frac{a}{c}|\theta| & -\frac{\pi}{2} \le \theta < 0 \\ -\frac{a}{c}\cos(\theta - \frac{\pi}{2}) & 0 \le \theta < \frac{\pi}{2} \end{cases} \tag{3.10}$$

where a is the radius of the head ($a = 0.0875\text{m}$ is commonly considered as an average radius of an adult human head). c is the ambient speed of sound and it is often assumed to be $c = 344\text{m/s}$.

$$\gamma_{\text{mod}} = \left(1 + \frac{\beta_{\min}}{2}\right) + \left(1 - \frac{\beta_{\min}}{2}\right)\cos\left(\frac{\theta - \frac{\pi}{2}}{\theta_{\min}} 180^{\circ}\right) \tag{3.11}$$

where the parameters are chosen as $\beta_{\min} = 0.1$ and $\theta_{\min} = 150^{\circ}$.

$$D_{\text{mod}}(\theta, \omega) = \frac{1 + j(\frac{\gamma_{\text{mod}}(\theta)\omega}{2\omega_0})}{1 + j(\frac{\omega}{2\omega_0})} e^{-j\omega\tau_{\text{mod}}(\theta)} \text{ with } \omega_0 = \frac{c}{a} \tag{3.12}$$

Taking the right front microphone as a reference, and then applying the head model to a certain target direction θ_s, we can have the HRTFs modeled by this Pole-Zero Spherical Model for each configuration :

- Monaural 1

$$D(\theta_s, k) = D_{\text{mod}}(\theta_s, \omega_k) \tag{3.13}$$

- Binaural 1+1

$$\begin{aligned} D_{RF}(\theta_s, k) &= D_{\text{mod}}(\theta_s, \omega_k) \\ D_{LF}(\theta_s, k) &= D_{\text{mod}}(\theta_s - \pi, \omega_k) \end{aligned} \tag{3.14}$$

We can also approximate the HRTFs of the rear microphones of the configuration Monaural 2 and Binaural 2+2 by an additional free field propagation from the front microphone on the same side, and assuming that the distance between microphones on the same side is $d = 0.0075\text{ m}$:

- Monaural 2

$$D_{RF}(\theta_s,k) = D_{\text{mod}}(\theta_s,\omega_k)$$
$$D_{RB}(\theta_s,k) = D_{RF}(\theta_s,k)e^{-j\omega\frac{d}{c}\cos\theta_s} \tag{3.15}$$

- Binaural 2+2

$$D_{RF}(\theta_s,k) = D_{\text{mod}}(\theta_s,\omega_k)$$
$$D_{LF}(\theta_s,k) = D_{\text{mod}}(\theta_s-\pi,\omega_k)$$
$$D_{RB}(\theta_s,k) = D_{RF}(\theta_s,k)e^{-j\omega\frac{d}{c}\cos\theta_s} \tag{3.16}$$
$$D_{LB}(\theta_s,k) = D_{LF}(\theta_s,k)e^{-j\omega\frac{d}{c}\cos\theta_s}$$

For simplicity, the Pole-Zero Spherical Head Model is given the abbreviation PZS in this thesis.

3.5.2 Range Dependent Spherical Head Model

In addition to the previous PZS model, the model from [DUD'98] also incorporates the distance of the source, so that the HRTFs modeled vary with range as well as with azimuth and elevation. As a result, this allows to better model the near-field effects and the ripples that they introduce in the frequency response. This is also a parametric model based on the classical spherical head model, so it is named Range Dependent Spherical Head Model, or RDS for short in this thesis.

Varying with three factors (range, frequency and direction), the HRTFs produced by the RDS model can be represented as:

$$H(\rho,\mu,\theta) = \frac{\rho}{i\mu}e^{-i\mu}\sum_{m=0}^{\infty}(2m+1)P_m(\cos\theta)\frac{Q_m(\frac{1}{i\mu\rho})}{\frac{m+1}{i\mu}Q_m(\frac{1}{i\mu})-Q_{m-1}(\frac{1}{i\mu})},\rho>1 \tag{3.17}$$

where $\rho = \frac{r}{a}$, r is the distance from the center of the sphere to the source and a is the radius of the sphere (m) (as before the radius is selected to be $a = 0.0875\text{m}$). The frequency factor is represented by $\mu = \frac{2\pi fa}{c}$, where c is the ambient speed of sound, set to $c = 344\text{m/s}$. The RDS model can be implemented by a recursive algorithm to compute the required $P_m()$ and $Q_m()$ factors [DUD'98].

Assuming that the right ear is located at $\theta_r = 100^\circ$, and the left ear is located at $\theta_l = -100^\circ$, and taking the right front sensor as a reference, we can produce the different HRTFs as:

$$D_{RF}(\rho,\theta,k) = H(\rho,\theta-\theta_r,\omega_k)$$
$$D_{LF}(\rho,\theta,k) = H(\rho,\theta-\theta_l,\omega_k) \qquad (3.18)$$

Similarly to the PZS, the same delay approximation as in equation (3.15) and (3.16) can be used to produce HRTFs for the rear microphones.

Due to the fact that under most circumstances, people are talking within a near-field distance, the range parameter was chosen as $\rho = 4$ in this thesis for this model, which means that the distance between the target speaker and the center of the hearing-aid user's head is assumed to be $r = 4 * a = 0.35\text{m} = 35\text{cm}$.

3.5.3 Measured HRTF Head Model

A third "model" uses the multichannel HRTFs from a hearing-aid manufacturer's binaural measurements. The measurements were obtained from two 3-channel behind-the-ear (BTE) hearing aids mounted on a KEMAR dummy head in an anechoic room. To measure the HRTFs, a dummy head was rotated towards different directions to receive the target source at a resolution of 5 degrees in the azimuth plane, and the target speech signal was produced by a loudspeaker at 1.5 meters from the dummy head. This is shown in Figure 3.6 with $\theta_{n+1} - \theta_n = 5^\circ$:

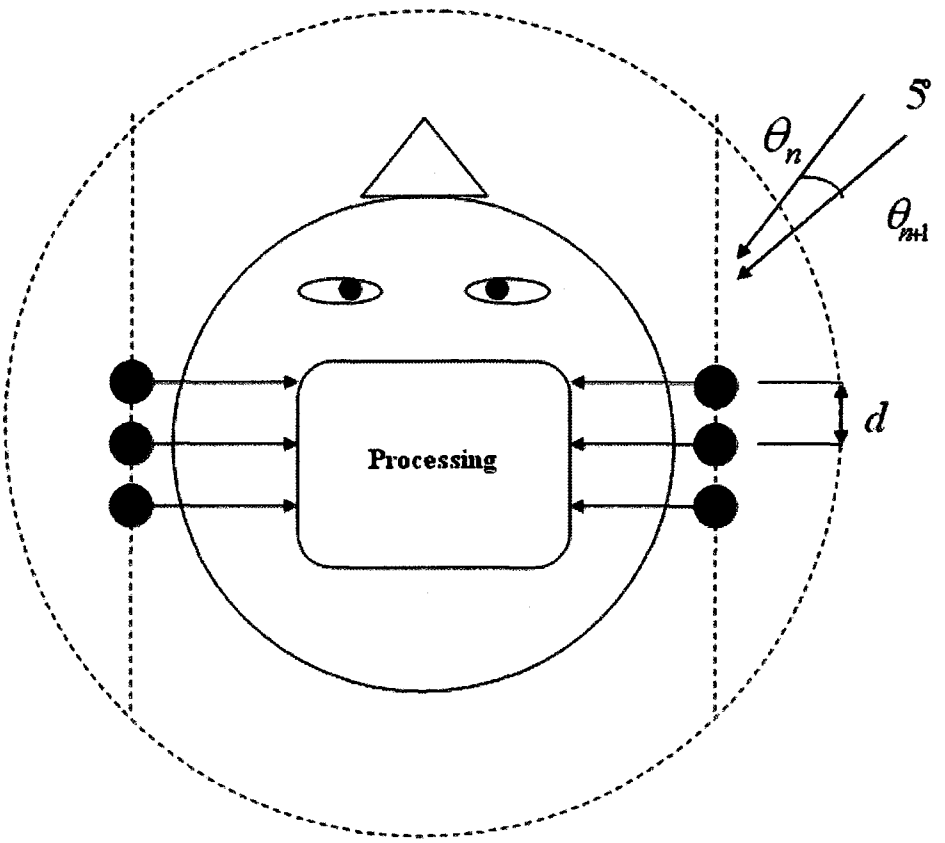


Figure 3.6 Measurement of HRTFs from a KEMAR dummy head

The HRTFs from the first two microphones on each side are directly used (as needed) to fill the required directivity vectors for the MVDR design. The measured HRTFs used for this thesis were built by using only the left half part of the measurements (for $-180^\circ \leq \theta \leq 0^\circ$ on the azimuth plane), and by using symmetry properties to generate the right half part (i.e. HRTFs on the right plane correspond to their "mirror" HRTFs on the left plane, with left and right channels inverted). This model will be referred to as MHRTF (Measured HRTF) in the rest of the thesis.

3.5.4 Free Field Head Model

The Free Field (FF) Head Model does not consider any head shadow effects at all. Strictly speaking, it is not a head model since it assumes in the equations that there is no head. With respect to the right front microphone and given array microphone distances

on the same side of $d = 0.0075$ m and between sides of $d_{LR} = 0.0175$ m, the directivity vector components for the right side and left side microphones can be obtained respectively by:

$$e^{-j\omega(\frac{(i-1)d}{c}\cos\theta_s)}$$
$$e^{-j\omega(\frac{d_{LR}}{c}\sin\theta_s+\frac{(i-1)d}{c}\cos\theta_s)} \qquad (3.19)$$

where i represents the i^{th} microphone in the right or left side. For each microphone configuration (Monaural 1, Monaural 2, Binaural 1+1, Binaural 2+2), the directivity vector can be filled in a straightforward way using (3.19).

3.5.5 Comparison between the Four Head Models

Among the four models introduced in the previous section, the MHRTF is a model based on experimental measures, and the other three are parametric models. Between PZS, RDS and FF, mathematically speaking FF is the simplest model among them, and it does not consider any head shadow effect at all. From the head models including the head shadow effects (i.e. PZS, RDS, and MHRTF), the MHRTF is supposed to be the one which includes the most information, since the others are only approximations. However, one problem with this model is that the HRTFs can vary greatly from person to person, and they also depend on the position of the hearing aid. In real life it is not feasible to obtain HRTFs from each hearing aid user (and for each hearing aid position). With the PZS and RDS models, some parameters such as head radius can be adjusted, which can be an advantage.

Figure 3.7 and Figure 3.8 provide a quick comparison between the magnitude directivity responses of the models for $\theta = 60^\circ$ and $\theta = -20^\circ$. For each model, the magnitude response in the figures is normalized by the magnitude of the frontal response (from each side).

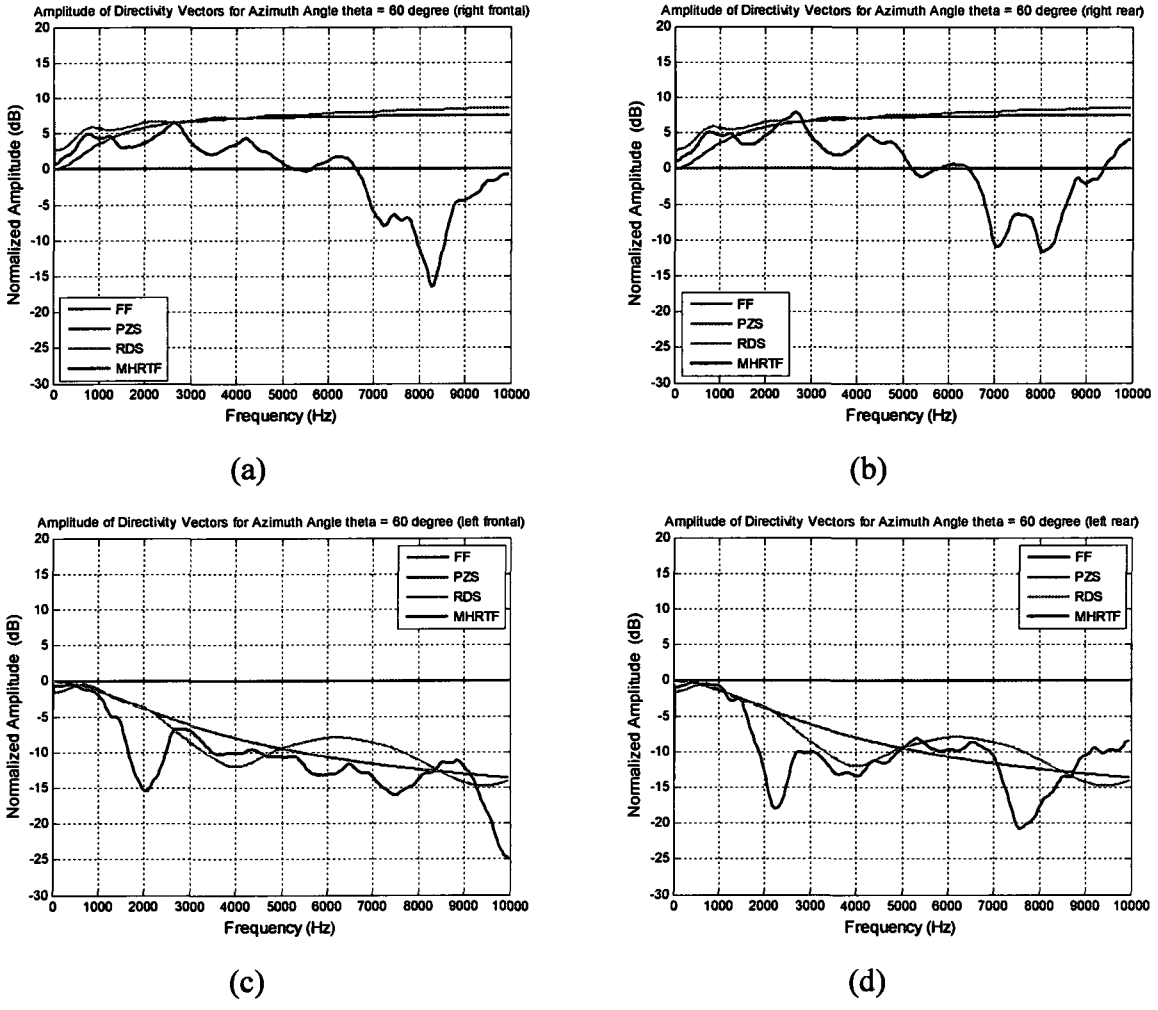


Figure 3.7 Magnitude of normalized directivity vectors for $\theta = 60^{\circ}$

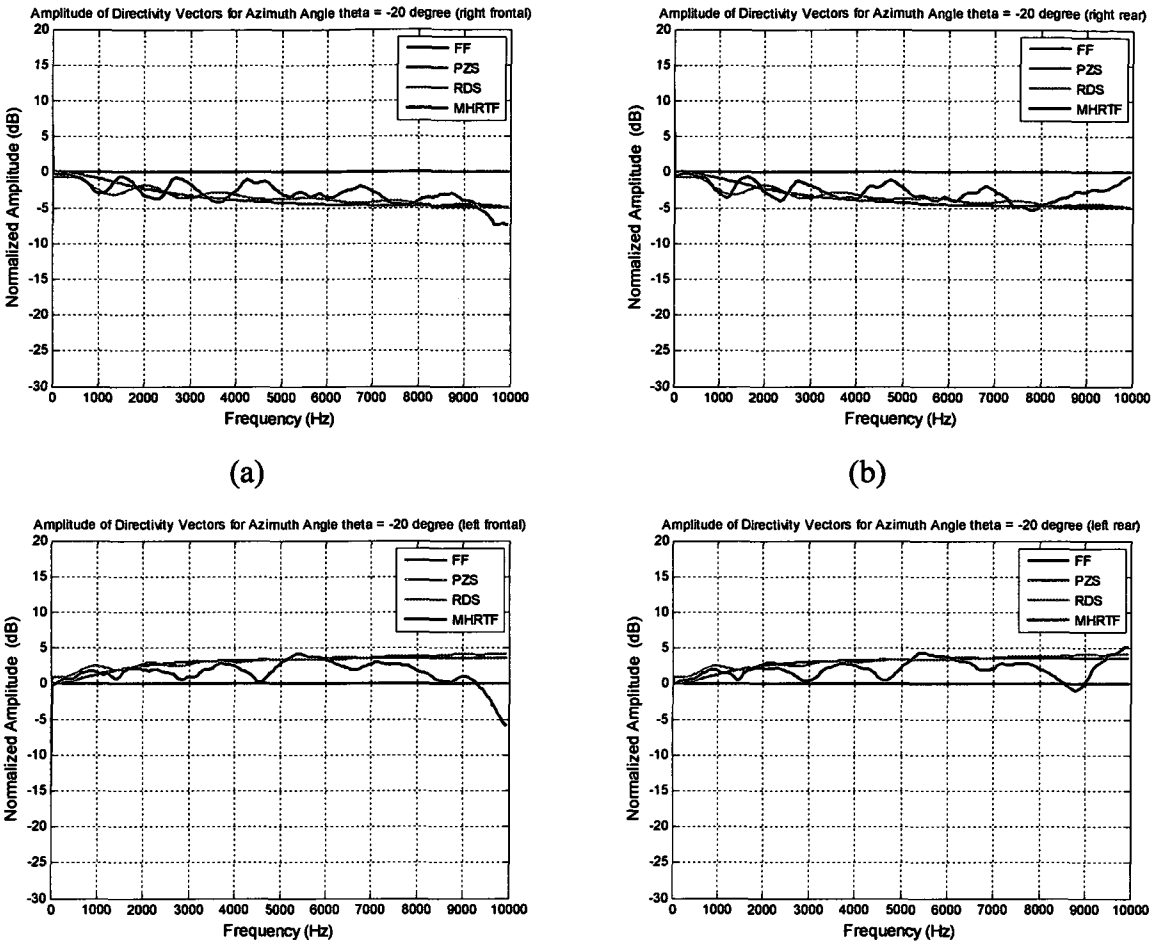


Figure 3.8 Magnitude of normalized directivity vectors for $\theta = -20^{\circ}$

From Figure 3.7 and Figure 3.8, it can be verified that the FF model does not include any head shadow effects, and its normalized amplitude is always the constant 1.0 (0dB). It can also be seen that the PZS and RDS model responses are much smoother than the MHRTF. For $\theta = 60^{\circ}$, there is a significant difference at high frequencies in the response of the MHRTF model compared to the PZS and RDS models, for the microphones on the right side (Fig. 3.7 (a) and (b)), thus clearly the measurements and the analytical head models do not always produce similar responses.

Chapter 4 Conversion to a Common Binaural Gain

4.1 Why a Common Binaural Gain?

It was previously mentioned that bilateral hearing aids users can locate an arriving sound source better without using hearing aids than with hearing aids, and one of the reason form this is that the binaural cues are being distorted by the signal processing algorithms. Under some environments it can be quite important to preserve the binaural cues of both the target source and the directional interference sources, and a few papers have been published where the preservation of the binaural cues was considered. In [KLA'06],[KLA'07],[KLA'05] and [BOG'07], a cost function aiming to minimize the cues distortion while maximizing noise reduction was introduced, and a tradeoff between those two parts was achieved. There is a cost however in terms of the noise reduction that can be achieved with this approach, and these papers using the Multichannel Wiener Filter are mostly suitable for stationary noise environments with only a few directional interferences [KAM'09C]. They also require a VAD or a classifier to distinguish between the target and the interferences, which can become a problem in practice (e.g. for low SNRs or several interfering sources).

For a MVDR beamformer, it should be noted that regardless of the configuration of the microphone or sensor array (monaural, binaural, endfire, broadside, etc.), the output of the MVDR beamformer is a monaural output, minimizing for each frequency the output signal power with the constraint of unit gain for a given target source direction. That is to say, whether the sensor signal vector $\underline{X}$ represents the Monaural 2 case or the Binaural 1+1 or 2+2 cases, the output $Y(\omega) = \underline{W}^H(\omega)\underline{X}(\omega)$ remains a scalar, thus using classical MVDR beamforming obviously removes all the spatial cues information. For a binaural microphone array to be used in hearing aids, it is desirable to produce binaural outputs

rather than a monaural output, where the binaural cues of all sources (target and interferences) will be preserved.

One approach to achieve the above requirements is to convert the beamforming output to a common gain $G(\omega)$, which is to be applied to signals coming from both sides of the head. For instance, we can have a common gain $G(\omega)$ which is applied to the frontal microphone signal on each side; or a common gain $G(\omega)$ applied to the average signal from the different microphones on each side. The latter is typically not preferred since it may disturb the binaural cues to some degree and bring up some phase mismatch problem. On top of having a common gain $G(\omega)$ which guarantees preservation of the binaural cues, the gain $G(\omega)$ should be:

1) real valued, so that the phase response is zero, the group delay is zero, and no frequency dependant dispersion is introduced;

2) scaled so that the received target source binaural signals $\underline{X}_s(\omega)$ keep the same level when they are processed to become the enhanced binaural signals $\underline{\hat{X}}_s(\omega)$.

If a common gain is applied to the frontal microphone signal on each side, the resulting structure is shown in Figure 4.1 below:

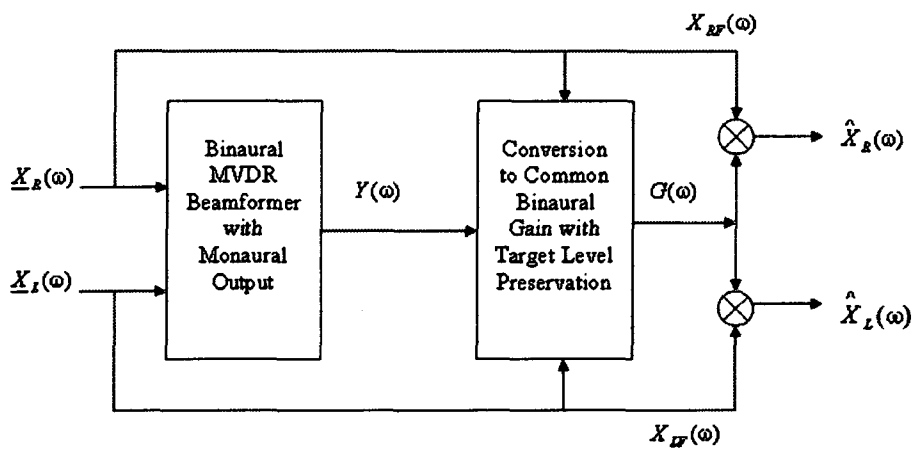


Figure 4.1 Using a common gain to preserve the binaural cues

The application of a common real valued gain to the input sensor signals $\underline{X}(\omega)$ from the front microphones of both sides can be written as:

$$\underline{\hat{X}}(\omega) = G(\omega)\underline{X}_F(\omega) \qquad (4.1)$$

with:

$$\underline{\hat{X}}(\omega) = [\hat{X}_L \hat{X}_R]^T$$

$$\underline{X}_F(\omega) = [X_{LF} X_{RF}]^T$$

By applying the common gain approach above, three types of binaural cues including the ITDs, ILDs and the onset time of speech (precedence effect) [HAR'99] are perfectly preserved. Also, ideally the common gain $G(\omega)$ is designed to preserve the level of the target source for each frequency, therefore for the target source it should not effect the monaural spatial spectral shaping from the pinnae [HAR'99], which is the 4th type of binaural cue. Head model mismatch and hearing aid position mismatch will affect this ideal behavior in practice though, up to some point. But in any case, it is clear that for the noise or interference components, the frequency dependent gain $G(\omega)$ will introduce some "coloring" of the spectrum, and thus the monaural spatial spectral shaping performed by the pinnae will be affected. Being able to preserve perfectly three out of the four binaural cues, and most of the fourth one for the target source, can achieve a satisfactory level of cues preservation, as found from informal listening of the sound files from Chapter 7. Note that this is no longer a "real" beamforming operation such as $Y(\omega) = \underline{W}^H(\omega)\underline{X}(\omega)$, because each output in $\underline{\hat{X}}(\omega)$ is produced directly from a single sensor input in $\underline{X}_F(\omega)$. Moreover, as we will see the gain $G(\omega)$ is time-varying even if the position of all sources remains fixed, as opposed to the beamforming weights which would remain constant in such a case. However, beamforming is still at the core of the system to generate the common gain. It should also be noted that this common gain approach in (4.1) becomes similar to gain-based speech enhancement methods (Wiener filtering, etc.), and therefore its time-varying behavior introduce similar types of distortion such as musical noise, which are typically not found in normal beamforming filtering.

4.2 Methods to Convert to a Common Gain

For simplicity, in this section we will use $\underline{X}(\omega)$ as opposed to $\underline{X}_F(\omega)$ to represent the two frontal microphones components, one from each side. There are several solutions for the gain $G(\omega)$ that meet the conditions mentioned in the previous section. It is natural for $G(\omega)$ to be proportional to the beamforming monaural output level $|Y(\omega)|$, because the level of this output is normally proportional to the level of the target input signal. But since $|Y(\omega)|$ in turn is also proportional to the level of all the components in the input sensor signals $\underline{X}(\omega) = \underline{X}_s(\omega) + \underline{X}_v(\omega)$, it is also natural to normalize the gain $G(\omega)$ by the level of those components (to avoid a square gain effect, because the sensor signals $\underline{X}(\omega)$ are also to be multiplied by the gain $G(\omega)$ to produce the enhanced signals $\underline{\hat{X}}(\omega)$ using (4.1)). Thus a sensible general form is:

$$G(\omega) = \frac{|Y(\omega)|}{f(\underline{X}(\omega))} \tag{4.2}$$

where $f(\underline{X}(\omega))$ is a real-valued function of the level of the sensor signals $\underline{X}(\omega)$. Note that some other forms could be $\sum_{k=L,R} \frac{|Y(\omega)|}{f(X_k(\omega))}$, $\prod_{k=L,R} \frac{|Y(\omega)|}{f(X_k(\omega))}$, $\max_{k=L,R} \frac{|Y(\omega)|}{f(X_k(\omega))}$, $\min_{k=L,R} \frac{|Y(\omega)|}{f(X_k(\omega))}$, etc..

If we recall our constraint that for the target source received signals $\underline{X}_s(\omega)$ the binaural outputs $\underline{\hat{X}}_s(\omega)$ must keep the same levels, the general form would then become:

$$G_s(\omega) = 1 = \frac{|Y_s(\omega)|}{f(\underline{X}_s(\omega))} \tag{4.3}$$

and therefore

$$\begin{aligned} |Y_s(\omega)| &= f(\underline{X}_s(\omega)) \\ \left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right||S(\omega,\theta_s)| &= f(\underline{X}_s(\omega)) \\ \left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right| &= \frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|} \end{aligned} \tag{4.4}$$

This shows that the gain $G(\omega)$ can be unity for the target source received sensor signals $\underline{X}_s(\omega)$ if the term $\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|$ has the general form $\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}$ (some more specific examples will be provided below). Recognizing that the MVDR design normally has the following constraint from equation (2.3): $\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)=1$, and thus:

$$\left|\underline{W}^H(\omega)\underline{D}(\omega,\theta_s)\right|=1 \qquad (4.5)$$

this means that to achieve the desired gain $G(\omega)$, the coefficients obtained by the "normal" MVDR design should be scaled by a scalar of the general form $\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}$ at each frequency. The specific choice of $\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}$ depends on the form chosen for the gain $G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}$.

To illustrate and clarify this, some specific examples for the gain $G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}$ and the corresponding term $\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}$ used to scale the "normal" beamformer MVDR weights will now be presented. In previous work in the literature such as in [LOT'06], the following specific normalization for $G(\omega)$ has been used:

$$G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}=\frac{|Y(\omega)|}{\|\underline{X}(\omega)\|_1}=\frac{|Y(\omega)|}{\sum_{k=L,R}|X_k(\omega)|}$$

$$\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}=\frac{\|\underline{X}_s(\omega)\|_1}{|S(\omega,\theta_s)|}=\frac{\sum_{k=L,R}|X_{s,k}(\omega)|}{|S(\omega,\theta_s)|}=\frac{\sum_{k=L,R}|D_k(\omega,\theta_s)||S(\omega,\theta_s)|}{|S(\omega,\theta_s)|}=\sum_{k=L,R}|D_k(\omega,\theta_s)|$$

(4.6)

Should the overall level of the experimental $\underline{D}(\omega,\theta_s)$ vector be different from the level of the vector $\underline{D}(\omega,\theta_s)$ used in the MVDR design, this would still lead to the same value of the ratio $G(\omega)=\dfrac{|Y(\omega)|}{f(\underline{X}(\omega))}$, as both the numerator and denominator terms would be increased. Consequently, for the target source received signals $\underline{X_s}(\omega)$, the binaural outputs $\underline{\hat{X}}_s(\omega)$ would still keep the same levels as $\underline{X_s}(\omega)$ (and the same phase too, since $G(\omega)$ is real-valued).

The general development of this section clearly shows that equation (4.6) is just a particular choice and there are several other choices possible for the normalization of $G(\omega)$. This has not appeared yet in the literature. Some of these choices could be:

$$G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}=\frac{|Y(\omega)|}{\|\underline{X}(\omega)\|_2}=\frac{|Y(\omega)|}{\sqrt{\sum_{k=L,R}|X_k(\omega)|^2}}$$

$$\frac{f(\underline{X_s}(\omega))}{|S(\omega,\theta_s)|}=\frac{\|\underline{X_s}(\omega)\|_2}{|S(\omega,\theta_s)|}=\frac{\sqrt{\sum_{k=L,R}|X_{s,k}(\omega)|^2}}{|S(\omega,\theta_s)|}=\frac{\sqrt{\sum_{k=L,R}|D_k(\omega,\theta_s)|^2}|S(\omega,\theta_s)|}{|S(\omega,\theta_s)|}=\sqrt{\sum_{k=L,R}|D_k(\omega,\theta_s)|^2}$$

(4.7)

$$G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}=\frac{|Y(\omega)|}{\sqrt{\prod_{k=L,R}|X_k(\omega)|}}$$

$$\frac{f(\underline{X_s}(\omega))}{|S(\omega,\theta_s)|}=\frac{\sqrt{\prod_{k=L,R}|X_{s,k}(\omega)|}}{|S(\omega,\theta_s)|}=\frac{\sqrt{\prod_{k=L,R}|D_k(\omega,\theta_s)|}|S(\omega,\theta_s)|}{|S(\omega,\theta_s)|}=\sqrt{\prod_{k=L,R}|D_k(\omega,\theta_s)|}$$

(4.8)

$$G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}=\frac{|Y(\omega)|}{\|\underline{X}(\omega)\|_{\infty}}=\frac{|Y(\omega)|}{\max_{k=L,R}|X_k(\omega)|}$$

$$\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}=\frac{\|\underline{X}_s(\omega)\|_{\infty}}{|S(\omega,\theta_s)|}=\frac{\max_{k=L,R}|X_{s,k}(\omega)|}{|S(\omega,\theta_s)|}=\frac{\max_{k=L,R}|D_k(\omega,\theta_s)||S(\omega,\theta_s)|}{|S(\omega,\theta_s)|}=\max_{k=L,R}|D_k(\omega,\theta_s)|$$

(4.9)

$$G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}=\frac{|Y(\omega)|}{\min_{k=L,R}|X_k(\omega)|}$$

$$\frac{f(\underline{X}_s(\omega))}{|S(\omega,\theta_s)|}=\frac{\min_{k=L,R}|X_{s,k}(\omega)|}{|S(\omega,\theta_s)|}=\frac{\min_{k=L,R}|D_k(\omega,\theta_s)||S(\omega,\theta_s)|}{|S(\omega,\theta_s)|}=\min_{k=L,R}|D_k(\omega,\theta_s)|$$

(4.10)

We note that there are several possible solutions, leading to different noise reduction versus target distortion tradeoffs (more on that in the next section). Recall also that the general form $G(\omega)=\frac{|Y(\omega)|}{f(\underline{X}(\omega))}$ is not the only possible one (other forms could be $\sum_{k=L,R}\frac{|Y(\omega)|}{f(X_k(\omega))}$, $\prod_{k=L,R}\frac{|Y(\omega)|}{f(X_k(\omega))}$, $\max_{k=L,R}\frac{|Y(\omega)|}{f(X_k(\omega))}$, $\min_{k=L,R}\frac{|Y(\omega)|}{f(X_k(\omega))}$, etc.), and thus other gains $G(\omega)$ could be developed in a similar way for those forms.

4.3 Methods to Combine Gains and Algorithms

The previous section presented the case of a binaural MVDR beamformer making use of sensor inputs from both sides (thus the name binaural), generating a monaural output $Y(\omega)$, converting to a common gain $G(\omega)$, which in turn generates binaural outputs with spatial cues preservation. Similarly, binaural outputs with spatial cues preservation could also be obtained using monaural noise reduction algorithms on the signals from different sensors. For example, this could be separate monaural MVDR beamformers operating independently on each side, or one-channel speech enhancement algorithms (e.g. spectral

subtraction, Wiener filtering) being applied separately on both sides. To combine different monaural noise reduction components from the left and right side, it is convenient for each noise reduction component to produce a gain $G(\omega)$ as its output, so that the different gains can then be combined into a global common gain, preserving the binaural cues.

For example, if monaural noise reduction algorithms are applied on each side of a Binaural 1+1 configuration, the right and left monaural noise reduction algorithms would produce the gains $G_R(\omega)$ and $G_L(\omega)$, and to obtain a common binaural gain (to be used as in (4.1)) the following four methods (and several others) could be used:

$$G_{\max}(\omega) = \max(G_R(\omega), G_L(\omega)) \tag{4.11}$$

$$G_{\min}(\omega) = \min(G_R(\omega), G_L(\omega)) \tag{4.12}$$

$$G_{sqr}(\omega) = \sqrt{G_R(\omega) G_L(\omega)} \tag{4.13}$$

$$G_{ave}(\omega) = average(G_R(\omega), G_L(\omega)) \tag{4.14}$$

Among the four methods, in term of noise reduction (4.11) is the most conservative one, (4.12) is the most aggressive one, and (4.13), (4.14) are in between. (4.13) and (4.14) can behave quite similarly when the values of $G_R(\omega)$ and $G_L(\omega)$ are close to each other. For instance, if $G_R(\omega) = 0.5$ and $G_L(\omega) = 0.5$, then $G_{sqr}(\omega) = \sqrt{G_R(\omega) G_L(\omega)} = 0.5$ and $G_{ave}(\omega) = average(0.5, 0.5) = 0.5$. However, by contrast, when there is a big difference between the values of $G_R(\omega)$ and $G_L(\omega)$, these two results will be more different. Considering $G_R(\omega) = 0.1$ and $G_L(\omega) = 0.9$, then $G_{sqr}(\omega) = \sqrt{G_R(\omega) G_L(\omega)} = 0.3$, and $G_{ave}(\omega) = average(0.1, 0.9) = 0.5$. Thus (4.13) is more aggressive than (4.14) in terms of noise reduction.

Another option to control the aggressiveness of the common gain method is to apply a "spectral floor" to the gain, in order to keep the gain above a threshold. This allows to limit the level of noise reduction and avoids producing too much distortion such as musical noise. It will be further discussed in the next chapter. It should also be noted that the final common gain G_{final} may need to be upper- limited by 1.0 as follows:

$$G_{final}(\omega) = \min(G(\omega), 1) \qquad (4.15).$$

Chapter 5 Combination of MVDR Beamformer with Other Algorithms

5.1 Wiener Post-Filter

In addition to a fixed binaural MVDR beamformer tuned for diffuse noise using equation (2.5) and a conversion to a common gain using equation (4.6), a Wiener post-filter was also proposed in [LOT'06] to produce enhanced signals optimal in the MMSE sense. In this case the overall common gain can be written as the multiplication of two common gains:

$$G_{Wiener}(\omega) = G_{pre}(\omega)G_{post}(\omega) \tag{5.1}$$

where $G_{pre}(\omega)$ is the common gain converted as in Chapter 4, and $G_{post}(\omega)$ is an additional Wiener post-filter common gain obtained according to:

$$G_{post}(\omega) = \frac{|Y(\omega)|^2}{\sum_{k=L,R}|X_k(\omega)|^2} \frac{\sum_{k=L,R}|D_k(\omega,\theta_s)|^2}{\left(\sum_{k=L,R}|D_k(\omega,\theta_s)|\right)^2} \tag{5.2}$$

As before, the outputs for the right and left ears are produced by multiplying the right-frontal and left-frontal sensor inputs with the overall common gain. The resulting system structure is shown in Figure 5.1.

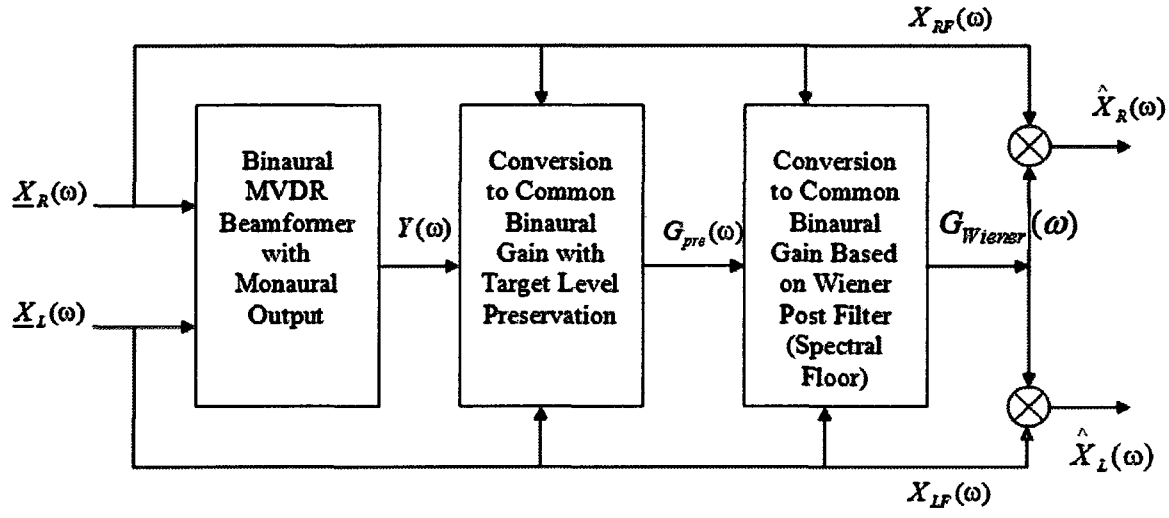


Figure 5.1 MVDR beamformer with common gain, followed by a Wiener post-filter

It should be noted that to better control the aggressiveness of the overall algorithm (and the tradeoff between noise reduction and speech distortion / musical noise), a limiting "spectral floor" is normally applied to the gain. If the spectral floor is represented by $g_{Wiener-sp}$ (value between 0 and 1), then the resulting common gain for the overall algorithm is obtained by:

$$G_{Wiener-sp}(\omega) = \max(G_{Wiener}(\omega), g_{Wiener-sp}) \qquad (5.3)$$

The spectral floor ensures that the signal attenuation will never be below $g_{Wiener-sp}$, and it can limit the level of signal distortion and musical noise introduced.

5.2 Minimum Mean Square Short-Time Spectral Amplitude Estimator (MMSE-STSA)

With the approach of using a common real-valued spectral gain to produce binaural outputs from a fixed binaural MVDR beamformer, the overall system is no longer a 'normal' beamformer, and these modifications can lead to musical noise or speech distortion, as previously mentioned. The Minimum Mean Square Short-Time Spectral Amplitude estimator (MMSE-STSA) proposed in [EPH'84] is a one-channel noise

reduction scheme known to produce low musical noise distortion [CAP'94]. In order to eliminate the musical noise generated in the previously described binaural MVDR beamformer with cues preserved, the MMSE-STSA can be considered as a post-filter following the binaural MVDR beamformer. The MMSE-STSA is a SNR-type amplitude estimator, and it is a monaural algorithm in its standard derivation which we use here. In this case, we need to process each of the channels (left and right) produced by the binaural MVDR beamformer with cues preservation. Detail about the MMSE-STSA can be found in [EPH'84], and a flow chart describing the implementation that we have used can be found in [KAM'09A]. A strength control or spectral floor can also be applied to this algorithm to make sure that the common gain will not drop below a minimum value, so that the noise reduction will not be too aggressive. Equation (5.4) shows how to incorporate the strength control $g_{mmse-sp}$:

$$G_{mmse-sp}(\omega) = \max(G_{mmse}(\omega), g_{mmse-sp}) \qquad (5.4)$$

where $G_{mmse}(\omega)$ denotes the combined gain derived from the MMSE-STSA algorithm. Combining the MMSE-STSA with the binaural MVDR beamformer and the Wiener post-filter from Figure 5.1, the resulting overall system is shown in Figure 5.2.

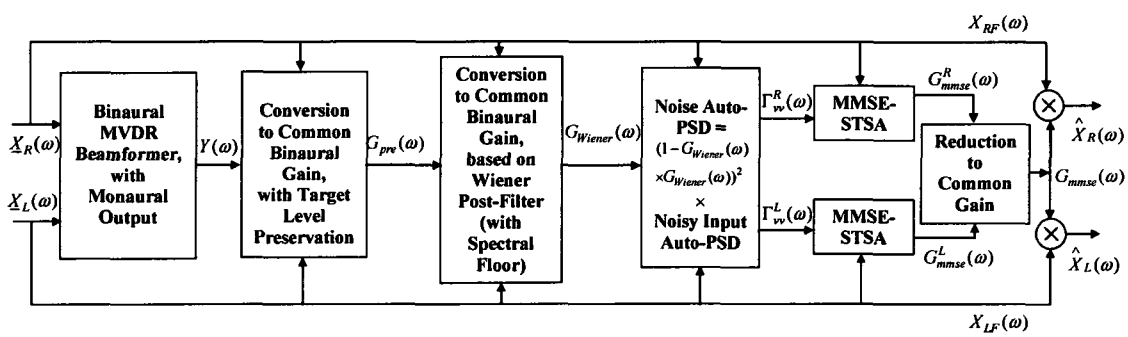


Figure 5.2 Binaural MVDR beamformer with Wiener post-filter, followed by MMSE-STSA

In Figure 5.2, the noise auto-PSD estimation required by the MMSE-STSA algorithm is computed as follows:

$$\text{Noise auto-PSD} = (1 - G_{Wiener}(\omega)G_{Wiener}(\omega))^2 \times \text{Noisy Input auto-PSD} \qquad (5.5)$$

Some details of the MMSE-STSA procedure are provided below in Figure 5.3, with the parameters chosen as $q = 0.2$, $\sigma = 0.98$ and $NFFT = 512$ in our implementation:

FOR j = RF:RB:LF: LB

Step 1— $V_j(i,\omega) = \sqrt{\mathrm{T}_{vv}^{j}(i,\omega) NFFT}$

Step 2— $\xi_j(i,\omega) = \dfrac{\left|X_j(i,\omega)\right|^2}{\left|V_j(i,\omega)\right|^2} - 1$

Step 3— $\gamma_j(i,\omega) = (1-\sigma)P\left[\xi_j(i,\omega)\right] + \sigma \dfrac{\left|G_{mmse}^{j}(i-1,\omega)X_j(i-1,\omega)\right|^2}{\left|V_j(i,\omega)\right|^2}$

Step 4— $\vartheta = \left(1+\xi_j(i,\omega)\right)\left(\dfrac{\gamma_j(i,\omega)}{1+\gamma_j(i,\omega)}\right)$

Step 5— $M[\vartheta] = e^{-\left(\frac{\vartheta}{2}\right)}\left((1+\vartheta)I_0\left(\dfrac{\vartheta}{2}\right) + \vartheta I_1\left(\dfrac{\vartheta}{2}\right)\right)$

Step 6— $\hat{\gamma}_j(i,\omega) = (1-q)\gamma_j(i,\omega)$

Step 7— $G_{mmse}^{j}(i,\omega) = \dfrac{\sqrt{\pi}}{2}\sqrt{\left(\dfrac{1}{1+\xi_j(i,\omega)}\right)\left(\dfrac{\hat{\gamma}_j(i,\omega)}{1+\hat{\gamma}_j(i,\omega)}\right)} M[\vartheta]$

Step 8— $\Lambda = \left(\dfrac{1-q}{q}\right)\left(\dfrac{1}{1+\hat{\gamma}_j(i,\omega)}\right) e^{\left(\frac{\hat{\gamma}_j(i,\omega)}{1+\hat{\gamma}_j(i,\omega)}\right)\cdot\left(1+\xi_j(i,\omega)\right)}$

Step 9— $G_{mmse}^{j}(i,\omega) = \dfrac{\Lambda}{1+\Lambda} G_{mmse}^{j}(i,\omega)$

Step 10— $i = i+1$

END

Figure 5.3 Implementation of MMSE-STSA

5.3 Binaural Target Estimator – Noise Reduction (PBTE-NR)

Since in this thesis we mostly use the fixed MVDR beamformer in Chapter 2 tuned for diffuse noise, it is expected that it will perform quite well under diffuse noise environments or reverberant environments including diffuseness. However, in complex environments where both diffuse noise and directional interferences coexist, we may consider adding some other post-processing algorithms which are more targeting directional noise reduction. The post-processing would then operate on the binaural output signals with cues preserved produced from the modified beamforming stage. An advanced noise reduction scheme called Proposed Binaural Target Estimator – Noise Reduction (PBTE-NR, or PBNR) was recently presented in [KAM'09A], and it is suitable for noise reduction in complex acoustic environments made of time-varying diffuse noise, multiple transient and non-stationary directional interferences, and reverberation. The overall structure in [KAM'09A] makes use of the recently proposed binaural estimators in [KAM'09B] and [KAM'09C]. The PBNR does not require the knowledge of the direction of the interfering sources, and it does not rely on any voice activity detection (VAD) either. It preserves the binaural cues by using the same common gain approach as the one described in Chapter 4. The performance of combining fixed binaural beamforming with PBNR post-processing will thus be evaluated in this thesis.

5.4 Combination of Monaural Beamformers and Monaural or Binaural Enhancement with Common Gain

In the previous sections, the use of a binaural MVDR beamformer with a common gain being applied to signals from both sides was described. As an alternative, below we first present a structure where monaural MVDR beamforming is performed separately on each side, followed by a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator (MMSE-STSA) monaural speech enhancement algorithm on each side. The resulting gains from each side are combined to produce a global binaural common gain,

applied to the noisy speech signals on each side to produce the enhanced signals. This structure is shown in Figure 5.4:

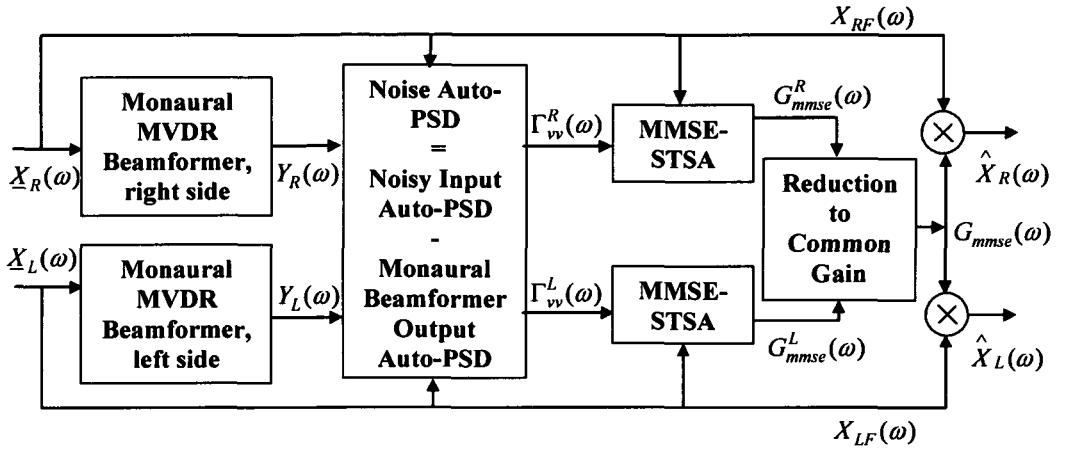


Figure 5.4 Combination of monaural beamformers with common gain

The noise auto-PSD estimations required for the MMSE-STSA algorithms in Figure 5.4 are now obtained by:

Noise auto-PSD = (Noisy Input auto-PSD) – (Monaural Output auto-PSD) (5.6)

where the monaural output of the Monaural 2 MVDR beamformers on each side is used to estimate the clean target speech input component on each side.

The binaural output signals from Figure 5.4 can then be used as either the final outputs, or they can become the input of a post-processing by the binaural PBNR algorithm, as described in the previous section. Alternatively, another structure which will be investigated in this thesis and which also involves monaural beamformers combined with the PBNR algorithm is the following structure :

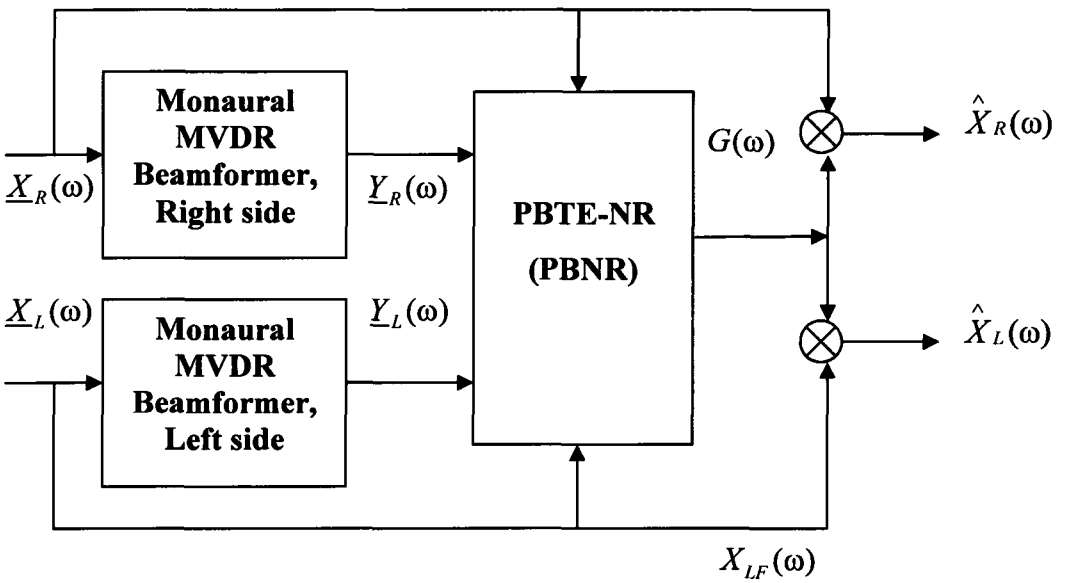


Figure 5.5 Combination of monaural beamformers with PBNR, without cues preservation

Unlike the other algorithms and structures presented in this chapter, the above structure does not guarantee the preservation of the binaural cues. Since the binaural PBNR algorithm assumes for some of its processing that the binaural input signals that it receives have the original spatial cues, this structure can be sub-optimal. However, since the binaural PBNR algorithm has a Binaural 1+1 configuration with limited capabilities for canceling sources from the back (please refer to Figure 6.1 in the next chapter for an illustration) and since the monaural MVDR beamformers are suitable for canceling sources coming from the back (again please refer to Figure 6.1), the combination of the two algorithms makes sense, as long as some cues distortion is allowed.

In order to preserve the cues when combining the monaural beamformers and the PBNR algorithm, another structure is proposed in Figure 5.6. It adds a common gain reduction step between the monaural beamformers and the PBNR. The system structures in Figure 5.5 and Figure 5.6 will be referred to as M2+PBNR and M2C+PBNR, respectively.

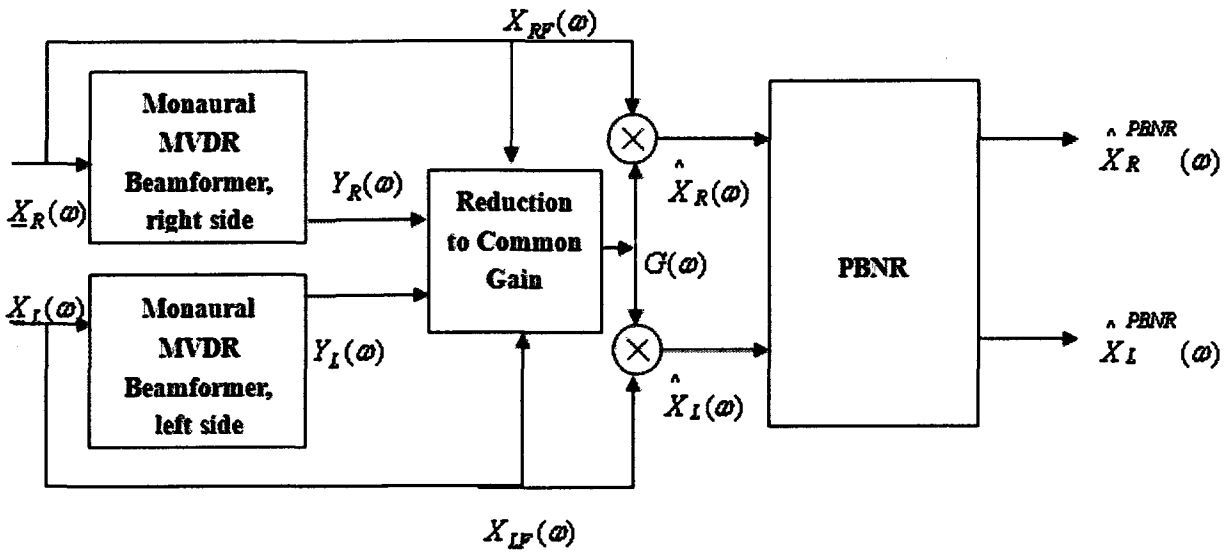


Figure 5.6 Combination of monaural beamformers with PBNR, with cues preservation

Chapter 6 Simulation Results with Classical Beamforming Performance Measures

6.1 Experimental Setup

The objective of the first simulation results in Section 6.2 is to compare the performance of different array configurations for a certain head model. Throughout the chapter, the fixed MVDR beamformers will be designed for diffuse noise and using the normalized design of Section 2.1.4. Since the MHRTF head model (from the measured HRTFs) is the one closest to practical conditions, it is the one used in that section. Both frontal target and non-frontal target situations are considered, where frontal target means that the direction of the target source is at 0 degree in azimuth, i.e. $\theta_s = 0^\circ$ in Figure 2.1. By contrast, a non-frontal target is a target coming from an angle where $\theta_s \neq 0^\circ$. In practice when people are talking (especially when more than two people are talking together), they do not necessarily face exactly each other ($\theta_s = 0^\circ$), and an azimuth angle within the range of $-20^\circ \leq \theta_s \leq 20^\circ$ is a more reasonable assumption for a target speaker. In order to assess the performance of a beamformer when a source is coming from a different direction than the angle for which the MVDR beamformer was designed, Section 6.2.3 will show the results for the situation where the beamformer is designed for $\theta_s = 0^\circ$ but the source target speech is actually at $\theta_s = 20^\circ$.

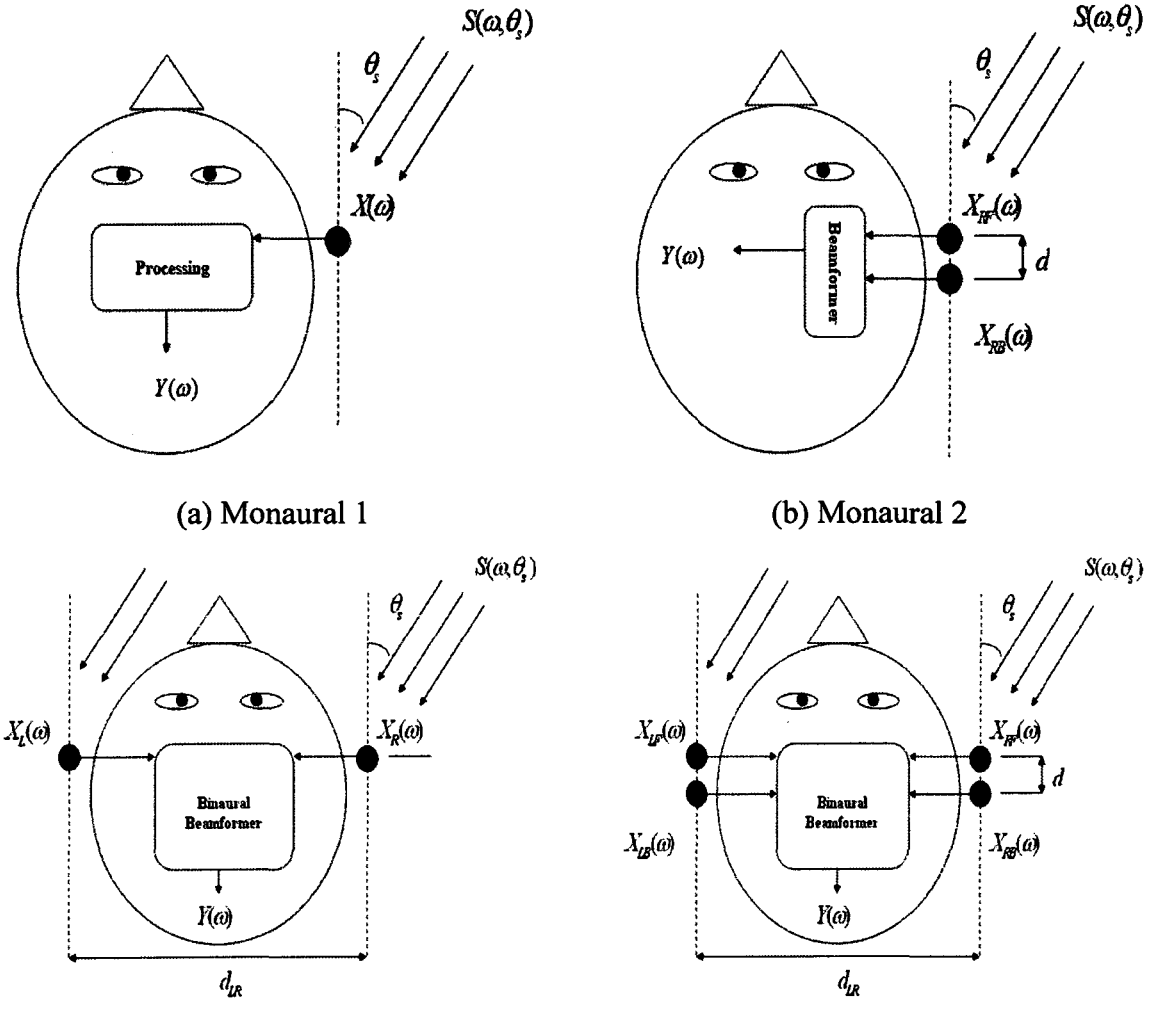
Section 6.3 will investigate the effect of head model mismatch in the design of fixed MVDR beamformers. MVDR beamformers will be designed using the HRTFs generated by each of the four head models. However the performance will always be evaluated by using the MHRTF model, to simulate real-life testing and to produce some model mismatch. As above, both frontal target and non-frontal targets will be considered.

All of the evaluations in this chapter are done using the classical measures discussed in Chapter 2 such as the Beampattern, the Array Gain and the Noise Gain. The model parameters are set as: intra-array distance $d = 0.75cm$; inter-array distance $d_{LR} = 17.5cm$ (radius of the head $a = 8.75cm$); sampling frequency $f_s = 20$ kHz. In the case of monaural beamformers, the results are only shown for the right ear (the results for the left ear would be symmetrical).

6.2 Comparison of Different Array Configurations

6.2.1 Frontal Target

Four different microphone array configurations are considered, as in Section 3.4:



(a) Monaural 1 (b) Monaural 2

(c) Binaural 1+1 (d) Binaural 2+2

Figure6.1 Four different microphone array configurations

For a frontal target, the corresponding classical measures of Beampattern, Array Gain and Noise Gain are as follows:

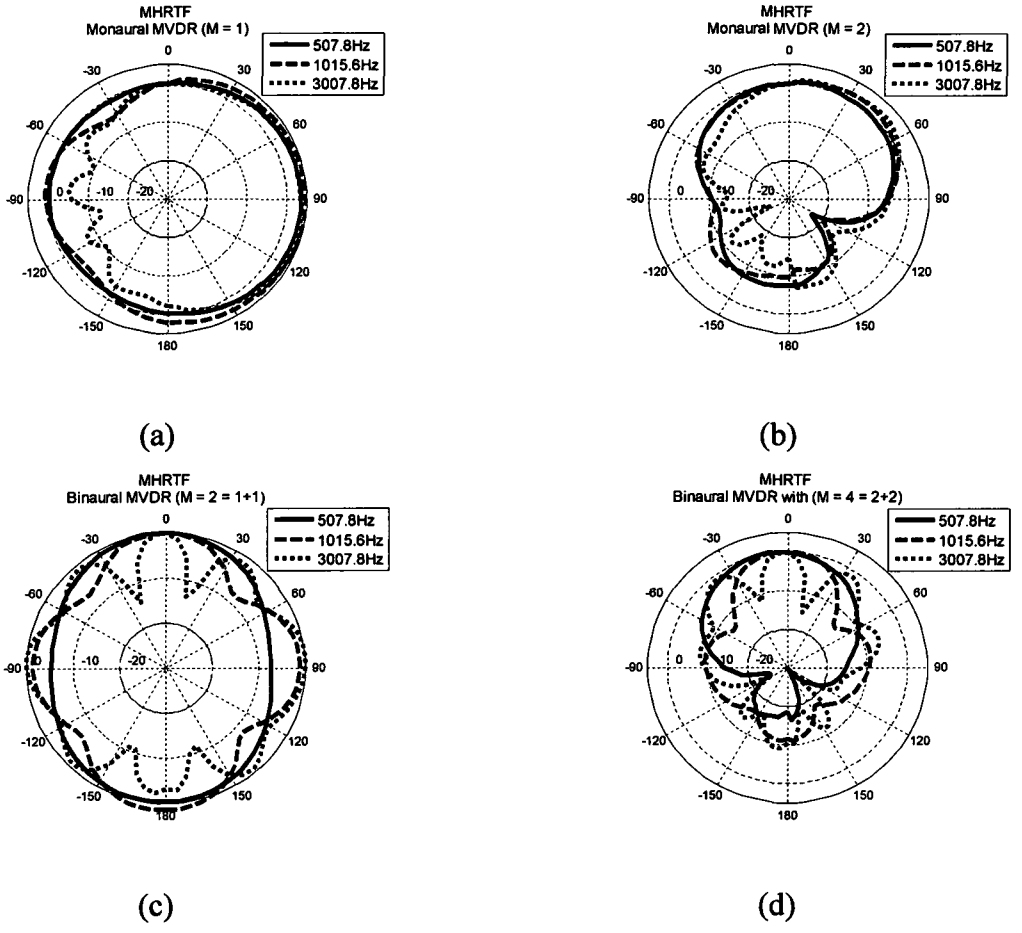
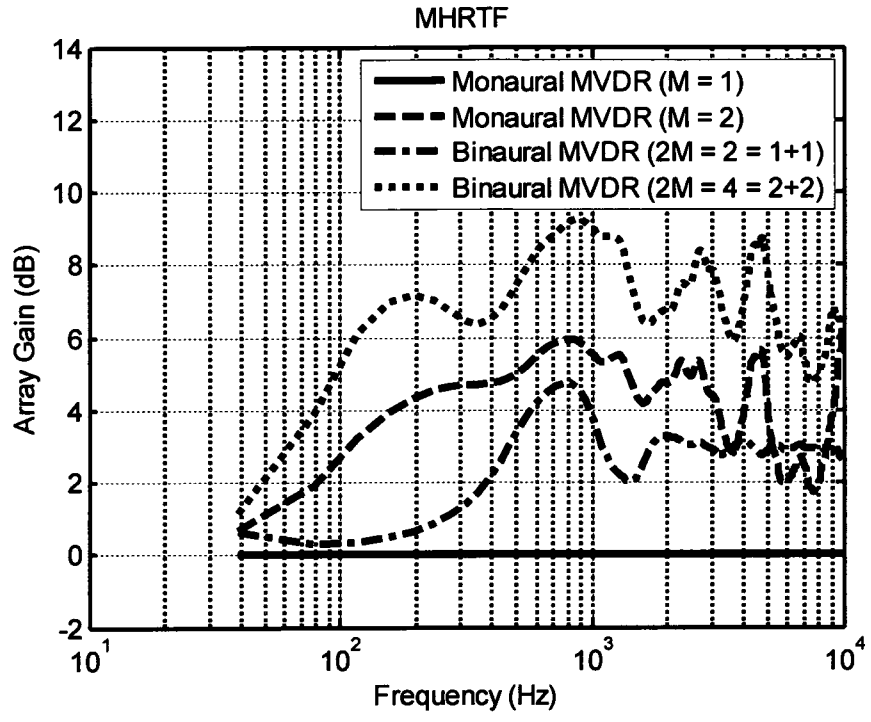
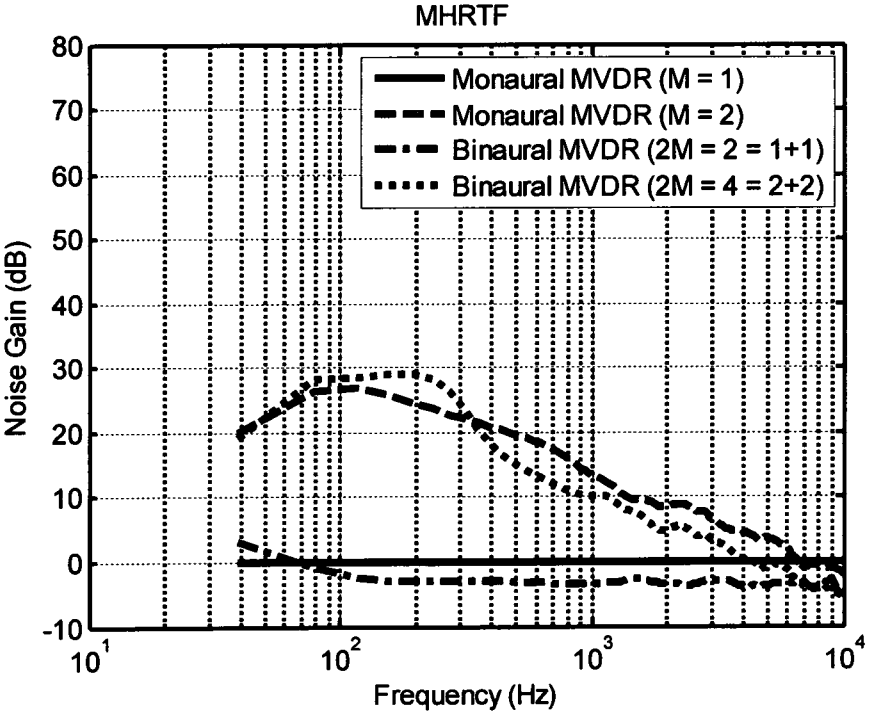


Figure 6.2 Beampatterns of different array configurations using MHRTF for $\theta_s = 0^\circ$



(a)



(b)

Figure 6.3 Array Gain and Noise Gain using MHRTF for $\theta_s = 0^\circ$

From Figure 6.1 (a) and (b), the head shadow effects can be easily perceived, since for these two monaural microphone array configurations the results are shown for the right side array, and the gains for the left azimuth plane where $-180^\circ < \theta < 0^\circ$ are smaller than for the right plane. The corresponding results for the left side array would be symmetrical about the $\theta = 0^\circ$ axis. The Monaural 1 configuration in Figure 6.1 (a) actually shows no beamforming but only head shadow effects. A front-back symmetry can be observed in the results for the Binaural 1+1 microphone array in Figure 6.1 (c), and it can be observed visually that the best beampattern is the one for the Binaural 2+2 microphone array in Figure 6.1 (d), as expected given the increased number of microphones.

The Array Gains in Figure 6.2 (a) compare the performance of different microphone array configurations for a frontal target. First, the effect of merging monaural arrays from both sides into a larger binaural array can be observed. The Binaural 1+1 configuration produces around 3 dB of increase over the Monaural 1 case (i.e. no beamforming), and the Binaural 2+2 also produces about 3 dB of improvement over the Monaural 2. Then, comparing the binaural beamforming configurations, we can also observe the benefit of moving from a Binaural 1+1 scenario to a Binaural 2+2 scenario (approximately 3-4 dB of improvement). Finally, for the same total number of microphones, we can also compare the orientation of the array, endfire and broadside, although this is not likely a choice that a designer would need to make in practice. For a frontal target, the Monaural 2 case is in a endfire configuration, while the binaural Binaural 1+1 case is in a broadside configuration. It is well known in the field of beamforming that for a linear array the endfire configuration is favorable compared to the broadside one, and this can be observed here by the superior performance of the Monaural 2 case over the Binaural 1+1 case (approximately 2-3 dB) for this particular setup.

The noise gain results in Figure 6.2 (b) show that for a frontal target the binaural configurations have typically a lower noise gain than their monaural counterparts (often a 3 dB difference). It also shows that the broadside configuration (Binaural 1+1 for this target direction) has a particularly low noise gain.

6.2.2 Non-Frontal Target

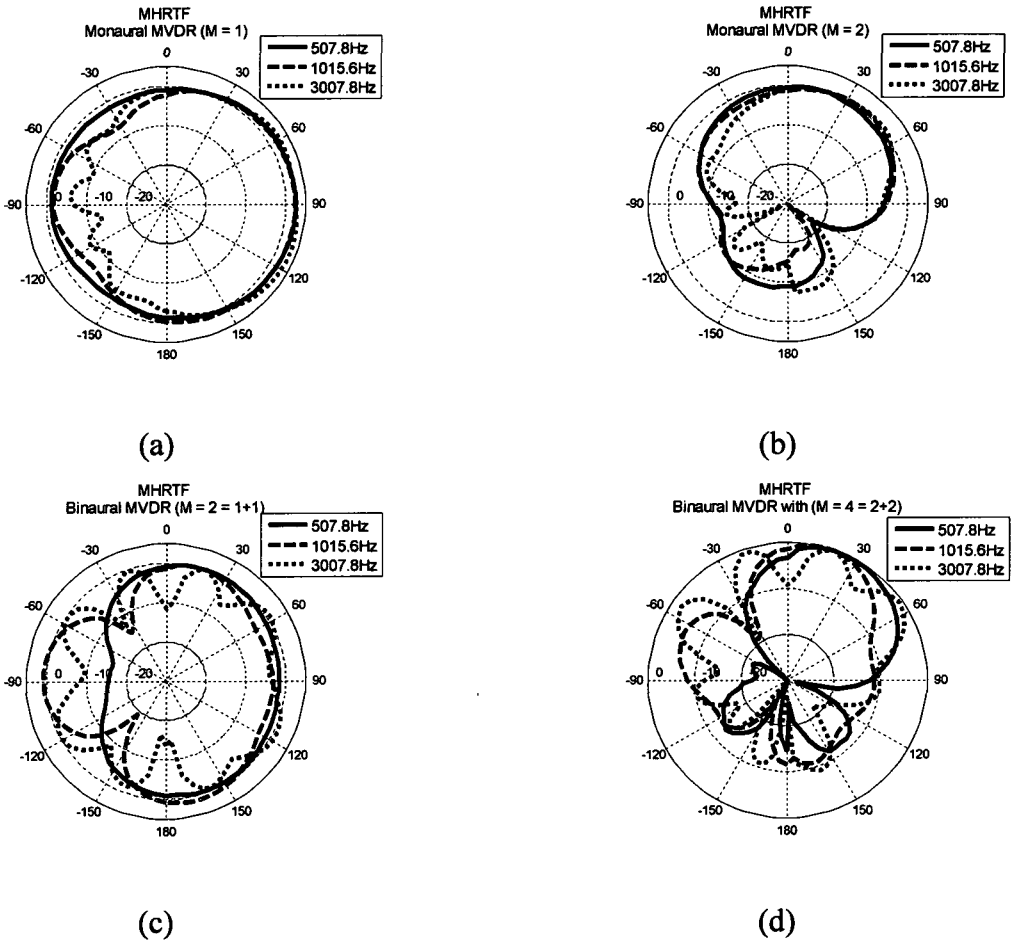
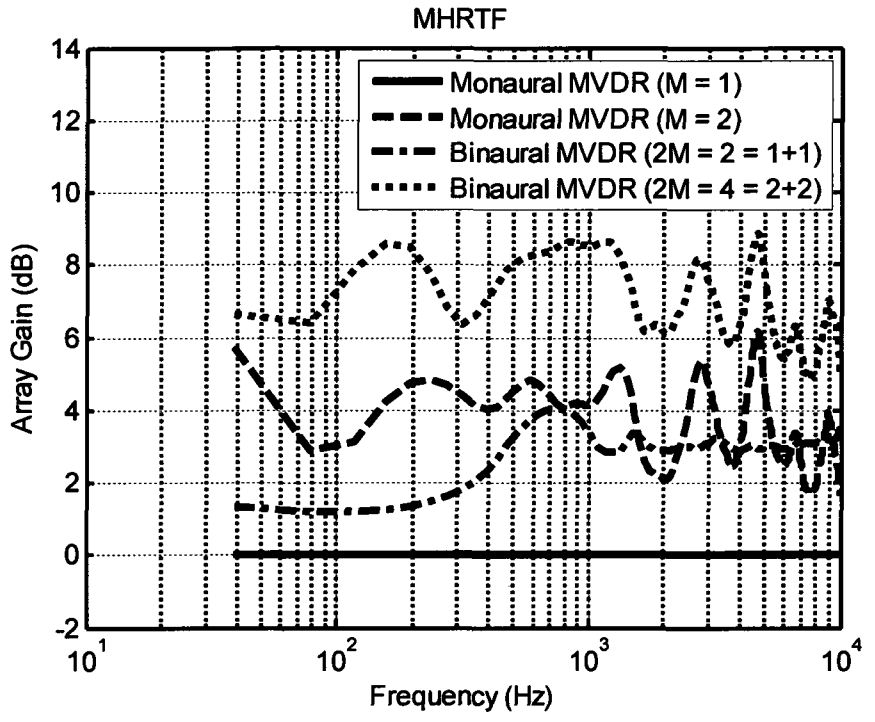
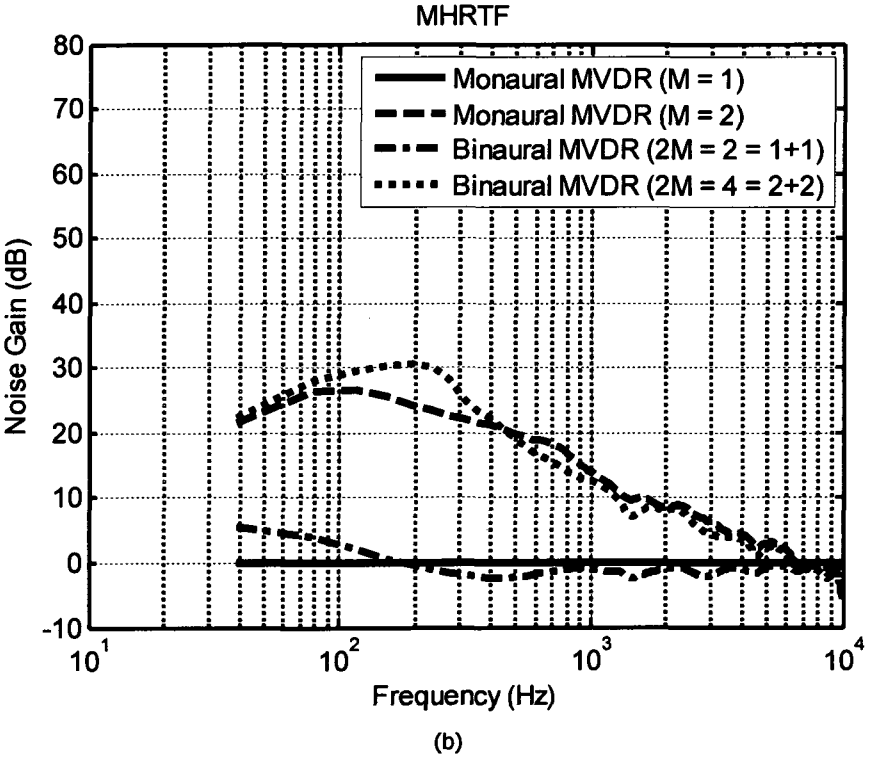


Figure 6.4 Beampatterns of different array configurations using MHRTF for $\theta_s = 20^\circ$

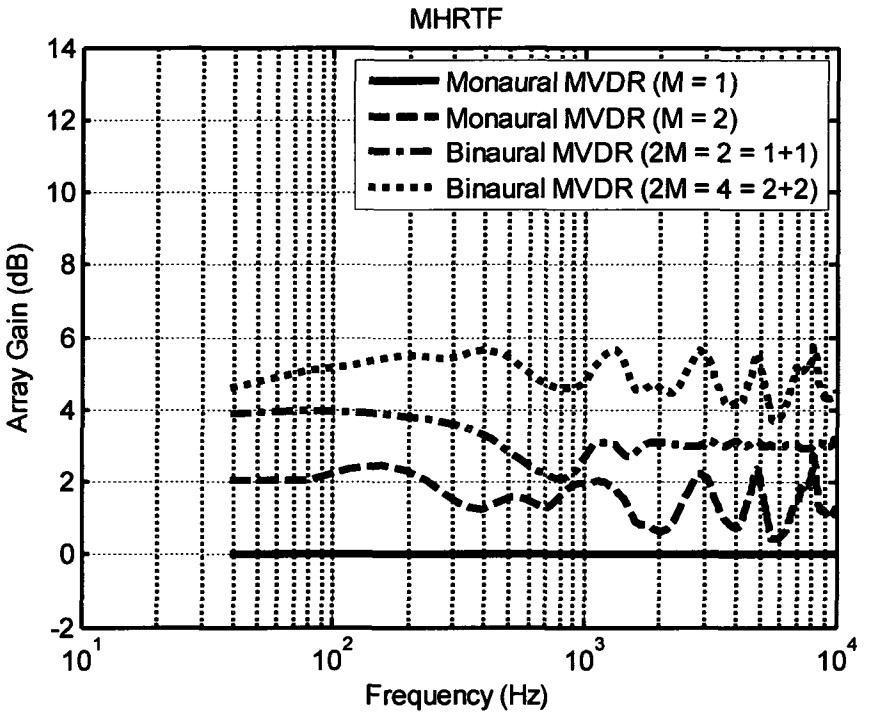


(a)

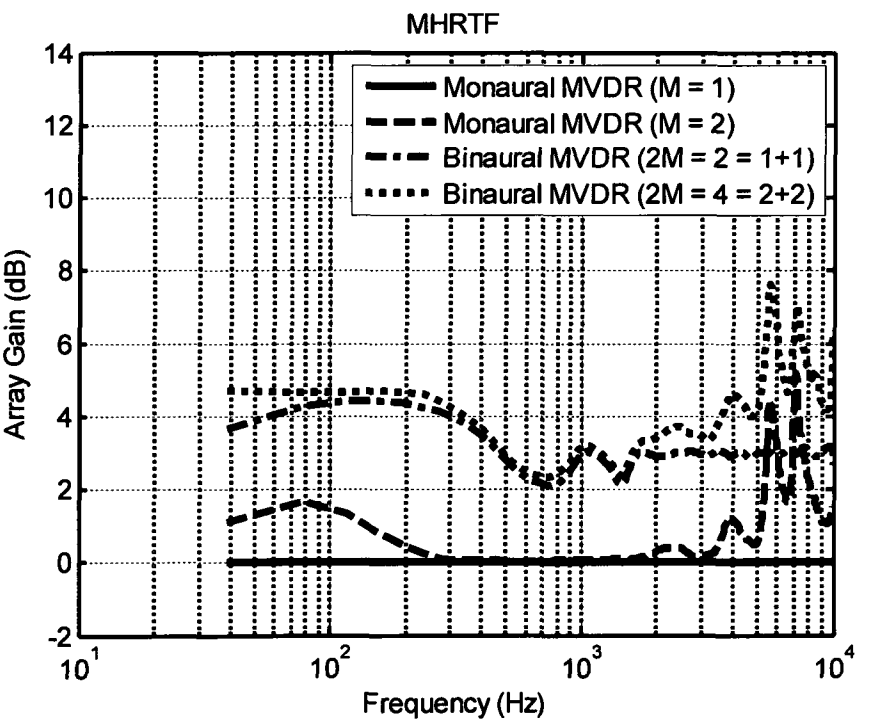


(b)

Figure 6.5 Array Gain and Noise Gain using MHRTF for $\theta_s = 20^\circ$



(a) $\theta_s = 60^\circ$



(b)

Figure 6.6 Array Gain using MHRTF for $\theta_s = 60^\circ$ and $\theta_s = 90^\circ$

From the Array Gain results of Figures 6.3 to 6.4, we can see that for a non-frontal target at $\theta_s = 20^\circ$ the binaural configurations still produce approximately 3dB of improvement over their respective monaural counterparts, and that the Binaural 2+2 configuration still produces approximately 3-4dB of improvement over the Binaural 1+1 configuration. For the Noise Gain results in Figure 6.4 (b), the binaural configurations no longer have a lower Noise Gain than their respective monaural counterparts, and they now have a more similar Noise Gain over most frequencies.

For the Array Gain results with a target at $\theta_s = 60^\circ$ and $\theta_s = 90^\circ$ in Figure 6.5, the binaural configurations still produce approximately 3dB of improvement over their respective monaural counterparts, and the Binaural 2+2 configuration produces approximately 1-2 dB of improvement over the Binaural 1+1 configuration. This reduction of the improvement for the Binaural 2+2 over the Binaural 1+1 is because in this case the Binaural 1+1 array is in an endfire configuration for $\theta_s = 90^\circ$, and this also explains why in this case the Binaural 1+1 starts to outperform the Monaural 2 case (which becomes a broadside configuration for $\theta_s = 90^\circ$).

6.2.3 DOA Mismatch

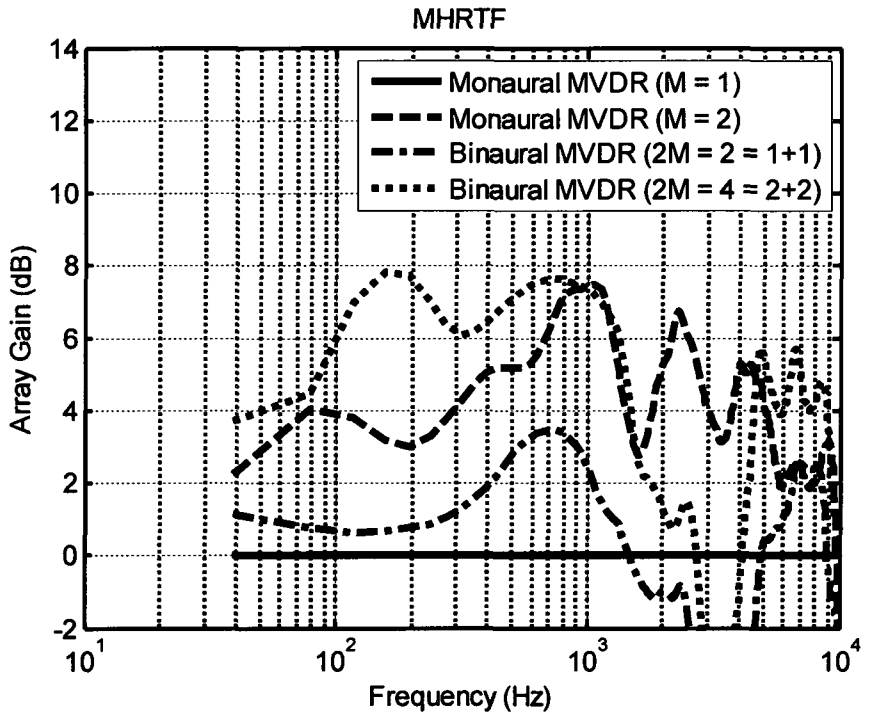


Figure 6.7 Array Gain and Noise Gain for case of DOA mismatch, using MHRTF

Figure 6.6 shows the Array Gain results when the MVDR beamformer is designed for $\theta_s = 0^\circ$ while the Array Gain performance measure is evaluated for a target at $\theta_s = 20^\circ$. Thus this is a significant DOA mismatch of 20 degrees in this case, and we can observe a drop in the performance of the Binaural 1+1 and Binaural 2+2 beamformers over the range of frequencies $2000\text{Hz} < f < 4000\text{Hz}$, although the overall Array Gain will remain positive. This stresses the importance of having relatively good DOA information about the target source. The performance of the Monaural 2 configuration was not affected as much in this particular scenario, although we will see later that on average (over different angles) it degrades similarly to the binaural cases.

6.3 Comparison of Different Head Models and Model Mismatch

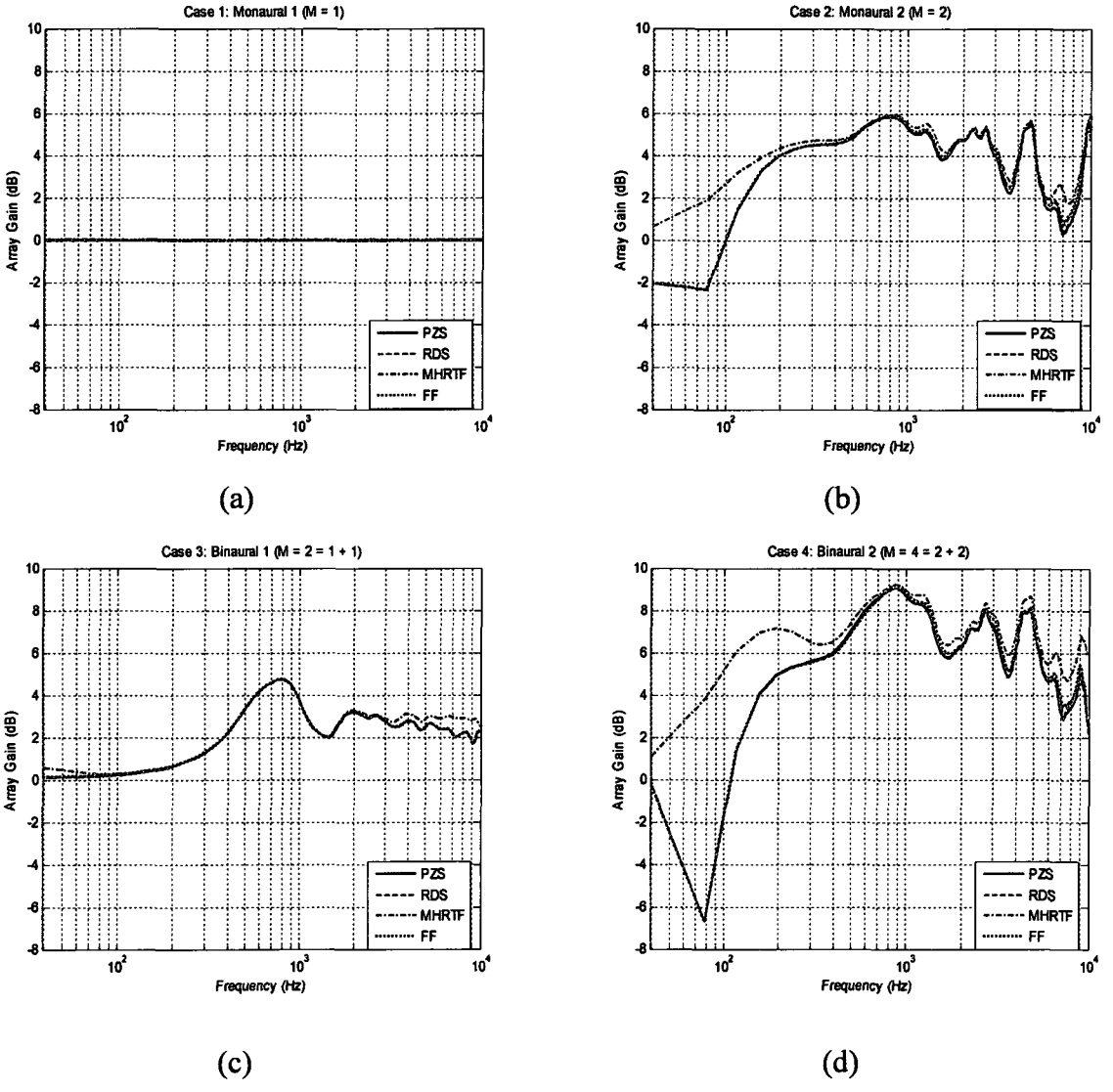


Figure 6.8 Array Gains of different head models for $\theta = 0^\circ$, with model mismatch

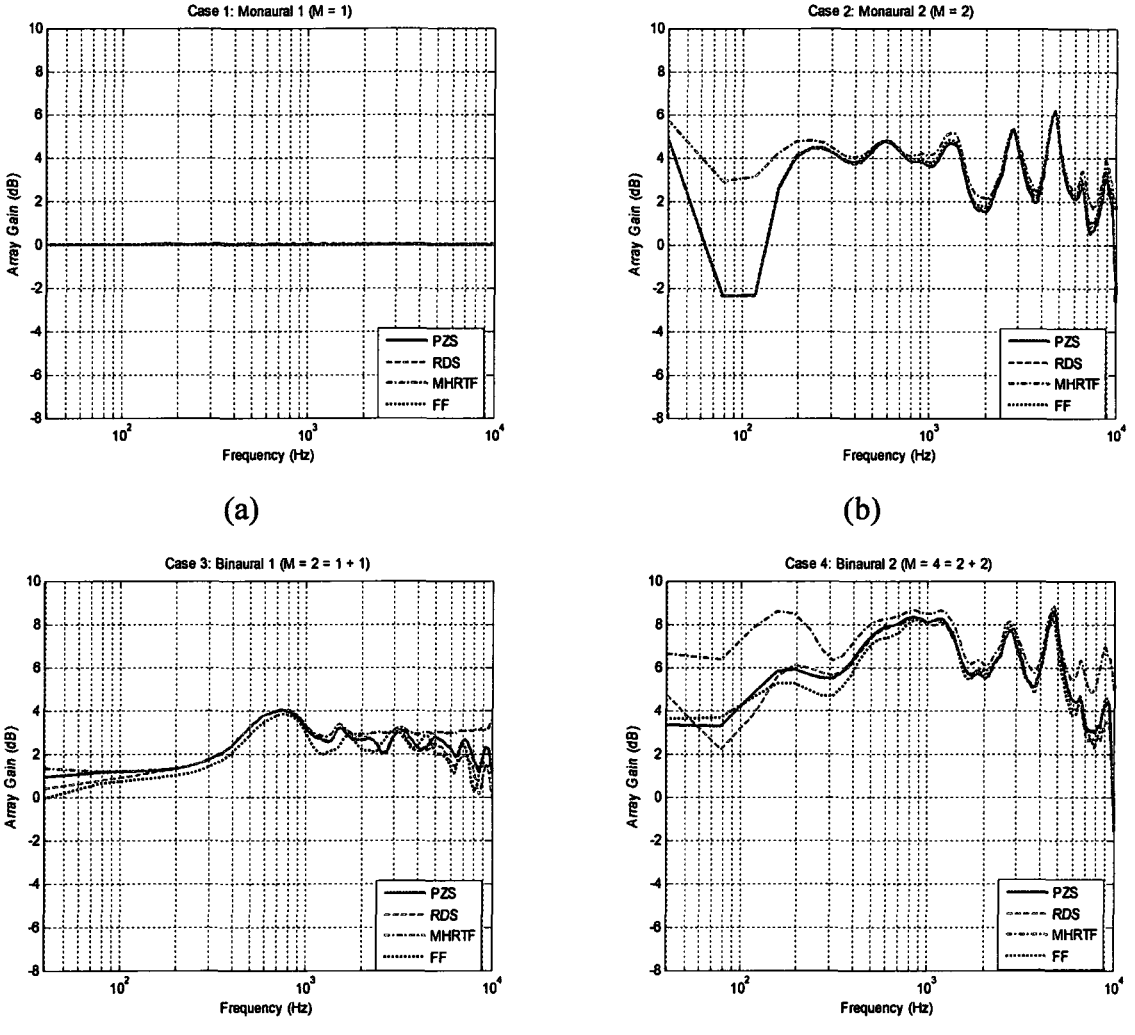


Figure 6.9 Array Gains of different head model for $\theta = 20^{\circ}$, with model mismatch

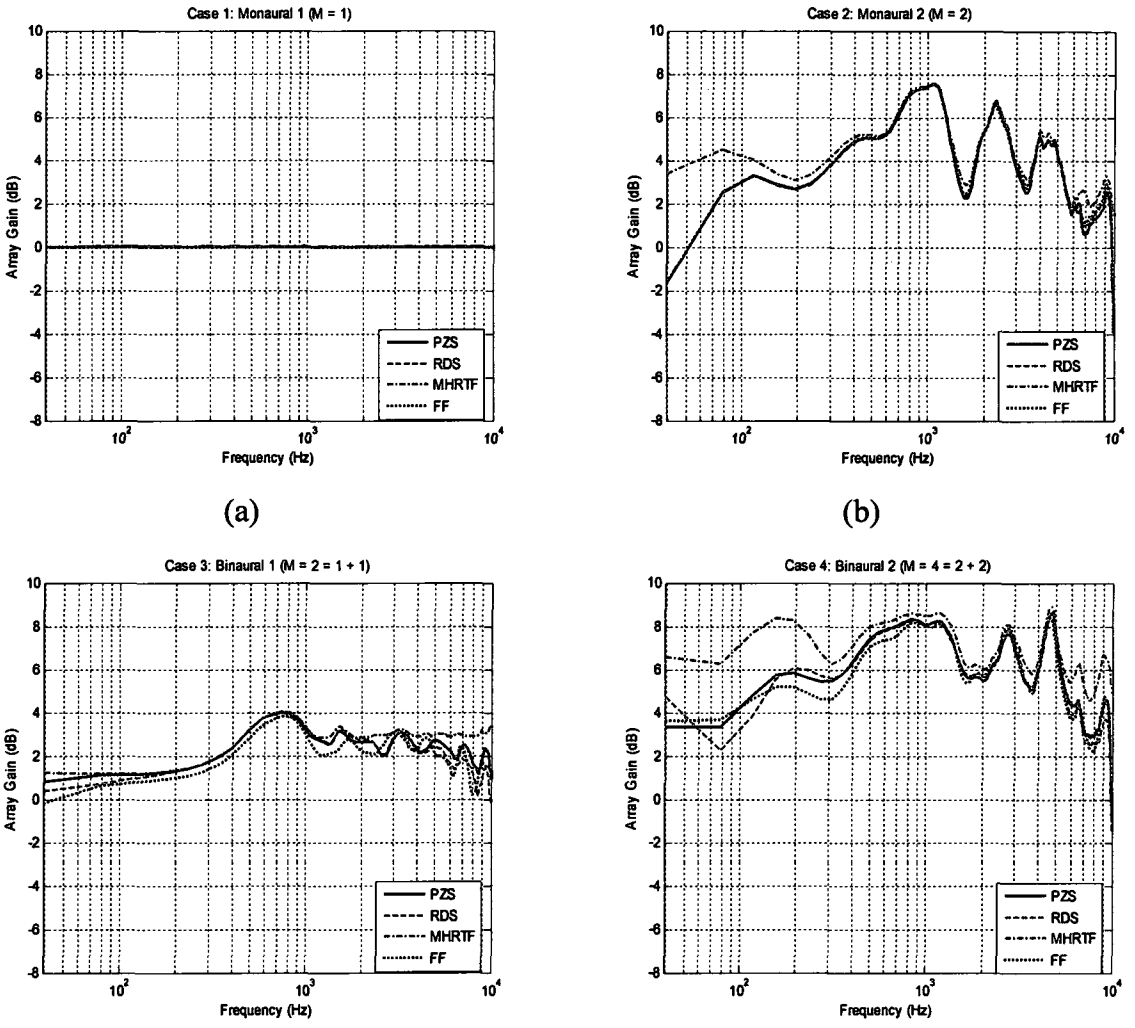


Figure 6.10 Array Gains of different head model for $\theta = -20^{\circ}$, with model mismatch

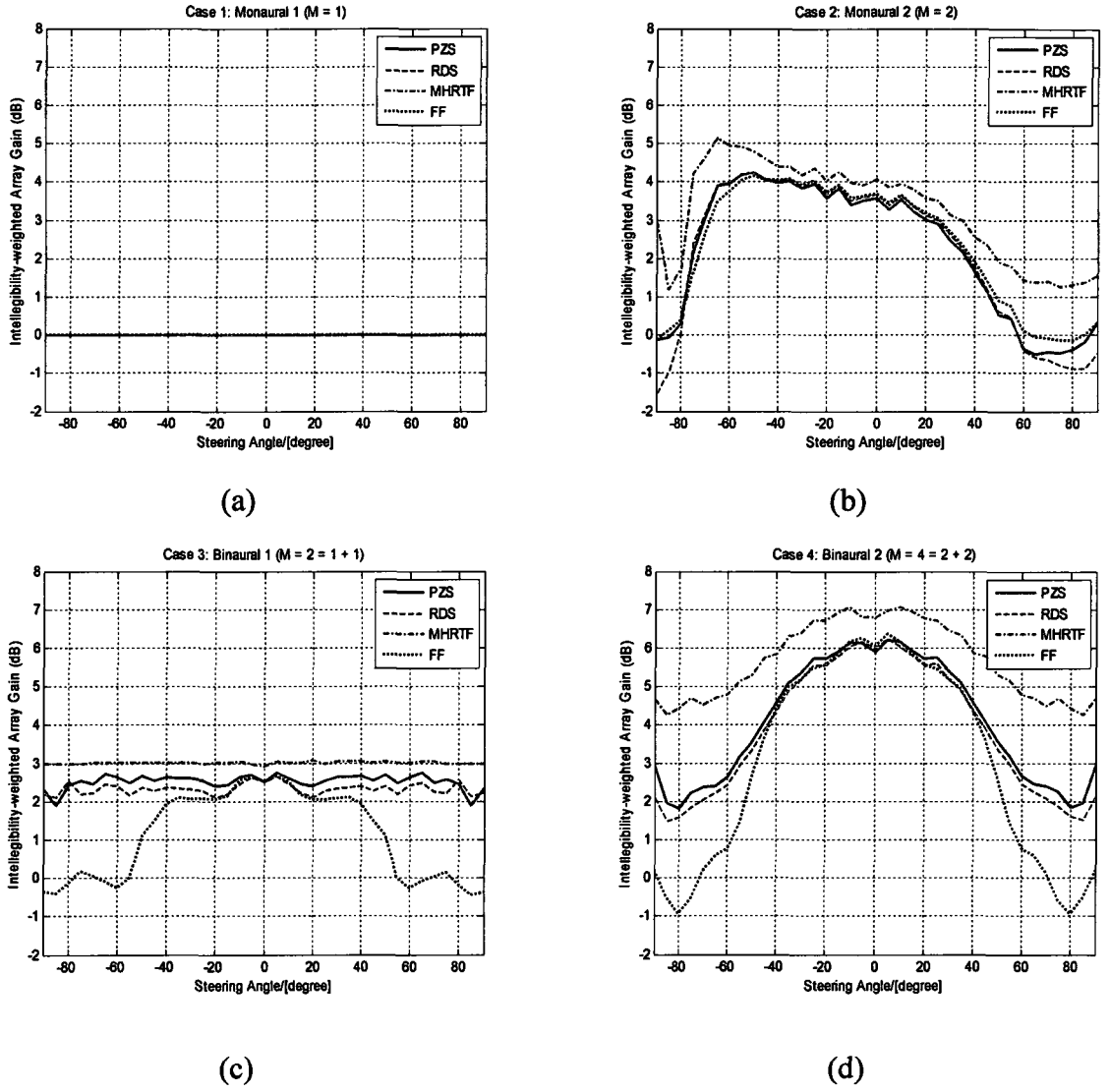


Figure 6.11 Average Array Gain of different models for different steering directions, with model mismatch

The results shown in Figures 6.7 to 6.10 show the performance when the MVDR design was made using different head models, while the Array Gain was always evaluated using the measured MHRTF model. Thus the ideal case is for the MVDR design using the MHRTF model, because in this case there is no model mismatch between the design and the performance evaluation. In addition, Figure 6.10 shows the Average Array Gain for different steering directions ($\text{Aver\_AG}(\theta_{si})$), where the Average Array Gain is obtained

by simply averaging frequency dependent Array Gain values over uniformly distributed frequency values. This can be illustrated by the following equation:

$$\text{Aver\_AG}(\theta_{si}) = 10\log_{10}\left(\frac{\sum_{j=1}^{N_{freq}} \text{AG}(\theta_{si}, F_j)}{N_{freq}}\right)(\text{dB}) \qquad (6.1)$$

where $N_{freq} = 256$ is the number of frequencies for which the Array Gains are calculated over $F = 1 \sim 10000\text{Hz}$; and $\text{AG}(\theta_{si}, F_j)$ is the array gain of the frequency F_j for each steering direction $\theta_{si} = -90°, -85°, \cdots -5°, 0°, 5°, \cdots 85°, 90°$. It should be noted that this approach does not include critical band weighting.

Generally speaking, with head models that are not particularly matched to fit the measured MHRTF, a performance within 1-2 dB of the ideal performance (i.e. with no model mismatch) was reached for a wide range of frequencies. In general, the performance of the PZS head model was the one which produced the closer results to the ideal case of no mismatch with the MHRTF model. This indicates that in practice the head model used for the fixed MVDR beamformer tuned for diffuse noise would not need to be overly accurate to provide a decent performance. Figure 6.10 also shows that overall the Monaural 2 and the binaural configurations have a similar behavior in terms of performance degradation with head model mismatch.

6.4 Conclusion

The results of this chapter have shown that if a reasonably good DOA estimate is available for a target source, then fixed binaural MVDR beamformers tuned for diffuse noise reduction outperform their monaural counterparts, and there is also a benefit to have a Binaural 2+2 configuration over a Binaural 1+1 configuration. The results are fairly robust to head model mismatch, especially for frontal or near-frontal targets, which indicates that the head model does not need to be overly accurate in practice for the design of the fixed MVDR beamformers. However, it was observed that DOA mismatch can significantly reduce the performance of fixed beamforming, thus it is important to

have a reasonably good DOA estimate of the target direction (or, equivalently, if we do not wish to have DOA estimation, it is important to force the target to be in a more narrow range of target angles, for example near-frontal). It should be noted that adaptive beamformers (i.e. with adaptive time-varying noise statistics), which we do not use here, have been found to be even less robust to DOA errors [ROH'07].

However, the classical beamforming performance results of this chapter have some limitations:

- the performance metrics of this chapter are useful to evaluate MVDR beamformers (monaural or binaural) having a single monaural output, but they may not be representative of the performance achieved by the modified binaural beamformer with a common binaural gain to preserve binaural cues, presented in Chapter 4. They are also not representative of the performance achieved by the different structures described in Chapter 5 and combining the beamforming algorithms with speech enhancement or noise reduction algorithms.
- the performance metrics of this chapter do not directly address speech quality or intelligibility, especially in the context of complex acoustic environments with time-varying diffuse noise, time-varying directional interferences and reverberation.

For those reasons, another set of experiments is performed in the next chapter, under a complex acoustic environment and using objective measures related to speech quality and intelligibility. Enhanced sound files are also produced for each method.

Chapter 7 Simulation Results with Speech Quality and Intelligibility Objective Measures

7.1 Experimental Setup for Complex Acoustic Environments

In this chapter, simulations are performed from real-life recordings. The data were recorded by a hearing aid manufacturer using two 3-channel 'Behind the Ear' (BTE) hearing aid shells mounted separately on the left and right ear of a KEMAR dummy head. On the advice of the hearing aid manufacturer, it was decided in this work to consider at most 2 microphones on each side of the head, so only up to 4 channels among 6 were selected: right-front channel, right middle channel, left front channel and left middle channel. The distance between both ears was 17.5 cm, and the distance between microphones from the same ear was 0.75 cm. To collect different real-life binaural data, the dummy head was placed in different environments such as in offices of various sizes, in a university cafeteria, near a road or in a car. The directional signals used in this chapter were separately recorded in a cafeteria environment with an average reverberation time of 1.76 seconds. The target speech signals used in this chapter were recorded at 0.75 m from the dummy head, while the directional speech interferences were recorded a 1.5 m from the dummy head. Diffuse-like binaural noise was also recorded by the hearing aid manufacturer under different environments, and binaural diffuse-like babble noise recorded from a cafeteria environment at lunch time was used in this chapter. The length of the recordings fed to the experiments was 8.5 seconds, at 20 kHz.

The following is the description of the complex acoustic scenario considered: a female target speaker is exactly in front of a hearing aid user (target azimuth = 0 degree) with a distance of 0.75 meters from the user. Two interfering speeches from male speakers both at 1.5 meters from the user were added as directional noise at -90 degrees and 120

degrees of azimuth. Transient noises such as dishes clattering were also added as lateral interferences (these were synthesized by using experimentally measured HRTFs, and thus they are the only components not including reverberation). Moreover, diffuse-like babble noise was added to the background. The level of the diffuse noise was in addition forced to be time-varying (with a sudden jump by 12 dB in the level, followed later by a sudden decrease in level), to make the scenario more challenging and to simulate the scenario when the hearing-aid user is entering a very noisy environment from a less noisy environment (e.g. user entering the cafeteria).

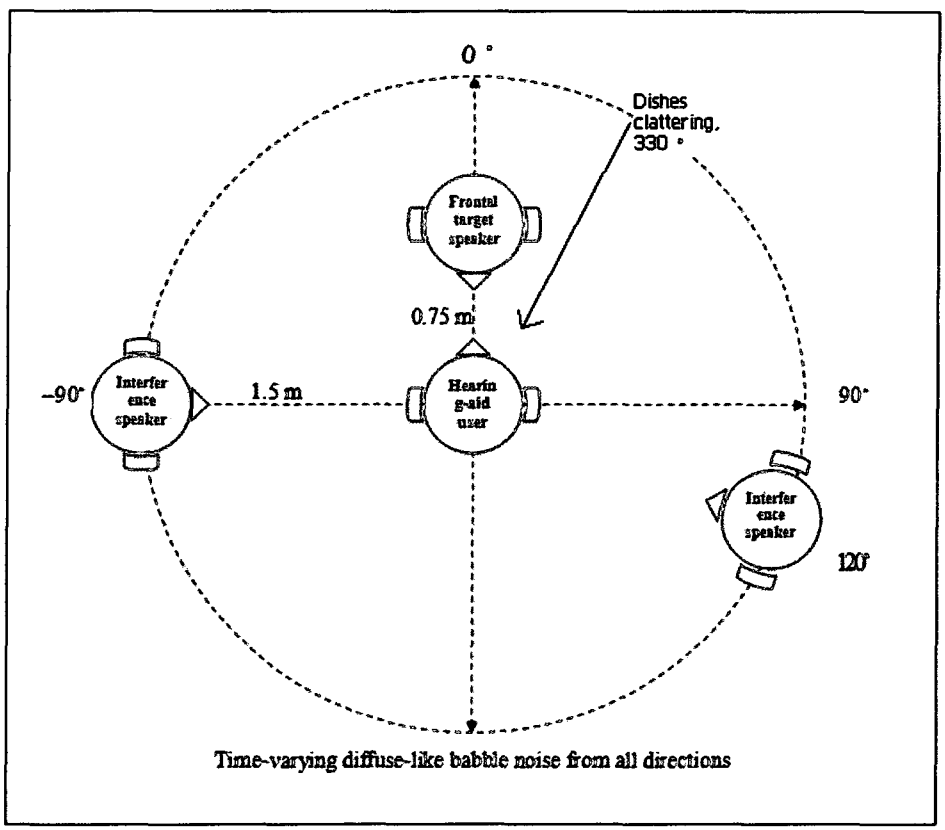


Figure 7.1 Experimental setup for complex acoustic environment

Scenario Number	Input SNR (dB)	
	Left Ear	Right Ear
Scenario-1	1.47	4.15
Scenario-2	-4.55	-1.87
Scenario-3	-14.09	-11.42

Table 7.1 Three levels of input SNR at both ears

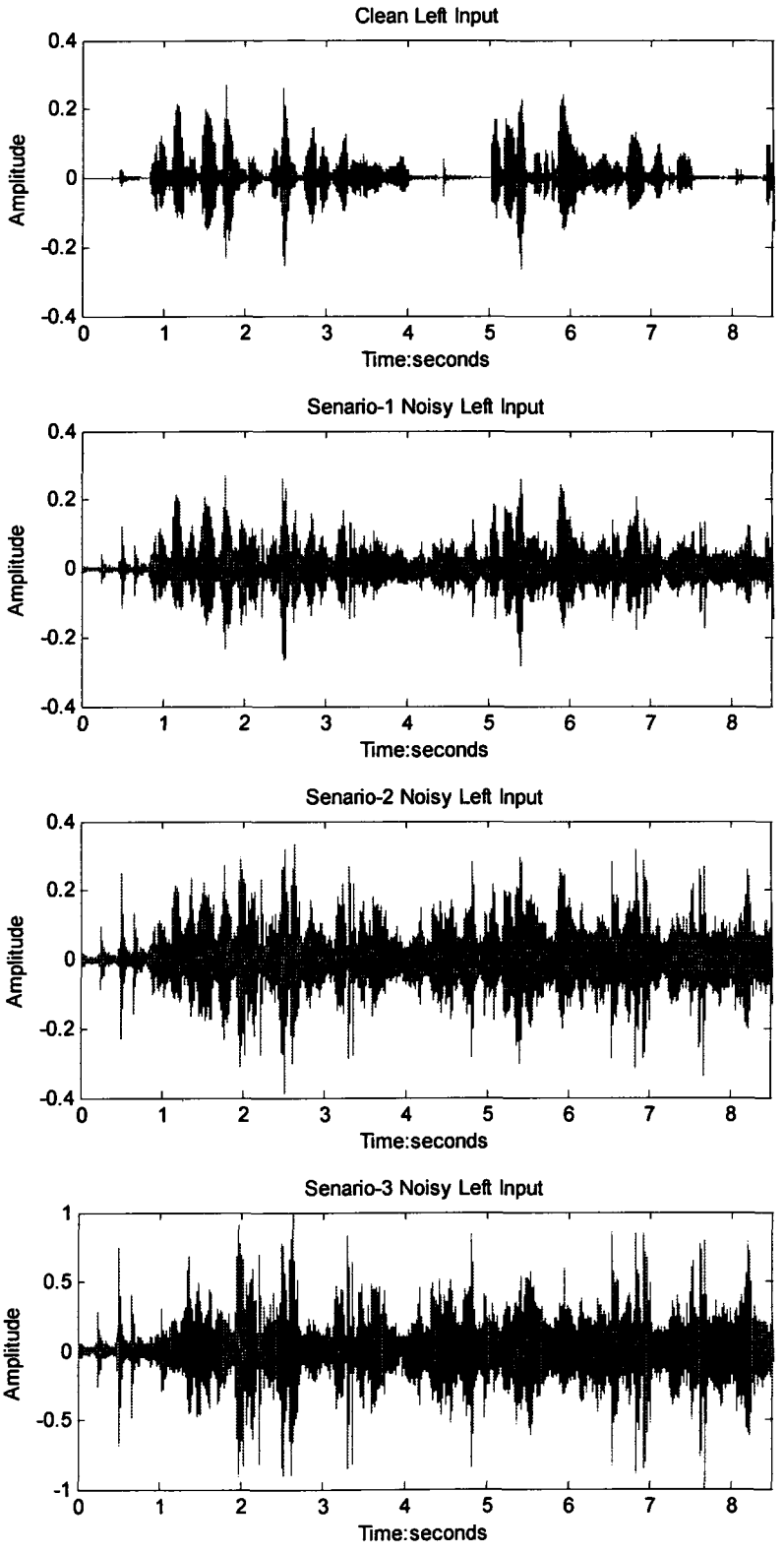


Figure 7.2 Clean and noisy input signals received by the left (front) sensor

For the acoustic scenario described in the previous paragraphs, three different levels of input SNRs will be considered in the simulations, as shown in Table 7.1. The clean speech input signal and the resulting noisy input signals received by the left (front) sensor are shown in Figure 7.2. In all the simulations, the noisy inputs were processed on a frame-by-frame basis, and the frame length was 25.6 ms (512 samples at sampling frequency $fs = 20$ kHz), with 50% overlap. A Hann window was applied to each frame. The FFT-size was set to $N = 512$. After processing the frames of the noisy input signals, an Overlap-and-Add (OLA) method was used to reconstruct the enhanced output signals. Except for Section 7.4.5 Head Model Mismatch, the design of the MVDR beamformers in this chapter were performed using the MHRTF model, which corresponds to the HRTFs measured from the same KEMAR mannequin with the same BTE units as the ones used for the binaural recordings. Taking the microphone configuration Binaural 1+1 with a common gain approach ("B1C" for short), an example of the whole processing can be illustrated in Figure 7.3:

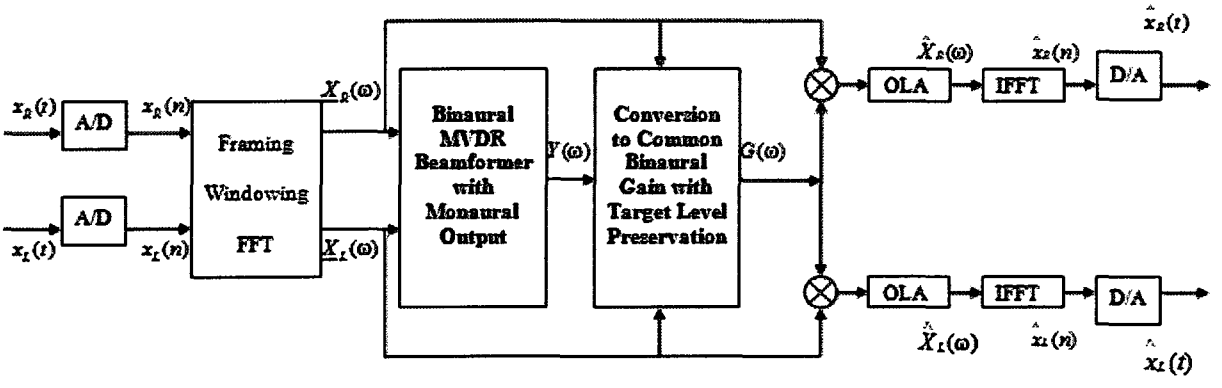


Figure 7.3 Applying binaural beamforming (B1C) to real-life recordings

7.2 Speech Quality and Intelligibility Objective Measures

To evaluate the noise reduction performance of the cases considered in this chapter, a variety of objective measures such as the Signal-to-Noise ratio (SNR), the Segmental SNR (segSNR), the Coherence Speech Intelligibility Index (CSII), and some composite

objective measures will be used in this chapter. A brief introduction to each of those objective measures is presented below.

The Signal-to-Noise Ratio (SNR) is one of the most classical measurements to evaluate the performance of a signal enhancement algorithm. It is defined as the ratio of a signal power to the noise power corrupting the signal. That is to say, the higher the ratio, the less obtrusive the background noise is. If we write the discrete time as n and the noisy signal and enhanced output as $x(n)$ and $\hat{x}(n)$ respectively, then in the logarithmic scale, the SNR can be defined as

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_n x^2(n)}{\sum_n (x(n)-\hat{x}(n))^2}\right)(\mathrm{dB}) \tag{7.1}$$

However, the SNR may not be the best choice to evaluate the performance of speech enhancement algorithms [HAN'98], since it is known to be one of the metrics least correlated with the average subjective opinion of quality, as the noise reduction that it indicates is in the mean-square sense rather than in a perceptual sense [MUS'08].

The segmental SNR (segSNR) is formed by averaging the SNRs of each frame:

$$\begin{aligned}&\mathrm{segSNR} = \frac{1}{L}\sum_{l=1}^{L}\mathrm{SNR}(l)\\ &\text{If segSNR} > 35\text{dB, then segSNR} = 35\text{dB;}\\ &\text{If segSNR} < -10\text{dB, then segSNR} = -10\text{dB.}\end{aligned} \tag{7.2}$$

where L is the number of frames, and $\mathrm{SNR}(l)$ is the SNR (in dB) of the l^{th} frame of the windowed speech, defined in (7.1). An upper threshold is used, because frames with SNRs above 35 dB do not produce large additional perceptual differences. Similarly, SNR values can be negative during the silence periods, since speech signal energy is very small then, and these frames do not reflect the true perceptual quality of the speech signal [HAN'98], so a lower SNR bound is also set at -10 dB.

The Coherence and Speech Intelligibility Index (CSII) is an extension of the Speech Intelligibility Index (SII) (ANSI S3.5-1997). SII is a mean to estimate speech

intelligibility under conditions of additive stationary noise or bandwidth reduction [KAT'05]. However, the outputs of hearing aids can be corrupted by distortion as well as noise, for example non-linear distortions such as peak-clipping and center-clipping which may be caused by the different speech enhancement algorithms. Using the coherence between the input and output signals to estimate the combined effects of noise and distortion, CSII further includes broadband peak-clipping and center-clipping distortion. The speech signal is first divided into low-, mid- and high-level regions, and then in each region a signal-to-distortion ratio (SDR) is computed by a mean-square coherence function. The advantage of SDR over SNR is that it includes both noise and distortion, while SNR considers noise only. Afterwards, weighted summing the SDRs across the frequencies for each level results in indexes for the three regions. Finally, the intelligibility index is obtained through a linear combination of the CSII results from each region [KAT'05]. CSII provides a score between "0" and "1", where "0" means that the signal is not intelligible at all, and "1" means that it is perfectly intelligible. It should be noted that CSII is applied to 16 kHz wideband speech signals, so the enhanced speech signals at 20 kHz were downsampled to 16 kHz for the computation of the CSII.

In [HU'06], composite measures were proposed by weighting and combining existing objective measures such as the Weighted-Slope Spectral (WSS) distance, the Log Likelihood Ratio (LLR) and the Perceptual Evaluation of Speech Quality (PESQ), using nonlinear or nonparametric regression models. The composite measures are referred to as *Csig*, *Cbak* and *Covl*. Comparing with conventional objective measures, they have produced a higher correlation with subjective measures. Among them, Csig represents the degree of speech distortion, Cbak rates the intrusiveness of background noise, and Covl evaluates the overall quality of the speech signals. Covl thus follows the same scale as the mean opinion score (MOS), and the other measures are also represented by a five-point scale as follows:

Composite / Measures / Scale	Csig	Cbak	Covl
1	Very unnatural, very degraded	Very conspicuous, Very intrusive	Bad
2	Fairly unnatural, fairly degraded	Fairly conspicuous, somewhat intrusive	Poor
3	Somewhat natural, Somewhat degraded	Noticeable but not intrusive	Fair
4	Fairly natural, little degradation	Somewhat noticeable	Good
5	Very natural, No degradation	Not noticeable	Excellent

Table 7.2 Meaning and scale of the composite measures

The composite measures used in this work operate at 8 kHz, thus the signals from our simulations performed at 20 kHz were downsampled to 8 kHz for the computation of the measures.

Each objective measure has its own advantages and disadvantages. By using several objective measures to evaluate the performance of the considered algorithms, it is expected that overall improvements in all or most of the measures will correspond to real speech quality and/or intelligibility improvements.

7.3 Comparison of Algorithms

7.3.1 Binaural MVDR Beamforming with Common Binaural Gain and Post-Filter

First, the best gain conversion method from Section 4.2 should be found for the "basic" Binaural 1+1 and Binaural 2+2 MVDR beamforming configurations with a common binaural gain. In order to achieve this goal, for each scenario (i.e. each input SNR level), the results generated by using conversion methods from (4.6)-(4.10) are shown in Tables 7.3-7.5. The resulting algorithms with the different gain conversion methods from (4.6)-(4.10) have the suffix "_sum", "_sqr", "_mul", "_max" and "_min". In the tables, B1C and B2C represent the Binaural 1+1 and Binaural 2+2 MVDR beamformers with a conversion to a common gain. The simulation results in this Section 7.3.1 are obtained under the assumption of frontal target, without any DOA or Head Model mismatch.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
B1C_sum	3.50	6.28	-0.77	0.11	0.69	0.80	3.28	3.44	2.13	2.25	2.58	2.74
B1C_sqr	4.26	6.68	-0.49	0.37	0.71	0.83	3.33	3.47	2.17	2.28	2.62	2.77
B1C_mul	2.87	5.89	-0.98	-0.14	0.66	0.78	3.25	3.41	2.10	2.23	2.54	2.72
B1C_max	**5.57**	**7.58**	**0.04**	**0.93**	**0.74**	**0.85**	**3.40**	**3.51**	**2.24**	**2.34**	**2.68**	**2.81**
B1C_min	2.33	5.26	-1.33	-0.56	0.63	0.74	3.20	3.37	2.06	2.18	2.50	2.67
B2C_sum	6.10	7.27	0.46	0.69	0.85	0.90	3.59	3.62	2.39	2.41	2.88	2.92
B2C_sqr	6.40	7.64	0.69	0.95	0.87	0.90	3.60	3.63	2.41	2.43	2.90	2.93
B2C_mul	5.08	7.16	0.32	0.63	0.84	0.88	3.54	3.59	2.35	2.39	2.84	2.89
B2C_max	**7.11**	**7.99**	**0.88**	**1.19**	**0.89**	**0.93**	**3.62**	**3.64**	**2.43**	**2.45**	**2.92**	**2.95**
B2C_min	3.88	6.31	-0.50	0.02	0.75	0.83	3.41	3.51	2.24	2.31	2.70	2.82

Table 7.3 Comparison of different common gain conversions for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
B1C_sum	-2.33	0.59	-4.24	-3.68	0.35	0.37	2.70	2.86	1.60	1.70	2.05	2.20
B1C_sqr	-1.50	1.05	-3.97	-3.44	0.33	0.41	2.75	2.89	1.64	1.72	2.09	2.23
B1C_mul	-3.00	0.16	-4.42	-3.88	0.37	0.37	2.67	2.83	1.57	1.67	2.01	2.17
B1C_max	**0.07**	**2.26**	**-3.34**	**-2.82**	**0.39**	**0.45**	**2.82**	**2.95**	**1.71**	**1.79**	**2.15**	**2.28**
B1C_min	-3.58	-0.55	-4.75	-4.22	0.34	0.38	2.61	2.77	1.52	1.62	1.96	2.13
B2C_sum	0.97	2.32	-2.63	-2.69	0.61	0.65	3.09	3.11	1.90	1.90	2.41	2.44
B2C_sqr	1.20	2.56	-2.45	-2.50	0.61	0.65	3.11	3.12	1.93	1.92	2.44	2.46
B2C_mul	-0.42	1.94	-2.89	-2.84	0.57	0.57	3.03	3.07	1.86	1.87	2.36	2.41
B2C_max	**2.30**	**3.16**	**-2.12**	**-2.17**	**0.65**	**0.70**	**3.14**	**3.15**	**1.97**	**1.95**	**2.47**	**2.48**
B2C_min	-1.76	0.98	-3.74	-3.45	0.40	0.49	2.86	2.96	1.73	1.78	2.20	2.30

Table 7.4 Comparison of different common gain conversions for Scenario-2

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C_sum	-11.77	-8.78	-8.16	-7.82	0.14	0.11	2.00	2.04	1.03	1.04	1.46	1.47
B1C_sqr	-10.89	-8.29	-8.02	-7.68	**0.15**	0.12	1.95	2.07	0.98	1.06	1.37	1.49
B1C_mul	-12.47	-9.23	-8.25	-7.93	0.12	0.10	**2.20**	2.02	**1.20**	1.03	**1.74**	1.45
B1C_max	**-9.20**	**-6.93**	**-7.63**	**-7.27**	**0.15**	**0.15**	2.00	2.10	1.03	1.09	1.41	1.52
B1C_min	-13.08	-9.99	-8.41	-8.13	0.11	0.12	2.07	**2.17**	1.10	**1.15**	1.59	**1.67**
B2C_sum	-8.08	-6.62	-6.93	-6.89	0.26	0.21	2.25	2.25	1.20	1.19	1.65	1.65
B2C_sqr	-7.85	-6.43	-6.82	-6.80	0.27	0.22	2.28	2.27	1.21	1.20	1.67	1.67
B2C_mul	-9.65	-7.15	-7.21	-7.07	0.23	0.17	2.21	2.23	1.16	1.17	1.61	1.63
B2C_max	**-6.53**	**-5.70**	**-6.49**	**-6.54**	**0.29**	**0.23**	**2.29**	**2.28**	**1.24**	**1.23**	**1.68**	**1.68**
B2C_min	-11.08	-8.19	-7.78	-7.48	0.16	0.15	2.06	2.14	1.06	1.11	1.48	1.56

Table 7.5 Comparison of different common gain conversions for Scenario-3

Under most scenarios, the method which clearly produced the best noise reduction performance was to convert to a binaural common gain using equation (4.9) i.e. using the maximum of the left and right input signal levels in the computation of the common gain,

leading to a smaller (i.e. more agressive) gain. The same conclusion was obtained by listening to the resulting enhanced files (a link to some sound files is provided at the end of this chapter). So in all of the simulations to follow in this chapter, equation (4.9) will be used when a binaural common gain is to be produced from a binaural beamformer.

Next, to better compare the performance of the Binaural 1+1 and Binaural 2+2 MVDR beamformers with a common gain computed with equation (4.9), i.e. "B1C" and "B2C" respectively, Table 7.6 presents a summary of the results for the three SNR scenarios:

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
SE1-noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	**2.51**	3.39	2.16	2.70
B1C	5.57	7.58	0.04	0.93	0.74	0.85	3.40	3.51	2.24	2.34	2.68	2.81
B2C	**7.11**	**7.99**	**0.88**	**1.19**	**0.89**	**0.93**	**3.62**	**3.64**	2.43	**2.45**	**2.92**	**2.95**
SE2-noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	**1.97**	**2.78**	1.60	2.13
B1C	0.07	2.26	-3.34	-2.82	0.39	0.45	2.82	2.95	1.71	1.79	2.15	2.28
B2C	**2.30**	**3.16**	**-2.12**	**-2.17**	**0.65**	**0.70**	**3.14**	**3.15**	**1.97**	1.95	**2.47**	**2.48**
SE3-noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C	-9.20	-6.93	-7.63	-7.27	0.15	0.15	2.00	2.10	1.03	1.09	1.41	1.52
B2C	**-6.53**	**-5.70**	**-6.49**	**-6.54**	**0.29**	**0.23**	**2.29**	**2.28**	**1.24**	1.23	**1.68**	**1.68**

Table 7.6 Comparison of basic Binaural 1+1 and Binaural 2+2 MVDR with a common gain

From Table 7.6, it can be seen that the Binaural 1+1 MVDR with a common gain can improve all of the objective measures for all of the 3 noisy environments (i.e. 3 SNR scenarios in Table 7.1). The Binaural 2+2 MVDR with a common gain can further improve the results. The noisier the environment, the more improvement is achieved by doubling the number of microphones (i.e. going from the Binaural 1+1 MVDR with common gain to the Binaural 2+2 MVDR with common gain). The superiority of the Binaural 2+2 configuration also matches the results from Chapter 6, which were obtained for pure monaural output MVDR beamforming (i.e. no common gain) and for classic beamforming metrics. Listening to the resulting recordings, we find that the musical noise in the Binaural 2+2 case is less obvious than the one in the Binaural 1+1 case, although the presence of a different type of distortion can be felt, similar to background modulated low-frequency noise.

Next, the potential improvement of adding a Wiener Post-Filter (WPF) after the MVDR beamformer with common gain is to be evaluated, with and without a spectral floor (SF). Please refer to Figure 5.1. The results for each SNR scenario are shown in Tables 7.7-7.9, where WPF represents a Wiener Post-Filter without a spectral floor, and WPF (0.3) represents a Wiener Post-Filter with a spectral floor set to 0.3.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
B1C	5.57	7.58	0.04	0.93	0.74	0.85	3.40	3.51	2.24	2.34	2.68	2.81
B1C+WPF	**7.54**	**8.59**	**0.89**	**1.65**	**0.84**	**0.92**	3.41	3.46	2.30	2.36	2.72	2.78
B1C+WPF(0.3)	7.11	8.36	0.66	1.49	0.80	0.90	**3.49**	**3.58**	**2.33**	**2.41**	**2.78**	**2.88**
B2C	**7.11**	**7.99**	**0.88**	**1.19**	0.89	0.93	**3.62**	**3.64**	**2.43**	**2.45**	**2.92**	**2.95**
B2C+WPF	6.84	6.96	0.73	0.87	**0.90**	**0.95**	3.19	3.22	2.28	2.29	2.60	2.63
B2C+WPF(0.3)	6.27	6.77	0.13	0.63	0.81	0.91	3.53	3.62	2.36	2.42	2.83	2.92

Table 7.7 Wiener Post Filter with or without a spectral floor for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
B1C	0.07	2.26	-3.34	-2.82	0.39	0.45	2.82	2.95	1.71	1.79	2.15	2.28
B1C+WPF	**3.04**	**4.30**	**-2.18**	**-1.66**	**0.49**	**0.59**	2.85	2.92	1.80	1.86	2.20	2.29
B1C+WPF(0.3)	2.30	3.77	-2.49	-1.97	0.45	0.54	**2.92**	**3.04**	**1.82**	**1.89**	**2.25**	**2.37**
B2C	2.30	3.16	-2.12	-2.17	0.65	0.70	**3.14**	**3.15**	**1.97**	**1.95**	**2.47**	**2.48**
B2C+WPF	**3.82**	**3.51**	**-1.78**	**-1.91**	**0.77**	**0.86**	2.81	2.81	1.91	1.87	2.26	2.25
B2C+WPF(0.3)	2.38	2.86	-2.65	-2.44	0.55	0.63	2.98	3.08	1.87	1.92	2.33	2.42

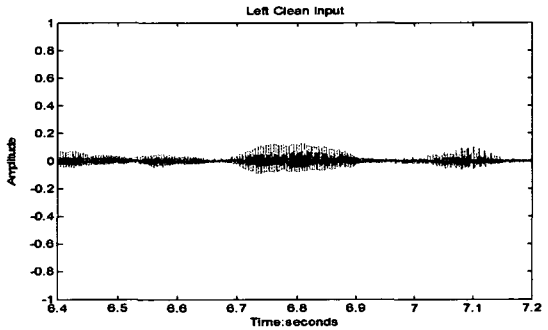
Table 7.8 Wiener Post Filter with or without a spectral floor for Scenario-2

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C	-9.20	-6.93	-7.63	-7.27	0.15	0.15	2.00	2.10	1.03	1.09	1.41	1.52
B1C+WPF	**-5.61**	**-4.21**	**-6.65**	**-6.35**	**0.16**	**0.19**	1.96	2.05	1.06	1.12	1.39	1.50
B1C+WPF(0.3)	-6.56	-4.99	-6.94	-6.62	0.14	0.16	**2.07**	**2.18**	**1.10**	**1.17**	**1.48**	**1.59**
B2C	-6.53	-5.70	-6.49	-6.54	0.29	0.23	**2.29**	**2.28**	1.24	**1.23**	**1.68**	**1.68**
B2C+WPF	**-3.61**	**-4.41**	**-5.55**	**-5.88**	**0.50**	**0.49**	2.04	2.02	**1.26**	1.20	1.57	1.54
B2C+WPF(0.3)	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	2.18	1.16	1.18	1.52	1.59

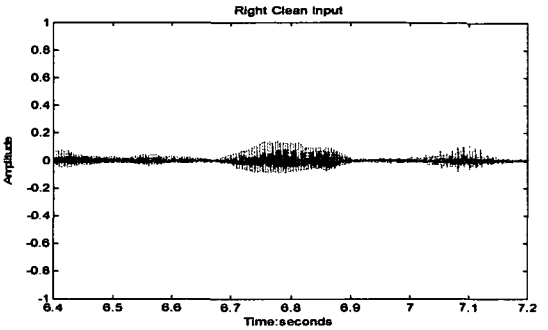
Table 7.9 Wiener Post Filter with or without a spectral floor for Scenario-3

From the data in the above three tables, it is found that for the Binaural 1+1 beamforming configuration in particular, the addition of a WPF post-filter (with or without spectral floor) helps to improve the performance in all scenarios and for all measures in general. For both beamforming configurations, adding a spectral floor to the WPF limits the level of noise reduction, but it also helps to limit the introduced musical noise distortion, as found by listening to the corresponding recordings. Consequently, since the WPF normally increases the level of musical noise introduced, if a WPF is used then a spectral floor should also be used. The benefits of using a WPF post-filter for the Binaural 2+2 beamforming configuration were less obvious, considering that a spectral floor is required. For example, in the high SNR scenario (Table 7.7) the WPF does not provide SNR benefits for the Binaural 2+2 beamforming configuration, and for all scenarios the composite measures often indicate a better performance for the Binaural 2+2 beamforming configuration without WPF. As before, the Binaural 2+2 beamforming configuration (with or without WPF) contained less musical noise than its Binaural 1+1 counterpart, as observed by listening to the corresponding sound files. Finally, it should also be noted that if a WPF with spectral floor is used, the performance of the Binaural 1+1 beamforming configuration then becomes as good (or even better at low SNR) than the performance of the Binaural 2+2 beamforming configuration. A link to some sound files is provided at the end of this chapter.

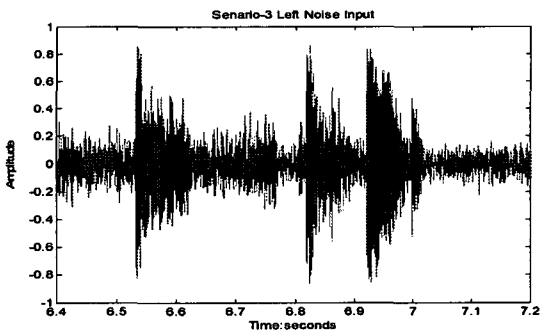
Taking the setup B2C+WPF(0.3) as an example, it is also of interest to observe the clean speech, noise, noisy speech and enhanced outputs in both the time and frequency domain. The corresponding results are shown in Figure 7.4 and 7.5, corresponding to a short segment of the signals (approx. 0.8 sec.):



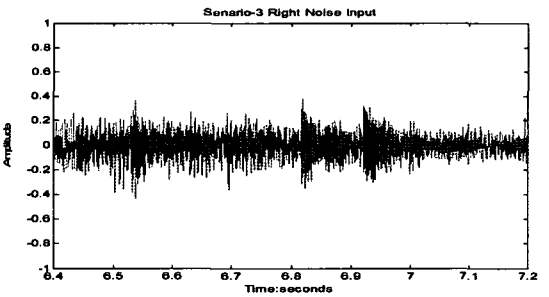
(a) Left Clean



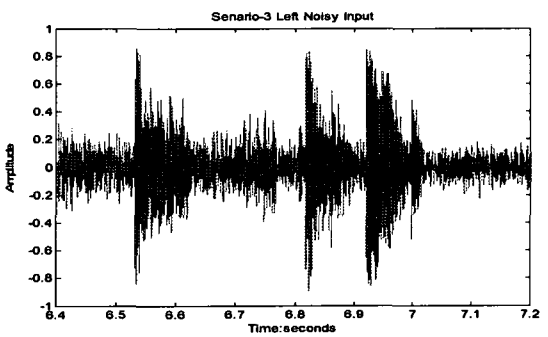
(b) Right Clean



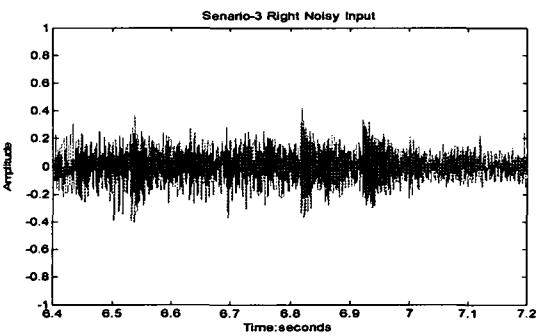
(c) Scenario-3: Left Noise



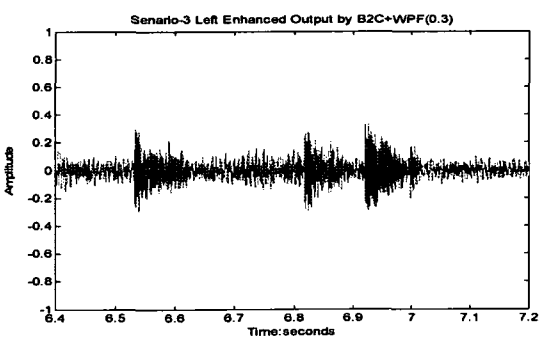
(d) Scenario-3: Right Noise



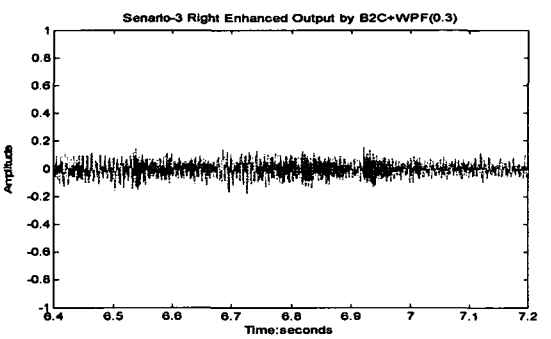
(e) Scenario-3: Left Noisy



(f) Scenario-3: Right Noisy

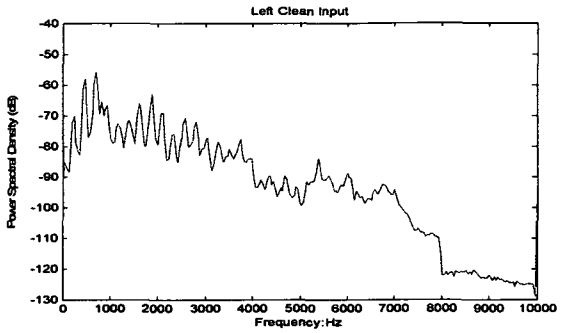


(g) Left Enhanced

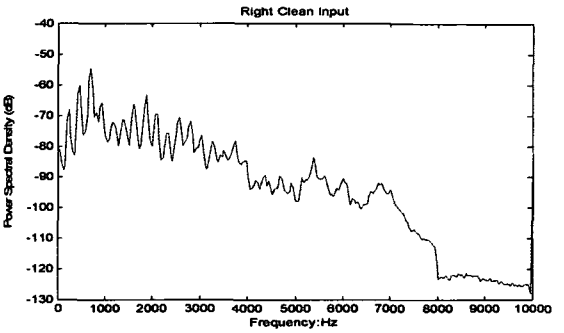


(h) Right Enhanced

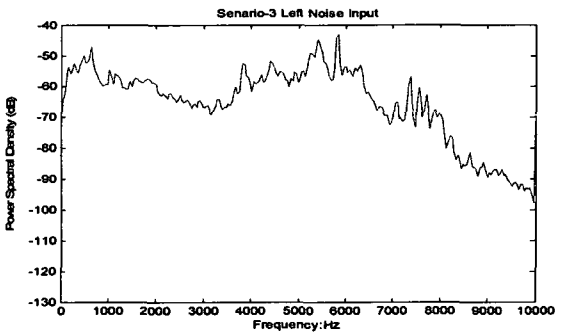
Figure 7. 4 Clean speech, noise, noisy signal and output enhanced by B2C+WPF(0.3) in the time domain



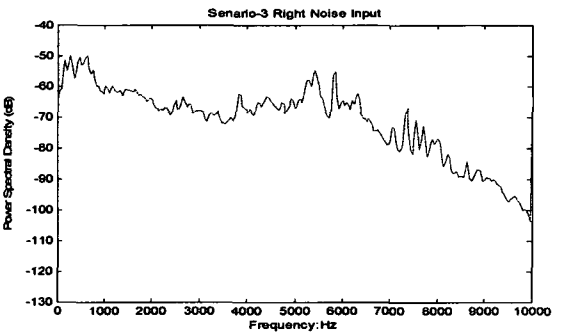
(a) Left Clean



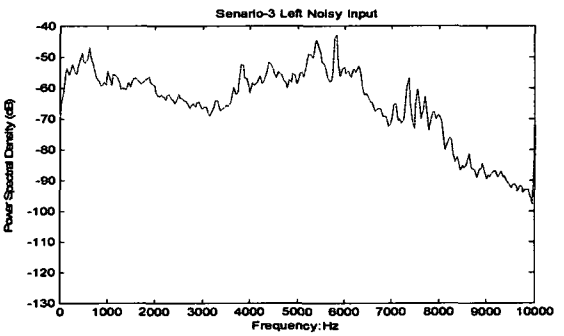
(b) Right Clean



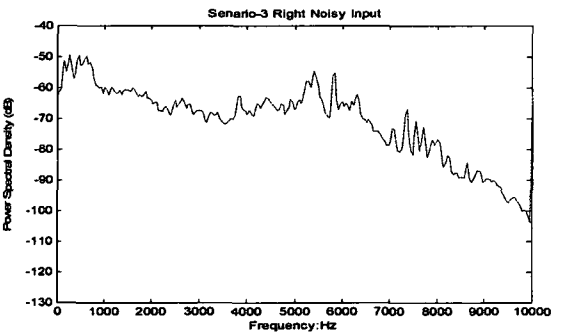
(c) Scenario-3: Left Noise



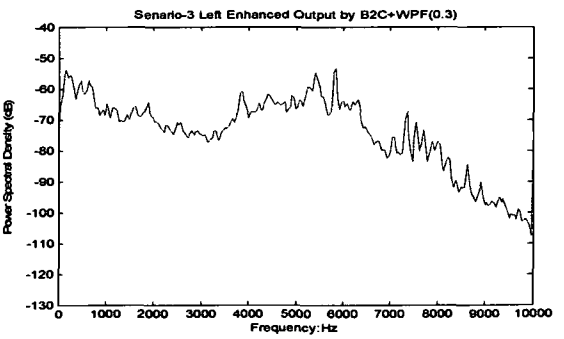
(d) Scenario-3: Right Noise



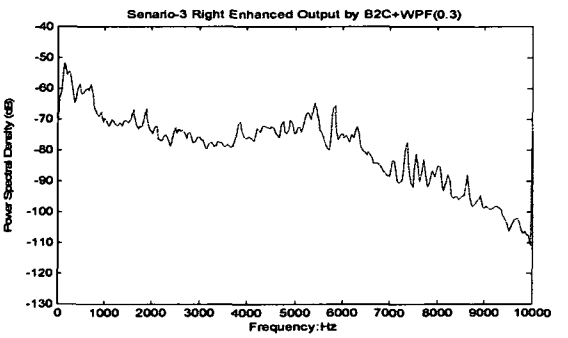
(e) Scenario-3: Left Noisy



(f) Scenario-3: Right Noisy



(g) Left Enhanced



(h) Right Enhanced

Figure 7. 5 Power Spectral Density of clean speech, noise, noisy signal and output enhanced by B2C+WPF(0.3)

7.3.2 Binaural MVDR Beamforming as a Pre-Processor for a Speech Enhancement Algorithm

As in the previous Section 7.3.1, the simulation results of this Section 7.3.2 are obtained under the assumption of frontal target, without any DOA or Head Model mismatch. Cases with DOA and Head Model mismatch will be discussed in Sections 7.3.4 and 7.3.5.

7.3.2.1 Combination with MMSE-STSA monaural enhancement

Due to the fact that applying a spectral floor to the Wiener Post Filter in a MVDR beamformer with common binaural gain can only limit the musical noise up to a certain amount, the addition of a MMSE-STSA enhancement is considered after the WPF (please refer to Figure 5.2). The "classic" MMSE-STSA is used to provide a monaural enhancement to signals on each of the left and right sides. Therefore, to preserve the binaural cues, the gains yielded by the left and right sides need to be combined into a common gain. Different possible combinations were discussed in Section 4.3 in equations (4.11)-(4.14), and we first need to determine which combination is more suitable. Simulation results corresponding to the use of equations (4.11)-(4.14) and for each SNR scenario are shown in Tables 7.10-7.12, where Eph (0.35) stands for a spectral floor of 0.35 added to the MMSE-STSA algorithm.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
B1C+WPF+Eph(0.35)_max	**5.43**	3.63	-0.62	-1.73	0.67	0.46	3.25	3.08	2.18	2.04	2.58	2.46
B1C+WPF+Eph(0.35)_min	5.30	6.07	-0.58	0.18	**0.73**	**0.85**	3.27	3.42	2.18	2.30	2.59	2.75
B1C+WPF+Eph(0.35)_sqr	**5.43**	**6.29**	**-0.55**	**0.21**	0.71	0.83	3.28	3.41	2.19	2.30	2.60	2.75
B1C+WPF+Eph(0.35)_aver	5.42	**6.29**	-0.57	0.19	0.70	0.83	3.27	3.41	2.19	2.30	2.60	2.74
B2C+WPF+Eph(0.35)_max	4.69	3.09	-1.15	-2.16	0.64	0.42	3.22	3.09	2.13	2.00	2.55	2.44
B2C+WPF+Eph(0.35)_min	4.57	4.94	-1.07	-0.50	0.68	**0.85**	3.24	3.39	2.14	2.23	2.57	2.71
B2C+WPF+Eph(0.35)_sqr	**4.74**	**5.14**	**-1.02**	**-0.47**	**0.69**	0.84	3.26	3.40	2.15	2.24	2.58	2.72
B2C+WPF+Eph(0.35)_aver	4.73	5.13	-1.04	-0.49	0.68	0.84	3.26	3.39	2.15	2.24	2.58	2.71

Table 7.10 Comparison of different common gain combinations for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
B1C+WPF+Eph(0.35)_max	1.63	0.79	-3.04	-3.73	0.36	0.25	2.68	2.57	1.72	1.64	2.07	2.00
B1C+WPF+Eph(0.35)_min	**2.02**	**3.59**	**-2.83**	**-2.25**	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
B1C+WPF+Eph(0.35)_sqr	1.93	3.59	-2.88	-2.30	**0.41**	**0.48**	2.70	2.84	1.74	1.84	2.08	2.23
B1C+WPF+Eph(0.35)_aver	1.89	3.56	-2.90	-2.34	0.39	0.47	2.70	2.84	1.74	1.83	2.08	2.22
B2C+WPF+Eph(0.35)_max	0.94	0.20	-3.65	-4.27	0.32	0.21	2.63	2.54	1.64	1.56	2.00	1.94
B2C+WPF+Eph(0.35)_min	**1.48**	**2.28**	**-3.31**	**-3.01**	**0.42**	**0.50**	2.63	2.77	1.66	1.73	2.00	2.13
B2C+WPF+Eph(0.35)_sqr	1.38	2.22	-3.38	-3.10	0.39	0.49	2.65	2.77	1.67	1.73	2.02	2.14
B2C+WPF+Eph(0.35)_aver	1.34	2.18	-3.41	-3.14	0.35	0.48	2.65	2.77	1.66	1.72	2.02	2.13

Table 7.11 Comparison of different common gain combinations for Scenario-2

Objective Measures / Algorithms	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C+WPF+Eph(0.35)_max	-6.61	-6.81	-6.88	-7.10	**0.12**	0.10	2.61	2.00	1.63	1.16	2.30	1.53
B1C+WPF+Eph(0.35)_min	**-5.67**	**-3.17**	**-6.47**	**-6.07**	**0.12**	**0.13**	**2.48**	**2.24**	**1.55**	**1.34**	**2.12**	**1.76**
B1C+WPF+Eph(0.35)_sqr	-6.02	-3.52	-6.62	-6.20	**0.12**	0.11	2.17	2.02	1.29	1.16	1.70	1.47
B1C+WPF+Eph(0.35)_aver	-6.09	-3.61	-6.66	-6.24	**0.12**	0.11	2.03	2.14	1.18	1.25	1.52	1.63
B2C+WPF+Eph(0.35)_max	-7.31	-7.49	-7.34	-7.53	0.12	0.10	2.05	2.38	1.16	1.42	1.57	2.04
B2C+WPF+Eph(0.35)_min	**-6.23**	**-4.94**	**-6.92**	**-6.61**	**0.14**	**0.19**	**2.55**	1.85	**1.58**	0.99	**2.25**	1.25
B2C+WPF+Eph(0.35)_sqr	-6.56	-5.32	-7.06	-6.74	**0.14**	0.17	2.45	2.07	1.49	1.15	2.11	1.55
B2C+WPF+Eph(0.35)_aver	-6.63	-5.39	-7.09	-6.78	0.13	0.16	2.23	**2.16**	1.31	**1.22**	1.80	**1.67**

Table 7.12 Comparison of different common gain combinations for Scenario-3

For the noisier cases of Scenario 2 and Scenario 3, overall the approach of equation (4.12) i.e. $G_{\min}(\omega) = \min(G_R(\omega), G_L(\omega))$ was found to perform best. Therefore, in the simulations to follow, the common gain of both sides will be combined using equation (4.12).

Next, a comparison is made for binaural MVDR beamforming with WPF for the following cases: without MMSE-STSA, with MMSE-STSA (no spectral floor), and MMSE-STSA with spectral floor.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
B1C+WPF	**7.54**	**8.59**	**0.89**	**1.65**	**0.84**	**0.92**	**3.41**	**3.46**	**2.30**	**2.36**	**2.72**	**2.78**
B1C+WPF+Eph	4.21	4.30	0.60	1.07	0.92	0.95	2.85	2.90	2.08	2.10	2.28	2.32
B1C+WPF+Eph(0.35)	5.30	6.07	-0.58	0.18	0.73	0.85	3.27	3.42	2.18	2.30	2.59	2.75
B2C+WPF	**6.84**	**6.96**	**0.73**	**0.87**	**0.90**	**0.95**	3.19	3.22	**2.28**	**2.29**	**2.60**	2.63
B2C+WPF+Eph	2.97	2.80	-0.88	-0.86	0.84	0.88	2.50	2.58	1.89	1.91	2.04	2.09
B2C+WPF+Eph(0.35)	4.57	4.94	-1.07	-0.50	0.68	0.85	**3.24**	**3.39**	2.14	2.23	2.57	**2.71**

Table 7.13 Comparison of MMSE-STSA with and without a spectral floor for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
B1C+WPF	**3.04**	**4.30**	-2.18	-1.66	0.49	0.59	**2.85**	**2.92**	**1.80**	**1.86**	**2.20**	**2.29**
B1C+WPF+Eph	2.87	3.17	**-1.00**	**-0.36**	**0.61**	**0.82**	2.38	2.48	1.73	1.79	1.87	1.95
B1C+WPF+Eph(0.35)	2.02	3.59	-2.83	-2.25	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
B2C+WPF	**3.82**	**3.51**	**-1.78**	**-1.91**	**0.77**	**0.86**	**2.81**	**2.81**	**1.91**	**1.87**	**2.26**	**2.25**
B2C+WPF+Eph	1.71	1.41	-2.67	-2.67	0.52	0.73	2.01	2.08	1.50	1.50	1.59	1.62
B2C+WPF+Eph(0.35)	1.48	2.28	-3.31	-3.01	0.42	0.50	2.63	2.77	1.66	1.73	2.00	2.13

Table 7.14 Comparison of MMSE-STSA with and without a spectral floor for Scenario-2

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C+WPF	-6.56	-4.99	-6.94	-6.62	0.14	0.16	2.07	2.18	1.10	1.17	1.48	1.59
B1C+WPF+Eph	**0.29**	**1.10**	**-3.62**	**-3.10**	**0.15**	**0.18**	1.57	1.67	1.17	1.21	1.17	1.22
B1C+WPF+Eph(0.35)	-5.67	-3.17	-6.47	-6.07	0.12	0.13	**2.48**	**2.24**	**1.55**	**1.34**	**2.12**	**1.76**
B2C+WPF	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	**2.18**	1.16	**1.18**	1.52	**1.59**
B2C+WPF+Eph	**-1.92**	**-2.92**	**-5.42**	**-5.59**	**0.21**	**0.34**	1.05	1.22	0.80	0.84	0.70	0.83
B2C+WPF+Eph(0.35)	-6.23	-4.94	-6.92	-6.61	0.14	0.19	**2.55**	1.85	**1.58**	0.99	**2.25**	1.25

Table 7.15 Comparison of MMSE-STSA with and without a spectral floor for Scenario-3

For the two higher SNR scenarios (Table 7.13 and 7.14), the use of the MMSE-STSA does not improve the noise reduction or the objective measures in general. However, the main reason for using the MMSE-STSA here is to reduce the musical noise or the level of

distortion introduced by the enhancement algorithms, so its use is nevertheless advisable for the Binaural 1+1 configuration (recalling from Section 7.3.1 that the Binaural 1+1 configuration suffers more from musical noise compared to the Binaural 2+2 configuration). Listening to the corresponding recordings, the MMSE-STSA can indeed reduce the musical noise significantly, especially for the Binaural 1+1 configuration in low SNR, i.e. Scenario-3. It was also found that using a spectral floor is needed to limit the aggressiveness of the MMSE-STSA algorithm, otherwise it will also introduce some distortion of its own and sound unnatural. Under low SNR conditions (Table 7.15), the MMSE-STSA algorithm with spectral floor can provide some additional noise reduction for the Binaural 1+1 configuration, but not as much for the Binaural 2+2 configuration. Therefore, the use of the MMSE-STSA with spectral floor is suitable for the Binaural 1+1 configuration with WPF, but not as much for the Binaural 2+2 configuration (for which the original Binaural 2+2 configuration without WPF might be more appropriate). A link to some sound files is provided at the end of this chapter.

7.3.2.2 Combination with PBNR binaural enhancement

As mentioned in Section 5.3, the MVDR beamformer with Binaural 1+1 and Binaural 2+2 configurations and binaural cues preservation can also be used as a pre-processor to a binaural enhancement scheme, such as the recently proposed binaural PBNR [KAM'09A], which is a Binaural 1+1 configuration and makes use of the binaural PSD estimates recently introduced in [KAM'09B], [KAM'09C]. The following tables present results for such a structure. The tables also include results for the original PBNR algorithm. From the results of the previous experiments in this chapter, four setups were selected to be combined with the PBNR algorithm:

- the Binaural 1+1 beamforming without WPF post-filter ("B1C"),
- the Binaural 1+1 beamforming with WPF and MMSE-STSA with spectral floor set to 0.35 ("B1C+WPF+Eph(0.35)"),
- the Binaural 2+2 beamforming without WPF post-filter ("B2C"),

- the Binaural 2+2 beamforming with WPF and spectral floor set to 0.3 ("B2C+WPF(0.3)").

The PBNR algorithm was also executed with two different spectral floor setups: "PBNR(0.1)" which is the normal setup and "PBNR(0.3)" which is a less aggressive setup.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	**2.51**	**3.39**	2.16	2.70
PBNR(0.1)	6.58	7.23	0.96	1.42	0.89	0.95	3.51	3.57	2.39	2.45	2.83	2.90
PBNR(0.3)	6.20	7.17	0.24	1.03	0.79	0.90	3.41	3.52	2.30	2.40	2.74	2.86
B1C	5.57	7.58	0.04	0.93	0.74	0.85	3.40	3.51	2.24	2.34	2.68	2.81
B1C+PBNR(0.1)	6.43	6.99	1.50	**1.82**	0.95	0.96	3.43	3.45	2.39	2.42	2.76	2.80
B1C+PBNR(0.3)	6.44	7.04	1.08	1.67	0.91	0.96	3.45	3.51	2.37	2.43	2.78	2.85
B1C+WPF+Eph(0.35)	5.30	6.07	-0.58	0.18	0.73	0.85	3.27	3.42	2.18	2.30	2.59	2.75
B1C+WPF+Eph(0.35) +PBNR(0.1)	4.08	4.29	0.83	1.15	**0.96**	**0.97**	3.39	3.43	2.33	2.38	2.71	2.76
B1C+WPF+Eph(0.35) +PBNR(0.3)	4.16	4.37	0.40	0.96	0.93	0.96	3.33	3.42	2.29	2.36	2.67	2.76
B2C	**7.11**	**7.99**	0.88	1.19	0.89	0.93	**3.62**	**3.64**	2.43	2.45	**2.92**	**2.95**
B2C+PBNR(0.1)	6.23	6.71	**1.64**	1.76	**0.96**	**0.97**	3.42	3.41	2.44	2.45	2.78	2.78
B2C+PBNR(0.3)	6.29	6.77	1.40	1.69	0.94	**0.97**	3.49	3.51	2.45	2.47	2.83	2.86
B2C+WPF(0.3)	6.27	6.77	0.13	0.63	0.81	0.91	3.53	3.62	2.36	2.42	2.83	2.92
B2C+WPF(0.3) +PBNR(0.1)	5.24	5.56	1.22	1.34	0.95	0.96	3.38	3.39	2.40	2.42	2.74	2.76
B2C+WPF(0.3) +PBNR(0.3)	5.28	5.61	0.98	1.26	0.93	0.96	3.43	3.49	2.40	2.45	2.78	2.85

Table 7.16 Combination of Binaural MVDR followed by WPF with PBNR for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
PBNR(0.1)	4.22	4.71	-1.52	-1.10	0.49	0.67	2.97	3.05	1.93	2.00	2.33	2.41
PBNR(0.3)	2.91	4.23	-2.26	-1.68	0.42	0.55	2.81	2.95	1.81	1.91	2.18	2.33
B1C	0.07	2.26	-3.34	-2.82	0.39	0.45	2.82	2.95	1.71	1.79	2.15	2.28
B1C+PBNR(0.1)	**4.63**	**4.96**	-0.91	-0.62	0.70	0.82	2.96	2.99	1.96	2.00	2.33	2.37
B1C+PBNR(0.3)	4.26	4.89	-1.39	-0.84	0.58	0.75	2.93	3.02	1.92	1.99	2.29	2.38
B1C+WPF+Eph(0.35)	2.02	3.59	-2.83	-2.25	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
B1C+WPF+Eph(0.35) +PBNR(0.1)	3.07	3.25	-0.85	**-0.48**	0.76	0.86	2.92	2.98	1.96	2.01	2.27	2.34
B1C+WPF+Eph(0.35) +PBNR(0.3)	3.01	3.28	-1.43	-0.82	0.54	0.76	2.80	2.91	1.88	1.97	2.18	2.29
B2C	2.30	3.16	-2.12	-2.17	0.65	0.70	**3.14**	**3.15**	1.97	1.95	**2.47**	**2.48**
B2C+PBNR(0.1)	4.62	4.78	**-0.70**	-0.61	**0.84**	**0.89**	3.04	3.02	**2.06**	2.06	2.42	2.41
B2C+PBNR(0.3)	4.49	4.78	-0.98	-0.73	0.78	0.87	3.10	3.09	**2.06**	**2.07**	2.46	2.47
B2C+WPF(0.3)	2.38	2.86	-2.65	-2.44	0.55	0.63	2.98	3.08	1.87	1.92	2.33	2.42
B2C+WPF(0.3) +PBNR(0.1)	3.90	4.01	-0.98	-0.90	0.83	0.87	2.99	3.00	2.02	2.03	2.38	2.39
B2C+WPF(0.3) +PBNR(0.3)	3.82	4.01	-1.29	-1.02	0.71	0.83	2.99	3.05	2.00	2.04	2.37	2.43

Table 7.17 Combination of Binaural MVDR followed by WPF with PBNR for Scenario-2

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	**1.91**	0.94	1.32
PBNR(0.1)	-1.87	-1.45	-5.34	-5.19	0.16	0.20	2.10	2.16	1.22	1.25	1.51	1.57
PBNR(0.3)	-4.65	-2.74	-6.14	-5.81	0.12	0.15	1.89	1.99	1.07	1.12	1.31	1.40
B1C	-9.20	-6.93	-7.63	-7.27	0.15	0.15	2.00	2.10	1.03	1.09	1.41	1.52
B1C+PBNR(0.1)	-0.27	-0.22	-4.53	-4.47	0.20	0.30	2.11	2.15	1.28	1.30	1.54	1.59
B1C+PBNR(0.3)	-1.60	-0.65	-5.03	-4.82	0.21	0.21	2.06	2.15	1.22	1.27	1.49	1.58
B1C+WPF+Eph(0.35)	-5.67	-3.17	-6.47	-6.07	0.12	0.13	**2.48**	2.24	**1.55**	1.34	**2.12**	**1.76**
B1C+WPF+Eph(0.35) +PBNR(0.1)	**1.01**	**1.18**	**-3.58**	**-3.27**	0.17	0.26	2.15	2.20	1.39	1.43	1.58	1.63
B1C+WPF+Eph(0.35) +PBNR(0.3)	0.17	0.92	-4.18	-3.77	0.13	0.17	2.00	2.11	1.30	1.37	1.46	1.56
B2C	-6.53	-5.70	-6.49	-6.54	0.29	0.23	2.29	**2.28**	1.24	1.23	1.68	1.68
B2C+PBNR(0.1)	0.20	-0.28	-4.12	-4.33	**0.45**	**0.55**	2.27	2.23	1.42	1.39	1.71	1.68
B2C+PBNR(0.3)	-0.50	-0.47	-4.36	-4.54	0.39	0.46	2.27	**2.28**	1.39	1.38	1.69	1.71
B2C+WPF(0.3)	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	2.18	1.16	1.18	1.52	1.59
B2C+WPF(0.3) +PBNR(0.1)	0.05	-0.51	-4.27	-4.42	0.38	0.52	2.15	2.17	1.35	1.34	1.58	1.60
B2C+WPF(0.3) +PBNR(0.3)	-0.48	-0.69	-4.61	-4.65	0.29	0.38	2.08	2.16	1.30	1.31	1.53	1.58

Table 7.18 Combination of Binaural MVDR followed by WPF with PBNR for Scenario-3

Comparing first the Binaural 1+1 PBNR algorithm under normal setup (0.1) with the previous algorithms from this thesis (i.e. B1C, B1C+WPF+Eph(0.35), B2C and B2C+WPF(0.3)), in terms of objective measures the PBNR algorithm outperforms the previous algorithms for mid-SNR and low-SNR situations (Scenarios 2 and 3, Tables 7.17-7.18). For high-SNR scenarios (Table 7-16), the previous B2C binaural beamforming with common gain algorithm is comparable to the PBNR and sometimes outperforms it. Listening to the corresponding files, it was found that a key point of the B2C algorithm is its good intelligibility over all scenarios (i.e. no attenuation of the target speech, unlike several of the other algorithms), even though the B2C algorithm may not always be the algorithm with the largest noise reduction. On the other hand, the PBNR algorithm normally provides a higher level of noise reduction for a tolerable level of distortion.

Next, the combinations of the previous algorithms from this thesis combined with the PBNR under normal setup (0.1) and under less aggressive setup (0.3) are compared to the original PBNR. As the SNR scenario goes from a higher SNR scenario (Table 7.16, Scenario 1) to lower SNR scenario (Table 7.18, Scenario 3), the level of additional noise reduction that can be provided by the combinations over the original PBNR steadily increases. Indeed, for high-SNR the additional noise reduction was small (when there was one), while additional noise reduction became very significant for low-SNR, as can be observed in Table 7.18. However, this additional noise reduction for lower SNR scenarios came with typically increased speech distortion, either as musical noise, muffling, clipping, or background low-frequency noise, as it was found when listening to the corresponding files. It is thus the traditional tradeoff of noise reduction versus speech distortion. When combining the binaural beamforming algorithms from this thesis and the PBNR, it should be noted that the two algorithms were tuned independently, which is certainly sub-optimal and there is be room for further performance improvement. On the other extreme of the spectrum, algorithms such as B2C and B2C+WPF(0.3) may not be spectacular in terms of noise reduction, but their intelligibility under low-SNR is quite good. A link to some sound files is provided at the end of this chapter.

7.3.3 Monaural MVDR Beamforming and MMSE-STSA Enhancement Converted to a Common Binaural Gain

This section first investigates the enhancement performance of combining monaural MVDR beamformers (M=2) followed by monaural MMSE-STSA algorithms on both sides, then converted to a common gain preserving the binaural cues (please refer to Figure 5.4). The monaural MVDR is a normal beamformer (i.e. no common gain involved) normalized by the approach of Section 2.1.4. A combination of the resulting structure with a binaural PBNR algorithm is also performed, as in the previous section. Then, the more basic approach of using directly the outputs of the monaural beamformers as the inputs of the binaural PBNR algorithm is considered and evaluated (please refer to Figure 5.5). Note that the above approach (Figure 5.5) does not guarantee to preserve the binaural cues, unlike all the other algorithms or structures that have been considered. By contrast, the combination in Figure

5.6 preserves the cues. All of the experiments in this section are implemented under the assumption of frontal target without any DOA or Head Model mismatch. In the tables, "M2" stands for the separate M=2 monaural MVDR beamformers (no common gain or cues preservation), and "M2+Eph(0.35)" stands for the case where MMSE-STSA algorithms with spectral floor 0.35 are applied on each side before the merging to a common gain (and cues preservation). For comparison, two competitive previous binaural beamforming algorithms are also included in the tables: "B1C+WPF+Eph(0.35)" and "B2C+WPF(0.3)", and the "PBNR (0.1)" is also included.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	**3.39**	2.16	2.70
B1C+WPF+Eph(0.35)	5.30	6.07	-0.58	0.18	0.73	0.85	3.27	3.42	2.18	2.30	2.59	2.75
B2C+WPF(0.3)	6.27	6.77	0.13	0.63	0.81	0.91	3.53	3.62	2.36	2.42	2.83	2.92
PBNR(0.1)	6.58	7.23	0.96	1.42	0.89	0.95	3.51	3.57	2.39	2.45	2.83	2.90
M2*	3.88	7.17	-1.09	0.39	0.84	0.84	3.51	3.62	2.25	2.40	2.80	2.92
M2+Eph	5.63	6.22	1.12	1.81	0.86	0.91	3.00	3.09	2.25	2.32	2.46	2.55
M2+Eph(0.35)	5.75	7.37	0.07	0.87	0.76	0.86	3.36	3.51	2.29	2.42	2.69	2.86
M2+PBNR(0.1)*	**6.60**	**7.70**	**1.67**	**2.65**	0.95	**0.97**	**3.64**	**3.71**	**2.55**	2.67	**2.97**	**3.07**
M2+PBNR(0.3)*	6.40	7.69	1.18	2.28	0.92	0.96	3.62	3.70	2.51	2.62	2.94	3.05
M2C+PBNR(0.1)	6.12	6.61	1.89	2.30	**0.96**	**0.97**	3.56	3.60	2.52	2.57	2.90	2.96
M2C+PBNR(0.3)	6.08	6.66	1.39	2.05	0.93	**0.97**	3.55	3.64	2.48	2.56	2.88	2.99
M2+Eph(0.35)+PBNR(0.1)	5.01	5.25	1.25	1.65	0.94	0.96	3.42	3.46	2.41	2.46	2.77	2.83
M2+Eph(0.35)+PBNR(0.3)	5.06	5.33	0.74	1.42	0.91	0.95	3.37	3.45	2.36	2.44	2.73	2.82

Table 7.19 Combination of Monaural MVDR (with and without MMSE-STSA) with PBNR for Scenario-1. Scenarios with "*" represent loss of cues preservation.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	**2.78**	1.60	2.13
B1C+WPF+Eph(0.35)	2.02	3.59	-2.83	-2.25	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
B2C+WPF(0.3)	2.38	2.86	-2.65	-2.44	0.55	0.63	2.98	3.08	1.87	1.92	2.33	2.42
PBNR(0.1)	4.22	4.71	-1.52	-1.10	0.49	0.67	2.97	3.05	1.93	2.00	2.33	2.41
M2*	-1.21	2.57	-3.77	-2.61	0.54	0.53	2.98	3.09	1.79	1.92	2.31	2.43
M2+Eph	2.83	4.37	-1.01	-0.41	0.63	0.72	2.52	2.64	1.85	1.93	2.03	2.13
M2+Eph(0.35)	1.39	4.05	-2.54	-1.94	0.46	0.47	2.78	2.94	1.80	1.92	2.16	2.33
M2+PBNR(0.1)*	**5.19**	**6.21**	-0.53	**0.36**	**0.82**	0.87	**3.26**	**3.34**	**2.18**	2.28	**2.61**	**2.70**
M2+PBNR(0.3)*	4.45	6.00	-1.01	-0.08	0.75	0.79	3.17	3.26	2.10	2.21	2.52	2.63
M2C+PBNR(0.1)	4.75	5.10	**-0.38**	0.02	0.77	**0.88**	3.15	3.21	2.13	2.19	2.51	2.58
M2C+PBNR(0.3)	4.27	5.08	-0.84	-0.34	0.74	0.81	3.09	3.20	2.07	2.15	2.45	2.57
M2+Eph(0.35)+PBNR(0.1)	3.75	3.93	-0.69	-0.25	0.74	0.86	2.96	3.02	2.02	2.07	2.34	2.40
M2+Eph(0.35)+PBNR(0.3)	3.57	3.95	-1.26	-0.64	0.55	0.74	2.86	2.96	1.94	2.03	2.25	2.36

Table 7.20 Combination of Monaural MVDR (with and without MMSE-STSA) with PBNR for Scenario-2. Scenarios with "*" represent loss of cues preservation.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	**1.91**	0.94	1.32
B1C+WPF+Eph(0.35)	-5.67	-3.17	-6.47	-6.07	0.12	0.13	**2.48**	2.24	**1.55**	1.34	**2.12**	1.76
B2C+WPF(0.3)	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	2.18	1.16	1.18	1.52	1.59
PBNR(0.1)	-1.87	-1.45	-5.34	-5.19	0.16	0.20	2.10	2.16	1.22	1.25	1.51	1.57
M2*	-10.43	-6.46	-7.71	-7.02	0.23	0.16	2.07	2.29	1.06	1.22	1.46	1.70
M2+Eph	-3.70	0.81	-4.10	-3.65	0.21	0.20	1.54	1.67	1.12	1.18	1.12	1.21
M2+Eph(0.35)	-7.18	-3.46	-6.46	-6.06	0.16	0.12	2.44	2.20	1.50	1.29	2.06	1.68
M2+PBNR(0.1)*	0.04	**1.53**	-4.22	-3.45	**0.34**	0.32	2.45	**2.56**	1.49	1.58	1.85	**1.95**
M2+PBNR(0.3)*	-1.91	0.68	-4.78	-3.99	0.33	0.23	2.18	2.42	1.30	1.49	1.56	1.84
M2C+PBNR(0.1)	**1.26**	1.49	-3.61	-3.59	0.29	**0.34**	2.37	2.42	1.50	1.51	1.80	1.84
M2C+PBNR(0.3)	-0.79	1.02	-4.05	-3.94	0.28	0.27	2.12	2.37	1.31	1.46	1.51	1.79
M2+Eph(0.35)+PBNR(0.1)	1.19	1.45	**-3.58**	**-3.30**	0.17	0.18	2.12	2.24	1.37	1.45	1.54	1.68
M2+Eph(0.35)+PBNR(0.3)	-0.21	1.11	-4.19	-3.81	0.14	0.17	2.04	2.11	1.32	1.35	1.50	1.54

Table 7.21 Combination of Monaural MVDR (with and without MMSE-STSA) with PBNR for Scenario-3. Scenarios with "*" represent loss of cues preservation.

Some conclusions can be made from the objective scores in Tables 7.19-7.21 and from listening to the corresponding files. Comparing first the new "M2", "M2+Eph" and "M2+Eph(0.35)" algorithms with the previous algorithms, the results of the "M2+Eph" without spectral gain show good objective scores but must be discarded, since additional distortion is added when the MMSE-STSA algorithm is used without spectral floor, as discussed in a previous section. The results for the new monaural "M2" and "M2+Eph(0.35)" algorithms are competitive with the previous binaural "B1C+WPF+Eph(0.35)" and "B2C+WPF(0.3)" algorithms, for all SNR scenarios, in terms of objective measures or when listening to the corresponding files. Listening to the corresponding files, one characteristic of the new algorithms ("M2", "M2+Eph(0.35)") is that they leave a bit more reverberation in the output, however they normally sound quite natural, with good intelligibility. However, the "M2" algorithm does not preserve the binaural cues, and its performance may not justify its use alone. Yet it is completely free of musical noise, as in all "true" fixed beamforming designs (i.e. no common gain conversion).

However, evaluating the performance when combined with the binaural PBNR, in all scenarios from Table 7.19 to Table 7.21, the combination "M2+PBNR(0.1)" proved to outperform the original "PBNR (0.1)" performance, which is a good achievement. Moreover, listening to the corresponding files also confirmed that the combination "M2+PBNR(0.1)" produces typically more intelligible outputs that the original "PBNR (0.1)", without the additional distortion that was experimented when other algorithms were combined with the PBNR previously. Thus this is an interesting result which shows a good potential for a combination of beamforming pre-processing with the binaural PBNR. However, the combination "M2+PBNR(0.1)" does not guarantee to preserve the binaural cues. The combination of the "M2+Eph(0.35)" with the "PBNR" was not as fruitful, as it introduced some additional distortion. A link to some sound files is provided at the end of this chapter.

7.3.4 DOA Mismatch

In all of the previous sections, no DOA mismatch was considered, but in reality it is a very common problem for the use of MVDR beamforming in microphone arrays. Simulations

were performed to assess the robustness of binaural beamformers with common gain over DOA mismatch, and the following three tables show the results for some algorithms discussed earlier: Binaural 1+1 MVDR followed by WPF plus MMSE-STSA with a spectral floor of 0.35 ("B1C+WPF+Eph(0.35)"), and Binaural 2+2 MVDR followed by WPF with a spectral floor of 0.3 ("B2C+WPF(0.3)"). In the comparison, the MVDR design is performed for a target at either 0, +10 or -20 degrees, while the input binaural files used for actual processing are the ones corresponding to a frontal target. The same complex acoustic environment as before and the same corresponding three SNR scenarios are used for this section.

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
0_B1C+WPF+Eph(0.35)	5.30	6.07	-0.58	0.18	0.73	0.85	3.27	3.42	2.18	2.30	2.59	2.75
Right_10_B1C+WPF+Eph(0.35)	4.52	5.06	-0.84	-0.11	0.73	0.84	3.26	3.40	2.17	2.28	2.58	2.73
Left_20_B1C+WPF+Eph(0.35)	4.17	4.62	-1.09	-0.36	0.73	0.82	3.21	3.37	2.12	2.24	2.52	2.70
0_B2C+WPF(0.3)	6.27	6.77	0.13	0.63	0.81	0.91	3.53	3.62	2.36	2.42	2.83	2.92
Right_10_B2C+WPF(0.3)	4.95	5.36	-0.25	0.24	0.80	0.91	3.51	3.61	2.34	2.40	2.81	2.91
Left_20_B2C+WPF(0.3)	4.09	4.38	-1.36	-0.94	0.66	0.85	3.16	3.28	2.12	2.20	2.52	2.65

Table 7.22 DOA mismatch for Scenario-1

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
0_B1C+WPF+Eph(0.35)	2.02	3.59	-2.83	-2.25	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
Right_10_B1C+WPF+Eph(0.35)	1.79	3.11	-2.97	-2.42	0.42	0.49	2.68	2.83	1.72	1.83	2.05	2.21
Left_20_B1C+WPF+Eph(0.35)	1.56	2.88	-3.10	-2.54	0.42	0.42	2.63	2.78	1.68	1.79	2.00	2.16
0_B2C+WPF(0.3)	2.38	2.86	-2.65	-2.44	0.55	0.63	2.98	3.08	1.87	1.92	2.33	2.42
Right_10_B2C+WPF(0.3)	1.90	2.28	-2.90	-2.69	0.49	0.62	2.96	3.06	1.86	1.90	2.30	2.40
Left_20_B2C+WPF(0.3)	1.06	1.68	-3.56	-3.39	0.34	0.46	2.62	2.73	1.69	1.73	2.03	2.13

Table 7.23 DOA mismatch for Scenario-2

	SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy	-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
0_B1C+WPF+Eph(0.35)	-5.67	-3.17	-6.47	-6.07	0.12	0.13	2.48	2.24	1.55	1.34	2.12	1.76
Right_10_B1C+WPF+Eph(0.35)	-5.62	-3.23	-6.51	-6.12	0.13	0.13	2.49	2.15	1.56	1.27	2.14	1.65
Left_20_B1C+WPF+Eph(0.35)	-5.88	-3.28	-6.54	-6.14	0.11	0.13	2.13	1.93	1.28	1.10	1.68	1.36
0_B2C+WPF(0.3)	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	2.18	1.16	1.18	1.52	1.59
Right_10_B2C+WPF(0.3)	-5.94	-5.57	-6.80	-6.62	0.17	0.22	2.07	2.16	1.14	1.16	1.49	1.56
Left_20_B2C+WPF(0.3)	-6.83	-5.92	-7.09	-6.87	0.14	0.17	2.15	1.93	1.29	1.08	1.72	1.39

Table 7.24 DOA mismatch for Scenario-3

It can be seen from the three tables above that for both algorithms considered and in all three noisy scenarios, the achieved noise reduction decreases in the presence of DOA mismatch. As expected, the larger the mismatch is, the larger the decrease of noise reduction is in general. Comparing with the frontal target MVDR design case (i.e. no DOA mismatch), the decrease in noise reduction can be as large as 2.4 dB in SNR and 1.6 dB in segmental SNR for the higher SNR ratio (Table 7.22, Scenario 1), and as large as 0.9 dB in SNR and 0.4 dB in segmental SNR for the lower SNR ratio (Table 7.24, Scenario 3). In all cases, even with DOA mismatch, the resulting scores were normally still significantly better than scores from the original noisy file, which indicates a reasonable degree of robustness to DOA mismatch. Listening to the corresponding recordings, it was found however that in some cases DOA mismatch can also cause a specific directional interfering noise to be less attenuated, affecting the intelligibility of the target speaker. A link to some sound files is provided at the end of this chapter.

7.3.5 Head-Model Mismatch

In Sections 7.3.1 to 7.3.4, the MVDR beamformers were all designed using the head model MHRTF, which is based on the recorded data from a KEMAR mannequin, while the complex environment recordings were also from the same KEMAR mannequin setup. Therefore, as discussed in Chapter 6, the head model MHRTF should produce the best performance among all of the four head models considered in this thesis. Chapter 6 also showed that the difference in performance between the MHRTF head model and the other

three head models was not very large in term of classical measures. It is of interest to see whether the same conclusion can be achieved using the speech quality and intelligibility objective measures. Therefore, the MVDR beamformers in this section are designed using either the measured MHRTF head model (as before) or using the synthetic head model PZS. The same algorithms as in the previous were selected for the comparison, i.e. "B1C+WPF+Eph(0.35)" and "B2C+WPF(0.3)".

		SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy		1.47	4.15	-1.82	-1.11	0.58	0.69	3.20	2.04	2.51	3.39	2.16	2.70
B1C+WPF+Eph(0.35)	MHRTF	5.30	6.07	-0.58	0.18	0.73	0.85	3.27	3.42	2.18	2.30	2.59	2.75
	PZS	5.02	5.70	-0.71	0.03	0.71	0.84	3.26	3.41	2.17	2.28	2.58	2.74
B2C+WPF(0.3)	MHRTF	6.27	6.77	0.13	0.63	0.81	0.91	3.53	3.62	2.36	2.42	2.83	2.92
	PZS	6.74	7.47	0.35	0.94	0.82	0.92	3.58	3.67	2.41	2.47	2.89	2.98

Table 7.25 Head Model mismatch with MVDR design using MHRTF or PZS for Scenario-1

		SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy		-4.55	-1.87	-5.25	-4.75	0.29	0.30	2.60	1.50	1.97	2.78	1.60	2.13
B1C+WPF+Eph(0.35)	MHRTF	2.02	3.59	-2.83	-2.25	0.39	0.47	2.69	2.85	1.73	1.84	2.07	2.23
	PZS	1.96	3.42	-2.90	-2.32	0.42	0.48	2.68	2.84	1.72	1.83	2.06	2.22
B2C+WPF(0.3)	MHRTF	2.38	2.86	-2.65	-2.44	0.55	0.63	2.98	3.08	1.87	1.92	2.33	2.42
	PZS	2.45	3.11	-2.62	-2.38	0.61	0.66	3.06	3.17	1.93	1.97	2.41	2.51

Table 7.26 Head-model mismatch with MVDR design using MHRTF or PZS for Scenario-2

		SNR L(dB)	SNR R(dB)	segSNR L(dB)	segSNR R(dB)	CSII L	CSII R	Csig L	Csig R	Cbak L	Cbak R	Covl L	Covl R
Noisy		-14.09	-11.42	-8.64	-8.43	0.07	0.10	1.67	0.78	1.06	1.91	0.94	1.32
B1C+WPF+Eph(0.35)	MHRTF	-5.67	-3.17	-6.47	-6.07	0.12	0.13	2.48	2.24	1.55	1.34	2.12	1.76
	PZS	-5.63	-3.19	-6.50	-6.09	0.13	0.13	2.44	2.00	1.51	1.15	2.07	1.44
B2C+WPF(0.3)	MHRTF	-5.97	-5.53	-6.72	-6.54	0.16	0.22	2.10	2.18	1.16	1.18	1.52	1.59
	PZS	-6.12	-5.46	-6.85	-6.65	0.16	0.19	2.19	2.28	1.20	1.24	1.62	1.70

Table 7.27 Head-model mismatch with MVDR design using MHRTF or PZS for Scenario-3

From Tables 7.25-7.27, it can be found that for a frontal target the performance of binaural fixed MVDR beamforming with common gain is similar whether the design was made with the MHRTF head model (no mismatch) or the PZS head model (mismatch). Indeed,

comparing with the MHRTF head model design, the decrease in noise reduction with the PZS head model design is less than 0.4 dB in SNR and 0.3 dB in segmental SNR for the higher SNR ratio (Table 7.25, Scenario 1), and less than 0.15 dB in SNR and 0.15 dB in segmental SNR for the lower SNR ratio (Table 7.27, Scenario 3). This is providing an indication that the accuracy of the head model may not be critical in practice, especially for a frontal or near-frontal target.

7.4 Discussion

From the previous sections, it can first be concluded that among the methods discussed in Chapter 4 to convert to and combine binaural common gains, equation (4.9) and equation (4.12) produce among the best results. It was also found that for a basic binaural MVDR beamformer with a common gain (without post-filter), the Binaural 2+2 configuration is superior to the Binaural 1+1 configuration. The use of a Wiener Post-Filter improved the noise reduction (mostly for the Binaural 1+1 configuration), and adding a spectral floor to it helped to reduce the musical noise that the post-filter also introduces. The addition of a MMSE-STSA stage was shown to help to further reduce the musical noise, and a spectral floor to the MMSE-STSA can also limit its aggressiveness to avoid producing distorted unnatural speech. This additional MMSE-STSA was found to be mostly useful for the Binaural 1+1 configuration. With a Wiener post-filter plus MMSE-STSA with spectral floor, the Binaural 1+1 configuration becomes competitive with the basic Binaural 2+2 configuration (i.e. with no post-filter, which is a good setup for the Binaural 2+2 configuration).

Structures making use of two independent Monaural M=2 MVDR beamformers were also considered. One approach was to merge them with monaural MMSE-STSA (with spectral floor for better performance) and then convert the result into a common gain, preserving the binaural cues. The other approach was to simply use the outputs of the monaural beamformers separately, not preserving the binaural cues. The best results produced by these structures were competitive with the best results produced by the previous binaural beamformers with common gain algorithms.

Both the binaural beamforming algorithms and the above structures making use of independent monaural beamformers were then combined with a recent binaural noise reduction algorithm called PBNR. The PBNR by itself was found to typically outperform the beamforming algorithms from this thesis. While for most cases it was possible to obtain further noise reduction with the combinations of algorithms from this thesis and the PBNR, it was also typically at the cost of increased distortion (typically increased target speech muffling or attenuation). The exception was the combination of the straight monaural beamformer outputs (no merging to common gain) with the PBNR, which produced highly intelligible results with also improved objective scores, outperforming the PBNR algorithm, with no guarantee of binaural cues preservation however. As previously mentioned in Chapter 5, since the binaural PBNR algorithm has a Binaural 1+1 configuration with limited capabilities for canceling sources from the back (please refer to Figure 6.1) and since the monaural MVDR beamformers are suitable for canceling sources coming from the back (again please refer to Figure 6.1), the combination of the two algorithms makes sense, as long as some cues distortion is allowed.

From the results of this chapter, it can also be concluded that a DOA mismatch between -20 degrees and 20 degrees can significantly affect the overall performance of fixed binaural MVDR beamformers tuned for diffuse noise, in terms of speech quality and intelligibility objective measures. However, it can also be said that the performance with DOA mismatch was nevertheless found to be fairly robust in the sense that even with mismatch the performance was still quite better than for the original noisy file, and the degradation in terms of SNR was never greater than 2.5 dB. It should be noted that this sensitivity of fixed beamformers is not as bad as for the case of adaptive beamformers [ROH'07]. As far as head model mismatch is concerned, it was found for the case of a frontal target that the binaural beamformer is quite robust to it (i.e. it is less sensitive to head model mismatch than to DOA mismatch), with a decrease of performance of less than 0.4 dB.

This chapter made use of speech quality and intelligibility objective measures, but some of the conclusions and observations in this chapter were also made from listening to the output sound files produced by the different algorithms. Therefore, some sound files representative

of the experiments made in this chapter as well as some explanatory notes can be found at the following address:

http://www.site.uottawa.ca/~bouchard/papers/thesis_Zhengwei_Luo.zip

Chapter 8 Conclusion and Future Work

8.1 Summary and Review of Contributions

Some of the objectives of this thesis were to compare the performance of different microphone array configurations for hearing aids (monaural, binaural, with different number of microphones), to investigate the performance of using different head models for fixed MVDR beamformer design, to investigate the effects of head model mismatch and DOA mismatch for fixed MVDR beamformer design, and to develop and evaluate new general methods of converting the binaural beamforming output to a common gain in order to preserve the binaural cues. The experimental conclusions in the thesis are based on both the classical beamforming performance measures such as Beampattern, Array Gain, and Noise Gain, and on speech quality and intelligibility objective measures computed for complex noisy acoustic environments including time-varying diffuse-like noise, multiple directional interfering sources (speeches and transient sounds), and reverberation. Different combinations of algorithms using fixed binaural beamformers as a pre-processor followed by other noise reduction or speech enhancement algorithms were investigated as well. Structures making use of existing monaural beamformers and combined with binaural processing algorithms were also presented.

Using classical measures, Chapter 6 presented the performance of fixed MVDR beamformers tuned for diffuse noise for the following comparisons: different array configuration, different head models, DOA mismatch and head model mismatch for frontal and non-frontal target sources. It showed the benefits of the Binaural 2+2 configuration over the Binaural 1+1 configuration, the benefits of the Binaural 2+2 configuration over the Monaural 2 configuration, and the benefits of the Binaural 1+1 configuration over the Monaural 1 configuration (i.e. no beamforming). It also illustrated that the MVDR binaural beamformer tuned for diffuse noise is robust to head model mismatch, especially for frontal targets. This is an important feature since in practice mathematical head models must be used as opposed to individual recording of HRTFs. However, Chapter 6 also showed that the performance of

the MVDR binaural beamformer tuned for diffuse noise can be more sensitive to DOA mismatch, in terms of classic Array Gain measures.

By contrast, Chapter 7 used some speech quality and intelligibility objective measures for the assessment of the performance of fixed MVDR beamformers tuned for diffuse noise and converted to a common gain, operating in complex acoustic environments. Different alternatives to convert to a common gain or to merge common gains were compared. Different setups of Binaural 1+1 and 2+2 configurations for binaural beamformers were compared: with and without Wiener post-filter, with and without spectral floor, with and without MMSE-STSA, etc. The best setups for the Binaural 1+1 configuration (e.g. with post-filter, with MMSE-STSA and spectral floor) are competitive with the best setups for the Binaural 2+2 configuration (e.g. with no post-filter), each with their own advantages and disadvantages (e.g. more noise reduction versus more natural sounding). Structures making use of two independent Monaural M=2 MVDR beamformers were also considered, also leading to comparable results under the best setups.

Combinations of the above algorithms with a recent binaural noise reduction algorithm called PBNR were evaluated, in most cases leading to further noise reduction but also increased distortion (typically increased target speech muffling or attenuation). The notable exception was the straight combination of the monaural beamformer outputs (i.e. no merging to a common gain) with the PBNR, which produced highly intelligible results with also improved objective scores over the PBNR algorithm, at the cost of not guaranteeing cues preservation however. DOA mismatch was found to significantly affect the overall performance of fixed binaural MVDR beamformers tuned for diffuse noise, in terms of speech quality and intelligibility objective measures, but nevertheless the performance with DOA mismatch was found to be reasonably robust. Regarding head model mismatch, the performance of fixed binaural MVDR beamformers was found to be even more robust.

8.2 *Suggestions for Future Research Work*

Due to the fact that a fixed MVDR beamformer tuned for diffuse noise can be sensitive to DOA mismatch, one approach to improve its performance would certainly be to investigate the design of an accurate and robust DOA estimator under the typical acoustic environments found in hearing aids and for the typical target directions of interest (e.g. between -20 to 20 degrees of azimuth). A joint tuning of fixed MVDR algorithms as a pre-processor and another binaural algorithm such as the PBNR as a post-processor could likely lead to an improved performance, compared to the performance observed in Chapter 7, and this should be further investigated. While the thesis has been making use of several objective measures (classical beamformer performance metrics and speech quality and intelligibility measures), some of the conclusions from the thesis were found by informal listening of the resulting sound files. More structured listening tests would allow a further validation of the results reported in the thesis.

List of References

[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.

[BEN'05] J. Benesty, J. Chen, Y, Huang, and S. Doclo, "Study of the Wiener filter for noise reduction", in Speech Enhancement, J. Benesty and et al eds., Chapter 2, pp. 9-41, Springer-Verlag: Berlin, 2005

[BOD'94] M. Bodden, "Binaural hearing and future hearing-aids technology", Journal de Physique IV, Colloque C5, supplément au Journal de Physique 111, Volume 4, pp. 411-414, May 1994

[BOG'07] T. Van den Bogaert, J. Wounters, S. Doclo, M. Mooneen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter", ICASSP 2007, pp 565-568, April 2007.

[BRO'98] C. P. Brown, and R. O. Duda, "A structural model for binaural sound synthesis", IEEE Trans. Speech, and Audio Processing, vol. 6, no. 5, pp. 476 - 488 , September 1998

[CAM'03] D.R. Campbell and P. W. Shields, "Speech enhancement using sub-band 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, April 1994.

[COX'87] H. Cox, R. M. Zeskind and M.M. Owen, "Robust adaptive beamforming", IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. asp-35, no. 10, pp. 1365-1376 October 1987.

[DIL'01]H. Dillon, Hearing Aids, Stuttgart: Thieme, 2001

[DUD'98] R.O. Duda and W.L. Martens, "Range dependence of the response of a spherical head model," J. Acoust. Soc. Am. 104(5), pp. 3048 – 3058, Nov. 1998.

[ELK'01] G. Elko, "Superdirectional microphone arrays", Chapter 10 in *Acoustic signal processing for telecommunication,* Kluwer, second printing 2001, pp.181-238.

[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

[GAN'09] A. Ganse, "An Introduction to beamforming", Applied Physics Laboratory, University of Washington, Seattle http://staff.washington.edu/aganse/beamforming/beamforming.html

[GOR'08] J. D. Gordy, M. Bouchard, and T. Aboulnasr, "Beamformer performance limits in monaural and binaural hearing aid applications", CCECE 2008, pp. 381-386, May 2008

[GRI'82] L. J. Griffiths, C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming", IEEE Transactions on Antennas and Propagation, vol. AP-30, n. 1, pp. 27 - 34, January 1982

[HAM'02] V. Hamacher, "Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids", ICASSP 2002, pp. IV-4008- IV-4011, May 2002

[HAN'98] J. Hansen, B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms", International Conference on Spoken Language Processing (ICSLP), vol. 7, Sydney, Australia, pp. 2819–2822, December 1998

[HAR'99] B. Hartmann, "How We Localize Sound", Physics today on the web, November 1999, http://www.aip.org/pt/nov99/locsound.html (last accessed July 2nd, 2009)

[HU'06] Y. Hu and P. C. Loizou , "Subjective comparison of speech enhancement algorithms", ICASSP 2006, vol. 1, pp. 153-156, May 2006

[KAM'09A] A.H. Kamkar-Parsi and M. Bouchard, "Advanced binaural noise reduction scheme for binaural hearing aids operation in complex noisy environments", submitted for publication in IEEE Trans. on Audio, Speech and Language Processing

[KAM'09B] A. H. Kamkar-Parsi, and 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

[KAM'09C] A.H. Kamkar-Parsi, and M. Bouchard, "Instantaneous target speech power spectrum estimation for binaural hearing aids and reduction of directional interference with preservation of interaural cues", submitted for publication in IEEE Trans. on Audio, Speech and Language Processing

[KAT'05] J. M. Kates, K. H. Arehart, "Coherence and the speech intelligibility index", J. Acoust. Soc. Am., 117 (4), pp. 2224 - 2237, April 2005

[KLA'05] T.J. Klasen, M. Moonen, T. Van den Bogaert, J. Wouters, 'Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction', ICASSP 2005, vol. 33, pp. 29-32, Mar. 2005

[KLA'06] T. J. Klasen et al, "Binaural multi-channel Wiener filtering for hearing aids: preserving interaural time and level differences," ICASSP 2006, vol. 5, pp. 145 – 148, May 2006.

[KLA'07] T. J. Klasen, T. Van den Bogaert, M. Moonen and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues", IEEE Transaction on Signal Processing, vol. 55, no. 4, pp. 1579 - 1585, April 2007

[LEU'06] S. Leukimmiatis and P. Maragos, "Optimum post-filter estimation for noise reduction in multichannel speech processing", 14th European Signal Processing Conference (EUSIPCO), Florence, Italy, Sept. 4-8, 2006

[LOT'06] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming", EURASIP Journal on Applied Signal Processing, vol. 2006, no.1, pp. 1-14, 2006

[MOO'89] B.C.J. Moore, An Introduction to the Psychology of Hearing, New York: Academic Press, 1989

[MOZ'80] R. A. Monzingo and T. W. Miller, Introduction to Adaptive Arrays, John Wiley & Sons, New York, NY, USA, 1980.

[MUS'08] F. Mustière, M. Bouchard, M. Bolic, "Low-cost modifications of Rao-Blackwellized particle filters for improved speech denoising", Signal Processing, Volume 88, Issue 11, pp.2678– 2692, November 2008

[PUD'06] H. Puder, "Adaptive signal processing for interference cancellation in hearing aids", Signal Processing, vol. 86, no. 6, pp.1239-1253, June 2006

[ROH'07] T. Rohdenburg, V. Hohmann and B. Kollmeier, "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 2007, pp. 315-318, Oct. 2007

[ROH'08] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays" , Icassp 2008, pp. 2449 - 2452, April 2008

[SIM'01] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in Microphone Arrays: Signal Processing Techniques and Applications, M. S. Brandstein and D. B. Ward, Eds., Chapter 3, pp. 39–60, Springer, Berlin, Germany, 2001.

[SPR'05] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications", IEEE Trans. Speech Audio Process., vol. 13, no. 4, pp 487-503, Jul. 2005

[TEU'07] H. Teutsch, G. W. Elko, "First- and second-order adaptive differential microphone arrays", in Proc. 7th IWAENC, pp. 35-38, Sep. 2007

[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.