uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Lihong Zhou

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Electrical and Computer Engineering)

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Blind Source Separation Systems for Hearing Aids

TITRE DE LA THÈSE / TITLE OF THESIS

M. Bouchard

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

T. Aboulnasr

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

R. Goubran

W. Gueaieb

Gary W. Slater

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

Blind Source Separation Systems for Hearing Aids

by

Lihong Zhou

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

Master of Applied Science

in Electrical and Computer Engineering

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

Faculty of Engineering
University of Ottawa
December 2009

Your file Votre référence
ISBN: 978-0-494-61302-3
Our file Notre référence
ISBN: 978-0-494-61302-3

Abstract

For many real-life situations, there is more than one speaker at a given time and people need to concentrate on a target sound signal to extract it. This process happens naturally for people with a normal hearing ability, but it is very difficult for hearing impaired persons. In this thesis, we present a system for enhancing the quality of the signal produced by a hearing aid. The proposed system combines spatial information with blind source separation (BSS) to extract the target signal. Results show that the proposed system can locate a target signal in different environments, with a good learning ability. The problem of locating and extracting a target source signal is first investigated. By applying a time-frequency masking method, it is then shown that the performance can be improved. Finally, the problem of underdetermined BSS is investigated and solved by combining a MVDR beamformer with a determined BSS system.

Acknowledgement

I would like to express my gratitude to all those who have given me the possibility to complete this thesis.

I would first give my thanks to Prof. Tyseer Aboulnasr, who has consistently given me advice, support and encouragement. She lead me all the way to do a self-motivated research work on a trial and error basis. Her great insight into my research work ensured a right direction for my exploration while allowing me a great freedom of trying different approaches. She has influenced me in various ways being more than a research supervisor.

I am also thankful to Prof. Martin Bouchard. He is such a diligent researcher and a patient professor in working with his students. He did his own programming and experiments to help me to get out of some confusions. I am also impressed by his attitude of getting every detail taken care of. Being a hard-working researcher towards being perfect, Prof. Bouchard becomes my role model as well.

I would also like to thank the committed members of my thesis defence. Thanks to your comments and suggestions, I am now able to improve my thesis and make it a complete and satisfactory one.

It has been a pleasant experience working in the Signal Processing Oriented Technologies (SPOT) lab. I have benefited from group discussions and have had a good time there with lab members, e.g., Nicholas, Rana, Hisham, to name a few.

Last but not least, my thanks go to my family, especially my husband Pengcheng. Their encouragement and support greatly helped me to finish my study and importantly, to prepare for my future work and life in Canada.

Thank you all / Merci beaucoup!

Table of Contents

List of acronyms and mathematical symbols

BSS	Blind Source (Signal) Separation
BF	Beam Forming
BM	Binary Mask
DOA	Direction of arrival
DP	Directivity Pattern
ERB	Equivalent Rectangular Bands
GCC	Generalized Cross-Correlation
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
ICA	Independent Component Analysis
ITD	Interaural Time Difference
LMSE	Least Mean Square Error
MIMO	Multiple-Input Multiple-Output
MVDR	Minimum Variance Distortionless Response
PCA	Principal Component Analysis
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
T-F	Time-Frequency

Chapter 1 Introduction

1.1 Background

In daily life, a lot of information is obtained from acoustic signals. However, in most occasions, the information received needs to be processed by the auditory system in order for it to be decoded or perceived.

One of the most typical examples is that in a noisy environment, such as in a cafeteria, or in a party, people need to focus on listening to someone's talk. During this process, the human auditory system has several tasks to accomplish. It locates the audio source and focuses the ears in the direction of the audio source, and then the brain separates the noisy acoustical signals, removing the interference signals and only transferring the target signals of interest.

This process comes natural to humans with a normal hearing ability. However, as humans age or because of variety of their causes, their hearing ability can degrade. The auditory signal which is received from the environment is then different from what people with normal hearing receive. Not only can the hearing threshold level and the discomfort hearing level change, but also the time and frequency selectivity may degrade. The resulting output signals for those hearing impaired persons are distorted at different levels. As a result, after the processing by a impaired auditory systems, much information is not available.

Hearing aids were developed to mitigate the impact of hearing impairment. Ideally, according to each person's hearing impairment profile, hearing aids parameters can be optimally set and according to different acoustic environments the parameters will automatically adjust to track a target signal in the given environment to optimize the speech intelligibility or in general signal quality. For example, the volume of the hearing aid output can be tuned according to the user's preference and environment. The hearing aid processing can change in different noisy situations, and since the speakers will change position and time of talking from time to time, it is highly desirable that the system have

the ability to locate and separate the target signal. Figure 1.1 shows a general diagram of a hearing aid.

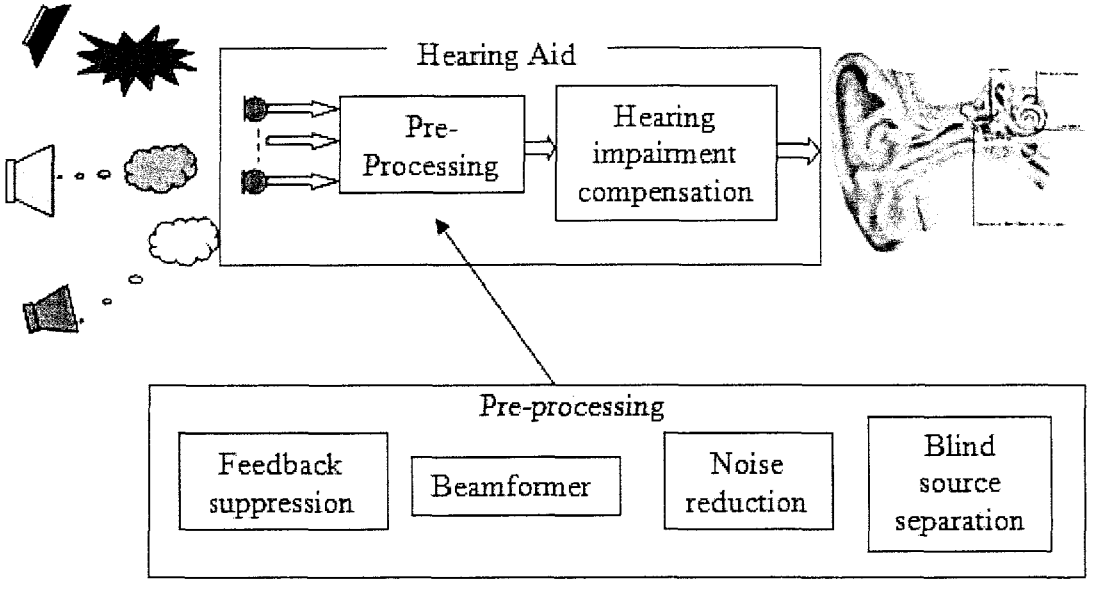

Figure 1.1: Diagram of a general Hearing Aid

1.2 Statement of problem

According to the diagram of a general hearing aid in Figure 1.1, a pre-processing stage is found before the hearing compensation stage (i.e. gain, amplification). In the pre-processing stage, the hearing aid needs to remove interferences and extract the target speech for the user by doing several tasks: feedback suppression, beamforming, noise reduction, blind source separation, etc. One of the most challenging acoustic environments for the auditory system is when multiple speakers are talking simultaneously, which is shown in Figure 1.2. Many studies have shown that the speech intelligibility in a multi-speaker environment is greatly affected by the number of competing talkers as well as by the temporal character of the noise [Loi'07]. In this multi-speaker scenario, even humans with a normal hearing ability need to make an effort to follow the conversation. Because of their hearing impairment, hearing impaired persons

may find that it is particularly difficult for them to discriminate the interference speeches from the target speech. Moreover, they may not be able to track the target speaker.

Blind source separation (BSS) is one of the possible ways of helping a hearing impaired person to track and extract the source signal of interest in a noisy environment. BSS does not need any prior knowledge about the environment. Theoretically, BSS can recover N speech signals from M mixtures, where N is the number of source signals and M equals the number of microphones in a hearing aid [Hay'00].

As the hearing aid technology develops, binaural hearing aids are becoming available in the market, making use of a wireless link connecting the hearing aid of both ears [Jad'08]. In such cases, the number of microphones M increases and more signal processing can be performed on the received sensor signals. However, in many real-life acoustic environments, due to the size of a hearing aid, it is not possible to put as many microphones as we want into a hearing aid, and the number of sound sources exceeds the number of microphones. As the number of source signals increases, the acoustic environment becomes more and more complicated. The cross-talk between several speakers (i.e. the fact that people receive signals coming from all speakers at the same time) makes it more difficult for the pre-processing system in hearing aids to extract the information of interest and to track different audio sources.

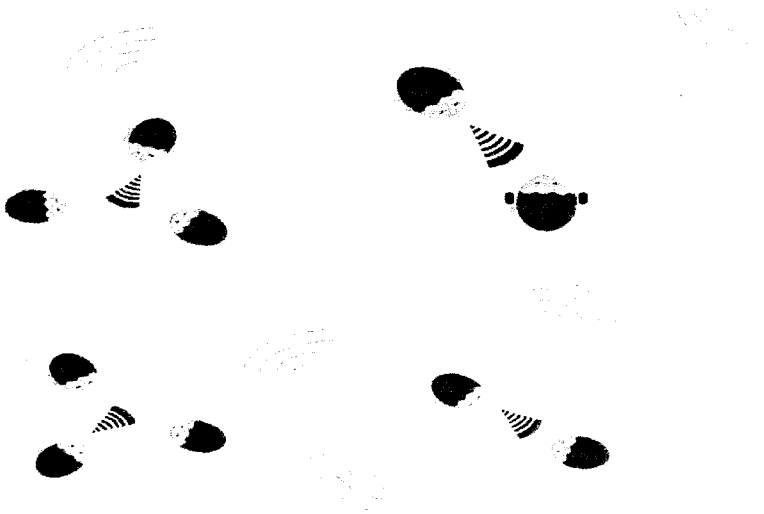

Figure 1.2: Multiple-speaker environment

1.3 Thesis Objective

The objective of this thesis is to develop a system which can attenuate directional interference speeches and extract the most dominant speech. This is basically an application of the blind source separation problem to the challenging scenario of a hearing aid application. The challenge includes: (1) Separating speech mixtures and adapting its coefficients under determined scenarios, i.e., in conditions where the number of microphones is equal to the number of source signals; (2) Dealing with the limited distance between the microphones given the size of the hearing aid, this results in the microphones having less diversity, which makes it harder for blind source separation systems to converge. In order to guarantee the quality of the separated speech signals, a new pre-processing step is needed for underdetermined scenarios, to remove some speech interferences by explicitly exploiting the spatial information of the speech mixtures. In this thesis, we aim to propose such a BSS system for the underdetermined case with closely placed microphones in hearing aid applications. By combining the spatial information and blind source separation, our proposed system aims to separate a speech signal from different noisy environment and track it.

1.4 Thesis outline

In this thesis, approaches to blind source separation of mixtures in hearing aids applications are reviewed, with emphasis on methods based on independent component analysis, time-frequency (T-F) masking and beamforming. Both determined and underdetermined blind source separation problems are analyzed.

In Chapter 2, some fundamental aspects of blind source separation theory are presented. Related concepts, such as the scaling problem and the permutation problems are explained as well. Some popular algorithms for both determined and underdetermined blind source separation are reviewed in this chapter. Some basic objective measures for blind source separation systems are also described.

In Chapter 3, the learning ability of blind source separation systems is analyzed and the blind source separation combined with T-F masking method is proposed to reduce the adaptation delay of the system. Different experiments are conducted to evaluate the adaptation performance of our proposed system under different acoustic environments.

In Chapter 4, the geometrical information of the speech mixtures is explored by applying beamforming as a pre-processing step to remove some interference signals in the underdetermined blind source separation problem. Direction-of-arrival (DOA) estimation methods are discussed and simulated in the chapter, in order to obtain robust and accurate DOA estimation for beamforming. The experimental results demonstrate that by combining spatial information with blind source separation, the underdetermined blind source separation problem can be solved effectively.

In Chapter 5, a summary of the work is presented and problems existing in the current DOA estimation algorithms are discussed. Future work including a possible method for improving the accuracy of the DOA estimator is also mentioned in the chapter.

Chapter 2 Overview of speech enhancement using blind source separation

Adaptive signal processing is based on a self-adjusting process, which uses a training sequence of data to adjust its parameters for extracting desired signals. If the desired signals are known, the process is called supervised adaptive signal processing. However, in real-life, the desired signals are usually unknown, and the adaptation is only based on the observed signals. This challenging adaptation process is thus referred to as unsupervised adaptive signal processing. Blind source separation is a typical example of unsupervised adaptive signal processing. It aims to recover the source signals from observations without either prior information about the source signals or knowledge of the process by which the sounds were mixed to produce the observations.

The human auditory system is a natural unsupervised adaptive signal processing system: significantly more advanced than any technical equipment so far. With a normal hearing ability, a human being can tell the direction of different speakers, different tones, voices and noises in a complex auditory scene. An even more important skill is that the human auditory system can analyze the relevance of different speeches and ignore unimportant ones. A blind source separation system tries to mimic some functions of the human auditory system to achieve the same goal.

The objective of this chapter is to review some basic knowledge about blind source separation algorithms.

2.1 Signal mixture model

We start with a simple acoustic scenario. If the environment has no reverberation, or very weak reverberation, the signal can be modeled as an instantaneous mixture model, which can be written as

$$x_j(n) = \sum_{i=1}^{N} h_{ji} s_i(n), j = 1,2,\cdots N \tag{1}$$

where s_i is the *ith* signal source, x_j is the *j-th* observation, h_{ji} is the attenuation parameter from source i to microphone j, N is the number of source signals and M is the number of microphones.

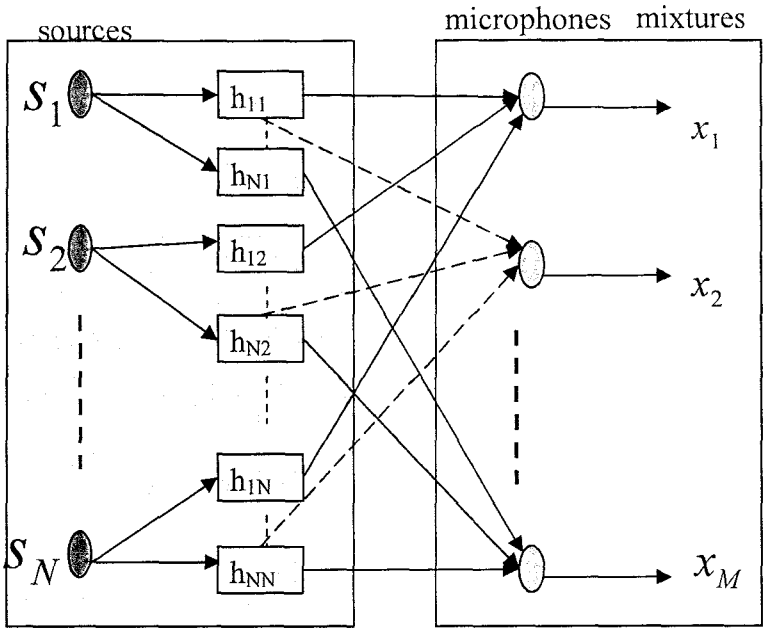

Figure 2.1: The signal mixing model

In a real acoustical environment, the audio sounds pass through several kinds of distortion before they are perceived by ears or microphones. Depending on the room situation, these distortions mainly contain room reverberation, attenuation and time delay.

Under these circumstances, N signals received by M microphones can be modeled as a convolutive mixture model, which can be written as

$$x_j(n) = \sum_{i=1}^{N}\sum_{k=1}^{p} h_{ji}(k)s_i(n-k+1), j = 1,2,\cdots M \tag{2}$$

where s_i is the *ith* source signal, x_j is the received signal by the *j-th* microphone, and $h_{ji}(k), k = 1\cdots p$ is a p tap impulse response modeling the transmission channel from source i to microphone j. In real situations, the acoustic environment, head shadow effect,

the position of the audio source relative to the head, etc, all have an effect on the mixing function h_{ji}.

From Figure 2.1, we can see that the real acoustic environment can be simplified as a signal mixing process. All the audio signals are convolved with the function representing the specific path from a given source to a given microphone. The outputs are then added and the signal mixtures are observed by the microphones located on the hearing aid.

Time domain convolution reduces to a multiplication in the frequency domain. The frequency domain speech mixtures $X(\omega) = [X_1(\omega), X_2(\omega), \cdots X_M(\omega)]^T$ can be written as a multiplication of the mixing matrix $H(\omega)_{M \times N}$ and the original sources $S(\omega) = [S_1(\omega), S_2(\omega), \cdots S_N(\omega)]^T$, which is similar to the time domain instantaneous model:

$$X(\omega) = H(\omega)S(\omega). \tag{3}$$

where $S_i(\omega), i = 1,2,\cdots N$ and $X_j(\omega), j = 1,2,\cdots M$ are the frequency transform of s_i and x_j , and $H(\omega)_{M \times N}$ is the M*N matrix of the frequency transform of $h_{ji}(k)$, i=1,...,N; j=1,...,M.

As a result, to simplify the process, many convolutive mixtures systems are processed in frequency domain.

It should also be noted that if signals are transmitted in a communication channel, like a telephone channel or a wireless channel, channel distortion is added (including possibly non-linear artefacts from coding and decoding of the speech or audio signals). Moreover, several kinds of additive noise contaminations inevitably exist in real environments.

2.2 Spatial information processing methods for hearing aids

To improve the speech intelligibility and speech quality, many speech enhancement algorithms have been proposed for hearing aids applications. The spectral-subtraction algorithms are mainly based on the idea that, by subtracting an estimate of the noise spectrum from the noisy speech spectrum, one can obtain an estimate of the clean signal spectrum with the assumption that the noise is additive. These methods are intuitive and the enhanced signal spectrum is not derived in an optimal way. Based on mathematically tractable error criterions, various linear or nonlinear estimators, such as the maximum-likelihood estimators and other Bayesian estimators, can produce the optimum parameters of interest (i.e. typically frequency dependent gain factors) to enhance the noisy speech. In this thesis, we focus on multi-microphone systems and consequently on spatial information processing for speech enhancement in hearing aids applications.

There have been several previous patents involving the use of spatial processing for hearing aids applications. A few of those patents are discussed below. The patent "Method for improving spatial perception and corresponding hearing apparatus" [Pat'04] demonstrates a method for improving the spatial perception of acoustic signals. In this patent, an input signal is received via the aid of a binaural hearing apparatus. After analysis, at least one of the variables that influence spatial perception of the binaural output signals will be changed in the hearing apparatus. For example, the distance or direction of a source at/from which it is perceived can, with the aid of a classifier or directional microphone, be varied automatically for corresponding input signals. As a result, an improved spatial perception can be achieved.

It is sometimes suitable to switch between different "directional characteristics" i.e. different algorithms to process the input signals from the microphones, in order to adapt to source positions, environmental changes, etc. However, this switching process can create artefacts. The patent "Hearing aid and operating method with switching among different directional characteristics" [Pat'08a] gives a solution so that the signal level of microphone signals which are respectively obtained from different microphones units with different-order directional characteristics (i.e. processing algorithms) are matched

with regard to a reference signal. The switching is then always carried out with the same signal level, so that the switching does not result in any sudden level changes.

The patent "System and method for adaptive multi-sensor arrays" [Pat'07] proposes the use of an adaptive differential microphone array method with the following steps: receiving a signal, estimating a measured signal spectral covariance matrix of the signal, and estimating a direction of arrival (DOA) of the signal based on the measured signal spectral covariance matrix of the signal. A fractional delay obtained from the signal spectral covariance matrix will produce the DOA estimation of the target signals, and it can finally be applied to a differential microphone array filter to process the signal mixture and produce the clean target signal from the estimated DOA.

Blind source separation algorithms are also popular for multi-speaker environments and have been used in hearing aids applications. The separated outputs of BSS contain more than one signal, which causes the problem of deciding which of the signals produced through BSS should be forwarded to the hearing aid wearer. The patent "Method for operating a hearing aid, and hearing aid" [Pat'08b] introduced a "speaker" mode, where the listener can track and select an acoustic speaker source in an ambient sound, according to a database of speech profiles of preferred speakers. For example, the listener may prefer a specific language or a frontal target, and only the output matching the speech profiles in the user's database will be tracked by the signal processor [Pat'08b].

As previously mentioned, this thesis focuses on spatial information processing for hearing aids systems involving the use of BSS algorithms. In the following sections, more details and background on BSS algorithms are provided.

2.3 Fundamentals of blind source separation and independent component analysis

An acoustic environment is considered where a given number of sensors receive unknown signal mixtures. At each receiver, the observation obtained from each sensor is a different combination of the source signals, and no information about this combination

is available. The field of BSS has developed a set of algorithms to recover the unobserved individual source signals from a group of observed signal mixtures. Here, the term "blind" stresses the lack of prior knowledge for the mixing process. Since there is no information about the source signals and the mixing process, the blind source separation can also be seen as unsupervised adaptive signal processing [Fra'98][Jad'08][Cao'96]. Among blind source separation algorithms, independent component analysis (ICA) is one of the most commonly used techniques.

2.3.1 Independent Component Analysis by maximization of nongaussianity

In order to analyze and process a signal mixture, it is necessary to make the statistically strong assumption of independence signals; a physically reasonable assumption. Based on the central limit theorem, we know that a sum of independent random variables tends to become a Gaussian distribution as the number of variables increases. As a result, the goal of BSS can be formulated as designing an adaptive filter to produce output signals as independent as possible, as evaluated by an information-theoretic cost function [Fra'98][Hay'00]. The problem of estimating one of the independent source signals can then be translated into maximizing the nongaussianity of outputs by adjusting the coefficients of the filter [Els'06][Pan'07].

Maximum Likelihood (ML) estimation based on the gradient descent method

Based on the instantaneous signal mixture model with N source signals $s=[s_1,s_2,\cdots s_N]^T$, the blind source separation system processes the source mixtures $x=[x_1,x_2,\cdots x_M]^T$ which is collected from M microphones ($M \le N$ here) and yields the outputs $y=[y_1,y_2,\cdots y_M]^T$. Assuming T observations $x=[x_1(t),x_2(t),\cdots x_M(t)]^T, t=1\cdots T$, the likelihood of the probability density function (pdf) of the observation is evaluated as

$$L = \prod_{t=1}^{T} p_x(x) \tag{4}$$

Given that $x = H \cdot s$, we have that $p_x(x) = \left|\det(H^{-1})\right| p_s(s)$, where H is the instantaneous mixing matrix in time domain [Car'97]. If the separation system works perfectly, the unmixing matrix can be thus written as $W = H^{-1}$ and $p_y(y) = p_s(s)$. Therefore, the above pdf of the observations can be rewritten as

$$p_x(x) = \left|\det(W)\right| p_y(y) = \left|\det(W)\right| \prod_{i=1}^{M} p_i(y_i(t)) \tag{5}$$

The likelihood function can then be rewritten as

$$L = \prod_{t=1}^{T} \prod_{i=1}^{M} \left|\det(W)\right| p_i(y_i(t)) \tag{6}$$

In order to estimate an unmixing matrix which maximizes the likelihood function, we use the normalized log-likelihood as a cost function as

$$J(W) = \frac{1}{T}\log(L) = \log\left|\det(W)\right| + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\log[p_i(y_i(t))]. \tag{7}$$

Using a basic gradient descent approach, we can obtain

$$\Delta W = \frac{\partial J(W)}{\partial W} = \frac{\partial}{\partial W}\{\log\left|\det(W)\right| + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\log[p_i(y_i(t))]\} \tag{8}$$

In the above equation, the first part can be simplified as

$$\frac{\partial}{\partial W}\{\log\left|\det(W)\right| = W^{-T} \tag{9}$$

The second part can be rewritten as

$$\frac{\partial}{\partial W}\{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{M}\log[p_i(y_i(t))]\} = E[\frac{\partial}{\partial W}\sum_{i=1}^{M}\log(p_i(y_i))], \tag{10}$$

where $E[\cdot]$ is expectation operator.

Since $y_i = \sum_{j=1}^{M} w_{ij}x_j$, we also have:

$$\frac{\partial}{\partial w_{ij}}\sum_{i=1}^{M}\log(p_i(y_i)) = \frac{\partial}{\partial y_i}\log(p_i(y_i))\frac{\partial y_i}{\partial w_{ij}} = \frac{1}{p_i(y_i)}\frac{\partial p_i(y_i)}{\partial y_i}x_j. \tag{11}$$

Let $g(y_i) = -\frac{1}{p_i(y_i)}\frac{\partial p_i(y_i)}{\partial y_i}$ and $g(y) = [g(y_1), g(y_2), \cdots, g(y_M)]^T$

then $\Delta W = W^{-T} - E[g(y)x^T]$.

The final *k-th* update unmixing matrix from ML estimation with a gradient descent method is then:

$$W_{k+1} = W_k + \mu(k)(W^{-T} - E[g(y)x^T]) \tag{12}$$

where $\mu(k)$ is a scalar update step at instant *k*.

Maximum Likelihood estimation with the natural gradient descent method

From a geometry information perspective, the mixing matrix is not invertible, and in practice the calculation of the inverse unmixing matrix can cause some problems. As a result, the above algorithm is often replaced by the so-called natural gradient descent algorithm [Ama'96]:

$$\Delta W = \frac{\partial J(W)}{\partial W} W^T W = \{W^{-T} - E[g(y)x^T]\} W^T W = \left\{I - E[g(y)]y^T\right\} \cdot W \quad . \tag{13}$$

The resulting *k-th* update of the unmixing matrix from ML estimation with the natural gradient descent method is:

$$W_{k+1} = W_k + \mu(k)(I - E[g(y)y^T])W_k \tag{14}$$

Maximum Likelihood estimation with the gradient descent method for convolutive BSS

Processing time domain convolutive mixtures with the previous blind source separation methods is very computationally complex and the convergence speed is low. However, when transferring to the frequency domain, the convolutive mixtures transfer into complex-valued instantaneous mixtures. Using this approach, many computationally efficient time-domain instantaneous blind source separation algorithms become applicable to the problem of convolutive mixtures.

Using the short-time Fourier transforms (furthermore using DFTs/FFTs in practice), the mixture in the frequency domain can be written as

$$X(f,t) = H(f,t)S(f,t) \tag{15}$$

where f is the frequency and t represents the frame number in the time domain.

The output of the unmixing system is

$$Y(f,t) = W(f,t)X(f,t) = W(f,t)H(f,t)S(f,t) \quad . \tag{16}$$

For each frequency bin, ML estimation with the natural gradient descent method is applied to the complex-valued mixtures, and we can obtain the *k-th* update of the unmixing matrix for this frequency bin as:

$$W_{k+1}(f) = W_k(f) + \mu(k)(I - E[g(y(f))y(f)^H])W_k(f) \quad (17)$$

Another advantage of using the frequency domain blind source separation is that the separation for each frequency bin can be performed in parallel i.e. independently. However, this also brings the permutation and scaling problems, as described next.

2.3.2 Scaling problem and permutation problem

In the frequency domain, the blind source separation updates the unmixing matrix and minimizes the nongaussianity, producing the outputs:

$$Y(f,t) = W(f)X(f,t) = W(f)H(f)S(f,t) \quad (18)$$

If the separation system works perfectly and all the source signals are separated from the mixtures one by one, then,

$$W(f) \cdot H(f) = I \quad (19)$$

However, in practice, no prior information is available for the original sources, and the order of the independent components cannot be determined. Thus, the recovered signals from a BSS system are a scaled and permuted version of the original sources, that is:

$$Y(f,t) = W(f) \cdot H(f) \cdot X(f,t) = P(f) \cdot D(f) \cdot S(f,t) \quad (20)$$

where $P(f)$ is a permutation matrix and $D(f)$ is a non-singular diagonal matrix.

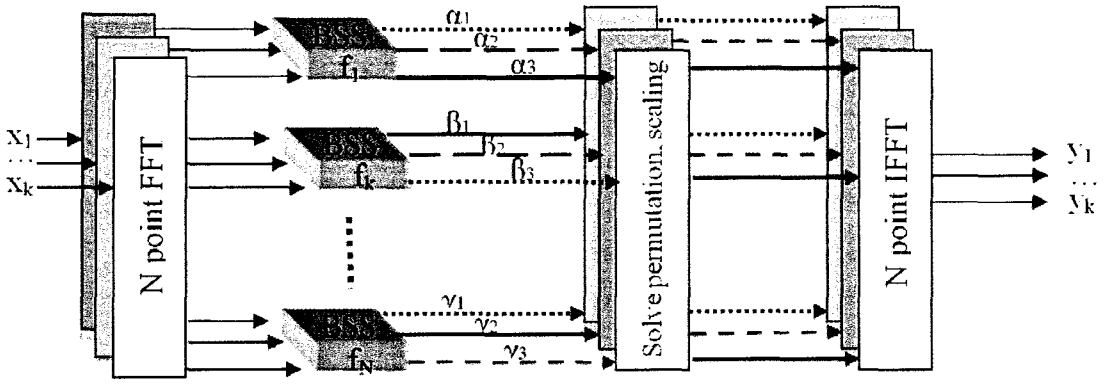

Figure 2.4: Illustration of Scaling and Permutation in frequency domain convolutive BSS

Figure 2.4 illustrates the scaling and permutation problem of frequency domain convolutive blind source separation. The scaling value and permutation order at each frequency bin can be different. For each of the independent component analysis output, the components at different frequency bins may not come from the same source signal and can have inconsistent scales. As a result, we need to align the order of the permutation and rescale the separated signals in each frequency bin, so as to make sure that each output comes from the same source signal for all frequency bins, and has consistent scaling values for each frequency bin.

In order to solve the scaling problem, many methods have been proposed. [Sma'98] solved the scaling ambiguity by simply normalizing the unmixing matrix $W(f)$ to ensure that the unmixing matrix does not scale the data for each frequency bin. Later, [Mur'01] solved the scaling problem through multiplying by a scalar which is calculated from the inverse of a separation matrix.

The permutation problem is more complex compared with the scaling problem. However, a lot of efforts have been devoted to this problem. In [Saw'04], the author proposed that the permutation problem can be solved through estimating a rough DOA from the blind source separation coefficients. Since the blind source separation has a similar structure as classic beamforming, for each frequency bin the directivity pattern of the unmixing

matrix can be obtained. A rough DOA of the source signal is then derived from the directivity pattern. These estimated DOAs are then used as flags for permutation alignment. However, this method cannot solve the permutation problem for all frequency bins, because DOA estimation methods do not work very well for high frequencies and low frequencies.

For speech signals, the envelopes of adjacent frequency bands are highly correlated. Using this property, the permutation problem can be solved [Mur'01]. However this method is based on local frequency information, thus one error in any frequency band may lead to consecutive misalignments in adjacent frequency bands.

2.4 BSS algorithms for overdetermined, determined and underdetermined systems

The most general BSS problem can be formulated as a "multiple-input multiple–output (MIMO) nonlinear dynamic system with its signal mixing system unknown. The objective is to find an adaptive inverse system, if it exists and is stable, to estimate the primary input signals" [Hay'00]. Usually, if we do not consider the time delay, the mixing system can be viewed as a scaled system. In this case, the BSS is referred to as instantaneous BSS. But if time delays are involved, the observed signals become the convolution of the source signals and some channels. In this case, convolutive BSS is used to process the convolutive mixture [Jad'08].

In some applications, it is beneficial to apply some kind of pre-processing, such as pre-whitening, principle component analysis (PCA) and preliminary noise reduction, either to eliminate redundancy (in the case of more sensors than sources), to reduce additive noise, or to improve convergence properties of adaptive algorithms by de-correlating mixing signals [Hay'00].

2.4.1 Blind source separation techniques for determined and overdetermined situations

In a real life acoustic environment, often the number of source signals N does not equal the number of microphones M. If $N < M$ or $N = M$, the resulting system is called respectively overdetermined or determined blind source separation. Otherwise, it is called underdetermined when $N > M$, which means that the microphones pick less information than needed to exactly recover the sources.

Several algorithms exist for solving the over-determined problem. A typical algorithm uses principle component analysis (PCA) as a pre-processor to transform the signal from M dimensions to N dimensions and to remove the redundancy, followed by a determined blind source separation system.

As explained in section 2.2, independent component analysis is a very common way to solve the blind source separation by maximizing the nongaussianity of the outputs. In determined blind source separation, the recovered signals are obtained from the unmixing matrix, and the difficulties lie in the permutation problem and the scaling problem.

2.4.2 Blind source separation techniques for underdetermined situations

In some real situations, a noisy environment may contain several audio sources. However, the number of microphones in hearing aids is limited due to the size of a hearing aid, which also limits the performance of the blind source separation. As a result, many efforts have been devoted to underdetermined blind source separation, where the number of sensors is less than the number of sources.

The problem of underdetermined blind source separation stems from separating N sources from M observations, where $N > M$. The noise-free model is written as:

$$X = HS \tag{21}$$

where the mixing matrix $H \in R^{M \times N}$ is unknown. When $N > M$, the inverse of the mixing matrix H does not exist. Even when the mixing matrix is known, its solution is not unique.

Ideally, the unmixing system gives the output Y, which can be written as

$$y_i = \sum_{j=1}^{M} w_{ij} x_j \tag{22}$$

where w_{ij} are the unmixing coefficients, and the observed speech mixtures are

$$x(t) = [x_1(t), x_2(t), \cdots x_M(t)]^T, t = 1 \cdots T \tag{23}$$

The original source signal is composed of N unknown sources, where the number of source N is also usually unknown: $s(t) = [s_1(t), s_2(t), \cdots s_N(t)]^T, t = 1 \cdots T$,

For instantaneous speech mixtures we have $x_i(t) = \sum_{l=1}^{N} h_{il} s_l(t)$

and then $y_i = \sum_{j=1}^{M} w_{ij} \sum_{l=1}^{N} h_{il} s_l = \sum_{l=1}^{N} s_l \sum_{j=1}^{M} w_{ij} h_{il}$

For convolutive speech mixtures we have $x_i(t) = \sum_{l=1}^{N} \sum_{\tau=0}^{T} h_{il}(\tau) s_l(t-\tau)$

and then

$$y_i = \sum_{j=1}^{M} w_{ij} \sum_{l=1}^{N} \sum_{\tau=0}^{T} h_{il}(\tau) s_l(t-\tau) = \sum_{l=1}^{N} \sum_{\tau=0}^{T} s_l(t-\tau) \sum_{j=1}^{M} w_{ij} h_{il}(\tau) \tag{24}$$

From this equation, each component of the original speech signals $s_l(t), l = 1,2,\cdots M$ will inevitably appear in the final output $y(t)$. Therefore, it is suitable to formulate the search in the mixing space rather than in the separation space [Jad'08]. The usual approach to separate signals in underdetermined BSS is by exploiting the sparseness of speech signals. Table 2.1 below lists several popular algorithms for underdetermined BSS.

Table 2.1 Algorithms for underdetermined BSS

Author	Restrictions (assumptions)	Method	Complexity
Joujine & al. [Jou'00]	Source signals are W-disjoint orthogonal*; Amplitude-delay model	Amplitude-delay parameter estimation	O(T*K) T: time resolution K: frequency resolution
Bofill & al. [Bol'00] [Bol'01]	The sources are sparse. If they are not sparse in the time domain, uses linear transformation to improve sparsity; Instantaneous mixting model.	Potential-function based clustering for estimating the mixing matrix; uses shortest-path method to estimate the sources	The difficulty of the separation depends more on the complexity of the sounds than on the number of sources present Two-step procedure.
Bofill [Bol'01b]	Independent Sources; the sources are sparse. Laplacian distribution of magnitudes with equal variances; Attenuate and delay model for speech mixtures.	Estimate the attenuation matrix by clustering of the magnitudes; estimate the delay matrix by maximizing clustering over phase; estimate the source components by maximum of a posteriori approach to source decomposition.	The difficulty of the separation depends more on the complexity of the sounds than on the number of sources present. Three-step prodecure.
Viste & al. [Vis'01]	Independent Sources; the transfer functions in the mixing matrix are similar (application in hearing aids) ; the sensor geometry is used to estimate the maximum delay and use it as the	Use the cross-correlation of the sensor signals to estimate the delay between the sensors for each source signal; apply the gradient method to the LMSE to estimate the de-mixing matrix	Two-step procedure.

	upper limit of reverberation delay;		
Li & al. [Li'03] [Li'06]	Source signal is sparse enough; if possible, apply wavelet packets transform or use T-F transformation pre-processing to improve sparsity; source signals are of Laplacian distribution	Use K-means clustering algorithm to estimate the mixing matrix; for a given mixing matrix, the source signals are estimated by maximizing posterior distribution	Two-step procedure
Araki & al. [Ara'04a] [Ara'04b]	When signals are sufficiently sparse, we can assume that the sources overlap at rate intervals; DOAs are known.	First remove N-M sources using T-F binary mask (in frequency domain or from a DOA view); then apply ICA to separate the source signals	Two-step procedure
Vincent & al. [Vin'04]	Source signal is sparse instantaneous musical mixtures; the source signals are independent and follow a Gaussian prior with known variance.	Add pre-processing step at the beginning and process the signals in ERB frequency scale, then estimate normalized mixing matrix based on sparse representation; finally, recover the source signal using the estimated mixing matrix	Three-step extraction procedure
Pedersen & al. [Ped'08]	Source signal is sparse; Each source signal comes from a distinct direction; the number of sources does not need to be known in advance (sources are iteratively extracted from the mixtures)	Iteratively tree-shaped method. Each time the branches with larger magnitudes are kept, smaller signals are masked out. If the masked output contains more than one signal, iterative procedure is followed until all masked outputs consist of only a single speech signal.	Very complex iterative algorithm, but theoretically this method could separate an arbitrary number of speech signals of equal power with only two microphones, for both instantaneous mixtures and convolutive mixtures.

* W-disjoint orthogonal: the supports of the (windowed) Fourier transform of any two signals in the mixture are disjoint sets

In general, when solving the underdetermined blind source separation problem, it is important to use the sparsity of speech signals. When the sources are sparse, at a certain

time t, if one of the sources is significantly larger, the remaining ones are likely to be close to zero. In the extreme, if only one source $s_i, i \in (1,n)$ is nonzero, the observation x^t is proportional to the column vector $a_i, i \in (1,n)$ of the mixing matrix, and all the data points in the mixture space are aligned along this direction. Therefore, the density of data shows a clear tendency to cluster along the direction of vector a_i. The mixing matrix is estimated from the directions of maximum data density [Jad'08]. The disadvantage of this approach is that when the sources are not sufficiently sparse, it is difficult to estimate the mixing matrix precisely. With the known mixing matrix, the usual approach for underdetermined BSS is to find the solution through minimization of an *l-p* norm.

A method for blind separation of any number of sources using only two mixtures was proposed in [Jou'00]. It is an effective and easy algorithm when sources are W-disjoint orthogonal where the supports of the (windowed) Fourier transform of any two signals in the mixture are disjoint sets [Jou'00]. However the mathematical assumption is strong, which requires that the supports of the windowed Fourier transform of any two signals in the mixture be disjoint sets. Based on this assumption, the parameters of the attenuation and delay model are estimated. Finally, the sources are estimated from the W-disjoint orthogonal model. Later, this strong assumption was relaxed by applying a time-frequency (T-F) mask to the mixture when there is an overlap between the source signals [Ped'08].

A two-step algorithm was proposed by Bofill and Zibulevsky [Bol'01]. In this algorithm, if the observation is sparse, the mixing matrix is estimated using a potential-function-based method without knowing the number of sources beforehand, which is built in a two-dimensional mixture space and parameterized in polar coordinates [Bol'01]. The computational complexity of this algorithms is $O(T \times K)$. In their paper, the authors assume that the data outside some radius ($l_t > h$) has less relevance, with h being an adjustable threshold. Therefore, by discarding these data outside radius l_t, the computational cost can be significantly reduced. Given the mixing matrix, the source signals are estimated using the minimal *l-1* norm representation (shortest path separation

criterion). If the observation is not sparse in the time domain, this algorithm is still effective in a sparser linear transformed domain [Bol'00]. [Bol'01] proved that the shortest-path algorithm in [Bol'00] indeed solves the maximum-likelihood conditions in the second step. Since the methods in [Bol'00][Bol'01] only solved the BSS problem with instantaneous model, [Bol'01b] proposed a new method which has good performance for observation signals modelled as attenuation and delay mixtures. The procedure includes three stages: first, estimate the attenuation matrix by clustering of the magnitudes of input mixtures; next, estimate the delay matrix by clustering maximization of the real and imaginary parts of inputs mixtures; finally, a posterior approach to the source decomposition is performed under the assumption of a Laplacian magnitudes of the input signals [[Bol'01b]].

Further, [Li'03] [Li'06] estimated the mixing matrix using a gradient type algorithm or a K-means clustering algorithm. The column vectors of the mixing matrix are the cluster centers of normalized data vectors. If the sources are sufficiently sparse in the time domain (in that paper, the *l-1* norm is used as a sparseness measure), blind separation can be carried out directly in the time domain. Otherwise, a wavelet packets transformation pre-processing is necessary, and the blind separation is implemented in the time-frequency domain [Li'03]. The paper represented and compared the three parts of the algorithms using three examples. The first example concerns the recoverability of sparse sources, which means that the mixing matrix is known. In this example, it uses a linear programming algorithm to get the sparse solution with minimum *1*-norm, which is shown to be unique. When the sources do not overlap in the time-frequency domain, a high-quality reconstruction can be obtained. However, in the example there is an overlap between the source signals, and the sources are not sparse enough. The second example concerns the blind source separation when sources are not sufficiently sparse. In this method, the wavelet packet transformation is applied to make the mixtures appear as sparse and then it processes the data using the two-step method. The third example uses the wavelet packet pre-processing to obtain a sparse representation and it produces good performance [Li'03][Li'06].

With only two microphones, it is impossible to separate more than two signals because at most only one null direction can be placed for each output. An alternative way is to explore masking out extra sources and transferring the underdetermined blind source separation to an ordinary blind source separation problem. In [Ara'04a][Ara'04b][Ara'04c], the separation is performed by first removing *N-M* signals using some masking method, and ICA is then applied to separate the remaining *M* signals. Time-frequency masking is typically applied as a binary mask. In the binary mask method, each T-F unit is thus multiplied by one or zero. [Ara'04a] proposed the binary-masking method, which is mainly based on the clusters in the frequency domain and masking out the interfering signals. This method can transfer the underdetermined blind source separation problem to a determined blind source separation problem. However, the original masking approach is based on finding only one active source point, and masking all the other signals out. This is a rough process and binary masking suffers from zero-padding in the time domain, which causes discontinuities in speech signals. To overcome this problem, [Ara'04b][Ara'04c] utilize a directivity pattern-based continuous mask instead of a binary mask at the source removal stage. By allowing a small gain for the directions of sources from unwanted directions and specifying a large gain for other wanted directions, experimental results from [Ara'04b] show better performance than the original method in [Ara'04c].

In order to achieve a better acoustical perceptual result, [Vin'04] added a pre-processing step at the beginning and it processed the signal in an Equivalent Rectangular Bands (ERB) frequency scale. In this approach, the first step is to pass the mixture through a filter bank to form the auditory-motivated ERB scale. The second step is to perform a linear separation in each frequency bin. Here, the author assumes that the source signals are independent and follow a Gaussian prior with known variance. Finally, the waveforms of the estimated sources are built by summing all the frequency bands. Since the ERB frequency scale gives more importance to low frequencies, this method results in a better separation performance than the usual linear frequency scale.

2.5 Objective measurement for blind source separation in this thesis

The signal to noise ratio (SNR) is one of the most commonly used metrics for performance measurement, and it is often computed by calculating the ratio of a clean signal power over the power of the "error" signal or difference signal between the system output (i.e. target signal estimate) and the clean signal.

However, for the BSS system being considered, the output signals are scaled and may be at different levels than the original source signals. Therefore, for our experiments, it is more convenient to use an alternative metric which is the "SNR gain", i.e. the ratio (or difference in dB) of the SNR at the output of the BSS system over the SNR at the input of the BSS system, where each SNR is defined as the ratio of the power of the target speech component over the power of the other components (interferences and/or noise).

We also analyzed and compared other objective measurements, such as PESQ and CSII. PESQ is ideally designed for speech coding, and it may not always effectively show the performance of speech enhancement systems under all types of interference, for example speech interference or music interference. It was also found that CSII was quite sensitive to the environment i.e. to the type of interferer being used. Compared with others, the SNR gain produced a performance that more effectively matched our subjective impression. As a result, we selected the SNR gain for the objective measurement in this thesis.

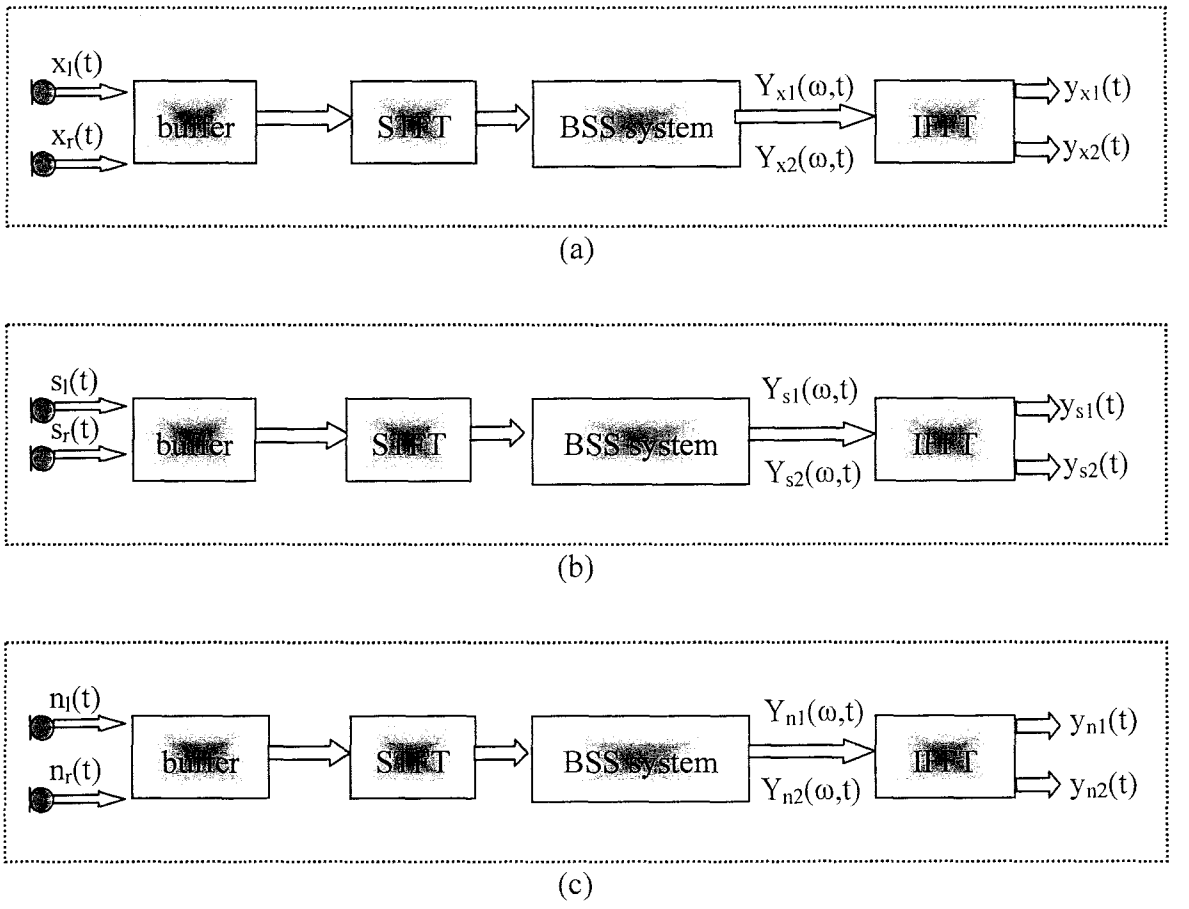

Figure 2.4.1: Diagram of system to calculate the SNR gain

In figure 2.4.1 (a), two speech mixture observations $x_l(t)$, $x_r(t)$ are processed by the BSS system. After BSS convergence, the two synthesized output signals $y_{x1}(t)$ and $y_{x2}(t)$ are assumed to respectively be scaled estimates of the two original source signals: the target speech $s(t)$ and interference $n(t)$ (e.g. speech, noise).

In simulated environments, it is possible to apply separately to the BSS system the sensor input signals received from the source $s(t)$ (i.e., $s_l(t), s_r(t)$ in figure 2.4.1 (b)) and the sensor input signals received from the source $n(t)$ (i.e. $n_l(t), n_r(t)$ in figure 2.4.1 (c)). We can thus write:

$$x_l(t) = s_l(t) + n_l(t) \qquad x_r(t) = s_r(t) + n_r(t). \tag{25}$$

The signal to noise ratio (or signal to interference ratio) at the input of the BSS system can be written as:

$$SNR_{in} = 10\log_{10}\frac{\sum s_l^{\,2}(t) + \sum s_r^{\,2}(t)}{\sum n_l^{\,2}(t) + \sum n_r^{\,2}(t)}\,(\text{dB}). \tag{26}$$

Note that with T-F masking, the BSS system is not linear, and therefore the principle of superposition does not apply for the output signals:

$$y_{x1}(t) \neq y_{s1}(t) + y_{n1}(t) \qquad y_{x2}(t) \neq y_{s2}(t) + y_{n2}(t). \tag{27}$$

Considering this, for the SNR ratio at the output of the system, we can build an "error signal" as $y_{s1}(t) - y_{x1}(t)$, and compute the following ratio:

$$SNR_{out} = 10\log_{10}\frac{\sum y_{s1}^{\,2}(t)}{\sum\left(y_{s1}(t) - y_{x1}(t)\right)^2}. \tag{28}$$

As mentioned earlier, the SNR gain is then simply the ratio of output and input SNRs:

$$G = \frac{SNR_{out}}{SNR_{in}} \tag{29}$$

$$G(dB) = SNR_{out}(dB) - SNR_{in}(dB). \tag{30}$$

Although the Array Gain can provide a good assessment of performance, it is not always representative of speech quality or intelligibility for speech and audio signals. Therefore, most of the outputs from the experiments in this thesis were also compared by informal listening, to get a more perceptual assessment of the performance. And from informal listening tests, we confirmed that the subjective impression roughly matched the objective scores.

Chapter 3 Blind source separation combined with T-F masking method

3.1 Fundamentals of T-F masking method

In real acoustic environments, there is often one dominant target speech. The target speech is usually louder than other sounds, and the other audio signals can be regarded as interfering speeches or noises. Most of the speech signal can be viewed as a sparse signal. From a signal processing view, when we transform the audio mixtures into the frequency domain, the time-frequency (T-F) units become sparser than those in the time domain, and each T-F unit may belong to either the target speech signal or the interference signals. Under this sparse assumption, theoretically, we can use an ideal binary mask $BM_{id}(f,t)$ to separate the target speech signal from the interference signal.

$$BM_{id}(f,t)=\begin{cases}1, & X(f,t)\subset T_a\\ 0, & X(f,t)\subset T_b\end{cases} \tag{31}$$

where T_a denotes a set for the target speech signal, and T_b denotes a set for the interference signal.

A common way to judge whether a T-F unit $X(f,t)$ belongs to the target speech T_a or to the interference speech T_b is to compare the amplitude of the target signal $S_i(f,t)$ and the amplitude of the interfering signal. In this case, the binary mask for the target source $S_i(f,t)$ can be rewritten as

$$BM_{id}(f,t)=\begin{cases}1, |S_i(f,t)|\geq|X(f,t)-S_i(f,t)|\\ 0, |S_i(f,t)|<|X(f,t)-S_i(f,t)|\end{cases} \tag{32}$$

In theory, each original source in the mixture can be obtained from T-F masking, but it requires that the T-F mask be complex-valued, and the quality of the output depends on the overall SNR. If the noise is stronger than the target, the ideal binary mask is very sparse and the output signal has large artificial noise.

However, in real applications, the target source $S_i(f,t)$ is unknown. Therefore, in many applications, an estimated target source was used instead of the real one to find the binary masks. Depending on the quality of the estimation of the target source, the binary masks which are obtained may vary considerably.

3.2 Review of T-F masking and ICA

In some audio environment where different speakers do not talk simultaneously, the speech mixtures gathered from the microphones are very sparse. As a result, an estimated target source is no longer needed for estimating the binary masks. [Rej'08] used the sparse observation instead of the target source to estimate the binary mask. At each time frequency bin, the author assumes that there is only one dominant signal.

$$BM(f,t) = \begin{cases} 1, & |X_i(f,t)| > |X_j(f,t)| \\ 0, & |X_i(f,t)| < |X_j(f,t)| \end{cases} \tag{33}$$

However, in real acoustical environment, there may be some simultaneous interference from other speakers or from the noisy environment. When the source signals are overlapping, it is hard to apply the strong assumption of sparsity. If we use the above method to estimate the binary mask, the performance will be unacceptable.

Blind source extraction by combining BSS and binary masking applied to an underdetermined 2-by-N mixture has been successfully applied to hearing aids. In [Ped'08], the binary masks are determined for each T-F unit by comparing the amplitudes of two spectrograms of input signal mixtures. This non-linear method is processed in an iterative way. Each time the binary masking is applied, some of the speech signals are removed. In order to increase the probability that all the sources are separated and that no source has been separated more than once, a merging step is followed to identify the binary masks that may contain the same source signal. If the masked output contains more than one signal, an iterative procedure is followed until all the masked outputs consist of only a single speech signal. In principle, this method can separate an arbitrary number of mixed speech signals. In [Ped'08], experimental results show the performance

of separating seven speech signals under instantaneous mixing conditions (i.e. non-convolutive mixing) [Ped'08]. The convolutive speech mixtures model is more applicable for speech signals because it takes early reflections and reverberation into account. The method proposed in paper [Ped'08] was found to have good performance both for instantaneous mixtures and convolutive mixtures.

In the separation step, the separation matrix (the ICA solution) tends to place the null towards sources spatially close to each other, so that each of the two outputs (y_1 and y_2) from the ICA solution represents a group of spatially close signals [Ped'08].

By comparing the amplitudes of the two spectrograms, the binary masks are determined for each T-F units:

$$\begin{aligned} BM_1'(\varpi,t) &= (|Y_1(\varpi,t)| > \tau |Y_2(\varpi,t)|) \,\&\, BM_a(\varpi,t) \\ BM_2'(\varpi,t) &= (|Y_2(\varpi,t)| > \tau |Y_1(\varpi,t)|) \,\&\, BM_a(\varpi,t) \end{aligned} \tag{34}$$

where τ is a parameter that controls how sparse the mask should be (how much of the interfering signals should be removed at each iteration), and $BM_a(\varpi,t)$ is a binary mask from the last iteration. As a result, it has an impact on the convergence speed. The complexity and delay (convergence speed) can be made adaptive through varying the parameter τ, with a corresponding change in the performance as well:

- $\tau = 1$: The two estimated masks together contain the same number of retained T-F units as the previous masks;
- $\tau > 1$: Fewer units retained than the previous binary mask. The combination of the two estimated masks is sparser. The convergence is faster at the expense of a sparser resulting mask.
- $0 < \tau < 1$: no convergence, case not to be considered.

The two estimated masks are applied in the T-F domain to the original signals, and four masked outputs are obtained:

$$\left.\begin{matrix} X_{1a}(\varpi,t) \\ X_{1b}(\varpi,t) \end{matrix}\right\} = BM_1(\varpi,t)\left\{\begin{matrix} X_1(\varpi,t) \\ X_2(\varpi,t) \end{matrix}\right.$$
$$\left.\begin{matrix} X_{2a}(\varpi,t) \\ X_{2b}(\varpi,t) \end{matrix}\right\} = BM_2(\varpi,t)\left\{\begin{matrix} X_1(\varpi,t) \\ X_2(\varpi,t) \end{matrix}\right. \quad (35)$$

The stopping criteria to evaluate the masked outputs are:

1) The masked signal is of poor quality: store for later use;

2) The masked signal consists of mainly one source signal: store as a candidate for a separated source;

3) The masked signal consists of more than one source signal: do further separation.

After the stopping criterion, the evaluation for each output signal is known. Then a merging stage follows. The merging decision is made by calculating a cross-correlation coefficient. In the merging stage:

1) Merging of the separated sources: if the cross-correlation coefficient between the output signals is larger than a threshold;

2) Merging of the low-quality source and the separated sources: if the cross-correlation coefficient between the output signal and the poor-quality signal is larger than a threshold;

After the merging stage, the binary masks are all updated and the signals are re-estimated from the updated binary masks. The remaining source signals are used to estimate the background mask.

$$BM_{background} = NOT(BM_{\hat{s}_1} \cup BM_{\hat{s}_2} \cup \cdots \cup BM_{\hat{s}_N}) \quad (36)$$

This mask is used to execute the main algorithm again. If the background mask has not changed, the separated signals are not changed any further and the algorithm stops.

With this method, the authors in [Ped'08] have used binaural instantaneous speech mixtures for simulations, and the method was found to iteratively separate the most significant speeches.

3.3 Adaptation of binaural blind source separation

For the application of any speech enhancement algorithm in a real-life hearing aid, it is important to figure out how fast an algorithm can adapt to the statistics of sources. The training of our considered BSS algorithm is essentially in "batch-mode" or on a block by block basis. It is not completely "offline", as the amount of data required for training is not in the order of minutes or hours, and the number of iterations required for convergence is not too large. The training cannot be considered truly "online" either, as the adaptation of the algorithm is not performed on a sample by sample basis. In this section we consider the adaptation or training to be used later for the proposed algorithm, and test it under both an anechoic environment and a reverberating room environment.

3.3.1 Anechoic environment

In many real-life situations, the audio sources may move from time to time, or listeners may need to switch from one target audio source to another one. As a result, our BSS speech enhancement module for hearing aids needs to quickly adapt the coefficients for the evolving environment. The adaptation time should be shortened as much as possible to make the listeners feel comfortable, and the parameter switching process should be smooth. Figure 3.1 shows the diagram of the adaptation process for the proposed blind source separation.

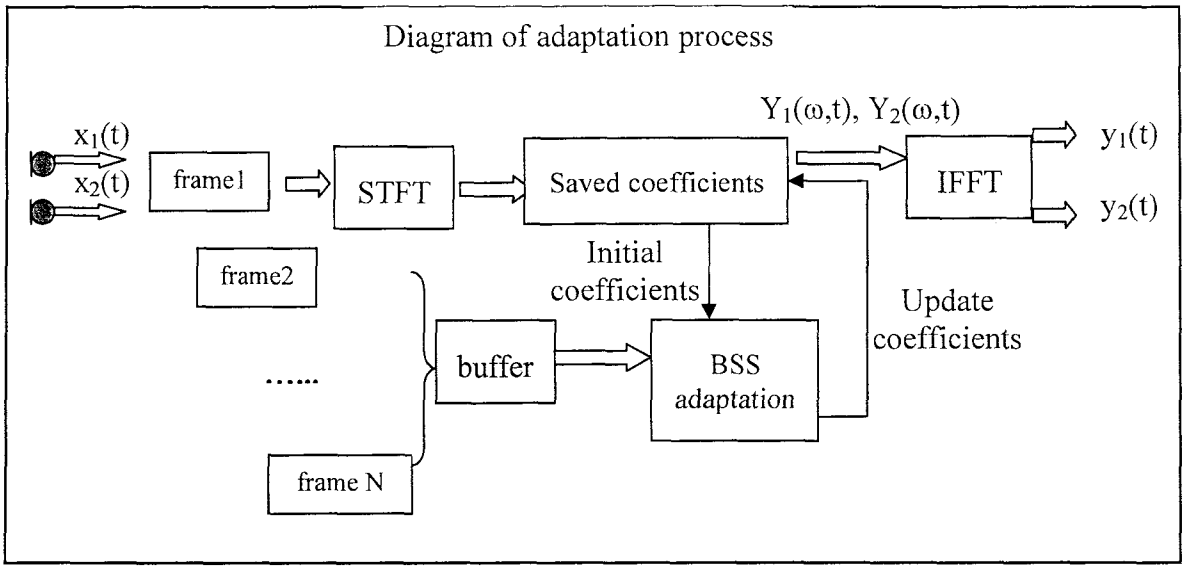

Figure 3.1: Diagram of processing

As illustrated in Figure 3.1, when the microphones in the hearing aid receive the input signals which were produced using experimental HRIR functions, the received signals are windowed and transferred into frequency domain using STFTs. Each frame is sent to the processing system, filtered using saved coefficients to produce the separated signal $Y_i(\omega,t), i = 1,2$. Finally, the signals are transferred and synthesized back to the time domain, continuously. At the same time, the BSS adaptation system is working in parallel. It buffers a period of signal recordings for adaptation, which contains several frames of buffered signal. Blind source separation adaptation is then conducted in the frequency domain, to separate the buffered speech mixtures. As there is similarity and correlation between adjacent frames of an audio recording, in this processing step the previous saved coefficients can be used as initial coefficients of processing for the current buffered data, which can speed up the convergence of the current buffered frames. Moreover, for the hearing aid wearer, the system adaptation will sound quick and smooth. In the whole system, the system time delay (or latency) is mainly caused by the first frame of data buffering and processing time. The adaptation delay, which stands for the learning ability of this adaptation system, mainly depends on the amount of buffering data used for BSS adaptation training.

3.3.2 Combined T-F masks with blind source separation for anechoic environments

For an anechoic environment, the speech mixtures are sparse compared with a reverberating room environment. As a result, for each frequency bin, there is likely a dominant speech signal in the mixtures. In this case, time-frequency masks can help to remove the interference speech while keeping the dominant speech signal untouched for each time-frequency bin. Moreover, the T-F masks do not require the two input signals to be very well separated. From this point of view, the system can solve the separation task without waiting for the outputs of the blind source separation to reach their optimum value. This only requires that the outputs of the blind source separation system be close to the optimum to yield correct masks.

For a binaural system, let $Y_1(\varpi,t)$ and $Y_2(\varpi,t)$ be the outputs of the blind source separation after N iterations. Then, for each time-frequency bin, it can be assumed that only one of the output signals is dominant. By comparing the magnitudes, two dominant speech signals $Y^{new}{}_1(\varpi,t)$, $Y^{new}{}_2(\varpi,t)$ are easily separated out by the T-F binary masks $BM_1(\varpi,t)$ and $BM_2(\varpi,t)$.

$$\begin{aligned} BM_1(\varpi,t) &= (|Y_1(\varpi,t)| > |Y_2(\varpi,t)|) \\ BM_2(\varpi,t) &= (|Y_2(\varpi,t)| > |Y_1(\varpi,t)|) \end{aligned} \tag{37}$$

$$\begin{aligned} Y^{new}{}_1(\varpi,t) &= BM_1(\varpi,t) \cdot Y_1(\varpi,t) \\ Y^{new}{}_2(\varpi,t) &= BM_2(\varpi,t) \cdot Y_2(\varpi,t) \end{aligned} \tag{38}$$

While the binary masks remove the interference and cross-talk, the discontinuities of the binary masks bring artificial noise to the outputs. As a result, some modifications are made to the masks and the equations of the outputs signals. The final outputs can be rewritten as

$$Y^{new}{}_1(\varpi,t) = BM_1(\varpi,t)\cdot Y_1(\varpi,t) + \lambda \cdot BM_2(\varpi,t)\cdot Y_1(\varpi,t)$$
$$Y^{new}{}_2(\varpi,t) = BM_2(\varpi,t)\cdot Y_2(\varpi,t) + \lambda \cdot BM_1(\varpi,t)\cdot Y_2(\varpi,t) \quad , \tag{39}$$

where λ is a small value scaling factor.

In the determined blind source separation case, these two outputs are the targets. In an underdetermined situation, further separation is needed for separating all the signals. However, in real-life environments, hearing aid users may not be interested in hearing all these speech signals. Only the dominant or target speech signal (or alternatively the signal from a given direction) may be the most interesting one.

Simulations and results

Simulation environment #1: we consider an anechoic environment, with two speakers from two different directions. The speech mixtures were produced using the measured head related impulse response (HRIR) from microphone recordings, by convolving the clean speech with HRIR #4 (from data provided by hearing aid manufacturer, corresponding to the first microphone at the left ear) and with HRIR #1 (from the first microphone at the right ear). The distance between the two microphones was 17.5 cm. The experiments were repeated using different speaker mixtures with different directions of arrival (DOA). Each frame of data contained 511 samples at a sampling rate of 20 KHz. For adaptation or training, 30 frames buffers (396 ms of recording) with half overlapping were used for this experiment, with 60 iterations and a step size of 0.0001 for each natural gradient ICA adaptation. The small scaling factor for T-F mask was $\lambda = 0.001$. The results compare the output performance of BSS with and without T-F masks.

Table 3.1: Female speech #9 from 60 degree direction, female speech #10from 355 degree direction

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: female speech #10, DOA:355 degree Interfering speech: female speech #9, DOA: 60 degree			
BSS with T-F masks	-2.86	8.98	11.85
BSS without T-F masks	-2.86	3.61	6.47
Target speech: female speech #9, DOA: 60 degree Interfering speech: female speech #10, DOA:355 degree			
BSS with T-F masks	2.86	13.11	10.25
BSS without T-F masks	2.86	6.02	3.15

Table 3.2: Female speech #10 from 80 degree direction, another male speech #13 from 270 degree direction

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #13, DOA:270 degree Interfering speech: female speech #10, DOA 80 degree			
BSS with T-F masks	-0.74	13.99	14.74
BSS without T-F masks	-0.744	6.75	7.49
Target speech: female speech #10, DOA: 80 degree Interfering speech: male speech #13, DOA 270 degree			
BSS with T-F masks	0.74	15.00	14.25
BSS without T-F masks	0.74	6.8	6.14

Table 3.3: Female speech #9 from 10 degree direction, male speech #5 from 350 degree direction

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:350 degree Interfering speech: female speech #9, DOA 10 degree			
BSS with T-F masks	4.18	10.52	6.33
BSS without T-F masks	4.18	5.43	1.24
Target speech: female speech #9, DOA: 10 degree Interfering speech: male speech #5, DOA 350 degree			
BSS with T-F masks	-4.18	3.58	7.77
BSS without T-F masks	-4.18	-3.39	0.78

From Tables 3.1, 3.2 and 3.3, we can see that the blind source separation improves the output performance, especially when the T-F masks are applied. The SNR gains for BSS with T-F masks are much larger than BSS without T-F masks. By listening to the output speeches, we also can confirm this conclusion.

3.3.3 Proposed combined T-F masks with blind source separation for real acoustic environments with reverberation

In real acoustic environments, no matter what the size of a room is, there is some amount of reverberation. To test the performance of our blind source separation processor, recordings from a real acoustic environment (provided by a hearing aid manufacturer) were used for simulations.

As the reverberation is made of delayed and scaled versions of the original speech signal, there are more sources in the mixtures and it brings more difficulties to the blind source separation system. As a result, more signal samples are needed for adaptation training. However, in real situations, the adaptation time should be as short as possible. Therefore, considering both requirements, interpolation was used as a mean to increase the training data. Figure 3.2 shows the diagram of the adaptation process for a room environment with reverberation.

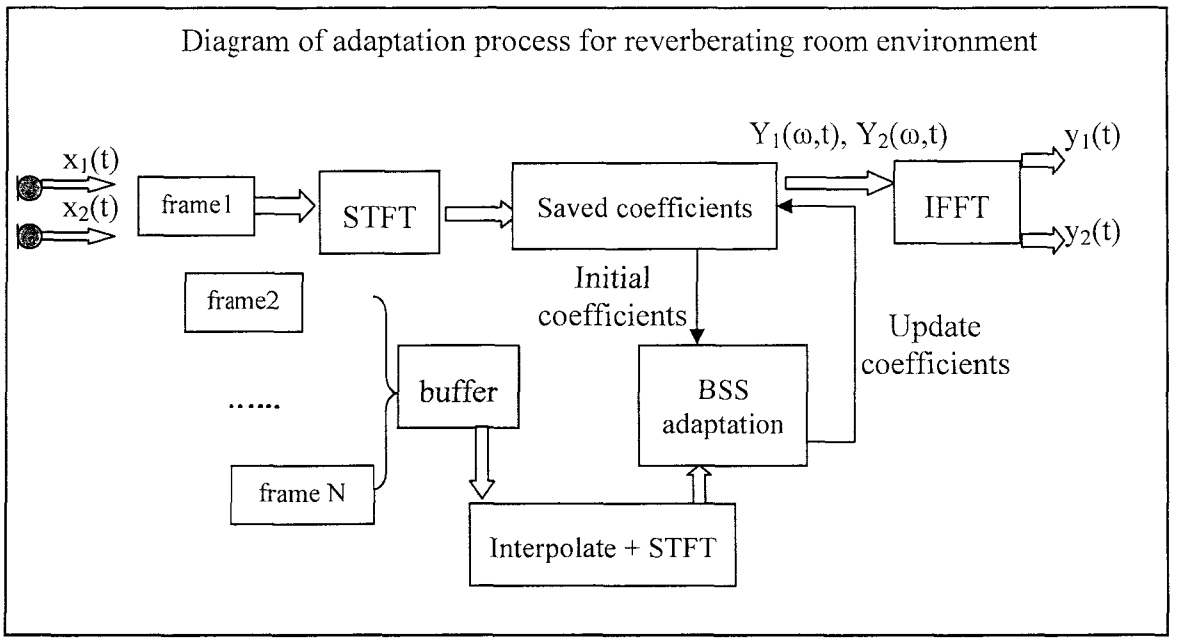

Figure 3.2: Diagram of adaptation processing with interpolation

Very similar to the adaptation process illustrated in figure 3.1, the only difference of this adaptation process system for reverberating room environment in figure 3.2 is in its BSS adaptation part. The processing system buffers a period of signals received from the microphones in the hearing aid device, which usually can be divided into several overlapped signal frames. This period of buffered frames is only used for adaptation training. In the adaptation system, the buffered recordings are firstly interpolated into more samples (*N* times linear interpolation, between consecutive time domain frames, before windowing). Each frame is transferred into the frequency domain using a STFT. After that, blind source separation training is conducted in the frequency domain. The blind source separation coefficients are saved as initial coefficients to process the next buffered data, to speed up the convergence of the next buffered data. Finally, the output speeches are transferred back into the time domain, continuously. This is very similar to the previous adaptation process for anechoic environments.

Simulations and results

Simulation environment #2: anechoic room recordings with two speakers from two different directions, containing more reverberation than the synthetic mixtures generated

by HRIR filtering in the previous simulations. The speech mixtures were measured by binaural microphones with one microphone on the left ear (microphone #4) and one on the right ear (microphone #1), and the distance between the two microphones was 17.5 cm. The experiments were repeated using different speaker mixtures with different directions of arrival (DOA). The recordings were sampled at a sampling rate of 20 KHz and windowed by Hann window of length 511. For adaptation or training, 120 frames of signals which equals to 1.546 seconds of recordings were interpolated by 3. This means that 360 frames of training data with half overlapping were available for adaptation and training. In each ICA adaptation process, 60 iterations with step size of 0.0001 in the natural gradient method were used and the small scaling value for the T-F mask was $\lambda = 0.001$. The results below compare the output performance of BSS with and without T-F masks.

Table 3.4: Female speech #9 from 0 degree direction, male speech #5 from 60 degree direction

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:60 degree Interfering speech: female speech #9, DOA: 0 degree			
BSS with T-F masks	-3.87	3.87	7.75
BSS without T-F masks	-3.87	0.73	4.60
Target speech: female speech #9, DOA: 0 degree Interfering speech: male speech #5, DOA:60 degree			
BSS with T-F masks	3.87	8.20	4.32
BSS without T-F masks	3.87	3.66	-0.20

Table 3.5: Female speech #9 from 30 degree direction, male speech #5 from 300 degree direction

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:300 degree Interfering speech: female speech #9, DOA 30 degree			
BSS with T-F masks	-4.79	10.57	15.36
BSS without T-F masks	-4.79	0.18	4.97
Target speech: female speech #9, DOA: 30 degree Interfering speech: male speech #5, DOA 300 degree			
BSS with T-F masks	4.79	13.82	9.02
BSS without T-F masks	4.79	8.68	3.89

Simulation environment #3: consider a living room (reverberating time 0.30 - 0.45 second) with two speakers from two different directions. The speech mixtures were measured by binaural microphones with one microphone on the left (microphone #4 from the data provided by the hearing aid manufacturer) and one on the right (microphone #1 from the provided data), and the distance between the two microphones was 17.5 cm. The experiments were repeated using different speaker mixtures with different directions of arrival (DOA). The following results compare the output performance of BSS with and without T-F masks.

Table 3.6: Female speech #9 from 30 degree direction, male speech #5 from 300 degree direction, living room environment with reverberating time of 0.30 - 0.45 second.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:300 degree Interfering speech: female speech #9, DOA 30 degree			
BSS with T-F masks	-7.04	2.33	9.38
BSS without T-F masks	-7.04	-3.57	3.46
Target speech: female speech #9, DOA: 30 degree Interfering speech: male speech #5, DOA 300 degree			
BSS with T-F masks	7.04	13.89	6.85
BSS without T-F masks	7.04	10.92	3.88

Simulation environment #4: Consider the room as the Cafeteria of the University of Oldenburg (0.75-2.25 sec reverb time), which is larger in room size and has a longer reverberating time compared with the living room case in simulation environment #3. The experiment settings are very similar to the experiments for simulation environment #3, with the only exception that more data is used to improve the overall adaptation performance. Here, 150 frames of signals corresponding to 1.929 seconds of recording are interpolated by 3. This means that 450 frames of training data with half overlapping are available for adaptation training. The results below compare the output performance of BSS with and without T-F masks.

Table 3.7: Female speech #10 from 90 degree direction, male speech #13 from 240 degree direction, Cafeteria of the University of Oldenburg with reverberating time of 0.75-2.25 seconds.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #13, DOA:240 degree Interfering speech: female speech #10, DOA 90 degree			
BSS with T-F masks	1.33	8.24	6.90
BSS without T-F masks	1.33	5.86	4.52
Target speech: female speech #10, DOA: 900 degree Interfering speech: male speech #13, DOA 240 degree			
BSS with T-F masks	-1.33	4.60	5.94
BSS without T-F masks	-1.33	2.78	4.12

From all the experiments in simulation environments #1, #2, #3 and #4, we can conclude that:

- The T-F masks always improve the performance;
- Typically the performance of BSS is better for lower reverberation environments, but BSS provides a SNR gain under all the considered environments.

As a result, for scenarios with a larger amount of reverberation, combining BSS + T-F masking with some other spatial information processing may be helpful or even required, to reduce some reverberation effects and improve the blind source separation performance. Combining blind source separation with beamforming will be introduced in Chapter 4, in the context of underdetermined BSS systems.

3.3.4 Conclusion

Blind source separation is an efficient method to separate two speech signals without knowing a priori knowledge about the speech mixtures. However, classical blind source separation needs a significant amount of training data to obtain a good convergence, and it may result in a significant adaptation delay in a real system. As a result, the adaptation

problem of classical blind source separation becomes an issue of reducing the training data while keeping a similar separation performance. From the previous sections, it can be seen that the T-F masking method can be used as a quick way to separate two signals which are disjoint in the frequency domain, and it does not need additional training data. However, in most cases, the speech mixtures can be sparse but not completely disjoint in the frequency domain. As a result, a pre-processing step is needed to make the mixtures sparser. In [Ped'08], the author uses blind source separation as a pre-processing method to estimate the T-F masks. In this chapter, we proposed using the T-F masking method as a post-processing step to further separate the outputs from blind source separation. Thus in the blind source separation step, we do not need as much training data for the ICA to get a good convergence. In our experiments, we selected a small value of step size for the adaptation process to make sure that the convergence procedure will reach a low value of the cost function after convergence. Then further separation work was handled by the T-F masking method.

In a real room environment, some reverberation exists almost everywhere, and the reflections from talkers will not be independent. Since the BSS/ICA is built based on the independence assumption, the reverberation will limit the performance of BSS/ICA. So in the next chapter, we will introduce a method of combining spatial information with BSS to reduce the reverberation effect and improve the performance of BSS/ICA .

Chapter 4 Combining spatial information with Blind Source Separation

This chapter will show how spatial processing such as beamforming can be used as a pre-processor to convert an underdetermined BSS problem into a determined one. The beamformers make use of Direction of Arrival (DOA) estimates, which are first described in the following section.

4.1 DOA estimator

One of the reasons why the human auditory system is much better than any other hearing equipment is that human ears can detect the direction of incoming audio sources and track the movement of the sources. During this process, the human ears play the role of DOA detector and audio source locator. In order to use a beamformer to mimic the human ears, DOA estimation is the first step required to obtain the geometry information. Two popular methods of DOA estimation are: time delay estimation [Roh'08] [Vis'01] [Kna'76] and directivity pattern estimation from an unmixing matrix [Pan'07] [Saw'04].

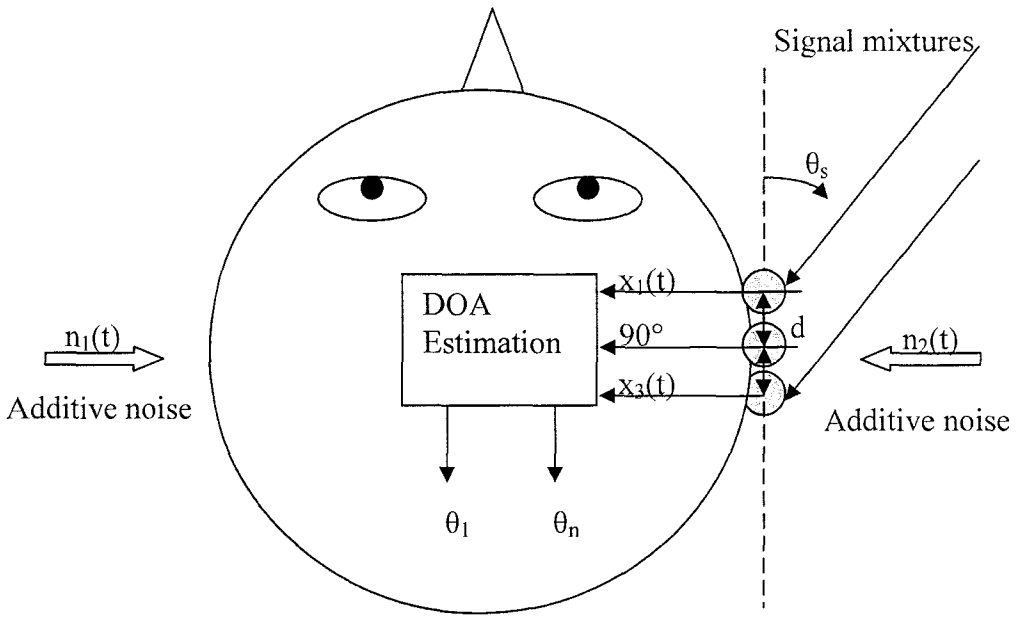

Figure 4.1: Block diagram of a monaural DOA estimator

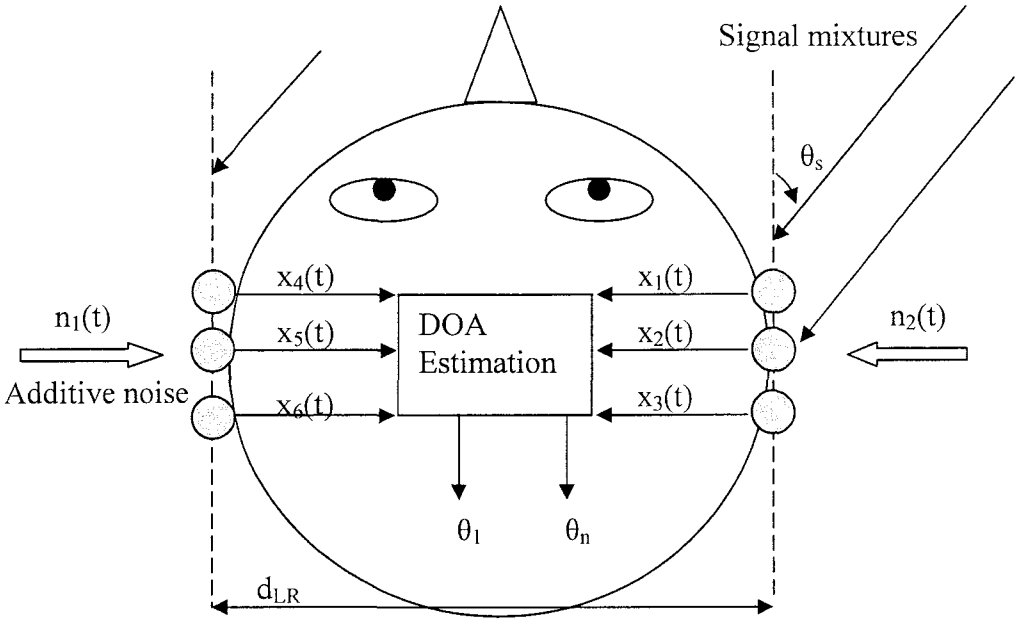

Figure 4.2: Block diagram of a binaural DOA estimator

4.1.1 DOA estimation using a time delay method

To benefit from using beamforming in hearing aids, we need to determine the direction of the target source and occasionally also the direction of the interference signals. The Generalized Cross-Correlation (GCC) method is one of the existing methods for estimating direction of arrivals [Kna'76]. By neglecting the effect of room reverberation, reflections, and the frequency dependent head shadow effect, a remote source signal $s(t)$ received by M spatially separated sensors in the presence of additive noise can be mathematically modeled as

$$x_i(t) = \alpha_i s(t-\tau_i) + n_i(t), i = 1, \cdots M \tag{40}$$

Without loss of generality, $x_i(t), i = 1 \cdots M$ is the observation from the i^{th} microphone in hearing aids, $n_i(t), i = 1,2$ is the additive noise signal received by the i^{th} sensor, and α_i is transmission attenuation (or a gain) from the source signal to the i^{th} microphone.

For an endfire microphone array (Figure 4.1), the time delay between signals from microphones on the same side of the ear is $\tau_i = \frac{(i-1)d}{c}\cos\theta_s$. Then, the observations from the M microphones are

$$x_i(t) = \alpha_i s(t - \frac{(i-1)d}{c}\cos\theta_s) + n_i(t), i = 1, \cdots M \quad . \tag{41}$$

For the binaural microphone array in Figure 4.2, the time delay between the microphones on the left ear and the microphones on the right ear is $\tau = \frac{d_{LR}}{c}\sin\theta_s$. For microphones located on the left side, the observations from the binaural microphone array in Figure 4.2 are then:

$$x_i(t) = \alpha_i s(t - \frac{(i-1)d}{c}\cos\theta_s - \frac{d_{LR}}{c}\sin\theta_s) + n_i(t), i = 1, \cdots M \quad . \tag{42}$$

The direction of arrival (DOA) is estimated through the time delay τ_d between two signals from the sensors. The time delay is obtained by finding the time-lag that maximizes the cross correlation between the two signals, which is written as:

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t-\tau)] \, . \tag{43}$$

The parameter τ_d is estimated from the time delay τ where the maximum value of the cross-correlation is achieved:

$$\tau_d = \arg(\max\left(R_{x_1x_2}(\tau)\right)) \quad . \tag{44}$$

Estimating the time delay and obtaining the DOA estimation through the basic cross-correlation method is easy and effective for clean speech environments. However, if the speech signals are corrupted by noise, the performance of the cross-correlation method deteriorates quickly. To improve the peak detection, some pre-filters or weighting functions are used in the cross-correlation estimation to produce sharp peak in $R_{x_1x_2}(\tau)$.

There are many popular weighting functions, such as the Roth Processor, the Smoothed COherence Transform (SCOT), the PHAse Transform (PHAT), the Eckart Filter and the Maximum Likelihood (ML) method [Kna'76][Gao'06] [Fou'08]. The SCOT method can be interpreted as pre-whitening before the cross-correlation (i.e. normalizing the level of each frequency), which is written as:

$$R_{x_1x_2}(k)=\frac{1}{L_{DFT}}\sum_{m=0}^{L_{DFT}-1}\frac{\hat{\Phi}_{x_1x_2}(m)}{\sqrt{\hat{\Phi}_{x_1x_1}(m)\hat{\Phi}_{x_2x_2}(m)}}e^{j\frac{2\pi}{M}mk},k=1\cdots L_{DFT}-1 \tag{45}$$

where $\hat{\Phi}_{x_1x_1}(m)$ and $\hat{\Phi}_{x_2x_2}(m)$ are auto spectra (auto-PSD estimate), and $\hat{\Phi}_{x_1x_2}(m)$ is a cross-spectrum (cross-PSD estimate), L_{DFT} is the DFT length.

The SCOT method is developed based on the ROTH method. In the ROTH method, the cross spectrum is normalized by the auto spectrum of one of the signals. In the PHAT method, the cross spectrum is normalized by the magnitude of the cross spectrum. The PHAT method is known for its ability to avoid causing spreading of peaks in the correlation function. It can be written as

$$R_{x_1x_2}(k)=\frac{1}{L_{DFT}}\sum_{m=0}^{L_{DFT}-1}\frac{\hat{\Phi}_{x_1x_2}(m)}{\left|\hat{\Phi}_{x_1x_2}(m)\right|}e^{j\frac{2\pi}{M}mk},k=1\cdots L_{DFT}-1 \tag{46}$$

Under ideal conditions (i.e. no noise or spatially uncorrelated noise, with perfect estimate of $\hat{\Phi}_{x_1x_2}(m)$), we would get $\frac{\hat{\Phi}_{x_1x_2}(m)}{\left|\hat{\Phi}_{x_1x_2}(m)\right|}=e^{j\phi(f)}=e^{j2\pi f\tau_d}$. Only the phase information is preserved in this method. In the time domain, it leads to: $R^{ideal}{}_{x_1x_2}(\tau)=\delta(t-\tau_d)$. However, in practice because the noise signals may not be totally jointly uncorrelated ($\Phi_{n_1n_2}(m)\neq 0$) and the estimation of the cross power spectral density is not perfect ($\hat{\Phi}_{x_1x_2}(m)\neq\Phi_{x_1x_2}(m)$), the cross correlation in the time domain is not a delta function.

As head shadow effects exist in hearing aid applications, it is better to take into account the diffraction of a harmonic plane wave on a sphere. Knowing the interaural time difference (ITD), through the relation of ITD and the DOA θ, it is possible to approximate the θ using the equation from [Kuh'77]:

$$ITD = a \cdot d_{LR} / c \cdot \sin\theta \tag{47}$$

where the coefficient a is a variable which is frequency independent below approximately 500Hz or above approximately 3kHz, d_{LR} is the distance between left and right microphones placed at both side of head and c is the speech of sound. According to [Kuh'77], for low frequencies (below 500Hz), the coefficient a is around 1.5. This means that the effective head radius at low frequencies is approximately 9.3 cm, which is larger than the parameter used for the recordings that were provided, which is 8.75cm. This larger radius may be due to the fact that the sound wave must travel over the protruding pinnae and nose to reach the ear canal [Kuh'77]. For high frequencies (above 200kHz), waves that crept around the head one or more times have been attenuated sufficiently to be neglected relative to the direct waves reaching the ear [Kuh'77]. As a result, the coefficient a is equal to 1. For the mid-frequencies (between 500Hz to 2kHz), there is a continuous and smooth transition from 1 to 1.5, but the mid-frequencies yield poor estimation of ITD. In a reverberating room, the high frequencies of direct sound become a significant fraction of the total sound in the sound mixture. In that case, it is better to use a coefficient a equals to 1.

Simulations and results:

The first experiment simulates an anechoic audio environment. Clean speeches were convolved with head related impulse responses (HRIR) measured from a KEMAR model [Gar'94] to generate the source speeches from specific directions of arrival. The cross correlation based time delay estimator analyzes the time cross correlation between the speech mixtures. The time delay estimate is obtained from the cross correlation function.

From equation $\tau_d = \arg(\max(R_{x_1x_2}(\tau)))$ we can see that the time resolution of the time delay τ_d is related to the sampling rate of the observation sequences $x_1(t)$ and $x_2(t)$: $\Delta t = \frac{1}{f_s}$. For a sampling rate of 20 kHz, the time resolution is thus $\Delta t = 5 \times 10^{-5}$. For binaural hearing aids with one microphone on each side, the distance between the two microphones is 17.5 cm, and the maximum time delay between the two microphones is $\tau_d = \frac{d_{LR}}{c}\sin(\theta_s) = \frac{d_{LR}}{c}\sin(90^\circ) = \frac{0.175}{340} = 5.147 \times 10^{-4}$ second.

For reference, table 4.1 lists the time delay for each DOA angle θ_s for the binaural hearing aid with one microphone on each side.

Table 4.1: The time delay τ_d for each DOA angle θ_s from the cross correlation function $\tau_d = \frac{d_{LR}}{c}\sin(\theta_s)$ for binaural hearing aids with one microphone on each side and $d_{LR} = 0.175$ cm,

θ	τ_d	θ	τ_d
5° or 175°	0.44e-4	10° or 170°	0.89e-4
15° or 165°	1.33e-4	20° or 160°	1.76e-4
25° or 155°	2.17e-4	30° or 150°	2.57e-4
35° or 145°	2.95e-4	40° or 140°	3.30e-4
45° or 135°	3.63e-4	50° or 130°	3.94e-4
55° or 125°	4.21e-4	60° or 120°	4.45e-4
65° or 115°	4.66e-4	70° or 110°	4.83e-4
75° or 105°	4.97e-4	80° or 100°	5.06e-4
85° or 95°	5.12e-4	90°	5.14e-4
0° or 180°	0		

The sampling rate of the signals from the recordings provided by Siemens is 48 kHz, and the recordings (and HRTFs) were downsampled to 20 kHz to match the operating sampling rate of the hearing aids. The discrete time cross power spectral density functions were estimated using Welch's averaged, modified periodogram method of spectral estimation.

Simulation and experiment # 1

One speaker in a real anechoic room is coming from 120 degrees. The distance between the speaker and the listener is 150 cm. The ITD for this scenario is estimated and compared using the generic cross-correlation method (Figure 4.3), the SCOT method (Figure 4.4), and the PHAT method (Figure 4.5).

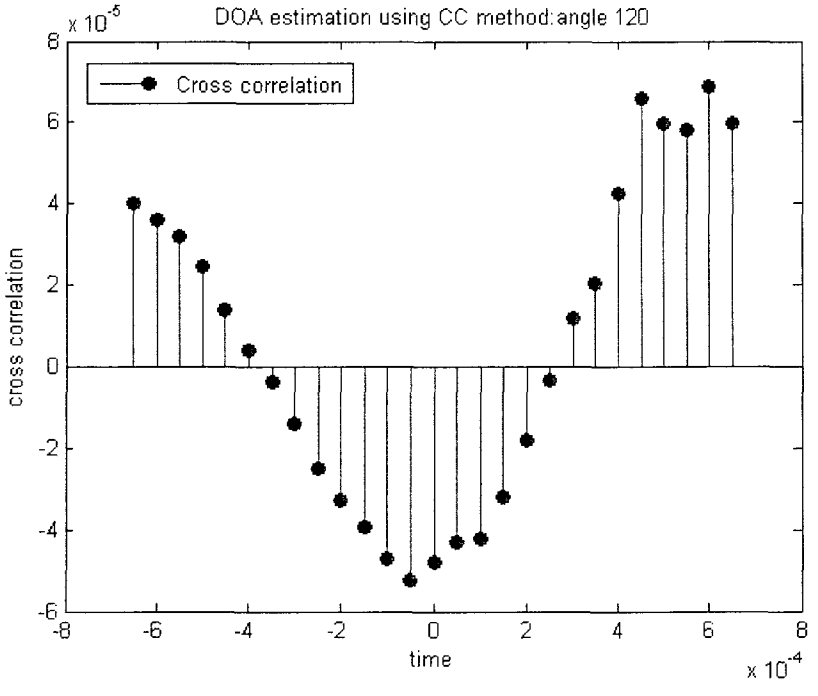

Figure 4.3: DOA estimator using the cross correlation method with a target source coming from 120 degrees. The maximum ITD from the cross correlation function is at 6.00e-004 second, which is larger than the expected ITD 4.45e-004 second in Table 4.1.

From Figure 4.3, we can see that the peak of the cross correlation is hard to discriminate. There are two peaks which appear very closely around the expected ITD (4.45e-004 second). The largest peak is 6.00e-004 second, which is larger than the maximum ITD (5.14e-004 second).

In order to improve the peak detection, generalized cross-correlation methods are conducted to repeat the ITD estimation experiment.

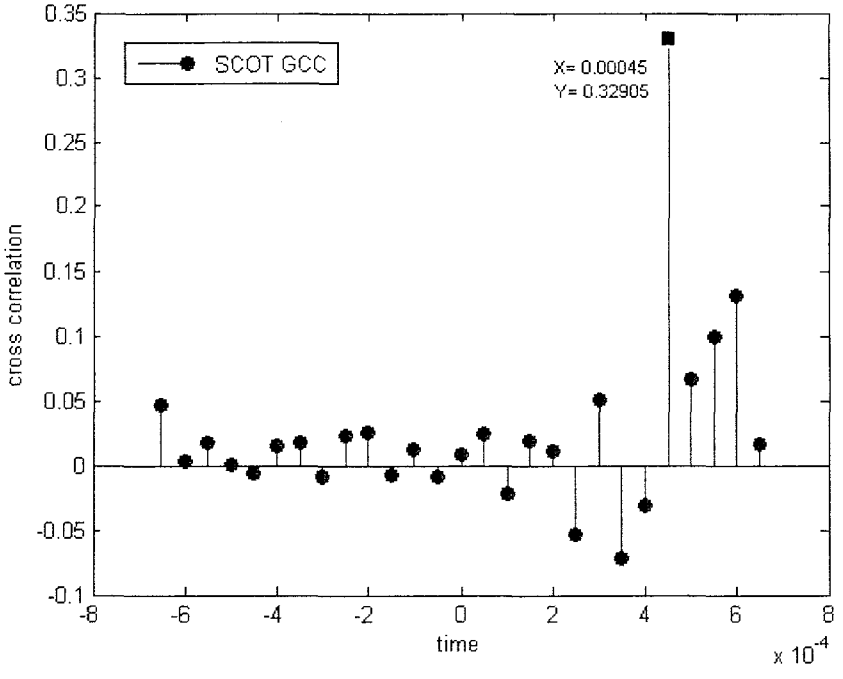

Figure 4.4: DOA estimator using the SCOT generalized cross correlation method with a target source coming from 120 degrees. The ITD from the cross correlation function is 4.5e-004.

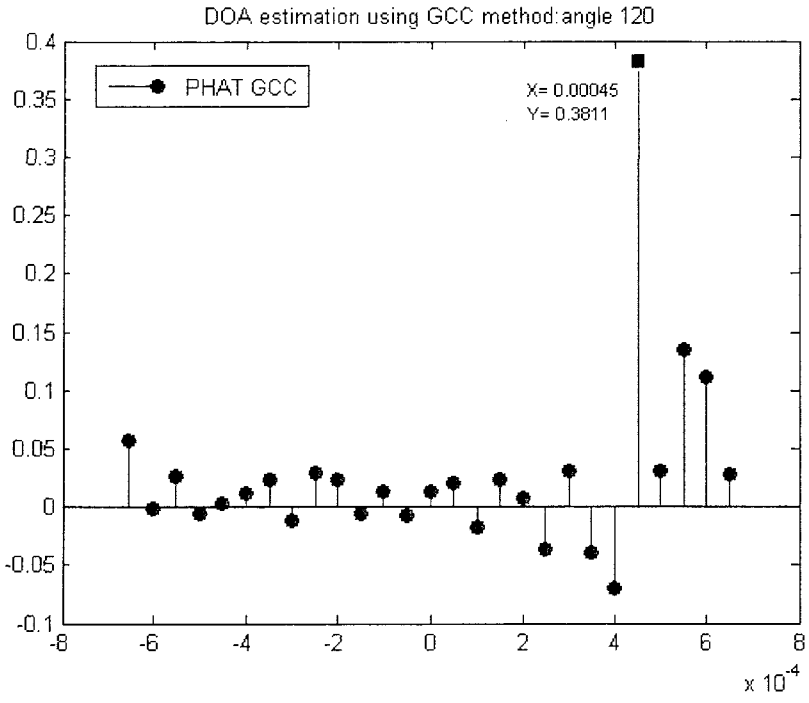

Figure 4.5: DOA estimator using the PHAT generalized cross correlation method with a target source coming from 120 degrees. The ITD from the cross correlation function is 4.5e-004.

After generalizing the cross-correlation using the SCOT and PHAT methods, the peak is much sharper than for the cross-correlation method. From Figure 4.4 and 4.5, we can see that the largest peaks give the ITD estimation of 4.5e-004, which yield the DOA estimation of 60.96 or equivalently 119.03 degrees because of the front back symmetry. This estimation value of 119.03 degrees is very close to the real input signal of 120 degrees.

Simulation and experiment # 2

In most cases, real acoustic rooms are not anechoic, and reverberation exists to some extent. In our second experiment, we test the ITD estimation method in a room with more reverberation.

The room we used for testing is a large cafeteria room from the University of Oldenburg. The reverberating time of this room is about 0.75 - 2.25 seconds (frequency dependent). The source speaker is located at 30 degrees and is 150 cm away from the listener.

The ITD for this scenario is estimated using different estimation methods, and the experimental results are shown in Figure 4.6, Figure 4.7 and Figure 4.8.

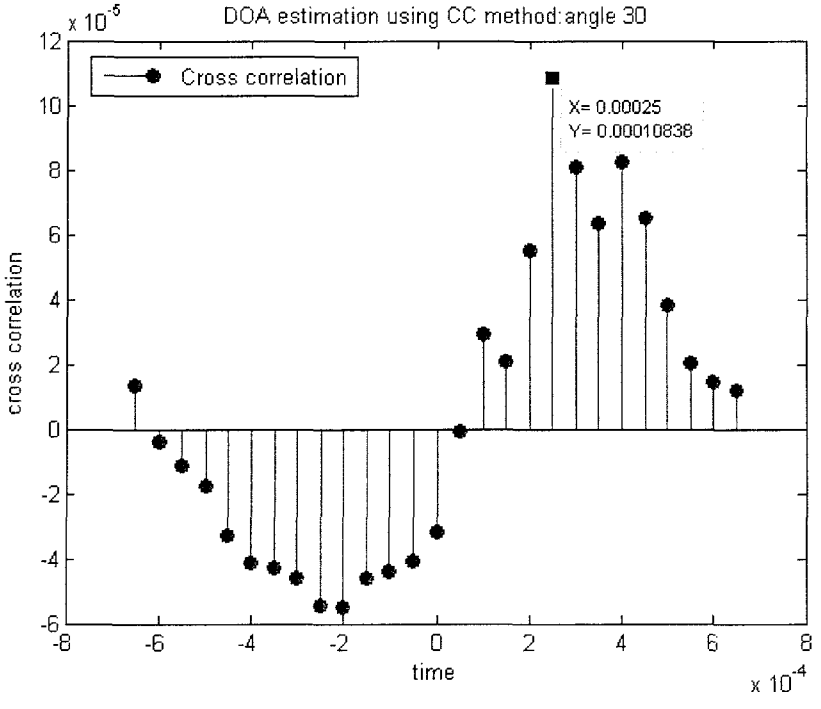

Figure 4.6: DOA estimator using the cross correlation method with a target source coming from 30°. The maximum ITD from the cross correlation function is 2.5e-004, and the estimated DOA is 29.05°.

In Figure 4.6, the peak is clear in this case, the ITD from the cross-correlation method is 2.5e-004, and the estimated DOA is 29.05 degrees.

In Figure 4.7 and 4.8, we can see that the peak is sharper and clearer than in Figure 4.6, after the cross-correlation function is normalized. In these two figures, it is very clear that the estimated ITDs are 2.50e-004, and the estimated DOA is 29.05 degrees. Compared with the target DOA of 30 degrees, these cross-correlation based ITD estimation methods

provide good estimation in this simulation, and these methods are robust in reverberating room environments.

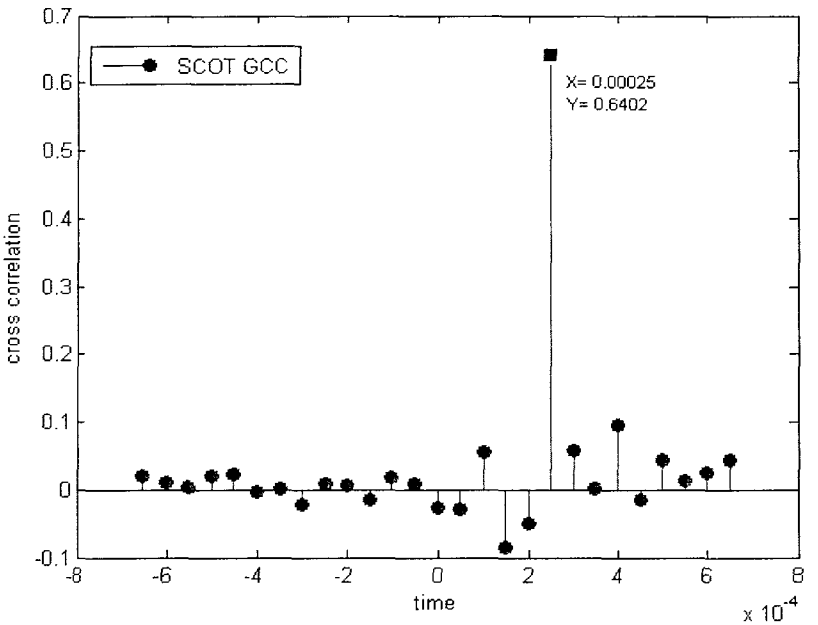

Figure 4.7: DOA estimator using the SCOT generalized cross correlation method with a target source coming from 30°. The maximum ITD from the cross correlation function is 2.5e-004, and the estimated DOA is 29.05°.

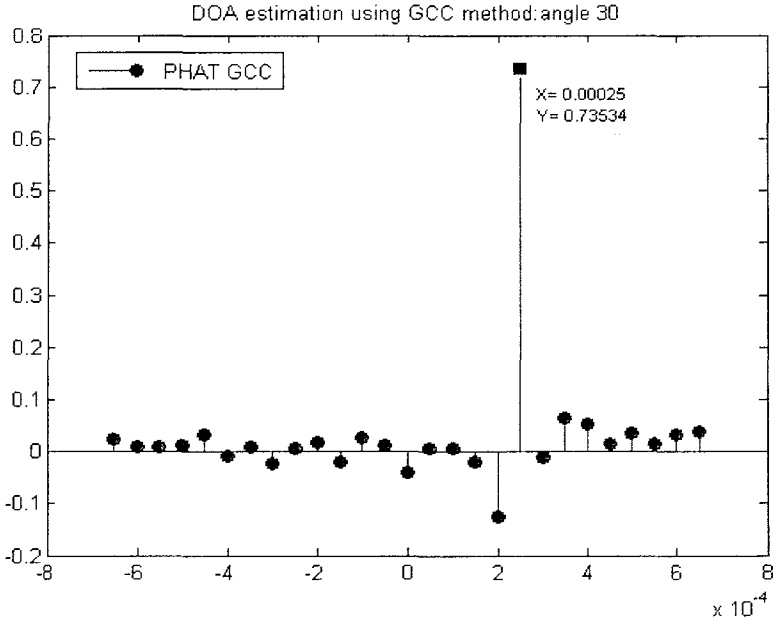

Figure 4.8: DOA estimator using the PHAT generalized cross correlation method with a target source coming from 30°. The maximum ITD from the cross correlation function is 2.5e-004, and the estimated DOA is 29.05°.

Experiment environment # 3

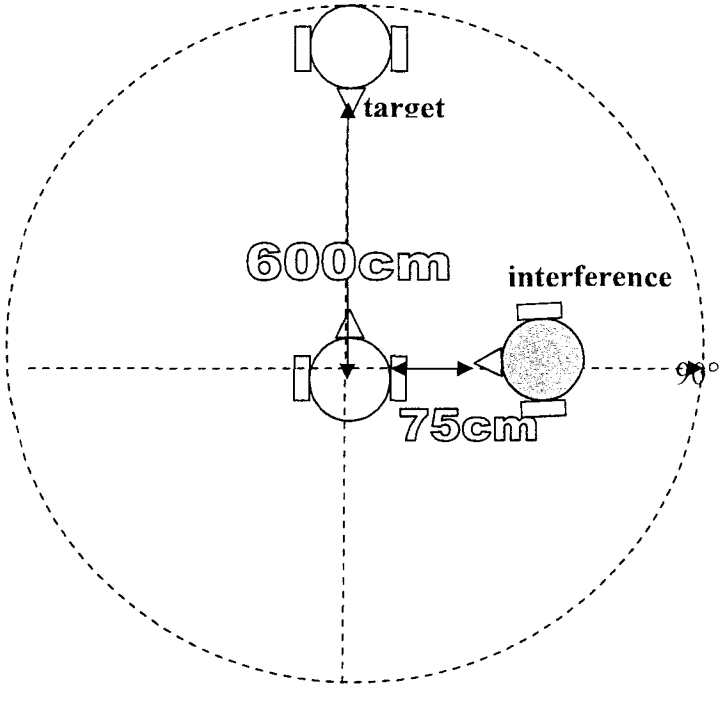

Figure 4.9: The acoustic environment of the cafeteria for experiment #3

Often, there is more than one sound source in a large room such as a cafeteria, and the interference sound signals can be directional interference or babble noise. In the following experiment, we use a female speaker at 0 degree and 600 cm away from the listener in a large cafeteria room from the University of Oldenburg, with a reverberating time of about 0.75 - 2.25 seconds (frequency dependent). At the same time, a male speaker is talking sitting at the listener's right side (90 degree of DOA), at 75 cm distance. As shown in Figure 4.9, in experiment #3 the target is the female speaker at the front.

The SNR calculated at the fist microphone on the left hearing aid is -3.27 dB, and the SNR at the first microphone on the right hearing aid is -4.73 dB. The ITD for this scenario is estimated using different estimation methods, and the experimental results are shown in Figure 4.10, Figure 4.11 and Figure 4.12.

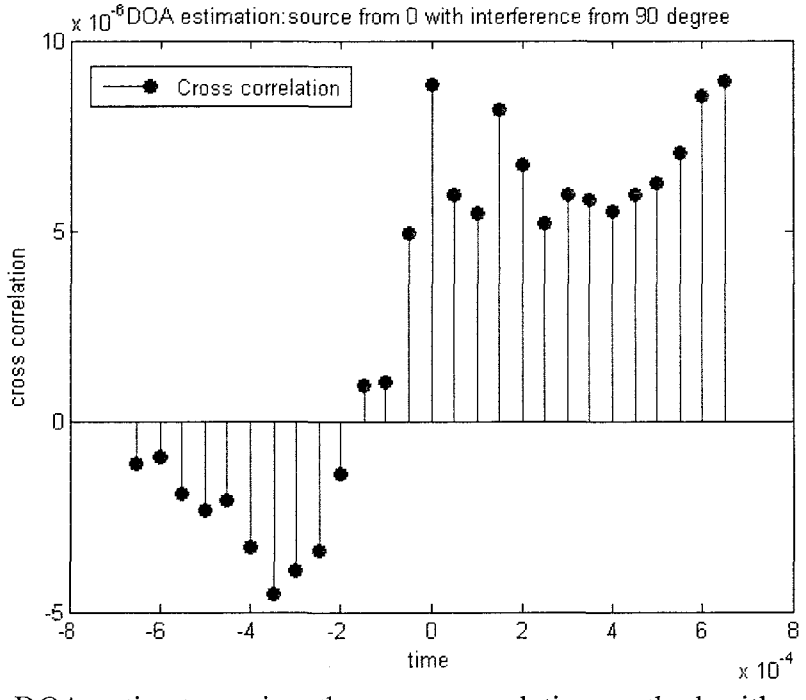

Figure 4.10: DOA estimator using the cross correlation method with a target source coming from 0° and a directional interference from 90°. The maximum ITD from the cross correlation function is 2.5e-004, and the estimated DOA is 29.0593°.

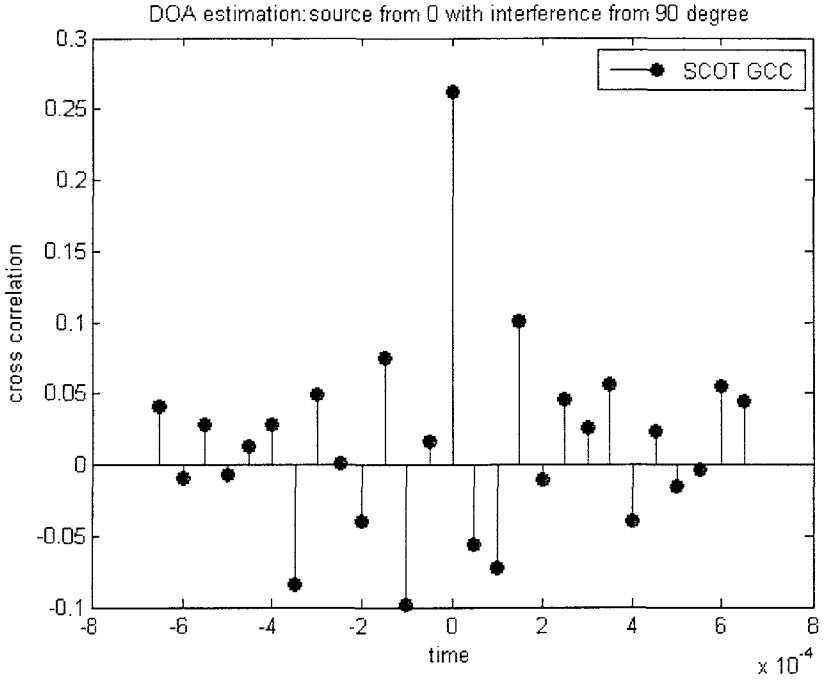

Figure 4.11: DOA estimator using the SCOT generalized cross correlation with a target source coming from 0° and a directional interference from 90°. The maximum ITD from the cross correlation function is 0, and the estimated DOA is 0°.

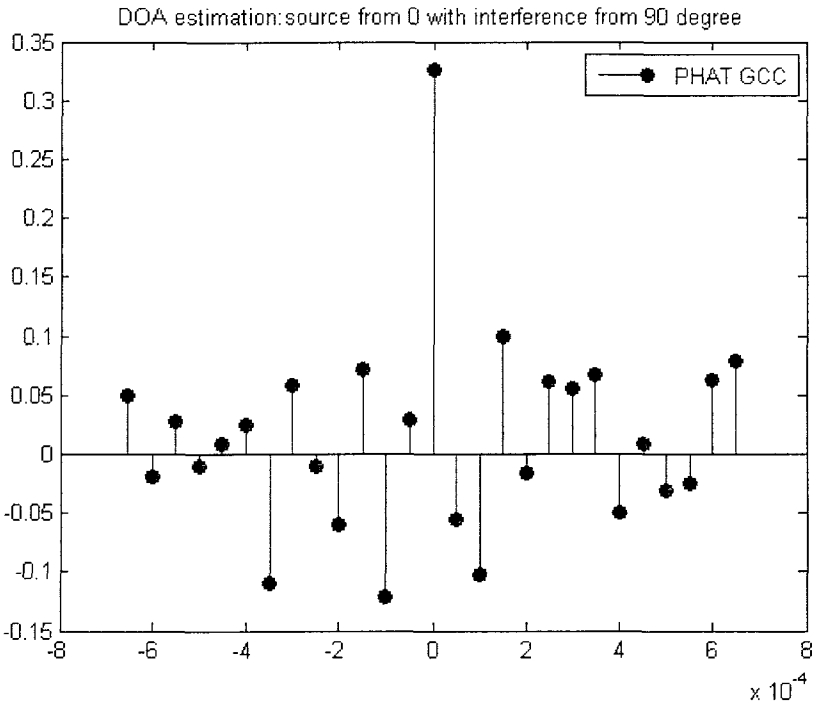

Figure 4.12: DOA estimator using the PHAT generalized cross correlation with a target source coming from 0° and a directional interference from 90°. The maximum ITD from the cross correlation function is 0, and the estimated DOA is 0°.

In Figure 4.10, the peak is hard to discriminate because there are three peaks and two of them are pretty close in magnitude. One is at time lag 0 and the other is greater than the maximum time lag (5.14e-4 second).

The SCOT and PHAT GCC methods show good estimation of target ITD and yield a correct DOA at an angle of 0 degree, as shown in Figure 4.11 and 4.12. The only peak in Figure 4.11 and 4.12 is much higher than the other estimating points and is easy to be discriminated.

Experiment environment # 4

When the source signal is corrupted with interference signal and noise, the advantage of the generalized cross-correlation method is obvious. In simulation # 4, we added a strong directional interference signal and a strong background noise. We again conduct this

experiment on the large cafeteria room recordings from the University of Oldenburg in Figure 4.13. The target male speaker comes from 30 degrees and is 150 cm away from the listener. The directional interference speaker comes from 240 degrees and is 3 meters away from the listener, and the Cafeteria is full of people having dinner, which makes a loud Cafeteria noise.

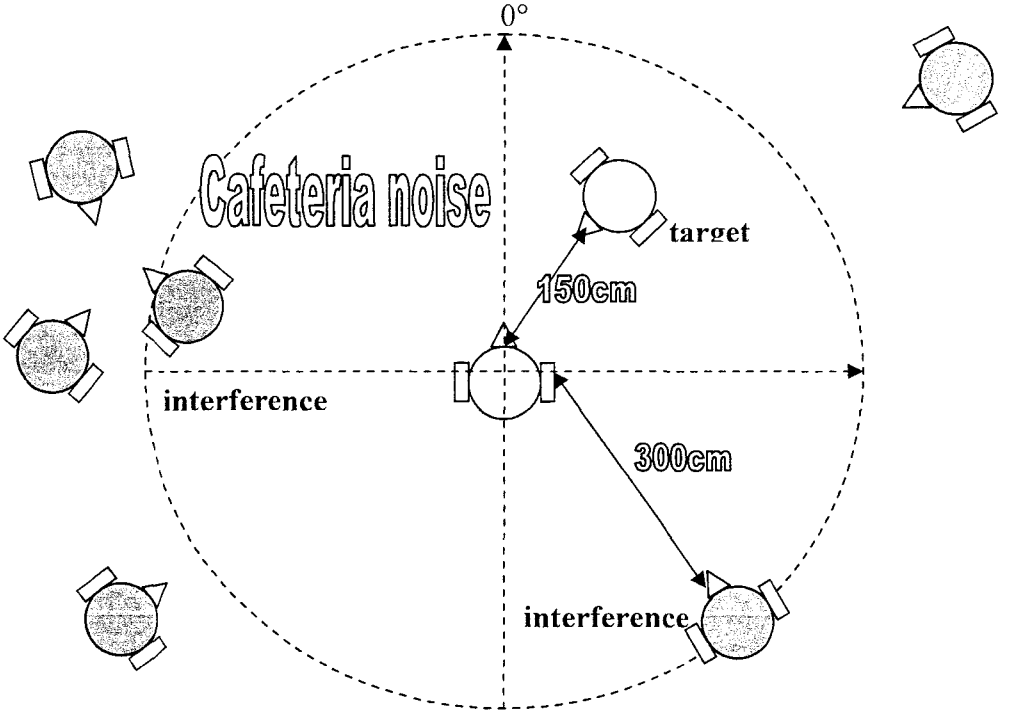

Figure 4.13: The acoustic environment of the cafeteria with loud noise for experiment # 4

The SNR from the first microphone at the left ear hearing aid is -13.36 dB and from the first microphone at the right ear hearing aid it is -8.36 dB.

The ITD for this scenario is estimated using different estimation methods, and the experimental results are shown in Figure 4.14, Figure 4.15 and Figure 4.16.

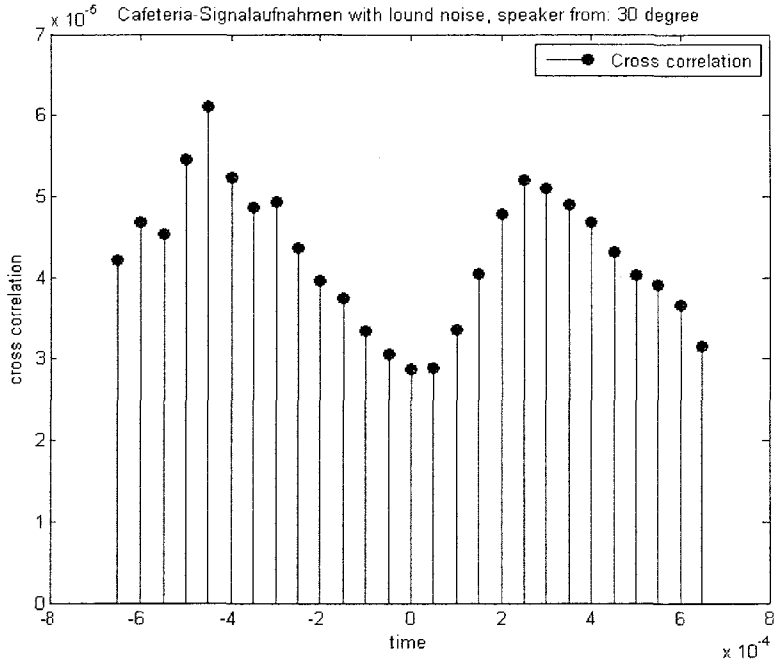

Figure 4.14: DOA estimator using the cross correlation method with a target source coming from 30°, a directional interference from 240° and loud background noise. The maximum ITD from the cross correlation function is -4.5e-004.

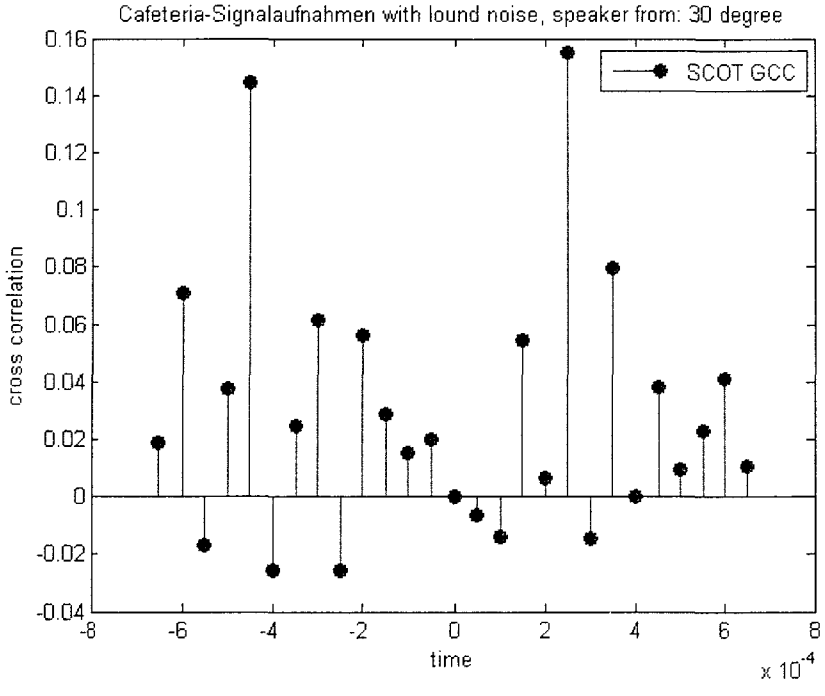

Figure 4.15: DOA estimator using the SCOT generalized cross correlation with a target source coming from 30°, a directional interference from 240° and loud background noise.

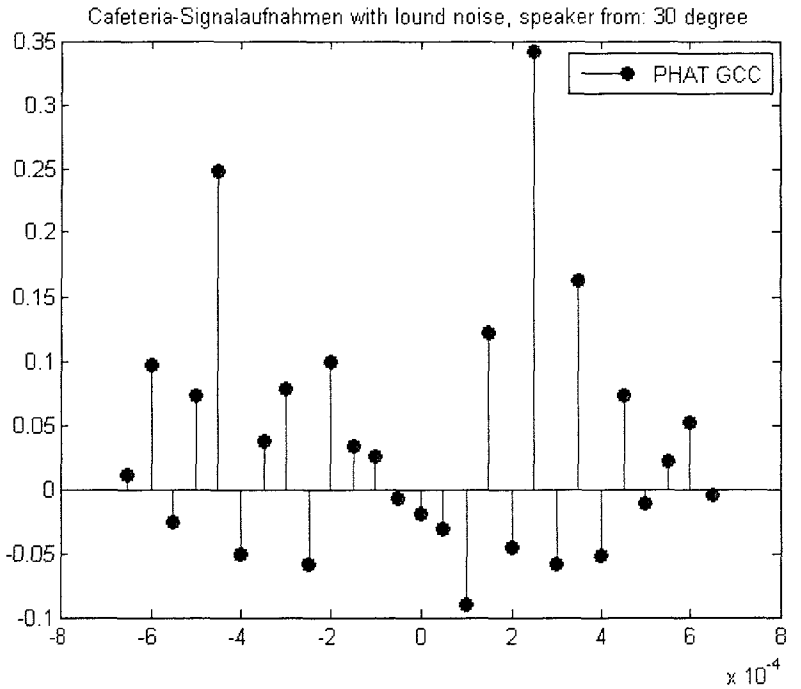

Figure 4.16: DOA estimator using the PHAT generalized cross correlation with a target source coming from 30°, a directional interference from 240° and loud background noise.

The ITD estimation in Figure 4.14 is 4.5e-4 second, which yields a DOA estimation of 299.03 (or 240.96 degrees because of front back symmetry). In this figure, we can also find another peak located at a time delay of 2.5e-4 second, which is lower than the first peak. This second highest peak corresponds to a source signal coming from an angle of 29.05 degrees or 150.94 degrees.

The target source signal and the directional interference signal are all strong, because in Figure 4.15 and 4.16, we can see two peaks in each plot. Differing from the situation in Figure 4.14, the largest peak is located at a time delay of 2.5e-4 second and the second peak is located at a time delay of -4.5e-4 second. However, as the target source signal is closer to the speaker, it is clearer than the interference signal. As a result, in the SCOT and PHAT GCC estimation output, the first peak is generated by the target source signal, and the estimation value from these methods is 29.05 degrees. The second peak is brought by the directional interference signal, and the estimation value is 299.03 (or 240.96) degrees.

Experiment environment # 5

To further demonstrate the advantage of generalized cross-correlation ITD estimation methods, we conducted another experiment with 2 directional interference signals in the same cafeteria room from the University of Oldenburg.

As is shown in Figure 4.17, the target male speaker is at 30 degrees and 150 cm away. The first female interference signal is at 240 degrees and 3 meters away. The second directional female interference speaker is at the listener's right side (DOA of 90 degrees) and 150 cm away.

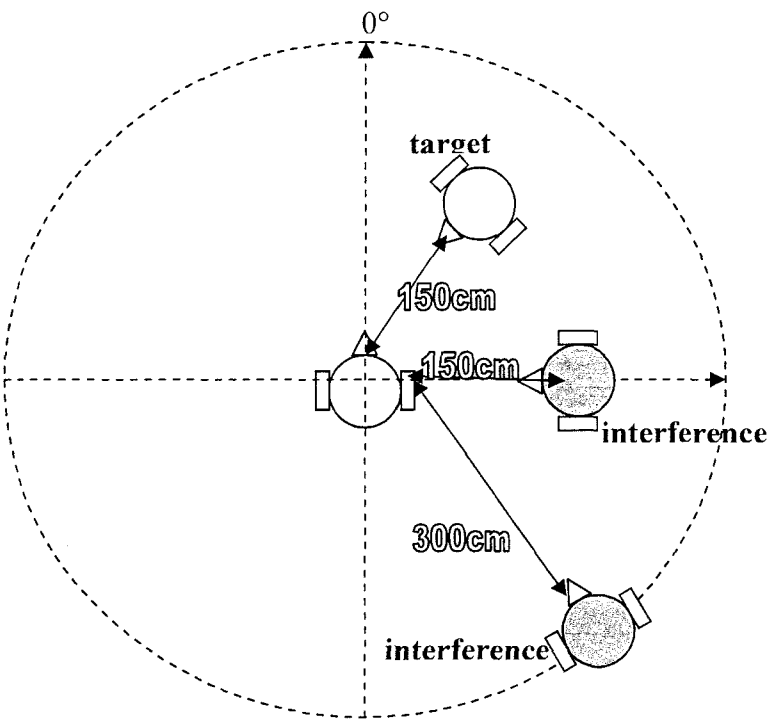

Figure 4.17: The acoustic environment of a large cafeteria for experiment # 5

The resulting SNR from the first microphone on the left ear hearing aid is -8.68 dB and from the first microphone at the right ear hearing aid it is -5.93 dB.

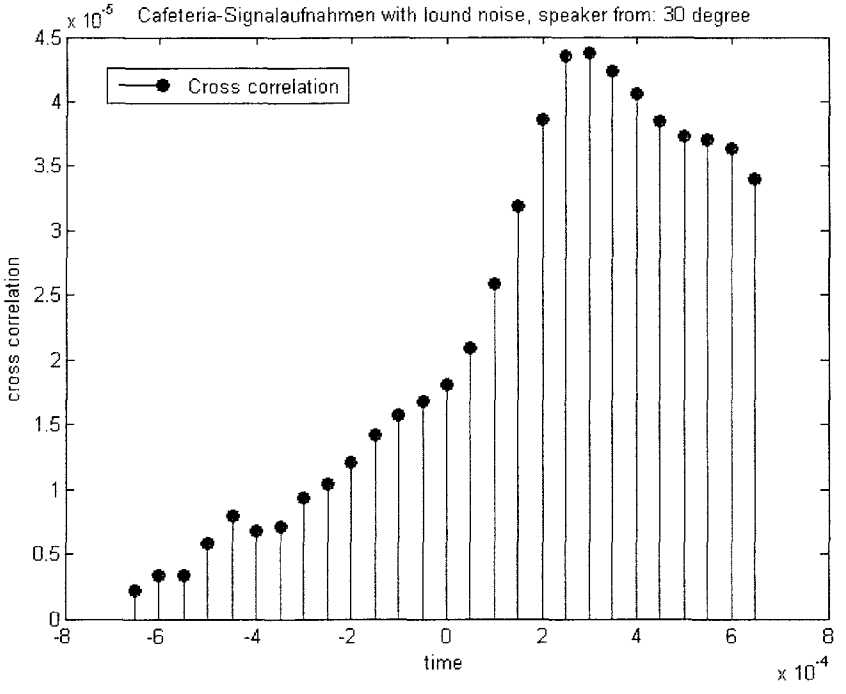

Figure 4.18: DOA estimator using the cross correlation method with a target source coming from 30°, and directional interferences from 240° and 90°.

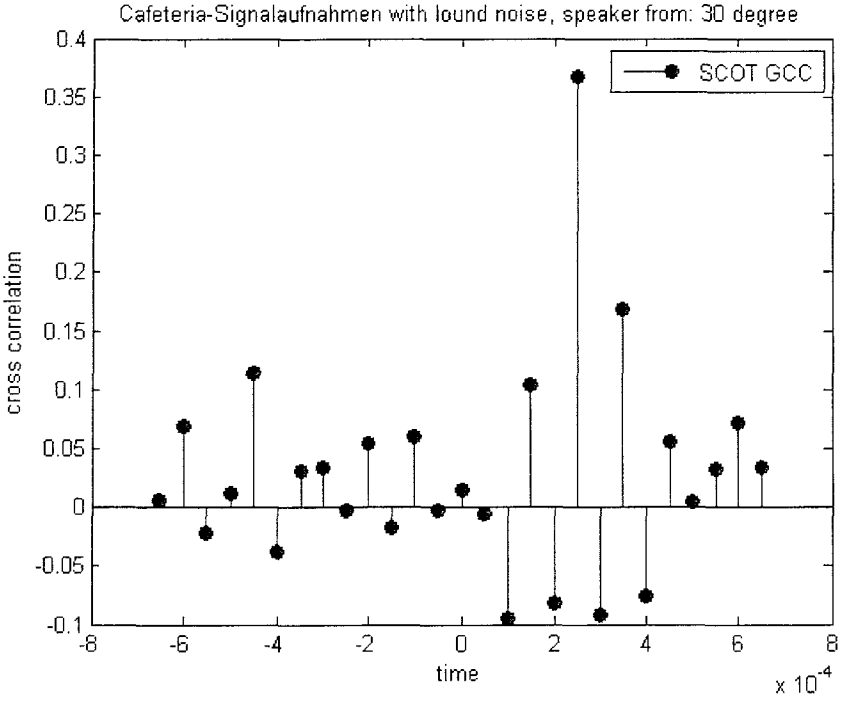

Figure 4.19: DOA estimator using the SCOT generalized cross correlation with a target source coming from 30°, and directional interference from 240° and 90°.

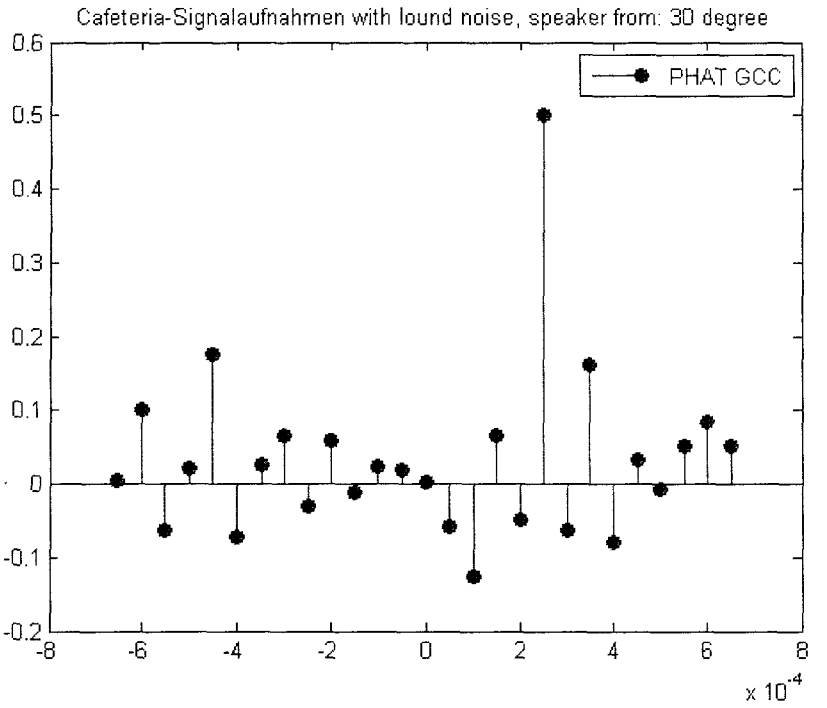

Figure 4.20: DOA estimator using the PHAT generalized cross correlation with a target source coming from 30°, and directional interference from 240° and 90°.

The ITD for this scenario is estimated using different estimation methods, and the experimental results are shown in Figure 4.18, Figure 4.19 and Figure 4.20.

In this scenario, the target signal is dominant in the speech mixtures observed by the microphones. As a result, in Figure 4.18 the cross-correlation has only one dominant peak in the output, and the estimated DOA is 35.65 degrees.

Similar to the previous basic cross-correlation experiment, for the SCOT and PHAT GCC in Figures 4.19 and 4.20, the estimated DOAs of the dominant target speech are both equal to 29.05 degrees. But being different from the previous basic cross-correlation experiment, using the SCOT and PHAT GCC the first interference signal is weak. As a result, it is hard to tell whether the second peak which is close to the time delay of -4.5e-4 second in Figures 4.19 and 4.20 results from the first interference signal. This means that the SCOT and PHAT GCC methods can be used for estimating the DOA of a dominant

target signal, but they cannot guarantee to get the DOA estimation for the second source signal.

Conclusion:

From the previous experiments, we can see that the generalized cross-correlation methods improve the peak detection ability and are also robust in reverberation and noisy acoustical environments. However if the speech mixture is composed of more than two signals, they are not robust for estimating the DOA of the second source signal. In the next section, we apply another DOA estimation method using ICA and directivity pattern based methods to estimate not only the DOA of a target source signal, but also the DOA of the interference signals.

4.1.2 DOA estimation using directivity pattern based method

It has been proposed by researchers that the unmixing matrix in frequency domain blind source separation (or independent component analysis, ICA) has directivity patterns which are similar to a null beamformer [Sar'06]. The DOA information of source signals could then be estimated from those directivity patterns. In [Saw'04], the author proposed a robust but somewhat inaccurate method of estimating the direction of arrival. This method was developed for solving the permutation problem of frequency domain blind source separation. In the frequency domain, due to permutation and scaling ambiguities, the frequency response $H(f)$ of the mixing matrix can be written as

$$H = W^{-1}P^{-1}\Lambda^{-1} \tag{48}$$

where W is the unmixing matrix, P is a permutation matrix, and Λ is a scaling matrix. Let θ_k be the direction of source k. The frequency response of the element h_{lk} in the mixing matrix can be written as:

$$H_{lk}(f) = A_{lk}e^{j\varphi_k}e^{j2\pi f\tau_{lk}} \tag{49}$$

where τ_{lk} is the delay from source k to mixture (or sensor) l. The magnitude A_{lk} and phase φ_k, which can be arbitrary values, are introduced due to the frequency permutation in the ICA. Then, by calculating the ratio between two elements of the same of column $H(f)$, we obtain:

$$\frac{H_{lk}}{H_{jk}} = \frac{\left[W^{-1}P^{-1}\Lambda^{-1}\right]_{lk}}{\left[W^{-1}P^{-1}\Lambda^{-1}\right]_{jk}} = \frac{\left[W^{-1}P^{-1}\right]_{lk}}{\left[W^{-1}P^{-1}\right]_{jk}} = \frac{A_{lk}e^{j2\pi f(\tau_{lk}-\tau_{jk})}}{A_{jk}} \tag{50}$$

where $\left[W^{-1}P^{-1}\right]$ is the unmixing matrix obtained from the previous source separation step, and the scaling matrix values are cancelled out.

For binaural hearing aids with one microphone on each side, the time delay between the two microphones is $\tau_k = \tau_{lk} - \tau_{jk} = \frac{d_{LR}}{c}\sin\theta_k$. Then from the unmixing matrix W, the directivity pattern can be obtained from the column components of the matrix W:

$$\theta_k = \arcsin\frac{\arg\left(\frac{[W^{-1}P^{-1}]_{lk}}{[W^{-1}P^{-1}]_{jk}}\right)}{2\pi f c^{-1} d_{LR}} = \arcsin\frac{\arg\left(\frac{H_{lk}}{H_{jk}}\right)}{2\pi f c^{-1} d_{LR}} \tag{51}$$

Similarly, for a linear endfire microphone array (e.g. monaural, as in Figure 4.1), the time delay between microphones located at positions d_l and d_j on the same side is $\tau_k = \tau_{lk} - \tau_{jk} = \frac{d}{c}\cos\theta_k$, and the DOA estimate can be calculated from:

$$\theta_k = \arccos\frac{\arg\left(\frac{[W^{-1}P^{-1}]_{lk}}{[W^{-1}P^{-1}]_{jk}}\right)}{2\pi f c^{-1}(d_l - d_j)} = \arccos\frac{\arg\left(\frac{H_{lk}}{H_{jk}}\right)}{2\pi f c^{-1}(d_l - d_j)} \tag{52}$$

Simulations and results:

In our experiments, we begin with DOA estimation for monaural hearing aids. From the literature we know that each frequency can yield a DOA estimation independently and that the low frequency bands and the high frequency bands typically do not provide good estimation using the directivity pattern (DP) estimation method. As a result, we only use a selected frequency band to get estimation, and this can also possibly save some computations. In the following experiments #6 to #10, we can observe that each frequency band gives different estimate values. By comparing the performance from each frequency band, we are able to select one frequency band which gives the best estimation performance.

Simulation Experiment # 6:

In this experiment, different DOA estimations are conducted to evaluate their preciseness using different estimation frequency bands. In the experiments, the target source speech signals come from the right side of the head. The target DOAs are 80 degrees and 110 degrees, and ICA parameters used in this experiment (and the other experiments to follow) are a step size of $\lambda = 0.001$ with the number of adaptation iterations $\kappa = 60$. In order to get a good estimation of the target DOAs, we use the microphone pair #1 and #3 in the provided recordings, which are the first and third microphones at the right ear hearing aid. The distance between these two microphones is 1.5 cm. In Table 4.2, we compare the DOA estimates within different frequency bands. We see that the frequency ranges (1k – 3kHz) and (5k -8kHz) are the two best frequency ranges which give good DOA estimations.

Table 4.2: DP based DOA estimation for two target sources from 80° and 110°

DOA	0 – 1kHz	1k – 3kHz	3k – 5kHz	**5k – 8kHz**	9k – 10kHz
Target1 (110°)	86.74°	106.69°	110.52°	**104.34°**	90.70°
Target2 (80°)	71.07°	79.84°	92.50°	**89.90°**	89.66°

Simulation Experiment # 7

In the experiment #7, the target sources are from 115° and 60° degrees, and the experimental results are shown in Table 4.3. We see that the frequency range (1k – 3kHz) and (5k -8kHz) are the two best frequency ranges which give good DOA estimations.

Table 4.3: DP based DOA estimation for two target sources from 60° and 115°

DOA	0 – 1kHz	1k – 3kHz	3k – 5kHz	**5k – 8kHz**	9k – 10kHz
Target1 (115°)	88.08°	113.12°	110.25°	**114.75°**	89.38°
Target2 (60°)	53.51°	57.65°	67.15°	**65.37°**	87.51°

Simulation Experiment # 8:

In the experiment #8, both target signals are from quadrant 1: one is from 45° and the other is from 75°. The results are shown in Table 4.4. We see that the frequency range (5k -8kHz) and (1k – 3kHz) are the two best frequency ranges which give good DOA estimations.

Table 4.4: DP based DOA estimation for two target sources from 45° and 75°

DOA	0 – 1kHz	1k – 3kHz	3k – 5kHz	**5k – 8kHz**	9k – 10kHz
Target1 (75°)	58.98°	70.64°	67.97°	**81.40°**	86.67°
Target2 (45°)	35.03°	33.34°	41.84°	**40.89°**	84.72°

Simulation Experiment # 9:

In the experiment # 9, one of the target sources is at 30 degree, which is more frontal and more co-linear to the axis of the monaural microphones pair. The other source is at 110

degrees. The results are shown in Table 4.5. We see that the frequency range (5k -8kHz) and (3k – 5kHz) are the two best frequency ranges which give good DOA estimations.

Table 4.5: DP based DOA estimation for two target sources from 30° and 110°

DOA	0 – 1kHz	1k – 3kHz	3k – 5kHz	**5k – 8kHz**	9k – 10kHz
Target1 (110°)	88.98°	111.51°	104.40°	**113.53°**	83.60°
Target2 (30°)	51.32°	43.69°	37.35°	**26.00°**	79.90°

In experiment # 9, we notice that there are some large estimation errors for target source 2, because it is more co-linear to the axis of the monaural microphones pair.

As we know that $\cos\theta_k = \frac{\arg\left(\frac{\left[W^{-1}P^{-1}\right]_{lk}}{\left[W^{-1}P^{-1}\right]_{jk}}\right)}{2\pi fc^{-1}(d_l - d_j)} \in [-1,1]$ and each frequency bin yields a DOA estimation independently. For target source 2 in some frequency bins the $\cos\theta_k$ value which is calculated from the coefficients of ICA is out of this range. In order to minimize the estimation error, these points have been ignored in the computation of the overall DOA estimate. An alternative would have been to use saturation or limiting function to force the measured $\cos\theta_k$ to always be in the range $[-1,1]$.

Simulation Experiment # 10

In this experiment the sources are located at 10° and 55°, and the results are shown in Table 4.6.We see that the frequency range (5k -8kHz) and (3k – 5kHz) are the two best frequency ranges which give good DOA estimations.

Table 4.6: DP based DOA estimation for two target sources from 10° and 55°

DOA	0 – 1kHz	1k – 3kHz	3 k – 5 kHz	**5 k – 8 kHz**	9k – 10 kHz
Target1 (55°)	41.34°	44.36°	47.01°	**59.22°**	83.23°
Target2 (10°)	31.46°	18.54°	22.15°	**10.87°**	79.87°

In experiment # 10, the second target estimation (source at 10°) has also encountered frequency bins for which the measured value of $\cos\theta_k$ was out of the range [-1, 1], as in experiment # 9. In the frequency band 0–1 kHz, 73% of the frequency bins had this problem. In the frequency bands 1kHz - 3kHz and 3kHz - 5kHz, the problem occurred in 41% and 44% of the bins, respectively. In the frequency band 5kHz - 8kHz, 90% of the frequency bins suffered from this problem.

While for the target source 2 there are a lot of frequency bins that cannot yield a correct DOA estimation, this does not affect the DOA estimation of the other target source. The DOA estimations of target source 1 and target source 2 are independent. The main factor that influences the DOA estimation is the relative position of the source signal with respect to the microphone array.

Simulation Experiment # 11

In this experiment, a three signals mixture is composed by a target signal (speech #9 from the provided recordings) coming from 30 degrees, an interference signal (speech #10) coming from 65 degrees, and another interference signal (speech #7) coming from 330 degrees. Their relative magnitude is 1:1:0.3, and the speech mixtures are received from microphones #1 and #4 of the binaural hearing aid recordings. Using BSS to separate the signals and estimating the DOA in the frequency range of 5 kHz to 8 kHz, we get the first DOA estimate of 26.82 degrees, and the second DOA estimate of 68.49 degrees.

20 estimate points for target 1 were out of range and were discarded as previously explained, while the rest of the 57 frequency bins yielded valid estimates. The estimates for interference 1 had no out of range errors.

Simulation Experiment # 12

A three-signal mixture is composed of a target signal (speech #9) coming from 10 degrees, an interference signal (speech #10) coming from 45 degrees, and another interference signal (speech #7) coming from 300 degrees. Their relative magnitude is 1:1:0.3, and the speech mixtures are received from microphones #1 and #3 from the

monaural hearing aid recordings. Using BSS to separate the signals and estimating the DOA in the frequency range of 5 kHz to 8 kHz, we get a first DOA estimate of 10.23 degrees and a second DOA estimate of 46.13 degrees.

Discussion:

By comparing the estimation results from experiment #6 to experiment #10, we find that overall the frequency range of 5kHz to 8kHz gives the best estimation performance, and we highlighted in bold the results for this range in the different tables of result.

The DOA estimation method of this section is based on BSS results and can lead to larger DOA estimation errors. From experiments #6-#12, it is clear that the DP based DOA estimation method gives a large estimation error in the first frequency band and the last frequency band. However, in the other frequency bands, it yields better estimation results. By comparing the results in tables 4.2-4.6, we can also see that the band 5 kHz to 8 kHz gives the closest DOA estimation. A screening step was added into the DP based DOA estimation method to eliminate the frequency bins for which the measured $\cos\theta_k$ was not in the range $[-1,1]$.

Comparing this DP based DOA estimation method with the cross-correlation based DOA estimation methods of the previous section, each of them has their strength and weaknesses. The cross-correlation based DOA estimation methods are robust in reverberation and noisy room environment, but they can only produce one robust target DOA estimation (i.e. the dominant one) and the precision of the estimated DOA is directly related to the inverse of the sampling frequency. The directivity pattern based method can estimate the DOA of both a target source signal and the main interference signal at the same time; however, the accuracy of the estimation results is typically weaker.

4.2 Null Beamformer

A null beamformer is a spatial filter to attenuate signals from interfering directions while keeping a target signal untouched. Two classical techniques for null beamformers are the delay-and-sum method and the minimum variance distortionless-response (MVDR) method.

A general delay-and-sum beamformer is obtained by linearly multiplying the microphone signals with weighting coefficients w. In a free field environment and a monaural array (endfire orientation), the steering vector is defined as

$\underline{D}(\omega,\theta) = [d_1(\omega,\theta) \quad d_2(\omega,\theta) \quad \cdots \quad d_N(\omega,\theta)]^T = [e^{-j\omega_1} \quad e^{-j\omega_2} \quad \cdots \quad e^{-j\omega_{(N-1)}}]^T$, where $\omega_i = 2\pi f d_i \sin(\theta)/c$, θ is the angle from the frontal direction, d_i is the position of the *i-th* microphone, f is the sampling frequency, and N is the number of the sensors.

If M source signals come from M directions, $\theta = \theta_1, \theta_2, \cdots \theta_M$, in order to steer nulls to the direction of the interference signals $\theta_n \in \{\theta_1, \theta_2, \cdots \theta_M\}$ and put a unit amplitude in the directions of the target sources $\theta_s \in \{\theta_1, \theta_2, \cdots \theta_M\}$, for a null beamformer we have the weighting coefficients w so as to make

$$w^H(\omega)D(\omega,\theta) = I, \tag{53}$$

where $D(\omega,\theta) = [d(\omega,\theta_s) \quad d(\omega,\theta_n)]$.

Then, we get the null beamformer weighting coefficients as

$$w^H = D^{-1}. \tag{54}$$

The beam-pattern of this null beamformer is written as

$$W(\omega,\theta) = w^H(\omega)d(\omega,\theta). \tag{55}$$

Finally, the beamformer output can be written as

$$Y(\omega) = W^H(\omega)X(\omega) \quad (56)$$

In binaural hearing aids, the head shadow effect is another factor which needs to be considered. In such cases, the steering vector $d(\omega,\theta) = [d_1(\omega,\theta) \quad d_2(\omega,\theta) \quad \cdots \quad d_N(\omega,\theta)]^T$ is replaced by the frequency response of the head related transfer function (HRTF) from the direction of θ.

The Minimum Variance Distortionless Response (MVDR) is another popular method for designing a beamformer:

$$W(\omega) = \frac{(D(\omega,\theta_i)D^H(\omega,\theta_i) + \mu' I)^{-1} D(\omega,\theta_s)}{D^H(\omega,\theta_s)(D(\omega,\theta_i)D^H(\omega,\theta_i) + \mu' I)^{-1} D(\omega,\theta_s)} \quad (57)$$

In the above equation, $D(\omega,\theta_i)$ is the steering vector of an interference signal coming from an angle of θ_i, $D(\omega,\theta_s)$ is the steering vector of a target source signal coming from an angle of θ_s, and the term $\mu' I$ is a trade-off factor which is introduced to improve numerical conditioning, to avoid the loss of directivity by microphone mismatch, or to reduce the white noise gain.

4.3 Preservation of the spatial cues using a common gain

Both classical beamformers and classical blind source separation algorithms destroy the spatial (i.e. interaural) cues in their processing (even for binaural hearing aids). However, in real life situations, the human auditory system relies on the spatial impression to distinguish directional information. As a result, it is necessary to recover or compensate the interaural amplitude and phase differences in order to keep the spatial cues for the listener.

In [Lot'06], the author proposed a modified beamforming method preserving the interaural amplitude and phase differences of the original signal. It is proposed to use one common gain at each frequency $G(f,t)$ for both the left and right ear side and to apply this common gain to the input signals of both sides to perform enhancement while preserving the binaural cues. An easy way to calculate the weights is by comparing the spectral amplitudes of the "normal" or classical beamformer output $Y(f,t)$ to the sum of both input spectral amplitudes, $X_l(f,t)$ and $X_r(f,t)$,

$$G(f,t) = \frac{|Y(f,t)|}{|X_l(f,t)| + |X_r(f,t)|} \tag{58}$$

The final binaural outputs for the left and right ears are then:

$$Y_l^{final}(f,t) = G(f,t) \cdot X_l(f,t) \tag{59}$$

$$Y_r^{final}(f,t) = G(f,t) \cdot X_r(f,t) \tag{60}$$

More details on the computations of the common gain $G(f,t)$ can be found in [Lot'06].

4.4 Proposed Combination of Beamformer with Blind Source Separation

We have previously introduced the method from [Ara'04a] [Ara'04b] [Ara'04c], where an algorithm was designed to mask out extra sources and to convert an underdetermined blind source separation to an ordinary or determined blind source separation problem. The process of the algorithm from [Ara'04a] [Ara'04b] [Ara'04c] is summarized here as follows:

- Remove *N-M* sources with the directivity pattern based continuous mask in each frequency bin. In this step, a null beamformer is formed by using *N-M+1* virtual

microphones, which creates nulls towards the given *N-M* directions. The directivity pattern is used to form the T-F masks.

- Separate the remaining sources by ICA. In this step, since the system has only two outputs, we need several masks with nulls towards different directions, if we want to obtain all the separated *N* signals.

Alternatively, we propose in this thesis to use a combination of beamforming as a pre-processor and ICA as a post-processor, in order to convert an undetermined blind source separation problem into a determined source separation problem. Figure 4.21 illustrates our proposed method.

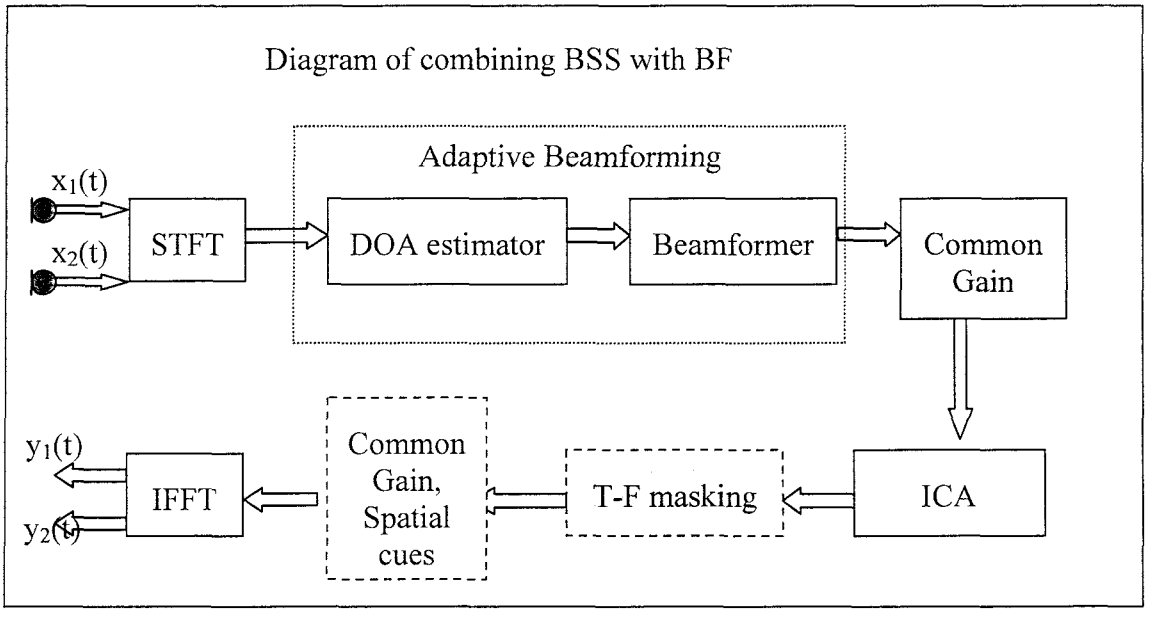

Figure 4.21: Block diagram of blind source separation combining beamforming

As shown in Figure 4.21, the signal mixture $x_1(t)$ and $x_2(t)$ observed by two microphones in a hearing aid are firstly passed to an adaptive beamforming procedure. This beamforming process part is composed of a DOA estimator and a beamformer. The adaptive beamformer removes one main interference signal and keeps the target signal and other interferences. If the original source mixture is composed of three speech signals, in this case, by removing one interference speech signal from the mixtures using an adaptive beamformer, the underdetermined blind source separation is transferred into a

determined blind source separation problem. At the output of adaptive beamforming there is only one output (and obviously no binaural cues preserved). But the ICA process needs two channels of input signals from the beamforming. In order to recover two channels of signals to be used for blind source separation, the procedure of applying a common gain (as for the preservation of the spatial cues) is added in the diagram. Then blind source separation is applied (i.e. ICA) and it may also include a T-F masking process. Finally, for binaural hearing aids applications, the preservation of the spatial cues is preferable, so it is necessary to add again the procedure of applying a common gain after the blind source separation and before forwarding the final output to the listener. But for monaural hearing aid i.e. with microphones on one side of the head only, this step can be removed.

4.5 Simulation for underdetermined blind source separation

To further illustrate the performance of our method, we include below two typical experiments which demonstrate the speech enhancement ability of our system.

Simulation experiment #13

In our simulation, we use a three speech signals mixture as the input of our hearing aids, which is the same as the experimental setting in the previous DOA estimator section. The acoustic environment is shown in Figure 4.22, where the source signals are composed of one target male speech coming from an angle of 45 degrees, a strong interference speech which is a female speech coming from an angle of 10 degrees, and a relatively weak female interference speech coming from an angle of 300 degrees. The magnitude ratio of these three speeches was set to 1:1:0.3. A monaural hearing aid at the right side of the head was used. The signal mixtures were collected from microphones #1 and #3 (from the provided recordings), with a distance of 1.5 cm.

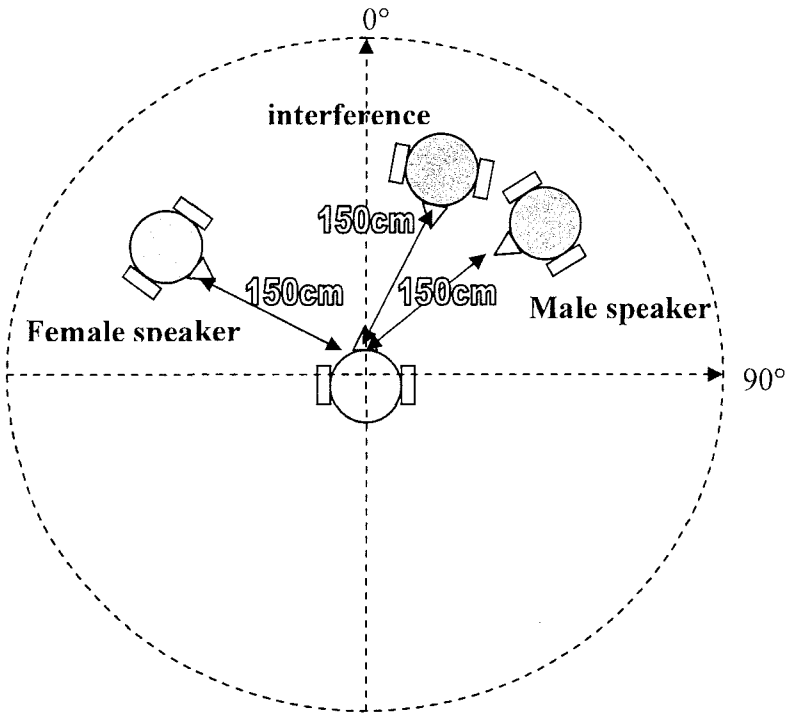

Figure 4.22: The acoustic environment for simulation # 13

This whole problem can be described as an underdetermined blind source separation problem. So, in our solution, we use a beamformer to remove the extra interference and transfer this underdetermined problem into a determined blind source separation problem. From the previous section on beampattern-based DOA estimation, the estimated DOA of the target male speaker is 46.13 degrees, and the DOA of first interference is 10.23 degrees. Since we do not have measured HRTFs for these two angles (required for the MVDR beamforming design), in this experiment we first assume that our DOA estimator has no estimation error, i.e that it gives 45 degrees for the target source and 10 degrees for the interference. MVDR beamforming is then conducted for removing the female interference at 10 degrees and one output is obtained. Before ICA processing, we use the spatial cues preservation method described in section 4.3 and recover two source signals, in which the main interference is almost gone. The ICA can then easily separate the target source and the remaining weak interference signals. Table 4.7 shows the evaluation of the final output, compared with or without the last T-F masking step.

Table 4.7: Underdetermined BSS using MVDR BF, target speech from 45°, interference speech from 10° and 300°. The two microphones are placed at the right side of the head.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:45° Interfering speech: female speech #9 DOA 10° and female speech #10 DOA 300°			
BSS with T-F masks	-1.90	12.13	14.03
BSS without T-F masks	-1.90	9.84	11.74
Target speech: female speech #10, DOA: 300° Interfering speech: female speech #9 DOA 10° and male speech #5 DOA 45°			
BSS with T-F masks	-15.50	2.58	18.09
BSS without T-F masks	-15.50	2.88	18.39

Simulation experiment # 14

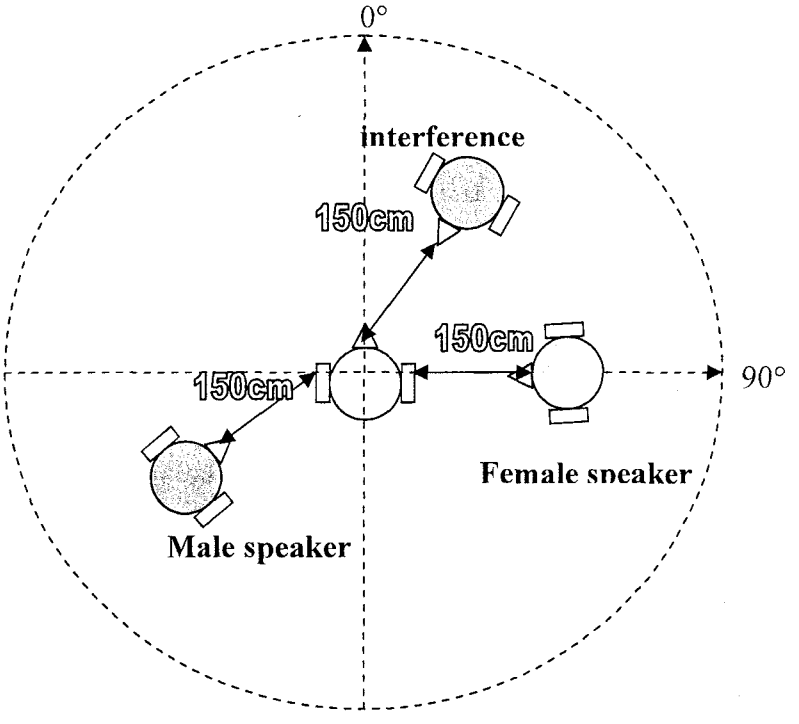

Figure 4.23: The acoustic environment for simulation # 14

Similar to simulation #13, in simulation #14 we still use three speech signals to compose the mixtures, as shown in Figure 4.23. But the two microphones (microphone #1 and #4 from provided recordings) are now from each side of the head. The target source comes from 240 degrees, the main interference signal comes from 30 degrees and the second interference comes from 90 degrees. The magnitude ratio of these three speeches is 1:1:0.3. The same processing procedures are conducted, and Table 4.8 shows the evaluation of the final outputs, with or without the last T-F masking step.

Table 4.8: Underdetermined BSS using MVDR BF, target speech from 240°, interference speech from 30° and 90°. The two microphones are placed at each side of head.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, DOA:240° Interfering speech: female speech #9 DOA 30° and female speech #10 DOA 90°			
BSS with T-F masks	-4.21	8.88	13.10
BSS without T-F masks	-4.21	9.88	14.10
Target speech: female speech #10, DOA: 90° Interfering speech: female speech #9 DOA 30° and male speech #5 DOA 240°			
BSS with T-F masks	-12.30	3.55	15.85
BSS without T-F masks	-12.30	3.00	15.30

Simulation experiment # 15:

In the previous experiments, we assume that the DOA estimator gives accurately the DOA of the target signal and the main interference. However, in real situations the DOA estimation contains errors. So in this experiment we use DOA estimation with an error in it and repeat the simulation # 13. The DOA estimation for the target speech used in this experiment is 50 degrees (real DOA is 45 degrees), and the DOA for the main interference speech is 15 degrees (real DOA is 10 degrees). Table 4.9 shows the resulting SNR gains for this experiment.

Table 4.9: Underdetermined BSS using MVDR BF, where the DOA estimate for BF includes an error. The target speech is from 45° while the estimated DOA is 50°. The main interference speech is from 10° while the estimated DOA is 15°. The second interference comes from 300°. The two microphones are placed at the right side of the head.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, estimated DOA:50°, real DOA 45° Interfering speech: female speech #9 estimated DOA 15° (real DOA10°) and female speech #10 DOA 300°			
BSS with T-F masks	-1.90	4.98	6.88
BSS without T-F masks	-1.90	4.86	6.76
Target speech: female speech #10, DOA: 300° Interfering speech: female speech #9 estimated DOA 15° (real DOA10°) and male speech #5 estimated DOA:50°, real DOA 45°			
BSS with T-F masks	-15.50	2.58	18.09
BSS without T-F masks	-15.50	2.99	18.50

Simulation experiment # 16:

Similarly, in this experiment we use DOA estimation with an error in it and repeat the simulation # 14. The DOA estimation for the target speech used in this experiment is 245 degrees (the real DOA is 240 degrees), the DOA for the main interference speech is 25 degrees (the real DOA is 30 degrees) and the second interference comes from 90 degrees. Table 4.10 shows the resulting SNR gains for this experiment.

Table 4.10: Underdetermined BSS using MVDR BF, where the DOA estimate for BF includes an error. The target speech is from 240° while the estimated DOA is 245°. The first interference speech is from 30° while the estimated DOA is 25°. The second interference comes from 90°. The two microphones are placed on each side of the head.

	SNR_in (dB)	SNR_out (dB)	SNR gain (dB)
Target speech: male speech #5, estimated DOA:245° (real DOA: 240°) Interfering speech: female speech #9 estimated DOA 25° (real DOA: 30°) and female speech #10 DOA 90°			
BSS with T-F masks	-4.21	4.23	8.45
BSS without T-F masks	-4.21	4.86	9.07
Target speech: female speech #10, DOA: 90° Interfering speech: female speech #9 estimated DOA 25° (real DOA: 30°) and male speech #5 estimated DOA:245° (real DOA: 240°)			
BSS with T-F masks	-12.30	-0.36	11.93
BSS without T-F masks	-12.30	-0.26	12.03

4.6 Conclusion

From the previous sections, it can be concluded that the underdetermined blind source separation problem can be solved effectively by combining beamforming with ICA. The estimation of DOA plays an important role in this pre-processing step. Time delay based DOA estimation methods which were introduced in section 4.1.1 are one of the most popular DOA estimation methods. They are robust in noisy or reverberating room environments compared with DP-based DOA estimation methods. However, their resolution directly depends on the sampling rate, and time delay based methods are not the best methods to provide the DOA of both a target and the first dominant interference. In section 4.1.2, a DP-based DOA estimation method was introduced, which has typically a larger estimation error than the time delay based DOA methods, but which is more suitable to provide the DOA of both a target and a dominant interference. This DOA estimation method relies on BSS outputs (i.e. de-mixing coefficients), and from our

simulations we found that the high frequency band (5kHz to 8kHz) yields the best estimations for this method. Using only this 5kHz to 8kHz band can also reduce the calculation complexity [Pan'07].

With correct or approximately correct DOA estimation, a MVDR beamformer can remove an extra interference signal efficiently. If the DOA estimate had some estimation error (e.g. 5 degrees), the MVDR beamformer was still able to largely attenuate the interference while keeping the target signal mostly untouched. As beamforming generates only a single output (with no binaural cues preserved), a simple step from section 4.3 i.e. the use of a common gain was applied to generate two channels of inputs for the BSS processing. ICA is then performed to separate the target source signal and the remaining interference(s). The experimental results have showed that the separated target typically sounds quite audible and clear. Depending on the performance of the beamforming, the separated interference may still contain a part of the other interferences. Although from the objective evaluations in section 4.5, the T-F masks did not always make an improvement on the SNR gain, from informal listening it appears that the T-F masks help to remove the remaining weak interference noise in the background sound, while at the same time possibly also introducing some distortion. The last step in the proposed system diagram of Figure 4.21 is for the preservation of the spatial cues. For binaural hearing aids, this step is important for the user to keep the interaural time and level differences. But for monaural hearing aids with microphones on one side of the head only, this step can be removed.

Chapter 5 Summary and Future work

When the acoustic environment is noisy, our human auditory system can steer our ears to the direction of a target source, and then the human brain analyzes the sound mixture and extracts the important target component from it. In this thesis, we proposed a system for enhancing the quality of the signal produced by the hearing aid in a multiple speaker environment to mimic the behaviour of the human brain when analyzing the audio mixtures. In hearing applications, the number of source signals is variable while the number of microphones on the hearing aids is fixed. The proposed system combines the spatial information with blind source separation. By removing one or more interference signals from the mixtures, the underdetermined BSS problem was converted into a determined BSS problem, and all the algorithms for determined BSS can be applied. Results show that our system can locate the target signal in different kinds of acoustical environment and have good learning ability. Compared with the work in [Ped'08], our method produces a good performance and it has a low complexity for convolutive speech mixtures. And compared with the work in [Pan'07], our method is more robust in noisy and reverberated room environments.

In the first stage of our work, we studied the adaptation (learning) ability of determined BSS systems and compared the learning ability of BSS system with or without T-F masks.

In the second stage of our work, we investigated the approach of using spatial processing (i.e. beamforming) to remove the extra interference(s) and convert an underdetermined BSS to a determined BSS. Given accurate DOA of the interference signal, an MVDR beamformer can effectively remove that interference. Therefore, the estimation of the DOA is important. The two DOA estimation methods which were used in the thesis both have their strengths and weaknesses. The cross-correlation based DOA estimation methods are robust in reverberation and noisy room environment, but they can only produce one robust target DOA estimation (i.e. the dominant one) and the precision of the estimated DOA is directly related to the sampling frequency. The directivity pattern

based method can estimate the DOA of both a target source signal and the main interference signal at the same time; however, the accuracy of the estimation results is more affected by the reverberation and other environment noise. [Muk'06] gives an overall evaluation of the directivity pattern based DOA estimation method and shows that this method is sensitive to source locations. The estimator can provide a robust DOA estimation when the source signals are perpendicular with a microphone pair (broadside configuration), but there is a large error when the source signals are nearly co-linear to the axis of the microphone pair (endfire configuration). The author proposes to use multiple sensor pairs with various axis directions to avoid these unreliable cases. In binaural hearing aids with more than one microphone on each side, it could be possible to use two dimensional microphone arrays to get a more reliable DOA estimation. However, there would be head shadow effects to consider between the different microphone pairs. As shown in Figure 4.18, for the frontal direction which is shown in DOA area 1, the DOA estimation could be computed with the signals from the microphones located on each side of the head. And if the source signal comes from the side, for example the DOAs which are shown in DOA area 2, the signals from the microphones located on the same side could be used to achieve a better estimation.

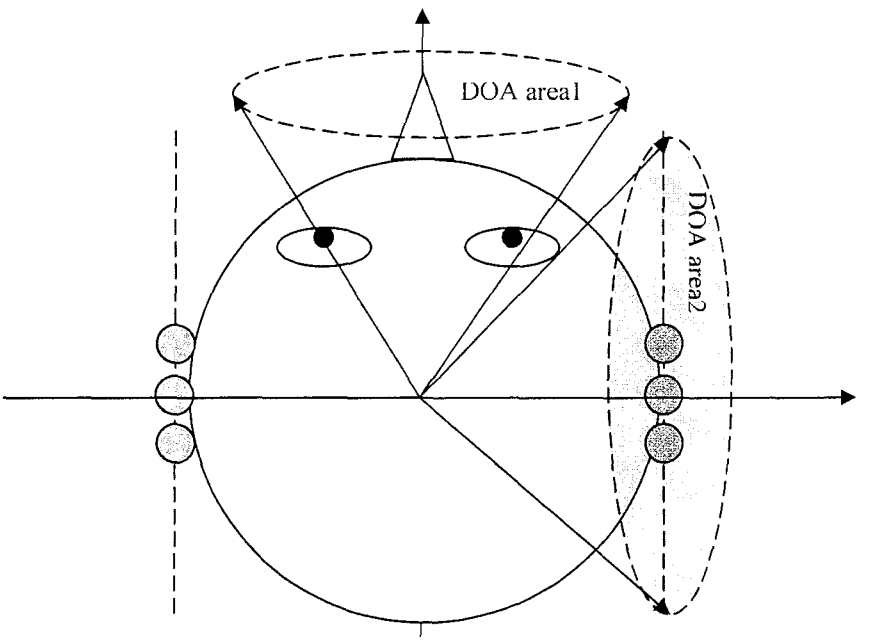

Figure 5.1: Two dimensional microphone array for estimation of DOA

Other obvious aspects that would require more work for our topic are the development of efficient implementations for the proposed solutions, with an evaluation of the resulting complexity. In particular, the use of on-line BSS algorithms would be of interest for reduced complexity, reduced latency, and reduced memory requirements. With such efficient realizations, a real-time implementation of the proposed methods could become possible, allowing further tuning and testing of the proposed methods under different real-life environments.

In this thesis, we have used the method of the common gain introduced in section 4.3 for two different purposes, as can be seen in figure 4.21: 1) to generate a pair of signals from the single channel output of the classic beamformer pre-processor in Chapter 4, because the ICA stage needs a two-channel input; and 2) in the case of binaural hearing aids, to re-generate the binaural cues at the final output of the system. However, applying this common gain to the original input mixtures causes a part of the interference noise to remain in the output signals, and introduces some distortion in the output signals. This can affect the performance of the ICA system, and/or lead to audible effects in the final outputs. Therefore, introducing a better or alternative method to generate a 2-channel input for the ICA or to preserve the binaural cues in the final output (in the case of binaural hearing aids), while at the same time keeping the interference noise to the minimum level, would be an interesting challenge for future work.

References

[Ama'96] S. Amari and A. Cichocki, "A new learning algorithm for blind signal separation," *In advances in neural information processing systems 8*, pp. 757-763, MIT press, 1996.

[Ara'04a] S. Araki, S. Makino, A. Blin, R. Mukai, H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.(ICASSP), 2004, vol.3, pp. 881 - 884.*

[Ara'04b] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," ," *in Proc. Int. Conf. on Ind. Compon. Anal. (ICA 2004)*, 2004, pp. 898-905.

[Ara'04c] S. Araki, S. Makino, H. Sawada and R. Mukai, "Underdetermined blind separation with directivity pattern based mask and ICA," *in Proc. European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, September 6-10, 2004.

[Bol'00] P. Bofill and M. Zibulevsky, "Blind source separation of more sources than mixtures using sparsity of their short-time fourier transform," *in Proc. 5th Int. Conf. on Ind. Compon. Anal. (ICA 2000)*, 2000, pp. 87-92.

[Bol'01] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Process.*, vol. 81, pp. 2353-2362, 2001.

[Bol'01b] P. Bofill, "Underdetermined blind source separation of delayed sound sources in the frequency domain," technical report, UPC-DAC-2001-14, http://www.ac.upc.es/homes/pau/.

[Cao'96] X. Cao, R. Liu, "General Approach to Blind Source Separation", *IEEE Trans. on Signal Processing*, Vol. 44, No. 3, 1996.

[Car'97] J. F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE letters and Signal Processing*, Vol. 4, pp. 112-114, 1997.

[Els'06] M. Elsabrouty, "Riemannian geometry based blind signal separation using independent component analysis", PH.D thesis, University of Ottawa, 2006.

[Fou'08] J. Foutz, A. Spanias, and M.K. Banavar, *Narrowband direction of arrival estimation for antenna arrays*, Morgan and Claypool Publishers, 2008.

[Fra'98] J.-F. Cardoso, "Blind Signal Separation: Statistical Principles", *Proc. of IEEE*, Vol. 86, No. 10, pp 2009-2025, 1998.

[Gao'06] Y. Gao, M.J. Brennan, P.F. Joseph, "A comparison of time delay estimators for the detection of leak noise signals in plastic water distribution pipes," *Journal of sound and vibration*, vol 292, Issues 3-5, 9 May 2006, pp.552-570.

[Gar'94] B. Gardner, K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *MIT Media Lab Perceptual Computing – Technical report #280*, May 1994.

[Har'05] W.M. Hartmann, B. Rakerd, A. Koller, "Binaural coherence in rooms," *Acta Acustica United with Acustica*, vol. 91, 2005, pp. 451-462.

[Hay'00] S. Haykin, *Unsupervised adaptive filtering, volume I: Blind source separation*, John Wiley &. Sons, Inc., 2000.

[Jad'08] S. D. Jadhav, A. S. Bhalchandra, "Blind Source Separation: Trend of New Age - a Review", *IET International Conference on Wireless, Mobile and Multimedia Networks*, 2008.

[Jou'00] A. Jourjine, S. Rickard, O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.(ICASSP), 2000, vol.5, pp. 2985 - 2988.*

[Kna'76] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. On Acoustics, Speech and Signal Processing,* vol. 24, no. 4, pp. 320-327, 1976.

[Kuh'77] G.F. Kuhn, "Model for the interaural time difference in the azimuthal plane," *Journal of Acoustic Society America*, vol. 82, No. 1, July 1977.

[Li'03] Y. Li, A. Cichocki, S. Amari, "Sparse component analysis for blind source separation with less sensors than sources," *in proc. 4th Int. Symp. on Ind. Compon. Anal. and Blind Signal Separation (ICA 2003)*, Nara, Japan Apr. 2003.

[Li'06] Y. Li, S. Amari, A. Cichocki, D. Ho, and S. Xie, "Underdetermined blind separation based on sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, No. 2, Feb. 2006.

[Loi'07] P.C. Loizou, *Speech enhancement theory and practice*, CRC Press, June 2007.

[Lot'06] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *EURASIP Journal on Applied Signal Processing*, Vol. 2006, pp. 1 – 14.

[Muk'06] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models," *EURASIP Journal on Applied Signal Processing*, pp. 1-13, Vol. 2006.

[Mur'01] N. Murata, S. Ikeda and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, Vol. 41, No. 1-2, pp. 1-24, Oct. 2001.

[Pat'04] E. Fisher, R. kasanmascheff, "Method for improving spatial perception and corresponding hearing apparatus," United States Patent, Patent class: 381 231.

[Pat'07] R.V. Balan, J. Rosca, L. Hong, V. Hamacher, E. Fischer, "System and method for adaptive multi-sensor arrays," United States Patent, Patent No.: US 7218741 B2, May 15, 2007.

[Pat'08a] E. Fisher, "Hearing aid and operating method with switching among different directional characteristics", United States Patent, Patent No.: US 7340073 B2, Mar. 4, 2008.

[Pat'08b] E. Fisher, M. Frohlich, J. Hain, H. Puder and A. Steinbubeta, "Method for operating a hearing aid, and hearing aid," United States Patent Application, application No.: 20080107297 A1, May 8, 2008.

[Pan'07] Q. Pan, "Efficient Blind Speech Signal Separation Using Independent Component Analysis", Ph.D. thesis, University of Ottawa, 2007.

[Ped'08] M. Pedersen, D. Wang, J. Larsen and U. Kjems, "Two microphone separation of speech mixtures," *IEEE Trans. on Neural Networks*, vol. 19, No. 3, Mar. 2008.

[Rej'08] V.G. Reju, S.N. Koh, and I.Y. Soon, "Partial separation method for solving permutation problem in frequency domain blind source separation of speech signals," *Neurocomputing for Vision Research; Advances in Blind Signal Processing*, vol. 71(10-12), pp. 2098-2112 , June 2008.

[Roh'08] T. Rohdenburg, S. Goetze, V. Hohamann, K. Kammeyer and B. Kollmeier, "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.(ICASSP), 2008, pp. 2449-2452.*

[Saw'04] H. Sawada, R. Mukai, S. Araki and S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no.5, Sep. 2004.

[Sar'06] H. Saruwatari, T. Kawamura, T Nishikawa, A. Lee and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 2, pp. 666-678, 2006.

[Sma'98] P. Smaragdis, "Blind separation of convolved sound mixtures in frequency domain," *Proc. Int. Workshop on Independence and Artificial Neural Networks*, Tenerife, Spain, Feb. 1998.

[The'02] F.J. Theis and E.W. Lang, "Formalization of the two-step approach to overcomplete BSS," *in Proc. SIP.*, pp. 207-212, 2002.

[Vee'88] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Trans. Acoustics, Speech Signal Processing*, pp. 4-24, April 1988.

[Vin'04] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," *in proc. Int. Symp. on Ind. Compon. Anal. and Blind Signal Separation (ICA 2004)*, Granada, Spain, Sep. 2004, pp. 327-334.

[Vis'01] H. Viste, G. Evangelista, "Sound source separation: preprocessing for hearing aids and structures audio coding," *in Proc. The COST G-6 Conf. on Digital Audio Effects (DAFX-01)*, Limerick, Ireland, Dec. 6-8, 2001.

Appendix

Female speech # 9 (tr9c1.wav): She had your dark suit in greasy wash water all year. Don't ask me to carry an oily rag like that. His captain was still in Hager his beautiful boots were worn in shabby. The reason for the style seems foolish now. Production may fall far below expectations. Pizza meal is convenient for a quick lunch. Put the butcher on a black table in the garage. Drop five forms in the box before you go out.

Female speech # 10 (tr10c1.wav): The wagons were burning fiercely. He may have a point urgent that Dickinson's have been given fewer prizes. The Honda has a hit due to an outstanding audio visual effect. A lawyer was appointed to execute her will. The toddler found a clamshell near the camp site. We saw a tiny icicle below the roof. Collaborations along with understanding alleviate dispute. She had your dark suit in greasy wash water all year. Don't ask me to carry an oily rag like that.