Binaural Beamforming with Spatial Cues Preservation

By

Hala As’ad

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 Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa
August 2015

Abstract

In binaural hearing aids, several beamforming algorithms can be used. These beamformers aim to enhance the target speech signal and preserve the binaural cues of the target source (e.g. with constraints on the target). However, the binaural cues of the other directional sources as well the background noise are often lost after processing. This affects the global impression of the acoustic scene, and it limits the perceptual separation of the sources by the hearing aids users. To help the hearing aids users to localize all the sound sources, it is important to keep the binaural cues of all directional sources and the background noise. Therefore, this work is devoted to find the best trade-off between the noise/interferers reduction and the cues preservations not only for the directional interferers but also for the background noise based on selection and mixing processes. In this thesis, some classification decision algorithms, which are based on different criteria such as the power, the power difference, and the coherence, are proposed to complete the selection and mixing processes. Simulations are completed using recorded signals provided by a hearing aid manufacturer to validate the performance of the proposed algorithm under different realistic acoustic scenarios. After detailed testing using different complex acoustic scenarios and different beamforming configurations, the results indicate that some of the proposed classification decision algorithms show good promise, in particular the classification decision algorithm based on coherence.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Professor Martin Bouchard for his continued guidance, invaluable support, and thoughtful suggestions during my whole research period. His passion and dedication to his work always encourage me to excel in my work. Thank you for giving me a chance to be a part of your research team.

I would like to thank Professor Eric Dubois and Professor Richard Dansereau for their interesting courses in Adaptive Signal Processing and Digital Signal Processing. These courses helped me to build solid foundations that made this work possible. I would like also to thank Dr. Homayoun Kamkar-Parsi for his continued technical feedback.

I would like to acknowledge Siemens Audiologische Technik GmbH Group (now Sivantos Inc.), the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Ottawa Admission Scholarship for the financial support that I have received during my Master's studies.

I am very grateful to my parents, my sister Heba, and my brother Omar for their continued support. I owe a debt of gratitude to my father Abdullatif and my mother Lina who always give and have never asked for something in return. I have not ever fully appreciated how difficult it is to be a parent until I became one myself. Thank you for everything!

I would like to thank my husband Na'el who is my main source of support and encouragement. He always stands by me no matter what happens. Thank you for all what you are doing for me. Finally, to the light of my life, my little son Nabeel, this work is dedicated.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ANC	Active Noise Canceller
BLCMV	Binaural Linear Constrained Minimum Variance
BTE	Behind-The-Ear
CIC	Completely-In-the-Canal
DAS	Delay-And-Sum
GEVD	Generalized Eigenvalue Decomposition
GSC	Generalized Side-Lobe Canceller
HRTFs	Head-Related Transfer Functions
IC	Interaural Coherence
ICA	Independent Component Analysis
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITC	In-The-Canal
ITD	Interaural Time Difference
ITE	In-The-Ear
ITF	Interaural Transfer Function
LCMV	Linearly Constrained Minimum Variance
LMS	Least Mean Square
LS	Least-Squares
MMSE	Minimization of the Mean Squared Error
MSC	Magnitude Squared Coherence
MVDR	Minimum Variance Distortion Response
MWF	Multichannel Wiener Filter
MWF-IC	Multichannel Wiener Filter with Interaural Coherence preservation
R1-MWF	Rank-1 Multichannel Wiener Filter (R1-MWF)
RITE	behind-the-ear hearing aid with the Receiver-In-The -Ear
SDR	Signal to Distortion Ratio
SDW-MWF	Speech Distortion-Weighted Multichannel Wiener Filter
SDW-MWF-n	Speech Distortion-Weighted Multichannel Wiener Filter with partial Noise estimation
SNR	Signal to Noise Ratio
SP-MWF	Spatial Prediction Multichannel Wiener Filter
STFT	Short Time Fourier Transform
TDOA	Time Differences Of Arrival
T-F plane	Time-Frequency plane
VAD	Voice Activity Detection

Chapter 1 Introduction

1.1 Motivation and Previous Work

Sensorineural hearing loss, conductive hearing loss, mixed hearing loss and central hearing loss are some of the most common types of hearing loss. However, sensorineural hearing loss is the most common among those types of hearing loss. Aging of the population, exposure to high levels of sound for a long period of time because of the growing use of personal listening devices such as the iPod, and other physiological disorders such as the loss of inner hair cell functions and reduced electrical potential in the cochlea are some of the main reasons for sensorineural hearing loss (Dillon, 2012; Glyde, Cameron, Dillon, Hickson, & Seeto, 2013).

Hearing-impaired people can face a decrease in audibility where they miss the high frequency information, which causes difficulties in speech understanding. Furthermore, decrease in the frequency resolution is another problem that the hearing-impaired people can face which leads to a weakness in the ability to distinguish between the sounds of close frequencies; therefore, sometimes the background noise masks parts of the speech if both of them have close frequencies. Decrease in the temporal resolution and decrease in the dynamic range are other problems that the hearing-impaired people may face (Dillon, 2012). In addition, some of the hearing-impaired people face difficulties in localizing simultaneous sound sources. Several studies have shown the poor discrimination abilities of the binaural cues such as the Interaural Level Difference (ILD), Interaural Phase difference (IPD) and Interaural Time difference (ITD) among people with hearing difficulties. The ILD is based on the intensity differences between the left ear and right ear signals. The IPD and ITD are based on the phase differences and the relative time differences, respectively, between the left ear and right ear signals. Several studies have also reported that the spatial separation abilities of speech and masking noise is smaller for the listener with impaired hearing compared with the normal hearing listener (Celesia & Hickok, 2015).

Consequently, hearing aids with advanced signal processing systems have been developed to overcome some of the hearing deficiencies. Many styles of hearing aids have been introduced

such as Completely-In-the-Canal (CIC), In-The-Canal (ITC), In-The-Ear (ITE), Behind-The-Ear (BTE), and behind-the-ear hearing aid with the Receiver-In-The -Ear (RITE) canal. Various digital signal processing algorithms are used in hearing aids. These algorithms can be divided into four main categories. The first category is the digital amplification algorithms. Hearing loss differs in the shape of the frequency response (the amount of hearing loss across frequencies), for example, they can have flat shapes, sloping shapes, raising shapes, etc. Therefore, there is a necessity to amplify some frequencies more than other frequencies, to compensate the hearing loss of some sounds at specific frequencies and avoid abnormal loudness growth. For example, in some cases, hearing impaired persons receive sounds at high frequencies louder than the normal cases, consequently, less amplification is needed for these sound components. The second category is the speech extraction and noise reduction algorithms using directional microphones or microphone arrays. These speech extraction algorithms are capable to separate the desired signal from a noisy environment with interferers and diffuse noise; in addition, they have the capability to improve the speech intelligibility of the desired signal. The third category is the feedback cancellation algorithms that are responsible of adaptively reducing the acoustic feedback, i.e., hearing aid generated sound being picked up by the hearing aid microphone. Acoustic feedback is generated from many factors such as poorly fit of the instrument case, large vent, high gain, small physical size, etc. Finally, the fourth category is the sound classification techniques that classify acoustic signal into speech, music, or environmental sounds, and then adaptively select the best signal processing algorithms based on the type of the classified acoustic scenario (Schaub, 2008).

Over the years, a lot of work has been devoted to speech enhancement and noise reduction. Single channel speech enhancement is a research field that has been studied for forty years. A speech enhancement algorithm is based on the estimation of a clean speech signal from its noisy environment (Kawamura, Thanhikam, & Iiguni, 2012; Lotter & Vary, 2005). The single channel algorithms have several limitation under high levels of noise, highly non-stationary noise, or when the noise has similar statistics to the desired speech signal. Moreover, typical single channel algorithms have a low or null efficiency in terms of improving the speech intelligibility. Therefore, a lot of research has been conducted in multichannel beamforming algorithms, which use arrays of microphones in order to improve the ability to extract the desired signal, suppress the undesired

noise, and overcome the limitation of the single channel speech enhancement and de-noising algorithms in terms of speech intelligibility.

Multichannel beamforming algorithms, which attempt to coherently combine the desired signal from different microphone signals and attenuate the undesired signals that come from other directions, can be classified into three categories. The first category is the fixed beamforming algorithms such as the work in (Desloge, Rabinowitz, & Zurek, 1997). These kinds of beamforming algorithms are data independent beamformers, with fixed parameters for the processing (assuming static conditions of all sources). The second category is the adaptive beamforming algorithms such as the work in (Welker, Greenberg, Desloge, & Zurek, 1997) which adaptively steer the position of the nulls in order to attenuate the noise as much as possible based on the spatial characteristics of the input noisy signal; therefore, this kind of beamformers is called data dependent algorithms. Finally, multichannel blind source separation is the third category, which uses independent component analysis (ICA) or clustering approaches in order to extract the desired speech signal from the background noise, without any prior knowledge of the source directions (Nesta, Svaizer, & Omologo, 2011; Yu, Hu, & Xu, 2014). The response of the microphone array system, which gives a complete characteristic of the array's input and output behavior, is called the beampattern. The beampattern has a main lobe that allows the desired signal to pass without attenuation, side lobes with smaller gain, and nulls that are responsible of attenuating the undesired signals.

Delay-and-sum (DAS) beamforming is a basic beamforming algorithm, sometimes called "traditional beamforming". This kind of a beamformer consists of two steps. In the first step, the time differences of arrival (TDOA) between the target signal at the reference microphone and the target signal at all other microphones in the array are calculated, and the signal at each microphone is shifted based on the corresponding TDOA to produce time-shifted signals. In the second step, the time-shifted signals are added up in order to calculate the output of the DAS beamformer (Flanagan, Johnston, Zahn, & Elko, 1985) . A filter and sum beamformer is an extension of the DAS beamformer. While the DAS beamformer applies equal amplitude weights to all input signals, the filter and sum beamformer considers that the source components at each microphone have frequency dependent responses. Therefore, frequency dependent amplitude and phase

weights are applied to the input signal at each microphone (McCowan, 2001). The Least-Squares (LS) beamforming algorithm is another optimum approximation approach, which depends on both the direction of the desired signal and the characteristics of the array geometry, and is used to approximate the desired directional response based on the LS criterion as described in details in (Benesty, Chen, & Huang, 2008; Flanagan et al., 1985).

Adaptive beamforming algorithms can potentially produce a better performance than fixed beamforming algorithms in term of the noise reduction since they have the capability to adaptively estimate some of the information of the noise field and the direction of the dominant noise signals, and therefore they adaptively steer the direction of the nulls of the beampattern. In practical scenarios, there is a need to change the direction of the nulls adaptively. Assuming a linear 1-D microphone array with uniform distance between microphones, this can be done by adaptively changing the time delay for the DAS beamformer. However, this approach is not practical because of the need of adjusting the uniform distance between the microphones. Therefore, in (Elko, 1995), an adaptive back-to-back first-order differential microphones algorithm is introduced for a linear 1-D microphone array with uniform distances between microphones and a target in the endfire direction of the array. In this algorithm, a back-to-back combination of arrays with cardioid response has been implemented. The cardioid response is an array system response, which has a beampattern that attenuates (or nulls) the signals coming from a certain angle. The output of Elko's configuration is generated by multiplying the rear-cardioid signal by a factor β, and then subtracting the result from the front-cardioid output. The factor β can be found by minimizing the mean squared error of the back-to-back configuration output using the Least Mean Square (LMS) algorithm. Assuming that we have a 1-D array with two microphones in the endfire direction, the front-cardioid output is generated by delaying the rear microphone signal by d/c, where d is the distance between the two microphones and c is the speed of sound. After that, the delayed rear microphone signal is subtracted from the front microphone signal. The rear-cardioid output is generated by reversing the roles of the microphones, where the front microphone signal is delayed by d/c and then subtracted from the rear microphone signal.

Another beamforming algorithm is the Minimum Variance Distortion Response (MVDR), which was introduced by Capon in (Capon, 1969). The MVDR beamformer is based on a

constrained minimization, where the power of the non-look direction noise signals (i.e. signals outside of the desired signal direction) is minimized while the response of the beamformer in the direction of the desired signal is constrained. In (Frost III, 1972), Frost introduced the Linearly Constrained Minimum Variance (LCMV), which has a similar concept to the MVDR beamformer, but it is based on multiple constraints instead of a single one. In order to calculate the coefficients of either the MVDR beamformer or the LCMV beamformer, the knowledge of the target directivity vector (or target steering vector) is required. At each frequency, the target directivity vector is equivalent to the frequency responses from a far field target source to each microphone. This directivity vector can include the head shadow effects (it then becomes similar to the Head-Related Transfer Functions (HRTFs)), as well as reverberation from the environment. The HRTFs describe the complex changes in the spectrum of the sound before reaching the eardrums. These changes come from some physical effects such as the diffraction and the reflection of the sound that is caused by the human shoulders, torso, and pinnae, and these effects vary from person to person (Kistler & Wightman, 1992).

In (Capon, 1969), it is claimed that since the MVDR beamformer is based on constrained minimization for the desired direction, the power of the total array output signal could be minimized instead of the power of the undesired noise signals alone. This is of interest since the noise signals alone may not be readily available. However, Cox in his work (Cox, 1973) studied the effect of the mismatch on the aforementioned beamformers. The mismatch usually happens when we have inaccurate knowledge of the desired signal directivity vector. Cox demonstrated that when there is mismatch, using the input noisy correlation matrix instead of the noise-only correlation matrix when calculating the weights of the beamformer generates a significant decline in the output Signal to Noise Ratio (SNR). In (Doclo, Gannot, Moonen, & Spriet, 2008), the relationship between the MVDR beamformer and the Minimum Signal to Noise Ratio (MSNR) beamformer based on maximizing the output SNR was found, and it was proven that both MVDR and MSNR are equivalent if the target directivity vectors are known.

In (Griffiths & Jim, 1982), another widely used beamforming structure was introduced; called the Generalized Side-Lobe Canceller (GSC). Under some conditions this beamformer is theoretically equivalent to the LCMV beamformer but it is implemented with two branches (Breed

& Strauss, 2002). The first branch has a fixed beamformer which ensures that the constraints of the LCMV are met (e.g., in the case of a single constraint on the target, then a delay-and-sum beamformer can be used). The second branch consists of a "blocking matrix", which ensures that there is zero content in all the directions corresponding to the constraints of the LCMV (e.g. in the case of a single constraint on the target, the blocking matrix generates outputs where a null has been positioned in the direction of the desired speech signal, and the outputs can then be used as reference signals for the undesired signals, as in (Griffiths & Jim, 1982)). After that, the fixed beamformer output is used as a "desired signal" by a linear predictor, and the blocking matrix outputs are used as the input signals from which the linear prediction is made, resulting in a multichannel Active Noise Canceller (ANC) system. The system attempts to remove the undesired signal components in the fixed beamformer output, which are correlated with the outputs of the blocking matrix.

The Multichannel Wiener Filter (MWF) is another type of speech enhancement and noise reduction technique, which is based on Minimization of the Mean Squared Error (MMSE) between the speech components in the reference microphone and the output of the beamformer. In order to preserve the trade-off between the noise reduction and the speech distortion, the MMSE criteria can be generalized to introduce the speech distortion-weighted SDW-MWF techniques (Doclo et al., 2008). However, the well-known SDW-MWF method needs an accurate estimation of the speech correlation matrix, which is not always readily available in practice, especially in non-stationary environments with competing talkers. Consequently, any error in estimating the speech correlation matrix will reduce the practical performance of the SDW-MWF. In (Cornelis, Moonen, & Wouters, 2009; Cornelis, Moonen, & Wouters, 2011), alternative methods are proposed in order to increase the robustness against the estimation error of the speech correlation matrix, for the common case where the target signal consists of a single speech source. These methods are called the Rank-1 Multichannel Wiener filter (R1-MWF) and the Spatial Prediction Multichannel Wiener filter (SP-MWF). Based on (Cornelis et al., 2011), the SP-MWF provides a better performance than the R1-MWF under low-SNR conditions. The MWF and its extensions do not require prior knowledge of neither the microphones spatial configuration nor the direction of the desired speech signals (i.e., target steering vector). However, in practice they are based on the second order statistics of the noisy signals (estimated during target speech activity) and on the second order

statistics of the noise/interference signals (estimated during target speech inactivity), which may require a sophisticated Voice Activity Detection (VAD) system, especially for cases where the target speech and the interferences have similar spectral statistics with different activity patterns.

A hearing-impaired person often needs one hearing aid on each ear as the hearing loss frequently happens at both ears. Therefore, bilateral hearing aid systems have been introduced, which have a hearing aid on each ear and some limited common control between the ears such as overall gain and user preference settings. In hearing aid applications, there is an increased need to have better speech extraction and noise reduction algorithms in order to improve the speech intelligibility for the hearing aid user in complex acoustic scenarios such as for the cocktail party problem (Bronkhorst & Plomp, 1992; Cherry, 1953), where there can be multiple interference speakers, non-stationary background noise, transient noise, and a reverberant environment. The bilateral hearing aids, however, can struggle in some scenarios since they have a limited number of microphones at each side without significant exchange of signal information between them. As a result, binaural hearing aid models have recently been introduced. These binaural models have a wireless connection between the left and right hearing aids. In general the increased number of microphones and information available to each hearing aid in a binaural hearing aids system allows to achieve better performance than bilateral hearing aids, at least for some scenarios (e.g., target and interferer in the same frontal half-hemisphere). Ideally, either binaural hearing aids or bilateral hearing aids should produce binaural output signals with reduced background noise and interferers, while preserving the binaural spatial cues of the speech target signal. In addition, they should also preserve the binaural spatial cues of the interference signals and the background noise. As a result, hearing aid users would still be able to localize the sound sources (target, interferers, and background noise), which will preserve the naturalness of the perceived acoustic environment and can also protect the hearing aid user from some dangerous situations. For example, while driving a car, hearing aid users could determine the direction of arrival of any incidental sound, so that accidents can be avoided (Doclo et al., 2008; Jingdong Chen & Benesty, 2013). In addition, preserving the binaural cues for the acoustic scene (target, interferers, and background noise) helps to limit the effect of the “unmasking” phenomena, such that our brain can more easily differentiate between sounds coming from different directions

In the design of binaural beamforming or noise reduction algorithms, there is a need to use metrics that can be used to measure the preservations of the spatial cues for the acoustic scene. The most common performance measurements used to evaluate the cues preservation for the acoustic scene at the output of the noise reduction or beamforming algorithm are based on the relative differences between the signals at the two ears on the azimuth plane (Begault, Durand R and others, 1994). These measurements are the Interaural Level Difference (ILD), the Interaural Time Difference (ITD), the Interaural Phase Difference (IPD), and the Magnitude Squared Coherence (MSC). Other measurements can be used to evaluate the cues preservations in the elevation plane. ILD, IPD, and ITD are frequency-dependant measurements that are suitable for directional signals such as the target or interferences (Doclo et al., 2008). Similarly, the MSC is a frequency-dependent measurement; however, it is suitable for the diffuse-like background noise (e.g. babble), as it has different characteristic in comparison with the directional interferers.

Most of the aforementioned beamforming algorithms can be extended to binaural designs. As it was mentioned before, due to the inherent design methods in most beamformers (e.g. distortion-less target response) the binaural spatial cues of the target source are well preserved. However, the binaural spatial cues of the other directional interferers as well as the spatial impression of the diffuse noise are not well preserved. In particular, all the sources after processing can seem to come from the same direction as the target source. Some of the previous methods that have been developed to address this are described below.

In (Welker et al., 1997) , the left and right input signals are divided into low frequency components and high frequency components. The low frequency components are passed without any kind of filtration since the low frequency components are important in preserving the binaural cues for the acoustic scene, especially, in terms of IPD/ITD. On the other hand, the high frequency components are processed through adaptive beamformer, and then they are added back to the low frequency components. However, by passing the unaltered low frequency components, which consists of both the desired speech signal in addition to the noise, the performance of the noise reduction algorithm will be degraded and the best trade-off between the noise reduction and the cues preservation will not be achieved.

In (Benesty & Chen, 2012), the problem of binaural noise reduction is investigated using Multichannel Wiener Filter (MWF) and Minimum Variance Distortion Response (MVDR). In this approach, Benesty and Chen form a single complex signal from the binaural output signals using widely linear estimation theory. Even though the derived beamformer has the capability to reduce the undesired signals and preserve the spatial information of the desired signal, it is not capable to preserve the binaural cues for the noise sources / interferers. In (Chen & Benesty, 2013), Chen and Benesty extended their previous work in order to preserve the spatial perception not only for the desired signal but also for the noise sources / interferers. In the latter work, the widely linear filtering approach has also been used to convert the binaural outputs into a single complex output. After that, the speech and noise output components are each decomposed into two orthogonal components (those showing high coherence between microphones, and those with low coherence between microphones). These components are used in the constraints of a Linearly Constrained Minimum Variance (LCMV) algorithm to reduce the unwanted noise, while at the same time partially preserving the spatial perception of the noise sources. This approach has so far only been tested in an acoustic scenario with one moving target source, one directional noise source, and diffuse noise; therefore, testing this approach under more complex scenarios with multiple interferers would be useful. Furthermore, similar to the MWF and its extensions, this approach needs estimations of the second order statistics of the target-only and the noise-only components, which may not be readily available in real-life scenarios. Moreover, (Szurley, Bertrand, & Moonen, 2013) have shown that using the widely linear filtering method to convert the two real signals into a complex signal doubles the computational complexity in comparison to linear filtering approaches, while at the same time not bringing extra benefits in terms of the speech enhancement over using the linear filtering approaches.

In (Hadad, Gannot, & Doclo, 2012), the LCMV beamformer is extended to a Binaural Linear Constrained Minimum Variance (BLCMV) and implemented as a part of a GSC beamformer. In order to reduce the implementation complexity, shared common blocks between the two beamformers of the hearing aid apparatuses are used. The proposed algorithm has the capability to preserve the binaural cues of the constrained sources by design (i.e., for the target and some interferers); however, it does not have the ability to preserve the spatial perception of unconstrained noise sources and the background diffuse noise. This proposed algorithm is an

extension of the work in (Markovich, Gannot, & Cohen, 2009), which uses Generalized Eigenvalue Decomposition (GEVD) for the correlation matrixes of the desired speech signal, the interference signals, and the background noise to estimate the desired and interference speakers subspace. Therefore, there is a need to determine time segments when firstly the target and interferers are inactive to estimate the background noise subspace. Secondly, there is a need to determine time segments when the desired signal is inactive to estimate the interference signals subspace. Finally, there is a need to determine time segments when the interference signals are inactive to estimate the desired signal subspace. Therefore, this kind of algorithm must rely on the use of a sophisticated VAD system (as in the MWF algorithms previously described), which reduces the capability of these approaches to perform well in real-life complex scenarios. Moreover, the proposed method in (Hadad et al., 2012) requires the steering vector for both the target signal and the interferers, which may be hard to estimate in practical situations.

Other extensions have been introduced to the MWF and its family, which are summarized in (Doclo et al., 2008), in order to preserve the binaural cues not only for the desired signals, but also for the interferers and the background diffuse noise. The binaural SDW-MWF has been extended by modifying the cost function to include the ILD, IPD, or the Interaural Transfer Function (ITF). These approaches perform well for a single lateral noise source under a low reverberant environment and require knowledge of the ITF for the interferer, which is hard to estimate in practice. Furthermore, these approaches don't have the capability to preserve the spatial perception for the background diffuse noise, since the diffuse noise has different characteristic than the directional interferers. As a result, a binaural Multichannel Wiener Filter with Interaural Coherence preservation (MWF-IC) was introduced in (Marquardt, Hohmann, & Doclo, 2014). The latter extension considers the trade-off between the noise reduction and the IC preservation; however, it does not preserves the binaural cues for the directional interferers. A more robust extension has been proposed in (Klasen, den Bogaert, Moonen, & Wouters, 2007), which is basically based on adding a small portion of the original noise to the enhanced signal. This approach is called Speech Distortion weighted Multichannel Wiener Filter with partial noise estimation (SDW-MWF-n). However, the main drawback of MWF remains in these extensions, i.e. they require sophisticated VAD for estimating speech and noise correlation statistics.

In (Lotter & Vary, 2006), a MVDR beamformer is used in a binaural hearing aid and the output of the beamformer is converted to a real-valued common gain to be applied to the input of both sides. The real-valued common gain is computed by taking the ratio of the spectral amplitudes for the beamformer output over the sum of spectral amplitudes of both input noisy signals. Other alternatives to compute the common-gain can be found in (Mustière et al., 2013). This approach is not a "true" or classic beamforming approach since the real common gain is applied individually to each input noisy signal (it is no longer a multichannel filtering, and it no longer attempts to coherently sum a target component from different microphones). The algorithm in (Lotter & Vary, 2006) also combined the outputs of the common gain filtering with a post-filtering consisting of a single channel speech enhancement stage applied as a common gain to each output. This type of speech enhancement post-filtering can in general be applied to any beamforming algorithm introduced so far (although not necessarily in the form of a common gain). It produces a trade-off between the additional noise reduction produced by the speech enhancement post-filtering and the target speech distortion or unnaturalness introduced by the process. Overall, the method in (Lotter & Vary, 2006) can produce a theoretically perfect preservation of the binaural cues for the target, the interferers, and the background noise. If only very low amounts of distortion can be tolerated, the post-filter can either be eliminated or tuned for modest noise reduction. Overall the drawback of the common gain method is that the common gain being applied to the individual noisy inputs can only provide a limited noise/interferers attenuation compared to a "true" beamforming system with multichannel filtering.

In (Thiemann, Muller, & Van De Par, 2014), an algorithm that uses a common binary decision for the left and right outputs in the Time-Frequency (T-F) plane was introduced. Based on the binary decision, the binaural outputs are selected at each T-F bin for example using either the MVDR binaural output of each side or the attenuated noisy signal at the reference microphone of each side. Each MVDR beamformer use a different reference microphone on each side. The method attempts to enhance the target signal, while attenuating the directional interferers and the diffuse noise without changing most of their binaural cues. The binary decision at each T-F bin is based on the domination of the target signal energy vs. the noise/interferers energy at each T-F bin. In the absence of the localized target (in T-F domain) in the beamformer output, the user hears the unmodified noisy signal, except for a gain factor. Three variants of the algorithm were

introduced. For the three variants, the classification is based on comparing the energy of the MVDR output (for the side closer to the target) with the averaged energy of the input noisy signals at the left and right reference microphones, to determine the domination (or not) of the target signal at each T-F bin. In the first variant, the selections are done between the MVDR output on each side (when target dominates) and the attenuated version of the noisy signal at the reference microphone (when target does not dominate). In the second variant, the selections are done between the noisy signals at the reference microphone on each side (when target dominates) and the attenuated version of the noisy signals at the reference microphones (when target does not dominate). In other words, for the T-F bins where the target is classified as dominant, the second variant considers the noisy signals while the first variant considers the beamformer outputs. Finally, the third variant is based on the assumption that the Interaural Time Differences (ITDs) are essential directional cues for the human auditory system. Therefore, the phase modification is important in preserving the spatial location for the acoustic scene. In this variant, in case of target domination in a T-F bin, a phase shift is applied on the output of the MVDR used for the binary decision (from side closer to the target) to generate the other output signal. The applied phase shift is similar in the concept to the ITF filtering and relies on a geometric calculation of the time difference of arrival (TDOA), which depends on the distance between the two ears, the angle of arrival of the target signal, the center frequency, and the speed of sound in the air. Overall, the proposed algorithm in (Thiemann et al., 2014) (and especially the first variant) was found to be promising since it can provide a trade-off between noise reduction and cues preservation, while it does not rely on sometimes quite challenging estimation of statistics or parameters, unlike other methods.

There are several possibilities that have not yet been considered in the work of (Thiemann et al., 2014). The classification decision that is used to distinguish between the domination of the target vs. the noise/interference signals at each T-F is an essential part of the algorithm, and other signals as well as other criteria can be used in order to make this decision. In addition, based on the decision at each T-F bin, a weighted mixture of signals can be selected for the output, based on a wider choice of signals and mixing conditions than the ones considered so far in (Thiemann et al., 2014). Moreover, further studies should be done to assess the quality of the output signals under real-life challenging conditions. In addition to harsh acoustic conditions, this also includes

the impact of using realistic T-F decompositions for hearing aids with a small group delay, which essentially means that small frame sizes need to be used (or equivalently a filter bank with coarse frequency resolution or with fairly large downsampled bandwidths needs to be used). As a result, this thesis aims to further investigate the approach introduced in (Thiemann et al., 2014) in order to find the best trade-off between noise reduction and cues preservations, using real-life data and real-life algorithms used in today's state of the art binaural hearing aids, with data and feedback provided by our sponsor Siemens Audiologische Technik GmbH Group (now Sivantos Inc.).

1.2 Objectives and Organization

The objective of this thesis is to find the best trade-off between cues preservations of the acoustic scene and the noise/interferers reduction when beamforming algorithms are used in binaural hearing aids. To achieve this objective, the outputs of the beamforming algorithms such as MVDR or GSC, which have a good noise/interferers reduction, will be considered. In addition, the original noisy signals and the outputs of other processing algorithms which preserves most of the noise/interferers cues (e.g. common gain approach) will be also considered for the mixing and for making the selection/decision in each T-F bin (or more precisely each subband-time sample). At each subband-time sample, the best mixture of the available signals will thus be selected based on some selection criteria to generate outputs that preserve most of the spatial cues for the acoustic scene, and at the same time, have good noise/interferers reduction. The selections can be frequency dependent or independent, and based on several possible criteria such as the power, the power difference, the coherence, etc. The best possible criteria using the appropriate mixture of the available signals will be tested under different complex acoustic scenarios. Finally, some emphasis will also be put on comparing the performance of the best-selected algorithm using different configurations of microphones or beamformer designs. The evaluation of the proposed algorithm will be based on several frequency dependent metrics and some informal listening.

To achieve these objectives, this thesis is organized as the following:

Chapter 2 provides a review of the beamforming algorithms used in this thesis. In addition, this chapter gives detailed descriptions of the different configurations for the beamforming design and the system notation.

Chapter 3 describes the performance metrics that are used for evaluation. These metrics are classical metrics to measure the noise reduction and the speech distortion, and cues preservations metrics.

Chapter 4 provides a detailed explanation for the binaural beamforming algorithms that are based on the selection and mixing between some available signals, aimed to reduce the noise/interferers and at the same time preserve the cues for the acoustic scene.

Chapter 5 presents simulation results to evaluate the cues preservation metrics for two beamforming algorithms. One of these algorithms preserves the cues while the other algorithm changes most of the cues (except for the target).

Chapter 6 evaluates all the proposed selection and mixing algorithms in terms of cues preservations and noise/interferers reduction, and chooses the approaches that demonstrate the best trade-off between noise reduction and cues preservations for the acoustic scene. Moreover, this chapter presents the capability of the best-selected approaches in preserving the diffuse noise by putting more emphasis on the background noise. In addition, the best-selected approaches are evaluated under different acoustic scenarios and using different beamforming configurations.

Chapter 7 provides the thesis conclusions and some suggestions for possible improvements from the work of this thesis.

1.3 Contributions

This work is an extension to the work in (Thiemann et al., 2014). However, in this thesis we have introduced the following main contributions:

The work in (Thiemann et al., 2014) was based on a classification which attempts to detect target-dominated segments versus noise-dominated segments based on level information.

However, the classification decision is not robust under low SNR scenarios during active target frames. In such cases, a beamformer output having lower level than the average noisy input signal level for a given T-F bin (strong noise reduction during an active target frame) would be detected as a noise-only condition, leading to a mixing scenario of attenuated (leveled-down) input, which is unacceptable since it would reduce the target level (loudness) and not improve SNR. In our work, the mixing varies based on a classification decision which attempts to determine the T-F bins where the noise reduction is strong, versus the T-F bins where the noise reduction is weak. When the noise reduction is weak, signals that have more cues preservation are given more weight in the output mixtures. When the noise reduction is strong, signals that have more noise reduction are given more weight in the output mixtures (even if they have less cues preservation).

In (Thiemann et al., 2014), only the beamformer outputs and the noisy signals at the reference microphones have been considered for the classification, selection and mixing. However, in this thesis the outputs of a common gain approach (Lotter & Vary, 2006), which have a moderate noise reduction and preserve most of the cues of the acoustic scene (described in section 2.4), have also been considered for the classification, selection and mixing.

In (Thiemann et al., 2014), the classification decision that is used to distinguish between the domination of the desired target and the undesired noise/interference signals at each T-F is based only on comparing the energy of the output of the beamformer with the averaged energy of the noisy signals at the reference microphones on both sides. However, in this thesis, we used several classification decisions that are based on the power of the available signals, the power difference between the available signals, and the coherence between the available signals as it is explained in chapter 4. Moreover, in one of the best selected cases which is based on the coherence, two classification decisions have been introduced: one to select the magnitude of the output signals and the other to select the phase of the output signals.

In (Thiemann et al., 2014), left and right beamformers on each side with their own local reference microphone have been considered for mixing. However, for the classification decision, the output of the beamformer on the side that is closer to the direction of the target is considered. Therefore, the classification decision requires a knowledge of the direction of arrival for the target

signal. In this thesis, the classification decision can be based on either one or both beamformers outputs, and both the case of beamformers with the same common reference microphone and the case with each beamformer having its own reference microphone on its side are considered.

In this thesis, the option of keeping the phase components, the magnitude components or both components have been considered during the selection and mixing. In (Thiemann et al., 2014), both components were always processed jointly. Moreover, some of the proposed approaches are frequency dependent, where different processing (i.e. mixing and selection algorithms) are done on the low frequency components of the available signals (e.g. considering that the role of IPD/ITD is mostly up to 1500 Hz), as discussed in section 4.3. This was not found in the previous work of (Thiemann et al., 2014).

In (Thiemann et al., 2014), the beamforming algorithm considered was always the MVDR algorithm. However, in this work, we have not only considered the use of the MVDR beamformer but also the use of the GSC beamformer. The GSC beamformer is a more real-life algorithm that does not require an explicit estimation and inversion of any correlation matrix, and it has an adaptive ANC component, which enables the beamformer to deal with more complex and realistic acoustic scenarios.

In this work, different configurations for the beamforming algorithms with either a single signal or two signals transmitted in each direction through the binaural wireless link have been introduced and tested. Therefore, the best trade-off has been evaluated between the bandwidth and the power consumption on one hand and the cues preservations of the acoustic scene, the noise/interferers reduction, and the target distortion on the other hand.

In this thesis, the best-selected approaches have been tested under different acoustic scenarios with frontal / non-frontal target and with single / multiple interferers, using real-life recordings provided by our sponsor Siemens Audiologische Technik GmbH Group (now Sivantos Inc.).

In this work, an emphasis has been put to investigate the capability of the proposed selection algorithms not only to preserve the directional interferers but also to preserve the background noise.

In order to achieve our objectives MATLAB code has been developed. This code supports the MVDR/GSC beamformers in section 2.2 and section 2.3, the common gain approach in section 2.4, and the selected cases that show the best trade-off between noise/interferers reduction and cues preservation in Chapter 6. Moreover, the developed MATLAB code also supports different beamforming configurations, which are explained in 2.5.

To evaluate the performance of the proposed algorithms, performance metrics (described in Chapter 3) have been implemented to measure the cues preservation of the interferers and the background noise, the noise/interferers reduction, and the speech distortion of the target signal. A method has been introduced to calculate the IPD error for the directional interferers in order to solve the problem of the phase ambiguity that is usually generated from phase subtraction. In addition, several approaches have been evaluated for the ITD calculations.

Chapter 2 Overview of Directional Beamforming Algorithms and their Different Configurations

2.1 Definitions and System Notation

Binaural hearing aids will be considered where we will have two microphone arrays with *M* microphones at ear and a wireless links between them. These hearing aids are in noisy and reverberant environment. The left and the right input noisy microphone signals in the frequency domain after a Short-Time Fourier Transform (STFT) can be written, respectively, as in equations (2.1) and (2.2);

$$y_{m,l}(f,t) = x^{in}_{m,l}(f,t) + v^{in}_{m,l}(f,t) + n^{in}_{m,l}(f,t), \quad m = 1,..,M \tag{2.1}$$

$$y_{m,r}(f,t) = x^{in}_{m,r}(f,t) + v^{in}_{m,r}(f,t) + n^{in}_{m,r}(f,t), \quad m = 1,..,M \tag{2.2}$$

$x^{in}_{m,l}$: Target speech components at the left m^{th} microphone.

$x^{in}_{m,r}$: Target speech components at the right m^{th} microphone.

$v^{in}_{m,l}$: Sum of directional interference components at the left m^{th} microphone.

$v^{in}_{m,r}$: Sum of directional interference components at the right m^{th} microphone.

$n^{in}_{m,l}$: Background noise components (diffuse-like noise) at the left m^{th} microphone.

$n^{in}_{m,r}$: Background noise components (diffuse-like noise) at the right m^{th} microphone.

f : the frequency index

t : the time frame index

In beamforming calculations, either the front left or the front right microphone can be used as a reference channel. If the front left microphone is used as a reference, we will use $y_{ref,l,}\ x_{ref,l}, v_{ref,l}$, and $n_{ref,l}$ to refer to the noisy input signal and its components: the target component, the sum of directional interferers component, and the background noise component, respectively.

The "*f*" frequency index and the "*t*" time index will be omitted from now on for simplicity. If a filter bank is used for the frequency decomposition, "*f*" can also refer to the center frequency of a subband within the filter bank. All microphone signals will be stacked in 2*M* dimensional vectors. Therefore, the input signals to the left and right microphone arrays can be written as in equation (2.3).

$$\mathbf{y} = \mathbf{x} + \mathbf{v} + \mathbf{n} \tag{2.3}$$

Where, $\mathbf{y} = [y_{1,l}, y_{2,l}, ..., y_{M,l}, y_{1,r}, y_{2,r}, ..., y_{M,r}]^T$

The desired target components at the 2*M* microphones can be written in term of the directivity vector as equation (2.4) shows:

$$\mathbf{x}^{in} = \mathbf{d}(\theta_s)\mathrm{s}_x(\theta_s) \tag{2.4}$$

s_x: Target source signal at the far field plane.

$\mathbf{x}^{in}$: Target components at the 2*M* microphones which is equal to $[x_{1,l}^{in}, x_{2,l}^{in}, ..., x_{M,l}^{in}, x_{1,r}^{in}, x_{2,r}^{in}, ..., x_{M,r}^{in}]^T$

$\mathbf{d}(\theta_s)$: Target directivity vector, which is the frequency response between the far field target source and each microphone. This directivity vector includes the head shadow effect; therefore, it is similar to HRTFs.

The target signal at the reference microphone can also be defined as in equation (2.5):

$$x_{ref} = d_{ref}(\theta_s)\mathrm{s}_x(\theta_s) \tag{2.5}$$

Where;

$d_{ref}(\theta_s)$: Directivity vector at the reference microphone, which is the frequency response from the far field desired source to the reference microphone (usually front left or front right microphone).

From now on, x_{ref} and $d_{ref}(\theta_s)$ will be used in the general cases where either the front left microphone or the front right microphone can be used as a reference. However, in case it is necessary to clarify whether the front left or front right microphone is used, we will use $x_{ref,l}$ and $d_{ref,l}(\theta_s)$ if the front left microphone is used as a reference or $x_{ref,r}$ and $d_{ref,r}(\theta_s)$ if the front right microphone is used as a reference.

The correlation matrix for the desired target components can be defined as in equation (2.6):

$$\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^H\} = E\{\mathbf{d}\mathrm{s}_x\mathbf{d}^H s_x^*\} = \mathbf{d}\mathbf{d}^H E\{|\mathrm{s}_x|^2\} \tag{2.6}$$

The superscript *H* refers to "Hermitian" which is the complex conjugate transpose.

Similarly, the correlation matrices for the sum of the directional interference components, and the background noise components, respectively, can be defined as in equations (2.7) and(2.8).

$$\mathbf{R}_v = E\{\mathbf{v}\mathbf{v}^H\} \tag{2.7}$$

$$\mathbf{R}_n = E\{\mathbf{n}\mathbf{n}^H\} \tag{2.8}$$

By assuming that the target components, the sum of directional interference components, and the background noise components are uncorrelated, the correlation matrix for the input signal can be written as in equation (2.9):

$$\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v + \mathbf{R}_n \tag{2.9}$$

In equations (2.10) and (2.11) below, we perform the design of two beamformers. One beamformer attempts to extract the target signal as received at a front left microphone (i.e., using the front left microphone as a reference) and generates a left beamformer output, while the other beamformer attempts to extract the target signal as received at a front right microphone (i.e., using the front right microphone as a reference) and generates the right beamformer output. The output of the left and right beamformers, respectively, will be written as in equations (2.10) and (2.11):

$$z_l = \mathbf{w}_l^H \mathbf{y} = x_l^{out} + v_l^{out} + v_l^{out} \tag{2.10}$$

$$z_r = \mathbf{w}_r^H \mathbf{y} = x_r^{out} + v_r^{out} + n_r^{out} \tag{2.11}$$

z_r : Output of the beamformer when the front right microphone is a reference

z_l : Output of the beamformer when the front left microphone is a reference

$\mathbf{w}_l$: Beamformer coefficients when front left microphone is a reference, with dimension of $2M$x1

$\mathbf{w}_r$: Beamformer coefficients when front right microphone is a reference, with dimension of $2M$x1.

2.2 Basics of Minimum Variance Distortion Response (MVDR) Design

Minimum Variance Distortion Response (MVDR), which is introduced in (Capon, 1969), is based on a constrained minimization of the power of the undesired signals, where the response of the beamformer in the direction of the target signal is constrained to be equal to the desired source signal $s_x(\theta_s)$. Therefore, the constrained minimization problem can be written as in equation (2.12).

$$\min_{\mathbf{w}} \mathbf{w}^H (\mathbf{R}_v + \mathbf{R}_n)\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{d}(\theta_s) = 1 \tag{2.12}$$

Assuming that,

$$\mathbf{R}_{undesired} = \mathbf{R}_{v} + \mathbf{R}_{n} \tag{2.13}$$

to solve the constrained minimization problem in (2.12), a Lagrangian cost function is defined as in equation (2.14).

$$\mathrm{L}(\mathbf{w},\mathbf{w}^{H}) = \mathbf{w}^{H}\mathbf{R}_{undesired}\mathbf{w} + \lambda(\mathbf{w}^{H}\mathbf{d}(\theta_{s}) - 1) \tag{2.14}$$

By taking the gradient of $\mathrm{L}(\mathbf{w},\mathbf{w}^{H})$ with respect to $\mathbf{w}^{H}$ and treating $\mathbf{w}$ as a constant (Wirtinger's calculus):

$$\nabla L(\mathbf{w},\mathbf{w}) = \mathbf{R}_{undesired}\mathbf{w} + \lambda\mathbf{d}(\theta_{s}) \tag{2.15}$$

Therefore,

$$\mathbf{w} = -\lambda\mathbf{R}_{undesired}^{-1}\mathbf{d}(\theta_{s}) \tag{2.16}$$

By substituting equation (2.16) into the constraint in equation (2.12);

$$(-\lambda\mathbf{R}_{undesired}^{-1}\mathbf{d}(\theta_{s}))^{H}\mathbf{d}(\theta_{s}) = 1 \tag{2.17}$$

Then;

$$\lambda = (\mathbf{d}(\theta_{s})^{H}\mathbf{R}_{undesired}^{-1}\mathbf{d}(\theta_{s}))^{-1} \tag{2.18}$$

After plugging equation (2.18) into equation (2.16), the solution can be written as in equation (2.19):

$$\mathbf{w} = \frac{\mathbf{R}_{undesired}^{-1}\mathbf{d}(\theta_{S})}{\mathbf{d}^{H}(\theta_{S})\mathbf{R}_{undesired}^{-1}\mathbf{d}(\theta_{S})} \tag{2.19}$$

In practice, a regularization term $\mu\mathbf{I}$ (or diagonal loading term) normally needs to be used to avoid the inversion ill-conditioning. Consequently, the derived MVDR weights would be written as in equation (2.20) :

$$\mathbf{w} = \frac{(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)}{\mathbf{d}^H(\theta_S)(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)} \tag{2.20}$$

Please note that equation (2.20) is a general form when a single beamformer could use either the front left or the front right microphone as a reference. In addition, the form of (2.20) is applicable regardless of the geometry of the microphones (i.e. can be used for other geometries such as 1-D linear array, 2-D regular or irregular array, 3-D arrays, in either free field or with head shadows, etc.).

Using the "normal" MVDR design with the constraint $\mathbf{w}^H\mathbf{d}(\theta_s) = 1$ preserves the level and the phase of the target component at the beamformer output compared to the far field source as in equation (2.21):

$$y_x = \mathbf{w}^H\mathbf{x} = \mathbf{w}^H\mathbf{d}(\theta_s)s_x(\theta_s) = s(\theta_s) \tag{2.21}$$

However, in realistic acoustic scenarios the aim should be to keep the level and phase of the target components at the output of the beamformer so that they are the same as the target components at the reference microphone $y_x = x_{ref}$. Thus, a different constraint should be considered as in equation (2.22):

$$\mathbf{w}^H\mathbf{d}(\theta_s) = d_{ref}(\theta_s) \tag{2.22}$$

By using the constraint in equation (2.22) instead of (2.12), the coefficients of the "normal" MVDR beamformer $\mathbf{w}$ should be scaled (multiplied) by the reference directivity vector for the target sources $d_{ref}(\theta_s)$ to generate "normalized" MVDR coefficients. Or, alternatively, the directivity vector $\mathbf{d}(\theta)$ should all be normalized by $d_{ref}(\theta_s)$ before the MVDR design so that the resulting $d_{ref}(\theta_s)$ would be 1.0 after normalization.

For binaural hearing aids, the previous beamforming design is extended into two beamformer designs. The first beamformer uses the front left microphone as a reference

microphone and we will refer to its coefficients as $\mathbf{w}_l$, while the second beamformer design uses the front right microphone as a reference microphone and we will refer to its coefficients as $\mathbf{w}_r$. The constrained optimization problem, thus, can be written as in equations (2.23) and (2.24):

$$\min_{\mathbf{w}_l} \mathbf{w}_l^H(\mathbf{R}_{undesired})\mathbf{w}_l \quad \text{subject to} \quad \mathbf{w}_l^H\mathbf{d}(\theta_s) = 1 \tag{2.23}$$

$$\min_{\mathbf{w}_r} \mathbf{w}_r^H(\mathbf{R}_{undesired})\mathbf{w}_r \quad \text{subject to} \quad \mathbf{w}_r^H\mathbf{d}(\theta_s) = 1 \tag{2.24}$$

Consequently, the "normal" MVDR coefficients for both beamformers will be written as the following:

$$\mathbf{w}_l = \frac{(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)}{\mathbf{d}^H(\theta_S)(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)} \quad \left(= \frac{(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}_l(\theta_S)}{\mathbf{d}_l{}^H(\theta_S)(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}_l(\theta_S)}\right) \tag{2.25}$$

$$\mathbf{w}_r = \frac{(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)}{\mathbf{d}^H(\theta_S)(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}(\theta_S)} \quad \left(= \frac{(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}_r(\theta_S)}{\mathbf{d}_r{}^H(\theta_S)(\mathbf{R}_{undesired} + \mu\mathbf{I})^{-1}\mathbf{d}_r(\theta_S)}\right) \tag{2.26}$$

As earlier, in order to ensure that the phase and the level of the target components at the output of the beamformer are similar to the target components at the reference microphone signal, the constraint minimization problem becomes as in (2.27) and (2.28):

$$\min_{\mathbf{w}_l} \mathbf{w}_l^H(\mathbf{R}_{undesired})\mathbf{w}_l \quad \text{subject to} \quad \mathbf{w}_l^H\mathbf{d}_l(\theta_s) = d_{ref,l} \tag{2.27}$$

$$\min_{\mathbf{w}_r} \mathbf{w}_r^H(\mathbf{R}_{undesired})\mathbf{w}_r \quad \text{subject to} \quad \mathbf{w}_r^H\mathbf{d}_r(\theta_s) = d_{ref,r} \tag{2.28}$$

Therefore, the directivity vector $\mathbf{d}(\theta_s)$ is normalized by $d_{ref,l}(\theta_s)$ or $d_{ref,r}(\theta_s)$ before the MVDR design.

As mentioned in (Capon, 1969), the noisy correlation matrix $\mathbf{R}_y$ can be used instead of the correlation matrix of the undesired sources $\mathbf{R}_{undesired}$ as long as the desired response is constrained

to find the beamformer coefficients $\mathbf{w}_l$ and $\mathbf{w}_r$. In practical hearing aids applications, it may be difficult to compute $\mathbf{R}_{undesired}$ as it may require a sophisticated intelligent online VAD system, so the use of the noisy correlation matrix $\mathbf{R}_y$ can be attractive. However, using the noisy correlation matrix increases the sensitivity to the mismatch between the estimated target directivity vector, which is used in the MVDR design, and the actual directivity vector corresponding to the processed data (Cox, 1973).

The problem in equations (2.27) and (2.28) can be reformulated to create the Linearly Constrained Minimum Variance (LCMV) beamformer, introduced in (Frost III, 1972), as in (2.29) and (2.30). The LCMV is similar to the MVDR as both of them aim to extract the desired signal which is coming from specific direction(s) while minimizing the undesired signals from other directions.

$$\min_{\mathbf{w}} \mathbf{w}^H(\mathbf{R})\mathbf{w} \quad \text{subject to} \quad \mathbf{C}^H\mathbf{w} = \mathbf{f} \tag{2.29}$$

where **C** is a matrix including the directivity vector of each constraint direction and **f** is a vector including the conjugate of the desired gains for each constraint direction. Using the complex Lagrangian multiplier to solve (2.29) as in equations (2.14) to (2.18) we obtain:

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{C}\ [\mathbf{C}^H\mathbf{R}\mathbf{C}]^{-1}\ \mathbf{f} \tag{2.30}.$$

In the case of binaural beamformer coefficients, (2.30) would be computed separately for $\mathbf{w}_l$ and $\mathbf{w}_r$.

2.3 Basics of Generalized-Side Lobe Canceller (GSC) Design

As explained in the introduction, the Generalized-Side Lobe Canceller (GSC) (Griffiths & Jim, 1982) can be considered as an implementation of the LCVM, or in the case of a single constraint applied on the target direction, an implementation of the MVDR. An advantage of the GSC is that it does not require an explicit estimation of any correlation matrix, nor any matrix inversion. Also,

the ANC part of the algorithm is adaptive, which allows to deal naturally with moving interferers and non-stationary background noise conditions.

Figure 2.1 below shows the resulting structure for the case of a single constraint, where the fixed beamformer $\mathbf{w_0}$ needs to preserve the single constraint in the direction of the target and the blocking matrix **B** needs to zero the components from the direction of the single constraint (i.e., the target direction).

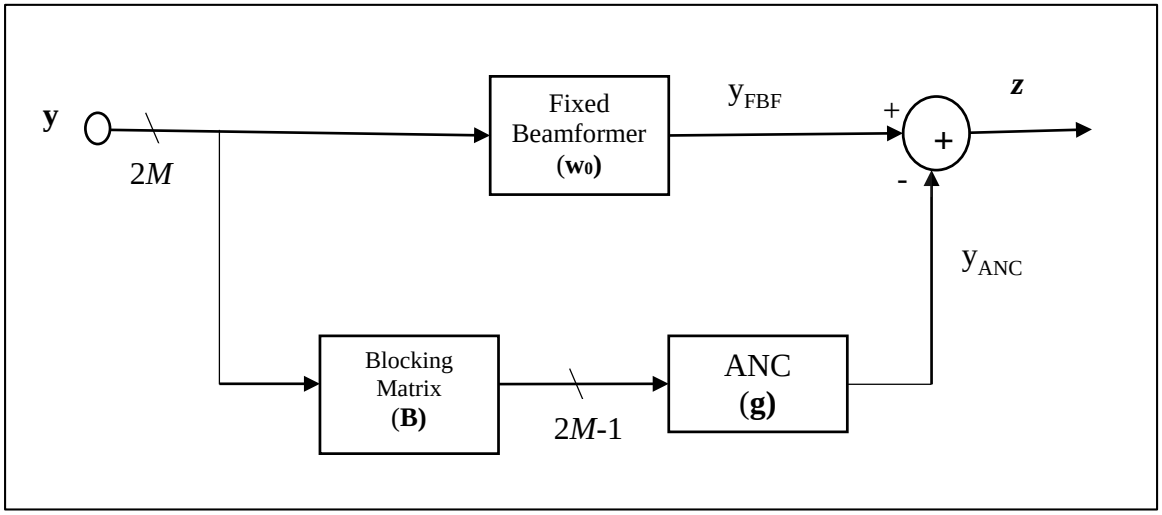

Figure 2.1: Generalized-Side Lobe Canceller

Assuming that we have *M* microphones at each side of the ear, so the total number of microphones is 2*M*, Figure 2.1 shows the structure of the GSC beamformer. The upper branch has a fixed beamformer, which is a classical non-adaptive beamformer such as a delay and sum beamformer, with a distortion-less response in the target direction. In the lower path of Figure 2.1, the blocking matrix creates outputs where a null has been steered in the direction of the target signal, creating 2*M-1* linearly independent estimates of the interferers and the background noise components or the so-called noise references.

After the blocking matrix, the Active Noise Canceller (ANC) system is an adaptive filtering system used to minimize the noise and interferer components at the output of the fixed beamformer, using the correlation with the noise and interference components in the 2*M-1* outputs

of the blocking matrix. Consequently, the filter **g** is adaptively adjusted to minimize the power of the output *z*.

2.4 Conversion to a Common Binaural Gain

Using classic beamforming algorithms for hearing aids on both sides of the head does not preserve the spatial cues of interfering directional sources and the background noise. In fact, it does not even preserve the frequency magnitude shape of the interfering directional sources and the background noise on each side. Therefore, methods to find trade-offs between cues preservation and noise/interference reduction have been investigated. One of these approaches is the common gain approach that is proposed in (Lotter & Vary, 2006). In Lotter and Vary's work, the MVDR beamformer is used with one microphone on each side of the ear, and one of the sides is used as the reference microphone side for the MVDR. The MVDR beamformer generates a monaural output, which minimizes the undesired signals while not distorting the target signal. If the same monaural output is played at both ears, this would significantly change the spatial cues, as everything (target, interferers and the background noise) would seem to come from the front. Instead, a common gain $G(f,t)$ can be applied to the signals from both sides of the head at the front left and front right microphone signal as it is shown in Figure 2.2, preserving the binaural cues of all components.

The design in Figure 2.2 is a general design for any kind of symmetric binaural array configuration such as 2+2 structure or 1+1 structure. These structures will be described in details in the next section.

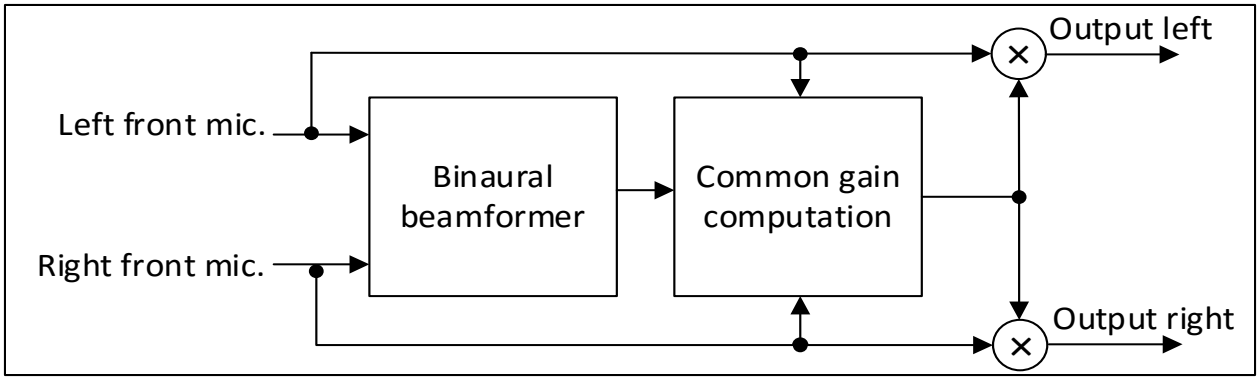

Figure 2.2: Common gain beamformer

The gain $G(f,t)$ should have positive real values, so that the phase response is zero, the group delay is zero, and no dispersion is introduced. Furthermore, the received target components at the front left and the front right microphones should be kept at the same level in the output signals. The front left and front right microphone signals are scaled by the gain $G(f,t)$ to generate the binaural signals $\hat{y}_{x,l}$ and $\hat{y}_{x,r}$. The left and right outputs in Figure 2.2 can be calculated as in equations (2.31) and (2.32)

$$\hat{y}_l = G\, y_{1,l} \tag{2.31}$$

$$\hat{y}_r = G\, y_{1,\mathrm{r}} \tag{2.32}$$

Where the frequency index f and the time index t have been omitted for simplicity.

The gain G is obtained from some normalized form of the beamformer output signal in order to force a unity gain for the target components. To explain this, the common gain values can first be written as in equation (2.33);

$$G = \frac{|z|}{f(\mathbf{y})} \tag{2.33}$$

where $f(\mathbf{y})$ is a real-valued function. In this section, $\mathbf{y}$ is equal to $[\mathrm{y}_{1,l}, \mathrm{y}_{1,r}]^T$.

In (Lotter & Vary, 2006)**,** the values of the gain G are calculated by taking the ratio between the spectral amplitudes of the beamformer output over the sum of the spectral amplitudes of the noisy signals at the front left and front right microphones as in equation (2.34):

$$G = \sum_{k=l,r} \frac{|z|}{f(\mathrm{y}_k)} = \frac{|z|}{|\mathrm{y}_{1,l}| + |\mathrm{y}_{1,r}|} \tag{2.34}$$

As it was mentioned previously, after applying the gain G the target components at the binaural outputs must have the same level and phase as the target components at the front left and front right microphones signals. Consequently, the gain values for the target components should be as in equation (2.35).

$$G_x = \frac{\left|x^{out}\right|}{|x_{1,l}^{in}| + |x_{1,r}^{in}|} = 1 \tag{2.35}$$

Therefore,

$$\left|x^{out}\right| = |x_{1,l}^{in}| + |x_{1,r}^{in}| \tag{2.36}$$

$$\left|\mathbf{w}^H \mathbf{d}(\theta_s)\right| \left|s_x(\theta_s)\right| = |x_{1,l}^{in}| + |x_{1,r}^{in}| \tag{2.37}$$

$$\left|\mathbf{w}^H \mathbf{d}(\theta_s)\right| = \frac{f(\mathbf{x}^{in})}{\left|s_x(\theta_s)\right|} = \frac{|x_{1,l}^{in}| + |x_{1,r}^{in}|}{\left|s_x(\theta_s)\right|} = \frac{|d_{1,l}(\theta_s)s_x(\theta_s)| + |d_{1,r}(\theta_s)s_x(\theta_s)|}{|s_x(\theta_s)|} \tag{2.38}$$

where $\mathbf{x}^{in} = [x_{1,l}^{in}, x_{1,r}^{in}]^T$ and $d_{1.l}(\theta_s)$ and $d_{1.r}(\theta_s)$ are the directivity vectors for the target components between the target source signal and the front left and front right microphone signals, respectively.

As a results, the constraint that is required in the MVDR minimization problem becomes as in equation (2.39).

$$\left|\mathbf{w}^H\mathbf{d}(\theta_s)\right| = |\, d_{1,l}(\theta_s)\,| + |\, d_{1,r}(\theta_s)\,| \tag{2.39}$$

Considering that the constraint for the "normal" MVDR beamformer is $\mathbf{w}^H\mathbf{d}(\theta_s) = 1$, the output of the beamformer z should be scaled by $|\, d_{1,l}(\theta_s)\,| + |\, d_{1,r}(\theta_s)\,|$ before applying the common gain, or more generally the output of the beamformer should be multiplied by the scalar $\frac{f(\mathbf{x}^{in})}{|s(\theta_s)|}$, to ensure a unity gain for the target components after applying the common gain. However, if a "normalized" MVDR beamformer with the constraint $\mathbf{w}^H\mathbf{d}(\theta_s) = d_{ref}(\theta_s)$ is used (to keep the level and phase of the reference microphone target component at the output of the beamformer) then the output of the "normalized" beamformer should be scaled by the scalar $\frac{|\, d_{1,l}(\theta_s)\,| + |\, d_{1,r}(\theta_s)\,|}{|\, d_{ref}(\theta_s)\,|}$.

Similarly, if the MVDR is replaced by a GSC implementation, the output of the fixed beamformer should also be scaled by $\frac{|\, d_{1,l}(\theta_s)\,| + |\, d_{1,r}(\theta_s)\,|}{|\, d_{ref}(\theta_s)\,|}$, assuming that the GSC fixed beamformer is preserving the level and phase of the reference microphone target component at the output.

Several choices of the gain values *G* are presented in (Mustière et al., 2013). These choices are based on different statistical criteria. The following simple examples are not statistically derived but they illustrate different approaches that can be used for the common gain calculations:

1. $$G = \frac{|z|}{f(\mathbf{y})} = \frac{|z|}{\sum_{k=l,r} |\, y_k\,|} \tag{2.40}$$

   The corresponding term that is used to scale the output of the "normal "beamformer is

$$\frac{f(\mathbf{x}^{in})}{|s(\theta_s)|} = \frac{\sum_{k=l,r} |x_{1,k}(\omega)|}{|s(\theta_s)|} = \frac{\sum_{k=l,r} |d_{1,k}(\omega,\theta_s)|\,|s(\theta_s)|}{|s(\theta_s)|} = \sum_{k=l,r} |d_{1,k}(\theta_s)| \tag{2.41}$$

For the "normalized' beamformer, the scale factor becomes $\dfrac{\sum_{k=l,r}\left|d_{1,k}(\theta_s)\right|}{|\,d_{ref}(\theta_s)\,|}$

2. $$G=\frac{|z|}{f(\mathbf{y})}=\frac{|z|}{\sqrt{\sum_{k=l,r}|\,y_k\,|^2}} \tag{2.42}$$

The corresponding term that is used to scale the output of the "normal" beamformer is

$$\frac{f(\mathbf{x}^{in})}{\left|s(\theta_s)\right|}=\frac{\sqrt{\sum_{k=l,r}\left|x_{1,k}(\omega)\right|^2}}{\left|s(\theta_s)\right|}=\frac{\sqrt{\sum_{k=l,r}\left|d_{1,k}(\omega,\theta_s)\right|^2}\left|s(\theta_s)\right|}{\left|s(\theta_s)\right|}=\sqrt{\sum_{k=l,r}\left|d_{1,k}(\theta_s)\right|^2} \tag{2.43}$$

For the "normalized' beamformer, the scale factor becomes $\dfrac{\sqrt{\sum_{k=l,r}\left|d_{1,k}(\theta_s)\right|^2}}{|\,d_{ref}(\theta_s)\,|}$

3. $$G=\frac{|z|}{f(\mathbf{y})}=\frac{|z|}{\sqrt{\prod_{k=l,r}|\,y_k\,|^2}} \tag{2.44}$$

The corresponding term that is used to scale the output of the "normal" beamformer is

$$\frac{f(\mathbf{x}^{in})}{\left|s(\theta_s)\right|}=\frac{\sqrt{\prod_{k=l,r}\left|x_{1,k}(\omega)\right|^2}}{\left|s(\theta_s)\right|}=\frac{\sqrt{\prod_{k=l,r}\left|d_{1,k}(\omega,\theta_s)\right|^2}\left|s(\theta_s)\right|}{\left|s(\theta_s)\right|}=\sqrt{\prod_{k=l,r}\left|d_{1,k}(\theta_s)\right|^2} \tag{2.45}$$

For the "normalized' beamformer, the scale factor becomes $\dfrac{\sqrt{\prod_{k=l,r}\left|d_{1,k}(\theta_s)\right|^2}}{|\,d_{ref}(\theta_s)\,|}$

4. $$G=\frac{|z|}{f(\mathbf{y})}=\frac{|z|}{\max_{k=l,r}|\,y_k\,|} \tag{2.46}$$

The corresponding term that is used to scale the output of the "normal" beamformer is

$$\frac{f(\mathbf{x}^{in})}{\left|s(\theta_s)\right|}=\frac{\max_{k=l,r}\left|x_{1,k}(\omega)\right|}{\left|s(\theta_s)\right|}=\frac{\max_{k=l,r}\left|d_{1,k}(\omega,\theta_s)\right|\left|s(\theta_s)\right|}{\left|s(\theta_s)\right|}=\max_{k=l,r}\left|d_{1,k}(\theta_s)\right| \tag{2.47}$$

For the "normalized' beamformer, the scale factor becomes $\dfrac{\max_{k=l,r}\left|d_{1,k}(\theta_s)\right|}{|\,d_{ref}(\theta_s)\,|}$

5. $$G=\frac{|z|}{f(\mathbf{y})}=\frac{|z|}{\min_{k=l,r}|\,y_k\,|} \tag{2.48}$$

The corresponding term that is used to scale the output of the "normal" beamformer is

$$\frac{f(\mathbf{x}^{in})}{|s(\theta_s)|} = \frac{\min_{k=l,r}|x_{1,k}(\omega)|}{|s(\theta_s)|} = \frac{\min_{k=l,r}|d_{1,k}(\omega,\theta_s)||s(\theta_s)|}{|s(\theta_s)|} = \min_{k=l,r}|d_{1,k}(\theta_s)|$$

For the "normalized' beamformer, the scale factor becomes $\frac{\min_{k=l,r}|d_{1,k}(\theta_s)|}{|d_{ref}(\theta_s)|}$

2.5 Overview of Different Configurations for the Beamformer Design

In this thesis, binaural Behind-The-Ear hearing aids (BTE) with four microphones (two microphones at the left side and two microphones at the right side) will be considered. Different possible configurations for the binaural beamformer design in BTE hearing aid will be implemented and tested. The first beamformer design can be achieved by having a direct access to the four microphones signals, as Figure 2.3 shows. We will refer to this configuration by 2+2 structure. Since we have a direct access to the four microphone signals, this configuration has access to all correlations between the four available signals. However, this structure requires each side to transmit two signals to the other side through a wireless link, which is considered a highly bandwidth-consuming and power-consuming operation. In this thesis, we have assumed an ideal wireless link. In practice, however, a non-ideal wireless link would cause delay, jitter and packet loss.

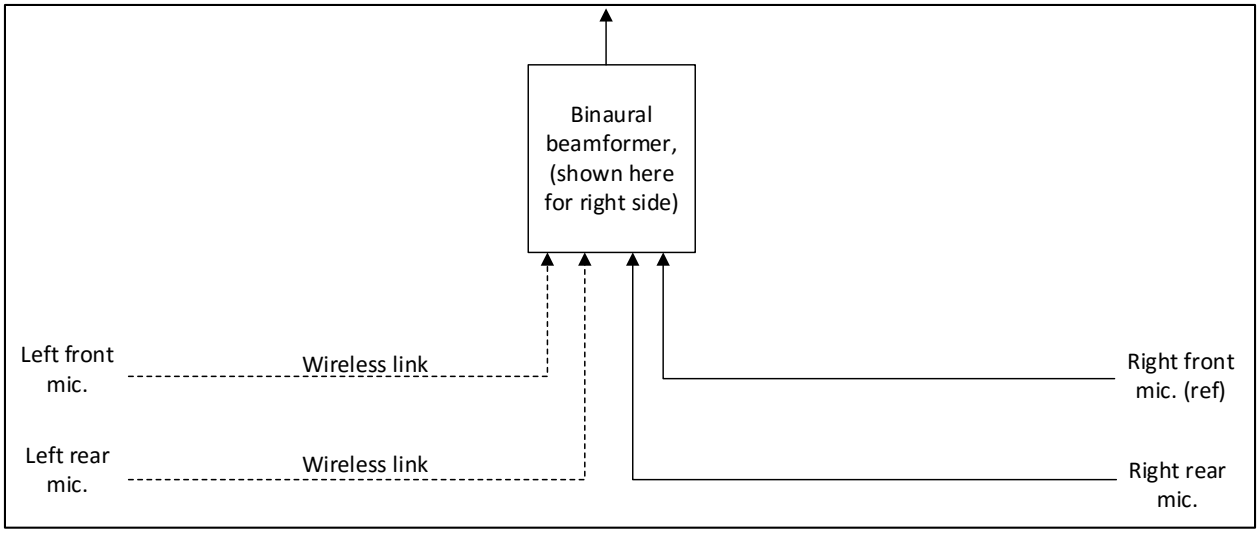

Figure 2.3: 2+2 structure with two wireless connection

In order to reduce the bandwidth requirements on the wireless link, another configuration is introduced as it is shown in Figure 2.4. In this configuration, which we will refer to as 2+1 structure without pre-processing, each side transmits only one signal to the other side. However, this structure does not take the advantage of having four microphones as it drops one microphone, which might lead to degradation in the performance.

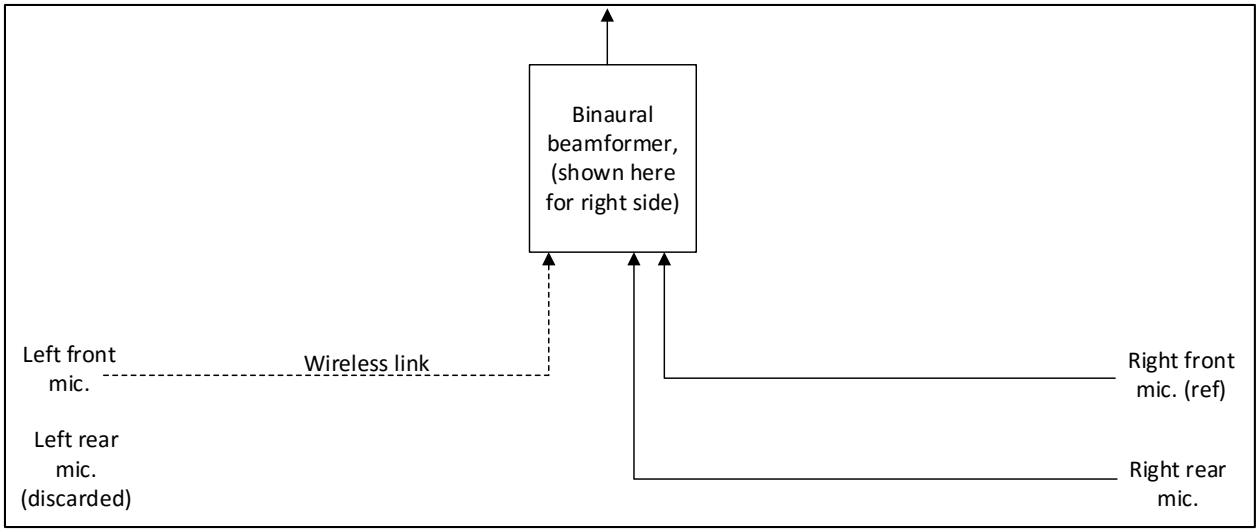

Figure 2.4: 2+1 structure without pre-processing

Consequently, the configuration in Figure 2.5, which we refer to as 2+1 structure with pre-processing on one side only, is introduced. As in the previous structure, each side only transmits one signal to the other side; however, the monaural beamformer output is computed on one side to be transmitted to the other side. In this configuration, we are taking the advantage of having four available microphones in contrast to the previous 2+1 structure without pre-processing.

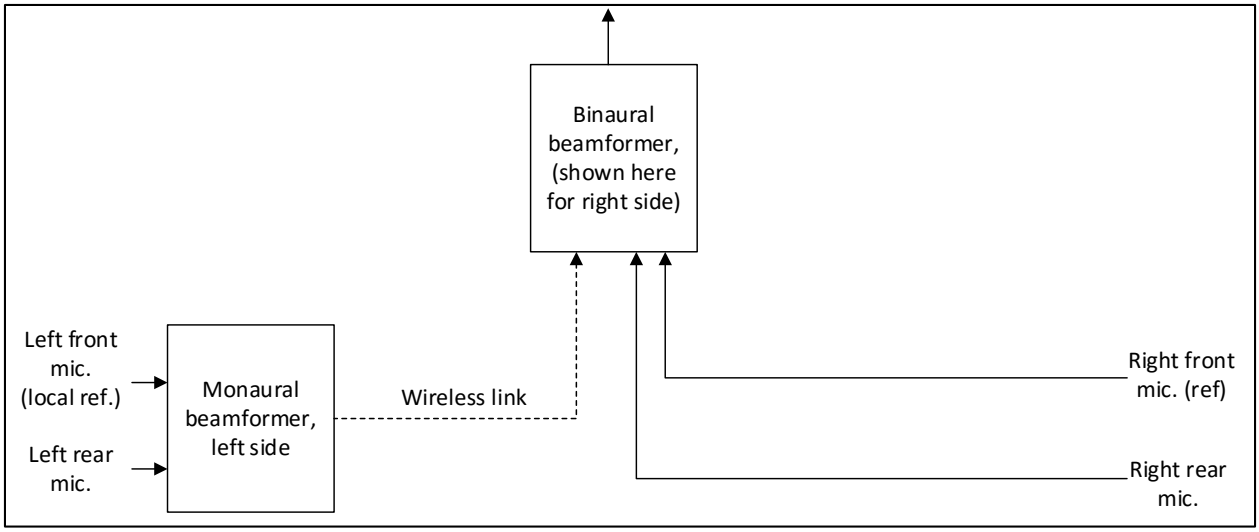

Figure 2.5: 2+1 structure with pre-processing on one side only.

Another configuration that needs only one wireless link and takes the advantage of the four microphones is shown in Figure 2.6. We will refer to this configuration as 1+1 structure with pre-processing on both sides. In this configuration, a monaural beamformer output is computed on each side, to be transmitted to the other side. The binaural beamformer on each side becomes simpler, with only two channels as inputs. On the other hand, the previous structure in Figure 2.5 with 3 inputs for the binaural beamformer on each side may allow to consider more correlations between the signals.

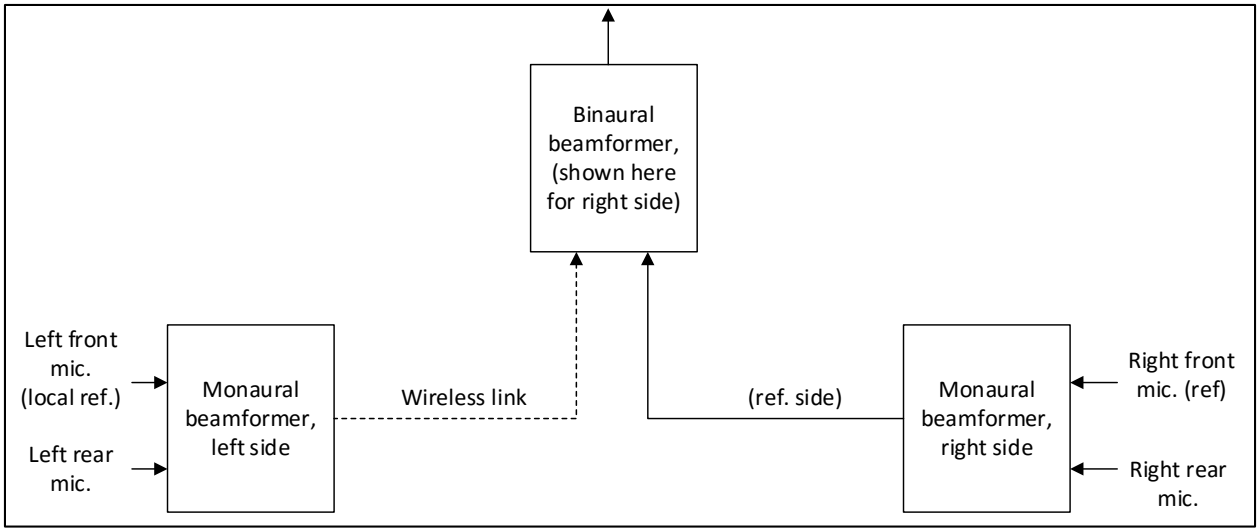

Figure 2.6: 1+1 structure with pre-processing on both sides

As it is noticeable from Figure 2.3 to Figure 2.6, these configurations produce one beamformer output. However, in binaural hearing aids it is required to generate two outputs, i.e., one output on each side. Therefore, the previous configurations must be duplicated and run in parallel, where the beamformer on the left side uses the front left microphone as a reference and the beamformer on the right side uses the front right microphone as a reference, as we will further describe in details in Chapter 5.

It is worth to mention that the common gain approach, which is discussed in section 2.4 , assumes that the same beamformer output is available on both sides. The common gain approach is thus (in principle) only doable in the configurations that have symmetric structures, i.e., where the same binaural beamformer input signals are available on both side, and the same microphone can be selected on both sides as the reference microphone. Consequently, the common gain approach is (in principle) only doable in the 2+2 structure and in the 1+1 structure with pre-processing on both sides, if they are configured such that they use the same reference microphone. And the common gain approach could not (in principle) be directly implemented using neither the 2+1 structure without pre-processing nor the 2+1 structure with pre-processing on one side only.

Chapter 3 Performance Measurements

This chapter presents several objective performance metrics used to evaluate the proposed algorithms. These performance metrics are classified into two categories. The first category is classical performance metrics that measure the noise reduction and speech distortion. The second category is cues preservation metrics that measure the capability of the proposed algorithms in preserving the binaural cues for the acoustic scene.

3.1 Classical Performance Metrics

3.1.1 Signal to Noise Ratio gain (SNR-gain)

The Signal to Noise Ratio gain (SNR-gain, or Array gain) is defined as the improvement in the SNR that the system introduces. SNR-gain is computed by taking the difference between the output SNR and the input SNR, and it is often measured in dB. The input SNR is computed by taking the ratio of the power of the target components over the power of the sum of the interferers and the background noise components at the reference microphone signal. The output SNR is computed by taking the ratio of the power of the target components over the power of the sum of the interferers and the background noise components at the beamformer's output. For the input SNR, the signal components at the front right or front left microphone can be used.

In human hearing, there are frequencies which are more important than other frequencies. Therefore, a frequency dependent measurement for the SNR–gain, which is computed as shown in equations (3.1) to (3.3), is required. In the SNR-gain calculations, we use the "raw" frequency-dependent SNR and we are not assuming anything about the perceptual importance of each frequency. Thus, the "raw" frequency-dependent SNR is more relevant than if we average the result of all frequencies, assuming equal weight for each frequency (which is not perceptually motivated).

$$SNR^{in} = 10\log\left(\frac{P_{x_{ref}}(f,t)}{P_{v_{ref}+n_{ref}}(f,t)}\right) \tag{3.1}$$

$$SNR^{out} = 10\log(\frac{P_{x^{out}}(f,t)}{P_{v^{out}+n^{out}}(f,t)}) \quad (3.2)$$

$$SNR-gain(dB) = SNR^{out}(dB) - SNR^{in}(dB) \quad (3.3)$$

Where;

x_{ref} : Target components at the reference microphone. It can be $x_{ref,l}$ when the front left microphone is used as a reference or $x_{ref,r}$ when front right microphone is used as a reference.

$P_{x_{ref}}$: Power spectrum density for the input target components at the reference microphone (front left or front right microphone).

$P_{x^{out}}$: Power spectrum density for the target components at output of the beamformer.

$P_{v_{ref}+n_{ref}}$: Power spectrum density for the sum of the interference components and the background noise components at the reference microphone (front left or front right microphone).

$P_{v^{out}+n^{out}}$: Power spectrum density for the sum of the interference components and the background noise components at the beamformer output.

f: Frequency index. It will be dropped throughout this chapter for simplicity.

t : Time frame index. It will be dropped throughout this chapter for simplicity.

SNR^{in}: Input signal to noise ratio (measured in dB).

SNR^{out}: Output signal to noise ratio (measured in dB)

$SNR-gain$: Signal to noise ratio gain (measured in dB).

In the case of binaural outputs, two SNR-gains are computed, using the beamformer output on both sides and the reference microphone signal on both sides.

For simulation purposes, the SNR-gain was computed with power spectrum densities using a Welch method with a Hamming window of size 2048 with 50% overlapping and FFTs of size 8192, with a sampling rate of 24 kHz (used throughout this thesis).

3.1.2 Signal to Distortion Ratio (SDR)

The Signal to Distortion Ratio (SDR) measures how the target components in the beamformer output are distorted compared with target components at the reference microphone. The SDR is computed by taking the power of the difference between the target component at the output channel and the target component signal at the reference microphone. After that, the computed power difference is normalized by the power of the target component at the reference microphone. The SDR is often measured in dB.

Similar to the SNR-gain, a frequency dependent measurement for the SDR can be calculated using equation (3.5):

$$x_{distorted} = x^{out} - x_{ref} \quad \text{(subtraction of time aligned signals)} \tag{3.4}$$

$$SDR = 10\log(\frac{P_{x_{distorted}}}{P_{x_{ref}}}) \tag{3.5}$$

Where;

$P_{x_{ref}}$: Power spectrum density for the input target components at the reference microphone.

$P_{x_{distorted}}$: Power spectrum density for the difference between the output target components at the output channel and the target components at the reference microphone.

In case of binaural outputs, two SDRs are also calculated, using the beamformer output on both sides and the reference microphone signal on both sides.

For simulation purposes, the SDR was calculated with power spectrum densities calculated using a Welch method with a Hamming window of size 2048, with 50% overlapping and FFTs of size 8192.

3.2 Cues Preservation Metrics

As it was mentioned previously, the performance metrics of the cues preservations for the acoustic scene are based on the relative differences of the signals at the two ears on the azimuth plane (Begault, Durand and others, 1994). The evaluations are done on the input signals at some reference microphones, usually the frontal left and frontal right microphones, and on the binaural output signals, left and right. The measurements are the Interaural Level Difference (ILD), the Interaural Time Difference (ITD), the Interaural Phase Difference (IPD), and the Magnitude Squared Coherence (MSC).

The ITD and the IPD are more suitable for the low frequency components, approximately below 1500 Hz as the wavelength of the low frequency components is longer than the diameter of the head. Consequently, it is possible for the auditory system to detect the phase differences and/or the time differences between the signals at the left and right sides of the ear. The signal components with frequencies higher than 1500 Hz, where the wavelengths are shorter than the size of the head, convey ambiguous information in terms of the phase differences and the relative time difference between the left and right side signals (Begault, Durand R and others, 1994; Blauert, 1997; Gilkey & Anderson, 2014). Based on (Begault, Durand R and others, 1994), the global ITD value for the average head is around 0.65 msec (averaged across frequencies).

On the other hand, the ILD is more suitable for the signal components with high frequencies, i.e., higher than 1500 Hz. In this case, the head acts as an obstacle and causes a head shadow effect, which creates a difference in the intensities between the left and right signals. As the frequency increases (i.e., as the wavelength decreases), the head shadow effect increases. However, for the low frequency components below 1500 Hz, the ILD does not have a significant effect on the spatial cues preservations, as these components have longer wavelengths that will bend around the head. As a result, the differences in the intensities between the left and right signals are small at low frequencies (Begault, Durand R and others, 1994; Blauert, 1997; Gilkey & Anderson, 2014).

The ITD envelope is another measurement that can be used to evaluate the binaural cues preservation for the high frequency components above 1500Hz (Begault, Durand R and others, 1994). The ITD envelope is based on the extraction of the time differences in the amplitude envelope of the signals. In other words, the ITD is different from the ITD envelope as the ITD extracts the time differences between two signals but not between the amplitude envelopes of these two signals. The amplitude envelope of a sound signal has lower frequency components than the original sound signal within the same envelope. The efficiency of the ITD envelope was proven experimentally in (Begault, Durand R and others, 1994) using high pass filtered noise at various cut-off frequencies above 1500 Hz. The evaluation was done by finding the interaural time difference for the amplitude envelope in each sub-band and then combining these sub-bands measurements to find a global measurement. However, further studies have shown that the ITD envelope measurements have limited efficacy for broadband sound sources, which are the most common sound signals in normal spatial hearing (Middlebrooks, 1992).

The aforementioned measurements, ILD, IPD, and ITD, are suitable to measure the binaural cues preservations for directional sources (target and interferers), but not for acoustic background noise which can be more diffuse. In order to measure the binaural cues preservations for the background noise in typical acoustic scene scenarios, the Interaural Coherence (IC) can be used. However, to avoid a complex analysis for the imaginary and the real parts of the IC, the magnitude squared coherence (MSC) is often used (Marquardt, Hohmann, & Doclo, 2013; Marquardt et al., 2014). Further explanations for this measurement will be presented later.

3.2.1 Interaural Level Difference (ILD)

The Interaural Level Difference (ILD) for the target speech signal and the sum of directional interference signals is defined as the difference (in dB) between the ratio of the power spectrum density of the left and right input signals at the reference microphones (front left and front right) and the ratio of the power spectrum density of the left and right binaural beamformer output components (Doclo et al., 2008).

$$ILD_x^{in} = 10\log_{10}\frac{P_{x_{ref,l}}}{P_{x_{ref,r}}} \tag{3.6}$$

$$ILD_v^{in} = 10\log_{10}\frac{P_{v_{ref,l}}}{P_{v_{ref,r}}} \tag{3.7}$$

$$ILD_x^{out} = 10\log_{10}\frac{P_{x_l^{out}}}{P_{x_r^{out}}} \tag{3.8}$$

$$ILD_v^{out} = 10\log_{10}\frac{P_{v_l^{out}}}{P_{v_r^{out}}} \tag{3.9}$$

$$\Delta\, ILD_x = ILD_x^{in} - ILD_x^{out} \tag{3.10}$$

$$\Delta\, ILD_v = ILD_v^{in} - ILD_v^{out} \tag{3.11}$$

Where;

$P_{x_{ref,l}}$, $P_{x_{ref,r}}$: Power spectrum density for the input target components at the front left and front right microphones.

$P_{x_l^{out}}$, $P_{x_r^{out}}$: Power spectrum density for the target components at left and right beamformer outputs.

$P_{v_{ref,l}}$, $P_{v_{ref,r}}$: Power spectrum density for the sum of the input interference components at the front left and front right microphones.

$P_{v_l^{out}}$, $P_{v_r^{out}}$: Power spectrum density for the sum of the interference components at the left and right beamformer outputs.

All the above power spectrum density estimates depend on the frequency index "*f*", however this is omitted throughout the document for simplicity.

ILD_x^{in} : Input interaural level difference between the target components at the front left and right microphones (measured in dB).

ILD_v^{in} : Input interaural level difference between the sums of interference components at the front left and right microphones (measured in dB).

ILD_x^{out} : Output interaural level difference between the target components at the beamformer outputs (measured in dB).

ILD_v^{out} : Output interaural level difference between the sums of interference components at the beamformer outputs (measured in dB).

$\Delta\, ILD_x$: Interaural level difference error for the target components (measured in dB).

$\Delta\, ILD_v$: Interaural level difference error for the sum of the interference components (measured in dB).

For simulation purposes, the ILDs were produced with power spectrum densities calculated using a Welch method with a Hamming window of size 192 and FFTs of size 2048. The input signal was divided into 2499 overlapping windows, each with 50% overlap.

In order to easily compare the capability of different beamformers in preserving the binaural spatial cues, a single scalar value for $\Delta\, ILD_x$ and $\Delta\, ILD_v$ is also calculated. As mentioned previously, the ILD is suitable for the high frequency components of the signals, above 1500 Hz. Therefore, the mean values for $\Delta\, ILD_x$ and $\Delta\, ILD_v$ are calculated for both the target components and the sum of interference components for frequency components higher than 1500Hz.

3.2.2 Interaural Phase Difference (IPD)

The Interaural Phase Difference (IPD) for the target and the sum of directional interference components is defined as the phase of the cross power spectrum density between the left and the right input signals at the reference microphones and between the left and the right signals at the beamformer outputs (Doclo et al., 2008).

$$IPD_x^{in} = \angle(P_{x_{ref,l},x_{ref,r}}) \tag{3.12}$$

$$IPD_v^{in} = \angle(P_{v_{ref,l},v_{ref,r}}) \tag{3.13}$$

$$IPD_x^{out} = \angle(P_{x_l^{out},x_r^{out}}) \tag{3.14}$$

$$IPD_v^{out} = \angle(P_{v_l^{out},v_r^{out}}) \tag{3.15}$$

Where;

$P_{x_{ref,l},x_{ref,r}}$: Cross power spectrum density between the input target components at the front left and from right microphones.

$P_{v_{ref,l},v_{ref,r}}$: Cross power spectrum density between the sums of the input directional interference components at the front left and front right microphones.

$P_{x_l^{out},x_r^{out}}$: Cross power spectrum density between the target components at left and right beamformer outputs.

$P_{v_l^{out},v_r^{out}}$: Cross power spectrum density between the sums of directional interference components at the left and right beamformer outputs.

Similar to the previous (auto-) power spectrum densities, the above cross power spectrum density estimates depend on the frequency index "*f*", however, this is omitted throughout the document for simplicity.

IPD_x^{in} : Input interaural phase difference between the target components at the front left and front right microphones (measured in radians, between $-\pi$ and π).

IPD_v^{in} : Input interaural phase difference between the sums of directional interference components at the front left and front right microphones (measured in radians, between $-\pi$ and π).

IPD_x^{out} : Output interaural phase difference between the target components at the left and right beamformer outputs (measured in radians, between $-\pi$ and π).

IPD_v^{out} : Output interaural phase difference between the sums of directional interference components at the left and right beamformer outputs (measured in radians, between $-\pi$ and π).

We also define the following:

$\Delta\ IPD_x$: Interaural phase difference error for the target components.

$\Delta\ IPD_v$: Interaural phase difference error for the sum of interference components.

For simulation purposes, the IPDs were produced with cross power spectrum densities calculated using a Welch method with a Hamming window of size 192 and FFTs of size 2048. The input signals were divided into 2499 overlapping windows, each with 50% overlap.

If the objective is to obtain an IPD error (ΔIPD) between –π and π, the error cannot be calculated directly from the wrapped or the unwrapped version of IPD^{in} and IPD^{out}. Therefore, to find a meaningful difference between the IPD^{in} and the IPD^{out}, equation (3.16) is used to calculate the ΔIPD. Figure 3.1 and Figure 3.2 show an example of the wrapped and unwrapped IPDs for the interference components at the inputs and outputs of a binaural beamformer and the corresponding error (ΔIPD).

$$\Delta IPD = modulo\left((IPD^{out} - IPD^{in}) + \pi,\ 2\pi\right) - \pi \tag{3.16}$$

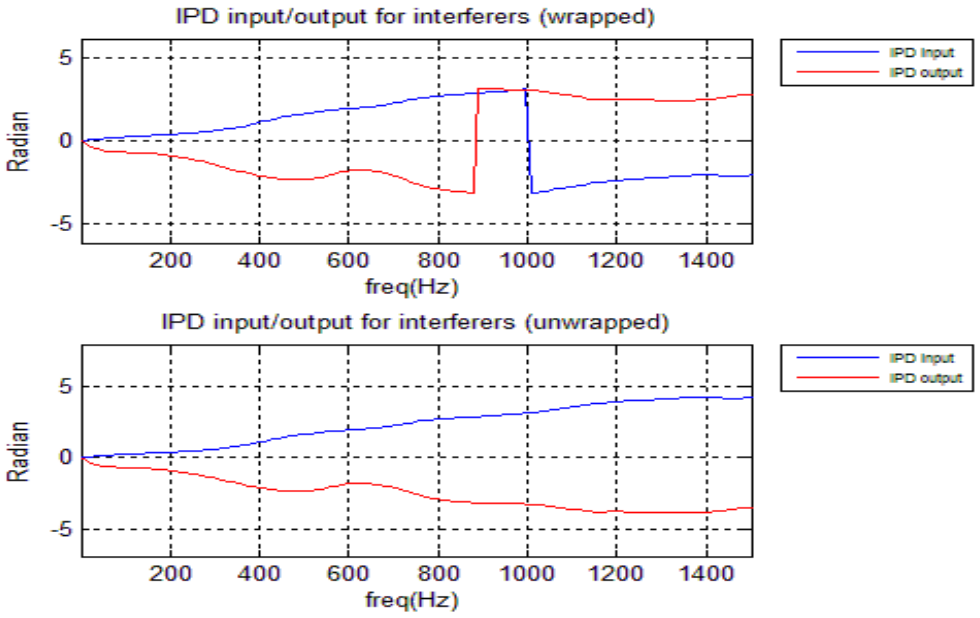

Figure 3.1: Wrapped and unwrapped *IPDs* for the interference components at the inputs and outputs of a binaural beamformer

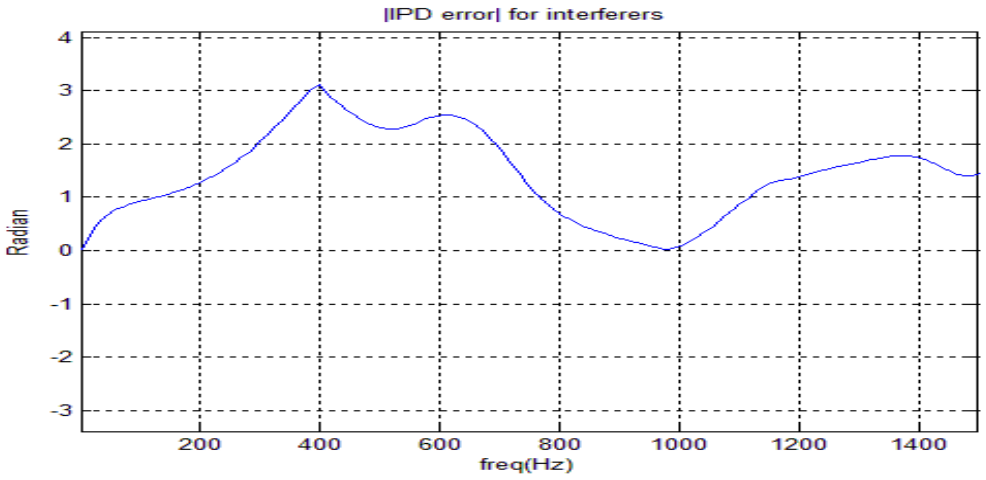

Figure 3.2: Δ*IPD* for the interference components of a binaural beamformer

To easily compare the capability of different beamformers in preserving the binaural spatial cues, single scalar values for $\Delta\, IPD_x$ and $\Delta\, IPD_y$ are also calculated. As mentioned previously, the IPD is suitable for the low frequency components, below 1500 Hz. Therefore, the mean values

for the $\Delta\, IPD_x$ and the $\Delta\, IPD_y$ are calculated for the signal components with frequencies lower than 1500Hz.

It is also possible to combine the frequency-dependent ILD and IPD to generate the Interaural Transfer Function (ITF), which is defined as the ratio of the signal components at the left and the right sides.

In case of multiple interferers, care must be taken when ILD and IPD/ITD measurements are made based of the sum of interference signals. The magnitude of the sum of interferers at each frequency is normally different from the sum of individual interferer magnitudes, and the phase of the sum of interferers at each frequency is normally different from the sum of individual interferer phases. Thus computing ILDs and IPDs/ITDs from the sum of interferer signals only makes true physical sense when there is a dominant interferer at some T-F bin or T-F region (sparseness assumption), but this assumption of the domination of one interferer over others may not always be valid.

In order to put more emphasis on the frequencies that have more magnitude content in the left and right signals, a weighted IPD error (*weighted-*ΔIPD) is also introduced. A global auto power spectrum density for the input interference components is computed by finding the auto power spectrum densities of the input interferer components at the front left and front right microphones, and then averaging them at each frequency. The *weighted-*ΔIPD is then calculated by multiplying (scaling) the ΔIPD by the global power spectrum density at each frequency. By using this measurement, less weight will be put on the frequency components that have low magnitude in terms of the auto power spectrum densities, while more weight will be put on the frequencies with high magnitude content. This weighted measurement is more suitable for the visualization purposes of the IPD in case of multiple interferers, as it was found to help reducing sudden peaks in ΔIPD (unweighted) generated from 2π jumps in the input or output IPD (these jumps in the IPD measurements for the sum of interference signals were caused by zero-crossings frequently observed in the magnitude of the cross power spectrum density between the sum of interference signals from the two sides). However, the unweighted ΔIPD gives us the real value for IPD error as the weighted ΔIPD is affected on the global auto power spectrum density values.

3.2.3 Interaural Time Difference (ITD)

The Interaural Time Difference (ITD) is another measurement that is used to evaluate the capability of beamforming or noise reduction algorithms in preserving the cues. The ITD is measured in samples or in seconds and is calculated from the unwrapped version of the IPD. Throughout this document, five approaches are used to calculate the ITD.

The first used approach is based on calculating the group delay from the unwrapped version of the IPD as it is shown in equations (3.17) to (3.22);

$$ITD_x^{in} = -\frac{d(unwrapped(IPD_x^{in}))}{dw} \tag{3.17}$$

$$ITD_v^{in} = -\frac{d(unwrapped(IPD_v^{in}))}{dw} \tag{3.18}$$

$$ITD_x^{out} = -\frac{d(unwrapped(IPD_x^{out}))}{dw} \tag{3.19}$$

$$ITD_v^{out} = -\frac{d(unwrapped(IPD_v^{out}))}{dw} \tag{3.20}$$

$$\Delta\ ITD_x = ITD_x^{in} - ITD_x^{out} \tag{3.21}$$

$$\Delta\ ITD_v = ITD_v^{in} - ITD_v^{out} \tag{3.22}$$

Where;

ITD_x^{in} : Input interaural time difference between the target components at the front left and front right microphones (measured in samples or second).

ITD_v^{in} : Input interaural time difference between the sums of directional interference components at the front left and right microphones (measured in samples or second).

ITD_x^{out} : Output interaural time difference between the target components at the left and right beamformer outputs (measured in samples or second).

ITD_v^{out} : Output interaural time difference between the sums of directional interference components at the left and right beamformer outputs (measured in samples or second).

$\Delta\, ITD_x$: Interaural time difference error for the target components (measured in samples or second).

$\Delta\, ITD_v$: Interaural time difference error for the sums of the directional interference components (measured in samples or second).

Similar to the *IPD*, to have a single scalar value for the $\Delta\, ITD_x$ and the $\Delta\, ITD_v$, the mean values for the aforementioned measurements are calculated for the frequency components below 1500 Hz.

The second approach used to calculate the ITD is based on the phase delay of the unwrapped IPD according to equations (3.23) to (3.28);

$$ITD_x^{in} = -\frac{unwrapped(IPD_x^{in})}{w} \tag{3.23}$$

$$ITD_v^{in} = -\frac{unwrapped(IPD_v^{in})}{w} \tag{3.24}$$

$$ITD_x^{out} = -\frac{unwrapped(IPD_x^{out})}{w} \tag{3.25}$$

$$ITD_v^{out} = -\frac{unwrapped(IPD_v^{out})}{w} \tag{3.26}$$

$$\Delta\, ITD_v = ITD_v^{in} - ITD_v^{out} \tag{3.27}$$

$$\Delta\, ITD_x = ITD_x^{in} - ITD_x^{out} \tag{3.28}$$

To have a single scalar value of the $\Delta\, ITD_x$ and the $\Delta\, ITD_v$, using the second approach, the mean values of the aforementioned measurements are calculated for the frequency components between 100 Hz and 1500 Hz.

The third approach that is used to calculate the ITD is based on finding local slope of the unwrapped IPD as a function of the frequency, using ten consecutive points on the curve. This is somewhat similar to the first approach, which measure the group delay, but it is more robust to measurement noise. Here again single scalar values of the ΔITD_x and the ΔITD_y, are found by taking the mean values of the ΔITD_x and the ΔITD_y, for the frequency components below 1500Hz.

Another frequency-independent method (i.e., producing a single scalar value) has been used to calculate values for ITD^{in} and ITD^{out}. In this case, ITD^{in} is calculated by finding the delay corresponding to the maximum cross correlation between the input signal (target or sum of directional interferers) at the front left and front right reference microphones. Similarly, ITD^{out} is calculated from the cross correlation of the left output components and the right output components for the target and the sum of directional interferers. This method is used to verify the results from the three previous approaches. However, the drawback of this approach is that it only produces integer ITD values. One way to address this problem would be to perform some curve fitting on the cross-correlation function to better approximate the peak position, or to up sample the data before computing the cross-correlation to improve the time resolution.

Alternatively, it is possible to compute a scalar ITD value directly from the phase response by performing a linear regression model on the whole IPD response up to 1500 Hz. A least-squares method is used to fit the unwrapped IPD into a linear model. Least-squares fitting is based on minimizing the sum of the squares of the residuals, where the residuals are the differences between the observed values and the predicted values by the model. Coefficient of determination R^2 is one of the measurements that are used to measure the goodness of the predicted model. It measures how closely the predicted values that are obtained from the fitting model match the real data (unwrapped IPD in this case) (Linear regression.2015).

$$R^2 = 1 - \frac{SS_{resid}}{SS_{total}} \tag{3.29}$$

SS_{resid} : Sum of the squared residuals from the regression.

SS_{total}: Sum of the squared differences from the mean of the dependent variable, calculated by multiplying the variance of the unwrapped IPD by the number of observations minus 1.

It is worth to mention that in our computation of ILD/ITD/IPD we compute the equivalent of the real practical HRTF ratio (including reverberant effects, head shadow, torso reflections, and direction-dependent filtering of the ear).

3.2.4 Magnitude Squared Coherence (MSC)

As it was mentioned previously, the Interaural Coherence (IC) is suitable for measuring the preservation of the binaural spatial cues for the background diffuse-like noise, i.e., noise that doesn't have the characteristics of a directional source. However, to facilitate the analysis, the Magnitude Squared Coherence (MSC) is often used by considering only the magnitude of the IC (Marquardt et al., 2013; Marquardt et al., 2014). The MSC depends on the cross power spectrum density between the left and the right signals, the auto power spectrum density of the left signal, and the auto power spectrum density of the right signal. The MSC is a frequency-dependent estimate with values between zero and one. The MSC is computed for the input background noise components at the front left and the front right reference microphones and for the left and right background noise components at the left and right beamformer outputs. The following equations show the mathematical descriptions for the input and the output IC and MSC.

The input Interaural Coherence IC^{in} for the input background noise at the front left and front right microphones is;

$$IC^{in} = \frac{P_{n_{ref,l},n_{ref,r}}}{\sqrt{P_{n_{ref,l}} P_{n_{ref,r}}}} \tag{3.30}$$

Where;

$P_{n_{ref,l},n_{ref,r}}$: Cross power spectrum density between the background noise components at the front left and front right microphones.

$P_{n_{ref,l}}$: Power spectrum density of the background noise components at the front left microphone.

$P_{n_{ref,r}}$: Power spectrum density of the background noise components at the front right microphone.

The input real-valued Magnitude Squared Coherence MSC^{in} is;

$$MSC^{in} = | IC^{in} |^2 \tag{3.31}$$

The output Interaural Coherence IC^{out} for the background noise components at the outputs of the beamformer is;

$$IC^{out} = \frac{P_{n_l^{out}, n_r^{out}}}{\sqrt{P_{n_l^{out}} P_{n_r^{out}}}} \tag{3.32}$$

Where;

$P_{n_l^{out}, n_r^{out}}$: Cross power spectrum density between the background noise components at the left and right beamformer outputs.

$P_{n_l^{out}}$: Power spectrum density of the background noise components at the left beamformer's output.

$P_{n_r^{out}}$: Power spectrum density of the background noise components at the right beamformer's output.

The output real-valued Magnitude Squared Coherence MSC^{out} is;

$$MSC^{out} = | IC^{out} |^2 \tag{3.33}$$

In order to measure the capability of the beamformers or the noise reduction algorithms to preserve the binaural cues for the background noise, we need to find the difference between MSC^{out} and MSC^{in} to find the ΔMSC as in equation (3.34);

$$\Delta\, MSC = MSC^{out} - MSC^{in} \tag{3.34}$$

For simulation purposes, the magnitude squared was calculated using Welch's averaged modified periodogram method with a Hamming window of size 192 and FFTs of size 2048. The input signal was divided into 2499 overlapping windows, each with 50% overlap.

As the MSC error is a broadband frequency dependent measurement, a single scalar measurement for the MSC error can also be calculated to allow an easier comparison between the developed beamformers. This single scalar value is calculated by taking the average of the MSC errors over the frequencies up to 1500 Hz, based on the perceptual capability of the human auditory system (Gabriel & Colburn, 1981; Walther & Faller, 2013).

3.3 Evaluation of the Different Components in the Beamformer Outputs

To calculate the SNR^{out}, SDR, ILD^{out}, IPD^{out}, ITD^{out}, and MSC^{out}, the target components, directional interference components, and background noise components of the left and right beamformer outputs are required. The "phase inversion technique" (Hagerman & Olofsson, 2004) can be used to separate the output signals into three components, the target, the sum of directional interferers, and the background noise. This is achieved by separately processing the "target + interferers + background noise", "target + interferers – background noise", and "target – interferers + background noise" signals as an input to the beamformer as in equations (3.35) to (3.37)

$$\mathbf{y} = \mathbf{x} + \mathbf{v} + \mathbf{n} \tag{3.35}$$

$$\mathbf{y}_n = \mathbf{x} + \mathbf{v} - \mathbf{n} \tag{3.36}$$

$$\mathbf{y}_i = \mathbf{x} - \mathbf{v} + \mathbf{n} \tag{3.37}$$

By assuming linearity and in order to find each output component separately, the following equations have been used;

$$v^{out} = 0.5z - 0.5z_i \tag{3.38}$$

$$n^{out} = 0.5z - 0.5z_n \tag{3.39}$$

$$x^{out} = 0.5z + 0.5z_n - v^{out} \tag{3.40}$$

Where;

z : Output of the beamformer when $\mathbf{y}$ is used as an input. Either the front left microphone or the front right microphone can be used as a reference.

z_i : Output of the beamformer when $\mathbf{y}_i$ is used as an input. Either the front left microphone or the front right microphone can be used as a reference.

z_n : Output of the beamformer when $\mathbf{y}_n$ is used as an input. Either the front left microphone or the front right microphone can be used as a reference.

Calculating SNR^{in}, ILD^{in}, IPD^{in}, ITD^{in}, and MSC^{in} is easier since we have an access to all input components (target, sum of directional interferers, and background noise) in the simulation. The main advantage of the "phase inversion technique" is that it is robust for the non-stationary scenarios such as variable activity patterns of sources, head movement, moving sources, etc. Moreover, it is robust to sign changes, gains or fractional delays introduced in the processing. The technique assumes linearity but it has also been found to work well in scenarios with some degrees of non-linearity such as compression and speech enhancement post-filtering in hearing aids. However, this technique requires three times of the processing compared with the normal beamforming processing time. But this only affects the simulations, as a real-time implementation of the beamforming algorithms would obviously not implement the "phase inversion technique" or compute the SNR^{out}, SDR, ILD^{out}, IPD^{out}, ITD^{out}, and MSC^{out} values.

Chapter 4 Binaural Beamforming Algorithms Based on Selection and Mixing

The common gain approach, which is discussed in section 2.4 and originally proposed in (Lotter & Vary, 2006), does not necessarily provide the best trade-off between cues preservation and noise/interference reduction. In (Thiemann et al., 2014), an algorithm that uses a binary decision in the time- frequency (T-F) plane was introduced. As mentioned in the introduction, this algorithm and its variants show a good potential because they do not require the sometimes unrealistic estimation of parameters as some other methods require, and there is flexibility for using different signals and different criteria in each T-F bin to produce the binaural output mixtures. Since realistic T-F decompositions for hearing aids need to be done with a very small group delay introduced, which means that small frame sizes need to be used (or equivalently a filter bank with coarse frequency resolution or with large down sampled bandwidths should be used), in this thesis we will use real-life data and real-life algorithms used in today's binaural hearing aids, provided by our sponsor Siemens Audiologische Technik GmbH Group (now Sivantos Inc.) in order to evaluate the effectiveness of the T-F decision method. In addition, there are several options that have not been considered yet in the work of (Thiemann et al., 2014). Consequently, several new scenarios are investigated in this chapter, in order to find the best trade-off between noise reduction and cues preservations.

4.1 Overview

In this section, we will present algorithms that are based on mixing the available signals at each subband-time sample (T-F plane in hearing aids applications using the filter bank) in order to find the best trade-off between the noise attenuation and the cues preservations of an acoustic scene. Figure 4.1 reveals the selection and the mixing algorithm of the available signals. Based on the outputs of the control unit, the left and right outputs are created. These outputs must have a good trade-off of enhanced target signal and preserved cues for the acoustic scene. One pair of signals available for mixing are the beamformer outputs which have good level of noise reduction; however, they have poor binaural cues preservation as they change most of the cues for the

interferers and the background noise. Another pair of signals available for mixing are the original noisy input signals from the left and right front microphones, which obviously preserve the cues of the acoustics scene, but have no SNR improvement. Finally, another pair of signals available for mixing are the binaural outputs of the common gain approach, which preserves the binaural cues while providing some intermediate level of noise reduction. For simplicity, the common gain signals and the weighted mixtures of the different signals are not shown in Figure 4.1.

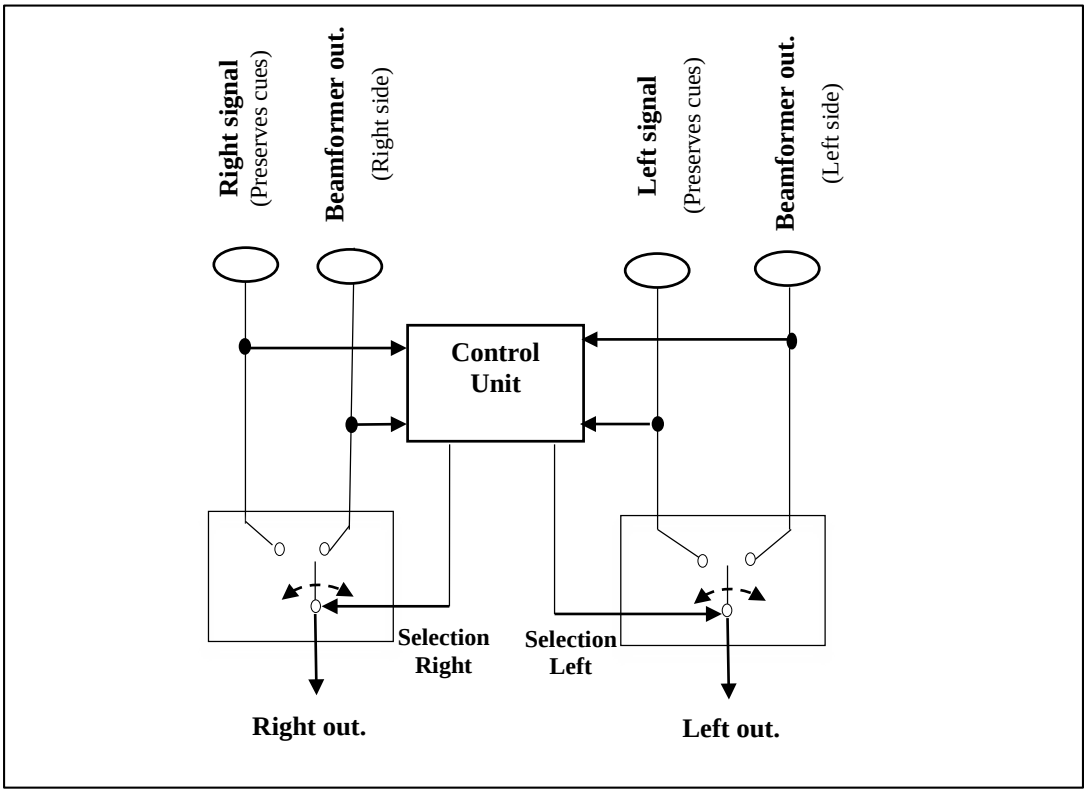

Figure 4.1: Selection and mixing of the available signals

Several classification criteria have been developed and tested in this chapter in order to mix the available signals at each subband-time sample. These criteria are the power of the beamformer outputs (whether from a "true" classical beamformer or a common-gain approach), the power of the noisy signals, the differences between those powers, and the coherence between the aforementioned signals. In each case, we have considered the option of keeping the phase components, the magnitude components or both components. Moreover, in some cases the selections/mixing are frequency independent while in other approaches the selections/mixing are

frequency dependent considering, in particular, that the role of IPD/ITD cues is mostly up to 1500 Hz. In other words, in the frequency dependent approaches, low frequency components (less than 1500Hz, or the first 8 subbands in the filter bank provided by our sponsor) and high frequency components are passed through a different kind of processing.

To find the best trade-off between noise reduction and cues preservations, Figure 4.2 shows different classification criteria that we have considered for classification, selection, and mixing. In order to generate output signals with enhanced desired target signal and preserved cues for the acoustic scene, the output signals will be selected/mixed at each subband-time sample.

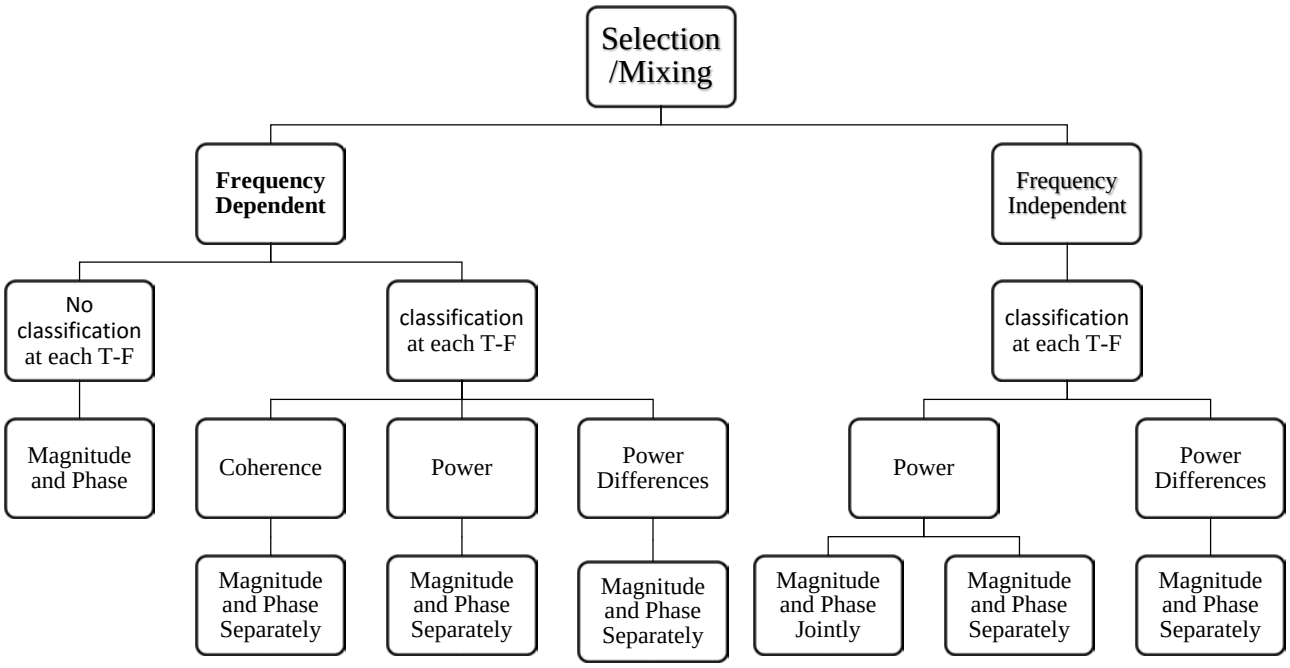

Figure 4.2: Classification, selection and mixing at each T-F bin or subband-time window

At the beginning, the powers of the available signals have been computed and compared as the classification criteria. The selections/mixing are based on the assumption that the signal with a lower power has a lower amount of residual noise; therefore, more emphasis will be put on keeping the signal from these low power bands. Further details will be introduced later.

Since the phase of the noisy signal plays an important role in preserving the spatial perception of the directional interferers in terms of IPD/ITD (we have only used ITD in case of single interference scenarios), we have also put emphasis on the phase of the noisy signals. Consequently, in our first design the powers of the signals are used to select and mix the magnitude of the available signals, whereas the phase of the noisy signal is kept. Our test results demonstrated the capability of these approaches in preserving the cues in terms of IPD /ITD for the interferers, and MSC for the background noise.

Based on the power differences between the available signals, other approaches have then been introduced to achieve a good trade-off between noise reduction and cues preservations. In these approaches, the range between the maximum power difference and the minimum power difference at each sub-band is quantized and used as criteria to adjust the weights of the available signals into the output signal.

The phases of the low frequency components have a crucial effect on the cues preservations of the directional interferers in terms of IPD/ITD, however, their effect decays for high frequency components. Moreover, the low frequency components (magnitude and phase) up to 1500Hz play an important role in preserving the spatial impression of the diffuse background noise. Consequently, the most promising frequency independent approaches have been extended to frequency dependent approaches where the phase of the original noisy signals are considered for the low frequency components of the selected/mixed output, and different processing is done on the high frequency components. In addition, since the coherence measures the similarity between two signals available for selection/mixing, frequency dependent approaches that are based on the coherence were also introduced.

Finally, just for comparison purposes and for further investigation of the trade-off between reducing the computation complexity on one hand and reducing the noise while preserving the spatial impression of the acoustic scene on the other hand, several frequency dependent approaches have been tested. These approaches do not do any sort of classifications at each subband-time window; they have just considered the phase and the magnitude of the noisy signal for some specific subbands.

In this chapter we will refer to the noisy signal at the front left microphone by y_l , the noisy signal at the front right microphone by y_r , the left and right beamformer outputs either with front left or front right reference microphone by z_l and z_r , the left common gain output by $\hat{y}_l$, and the right common gain output by $\hat{y}_r$. In addition, we will refer to the left output after selection/mixing by $z_{selected-left}$ and to the right output after selection/mixing $z_{selected-right}$.

In the test cases, case 1 and case 2 represent the common gain approach and the beamforming processing approach, respectively. These first two cases are not related with the selection and mixing approaches; however, we have used them for comparison purposes as the common gain approach preserves most of the cues while the beamforming processing introduces the best SNR gain.

4.2 Frequency Independent Selection and Mixing Approaches

4.2.1 Power as Classification Criterion

In this section, the power of the beamformer outputs, the power of the common-gain outputs, and the power of noisy signals are used for classification ,selection and mixing. In some cases, the whole signal (magnitude and phase) is considered for the selection/ mixing, while in other cases the magnitude and the phase components are considered separately.

Case 3

In this approach, the mixing are based on comparing the power of the left/ right common gain outputs with the power of the left/right beamformer outputs at each subband-time window. During the simulation for this case and the other cases 50 samples are used as the size of the time-window at sampling rate 24 kHz. So the power is estimated over a time window of size 50 ms at each subband. When the power of the beamformer output is lower than the power of the common gain output, a small portion β of the common gain output is added to a higher portion α of the beamformer output as the common gain output has more residual noise/interferers, and vice versa as equations (4.1) and (4.2) show. The beamformers and the common gain approach can be based on either the left or the right front microphone as the reference microphone.

$$z_{selected-left}(sb,t)=\begin{cases}\alpha z_l(sb,t)+\beta\hat{y}_l(sb,t)\\ if \quad power(z_l(sb,t))\le power(\hat{y}_l(sb,t))\\ \\ \beta z_l(sb,t)+\alpha\hat{y}_l(sb,t)\\ if \quad power(z_l(sb,t))> power(\hat{y}_l(sb,t))\end{cases} \tag{4.1}$$

$$z_{selected-right}(sb,t)=\begin{cases}\alpha z_r(sb,t)+\beta\hat{y}_r(sb,t)\\ if \quad power(z_r(sb,t))\le power(\hat{y}_r(sb,t))\\ \\ \beta z_r(sb,t)+\alpha\hat{y}_r(sb,t)\\ if \quad power(z_r(sb,t))> power(\hat{y}_r(sb,t))\end{cases} \tag{4.2}$$

Where $0.5<\alpha<1$ and $0<\beta<0.5$. In this case we used $\alpha= 0.7$ and $\beta=0.3$.

Case 4

In this approach, the unprocessed noisy signals at the front left and front right microphones are used instead of using the outputs of the common gain beamformer. The mixing in this approach is performed by comparing the power of the beamformer outputs with the power of the original noisy signals at the front left and front right microphones at each subband-time sample. The beamformers can be based on either the left or the right front microphones as reference microphone. When the power of the beamformer output is lower than the power of the front noisy signal, a strong portion "α" of the beamformer output is added to a lower portion "β" of the front left or right noisy signals and vice versa. In other words, we will choose the signal with the lower power (presumably with lower noise).

This classification is opposite to the one found in (Thiemann et al., 2014), which can be surprising at first glance. The approach in (Thiemann et al., 2014) is based on the following reasoning: if a noise-only segment is detected, the beamformer power is expected to be significantly lower than the input microphone signal (beamformer noise reduction), and in such cases a fraction of the (noise-only) input microphone signal can be kept as the output signal. This attenuates the noise while preserving the noise cues. If a noise-only segment is not detected, then

the beamformer outputs from both sides are used (with no noise cues preservation for those segments). However, this approach is not robust for the case of noisy inputs with low SNR, during target activity: in such cases the beamformer power is also expected to be significantly lower than the input microphone signal (beamformer noise reduction), and if a fraction of the (noisy) input signal is then kept as the output, this will attenuate the target signal and not provide any noise reduction (in SNR), thus it is not an acceptable scenario.

The approaches in this thesis are based on a different scheme: they put more emphasis on the beamformer output when it is detected that locally (in the T-F representation) the beamformer output provides more noise reduction compared to other signals (i.e., microphone input signal or other processed signals). And more emphasis is put on other signals (with better noise cues preservation) for segments where the beamformer output locally does not provide significant additional noise reduction compared to the other signals. It is worth noting that even for noise-only or low SNR scenarios where the beamformer is expected to provide significantly more noise reduction than the other signals, there are still a significant number of T-F bins where the noise reduction is not significantly greater in the beamformer output compared to the other signals, leading to some noise cues preservation since the other signals are given more emphasis for those T-F bins.

$$z_{selected-left}(sb,t)=\begin{cases}\alpha z_l(sb,t)+\beta y_l(sb,t)\\ if \quad power(z_l(sb,t)) \le power(y_l(sb,t))\\ \\ \beta z_l(sb,t)+\alpha y_l(sb,t)\\ if \quad power(z_l(sb,t)) > power(y_l(sb,t))\end{cases} \tag{4.3}$$

$$z_{selected-right}(sb,t)=\begin{cases}\alpha z_r(sb,t)+\beta y_r(sb,t)\\ if \quad power(z_r(sb,t)) \le power(y_r(sb,t))\\ \\ \beta z_r(sb,t)+\alpha y_r(sb,t),\\ if \quad power(z_r(sb,t)) > power(y_r(sb,t))\end{cases} \tag{4.4}$$

Where we have used α= 0.7 and β=0.3

In the aforementioned approaches, the selections/mixing have been done on the whole signals (magnitude and phase together). The following two cases (case 6 and case 7) are introduced to test the performance of considering the phase and the magnitude for the mixed signals separately.

Case 6

The phase of the noisy signal plays an important role in preserving the IPD/ITD cues for the directional interferers; in addition, it has an important effect in preserving the cues for the background noise. Therefore, only the phase of the noisy signal is considered here, while the magnitudes of both the beamformer output and common gain output are used as shown in equations (4.5) and (4.6). The magnitudes of the common gain outputs has an effect in preserving the ILD cues for the directional interferers and the MSC for the background noise, while the magnitude of the beamformer output plays an important role in improving the SNR-gain. The left and right beamformers and the common gain approach can be based on either the left or the right front microphones as reference microphone and the values of α and β used were 0.7 and 0.3, respectively:

$$z_{selected-left}(sb,t) = \begin{cases} (\alpha \mid z_l(sb,t) \mid + \beta \mid \hat{y}_l(sb,t) \mid e^{j\measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) < power(\hat{y}_l(sb,t)) \\ \\ (\beta \mid z_l(sb,t) \mid + \alpha \mid \hat{y}_l(sb,t) \mid e^{j\measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) > power(\hat{y}_l(sb,t)) \end{cases} \tag{4.5}$$

$$z_{selected-right}(sb,t) = \begin{cases} (\alpha \mid z_r(sb,t) \mid + \beta \mid \hat{y}_r(sb,t) \mid) e^{j\measuredangle y_r(sb,t)} \\ if \quad power(z_r(sb,t)) < power(\hat{y}_r(sb,t)) \\ \\ (\beta \mid z_r(sb,t) \mid + \alpha \mid \hat{y}_r(sb,t) \mid) e^{j\measuredangle y_r(sb,t)} \\ if \quad power(z_r(sb,t)) > power(\hat{y}_r(sb,t)) \end{cases} \tag{4.6}$$

Case 7

This approach is similar to case 6; however, this time the selections/mixing are done between the original noisy signals at the left and right front microphones, and the left and right beamformer outputs according to the following equations (4.7) and (4.8);

$$z_{selected\text{-}left}(sb,t) = \begin{cases} (\alpha \mid z_l(sb,t) \mid + \beta \mid y_l(sb,t) \mid) * e^{j \measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) < power(y_l(sb,t)) \\ (\beta \mid z_l(sb,t) \mid + \alpha \mid y_l(sb,t) \mid) * e^{j \measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) > power(y_l(sb,t)) \end{cases} \tag{4.7}$$

$$z_{selected\text{-}right}(sb,t) = \begin{cases} (\alpha \mid z_r(sb,t) \mid + \beta \mid y_r(sb,t) \mid) e^{j \measuredangle y_r(sb,t)} \\ if \quad power(z_r(sb,t)) < power(y_r(sb,t)) \\ (\beta \mid z_r(sb,t) \mid + \alpha \mid y_r(sb,t) \mid) e^{j \measuredangle y_r(sb,t)} \\ if \quad power(z_r(sb,t)) > power(y_r(sb,t)) \end{cases} \tag{4.8}$$

The beamformers can be based on either the left or right front microphones as reference microphone, and the values of α and β used were 0.7 and 0.3, respectively.

Since we have used the phases of the noisy signals in case 6 and case 7, these approaches would have similar performance in terms of IPD/ITD. Moreover, as the phase of the noisy is important in preserving the spatial perception of the background noise, these approaches would show comparable performance in terms of MSC (up to 1500 Hz). However, due to the use of the common gain outputs' magnitude instead of the original noisy signals' magnitude, the output of case 6 would have higher SNR (at the cost of some additional complexity required for the computation of the common gain approach).

4.2.2 Power Difference as Classification Criterion

In this section new approaches, which are based on the power differences as classification criterion, are introduced. Conceptually, the power difference is used to make decisions for the

selections/mixing between the available signals at each subband-time sample if the power of one signal is greater or lower than the power of the other signal by specific threshold values. This concept is similar to cases that use the power as classification criterion. However, the power difference as classification criterion uses different threshold values to take different decisions based on these threshold values.

In the cases where the power was used as classification criterion, fixed values of α and β were used to combine the available signals. Changing these values adaptively at each subband-time sample based on the power of the available signals might enhance the performance of these approaches; however, it needs more processing time. Therefore, the power difference approaches introduce a simplified version of the adaptively changed factors (e.g. α and β in case 3 to case 7) by using different threshold values, as it will be shown later. Moreover, in the previous cases based on the power, the power of one signal is sometime higher or lower than the power of the other signal by a very small value at each subband-time sample. These small differences should not have a significant effect in making the decision of which output is supposed to be selected in order to find the best trade-off between the noise reduction and the cues preservations. So, the power difference approaches using different threshold values can solve this problem.

The first approach based on the power difference selects between the left and right outputs of the common gain approach and the outputs of left and right beamformers, while the second approach selects between the front left and front right noisy signal and the output of the left and right beamformers.

Case 8

In this approach, the selection/mixing is based on the power differences between the common gain outputs and the outputs of the beamformers at each subband-time sample. The phase of the noisy signal is considered in order to preserve the cues for the directional interferences in terms of the IPD and for preserving the background noise in terms of MSC. The differences between the powers are divided into 4 regions based on 5 different threshold values calculated at each subband. The size of the time window used during the power difference calculation is 50 ms,

while the threshold values are calculated over 200 time-windows, each time-window with a size of 50 ms. Equation (4.9) below has a detailed description of this approach;

$$z_{selected-left}(sb,t)=\begin{cases}(a\mid z_l(sb,t)\mid+(1-a)\mid \hat{y}_l(sb,t)\mid)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{\max-mean,left}(sb)<diff_{left}(sb,t)\ \le threshold_{\max,left}(sb) \\ (b\mid z_l(sb,t)\mid+(1-b)\mid \hat{y}_l(sb,t)\mid)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{mean,left}(sb)<diff_{left}(sb,t)\le\ threshold_{\max-mean,left}(sb) \\ (c\mid z_l(sb,t)\mid+(1-c)\mid \hat{y}_l(sb,t)\mid)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{mean-\min,left}(sb)<diff_{left}(sb,t)\le\ threshold_{mean,left}(sb) \\ (d\mid z_l(sb,t)\mid+(1-d)\mid \hat{y}_l(sb,t)\mid)*e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{\min,left}(sb)\le diff_{left}(sb,t)\le\ threshold_{mean-\min,left}(sb)\end{cases} \tag{4.9}$$

Where;

$$0.5<d<c<b<a<1$$
$$diff_{left}(sb,t)=power(\hat{y}_l(sb,t))-\ \text{power}(z_l(sb,t))$$
$$threshold_{\max,left}(sb)=\max_t(diff_{left}(sb,t))$$
$$threshold_{\max-\text{mean},left}(sb)=(threshold_{\max,left}(sb)+threshold_{\text{mean},left}(sb))/2$$
$$threshold_{\text{mean},left}(sb)=mean_t(diff_{left}(sb,t))$$
$$threshold_{\text{mean}-\min,left}(sb)=(threshold_{\text{mean},left}(sb)+threshold_{\min,left}(sb))/2$$
$$threshold_{\min,left}(sb)=\min_t(diff_{left}(sb,t))$$

Where; a=0.9, b=0.8, c=0.7, d=0.6

Similarly, $y_{selected-right}(sb,t)$ is computed, but this time by using the magnitude of the right common gain output, the phase of the noisy signal at the front right microphone, and the output of the right beamformer. Both beamformers and the common gain approach can be based on either the left or the right front microphone as reference microphone.

When the differences between the power of the common gain outputs and the power of the beamformer output at each subband-time sample is low, the beamformer and the common gain

outputs have approximately similar noise reduction, and then high portion of the magnitude of the common gain output can be used (the red region in Figure 4.3), so that the cues are better preserved in term of ILD and MSC. However, if the power difference is large (the dark green region in Figure 4.3), then low portions of the magnitude of the common gain outputs need to be added to the beamformer output in order to only help partially preserving the cues in terms of ILD and MSC, while keeping good noise reduction level. The region between the maximum power difference and the minimum power difference is quantized into four regions, where the amount of the common gain outputs used is modulated based on the regions (the power differences). If the power difference is in the dark green region, 10% of the magnitude of the common gain is added to 90% of the magnitude of the beamformer output. On the other hand, 40 % of the magnitude of the common gain is added to 60% of the magnitude of the beamformer output if the power differences is in the red region, and so on. Small negative values for the power difference are possible, where the power of the beamformer output is higher than the power of the common gain outputs. Those cases fit in the red region, with more emphasis put on the common gain outputs.

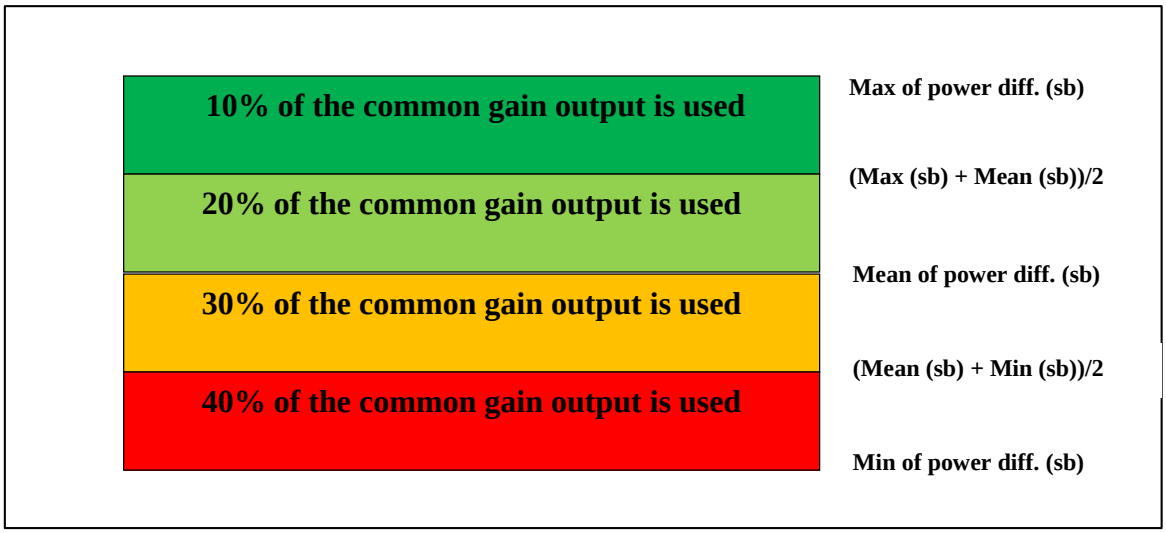

Figure 4.3: Power difference classification criterion at each subband-time window

Case 9

This approach is similar to case 8 , however, this time the selections/mixing are done between the front left or front right noisy signal and the outputs of the beamformers as equation (4.10) shows ;

$$z_{selected-left}(sb,t)=\begin{cases}(a\,|\,z_l(sb,t)\,|+(1-a)\,|\,y_l(sb,t)\,|)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{\max-mean,left}(sb)<diff_{left}(sb,t)\ \le threshold_{\max,left}(sb) \\ (b\,|\,z_l(sb,t)\,|+(1-b)\,|\,y_l(sb,t)\,|)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{mean,left}(sb)<diff_{left}(sb,t)\le\ threshold_{\max-mean,left}(sb) \\ (c\,|\,z_l(sb,t)\,|+(1-c)\,|\,y_l(sb,t)\,|)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{mean-\min,left}(sb)<diff_{left}(sb,t)\le\ threshold_{mean,left}(sb) \\ (d\,|\,z_l(sb,t)\,|+(1-d)\,|\,y_l(sb,t)\,|)e^{j\measuredangle y_l(sb,t)} \\ if \quad threshold_{\min,left}(sb)\le diff_{left}(sb,t)\le\ threshold_{mean-\min,left}(sb)\end{cases} \tag{4.10}$$

Where;

$$0.5<d<c<b<a<1$$
$$diff_{left}(sb,t)=power(y_l(sb,t))-\ power(z_l(sb,t))$$
$$threshold_{\max,left}(sb)=\max_t(diff_{left}(sb,t))$$
$$threshold_{\max-mean,left}(sb)=(threshold_{\max,left}(sb)+threshold_{mean,left}(sb))/2$$
$$threshold_{mean,left}(sb)=mean_t(diff_{left}(sb,t))$$
$$threshold_{mean-\min,left}(sb)=(threshold_{mean,left}(sb)+threshold_{\min,left}(sb))/2$$
$$threshold_{\min,left}(sb)=\min_t(diff_{left}(sb,t))$$

Where; a=0.9,b=0.8, c=0.7, d=0.6

Similarly, $y_{selected-right}(sb)$ is generated, but this time by using the noisy signal at the front right microphone and the output of the right beamformer. The beamforming algorithms can be based on either the left or right front microphones as reference microphone.

4.3 Frequency Dependent Selection and Mixing Approaches

In all previous cases the selections/mixing were frequency independent where the same processing was done on low frequency components (e.g. < 1500Hz) and on high frequency components (e.g.>1500Hz). In order to investigate the effect of doing different processing on the low frequency components and on the high frequency components (e.g. considering that the IPD/ITD is mostly important up to 1500 Hz), the following cases are introduced.

As mentioned previously, for the beamforming simulations the filter bank provided by our sponsor is used. This filter bank has 48 overlapped subbands, each with a bandwidth of 1 kHz and a spacing of 250 Hz between the center frequency of each band (12 kHz divided into 48 subbands, with downsampling from 24 kHz to 1 kHz).

4.3.1 Frequency Dependent Approaches without Classification

Case 10

Based on some previous literature such as (Wightman & Kistler, 1992), the IPD/ITD is more important than the ILD in preserving the cues at low frequencies. Therefore, only the phase of the noisy signals at the front left and front right microphones in the subbands from 1 to 4 are considered (combined with the magnitudes of the beamformer outputs), while the magnitudes and the phases of the beamformer outputs are considered for the other subbands as shown in equations (4.11) and (4.12). In the filter bank provided by our sponsor, the first four subbands have center frequencies 0, 250, 500 and 750 Hz, each with a bandwidth of 1 kHz, thus covering all the frequency components up to 500 Hz. The frequency components up to 500 Hz are the most critical frequencies in preserving the cues in term of IPD/ITD perceptually. The beamformers can be based on either the left or the right front microphones as reference microphone.

$$z_{selected-left}(sb,t) = \begin{cases} |z_l(sb,t)|\, e^{j\angle y_l(sb,t)}, & sb = 1,2,3,4 \\ |z_l(sb,t)|\, e^{j\angle z_l(sb,t)}, & other\ sb \end{cases} \tag{4.11}$$

$$z_{selected-right}(sb,t) = \begin{cases} |z_r(sb,t)|e^{j\angle y_r(sb,t)}, & sb = 1,2,3,4 \\ |z_r(sb,t)|e^{j\angle z_r(sb,t)}, & other\ sb \end{cases} \quad (4.12)$$

Case 11

To improve the performance of case 10 in preserving both the cues of the directional interferers and the background noise, this approach is introduced. For the first eight subbands, the phase of the noisy signal is considered as it is important in preserving the cues for the directional interferers in terms of IPD/ITD. Moreover, the phase of the noisy signal at the first eight subbands are important in preserving the cues for the background noise, as we will see later in the results. On the other hand, the magnitude of the beamformer output is considered for the first eight subbands as the magnitude of the noisy signal does not play a significant role in preserving the ILD cues for the low frequency components. For high frequency components, a small portion β of the magnitude of the noisy signal is added to a larger portion α of the magnitude of the beamformer output. Only the phase of the beamformer output is considered for high frequency components since the phase is not important in preserving IPD/ ITD for the high frequency components. The beamformers can be based on either the left or right front microphones as reference microphone.

The choice of considering the phase of the noisy signal for the first eight subbands come from the design of the filter bank provided by our sponsor. The eighth subband has a center frequency of 1750 Hz and it is the last subband that has the frequency components around and below 1500 Hz. Therefore, by considering the first eight subbands, we consider all the frequency components up to 1500 Hz. As it was mentioned previously, the frequency components up to 1500 Hz are important in preserving the cues for the directional interferers in terms of IPD/ITD, and for the background noise in terms of MSC.

$$z_{selected-left}(sb,t) = \begin{cases} |z_l(sb,t)|e^{j\angle y_l(sb,t)}, & sb \le 8 \\ (\alpha|z_l(sb,t)| + \beta|y_l(sb,t)|)e^{j\angle z_l(sb,t)}, & sb > 8 \end{cases} \quad (4.13)$$

$$z_{selected-right}(sb,t) = \begin{cases} |z_r(sb,t)|e^{j\angle y_r(sb,t)}, & sb \le 8 \\ (\alpha|z_r(sb,t)| + \beta|y_r(sb,t)|)e^{j\angle z_r(sb,t)}, & sb > 8 \end{cases} \quad (4.14)$$

Where $0.5<\alpha<1$ and $0<\beta<0.5$. In this case, we used $\alpha=0.7$, $\beta=0.3$.

In this approach and in the previous one, the magnitude of the common gain could also be used instead of the magnitude of the noisy signal. If the magnitude of the common gain outputs is used, the ΔILD error will be increased since the magnitude of the common gain outputs generates more error in term of the ILD compared with using the magnitude of the noisy signals. While this statement may be surprising since in principle the common gain approach perfectly preserves the cues of all individual directional sources, in practice the fact that the common gain approach is implemented using a filter bank with limited frequency resolution results in non-ideal ΔILD errors when the ILD measurements are performed using a T-F analysis with a finer frequency resolution (as is the case in this thesis). Using the magnitude of the common gain outputs instead of using the magnitude of the noisy signals will improve the SNR-gain for the high frequency components. We have used the simplified version based on the noisy signal just for comparison and evaluation purposes.

Case 12

This approach is an extension of case 11. The main difference between this approach and the approach in case 11 is that for the first eight subbands a small portion β of the magnitude of the noisy signal is added to a larger portion α of the magnitude of the beamformer output. This approach is introduced in order to test the effect of the magnitude of the noisy signals in preserving the spatial perception of the background noise for the low frequency components. The beamformers can be based on either the left or the right front microphones as reference microphone.

$$z_{selected-left}(sb,t)=\begin{cases}(\alpha\mid z_l(sb,t)\mid+\beta\mid y_l(sb,t)\mid)e^{j\measuredangle y_l(sb,t)}, & sb\le 8\\ (\alpha\mid z_l(sb,t))+\beta\mid y_l(sb,t)\mid)e^{j\measuredangle z_l(sb,t)}, & sb>8\end{cases} \tag{4.15}$$

$$z_{selected-right}(sb,t)=\begin{cases}(\alpha\mid z_r(sb,t)\mid+\beta\mid y_r(sb,t)\mid)e^{j\measuredangle y_r(sb,t)}, & sb\le 8\\ (\alpha\mid z_r(sb,t)\mid+\beta\mid y_r(sb,t)\mid)*e^{j\measuredangle z_r(sb,t)}, & sb>8\end{cases} \tag{4.16}$$

α=0.7 , β=0.3

Similar to case 12, the magnitude of the common gain output could also be used instead of the magnitude of the noisy signal. Using the magnitude of the common gain would increase the ΔILD error but it would improve the SNR-gain.

4.3.2 Power Difference as Classification Criterion

The aforementioned frequency independent approaches that used the power differences between the signals as classification criterion are next extended to frequency dependent processing.

Case 13

Since case 8 provides a better trade-off in comparison with case 9 in terms of the noise reduction and cues preservation, as it will be shown later in the results, case 8 is extended to a frequency dependent approach. In this approach, the classification is based on the power differences between the left/right common gain outputs and the outputs of the left and right beamformers at each subband-time sample. The phase of the noisy signal is considered for the low frequency components (e.g., <1500Hz) in order to improve the ITD/IPD. On the other hand, the phase of the beamformer output is considered for the high frequency components (e.g. > 1500 Hz) as the IPD/ITD cues are not important for these frequency components. The differences between the powers are divided into 4 regions based on 5 different threshold values calculated at each subband according to the equations (4.17) and (4.18).

if $sb \le 8$

$$z_{selected-left}(sb,t) = \begin{cases} (a \mid z_l(sb,t) \mid +(1-a) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle y_l(sb,t)} \\ if \quad threshold_{\max-mean,left}(sb) < diff_{left}(sb,t) \le threshold_{\max,left}(sb) \\ \\ (b \mid z_l(sb,t) \mid +(1-b) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle y_l(sb,t)} \\ if \quad threshold_{mean,left}(sb) < diff_{left}(sb,t) \le threshold_{\max-mean,left}(sb) \\ \\ (c \mid z_l(sb,t) \mid +(1-c) \mid \hat{y}_l(sb,t) \mid) * e^{j \measuredangle y_l(sb,t)} \\ if \quad threshold_{mean-\min,left}(sb) < diff_{left}(sb,t) \le threshold_{mean,left}(sb) \\ \\ (d \mid z_l(sb,t) \mid +(1-d) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle y_l(sb,t)} \\ if \quad threshold_{\min,left}(sb) \le diff_{left}(sb,t) \le threshold_{mean-\min,left}(sb) \end{cases} \tag{4.17}$$

if $sb > 8$

$$z_{selected-left}(sb,t) = \begin{cases} (a \mid z_l(sb,t) \mid +(1-a) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad threshold_{\max-mean,left}(sb) < diff_{left}(sb,t) \le threshold_{\max,left}(sb) \\ \\ (b \mid z_l(sb,t) \mid +(1-b) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad threshold_{mean,left}(sb) < diff_{left}(sb,t) \le threshold_{\max-mean,left}(sb) \\ \\ (c \mid z_l(sb,t) \mid +(1-c) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad threshold_{mean-\min,left}(sb) < diff_{left}(sb,t) \le threshold_{mean,left}(sb) \\ \\ (d \mid z_l(sb,t) \mid +(1-d) \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad threshold_{\min,left}(sb) \le diff_{left}(sb,t) \le threshold_{mean-\min,left}(sb) \end{cases} \tag{4.18}$$

Where;

$$0.5 < d < c < b < a < 1$$
$$diff_{left}(sb,t) = power(\hat{y}_l(sb,t)) - \text{power}(z_l(sb,t))$$
$$threshold_{\max,left}(sb) = \max_t(diff_{left}(sb,t))$$
$$threshold_{\max-\text{mean},left}(sb) = (threshold_{\max,left}(sb) + threshold_{\text{mean},left}(sb))/2$$
$$threshold_{\text{mean},left}(sb) = mean_t(diff_{left}(sb,t))$$
$$threshold_{\text{mean}-\min,left}(sb) = (threshold_{\text{mean},left}(sb) + threshold_{\min,left}(sb))/2$$
$$threshold_{\min,left}(sb) = \min_t(diff_{left}(sb,t))$$

a=0.9, b=0.8, c=0.7, d=0.6

Similarly, $y_{selected-right}(sb,t)$ is computed, but this time by using the magnitude of the right output of the common gain approach, the phase of the noisy signal at the front right microphone, and the output of the right beamformer. The beamformers and the common gain approach can be based on either the left or the right front microphones as reference microphone.

4.3.3 Power as Classification Criterion

Case 14

An extension of case 6, which uses the power as classification criterion, is introduced. In this approach the phase of the noisy signals at the front left and front right microphones are considered for low frequency components, while the phases for the outputs of the beamformers are considered for the high frequency components. The magnitude is mixed in a way similar to case 6. Equation (4.19) gives the detailed description of this approach. The beamformers and the common gain approach can be based on either the left or the right front microphones as reference microphone. The choice of extending case 6 instead of extending the other frequency independent approaches that are based on the power is because case 6 demonstrates the best trade-off between noise reduction and cues preservations, as it will be shown later in the results.

$$sb \le 8$$

$$z_{selected-left}(sb,t) = \begin{cases} (\alpha \mid z_l(sb,t) \mid + \beta \mid \hat{y}_l(sb,t) \mid) * e^{j \measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) < power(\hat{y}_l(sb,t)) \\ \\ (\beta \mid z_l(sb,t) \mid + \alpha \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle y_l(sb,t)} \\ if \quad power(z_l(sb,t)) > power(\hat{y}_l(sb,t)) \end{cases} \tag{4.19}$$

$$sb > 8$$

$$z_{selected-left}(sb,t) = \begin{cases} (\alpha \mid z_l(sb,t) \mid + \beta \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad power(z_l(sb,t)) < power(\hat{y}_l(sb,t)) \\ \\ (\beta \mid z_l(sb,t) \mid + \alpha \mid \hat{y}_l(sb,t) \mid) e^{j \measuredangle z_l(sb,t)} \\ if \quad power(z_l(sb,t)) > power(\hat{y}_l(sb,t)) \end{cases}$$

Similarly, the output for the right channel is selected/mixed by using the right common gain output, the right beamformer output, and the phase of the noisy signal at the frontal right microphone. In this approach we used α=0.7 and β=0.3.

4.3.4 Coherence as Classification Criterion

In this section, another classification criterion is used for mixing and selection at each subband-time sample: the coherence between the two sides. The coherence function (Challis & Kitney, 1991) is a frequency dependent function that measures the similarity between two different signals. The coherence is defined to be a normalized cross-power spectrum; therefore, it is between zero and one. Two signals are uncorrelated when the coherence between them is zero. On the other hand, when the coherence is unity, the signals are fully correlated, meaning that one can be predicted by the other through linear filtering. The coherence generates complex values and can be calculated according to the equation(4.20) ;

$$C_{x,y} = \frac{P_{x,y}}{\sqrt{P_x P_y}} \tag{4.20}$$

$P_{x,y}$ is the cross-power spectrum, P_x and P_y are auto-power spectra.

Another commonly used measurement is the magnitude-squared coherence (MSC) (Challis & Kitney, 1991), which produces real values and is used to facilitate the analysis of the coherence.

$$MSC = \frac{|P_{x,y}|^2}{P_x P_y} \tag{4.21}$$

Therefore, two frequency dependent approaches are introduced below: one of them based on the MSC and the other approach based on the magnitude and the phase of the coherence between the signals.

Case 15

In this approach, the MSC is used for selections/mixing between the common gain outputs and the outputs of the left and right beamformers as the following:

1. Magnitude-Squared Coherences (MSC) between the left /right common gain outputs and the outputs of the left and right beamformers are calculated according to equations (4.22) and (4.23).

$$MSC_{\hat{y}_l, z_l}(sb,t) = \left| \frac{P_{\hat{y}_l, z_l}(sb,t)}{\sqrt{P_{\hat{y}_l}(sb,t) P_{z_l}(sb,t)}} \right|^2 \tag{4.22}$$

$$MSC_{\hat{y}_r, z_r}(sb,t) = \left| \frac{P_{\hat{y}_r, z_r}(sb,t)}{\sqrt{P_{\hat{y}_r}(sb,t) P_{z_r}(sb,t)}} \right|^2 \tag{4.23}$$

Where;

$P_{\hat{y}_l, z_l}$: Cross power spectrum density between the left common gain output and the left beamformer output.

$P_{\hat{y}_r, z_r}$: Cross power spectrum density between the right common gain output and the right beamformer output.

$P_{\hat{y}_l}$: Power spectrum density for the left common gain output.

$P_{\hat{y}_r}$: Power spectrum density for the right common gain output.

P_{z_l} : Power spectrum density for the output of the left beamformer.

P_{z_r} : Power spectrum density for the output of the right beamformer.

$MSC_{\hat{y}_l,z_l}$: Magnitude-Squared Coherence between the left common gain output and the left beamformer output.

$MSC_{\hat{y}_r,z_r}$: Magnitude-Squared Coherence between the right common gain output and the right beamformer output.

2. Left and right threshold values are calculated by taking the mean of $MSC_{\hat{y}_l,z_l}$ and $MSC_{\hat{y}_r,z_r}$ at each subband, respectively, as in equations (4.24) and (4.25).

$$threshold_{left}(sb) = mean(MSC_{\hat{y}_l,z_l}(sb,t)) \quad (4.24)$$

$$threshold_{right}(sb) = mean(MSC_{\hat{y}_r,z_r}(sb,t)) \quad (4.25)$$

Each “instantaneous” MSC is calculated in a time window of 50 ms at each subband. After that, the averages of the all MSCs are computed at each subband, over 200 windows of MSCs, where each window is 50 ms.

3. Left and right output signals are selected/mixed based on the following classification criteria;

$$
\begin{aligned}
&if\ \ sb \le 8 \\
&z_{selected-left}(sb,t) = \begin{cases} |z_l(sb,t)|\, e^{j\angle(y_l(sb,t))} \\ if \quad MSC_{\hat{y}_l,z}(sb,t) < threshold_{left}(sb) \\ (\beta\,|z_l(sb,t)| + \alpha\,|\hat{y}_l(sb,t)|)e^{j\angle(y_l(sb,t))} \\ if \quad MSC_{\hat{y}_l,z}(sb,t) \ge threshold_{left}(sb) \end{cases} \\
&if\ \ sb > 8 \\
&z_{selected-left}(sb,t) = \begin{cases} (\alpha\,|z_l(sb,t)| + \beta\,|\hat{y}_l(sb,t)|)e^{j\angle z_l(sb,t)} \\ if \quad MSC_{\hat{y}_l,z}(sb,t) < threshold_{left}(sb) \\ (\beta\,|z_l(sb,t)| + \alpha\,|\hat{y}_l(sb,t)|)e^{j\angle z_l(sb,t)} \\ if \quad MSC_{\hat{y}_l,z}(sb,t) \ge threshold_{left}(sb) \end{cases}
\end{aligned}
\tag{4.26}
$$

$$
\begin{aligned}
&if\ \ sb \le 8 \\
&z_{selected-right}(sb,t) = \begin{cases} |z_r(sb,t)|\, \mathrm{e}^{j\angle(\mathrm{y}_r(sb,t))} \\ \text{if} \quad MSC_{\hat{y}_r,z}(sb,t) < \text{threshold}_{right}(\text{sb}) \\ (\beta\,|z_r(sb,t)| + \alpha\,|\hat{\mathrm{y}}_r(sb,t)|) * \mathrm{e}^{j\angle(\mathrm{y}_r(sb,t))} \\ \text{if} \quad MSC_{\hat{y}_r,z}(sb,t) \ge \text{threshold}_{right}(\text{sb}) \end{cases} \\
&if\ \ sb > 8 \\
&z_{selected-right}(sb,t) = \begin{cases} (\alpha\,|z_r(sb,t)| + \beta\,|\hat{y}_r(sb,t)|)e^{j\angle z_r(sb,t)} \\ if \quad MSC_{\hat{y}_r,z}(sb,t) < threshold_{right}(sb) \\ (\beta\,|z_r(sb,t)| + \alpha\,|\hat{y}_r(sb,t)|)e^{j\angle z_r(sb,t)} \\ if \quad MSC_{\hat{y}_r,z}(sb,t) \ge threshold_{right}(sb) \end{cases}
\end{aligned}
\tag{4.27}
$$

Both the beamforming and the common gain approaches can be based on either the left or right front microphones as reference microphone, and the values of α and β were set to 0.7 and 0.3, respectively.

In this case, for low frequency components only the phase of the noisy signal is considered since it is important in preserving the IPD/ ITD cues. However, the magnitude of both the common gain outputs and the outputs of the beamformers are considered by comparing the values of MSC

with the threshold value at each subband. If the MSC is lower than the threshold value for a certain subband, this means that the beamformer and common gain outputs are not similar (beamformer likely produces more noise reduction), so only the magnitude of the output of the beamformer is considered. On the other hand, if the MSC is higher than the threshold value for a certain subband, then these two signals are highly correlated. Therefore, we can add a portion of the magnitude of the common gain output to the magnitude of the beamformer output to preserve the cues in terms of ILD and MSC. The phase and the magnitude of the common gain output are important in preserving the spatial perception for the background noise, especially for frequency components lower than 1500 Hz.

For high frequency components, the phase of the beamformer output is considered as the phase does not have an effect in preserving the IPD/ITD cues. On the other hand, both the magnitudes of the common gain output and the beamformer output are considered, as shown in equations (4.26) and (4.27) in order to improve the preservation of ILD cues while providing a good SNR gain.

Case 16

Another approach, which is based on the coherence instead of the magnitude-squared coherence in order to take the advantage of the phase of the coherence, is proposed. The phase of the coherence is the difference between the phases of the two signals being compared, and it is between $-\pi$ and π. If the phase of the coherence is zero or near zero then the signals are highly time-aligned and vice versa.

In this case, the classification criterion of the left and right output signals is based on the magnitude and the phase of the coherence between the outputs of the common gain approach and the outputs of the beamformers. The beamforming and the common gain approaches can be based on either the left or the right front microphones as reference microphone. The signal's magnitude selection/mixing in this approach is very similar to case 15. However, the main difference is in the phase selection. Since the phase is not important in preserving the IPD/ITD cues for high frequency components, we have considered only the phase of the beamformers outputs. On the other hand,

the phase of the noisy signal is important in preserving the IPD/ITD cues for low frequency components in addition to the spatial perception of the background noise; however, it has an effect in reducing the SNR-gain. Therefore, the outputs are selected/mixed based on the following algorithm;

1. The complex coherence between the left and right common gain outputs and the left and right beamformers outputs ($C_{\hat{y}_l,z_l}$ / $C_{\hat{y}_r,z_r}$) is calculated according to equations (4.28) and (4.29).

$$C_{\hat{y}_l,z_l}(sb,t)=\frac{P_{\hat{y}_l,z_l}(sb,t)}{\sqrt{P_{\hat{y}_l}(sb,t)P_{z_l}(sb,t)}} \tag{4.28}$$

$$C_{\hat{y}_r,z_r}(sb,t)=\frac{P_{\hat{y}_r,z_r}(sb,t)}{\sqrt{P_{\hat{y}_r}(sb,t)P_{z_r}(sb,t)}} \tag{4.29}$$

Where;

$P_{\hat{y}_l,z_l}$: Cross power spectrum density between the left common gain output and the left beamformer output.

$P_{\hat{y}_r,z_r}$: Cross power spectrum density between the right common gain output and the right beamformer output.

$P_{\hat{y}_l}$: Power spectrum density for the left common gain output.

$P_{\hat{y}_r}$: Power spectrum density for the right common gain output.

P_{z_l} : Power spectrum density for the output of the left beamformer.

P_{z_r} : Power spectrum density for the output of the right beamformer.

$C_{\hat{y}_l,z_l}$: Complex coherence between the left common gain output and the left beamformer output.

$C_{\hat{y}_r,z_r}$: Complex coherence between the right common gain output and the right beamformer output.

2. Left and right threshold values are calculated by taking the mean of the magnitude of the complex $C_{\bar{y}_l,z_l}$ and $C_{\bar{y}_r,z_r}$ at each subband, respectively, as it is shown in equations (4.30) and (4.31).

$$threshold_{left}(sb) = mean(|\, C_{\bar{y}_l,z_l}(sb,t)\,|) \tag{4.30}$$

$$threshold_{right}(sb) = mean(|\, C_{\bar{y}_r,z_r}(sb,t)\,|) \tag{4.31}$$

3. The magnitudes of the left and right output for the low frequency and high frequency components are selected/mixed according to equations (4.32) and (4.33).

$$\text{if } sb \le 8$$

$$|\, z_{selected-left}(sb,t)\,| = \begin{cases} |\, z_l(sb,t)\,| \\ if \quad |\, C_{\bar{y}_l,z_l}(sb,t)\,| < threshold_{left}(sb) \\ \\ \alpha\,|\, z_l(sb,t)\,| + \beta\,|\, \bar{y}_l(sb,t)\,| \\ if \quad |\, C_{\bar{y}_l,z_l}(sb,t)\,| \ge threshold_{left}(sb) \end{cases}$$

$$\text{if } sb > 8 \tag{4.32}$$

$$|\, z_{selected-left}(sb,t)\,| = \begin{cases} \alpha\,|\, z_l(sb,t)\,| + \beta\,|\, \bar{y}_l(sb,t)\,| \\ if \quad |\, C_{\bar{y}_l,z_l}(sb,t)\,| < threshold_{left}(sb) \\ \\ \beta\,|\, z_l(sb,t)\,| + \alpha\,|\, \hat{y}_l(sb,t)\,| \\ if \quad |\, C_{\bar{y}_l,z_l}(sb,t)\,| \ge threshold_{left}(sb) \end{cases}$$

$$
\begin{aligned}
&\text{if } sb \le 8 \\
&|z_{selected-right}(sb,t)| = \begin{cases} |z_r(sb,t)| \\ \text{if} \quad |C_{\tilde{y}_r,z_r}(sb,t)| < threshold_{right}(sb) \\ \\ \alpha|z_r(sb,t)| + \beta|\hat{y}_r(sb,t)| \\ \text{if} \quad |C_{\tilde{y}_r,z_r}(sb,t)| \ge threshold_{right}(sb) \end{cases} \\
&\text{if } sb > 8 \\
&|z_{selected-right}(sb,t)| = \begin{cases} \alpha|z_r(sb,t)) + \beta|\hat{y}_r(sb,t)| \\ \text{if} \quad |C_{\tilde{y}_r,z_r}(sb,t)| < threshold_{right}(sb) \\ \\ \beta|z_r(sb,t)| + \alpha|\hat{y}_r(sb,t)| \\ \text{if} \quad |C_{\tilde{y}_r,z_r}(sb,t)| \ge threshold_{right}(sb) \end{cases}
\end{aligned}
\tag{4.33}
$$

The value of α and β were set to 0.7 and 0.3, respectively. As it was mentioned previously, the magnitude of the common gain output does not play a significant role in preserving the cues of the directional interference for the low frequency components. Consequently, even though it plays a role in preserving the spatial perception for the background noise, we have put less emphasis on preserving the magnitude cues for the frequency components lower than 1500Hz so that we can improve the SNR.

4. The phase of the left and the right output is selected based on equations (4.34) and (4.35).

$$
\begin{aligned}
&\text{if } sb \le 8 \\
&\measuredangle(z_{selected-left}(sb,t)) = \begin{cases} \measuredangle(y_l(sb,t)), & |\measuredangle(C_{\tilde{y}_l,z_l}(sb,t))| > \mu\pi \\ \measuredangle(z_l(sb,t)), & |\measuredangle(C_{\tilde{y}_l,z_l}(sb,t))| < \mu\pi \end{cases} \\
&\text{if } sb > 8 \\
&\measuredangle(z_{selected-left}(sb,t)) = \measuredangle(z_l(sb,t))
\end{aligned}
\tag{4.34}
$$

$$
\begin{aligned}
&if\ \ sb \le 8 \\
&\measuredangle(z_{selected-right}(sb,t)) = \begin{cases} \measuredangle(y_r(sb,t)), & |\measuredangle(C_{\bar{y}_r,z_r}(sb,t))| > \mu\pi \\ \measuredangle(z_r(sb,t)), & |\measuredangle(C_{\bar{y}_r,z_r}(sb,t))| < \mu\pi \end{cases} \\
&if\ \ sb\ >\ 8 \\
&\measuredangle(z_{selected-right}(sb,t)) = \measuredangle(z_r(sb,t))
\end{aligned}
\tag{4.35}
$$

Where $\mu < 0.5$, and here we have used 0.1.

As the phase information does not play an essential role in preserving the cues for the directional interference for the high frequency components (e,g, >1500Hz), only the phases of the beamformers outputs are considered for those frequencies. On the other hand, for the low frequency components (e.g. <1500Hz), if the phase of the coherence is lower than a certain threshold value (in this case, 0.1π), then the phases of the beamformers outputs would be considered since they do not differ significantly with the phases of the original noisy signals. In case the phase of the coherence is higher than the threshold value (0.1π), the phases of the noisy signals would be considered as there are large differences between the phases of the noisy signals and the phases of the beamformers outputs.

Chapter 5 Simulation Results for the Cues Preservation Metrics and Discussion.

5.1 Experimental Setup

Two beamforming algorithms are used below to test their performance in terms of cues preservation measurements. Based on an informal listening of the binaural output signals, one of the beamforming algorithms is supposed to preserve the binaural cues for the interferers partially, while the other beamformer is supposed to significantly change the binaural spatial cues for the interferers.

In this section, a subband Generalized Sidelobe Canceler (GSC) algorithm, which was explained in section 2.3, is used. The fixed beamformer and the blocking matrix of the GSC were implemented by first aligning the signals from the different microphones such that the target component is aligned in all the signals. This is achieved by applying target equalization/alignment coefficients (or subband filtering) to the different microphone signals. The GSC beamformer used three coefficients for target equalization/alignment obtained directly from the target signal components at the microphones (ideal), and with a causality delay of one sample for the target equalization/alignment. Each sample delay in the subband processing at the downsampled rate of 1 kHz introduces a 1 ms delay. The equalization/alignment coefficients were computed from the ratio of the cross power spectrum density between the reference microphone target signal and the other microphones' target signal, over the auto power spectrum density of the other microphones' target signal. The causality delay was used to delay the reference microphone noisy signal, and the equalization/alignment coefficients were used to filter the non-reference microphones noisy signals to equalize/align the target component of those signals. By averaging all the target aligned/equalized noisy signals and the delayed reference microphone noisy signal, we obtain the output of the fixed beamformer (i.e., it is like a delay and sum beamformer), with unit gain in the target direction. And by subtracting the aligned/equalized target noisy signals from the delayed reference microphone noisy signal, we obtain the output signals of a blocking matrix (i.e., it is like a Griffiths-Jim blocking matrix). In the Active Noise Canceller (ANC) part of the GSC, the fixed

beamformer output was delayed by one sample causality delay. The ANC used three coefficients per subband. These ANC coefficients were estimated by a least-squares algorithm over the whole length of the processed signals (10 sec.). A noise/interferers estimate was generated by filtering the output signals of the blocking matrix by the ANC coefficients. Finally, the noise/interferers estimate was subtracted from the delayed fixed beamformer output to obtain the GSC output.

The first beamforming algorithm of this Chapter uses two parallel bilateral beamformers as shown in Figure 5.1 . Each beamformer uses the subband Generalized Sidelobe Canceler (GSC) algorithm, with two microphones, and each generates a monaural output (left and right monaural outputs). The first beamformer is designed to be on the left side (left ear) and uses the front left microphone as a reference channel, and the second beamformer is designed to be on the right side (right ear) and uses the front right microphone as the reference channel. There is no wireless link between the left and right channels in this scenario. For evaluation purposes, a 10 second recording is used sampled at 24 kHz sampling rate with two speech signals, i.e., a target from 45 degrees and an interference from 315 degrees, as well as some background diffuse-like noise (soft, 14 dB lower than the power of the target and the interferer). Informal listening of the results for this beamformer shows that the binaural cues for the target and the interference signals are well preserved under this design.

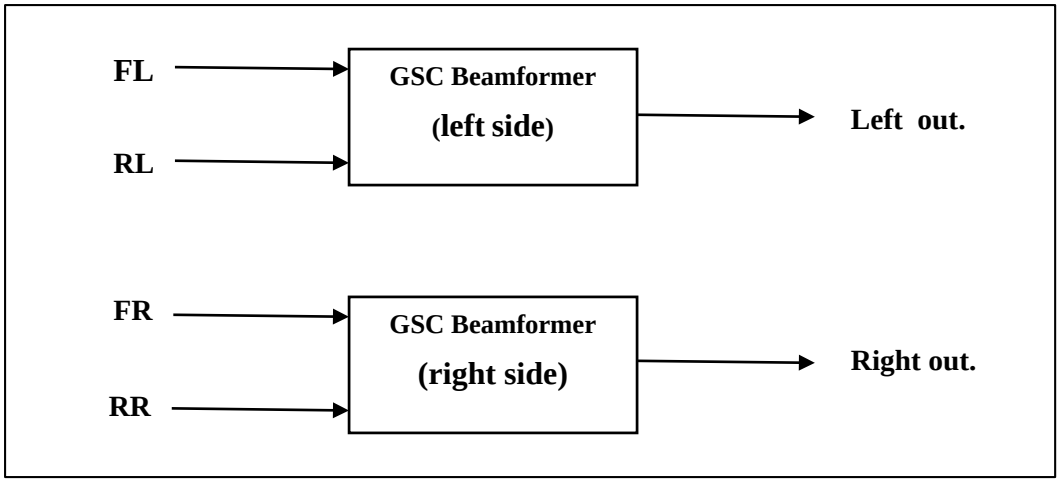

Figure 5.1: Two parallel bilateral beamformers (FL=front left, RL=rear left, FR = front right, RR=rear right)

The second beamforming algorithm is based on two parallel binaural beamformers as shown in Figure 5.2. Each beamformer also uses the subband Generalized Sidelobe Canceler (GSC) algorithm to generate a monaural output, and each has input signals from four microphones (2+2 configuration, two microphones from each ear, with a wireless link between them). The first beamformer uses the front left microphone as a reference channel, and the second beamformer uses the front right microphone as the reference channel. For evaluation purposes, the same scenario previously described is used. The informal listening of the outputs for this beamformer shows that the binaural cues for the interference signal are significantly changed. For the binaural input interference signals on the left and right reference microphones, the DOA appears to be from -45 degree; however, for the output binaural interference signals (from the left and right output signals), the DOA appears to be near 45 degree, i.e., the same direction as the target signal.

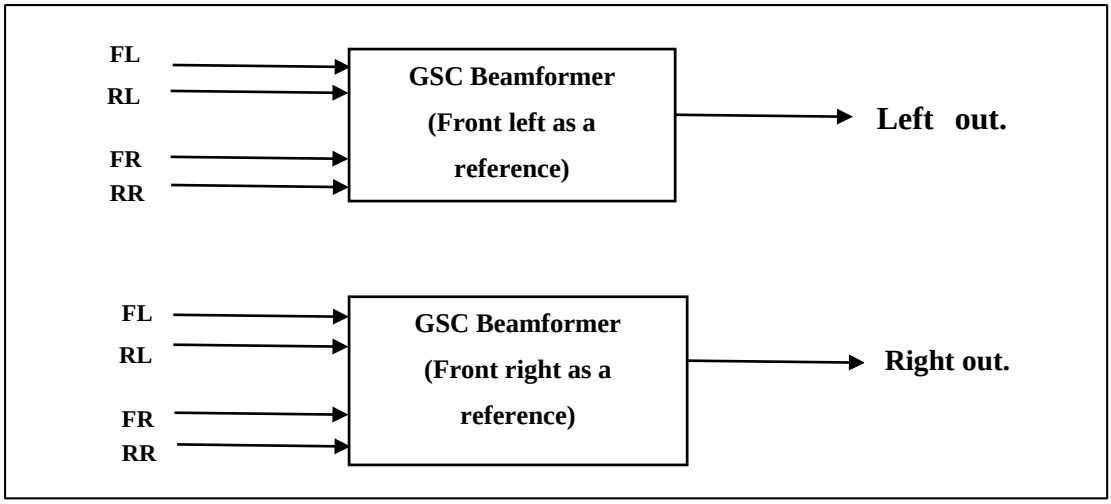

Figure 5.2**: Two parallel binaural beamformers**

5.2 Simulation Results

In the Figure 5.3 and Figure 5.4, the measured ILD_x^{in}, ILD_x^{out}, ILD_v^{in}, ILD_v^{out}, ΔILD_x, and ΔILD_v for the first and the second beamforming algorithms are shown, respectively. For the target signal, both beamformers are capable to preserve the ILD cues (left column of Figure 5.3 and Figure 5.4). The first beamforming algorithm, which uses two bilateral beamformers, is also capable to partially preserve the ILD of the interference signal at -45 degrees, as shown in the right column of Figure 5.3, with an ILD error mostly within 0 and 15 dB. However, the second beamforming algorithm, which has two binaural beamformers, changes the ILD cues for the interference signal as shown in the right column of Figure 5.4, with an ILD error often greater than 20 dB. As it is mentioned in section 3.2, the ILD is more suitable for the high frequency components, i.e., higher than 1.5 kHz. As the frequency increases (i.e. the wavelength decrease), the head acts as an obstacle and causes a head shadow effect, which creates a difference in the intensities between the left and right signals. However, the low frequency components below 1.5 kHz have longer wavelength that will bend around the head. Therefore, the differences in the intensities between the left and right signals are small at low frequencies and the ILD does not provide us with useful information. Throughout this thesis, we have represented the ILD (in frequency dependent plots) up to 12 kHz; however, the most important frequency components are higher than 1.5 kHz.

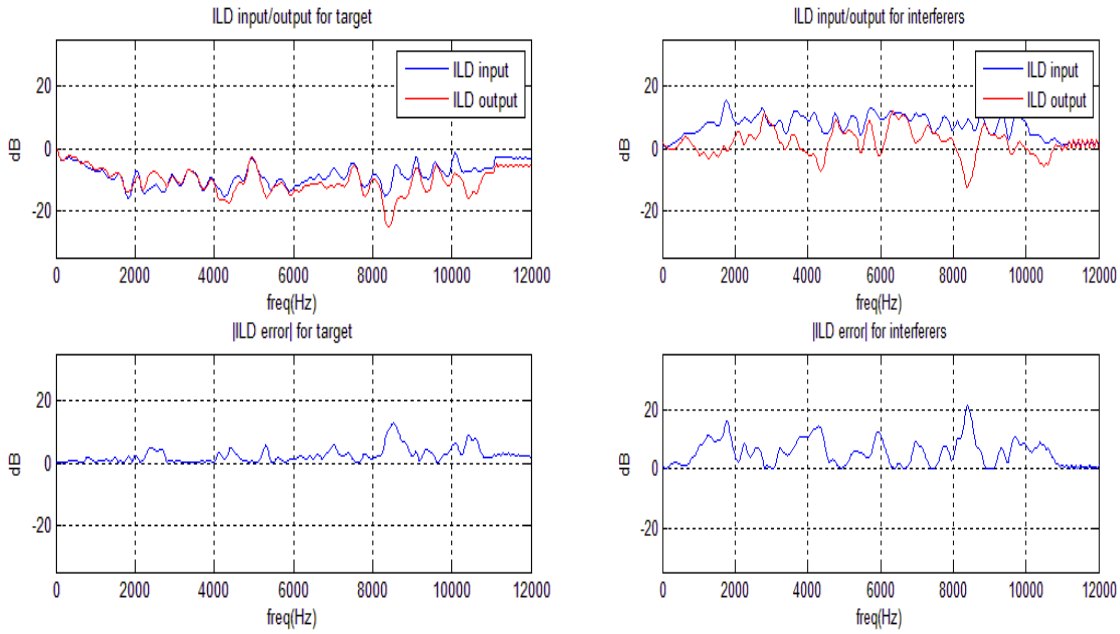

Figure 5.3: ILD for both the target and the interference components using two bilateral beamformers

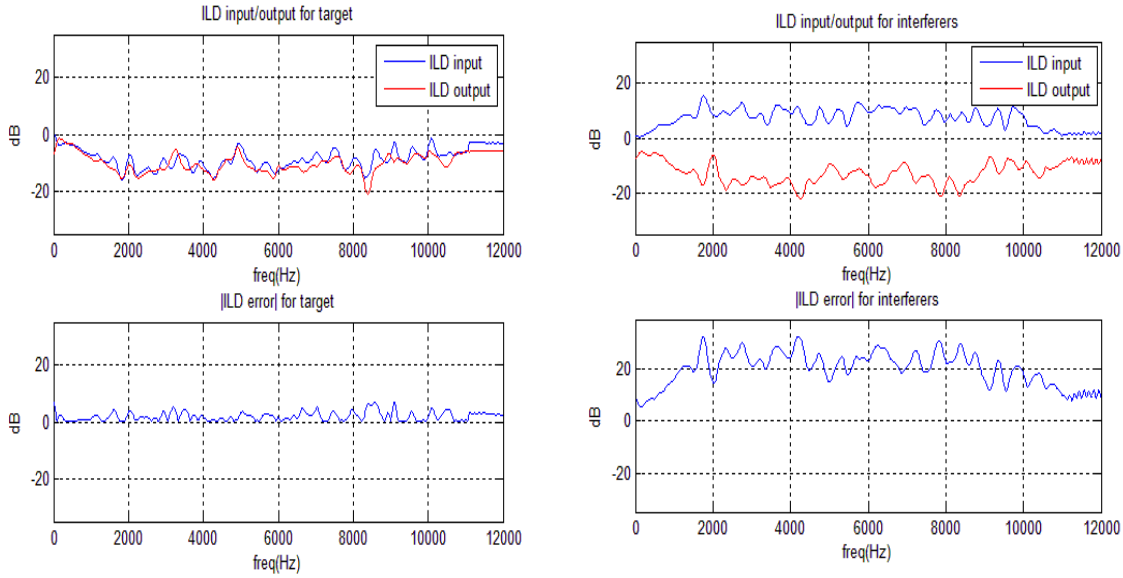

Figure 5.4: ILD for both the target and the interference components using two binaural beamformers

Figure 5.5 and Figure 5.6 show the wrapped and the unwrapped versions of IPD_x^{in}, IPD_x^{out}, IPD_v^{in}, and IPD_v^{out}, in addition to ΔIPD_x and ΔIPD_v for both the first and the second beamforming algorithms, respectively. Comparing Figure 5.5 and Figure 5.6, we see that the bilateral beamforming algorithm is capable to better preserve the IPD cues of the interference signal compared to the binaural beamforming algorithm. Both algorithms produce a similar performance for the IPD cues of the target. The weighted-ΔIPD, which is a version of ΔIPD with a frequency dependent scaling as previously described, is shown in Figure 5.7. In this acoustic scenario, using either the ΔIPD or the weighted-ΔIPD will provide us with similar information. This is because the tested acoustic scenario has only one interference, and the weighted-ΔIPD measurement shows some advantage over the ΔIPD only in cases of multiple interferers.

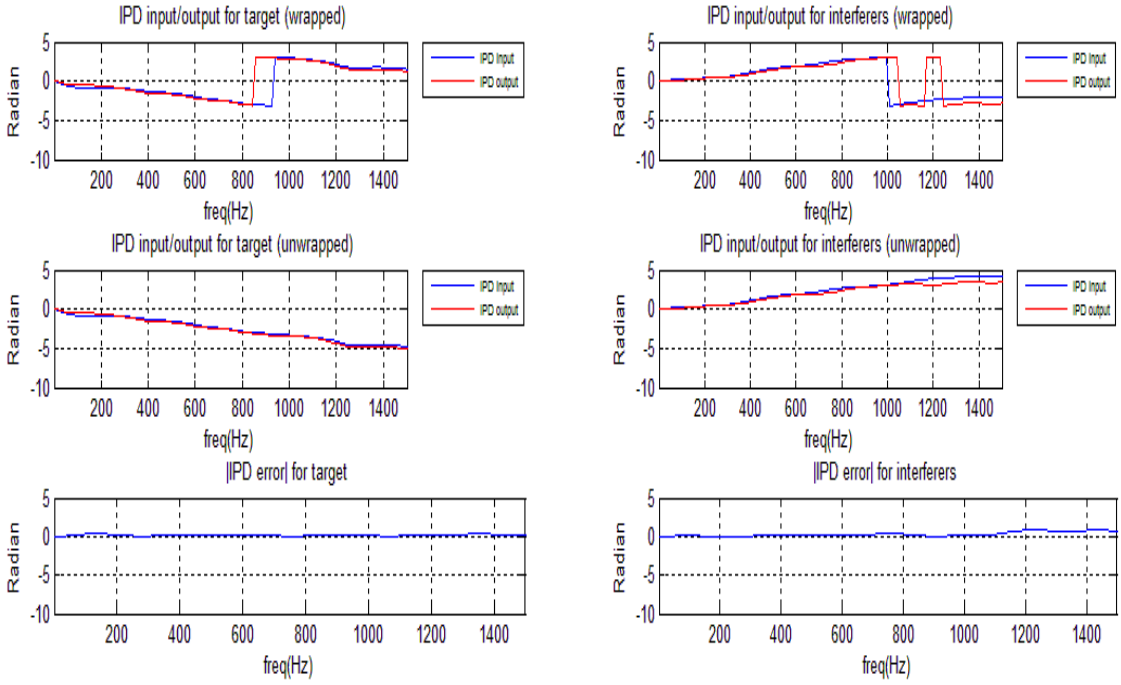

Figure 5.5: IPD for both the target and interference components using two bilateral beamformers

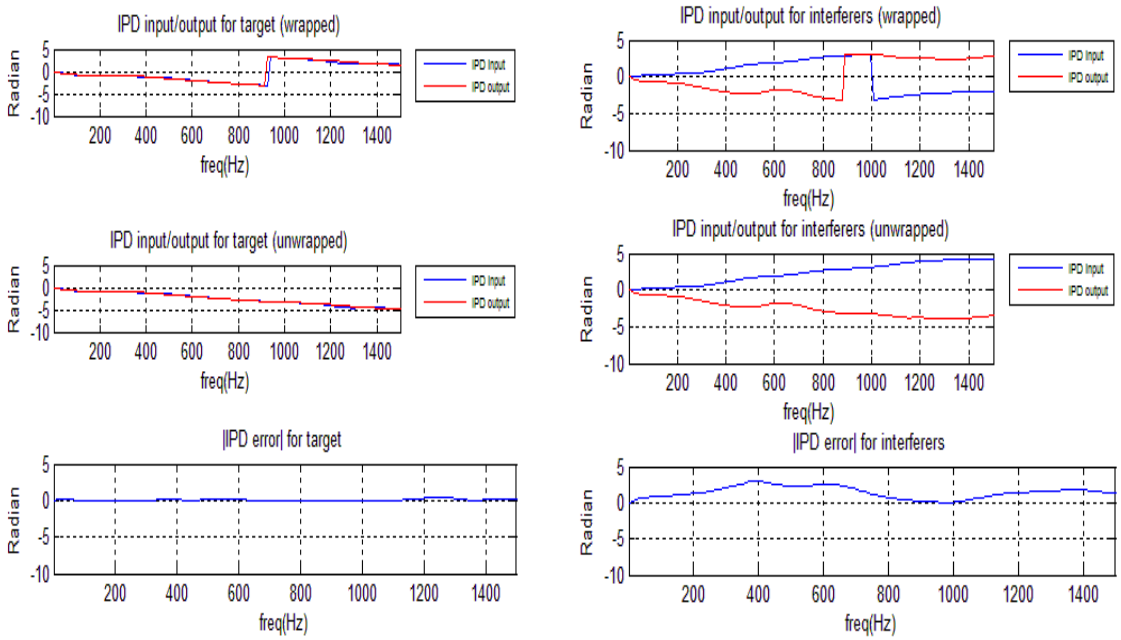

Figure 5.6: IPD for both the target and interference components using two binaural beamformers

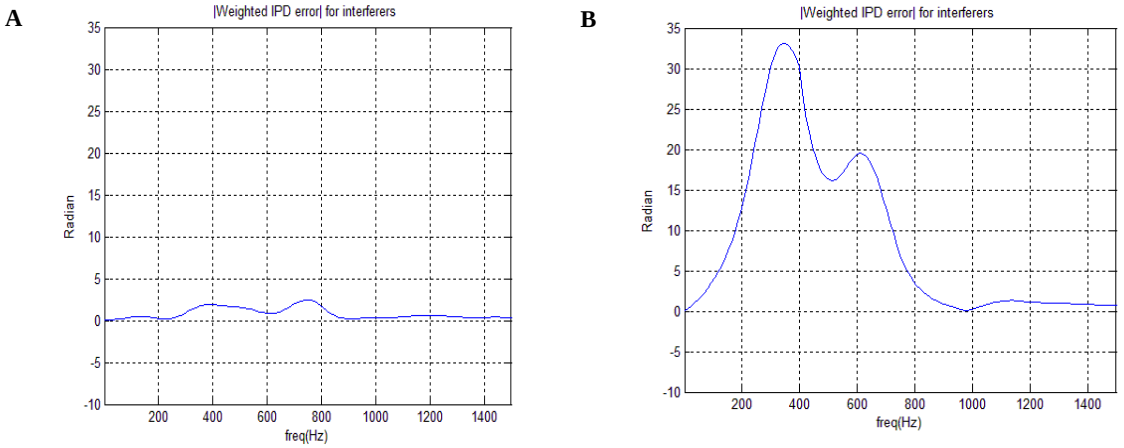

Figure 5.7: Weighted-ΔIPD of interference components using A) two bilateral beamformers B) two binaural beamformer

In Figure 5.8, Figure 5.9, and Figure 5.10, the ITD_x^{in}, ITD_x^{out}, ITD_v^{in}, ITD_v^{out}, ΔITD_x, and ΔITD_v curves for the two bilateral beamformers, using the three aforementioned frequency dependent approaches for ITD computation (group delay, phase delay, and local slope), are shown.

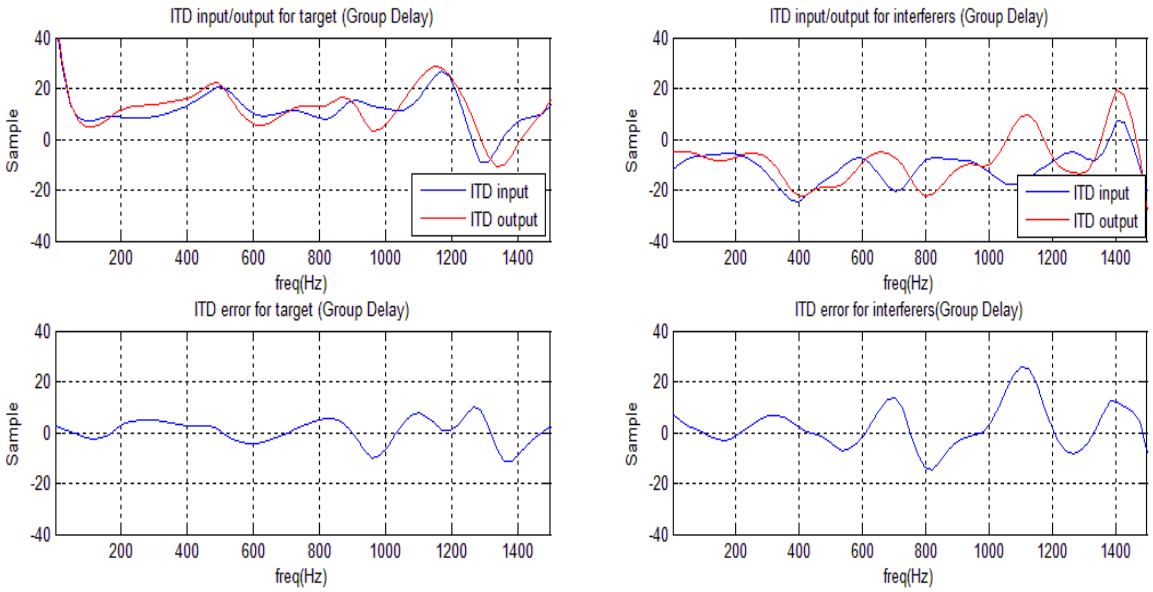

Figure 5.8: ITD using the group delay approach for both the target and interference components using two bilateral beamformers

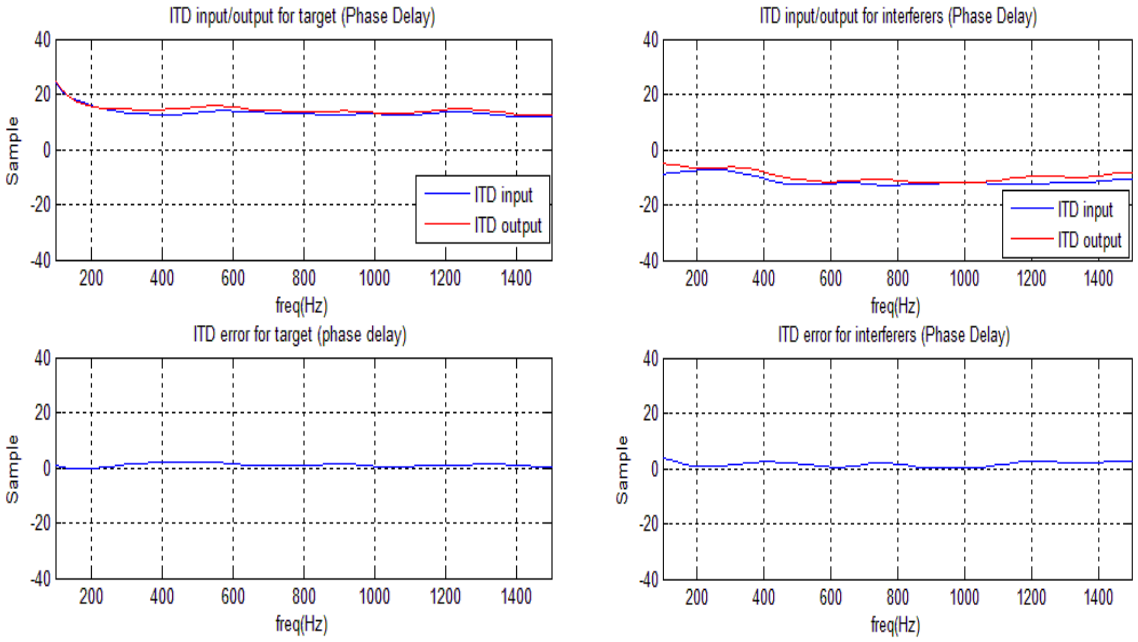

Figure 5.9: ITD using the phase delay approach for both the target and interference components using two bilateral beamformers

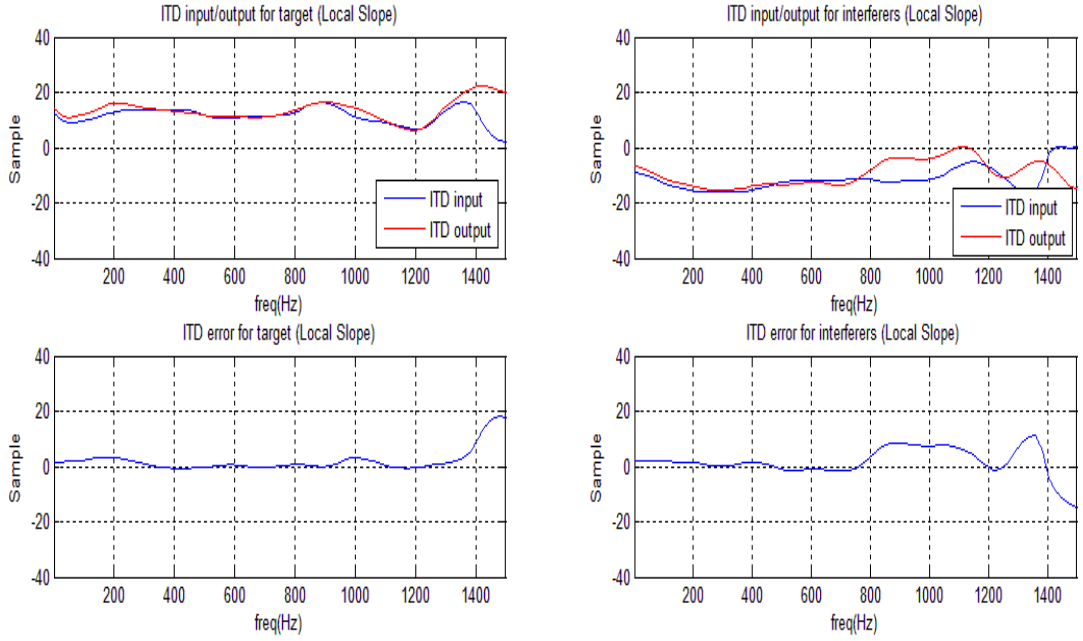

Figure 5.10: ITD using the local slope approach for both the target and interference components using two bilateral beamformers

Similarly, Figure 5.11, Figure 5.12, and Figure 5.13 show the ITD_x^{in}, ITD_x^{out}, ITD_v^{in}, ITD_v^{out}, ΔITD_x, and ΔITD_v curves for the binaural beamforming algorithm. Comparing with the previous figures, we see that for the interference signal the ITD is much better preserved with the bilateral beamformers than with the binaural beamformers. For the target signal the ITD is preserved similarly with both types of beamformers. There are also some differences in the results obtained using the three approaches of measuring the ITD. In Figure 5.11, it is noticeable that at a few frequencies the ITD^{in} and ITD^{out} are higher than 100 samples, which is much higher than the maximum possible delay between the left and right signals, even taking into account the head shadow effect and the changes in the speed of sound inside the head. These high ITD values are considered as outliers generated from calculating the derivative of the IPD numerically, and could be eliminated by an outlier removal algorithm. Furthermore, ΔITD_x and ΔITD_v can exceed the maximum lag between the left and the right channels since it is the difference between the input and the output ITDs, so opposite signs can increase the error.

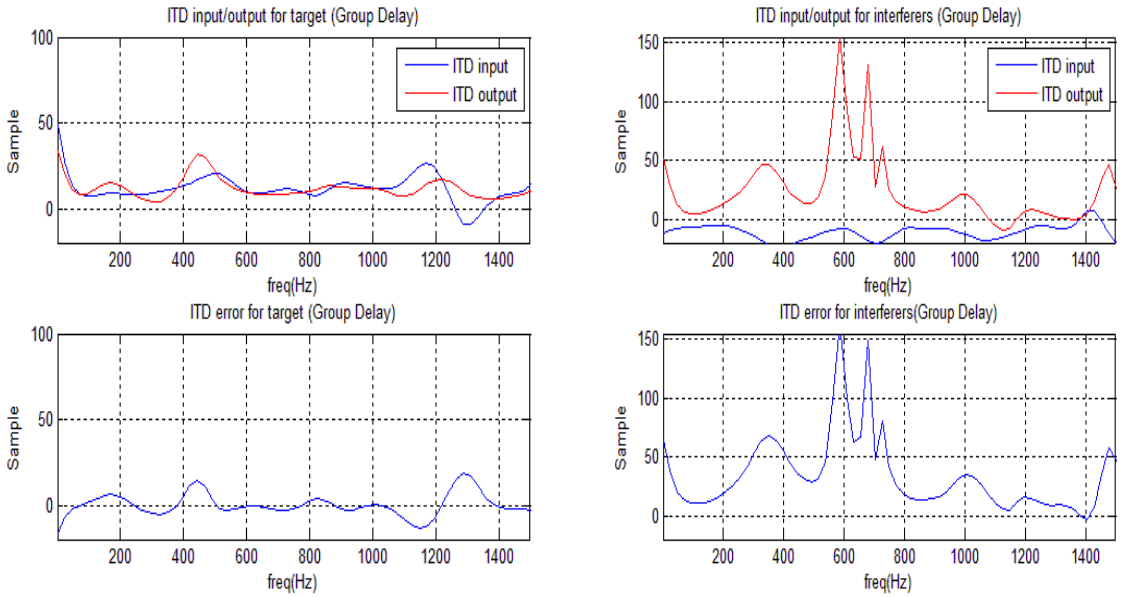

Figure 5.11: ITD using the group delay approach for both the target and interference components using two binaural beamformers

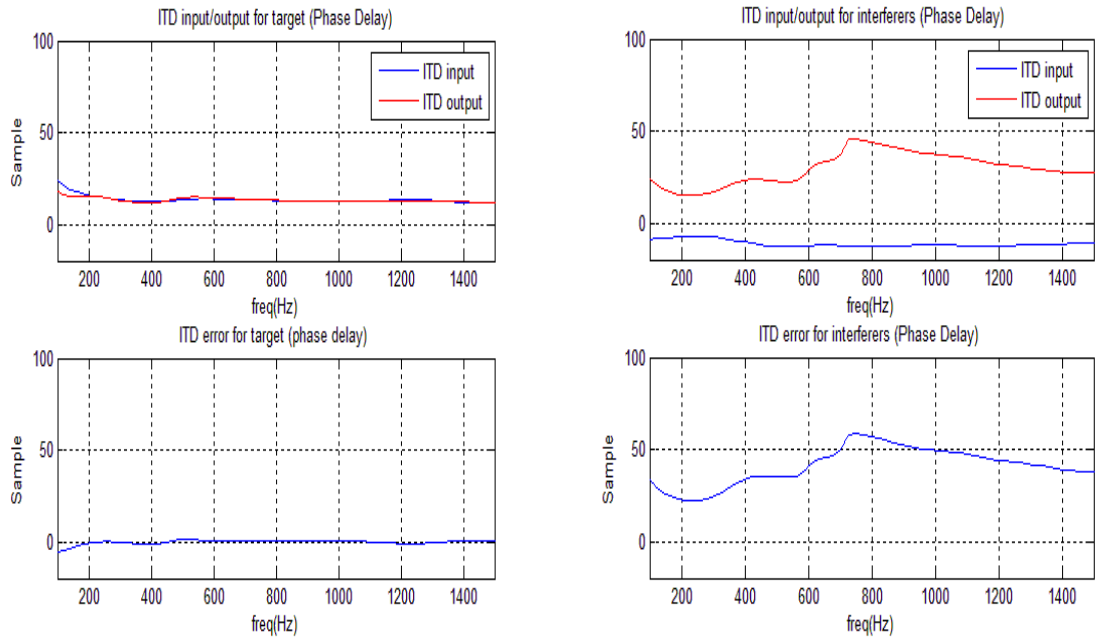

Figure 5.12: ITD using the phase delay approach for both the target and interference components using two binaural beamformers

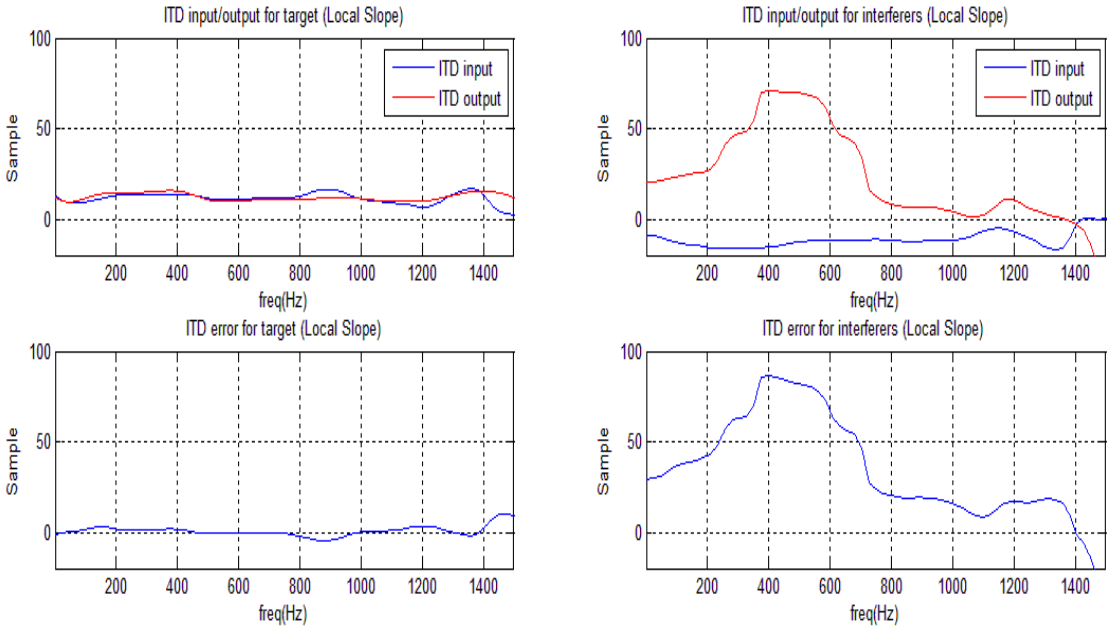

Figure 5.13: ITD using the local slope approach for both the target and interference components using two binaural beamformers

Overall, we can say that the ITD is a noisy measurement compared to the IPD, however it is more intuitive to analyze than the IPD. The computation of the ITD using the group delay is the noisiest of the three implementations and it should only be used with care.

Figure 5.14 and Figure 5.15 show the unwrapped IPDs and their fitted linear curves for the input and the output interference components in case of using two bilateral beamformers and two binaural beamformers. From the fitted curves in Figure 5.14 and Figure 5.15 the slopes are calculated in order to find scalar values for the ITD_v^{in} and the ITD_v^{out}, and then a scalar ΔITD_v is computed. As it is mentioned in section 3.2, the ITD and the IPD are more suitable for the low frequency components (below 1500 Hz) as the wavelength of the low frequency components is longer than the diameter of the head. Therefore, it is possible to detect the phase differences and/or the time differences between the signals at the left and right sides of the ear. However, the high frequency components convey ambiguous information in terms of the phase differences and the relative time difference between the left and right side signals (Begault, Durand R and others, 1994; Blauert, 1997; Gilkey & Anderson, 2014). Therefore, throughout this thesis, we have represent the IPD and the ITD up to 1500 Hz.

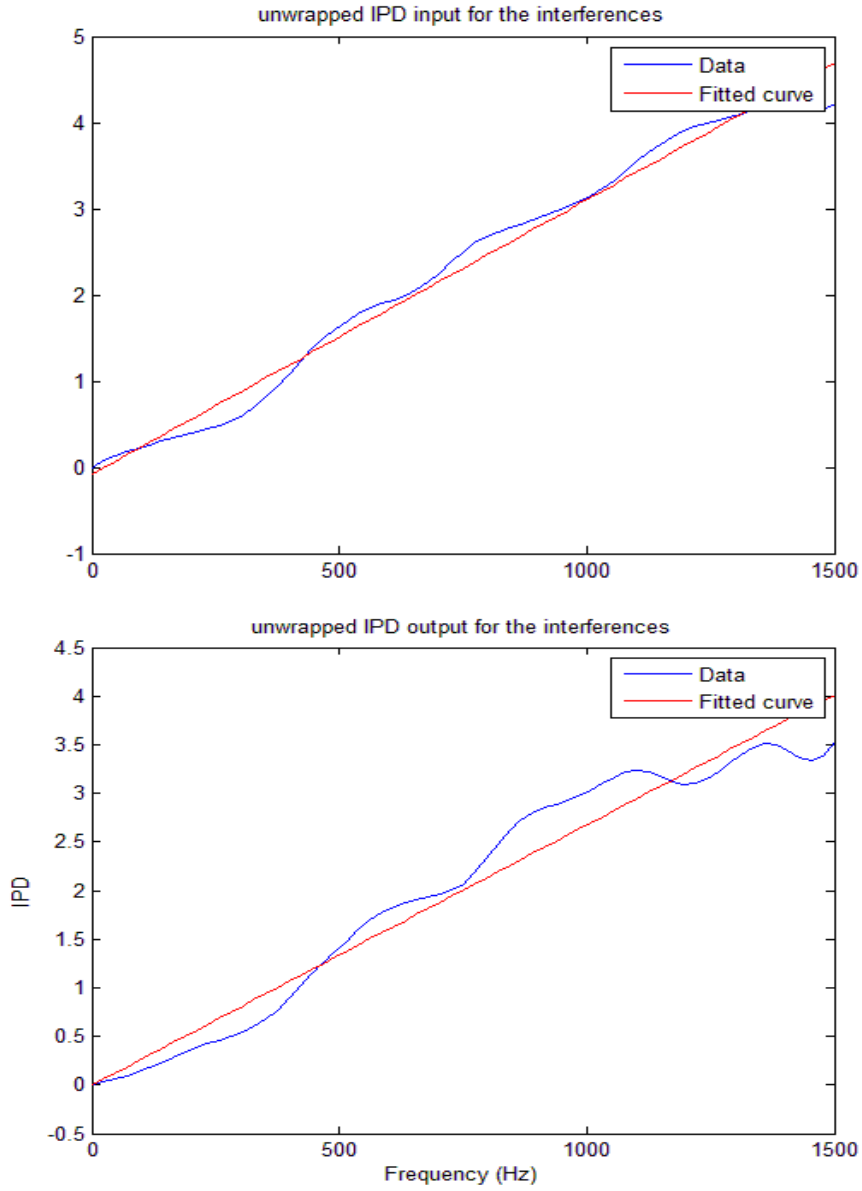

Figure 5.14: Unwrapped input and output IPDs and their fitted curve for the interference components using two bilateral beamformers

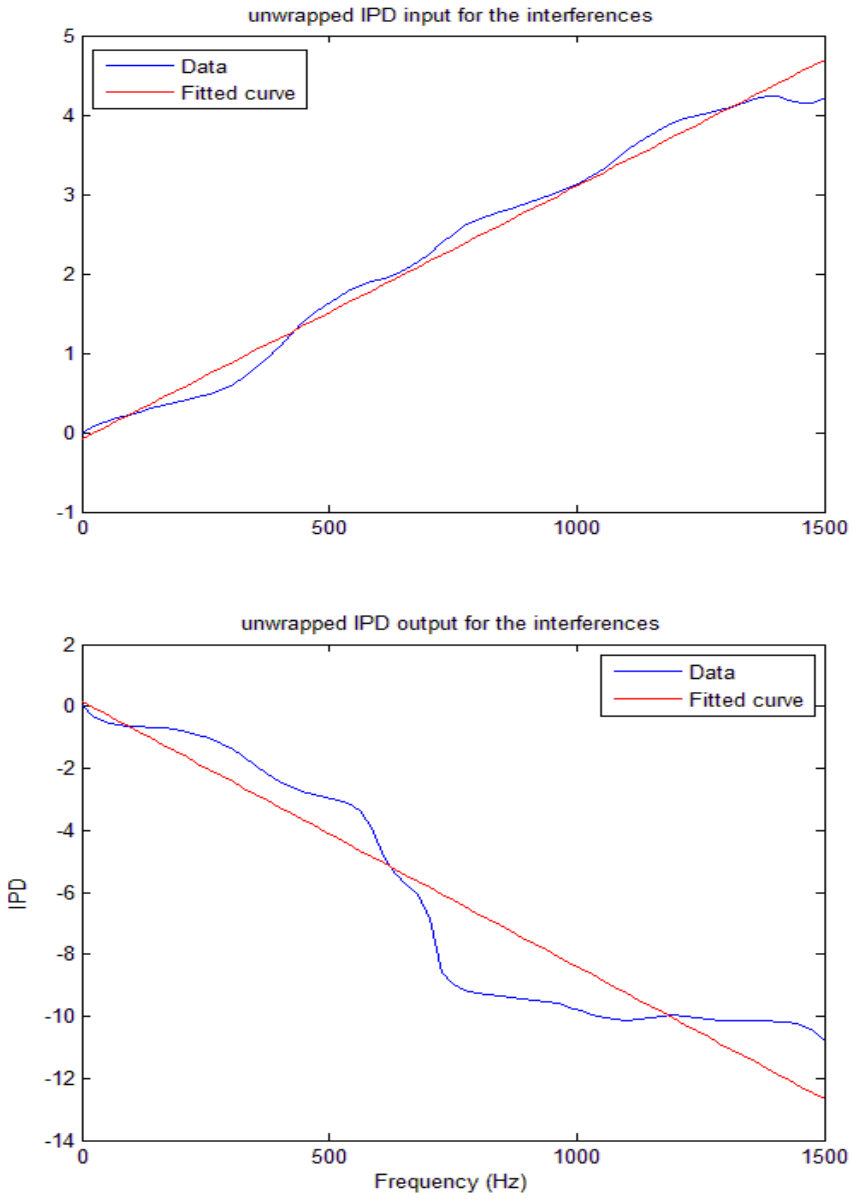

Figure 5.15: Unwrapped input and output IPDs and their fitted curve for the interference components using two binaural beamformer

Table 5.1 shows the ITD values for the interference components, in addition to the determination coefficient R^2 for the linear regression method. It is noticeable that the bilateral beamforming algorithm is better in preserving the spatial cues for the directional interferer.

Table 5.1: ITD measurements using linear regression

Metric	Two bilateral beamformers design	Two binaural beamformers design
ITD_v^{in}	-12.16	-12.16
ITD_v^{out}	-10.13	32.71
ΔITD_v	2.03	44.87
R_{in}^2 / R_{out}^2	0.99/0.94	0.99/0.89

Figure 5.16 and Figure 5.17 show the MSC^{in}, MSC^{out} and ΔMSC of the background noise component for the aforementioned beamforming algorithms. We can observe that the bilateral beamforming algorithm is also capable to better preserve the spatial impression of the background noise compared to the binaural beamforming algorithm.

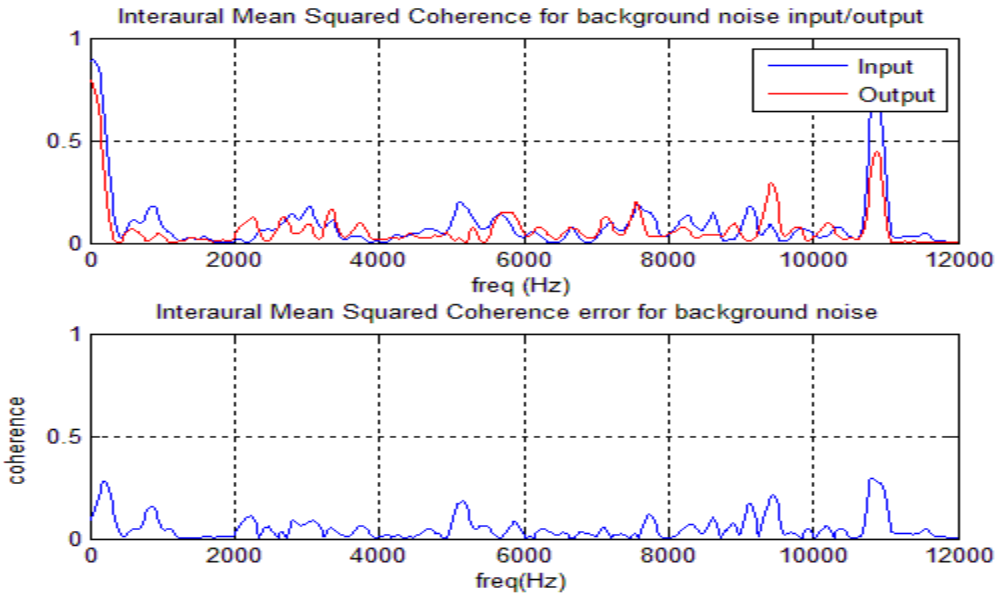

Figure 5.16: MSC for the background noise components using two bilateral beamformers

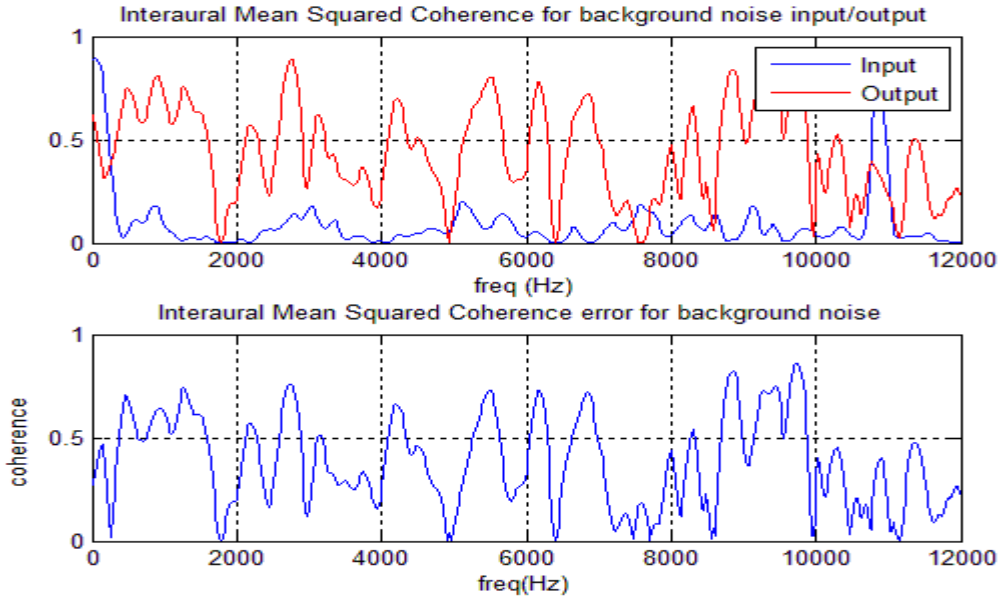

Figure 5.17: MSC for the background noise components using two binaural beamformers

Finally, Table 5.2 shows the scalar measurement values for all aforementioned cues preservation metrics for the interference and the background noise components. It is clear from

Table 5.2 that for the specific acoustic scenario considered here the bilateral beamforming design is better in preserving the spatial cues for the interference signal and the background noise than the binaural beamforming design.

Table 5.2: Summarizing the performance of the bilateral and binaural algorithms in terms of the cues preservation measurements, for specific acoustic scenario considered

Metric	Two bilateral beamforming design	Two binaural beamforming design
$\lvert \Delta ILD_v \rvert$	7.07 dB	21.16 dB
$\lvert \Delta IPD_v \rvert$	0.31 radians	1.28 radians
$\lvert \Delta ITD_v \rvert$ (group delay)	2.46 samples	34.40 samples
$\lvert \Delta ITD_v \rvert$ (phase delay)	1.50 samples	40.90 samples
$\lvert \Delta ITD_v \rvert$ (local slope)	1.71 samples	34.13 samples
$\lvert \Delta ITD_v \rvert$ (Time domain)	1.00 samples	38.00 samples
$\lvert \Delta ITD_v \rvert$ (linear regression)	2.03 samples	44.87 samples
ΔMSC	0.052	0.38

In case of multiple interferers, the previous ILD/IPD/ITD measurements can be calculated for the sum of directional interferers. However, as previously explained, the ILD/IPD/ITD may lose their physical sense when measured from the sum of directional interferers. Assuming that the interferers are speech sources, due to speech sparsity in the time-frequency plane the assumption that in each time-frequency "tile" there is a dominant interferer may often be valid. However, over time, the assumption that for each frequency there is a dominant interferer source may not necessarily be valid. Therefore, care must be taken when interpreting the ILD/IPD/ITD results for multiple interferers. To illustrate this, an acoustic scenario is used for testing with a frontal target speaker at 0 degree, three interferers at 45 degrees, 315 degrees and 90 degrees, and some background noise (soft, 14 dB lower than the power of each directional source). Table 5.3 shows the ILD/IPD/ITD for each individual directional interferer and for the sum of the directional

interferers using the common gain approach (which is supposed to preserve most of the spatial cues for the directional interferers as discussed in section 2.4). It is noticeable from Table 5.3 that using the cues preservation metrics for each interferer separately is much more robust than using them for the sum of interferers, especially in terms of ITD. For example, ΔITD using the local slope approach for each interference separately is less than one sample, while it is around 14 samples for the combined interferers! This result indicates that ITD measurement is more robust to be used for a single interference. However, using the "phase inversion method" the ILD/IPD/ITD metrics for each individual interference require (n+2) beamforming simulations, where n is the number of interferers, compared to three beamforming simulations if the ILD/IPD/ITD metrics are used for the sum of interferers. On the other hand, both ILD and IPD show more acceptable results when used for either individual interference or for the sum of interferers. As a result, we will rely more on ILD and IPD measurements for the analysis of acoustic scenarios with multiple interferers. Further results and discussion will be shown later in Chapter 6.

Table 5.3: Cues preservations metrics for each individual interference and for the sum of the interferers

Metric	Sum of interferers	Interf. # 1	Interf. # 2	Interf. # 3
$\lvert \Delta ILD_v \rvert$	3.50 dB	0.65 dB	1.08 dB	2.78 dB
$\lvert \Delta IPD_v \rvert$	0.238 radians	0.062 radians	0.109 radians	0.162 radians
$\lvert \Delta ITD_v \rvert$ (group delay)	5.53 samples	0.21 samples	0.32 samples	1.28 samples
$\lvert \Delta ITD_v \rvert$ (phase delay)	3.77 samples	0.07 samples	0.44 samples	0.12 samples
$\lvert \Delta ITD_v \rvert$ (local slope)	13.83 samples	0.71 samples	0.38 samples	0.11 samples
$\lvert \Delta ITD_v \rvert$ (Time domain)	0.00 samples	1.00 samples	0.00 samples	2.00 samples
$\lvert \Delta ITD_v \rvert$ (linear regression)	11.71 samples	0.75 samples	0.45 samples	0.70 samples

Chapter 6 Simulation Results of the Selection and Mixing Algorithm.

6.1 Experimental Setup

In the following section, we have first considered a 10 seconds recording of an acoustic scenario with one target speaker, three interference speakers (all at the same level as the target speaker) and diffuse babble noise. The frontal target is an English-speaking male at 0 degree (a non-frontal target is also considered in a later section). The interferers are a German-speaking female at 45 degrees, a French-speaking male at 315 degrees, and a German-speaking male at 90 degrees; in addition to background noise (soft, 14 dB lower than the power of each directional source). All the directional sources are at the same power. This acoustic scenario was recorded from a real-life environment with mild reverberation, which makes our simulations more realistic.

A 2+2 microphones structure has first been considered. A more detailed description of the 2+2 microphones structure is provided in section 2.5. For Case 1, the common gain approach without post-filtering proposed in (Lotter & Vary, 2006), has been used on each side (i.e., no mixing) with a common reference microphone (right front microphone) used on both sides for the built-in beamformer in the common gain approach. For Case 2 involving only beamformers on each side (i.e., no mixing, no common gain approach), the beamformer on each side uses the local front microphone as the reference microphone. For all the other cases involving the beamformer outputs or the common gain approach outputs in either a selection process or a mixing process at each T-F bin, for now we assume that the front right microphone was used as the reference microphone for the beamformers on both sides (including the built-in beamformer in the common gain approach). This is either to ensure that the two sides have access to the same information for the selection process or for the common gain processing, or to avoid the computation of two beamformers on one of the sides (one with each frontal microphone as the reference microphone). We will consider some alternatives later in this chapter.

For the simulations, the filter bank provided by our sponsor is used with input signals at 24kHz. This filter bank has 48 subbands down sampled to 1 kHz after analysis, with a bandwidth of 1 kHz in each subband and a spacing of 250 Hz between the center frequency of each subband.

First, a MVDR beamformer with the ideal noise correlation matrix was used to generate the beamformer outputs. After that, in order to have more realistic results as we usually do not have an access to the interferers and the background noise signals separately (we would need a very sophisticated VAD to estimate the background noise and the interferers correlation matrix), the correlation matrix of the noisy signals was used in the MVDR as in (Capon, 1969). For both MVDR beamformers, a good estimate of the target signal steering vector was used, computed directly from the target signals, so there is no effect of target signal steering vector mismatch in the simulations of this chapter. Finally, we have also used a GSC beamformer instead of the MVDR beamformer, which does not require any noise statistics estimation. For the GSC, we have used the same good estimate of the target signal steering vector in order to produce the blocking matrix output (coherent subtraction of target component between pairs of two channels) and the fixed beamformer output (coherent sum of target component from all channels).

6.2 Comparing Different Selection and Mixing Algorithms Using MVDR Beamformer with Ideal Noise Statistics

In Table 6.1, the performances of different selection and mixing approaches, which use the MVDR beamformer with ideal noise and interferers statistics, are compared in order to find the best trade-off between the cues preservations for the acoustic scene and the noise/ interferers reduction in terms of SNR-gain (left/right), ΔILD, ΔIPD, and ΔMSC.

Table 6.1: The performance of different selection and mixing algorithms and the common gain approach using MVDR beamformer with the ideal noise correlation matrix

Mixing/ Selection	Classification Criteria	Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| (dB)	|ΔIPD | (Radian)	|ΔMSC|
		Common Gain	4.09	4.76	3.50	0.238	0.078
		MVDR	6.19	7.89	4.67	1.792	0.756
Frequency Independent Approaches	Power	Case 3	5.00	5.90	4.76	1.648	0.178
		Case 4	3.50	4.01	2.93	1.385	0.130
		Case 6	4.52	5.43	3.87	0.293	0.085
		Case 7	2.79	3.50	2.45	0.124	0.057
	Power Difference	Case 8	4.50	5.55	3.97	0.315	0.100
		Case 9	3.07	3.90	2.55	0.306	0.055
Frequency Dependent Approaches	Without comparisons	Case 10	4.34	5.58	5.36	1.819	0.489
		Case 11	3.94	5.10	2.91	0.628	0.093
		Case 12	2.73	3.46	2.40	0.229	0.053
	Power Difference	Case 13	4.65	5.64	4.57	0.299	0.099
	Power	Case 14	4.60	5.51	4.38	0.292	0.085
	Coherence	Case 15	4.52	5.58	4.68	0.259	0.106
		Case 16	5.84	7.46	4.67	0.762	0.119

First, we will compare the performance of the first family of mixing algorithms that is based on the frequency independent processing and uses the power as classification criterion (case 3 to Case 7). Case 3 and Case 4 consider jointly the magnitude and phase of the signal in the selection and mixing. Case 3 mixes the common gain outputs with the left and right beamformers outputs, while Case 4 mixes the noisy signal at the front left and front right microphones with the beamformer outputs.

Table 6.1 reveals that Case 3 produces a better-enhanced signal compared with Case 4 since the common gain approach only introduces a moderate SNR improvement. In addition, both cases show comparable performance in terms of IPD, where Case 4 shows a slightly better performance by less than 0.3 radian. By also comparing the performance of Case 6 and Case 7, which separately consider the magnitude and phase of signals in the mixing and selection, Case 6 outperforms Case 7 in terms of the SNR-gain by around 2dB. On the other hand, in terms of ΔILD, Case 7 and Case 4 outperform Case 3 and Case 6.

By comparing Case 6 with Case 3, though Case 3 introduces additional 0.5 dB improvements in SNR-gain, Case 6 is better in preserving the IPD cues than Case 3 as it is shown in Figure 6.1 (since we are dealing with multiple interferers acoustic scenario, weighted-ΔIPD will be used for visualization purposes). In terms of the cues preservations for the background noise, Case 6 to Case 7 show comparable performance and slightly better than Case 3 (and all significantly better than the MVDR beamformer output). As a result, since the IPD is an important measurement in the perceptual preservations of the directional interferers (Wightman & Kistler, 1992), Case 6 could be selected as the best case from this family as it produces the best trade-off between noise reduction and cues preservations (especially IPD) for the acoustic scene. Moreover, Case 6 shows comparable performance with the common gain approach in terms of the cues preservations, while at the same time having a better ability to enhance the required outputs.

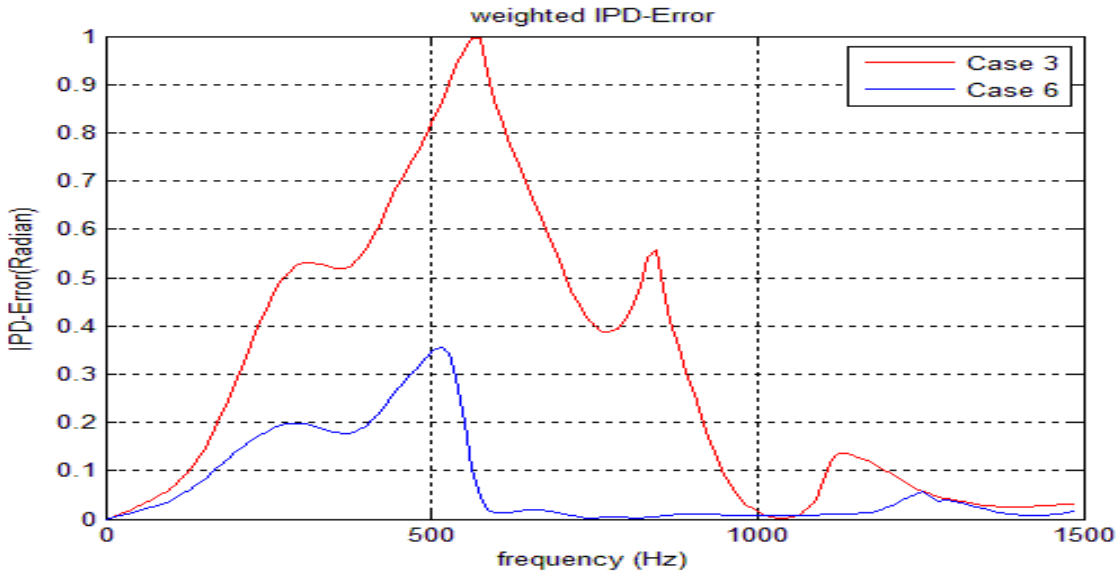

Figure 6.1: The performance of Case 3 and Case 6 in terms of weighted IPD error

Similarly, Case 8 and Case 9, which are frequency independent approaches using the power difference with different thresholds as classification criterion, show the same trend as Case 6 and Case 7 in terms of the noise reduction and the cues preservations. Case 9 generates outputs with lower SNR than Case 8 by around 2dB as Figure 6.2 and Table 6.1 show**.** On the other hand, Case 9 shows a better ability in preserving the ILD cues since it uses the magnitude of the original noisy signal instead of the magnitude of the common gain output. In terms of IPD, both Case 8 and Case 9 show comparable performance since both of them use the phase of the noisy signals. At the same time, Case 8 and Case 6 show comparable performance in terms of the SNR-gain. Consequently, Case 6 and Case 8 were the cases that have been chosen to be extended to the frequency dependent approaches.

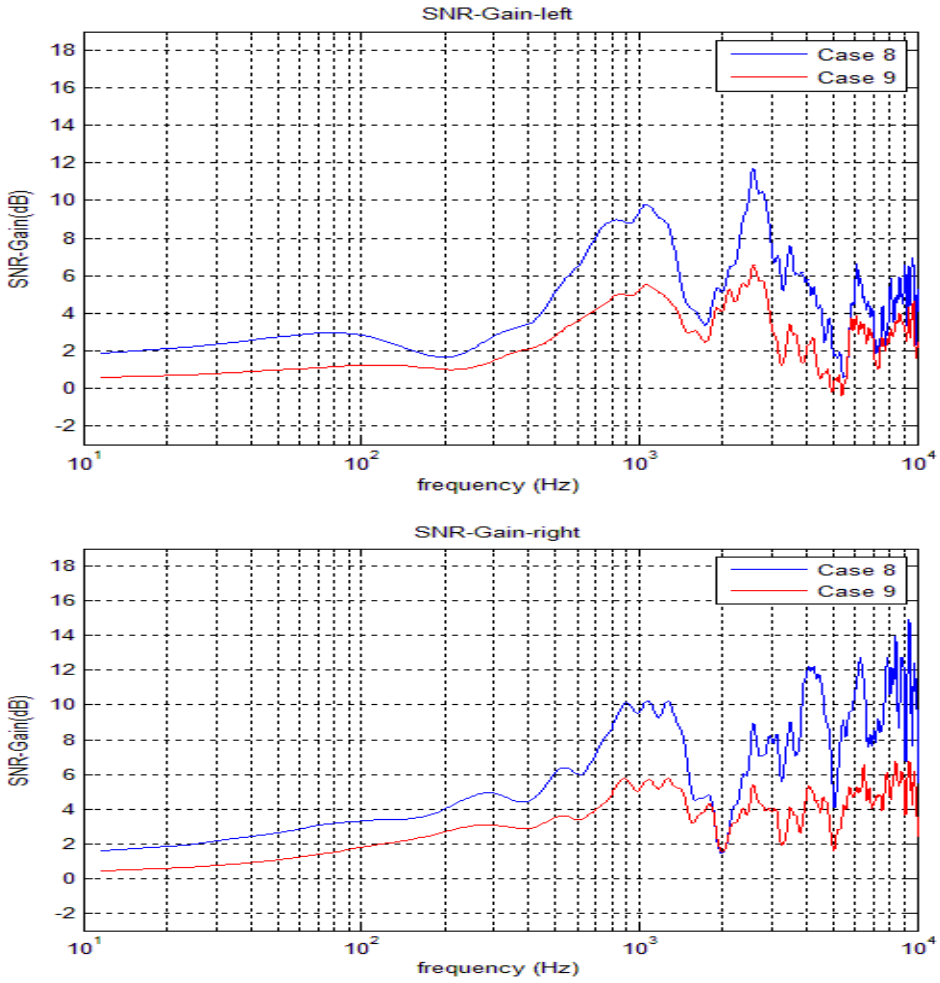

Figure 6.2: Comparison between the performances of Case 8, Case 9 in terms of SNR-gain

We now compare the frequency dependent processing of Case 10, Case 11, and Case 12. Case 10, which uses the phases of the noisy signal from the first to the fourth subbands, outperforms Case 11 in terms of SNR-gain by around 0.4 dB on average for the left and right channels. Case 12 falls behind Case 10 and Case 11 in terms of SNR-gain as it is shown in Figure 6.3 and Table 6.1.

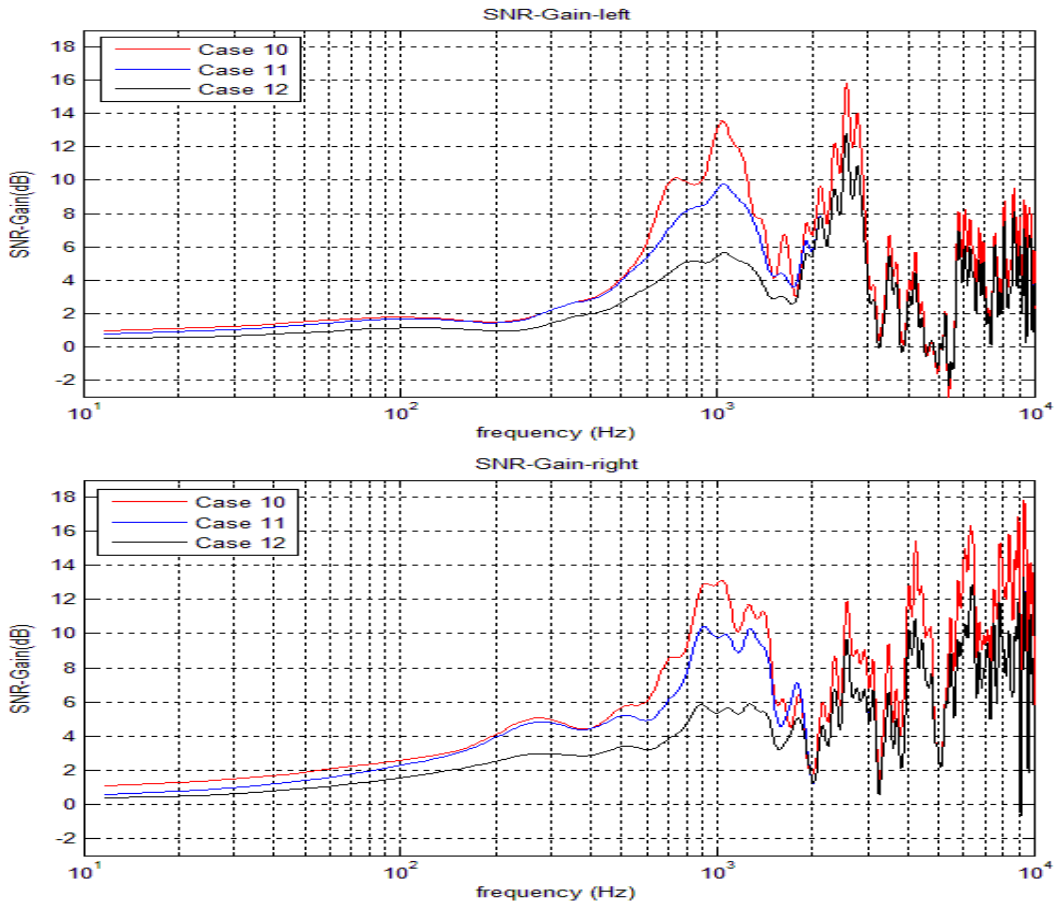

Figure 6.3: SNR-gain of the left and right channels for Case 10, Case 11, and Case 12

In terms of ILD, Case 10 falls behind Case 11 and Case 12 since Case 10 uses only the magnitude of the beamformer output for all subbands, which is not the case in Case 11 and Case 12. The latter cases use a small portion of the magnitude of the noisy signal for the high frequency components. In terms of IPD, Case 10 also falls behind Case 11 and Case 12 as it is shown in Figure 6.4, since Case 10 uses the phases of the noisy signal only for four subbands, while Case 11 and Case 12 use the phase of the noisy signal for the first eight subbands. Up to 500 Hz, Case 10 shows comparable performance with Case 11.

Regarding the preservation of the spatial perception of the background diffuse noise, based on the informal listening and MSC measurement in Figure 6.5, Case 10 changes most of the cues for the background noise, where the background noise seems to be aligned in the direction of the

target signal. However, Case 11 and Case 12 preserve the spatial impression of the background noise and reveal comparable performances since the phase of the noisy signal for the first eight subbands have been considered.

Figure 6.4: The performance of Case 10, Case 11, and Case 12 in terms of IPD

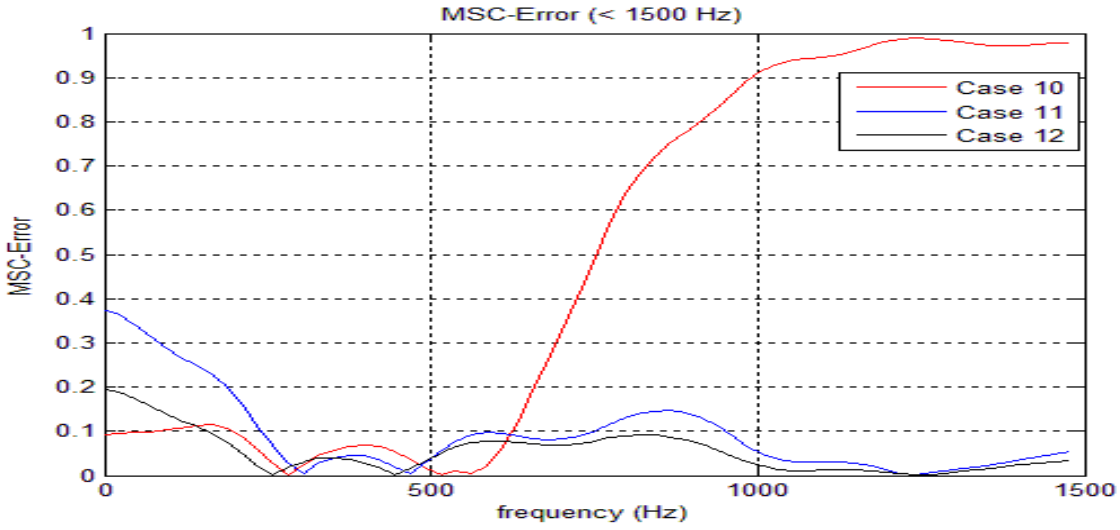

Figure 6.5: The performance of Case 10, Case 11, and Case 12 in terms of MSC

In Case 11, if 1.0 is assigned to α and 0.0 is assigned to β , which means we have not considered the magnitude of the noisy signal for high frequency components, the SNR-gain

increases only by around 0.1 dB while the ILD error increases by around 2dB. On the other hand, the IPD and MSC (< 1500Hz) measurements remain the same since these measures are suitable for the low frequency components. As a result, adding a small portion of the magnitude of noisy signal to magnitude of the beamformer output for the high frequency components produces a negligible reduction in the SNR-gain, while it improves the cues preservations in terms of ILD as Table 6.2 shows. Furthermore, adding a small portion of the magnitude of the noisy signal in low frequency components reduces the SNR-gain as in Case 12 by around 2 dB while introducing an improvement in IPD by around 0.4 radians. To sum up, Case 11 shows the best trade-off over Case 10 and Case 12 in terms of the cues preservations and noise reduction.

Table 6.2: The effect of changing α and β factors in the performance of Case 12.

Measurements	**Case 11 (α=0.7 and β=0.3)**	**Case 11 (α=1 and β=0)**
SNR gain left (dB)	3.94	4.01
SNR gain right (dB)	5.10	5.24
Δ ILD (dB)	2.91	4.78
Δ IPD (Radian)	0.628	0.628
Δ MSC (<1500 Hz)	0.093	0.092

Case 13 and Case 14 are frequency dependent extensions of Case 6 and Case 8. As Table 6.1 shows, both Case 13 and Case 14 have the same trade-off between noise reduction and cues preservations, while Case 6 and Case 8 fall behind Case 13 and Case 14 in terms of SNR-gain by less than 0.1 dB. However, considering that the phase of the noisy signals is just important for the low frequency components, and it does not play any role in preserving the IPD/ITD for the directional interferers at the higher frequency components, the frequency dependent cases are likely more robust. Therefore, Case 13 and Case 14 will be considered for further tests under different acoustic scenarios and different configurations.

Further comparisons are done next for the frequency dependent approaches which use the coherence as classification criterion. In Case 15, the magnitude-squared coherence (MSC) is used for the selections/mixing between the outputs of the common gain method and the outputs of the beamformers, compared to a threshold as previously explained. Case 16, which uses the phase and the magnitude of the coherence as the classification criterion to select between the common gain

outputs and the beamformer output, outperforms Case 15 in terms of SNR-gain by around 1.5-2.0 dB as Figure 6.6 shows. In terms of ILD, both Case 15 and Case 16 shows comparable performance. Though Case 16 falls behind Case 15 in terms of IPD because of the phase selection between the phase of the noisy signal and the phase of the beamformer output for low frequency components in Case 16, from informal listening both approaches show comparable capabilities in preserving the acoustic scene (in other words, they both significantly improve the IPD preservation compared to the MVDR beamformer). Furthermore, as shown by the MSC metric in Table 6.1, Case 15 and Case 16 have similar abilities in preserving the perceptual impression of the background noise.

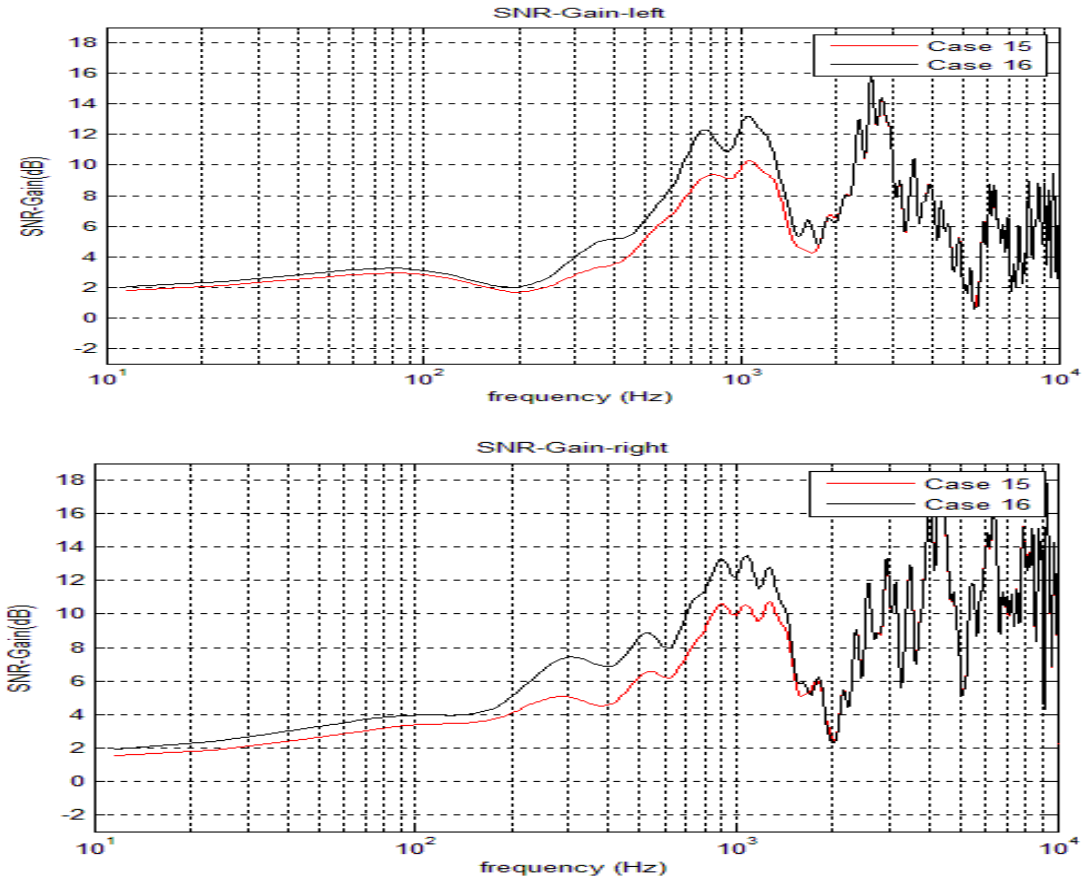

Figure 6.6: SNR-gain of the left and right outputs for Case 15, and Case 16

Based on the previous analysis, Table 6.3 shows the best-selected approaches in addition to the performance of the common gain approach and the beamformer-only approach. The

selection of the best approaches is based mainly on the amount of the SNR-gain or speech enhancement relative to the reference microphone, while considering at the same time their ability in keeping the cues for the directional interferences and the background babble noise unmodified. During the selection for the best approaches, ILD, IPD, and MSC have been considered as measurements for the cues preservation. However, we have put more emphasis on the IPD measurement since it is often more important than the ILD in preserving the spatial cues for the directional interferences perceptually. This assumption is suitable for the acoustic scenarios with frontal and non-frontal target signals (Wightman & Kistler, 1992).

In Table 6.3, all selected approaches are frequency dependent. Case 13 uses the power difference as classification criterion, case 14 uses power as classification criterion, while case 16 uses the coherence as classification criterion. From informal listening, all selected approaches, in addition to the common gain approach, have the capability of preserving the spatial impression of the acoustic scene. Nevertheless, Case 16 introduces the highest SNR-gain, and it has a comparable SNR-gain with the beamformer-only approach. The best-selected approaches in Table 6.3 will be considered in later sections for further investigation using different acoustic scenarios and different microphone or beamformer configurations.

Table 6.3: The performance of the best-selected approaches, the common gain approach, and the MVDR with the ideal noise correlation matrix

Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD} (dB)	|ΔIPD | (Radian)	|ΔMSC|
Common Gain	4.09	4.76	3.50	0.238	0.078
MVDR	6.19	7.89	4.67	1.792	0.755
Case 13	4.65	5.64	4.57	0.299	0.099
Case 14	4.60	5.51	4.38	0.292	0.085
Case 16	5.84	7.46	4.67	0.762	0.119

For more investigations of the behaviours of the best selected approaches at each subband-time sample, we have calculated the number of times when the power differences in Case 13 fall within different intervals based on different threshold values as in equation (4.17) and (4.18) in section 4.3. For the left channel and under the previous acoustic scenario, the selection algorithm considers the first interval (between the $\text{threshold}_{\text{max}}$ and the $\text{threshold}_{\text{max-mean}}$) 5.5% of the

selection time, the second interval (between the $threshold_{max\text{-}mean}$ and the $threshold_{mean}$) 35.1% of the selection time, the third interval (between the $threshold_{mean}$ and the $threshold_{mean\text{-}min}$) 57.8% of the selection time, and the fourth interval (between the $threshold_{mean\text{-}min}$ and the $threshold_{min}$) 2.3% of the selection time. For the right channel the power differences fall within the first, second, third, and fourth intervals 1.2%, 28.4%, 67.75%, and 2.2% of the selection time, respectively. In Case 14 and based on equation (4.19) in section 4.3, the powers of the beamformer output at each subband-time sample are 52% of the time lower than the power of the common output for the left channel and 55% of the time lower than the power of the common gain output for the right channel. In Case 16 and based on equations (4.32) and (4.33) in section 4.3, the coherences between the common gain and the beamformer output at each subband-time sample is 40% of the time lower than the threshold value for the left channel and 45% of the time lower than the threshold value for the right channel.

Table 6.4 shows also the performance of the best-selected approaches, in addition to the common gain and the MVDR beamformer, for the case where the noisy signal correlation matrix was used instead of the noise correlation matrix in the MVDR, to make the beamformer more practical and reduce the need for a VAD. It is noticeable that even for the current setup where we use very good estimates of the target steering vector in the MVDR algorithm, by switching to the noisy correlation matrix the SNR-gain for both channels (left /right) decline by around 1.0 to 1.5 dB. It is known from classic beamforming theory that if there were a significant mismatch between the target steering vector being used in the MVDR and the steering vector corresponding to the actual data being processed, the performance of the MVDR using the noisy correlation matrix would further degrade much faster than the performance of the MVDR using the noise correlation matrix. Nevertheless, all the tested cases show similar performance to the previous results in terms of the cues preservations for both the directional interferers and the background noise.

Table 6.4: The performance of the best-selected approaches, the common gain approach, and the MVDR beamformer with the noisy correlation matrix

Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD} (dB)	|ΔIPD | (Radian)	|ΔMSC|
Common Gain	3.22	3.96	3.40	0.307	0.055
MVDR	4.48	6.14	4.57	1.823	0.642
Case 13	3.65	4.70	4.49	0.402	0.085
Case 14	3.69	4.68	4.27	0.399	0.072
Case 16	4.36	6.00	4.49	0.933	0.165

6.3 More Emphasis on Preserving the Spatial Impression of the Diffuse Noise

In Case 10, only the phases of the noisy signals are considered for the first four subbands, which can lead to some perceptual preservation of the directional interferers, based on the IPD metric and from some informal listening. Nonetheless, this approach does not perform as well as the other approaches in its capability of preserving the spatial impression of the babble diffuse noise based on both the objective MSC evaluation and some informal listening. To put more emphasis on the preservation of the background noise cues, simulations were performed where the interferers' amplitudes were reduced by a factor of 5 (now each interferer is 14 dB below the target level) and the amplitude of the background babble diffuse noise was increased by a factor 5 (now at the same level as the target signal), for the same acoustic scenario that was tested before: one frontal target and three interferers at 45 degrees, 315 degrees, and 90 degrees.

Table 6.5 shows the performance of the selected approaches (Cases 13-16) in addition to the simplified approach of Case 10, as well as Case 11. In this test scenario, the focus is on the cues preservations of the diffuse noise. It is noticeable that most of the selected approaches Cases 13-16 as well as Case 11 perform well in preserving the cues for the background noise, while Case 10 changes most of the cues for the background noise. Similarly to the previous testing scenario, the MSC was computed from the background noise signal, and the ILD and IPD were computed from the sum of directional interferers.

Table 6.5: The performance of some selected cases with more emphasis on the diffuse noise using MVDR beamformer with ideal noise statistics

Tested Cases	SNR gain left (dB)	SNR gain right (dB)	ΔILD (dB)	|ΔIPD | (Radian)	|ΔMSC| < 1500Hz
Case 10	4.49	3.86	4.72	1.780	0.540
Case 11	4.00	3.47	2.76	0.543	0.097
Case 13	4.71	4.00	3.57	0.273	0.083
Case 14	4.46	3.83	1.26	0.206	0.069
Case 15	4.82	4.12	2.60	0.300	0.092
Case 16	6.03	5.39	2.67	1.533	0.384

As a result, through a direct comparison of Case 10 and Case 11 we can conclude that the first eight subbands are important in preserving the spatial impression of the background noise.

Furthermore, in Case 16, the phases of the first eight subband are selected at each subband-time sample based on the phase of the coherence between the common gain output and the beamformer output. Based the ΔMSC as well as some informal listening, this approach preserves some of the spatial impression of the background noise. Case 16 is better than Case 10 but it falls behind the other cases for this scenario.

6.4 Comparing the Best-Selected Approaches Using Different Acoustic Scenarios

6.4.1 Acoustic Scenarios with Frontal Target Speaker

Different acoustic scenarios have been considered to test the best-selected approaches. The first acoustic scenario has a frontal target speaker, one interference speaker at 315 degrees, and background babble noise (soft, 14 dB lower than the target's power and each interferer's power). To have more practical results, a GSC beamformer, which does not need a VAD system for noise correlation matrix estimation, has been used to test this acoustic scenario and for all the other acoustic scenarios in this section. A 2+2 microphones configuration has also been used for testing. In this section, we will refer to Case 13 by classification and mixing algorithm based on the power difference (or more simply the power difference approach), Case 14 by classification and mixing

algorithm based on the power (or more simply the power approach) , and Case 16 by classification, selection and mixing algorithm based on the coherence (or more simply the coherence approach). The beamformers for the common gain, the power difference, the power and the coherence approaches used a common reference microphone on both sides (the left front microphone in this scenario), while the case of GSC-only used the local front microphone on each side as the reference microphone.

Figure 6.7 and Table 6.6 show that the best-selected approaches, in addition to the common gain approach, outperform the GSC beamformer in terms of ILD and IPD. At the same time, Table 6.6 shows that the coherence approach outperforms the other approaches (except the GSC beamformer) in terms of SNR-gain. The GSC beamformer falls behind the other approaches in terms of MSC. As this acoustic scenario has only one interference signal, we will use the ITD measurement (based on two approaches: the local slope and the linear regression) as well as the ILD and IPD to measure the cues preservation for the directional interference signal. Table 6.6 shows that the GSC beamformers also significantly falls behind the other approaches in terms of ITD, while the other approaches show comparable performance. As a result, the coherence approach demonstrates the best trade-off between the noise reduction and the cues preservations for the directional interference signal and the background noise in this acoustic scenario.

Table 6.6: The performance of the best-selected approaches, the common gain approach and the GSC beamformer for an acoustic scenario of frontal target and one interferer at 45 degrees

Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| (dB)	|ΔIPD | (Radian)	|ΔITD (Local Slope) | (sample)	|ΔITD (Linear Regression) | (sample)	|ΔMSC|
Common Gain	5.65	3.57	3.13	0.102	0.280	0.500	0.072
GSC	8.23	4.75	8.19	1.651	23.880	22.510	0.600
Power Diff.	6.46	3.70	6.72	0.085	2.060	0.280	0.092
Power	6.39	3.74	6.06	0.077	1.910	0.090	0.086
Coherence	8.20	4.69	6.55	0.478	3.680	3.180	0.174

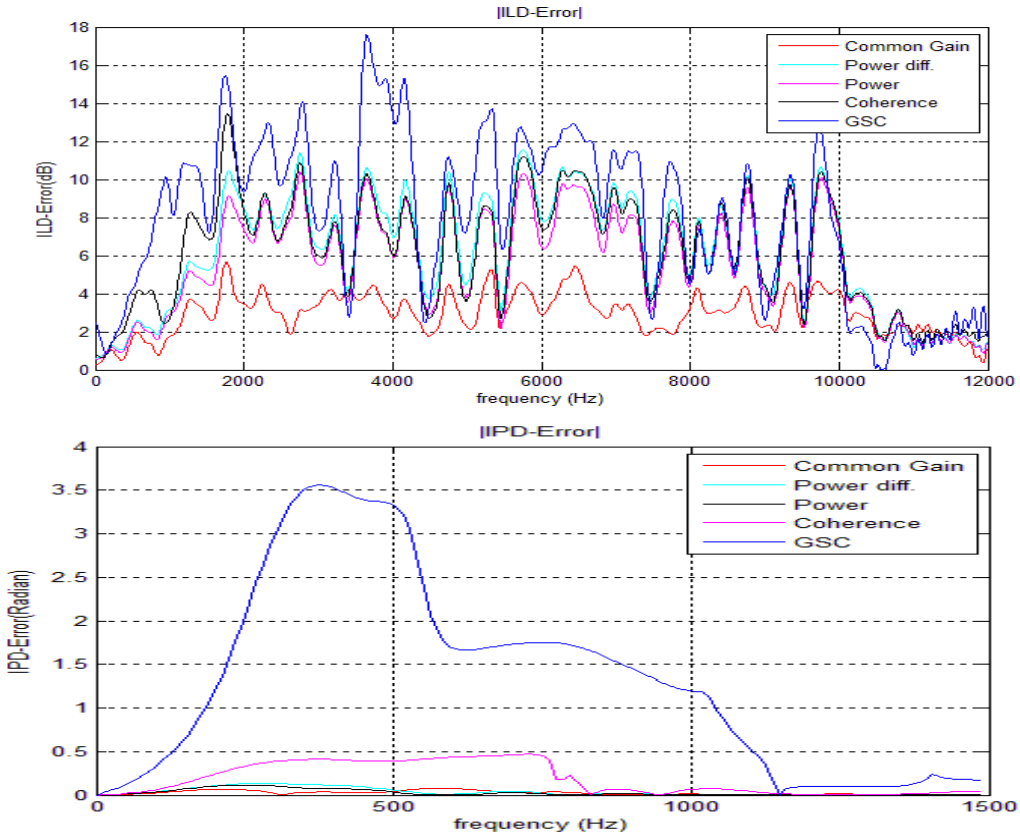

Figure 6.7: The performance of the best-selected approaches, the common gain approach and the GSC beamformer in terms of ILD and IPD for scenario of frontal target with one interference.

The best-selected approaches have also been tested under more complex acoustic scenarios with a frontal target speaker and multiple interferers. In the first acoustic scenario, we have used one frontal target speaker, one back interference speaker at 225 degrees, one interference speaker at 45 degrees, and one lateral interference speaker at 90 degrees, in addition to background noise (soft, 14 dB lower than the power of the target and each interferer). In the second acoustic scenario, we have used one frontal target speaker, two back interference speakers from 135 degrees and 225 degrees, and one lateral interference speaker from 270 degrees, in addition to background noise (soft, 14 dB lower than the power of the target and each interferer). In these scenarios, the front right microphone is used as a reference for all beamformers except the "GSC" scenario, where the beamformer on each side uses the local front microphone as the reference microphone. Table 6.7 shows the performance metrics of the signals that are used in the selection/mixing, i.e., the outputs

of the GSC beamformers and the common gain method, for the two acoustic scenarios. The ILD and IPD measurements for the target have been added to verify the capability of the proposed algorithms to keep the cues for the target speaker unchanged.

Table 6.7: The signals that are used in mixing/selections, frontal target acoustic scenarios

Tested Case	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Radian)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Radian)	|ΔMSC|
Frontal target, interferers at 225,45,90 degrees							
Common Gain	6.40	6.42	0.32	0.030	0.20	0.156	0.072
GSC	9.34	10.29	0.72	0.064	2.37	1.551	0.607
Frontal target, interferers at 135,225,270 degrees							
Common Gain	6.96	6.75	0.19	0.034	0.37	0.149	0.065
GSC	11.27	9.74	0.31	0.120	1.61	1.762	0.587

The results for the best-selected classification and mixing approaches in these complex acoustic scenarios are shown in Table 6.8. Table 6.8 shows that the coherence approach produces the best trade-off between noise reduction and cues preservations, by putting more emphasis on the noise reduction first and then on the cues preservation. Sometimes we might notice that the value of the ΔIPD in the coherence approach is significantly higher than its value in the power and the power difference approaches. However, by comparing the values of ΔIPD in the coherence approach (Table 6.8) with their values in the beamformer outputs (Table 6.7), which change most of the cues, we can notice significant differences between these values (always higher than 0.5 radians), which indicate the better ability of the coherence approach in preserving the cues compared to the beamformer outputs. These results have been also verified by informal listening, and in all cases the coherence approach shows a good capability in reducing the noise/interferers while at the same time preserving the cues.

Table 6.8: The performance of the best-selected approaches under frontal target acoustic scenarios

Tested Case	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Radian)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Radian)	|ΔMSC|
Frontal target, interferers at 225,45,90 degrees							
Power Diff.	6.96	7.41	0.22	0.030	1.242	0.156	0.091
Power	6.91	7.29	0.18	0.030	0.894	0.141	0.084
Coherence	8.75	9.88	0.06	0.097	0.752	0.938	0.178
Frontal target, interferers at 135,225,270 degrees							
Power Diff.	8.06	7.43	0.19	0.040	1.29	0.157	0.079
Power	7.92	7.40	0.20	0.050	0.94	0.131	0.070
Coherence	10.48	9.64	0.07	0.080	0.99	1.031	0.118

6.4.2 Acoustic Scenarios with Non-Frontal Target Speaker

For the different algorithms making use of either the common gain approach or an approach making classification, selection and mixing of signals for each T-F bin, it has been assumed so far that the beamformers on each side of a binaural system use the same reference microphone. The only exception has been for systems using only a beamformer on each side (no common gain, no selection or mixing), where the frontal microphone on each side has been used as the reference microphone. In all the acoustic scenarios considered so far in this chapter, i.e., with frontal target speaker, this assumption is reasonably valid since the front left and front right signals available for selection/mixing have similar levels and phases for the target component. Consequently, it is possible to use a common reference for the beamformers that are used in selection/mixing on both side as well as for the common gain approach, and the distortionless response for the target speaker is preserved as we have noticed in Table 6.7 and Table 6.8 (e.g., in terms of ILD and IPD for the target).

However, in acoustic scenarios with a non-frontal target, the left and right front microphone signals available for selection and mixing do not have the same level and phase for the target speaker. Therefore, it might not be valid to use a common reference for the beamformers on both sides in the selection/mixing approaches, because it means that the overall resulting beamforming system is no longer distortionless for the target component on one of the sides. For example, for

non-frontal targets it means that the amplitude and the phase of the target component will not be as expected on one of the sides.

To assess this, an acoustic scenario with a non-frontal target speaker at 45 degrees, an interference speaker at 315 degrees, and some background noise has been tested. In the selection/mixing approaches under this non-frontal acoustic scenario, we have used either:

- The output signals of the GSC beamformers which use a common reference (front left microphone), or
- The output signal of the GSC beamformers which use a local front microphone on each sides as a reference.

Table 6.9 shows results for the methods producing the signals involved in the selection and mixing processes. These signals are:

- The outputs of the common gain approach that uses the front left microphone as a reference for both sides, and
- The GSC beamformers' outputs, where each beamformer on each side of the binaural system uses the local front microphone as a reference.

Table 6.9: The signals that are used in mixing/selections, acoustic scenario with non-frontal target and interferer at 315 degrees

Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Rad)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Rad)	|ΔITD | Interf. (Local Slope) (sample)	|ΔITD | Interf. (Linear Reg.) (sample)	|ΔMSC|
Common Gain	7.21	4.86	0.89	0.055	2.80	0.120	0.060	0.590	0.089
GSC	12.94	3.52	1.72	0.091	21.14	1.307	34.210	44.850	0.530

The best-selected approaches using selection and mixing have been tested under the non-frontal acoustic scenario. Table 6.10 shows that using the outputs of the beamformers with a common reference microphone will change the ILD cues for the target, while still preserving most of the IPD cues for the target since the best-selected approaches put more emphasis on the phase

of the noisy signals in the first eight subbands. As it was mentioned previously, the first eight subbands in the filter bank that we use cover the frequency components up to 1500 Hz, so that keeping the phase of the noisy signal in those subbands is enough to preserve the cues in terms of IPD (and the IPD is perceptually more important than the ILD for the cues preservations). On the other hand, the best-selected approaches using the outputs of the beamformers with frontal local microphones as references (combined with the outputs of the common gain approach that uses a common reference microphone) guarantee a distortionless response for the targets in terms of IPD and ILD. However, on one of the sides of the binaural system this increases the computational complexity, as we need to have an additional beamformer for the common gain approach using a different reference signal than the one used for the GSC-only beamformer. By comparing the best-selected approaches using either a common reference signal or a local reference signal for the GSC beamformer component, overall similar performances have been noticed with a non-frontal target speaker, with the exception of the ΔILD for the interference signal which is significantly higher when the local reference signals are used. Table 6.10 shows that the coherence approach outperforms the other approaches in terms of SNR-gain.

Table 6.10: The performance of the best-selected approaches for an acoustic scenario with non-frontal target and interferer at 315 degrees.

Ref	Tested Cases	SNR gain left (dB)	SNR gain right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Rad)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Rad)	|ΔITD| Interf. (Local Slope) (sample)	|ΔITD| Interf. (Linear Reg.) (sample)	|ΔMSC|
Common ref. for GSC	Power Diff.	8.65	3.77	6.11	0.087	7.22	0.064	1.54	0.55	0.083
	Power	8.56	4.19	5.05	0.079	5.90	0.071	1.28	0.52	0.081
	Coherence	11.66	4.29	3.06	0.082	6.47	0.204	2.69	0.79	0.093
Local ref. for GSC	Power Diff.	8.87	3.68	0.25	0.043	15.16	0.074	2.18	0.37	0.097
	Power	8.67	4.33	0.10	0.042	12.24	0.072	1.58	0.57	0.088
	Coherence	12.00	4.38	0.81	0.071	12.82	0.480	1.61	0.10	0.197

In Figure 6.8 to Figure 6.12, a common reference (the front left microphone) is used in the beamformers of the best-selected approaches and the common gain approach. Figure 6.8 shows that the power difference, the power, and the coherence approaches introduce comparable performance in terms of ILD, while the common gain approach outperforms these cases in terms of ILD. The GSC beamformer outputs significantly change the spatial cues of the interference signal compared with the other approaches in terms of ILD. Figure 6.9 and Figure 6.10 show that the common gain, the power difference, the power, and the coherence approaches have comparable performance in terms of IPD and ITD (using local slope) and significantly outperform the GSC beamformer. Using the ITD measurement and from some informal listening, the algorithm with ΔITD values lower than 4 samples on average over the frequencies seem to preserve the cues for the directional interference signal in this considered scenario.

Figure 6.11 shows that the GSC beamformer significantly changes the cues for the background noise in terms of MSC while the other approaches demonstrate better abilities in preserving the spatial impression of the background noise.

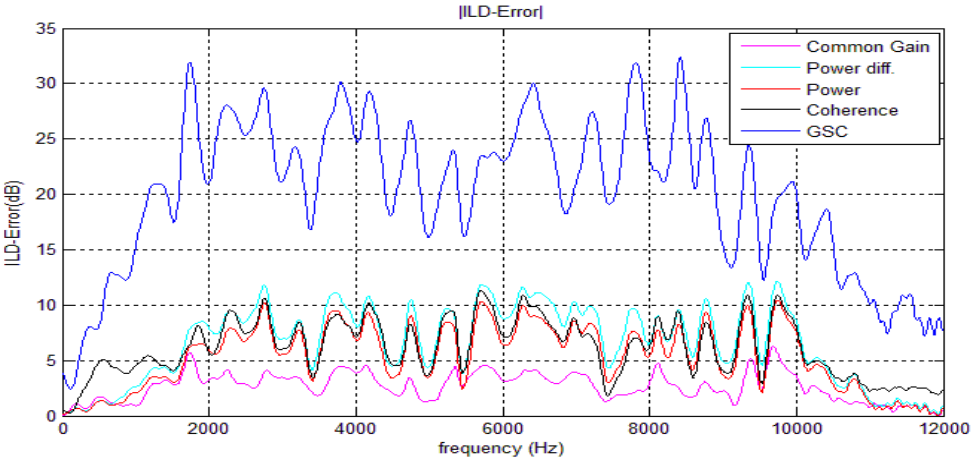

Figure 6.8: The performance of the best-selected approaches, the common gain approach, and the GSC beamformer in terms of ILD for non-frontal target with one interference

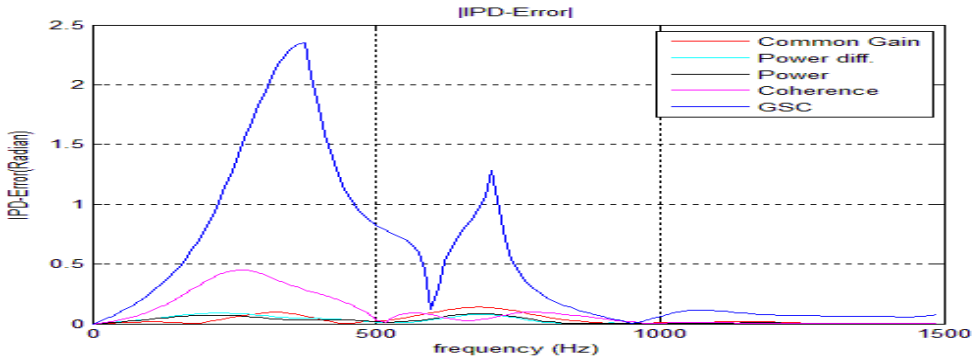

Figure 6.9: The performance of the best-selected approaches, the common gain approach, and the GSC beamformer in terms of IPD for non-frontal target with one interference

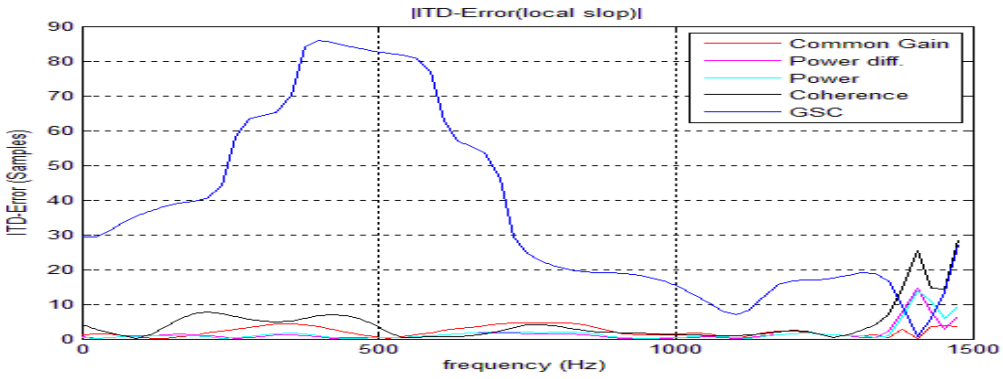

Figure 6.10: The performance of the best-selected approaches, the common gain approach, and the GSC beamformer in terms of ITD for non-frontal target with one interference

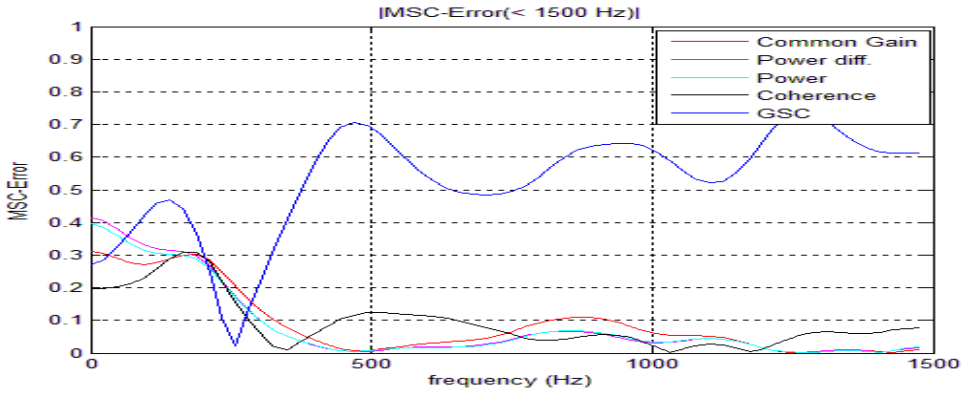

Figure 6.11: The performance of the best-selected approaches, the common gain approach, and the GSC beamformer in terms of MSC for non-frontal target with one interference

On the other hand, even though the coherence approach performs approximately as well as the common gain, the power difference, and the power approaches for the cues preservation, the coherence approach introduces an improvement in SNR by around 3 dB for the left channel as it is shown in Figure 6.12. As a result, the coherence approach demonstrates the best trade-off between cues preservation and noise reduction for this acoustic scenario.

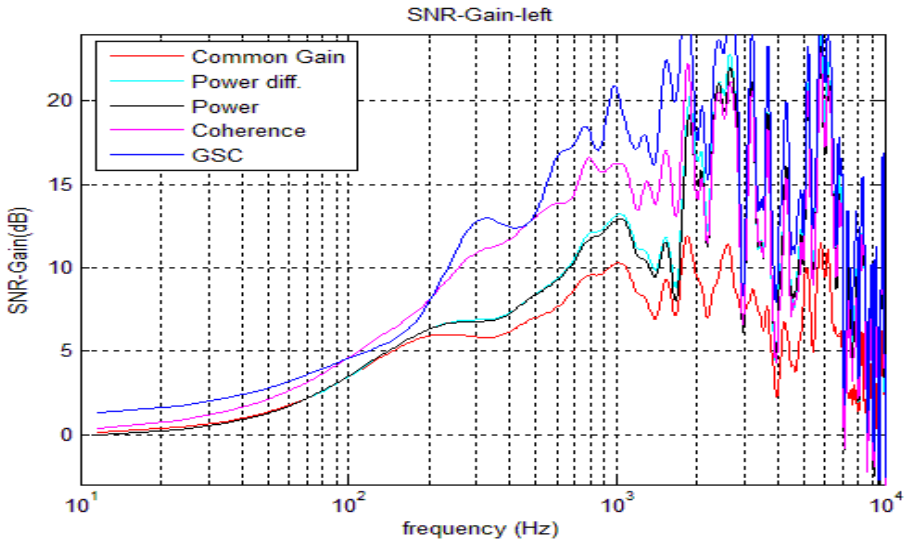

Figure 6.12: The performance of the best-selected approaches and the common gain approach in terms of SNR-gain for non-frontal target with one interference

Two additional acoustic scenarios with non-frontal target speakers and multiple interferers have been also evaluated. In the first acoustic scenario, we have used a target speaker at 45 degrees, a back interference speaker at 135 degrees, an interference speaker at 315 degrees, and a lateral interference speaker at 90 degrees. In the second acoustic scenario, we have also used one target speaker at 45 degrees, but this time with two back interference speakers at 135 degrees and 225 degrees, and one interference speaker at 315 degrees. In these acoustic scenarios, some background babble noise (soft, 14 dB lower than the power of the target and each interferer) has been also added. Table 6.11 shows the performance metrics of the signals that are used in the selection/mixing, i.e., outputs of the GSC beamformers (with each beamformer using the local front microphone on its side as a reference) and the common gain method with front right microphone as a reference, under the two acoustic scenarios.

Table 6.11: Results for signals used in mixing/selections, other non-frontal target acoustic scenarios

Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Radian)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Radian)	|ΔMSC|
Non-frontal target at 45 degrees, interferers at 135,315,90 degrees									
Common Gain	5.39	5.97	10.18	10.91	0.27	0.060	1.21	0.124	0.103
GSC	10.82	7.63	7.91	9.1	1.48	0.096	6.65	0.688	0.551
Non-frontal target at 45 degrees, interferers at 135,225,315 degrees									
Common Gain	7.14	7.12	9.85	10.99	0.09	0.054	0.23	0.121	0.081
GSC	14.19	8.98	7.69	11.41	2.19	0.099	15.49	1.151	0.633

As previously mentioned, in the selection/mixing approaches the use of GSC beamformers with a common reference microphone on both sides of the binaural system has an effect on the target distortion response for non-frontal target acoustic scenarios. Consequently, additional measurements have been added here to measure the speech distortion for the target signal, based on the SDR. Table 6.12 shows that the coherence approach provides the best trade-off between noise reduction and cues preservations, whether it uses a common reference or a local front microphone as reference for the beamformers in the selection and mixing approaches. It is also noticeable that the cues for the target speaker are preserved in terms of IPD, by using in the selections/mixing either the GSC beamformers with a common reference microphone (front right microphone) or the GSC beamformer with the local front microphone as a reference. However, using the GSC beamformers with a common reference microphone (front right microphone in these cases) in the selections/mixing changes the cues for the target in terms of ILD, as previously observed. We have chosen the frontal right microphone as a common reference in this case since it introduced a better trade-off between noise/interferers reduction and speech distortion than using the front left microphone. Further investigation of these kinds of trade-off will need to be considered as part of future work.

Table 6.12: The performance of the best-selected approaches under other non-frontal target acoustic scenarios

Ref	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Target (dB)	|ΔIPD | Target (Rad)	|ΔILD| Interf. (dB)	|ΔIPD | Interf. (Rad)	|ΔMSC|
Non-frontal target at 45 degrees, interferers at 135,315,90 degrees										
Local ref. for GSC	**Power Diff.**	8.51	6.23	8.51	10.2	0.36	0.042	3.64	0.152	0.104
	Power	7.58	6.41	8.57	10.08	0.20	0.044	2.45	0.127	0.098
	Coherence	10.18	7.58	7.87	9.28	0.50	0.054	1.77	0.493	0.094
Common ref. for GSC	**Power Diff.**	6.55	6.03	6.68	10.16	7.20	0.092	3.65	0.139	0.089
	Power	6.08	6.12	8.42	10.29	5.11	0.073	3.05	0.119	0.087
	Coherence	6.80	7.60	7.35	9.09	5.46	0.081	3.55	0.355	0.080
Non-frontal target at 45 degrees, interferers at 135,225,315 degrees										
Local ref. for GSC	**Power Diff.**	9.48	6.92	8.39	11.85	0.65	0.035	10.66	0.135	0.088
	Power	9.13	7.15	8.54	11.90	0.44	0.034	8.03	0.125	0.083
	Coherence	12.44	8.79	7.61	11.21	0.54	0.060	6.78	0.829	0.105
Common ref for GSC	**Power Diff.**	8.80	6.87	8.08	9.76	7.20	0.092	1.69	0.132	0.084
	Power	8.11	7.12	8.15	11.65	5.11	0.073	0.97	0.116	0.078
	Coherence	8.90	8.72	8.20	10.65	5.46	0.081	1.24	0.555	0.111

From our informal listening, SDR values above 7-8 dB can be considered to have fairly low distortion, thus all the results above can be considered as low distortion. Note that this distortion metric is measuring the (monaural) processing artifacts and it should not be confused with the measures such as ILD/IPD/ITD/MSC that are used to evaluate the distortion on the binaural cues in the output signals.

To conclude, based on the previous comparisons using different acoustic scenarios, and since we are interested in the approaches that generate the best enhanced output, with SNR-gain similar to the output of the beamformer while at the same time preserving most of the spatial

perception for the target, the interferers and the background noise, we believe that the coherence approach is a promising method that can provide a good trade-off. As it was described in details previously, the coherence approach is a frequency dependent approach using the magnitude and the phase of the coherence between the beamformer output and the common gain output as the classification criteria, and mixing both the magnitudes and the phases of these outputs at each subband-time sample.

6.5 Comparing the Best-Selected Approach Using Different Microphone Configurations

In all the previous tested cases, a 2+2 configuration with wireless transmission of two signals in both directions has been assumed. Even though the aforementioned configuration can be considered in principle as an optimal one for the beamformer design since we have direct access to all four microphones, in practical situations the 2+2 configuration consumes a lot of power and bandwidth because of the required transmission of 4 signals overall. Consequently, other configurations that were described in details in section 2.5 have been considered. As it was discussed in section 2.5, the common gain approach can be implemented using the 2+2 structure and the 1+1 structure with pre-processing on both side (we will refer to it by 1+1 in this section), as they have symmetric structures. These two configurations have been tested and compared under an acoustic scenario with one frontal target signal, one interferer at 315 degree, and soft background noise (14 dB lower than the level of the target and the interferer). Table 6.13 demonstrates the ability of the 1+1 structure in improving the SNR-gain by approximately 1 dB for both channels compared with the 2+2 structure for different algorithms: the GSC beamformer, the common gain approach, and the coherence mixing approach. Our tentative explanation for this perhaps surprising result is that improvement of performance is caused by practical considerations (e.g. performance of 2+2 structure is affected by diagonal loading used in least-squares algorithm in GSC-ANC, or finite number of coefficients used in the GSC, or finite causality delay used in the GSC, etc.). Both configurations generate an output with low target distortion. However, the 2+2 structure shows a better ability in preserving the cues for the acoustic scene in terms of ILD, IPD and MSC for the common gain approach and for the coherence approach, as this structure has a direct access to the frontal left and right noisy signals, which have all the cues of the interferers

and background noise. As it is also shown in Table 6.13 and based on the informal listening, the coherence approach using the 1+1 structure only preserves some of the spatial cues of the acoustic scene, however, it is enough to give the listener an impression of the direction of the interferers and background noise components. Similar to the results with the 2+2 structure, the common gain approach outperforms both the coherence approach and the GSC in terms of ILD, IPD, and MSC using the 1+1 structure. Moreover, in the 1+1 structure, the coherence approach preserves some of the cues and thus outperforms the GSC beamformer as Figure 6.13 to Figure 6.15 shows.

Table 6.13: The performance of the common gain approach, the coherence approach, and the GSC using different configurations for the acoustic scenario of a frontal target and interferer at 315 degrees

Configuration	Tested Cases	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| (dB)	|ΔIPD| (Radian)	|ΔMSC|
2+2	Common Gain	5.65	3.57	12.86	13.2	3.13	0.102	0.072
	GSC	8.23	4.33	9.75	10.61	8.19	1.651	0.600
	Coherence	8.20	4.69	10.01	9.93	6.55	0.478	0.174
1+1	Common Gain	6.83	4.65	10.35	11.42	4.34	0.206	0.100
	GSC	9.06	5.34	8.76	10.11	8.93	2.094	0.731
	Coherence	9.00	5.63	8.94	9.33	6.68	1.404	0.359

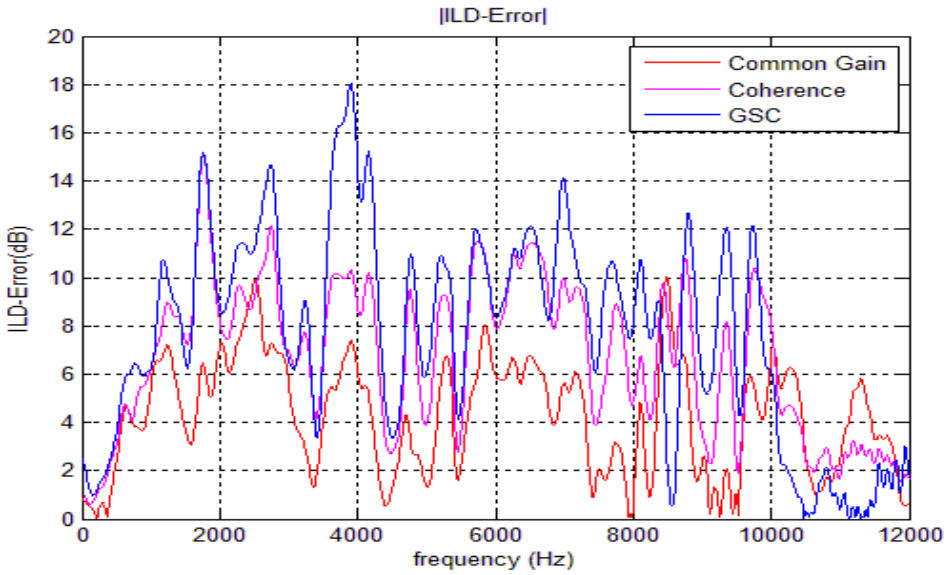

Figure 6.13: ILD error using the 1+1 structure with pre-processing in both side

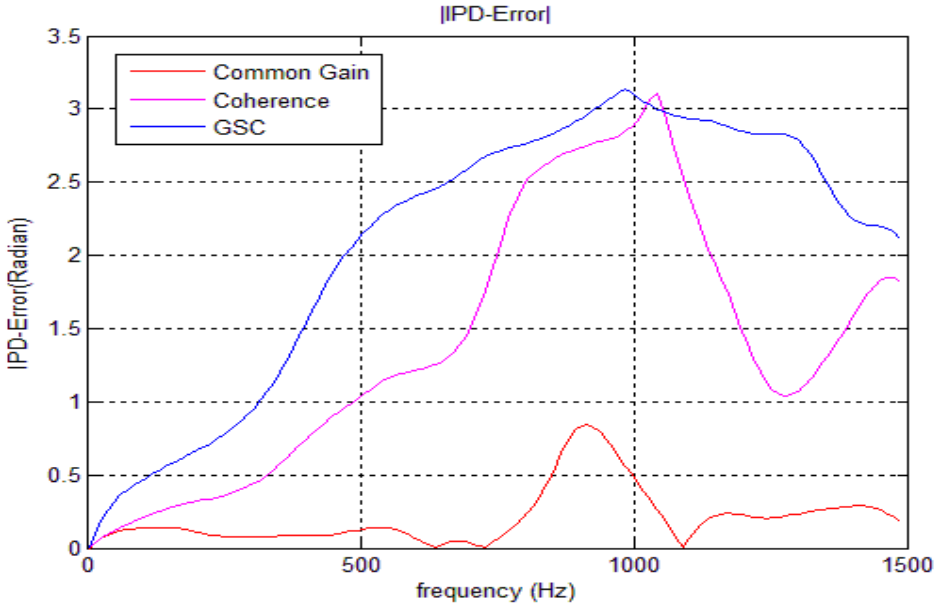

Figure 6.14: IPD error using 1+1 structure with pre-processing in both side

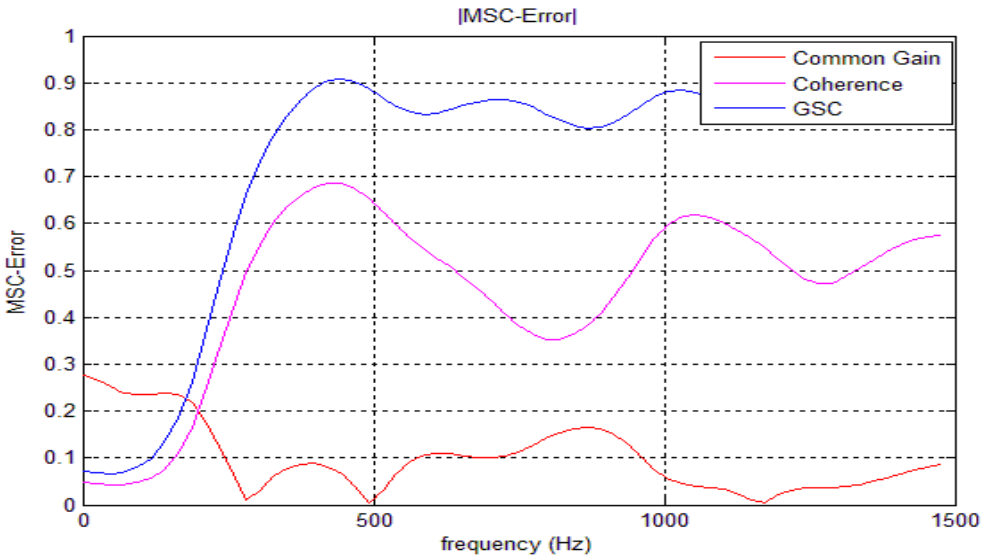

Figure 6.15: MSC error using 1+1 structure with pre-processing in both side

The same configurations which have been evaluated above have also been tested using more complex acoustic scenario with multiple interferers. This scenario has a frontal target at 0 degree, frontal hemisphere interferers at 45 degrees and 315 degrees, a lateral interferer at 90 degrees, and background babble noise (soft, 14 dB lower than the target and the interferers' level).

Table 6.14 shows that using the 1+1 structure introduces a small improvement in the SNR-gain for the outputs of the GSC beamformer, the common gain approach, and the coherence approach. In addition, based on the ILD/IPD/MSC metrics and some informal listening, both the 2+2 structure and the 1+1 structure in the common gain approach and the coherence approach have the capability to preserve the spatial impression of the acoustic scene.

Table 6.14: The performance of the common gain approach, the coherence approach, and the GSC using different configurations for the acoustic scenario target 0 interferers 315, 45, and 90 degrees

Configuration	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| (dB)	|ΔIPD| (Radian)	|ΔMSC|
2+2	Common Gain	3.28	4.44	12.49	11.84	3.85	0.422	0.068
	GSC	5.07	6.48	9.35	9.78	5.38	1.748	0.711
	Coherence	4.71	6.46	8.91	9.40	4.94	1.054	0.121
1+1	Common Gain	3.99	6.11	10.31	9.15	4.75	1.048	0.107
	GSC	4.97	6.51	9.24	9.01	5.44	1.777	0.758
	Coherence	4.87	6.50	9.89	9.04	5.10	1.350	0.418

In order to investigate the other configurations which have been discussed in section 2.5, the front left and front right noisy signals are now used instead of the common gain outputs (left/right) in the classification approach based on the coherence. These configurations are the 2+1 structure without pre-processing and the 2+1 configuration with pre-processing on one side. The same complex acoustic scenario with a target speaker at 0 degree, three interferers at 45, 90, and 315 degrees, in addition to some background noise is used.

Table 6.15 shows that the 1+1 structure again outperforms the other configurations in term of the SNR-gain. All the configurations generate outputs with low target distortion in terms of SDR. On the other hand, the 1+1 structure falls behind the other configurations in terms of the cues preservations of the acoustic scene for the coherence approach. Nevertheless, it has the ability to keep some of these binaural cues, which is enough to preserve the global impression of the acoustic scene.

Table 6.15: The performance of the common gain approach, the coherence approach (using noisy signals), and the GSC using different configurations for the acoustic scenario of frontal target and interferers 315, 45, and 90 degrees

Configuration	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| (dB)	|ΔIPD| (Radian)	|ΔMSC|
2+2	**Common Gain**	3.28	4.44	12.49	11.84	3.85	0.422	0.068
	GSC	5.07	6.48	9.35	9.78	5.38	1.748	0.711
	Coherence	3.98	5.72	9.80	9.87	2.35	0.814	0.101
1+1	**Common Gain**	3.99	6.11	7.21	9.15	4.76	1.048	0.107
	GSC	4.97	6.51	9.24	9.01	5.44	1.777	0.758
	Coherence	4.77	6.42	8.89	9.11	3.51	1.325	0.369
2+1 (one pre-processing)	**Common Gain**	Not Applicable						
	GSC	4.93	6.44	9.62	10.00	5.81	1.765	0.219
	Coherence	4.72	5.58	9.08	9.99	0.78	1.201	0.160
2+1 (without pre-processing)	**Common Gain**	Not Applicable						
	GSC	4.33	6.21	10.54	10.87	4.97	1.764	0.205
	Coherence	3.84	5.27	10.64	10.66	2.11	0.818	0.131

Further comparisons are done to evaluate the effect of using either the common gain outputs or the front microphone noisy signal for the selection/mixing in the coherence approach using both the 1+1 structure and 2+2 structure. In the 1+1 structure, mixing the common gain outputs and the beamformer outputs at each subband-time sample based on the coherence criterion introduces less than 0.4 dB improvement in terms of the SNR-gain under different acoustic scenarios, in comparison with mixing the frontal microphone noisy signals and the beamformer outputs as Table 6.16 shows. At the same time, using the noisy signal outperform the common gain approach in terms of ILD and MSC, while they demonstrate similar capability in preserving the cues for the directional interferers in terms of IPD. On the other hand, in the 2+2 structure,

mixing the common gain outputs with the beamformers outputs based on the coherence criteria introduces more than 1.5 dB improvement in the SNR-gain under some acoustic scenarios, compared with using the front microphone noisy signals and the beamformers outputs in the selection/mixing as Table 6.17 shows.

As a result, in the 1+1 structure, which requires only one binaural wireless link, using either the noisy signals or the common gain outputs in the selection/ mixing shows a comparable performance in terms of the speech enhancement, while using the noisy signals in the selection and mixing approach based on the coherence slightly outperforms using the common gain outputs in the selection and mixing in terms of the cues preservations. Using the noisy signals also reduces the computational complexity compared to using the common gain approach.

Table 6.16: The performance of the 1+1 structure with the coherence approach using the common gain outputs and the noisy signals in the selections and mixing

Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| (dB)	|ΔIPD| (Radian)	|ΔMSC|
Frontal target, interferers at 45,315,90 degrees, and some background noise							
Coherence using common gain	4.87	6.50	9.89	9.04	5.10	1.350	0.418
Coherence using noisy signals	4.77	6.42	8.89	9.11	3.51	1.325	0.369
Frontal target, interferers at 45,225,90 degrees, and some background noise							
Coherence using common gain	8.38	9.99	9.34	9.54	1.74	1.760	0.319
Coherence using noisy signals	8.29	9.48	9.41	9.52	1.12	1.718	0.286
Frontal target, interferers at 225,315,270 degrees, and some background noise							
Coherence using common gain	10.45	9.57	10.46	10.91	1.20	1.763	0.419
Coherence using noisy signals	10.12	9.42	10.43	10.93	0.73	1.760	0.393

Table 6.17: The performance of the 2+2 structure for coherence approach using the common gain outputs and the noisy signals in the selections and mixing

Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| (dB)	|ΔIPD| (Radian)	|ΔMSC|
Frontal target, interferers at 45,315,90 degrees, and some background noise							
Coherence using common gain	4.71	6.46	8.91	9.4	4.94	1.054	0.121
Coherence using noisy signals	3.98	5.72	9.80	9.87	2.35	0.814	0.101
Frontal target, interferers at 45,225,90 degrees, and some background noise							
Coherence using common gain	8.75	9.88	9.74	9.92	0.75	0.938	0.178
Coherence using noisy signals	7.35	8.41	9.95	10.09	0.40	0.816	0.148
Frontal target, interferers at 225,315,270 degrees, and some background noise							
Coherence using common gain	10.48	9.64	9.75	10.46	0.99	1.031	0.118
Coherence using noisy signals	8.36	8.12	9.66	10.66	0.70	0.951	0.095

From the previous comparisons, the 1+1 structure introduces the best improvement in the SNR and has the capability to preserve the global impression of the acoustic scene in the selection and mixing approaches based on the coherence. Thus, we have compared the performance of the selection and mixing approaches based on the coherence using the 1+1 structure, with the performance of the similar selection and mixing approach using the simpler microphone configuration with one signal transmitted from one side to the other side (the 2+1 structure without pre-processing, which requires low computation complexity with no monaural beamformer at each ear). The selection and mixing approaches based on the coherence using the 2+1 structure have been computed using the beamformer outputs and the front microphone noisy signals, as the common gain method is only doable in the microphone configurations that have symmetric structures, as it was mentioned previously.

As the selection and mixing method based on the coherence always includes the binaural beamformers outputs, the two previously mentioned set-ups have been considered. In the first set-up, the binaural beamformer on each side uses a local (e.g. front) microphone as reference signal, while in the second set-up the binaural beamformers on each side use a common microphone as reference signal (i.e., the front microphone of the side with the strongest target component). The first set-up corresponds to the normal distortion-less scenario, where the target component is preserved by the binaural beamformer on each side. The second set-up does not preserve the binaural cues of the target in the stand-alone beamformer, but the cues of the target are partially preserved in the final outputs of the selection and mixing algorithm based on the coherence. The reason for having this second set-up is to reduce complexity: using the first set-up in the 1+1 structure, two beamformers will be required on one of the sides: one for the common gain using a microphone from the other side as the reference signal, and one for the normal beamformer using a local microphone as the reference signal. Using the second set-up, only one beamformer would be required on each side, i.e., a common reference microphone would be used for the beamformers on each side. In the 2+1 structure, using the first set-up only requires one beamformer on each side, as we use the frontal microphone noisy signal instead of the common gain outputs. Thus, the second setup is in principle not required for the 2+1 structure.

To evaluate the proposed algorithm using the previously mentioned configurations, two acoustic scenarios have been used with a frontal target speaker, a single or multiple interferers (all at the same level as the target), and background babble noise (soft, 14 dB lower than the power for the directional sources). In addition, two acoustic scenarios have been considered with a non-frontal target speaker (45 degrees), a single or multiple interferers (all at the same level as the target), and background babble noise (soft, 14 dB lower than the power for the directional sources).

Table 6.18 compares the performance of the common gain method using a common reference microphone signal, the beamformer-only method using the local front microphone signal as a reference, the selection and mixing approach based on the coherence with a common front reference microphone (at the side of the strongest target component) for the binaural beamformers and common gain beamformer, and the mixing and selection approach based on the coherence with a local front microphone for the binaural beamformer and a common reference microphone

for the common gain beamformer. In Table 6.18 , the acoustic scenario with a frontal target signal and one interferer at 315 degrees has been used.

Table 6.18: Frontal target, one interferers at 315 degrees, and diffuse-like background noise

Configuration	Ref. mic.	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Interf. (dB)	|ΔIPD| Interf. (Radian)	|ΔMSC| Backgr. Noise
1+1 (both pre-processed)	FR	Common Gain	6.25	4.92	10.45	11.80	2.67	0.268	0.094
	local ref.	Beamformer-only	9.06	5.34	8.76	10.11	7.92	2.094	0.731
	FR	Coherence	7.42	4.66	8.67	10.84	5.84	1.090	0.207
	FR for common gain/ local ref. for beamfor.	Coherence	9.00	5.33	9.01	10.33	6.56	1.586	0.489
2+1 (without pre-processing)	local ref.	Beamformer-only	8.04	5.56	8.04	12.97	6.32	1.892	0.565
	FR	Coherence	7.40	5.39	10.39	13.84	4.10	1.039	0.340
	local ref.	Coherence	8.23	5.39	10.71	13.84	4.77	0.838	0.206

For this acoustic scenario, the following can be noticed:

- The binaural beamformers using the 1+1 microphone configuration outperform the binaural beamformers using the 2+1 microphone configuration in terms of SNR-gain by around 1 dB in average.
- In the 1+1 microphone configurations, the selection and mixing approach based on the coherence generally provides a better SNR gain than the common gain method.
- All approaches (except the beamformer-only) have the capability to preserve the global impression of the acoustic scene, as we can notice from the cues preservation measurements.
- In the 1+1 and 2+1 microphone configurations, the selection and mixing approach based on the coherence using the local reference microphone signal for the beamformers generates output signals with better SNR improvement compared with the selection and mixing approach based on the coherence using a common reference microphone signal (e.g. front microphone of the strong side).
- The selection and mixing approach based on the coherence with local reference microphone for the beamformers using the 1+1 structure outperforms the same set-up using

the 2+1 structure by less than 1 dB on average in terms of the SNR-gain. However, the 2+1 structure requires less computation (no monaural pre-processing, no common gain, no two binaural beamformers on one side) and does not suffer from a pre-processing delay.

Table 6.19: Frontal target, three interferers at 225, 45, 90 degrees , and diffuse-like background noise

Configuration	Ref. mic.	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Interf. (dB)	|ΔIPD| Interf. (Radian)	|ΔMSC| Backgr. Noise
1+1 (both pre-processed)	FL	Common Gain	8.16	7.81	11.18	9.98	0.52	0.530	0.114
	local ref.	Beamformer-only	8.38	9.87	10.36	9.55	2.52	1.703	0.770
	FL	Coherence	8.45	9.55	10.45	8.51	0.88	1.729	0.454
	FL for common gain/ local ref. for beamfor.	Coherence	8.45	9.95	10.45	9.54	1.48	1.665	0.355
2+1 (without pre-processing)	local ref.	Beamformer-only	8.84	8.66	11.42	11.06	1.48	1.561	0.500
	FL	Coherence	7.27	7.69	11.60	9.38	0.61	0.883	0.165
	local ref.	Coherence	7.27	8.31	11.60	8.31	0.33	0.831	0.145

In Table 6.19 , an acoustic scenario with a frontal target signal and three interferers at 225, 45, and 90 degrees is used. For this acoustic scenario, the following can be noticed:

- The binaural beamformer using the 1+1 microphone configuration outperforms the binaural beamformer using the 2+1 microphone configuration in terms of SNR-gain by less than 0.4 dB in average.
- For this acoustic scenario with the 1+1 microphone configuration, either the selection and mixing approach based on the coherence with a common reference microphone or the mixing and selection approach based on the coherence with a local reference microphone for the binaural beamformers have comparable performance in terms of the SNR-gain. At the same time, both approaches has the capability to preserve the global impression of the acoustic scene, as noticeable from the cues preservation metrics.
- Using the 1+1 microphone configuration with the selection and mixing approach based on the coherence generally produced better SNR improvement than using the 2+1 structure with the same proposed approach.

In Table 6.20, an acoustic scenario with a non-frontal target at 45 degrees and one interferer at 315 has been considered. The front microphone on the right side is used as a common reference, as it has a stronger target level.

Table 6.20: Non-frontal target at 45 degrees, one interferers at 315 degrees, and diffuse-like background noise

Configuration	Ref. mic.	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Target (dB)	|ΔIPD| Target (Radian)	|ΔILD| Interf. (dB)	|ΔIPD| Interf. (Radian)	|ΔMSC| Backgr. Noise
1+1 (both pre-processed)	FR	Common Gain	7.13	4.49	9.02	12.50	0.71	0.116	4.51	0.420	0.113
	local ref.	Beamformer-only	12.04	4.11	8.73	9.20	0.08	0.097	20.73	1.559	0.608
	FR	Coherence	8.02	4.63	6.97	9.72	4.93	0.054	6.91	1.076	0.248
	FR for common gain/ local ref. for beamfor.	Coherence	11.29	4.63	9.28	9.72	0.17	0.057	11.13	1.190	0.335
2+1 (without pre-processing)	local ref.	Beamformer-only	12.86	5.79	9.04	13.41	4.29	0.084	20.18	1.528	0.493
	FR	Coherence	7.27	5.79	6.37	14.28	4.99	0.075	3.77	0.692	0.119
	local ref.	Coherence	10.49	5.79	8.27	14.28	0.47	0.058	10.05	0.751	0.166

For this acoustic scenario, the following can be noticed:

- In the 1+1 and 2+1 microphone configurations, the selection and mixing approach based on the coherence using the local reference microphone signal for the beamformers generates output signals with better SNR improvement compared with the selection and mixing approach based on the coherence using a common reference microphone signal.
- In the 1+1 configuration, the selection and mixing approach based on the coherence generally provides a better SNR gain than the common gain method.
- All approaches (except the beamformer-only) have the capability to preserve the global impression of the acoustic scene, as we can notice from the cues preservation measurements.
- For this acoustic scenario, using either the 1+1 or the 2+1 microphone configuration in the proposed selection and mixing approach based on the coherence generates an output with comparable SNR-gain on average.

In Table 6.21, an acoustic scenario with a non-frontal target signal at 45 degrees and three interferers at 135, 225, and 315 degrees has been used. The front microphone on the right side (the stronger target side) has been used as a common reference.

Table 6.21:Non-frontal target at 45 degrees, three interferers at 135,225,315 degrees, and diffuse-like background noise

Configuration	Ref. mic.	Tested Case	SNR gain left (dB)	SNR gain right (dB)	SDR left (dB)	SDR right (dB)	|ΔILD| Target (dB)	|ΔIPD| Target (Radian)	|ΔILD| Interf. (dB)	|ΔIPD| Interf. (Radian)	|ΔMSC| Backgr. Noise
1+1 (both pre-processed)	FR	Common Gain	7.17	7.80	10.37	12.27	1.60	0.071	1.08	0.314	0.088
	local ref.	Beamformer-only	12.35	6.99	8.30	11.30	1.22	0.076	14.66	1.149	0.706
	FR	Coherence	8.29	7.44	6.77	11.43	4.65	0.048	2.25	1.086	0.141
	FR for common gain/ local ref. for beamformer	Coherence	11.63	7.44	8.38	11.43	0.29	0.052	6.90	1.134	0.240
2+1 (without pre-processing)	local ref.	Beamformer-only	10.57	6.89	8.26	12.62	4.19	0.062	9.71	1.072	0.387
	FR	Coherence	6.60	6.00	5.17	12.21	6.15	0.072	0.31	0.767	0.096
	local ref.	Coherence	8.15	6.00	7.41	12.21	2.40	0.048	3.47	0.808	0.114

For this acoustic scenario, the following can be noticed:

- Using the 1+1 microphone configuration with the proposed selection and mixing approach based on the coherence generally produced better SNR improvement than using the 2+1 configuration.
- In both microphone configurations 1+1 and 2+1, using the local front microphone as a reference for the binaural beamformer in the selection and mixing approach based on the coherence generates outputs with better SNR improvement compared with using a common reference microphone for the binaural beamformers.

To sum up, as the proposed selection and mixing algorithm based on the coherence seems to have reasonable capability to preserve the global impression of the acoustic scene, the selection of the best method then becomes based on the SNR improvement. Table 6.22 and Table 6.23 compare the performance of the proposed selection and mixing algorithm using the 1+1 microphone configuration and the 2+1 microphone configuration using different set-ups. From

Table 6.23 we can notice that using the 1+1 microphone configuration generally provide us with a better trade-off between noise reduction and cues preservation than using the 2+1 microphone configuration in most of the considered acoustic scenarios. In addition, the proposed selection and mixing algorithm using the 1+1 microphone configuration with a common reference microphone for the beamformers and the 2+1 microphone configuration with a local reference microphone for the beamformers show comparable performance in terms of the SNR-gain. However, the 2+1 microphone configuration requires less computation (no monaural pre-processing, no common gain) and generates a lower group delay, as it does not have pre-processing beamformers.

Table 6.22: the performance of the proposed selection and mixing algorithm based on the coherence with different configurations, under different acoustic scenarios

Acoustic scenario	SNR gain	1+1/Coherence/common	1+1/Coherence/local	2+1/Coherence/common	2+1/Coherence/local
target 0 interferer 315	SNR gain left	7.42	9.00	7.40	8.23
	SNR gain right	4.66	5.33	5.39	5.39
target 0 interferers 225,45,90	SNR gain left	8.45	8.45	7.27	7.27
	SNR gain right	9.55	9.95	7.69	8.31
target 45 interferer 315	SNR gain left	8.02	11.29	7.27	10.49
	SNR gain right	4.63	4.63	5.79	5.79
target 45 interferers 135,225,315	SNR gain left	8.29	11.63	6.60	8.15
	SNR gain right	7.44	7.44	6.00	6.00

Table 6.23: difference of performance between different configurations of the proposed selection and mixing algorithm based on the coherence, under different acoustic scenarios

Acoustic scenario	SNR gain	(1+1/ Coherence /local) - (1+1/ Coherence /common)	(1+1/Coherence/common) -(2+1/ Coherence /local)	(1+1/Coherence/common) - (2+1/Coherence/common)	(1+1/Coherence/ local)- (2+1/Coherence/ local)
target 0 interferer 315	SNR gain left	1.58	-0.81	0.02	0.77
	SNR gain right	0.67	-0.73	-0.73	-0.06
target 0 interferers 225,45,90	SNR gain left	0.00	1.18	1.18	1.18
	SNR gain right	0.40	1.24	1.86	1.64
target 45 interferer 315	SNR gain left	3.27	-2.47	0.75	0.80
	SNR gain right	0.00	-1.16	-1.16	-1.16
target 45 interferers 135,225,315	SNR gain left	3.34	0.14	1.69	3.48
	SNR gain right	0.00	1.44	1.44	1.44
	Mean	1.16	**-0.15**	0.63	1.01

6.6 The Effect of Using Different Reference Microphones

As it was mentioned previously, hearing aid users often have a hearing loss in both ears. Hearing aids, therefore, are then used in both ears and one binaural beamformer would be used on each side. These beamformers (e.g., no common gain, no mixing) would use their local front microphone as a reference. However, for the beamformers in the common gain approaches and for the algorithms making use of selection/mixing previously introduced and tested, a common reference for both sides would normally be used. This common reference microphone would be either the front microphone of the “poor-side” or the front microphone of the “good-side”. In this section, we will refer to the side with low input SNR at the local front microphone as “poor-side”, and we will refer to the side with higher input SNR at the local front microphone as “good-side”.

In order to investigate the effect of using the poor-side reference microphone versus using the good-side reference microphone for the beamforming processing, four acoustic scenarios have been tested: two of them with non-frontal target signals and two of them with frontal target signals. The 2+2 configuration was used in this section. Table 6.24 provides the results for those different cases.

For the beamforming processing on the poor-side (using the local front microphone of the poor-side as a reference), the SNR-gain is higher than the SNR-gain at the good-side (using the local front microphone of the good-side as a reference). This is mostly because of the low input SNR on the poor side, which leads to a higher SNR gain on that side if a similar output SNR is generated on both sides. Using the poor side front microphone as a common reference for the beamformers to be used within the common gain approach has in most cases generated higher SNR-gain values, but not always, and the difference was not always significant. So it is not possible to draw more general conclusions here on which side should be used as the reference microphone for the common gain method.

Table 6.24: The effects of using different reference microphones on the performance of the GSC beamformer and the common gain approach

Tested Cases	SNR- in left (dB)	SNR-out left (dB)	SNR-gain left (dB)	SNR-in right (dB)	SNR-out right (dB)	SNR-gain right (dB)
Non-Frontal target at 45 degrees, interferer at 315 degrees , and some background noise						
Sides	Poor-side			Good-side		
GSC	-3.96	8.99	12.94	3.25	6.77	3.52
Common gain (Poor side ref.)	-3.96	3.25	7.21	3.25	8.11	4.86
Common gain (Good side ref.)	-3.96	1.38	5.34	3.25	7.14	3.90
Non-Frontal target at 45 degrees, interferers at 315,135, and 90 degrees , and some background noise						
Sides	Poor-side			Good-side		
GSC	-5.85	4.96	10.82	-3.92	3.71	7.63
Common gain (Poor side ref.)	-5.85	0.81	6.66	-3.92	2.36	6.27
Common gain (Good side ref.)	-5.85	-0.47	5.39	-3.92	2.05	5.97
Frontal target at 0 degree, interferers at 315 degrees , and some background noise						
Sides	Poor-side			Good-side		
GSC	-0.23	8.00	8.23	3.25	7.57	4.33
Common gain (Poor side ref.)	-0.23	5.42	5.65	3.25	6.82	3.57
Common gain (Good side ref.)	-0.23	4.63	4.86	3.25	6.98	3.73
Frontal target at 0 degree, interferers at 315,45, and 90 degrees , and some background noise						
Sides	Good-side			Poor-side		
GSC	-2.48	2.59	5.07	-3.95	2.53	6.48
Common gain (Poor side ref.)	-2.48	0.80	3.28	-3.95	0.48	4.44
Common gain (Good side ref.)	-2.48	1.30	3.78	-3.95	0.27	4.22

Chapter 7 Conclusions and Future Work

7.1 Conclusions

The main purpose of this work was to develop a binaural beamformer that is able to preserve the binaural cues of the acoustic scene, while providing a good noise/interferers reduction, i.e., better than the noise/interferers reduction produced by the common gain approach without the post-filtering proposed in (Lotter & Vary, 2006). The proposed algorithms have to be able to work under real-life environments and were tested using real-life signals processed with a realistic low-delay filter bank, both provided by our sponsor Siemens Audiologische Technik GmbH Group (now Sivantos Inc.).

The experimental evaluations of this work were mostly based on objective measurements, as well as some informal listening. The objective metrics were introduced in Chapter 3 and tested under two types of beamformer configurations in Chapter 5. Several metrics to measure preservations of the cues have been introduced and implemented such as ILD, IPD, ITD, and MSC, and several approaches have been used to derive the ITD measurement. The detailed testing results show that the ITD measurements are only suitable for single directional sources (e.g. a single interference), and they do not have the ability to deal with a summation of multiple interferers unlike the ILD and the IPD measurements.

Different binaural selection and mixing algorithms have been introduced in Chapter 4. These algorithms have the capability of mixing different available signals based on some possible criteria such as the power, the power difference and the coherence. In section 6.2, all the proposed algorithms have been tested, and three approaches based on the power, the power difference, and the coherence were selected because they demonstrated the best trade-off between the cues preservations and the noise/interference reduction. Those best-selected approaches are frequency dependent approaches based on mixing the magnitude and the phase separately. In section 6.4, further testing are done on the best-selected approaches in addition to the common gain approach, under several realistic acoustic scenarios with frontal/ non-frontal targets and with single and multiple interferers from different azimuth angles. Moreover, in the best-selected approaches, the

effect of using a common reference microphone for the beamformers on both sides or a local frontal microphone as a reference for the beamformer on each side has been evaluated. For the best-selected approaches, the results show that in the acoustic scenarios with a non-frontal target using a common reference for the beamformers on both sides distorts the target in terms of ILD, while the IPD cues for the target are kept mostly unchanged. From the detailed results, we have concluded that the frequency dependent coherence approach, which is designed to select and mix the phases and the magnitudes of the common gain and the beamformer outputs at each subband-time sample, introduces the best trade-off between noise/interferers reduction and the ability to preserve the cues under different acoustic scenarios.

In section 6.5, the coherence approach has been further tested using different microphone configurations (e.g. 2+2 structure, 2+1 structure without pre-processing, 2+1 structure with pre-processing on one side, and 1+1 structure with pre-processing on both sides), for a frontal target. The preliminary results shows that the 1+1 structure generates the best noise/interference reduction compared with the other configurations, while at the same time preserving enough of the cues from the acoustic scene. The main advantage of the 1+1 structure over the 2+2 structure is that it requires a single binaural wireless link in each direction, which reduces the power and the bandwidth consumption. Further investigations of the coherence approach using the 1+1 structure showed that in this configuration the best selected methods using selection and mixing of signals can use the front microphones noisy signals instead of the common gain outputs while introducing better cues preservation and approximately similar SNR gain.

In this work, the experimental evaluations were based on objective measurements, as well as some informal listening. Therefore, some audio files which represent some of the experiments made in chapter 6 can be found at the following address:

https://www.eecs.uottawa.ca/~bouchard/papers/audio files_Hala_Asad_MASc_thesis.zip

7.2 Future work

In the coherence approach, some future work should be dedicated to enhance the threshold calculation method in order to enhance the trade-off between the SNR-gain and the IPD. Moreover, in the coherence approach (and also in all the best-selected approaches), fixed portions (α and β) of the magnitude of the available signals are currently added based on the classification decision. Further work should be devoted to adaptively adjusting these factors/portions so that they become frequency dependent and perhaps time-variant as well.

In the power and the power difference approaches, more work should be devoted to determine the optimal time window used to estimate the short time powers of the available signals, which are used in the calculations of the classification decision.

Ideal target steering vector have been considered during the simulations of the beamforming algorithms. However, in practical situations a (noisy) estimated target steering vector is used, which might generate a problem of mismatch between the actual and the estimated target steering vector, degrading the performance of the beamformer. Consequently, more studies should be done to design a robust online target steering vector estimator under realistic acoustic scenarios for hearing aid applications. Moreover, an adaptive GSC-ANC noise canceller should be considered instead of the non-adaptive least-squares GSC-ANC noise canceler that has been used in most of our simulations, to validate the different algorithms under non-stationary acoustic scenarios such as moving targets.

As the evaluation of the proposed algorithms was mostly based on several objective measurements as well as some informal listening, more structured listening tests with a proper number of listeners should be used for further validation of the reported results.

Further validation of the proposed algorithms should be done with hearing impaired people to verify whether they perceptually feel the same effects for the binaural cues or not. Moreover, validation of the proposed algorithms should be done considering their interaction with other components found in a hearing aid (non-linear compression/amplification, feedback cancelation, active noise control, etc.).

References

Begault, Durand R and others. (1994). *3-D sound for virtual reality and multimedia* AP professional Boston etc.

Benesty, J., & Chen, J. (2012). A multichannel widely linear approach to binaural noise reduction using an array of microphones. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, 313-316.

Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone array signal processing* Springer Science & Business Media.

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization* MIT press.

Breed, B., & Strauss, J. (2002). A short proof of the equivalence of LCMV and GSC beamforming. *Signal Processing Letters, IEEE, 9*(6), 168-169.

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America, 92*(6), 3132-3139. doi:http://dx.doi.org/10.1121/1.404209"

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE, 57*(8), 1408-1418. doi:10.1109/PROC.1969.7278

Celesia, G. G., & Hickok, G. (2015). *The human auditory system: Fundamental organization and clinical disorders* (Vol. 129).Elsevier,489-490

Challis, R., & Kitney, R. (1991). Biomedical signal processing (in four parts). *Medical and Biological Engineering and Computing, 29*(3), 225-241. doi:10.1007/BF02446704

Chen, J., & Benesty, J. (2013). On the time-domain widely linear LCMV filter for noise reduction with a stereo system. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(7), 1343-1354.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975-979.

Cornelis, B., Moonen, M., & Wouters, J. (2009). Comparison of frequency domain noise reduction strategies based on multichannel wiener filtering and spatial prediction. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, 129-132. doi:10.1109/ICASSP.2009.4959537

Cornelis, B., Moonen, M., & Wouters, J. (2011). Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1368-1381. doi:10.1109/TASL.2010.2090519

Cox, H. (1973). Resolving power and sensitivity to mismatch of optimum array processors. *The Journal of the Acoustical Society of America*, *54*(3), 771-785.

Desloge, J. G., Rabinowitz, W. M., & Zurek, P. M. (1997). Microphone-array hearing aids with binaural output. I. fixed-processing systems. *IEEE Transactions on Speech and Audio Processing, 5*(6), 529-542.

Dillon, H. (2012). *Hearing aids* (2nd Edition ed.) Thieme Publishers. doi:978-0-9578168-1-7

Doclo, S., Gannot, S., Moonen, M., & Spriet, A. (2008). Acoustic beamforming for hearing aid applications. *Handbook on Array Processing and Sensor Networks*, 269-302.

Elko, G. W. (1995). A simple adaptive first-order differential microphone. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995.* 169-172. doi:10.1109/ASPAA.1995.482983

Flanagan, J., Johnston, J., Zahn, R., & Elko, G. (1985). Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America, 78*(5), 1508-1518.

Frost III, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE, 60*(8), 926-935.

Gabriel, K. J., & Colburn, H. S. (1981). Interaural correlation discrimination: I. bandwidth and level dependence. *The Journal of the Acoustical Society of America, 69*(5), 1394-1401.

Gilkey, R., & Anderson, T. R. (2014). *Binaural and spatial hearing in real and virtual environments* Psychology Press.

Glyde, H., Cameron, S., Dillon, H., Hickson, L., & Seeto, M. (2013). The effects of hearing impairment and aging on spatial processing. *Ear and Hearing, 34*(1), 15-28.

Griffiths, L. J., & Jim, C. W. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation, 30*(1), 27-34.

Hadad, E., Gannot, S., & Doclo, S. (2012). Binaural linearly constrained minimum variance beamformer for hearing aid applications. *Proceedings of IWAENC 2012; International Workshop On Acoustic Signal Enhancement*, 1-4.

Hagerman, B., & Olofsson, A. (2004). A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. *Acta Acustica United with Acustica, 90*(2), 356-361.

Jingdong Chen, & Benesty, J. (2013). On the time-domain widely linear LCMV filter for noise reduction with a stereo system. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(7), 1343-1354. doi:10.1109/TASL.2013.2248719

Kawamura, A., Thanhikam, W., & Iiguni, Y. (2012). Single channel speech enhancement techniques in spectral domain. *International Scholarly Research Notices, 2012*

Kistler, D. J., & Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America, 91*(3), 1637-1647. doi:http://dx.doi.org/10.1121/1.402444"

Klasen, T. J., den Bogaert, T., Moonen, M., & Wouters, J. (2007). Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues. *IEEE Transactions on Signal Processing, 55*(4), 1579-1585.

Linear regression. (2015). Retrieved from

http://www.mathworks.com/help/matlab/data_analysis/linear-regression.html (accessed July 2015)

Lotter, T., & Vary, P. (2005). Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing, 2005*, 1110-1126.

Lotter, T., & Vary, P. (2006). Dual-channel speech enhancement by superdirective beamforming. *EURASIP Journal on Applied Signal Processing, 2006*, 175-175.

Markovich, S., Gannot, S., & Cohen, I. (2009). Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1071-1086.

Marquardt, D., Hohmann, V., & Doclo, S. (2013). Coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8648-8652.

Marquardt, D., Hohmann, V., & Doclo, S. (2014). Perceptually motivated coherence preservation in multi-channel wiener filtering based noise reduction for binaural hearing aids. *IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*, 3660-3664.

McCowan, I. (2001). Microphone arrays: A tutorial. *Queensland University, Australia*, 1-38.

Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, *92*(5), 2607-2624.

Mustière, F., Bouchard, M., Najaf-Zadeh, H., Pichevar, R., Thibault, L., & Saruwatari, H. (2013). Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization. *Signal Processing, 93*(1), 321-325.

Nesta, F., Svaizer, P., & Omologo, M. (2011). Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(3), 624-639.

Schaub, A. (2008). *Digital hearing aids*. New York: Thieme Medical Publisher. doi:978-1-60406-006-5

Szurley, J., Bertrand, A., & Moonen, M. (2013). On the use of time-domain widely linear filtering for binaural speech enhancement. *Signal Processing Letters, IEEE, 20*(7), 649-652.

Thiemann, J., Muller, M., & Van De Par, S. (2014). A binaural hearing aid speech enhancement method maintaining spatial awareness for the user. *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European,* 321-325.

Walther, A., & Faller, C. (2013). Interaural correlation discrimination from diffuse field reference correlations. *The Journal of the Acoustical Society of America, 133*(3), 1496-1502.

Welker, D. P., Greenberg, J. E., Desloge, J. G., & Zurek, P. M. (1997). Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system. *IEEE Transactions on Speech and Audio Processing, 5*(6), 543-551.

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America, 91*(3), 1648-1661. doi:http://dx.doi.org/10.1121/1.402445"

Yu, X., Hu, D., & Xu, J. (2014). Independent component analysis., 59-84. doi:10.1002/9781118679852.ch5