A Study on Binaural Sound Source Localization based on Equalization-Cancellation Theory

by

Duc Thanh CHAU

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato AKAGI

School of Information Science
Japan Advanced Institute of Science and Technology

August 6, 2014

Abstract

Simulating the human auditory system to deal with problems in sound signal processing is an interesting topic that has attracted a huge number of research recently. One of the important problems is binaural sound source localization (SSL), which plays a crucial role in binaural speech enhancement, binaural source separation and humanoid robot. Although previous research has achieved many impressive results in sound localization, the problem of binaural SSL in the presence of noise and reverberation has not been completely solved. This thesis aims at an effective SSL method based on the human hearing mechanism, which is able to work on binaural systems in practical noisy reverberant environments.

Binaural SSL is an important task in binaural signal processing field as it provides the location of sound source, commonly the direction of arrival (DOA) of the target sound. In the past decades, a large number of DOA estimation methods have been introduced, in which each one differs from others by the way of exploiting two main localization cues: the interaural time difference (ITD) and the interaural level difference (ILD). The well-known conventional GCC-PHAT method is based on only ITD and does not account well for noise. Therefore, there has been many research showing that it is not effective for binaural SSL. Azimuth-dependent models of binaural cues, such as joint estimations of ITD and ILD and DOA classification, have been presented. Although these research showed relatively good results by combining both ITD and ILD, their applicability in adverse noisy reverberant environments is still limited since there has been lack of methods accounting for the effect of interference signals efficiently. Methods directly based on head-related transfer functions (HRTFs) have also been studied, such as the inverse HRTF filtering and the cross-channels HRTFs. However, these methods highly depend on the HRTFs and suffer from reverberation because the HRTFs vary largely along the reverberation levels.

In psychoacoustic research field, binaural hearing has been studied for more than a century and several theoretical models of binaural processing have been developed. Among them, the equalization-cancellation (EC) model of Durlach has received a significant attention as its description is consistent with the human perception on binaural data. The EC model was originally proposed to explain the phenomenon of binaural masking-level

differences (BMLDs) in binaural detection. Due to its well performance on BMLD prediction, the EC model was further extended to selective hearing in the 'cocktail party' scenarios. This suggested that the EC model has great potential for sound localization and segregation in the presence of multiple interference signals.

Inspired by the EC model, this thesis investigates a binaural DOA estimation method based on the EC mechanism. The principle idea is that the EC procedures are first utilized to eliminate the sound signal component at each interest direction; the direction of sound source is then determined as the direction at which the residual energy is minimal. In order to make this idea applicable in practice, two approaches are proposed to accommodate it with the problem of SSL under the effect of noise and reverberation, resulting in two improved algorithms namely Adaptive EC-BEAM and Weighted EC-BEAM. The Adaptive EC-BEAM algorithm improves SSL performance by adapting the EC model to the level of reverberation in room, using the direct-to-reverberant energy ratio (DRR). The Weighted EC-BEAM algorithm deals with the problem in a contradict way, in which two weighted functions are applied to reduce the negative effects from the observed signals, without modifying the localization model. Improvement of the suggested algorithms is verified by experimental results in various noisy reverberant conditions.

The proposed Weighted EC-BEAM algorithm is then selected to apply in two binaural applications, speech enhancement and source separation, as its assumption is easier to be satisfied in practice. In the first application, the proposed method is employed to localize the meaningful sound signals for an intelligent speech enhancement system, which is able to extract and present the meaningful signals together with the target speech. The second application applies the proposed SSL method to estimate the DOAs of all sound sources before extraction (separation), resulting in a new blind source separation method. Experimental results showed that the Weighted EC-BEAM localized the desired sound sources correctly in both applications, from which the effectiveness of the proposed SSL method is confirmed.

Acknowledgments

No guide, no realization. I could not be able to have written this thesis without the significant helps of many people. The first one I would like to thank is my advisor Professor Masato Akagi of Japan Advanced Institute of Science and Technology (JAIST). His invaluable experience in Signal Processing has opened my mind to the world of this interesting field. Standing behind his advising, I always receive precious comments and constant encouragement which guided me through my most difficult time in research.

I would like to thank Associate Professor Masashi Unoki of JAIST for his supportive discussions and suggestions. Be confronted with his challenging questions has furthered my mature in scientific life. I would like to thank Professor Junfeng Li of Institute of Acoustic, Chinese Academy of Sciences for the insightful counsel he has given me throughout my research, which was especially kind of him. I would like to thank Associate Professor Mitsunori Mizumachi of Kyushu Institute of Technology for his precious comments on my research and my thesis so that this study has been improved significantly. I am really grateful to Professor Tu-Bao Ho of JAIST for his supervision of my minor research. More than that, his caring from the very first time that I came to Japan has helped me quickly get accustom to the unfamiliar environment.

I would like to thank my father, Thanh-Phu Chau, who I admire for his talent and intelligence. His kind-hearted behaving towards everyone serves as a shining example for me to follow.

I would like to thank my mother, Thi-The Nguyen, who does not know how to read but has successfully taught us to be strong and independent. Her courage and striving in life were the inspiration for me to never give up.

I give the deepest appreciation to my wife, Hai-Minh Nguyen. It was only her love and empathy that tolerant the strenuous work came to our lives.

I would like to acknowledge my sisters and brothers who have been always bringing me true happiness and great encouragements.

I wish to express my gratitude to the Graduate Research Program (GRP) of JAIST that supported me during my 3 years of doctoral study. The program has provided excellent conditions that allowed me to accomplish my research goals successfully. During my time as a graduate student, I was also supported by the A3 Foresight program that sponsors my presentations in many domestic and international conferences.

Last but not least, I would like to give special mention to Assistant Professor Ryota Miyauchi, Mr. Yasuhiro Hamada and Ms. Rieko Kubo for our conversations that I found completely inspiring. I appreciate so much all lab members for their helps during my time living in Japan. They set a tone in the lab of friendly cooperation, which makes it possible to get more work done and makes the work itself more enjoyable. I also wish to thank all of my trustworthy friends for their sharing and encouragements during my everyday life.

I devote my sincere thanks and appreciation to all of them.

Contents

List of Figures

List of Tables

Glossary

Notations

t	time index
ω	frequency bin index
$s(t)/S(\omega)$	target sound emitted at the source in the time / frequency domain
$x_i(t)/X_i(\omega)$	target sound observed at the microphone i
$m_i(t)$	masker component observed at the microphone i
$h_i(t)/H_i(\omega)$	transfer function at the microphone i
$r_i(t)/R_i(\omega)$	reverberation observed at the microphone i
$R_i^E(\omega)$	early reverberation at the microphone i
$R_i^L(\omega)$	late reverberation at the microphone i
$N_i(\omega)$	background noise observed at the microphone i
$y(t)/Y(\omega)$	total observed sound signal the presence of noise and/or reverberation
ϕ	direction of target source
$\widehat{\phi}$	estimation of ϕ
D	set of interest directions on the horizontal plane to look for the target source
θ	steering direction to look for sound source, $\theta \in D$
C	set of candidate directions, $C \subset D$
$W(\omega)$	equalizer
$W_{Adaptive}(\omega)$	adaptive equalizer in reverberant conditions
$C_X(\theta)$	cancellation output applying on signal X at direction θ
$C_X^N(\theta)$	cancellation output of signal X with consideration of noise
$C_X^R(\theta)$	cancellation output of signal X with reverberation of noise
$C_{Adaptive}(\theta)$	cancellation output of the Adaptive EC-BEAM
$C_{Weighted}(\theta)$	cancellation output of the Weighted EC-BEAM
T_R/T_{60}	time of reverberation
T_θ	threshold for minimal angle between two sources

T_E	threshold for maximal residual energy at a source
$\psi(\omega)$	impulse response representing reverberation based on the direct component
β	exponent presenting the corresponding equalizer in reverberant condition
$\widehat{\beta}$	approximation of β
$Q(\omega)$	conceptual component representing the effect of reverberation on equalizer
$\nu(\omega)$	coefficient to present $Q(\omega)$ based on the reverberation
$\kappa(\theta)$	weighting function characterizing for the distribution of noise
Γ	function presenting corresponding equalizer in reverberant condition
$Z_i(\omega, t)$	noise estimation at microphone i
$E[.]$	expectation/mean operator
Φ_{XX}	auto-power spectral density of X
Φ_{XY}	cross-power spectral density of X and Y
$G(\omega)$	gain function of speech enhancer
ξ	a priori SNR
$B_i(\omega)$	compensation factor for noise estimation at microphone i
$B_i^{opt}(\omega)$	optimal $B_i(\omega)$
arg	parameter operator
max/min	maximum/minimum operator
$\forall$	for all
*	complex conjugate
$\circledast$	convolution operation
j	imaginary unit: $\sqrt{-1}$
μ	step size for training equalizer

Acroyms and Abbreviations

ASA	auditory sense analysis
AEE	average estimation error
ANN	artificial neural network
BMLD	binaural masking-level difference
BSS	blind source separation
CASA	computational auditory scene analysis
CC	cross-correlation
CMA	constant modulus algorithm
DCA	dependent component analysis
DOA	direction of arrival
DRR	direct-to-reverberant energy ratio
DUET	degenerate unmixing estimation technique
EC	equalization-cancellation
GCC	generalized cross-correlation
HRIR	head-related impulse response
HRTF	head-related transfer function
ICA	Independent component analysis
IID	interaural intensity difference
ILD	interaural level difference
IPD	interaural phase difference
ITD	interaural time difference
JADE	joint approximate diagonalization eigen-matrices
LSD	log-spectral distance
MMA	minimal audible angle
MTR	meaningful-to-target energy ratio
NLMS	normalized least mean squared
PCA	principal components analysis
PESQ	perceptual estimation of speech quality
PHAT	phase transform
SE	standard error
SNR	signal-to-noise energy ratio
SSL	sound source localization
STFT	short-time Fourier transform
SVD	singular value decomposition

TS-BASE	two-stage binaural speech enhancement
VAD	voice activity detection
WSS	wide sense stationary

Chapter 1

Introduction

Human beings can perceive sound perfectly in many undesired environments, even in a high noise and high reverberant condition. When multiple speakers are active concurrently, we are able to selectively attend to an individual speaker, despite the disturbance of other speakers. A huge number of studies have been investigated to understand the processing mechanism underlying the human abilities, giving birth to the series of research on 'Auditory Sense Analysis' (ASA) and perception models [1, 3]. This has motivated a lot of research to simulate the human auditory model to deal with various problems in the field of sound signal processing and understanding.

One of the most important abilities of the human auditory system is sound source localization (SSL), which enables us to specify where the sound comes from. Source location information is important in daily life since it improves speech perception in communication and helps us react properly in each situation, e.g. avoiding a car when hearing its engine sound. In sound processing applications, sound localization is also needed as a front-end in a various systems, e.g. a humanoid robot which can communicate and mimic human behaviors. It is impressive that a human can localize sound sources correctly in various environments by using only two ears, while this task is still a challenge in simulating systems, especially when noise and reverberation are present. This inspires an idea to solve the problem of SSL by adopting the processing mechanism of human under this ability.

1.1 Binaural sound source localization

In daily life, we can specify locations of speakers easily by listening to their sounds, without looking around. This ability is referred to as the ability of sound source localization. Binaural SSL is the task of SSL using only two sensors (ears or microphones). The

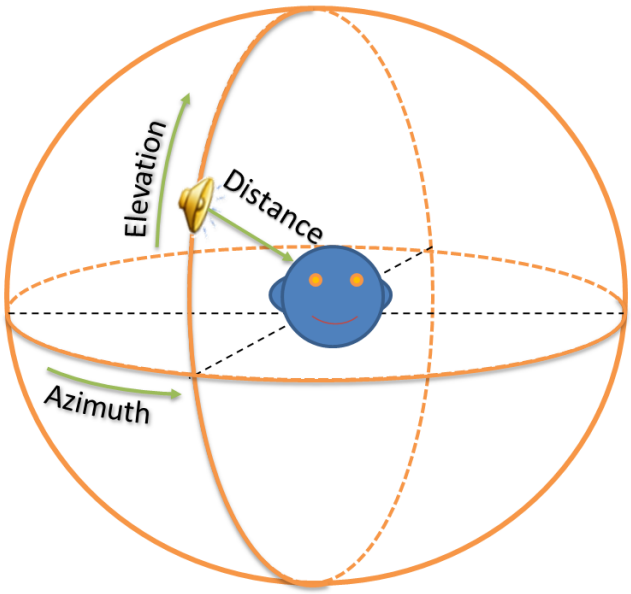

Figure 1.1: Binaural sound source localization.

location of a sound source can be represented by three factors: azimuth, elevation and distance (as illustrated in Fig. 1.1). Humans localize the azimuth based on two main cues: interaural time difference (ITD) due to the difference of distances from the source to two ears, and interaural level difference (ILD) caused by the difference in amplitude of sounds observed at both ears [2, 3]. ITD plays a major role for frequencies below 800 Hz, while in higher frequencies, ILD is more important. The elevation of sound sources is discriminated by sound spectra (monaural cue), which vary in the vertical plane because of the effect of head, outer ears and other parts to the propagation of sound. Binaural cues and monaural cues are individual-dependent due to the difference of the head's size, in which the spectral cue is quite sensitive to this difference. Distance of sound source is related to a number of factors, such as the loss of amplitude, the loss of high frequencies and the direct-to-reverberant energy ratio (DRR), in which DRR is supposed to be the most important cue in distance perception. The main factors related to SSL are briefly summarized in Table 1.1. Due to the great attention on the direction of sound sources in practical applications, sound localization is commonly understood as the task of determining the direction of arrival (DOA) of the observed sound. Consequently, this thesis is targeted at DOA estimation on the horizontal plane (azimuth), and the term SSL is also used with this meaning within this study.

In signal processing field, SSL is an important front-end processing of a number of applications. For speech enhancement on binaural and microphone-array systems, DOA of the target sound is required to perform enhancement [4]. If the DOA is specified incor-

Table 1.1: Summary of sound localization.

Dimension	*Main cues*
Azimuth (Horizontal)	-Interaural time/phase difference -Interaural level/intensity difference
Elevation (Vertical)	-Sound spectrum
Distance	-Direct-to-reverberant energy ratio -Loss of amplitude -Loss of high frequencies

rectly, noise may be enhanced instead of the target sound. In human-robot communication systems, DOA of the speaker is required to enable a robot to mimic some basic human-like behaviors. For instance, the robot is expected to face to the speaker when it is called [5]. So far, a large number of SSL methods have been introduced [3, 6], which may be generally categorized into three approaches: monaural, binaural and microphone-array. The monaural approach mainly localizes sound on the vertical plane using the spectral cue. Monaural localization on the horizontal plane has also been investigated [7], though its performance is till far from a practical application. The approach based on microphone-array achieves the highest precision and is quite robust against undesired factors in rooms, such as noise and reverberations since it has the advantage of spatial information. However, this advantage normally goes along with a large-size microphone-array and a high complexity system, which prevent its applicability from real-time applications. Furthermore, general microphone-array based methods are also not effective on binaural systems because they are not specialized for the binaural effect, such as the obstacle of head and outer ears. Nowadays, along with the development of binaural systems with real-time processing tasks, there has been a growing need of binaural sound localization as a front-end for these systems. Despite the limit of spatial information, by understanding and imitating human localization ability, the binaural approach for localization is promising increasingly.

1.2 Challenges of binaural SSL in noisy reverberant environments

Binaural SSL in practical environments is a challenging task because of various factors affecting sound propagation and perception, which can be generally divided into two main groups: internal effect and external effect.

The internal effect of a binaural system includes all the effects caused by the system

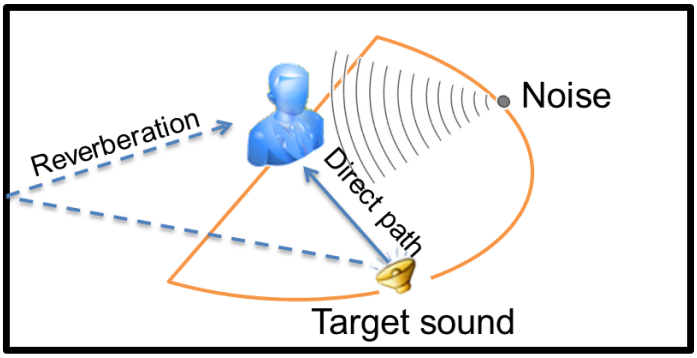

Figure 1.2: Challenges of binaural SSL in real environments.

itself, such as the shadow effect and internal noise. The internal noise may appear in movable systems and will not be considered in the cope of this study as it is not the general case. The shadow effect is generated by the obstacle of the pinna, head and shoulder, that prevents the sounds from directly reaching the receivers, and normally referred to as the *head related transfer functions* (HRTFs). This is one of the main factors making SSL on a binaural system more difficult than on a microphone-array, besides the limitation of input (only two receivers). For this reason, it is necessary to learn the shadow effect before performing SSL on a binaural system, such as directly measuring the HRTFs [8, 9] or indirectly learning this effect when building perception models [10, 11].

Two crucial factors that cause the external effects are noise and reverberation (Fig. 1.2). Reverberation is present in any enclosed space because of the reflection of sound on the walls, floor, ceiling and other objects. Reverberation can be conceptually divided into two components: early reverberation and late reverberation. Early reverberation includes strong reflections arriving the ears within a time t_0 (commonly $t_0 = 50ms$) after the direct path. Since sound can be reflected by any objects at any direction in rooms, the effect of early reverberation is similar to the effect of sounds observed from multiple sources. Late reverberation contains all reflections arriving after t_0. Each reflection of this component is weaker than that of early reverberation; however, as the reverberation time is high, the total late reverberation comparably affects sound localization and perception [12]. Total reverberation generally tends to make the sound diffuse and reduces sound localization ability. It is commonly observed that the sources in reverberant rooms seem to be larger and difficult to be localized correctly.

Noise is sound produced by any source other than the target source which is being localized. Noise exists in almost all practical environments. Generally, noise includes the background noise, whose location and power are stable in time and can be assumed as stationary (such as noise from fans and air-conditioners), and the foreground noise,

which is caused by the suddenly occurring sounds. SSL methods commonly deal with only background noise because the foreground noise is not easy to be distinguished from the target sound and is related to the problem of tracking and separation. When there are multiple noise sources with similar characteristics in a room, the background noise tends to diffuse. However, in some special case, a noise source may be observed with larger power and different characteristic in comparison with the other sources, it will be directional noise. This kind of noise makes the problem of SSL more challenging when its energy is comparable or bigger than that of the target source.

In summary, corrupting factors affecting localization ability include HRTFs, noise and reverberation. HRTFs make the localization cues more different from those observed in normal microphone array. Reverberation smears the target signals and tends to make it diffuse. Noise increases sound energy at directions other than the target source, especially a directional noise. So far, a large number of binaural SSL methods have been introduced, in which each one differs from the others by the assumptions of HRTF, noise or reverberations [3, 6]. However, an SSL method which fully considers all of these factors has been still being researched.

1.3 Motivation and research goals

The excellent performance of the human auditory system has attracted a large number of research. Many binaural interaction models have been proposed to simulate the mechanism of binaural hearing, in which two seminal models are the *coincidences* model of Jeffress [13] and the *equalization-cancellation* (EC) model of Durlarch [1, 14]. The other models are supposed to be derived from one (or in some cases both) of these models (see [3]). The coincidences model is commonly realized as *cross-correlation* (CC) model, which was the principle of the well-known *generalized cross-correlation* (GCC) method [15] and various GCC-based algorithms (e.g. [16, 17, 18, 19]). Although the GCC function is a very good ITD estimator, it is not able to account for the ILD of the target signal. This is the reason why GCC-based methods are less effective in the case of localization in binaural systems. The EC model was originally proposed to explain the mechanism of binaural masking level differences (BMLDs) applied in binaural detection in noise [14]. It imitates human auditory system in a simple way, in which when a subject is presented with a binaural stimulus (target) masked by another one (masker), the auditory system attempts to eliminate the masking components by first transforming the stimuli presented to the two ears so as to equalize the masking components related to ITD and ILD (the E operation), then subtracting the equalized stimulus at an ear to that at the other ear

(the C operation). In this procedure, both the ITD and ILD contribute to the process of masker elimination. Due to its ability to explain the BMLDs phenomenon, the EC model has been extended to selective hearing in the presence of multiple sound signals, which is usually referred to as the 'cocktail party effect' [20]. This revealed the potential of the EC model for the problem of SSL in the presence of multiple interference signals, especially in the manner of binaural systems, where the ITD and ILD are both important to specify the direction of sound source. However, there has been no research considering these advantages into the issue of binaural SSL so far.

Inspired by the EC model, this thesis investigates a binaural SSL approach based on the EC mechanism. With the goal of SSL in a practical binaural system, the main focus of this study will be put on the engineering aspect to adopt the EC model to tackle the existing problems of binaural SSL, i.e. HRTF effects, noise and reverberation, rather than to develop the model as a theory for SSL in spychoacoustic field. In order to achieve this goal, the proposed approach is expected to satisfy two following objectives:

(1) Effectiveness with binaural systems, and

(2) Robustness under noisy reverberant environments.

The first objective is related to the issue of how an SSL method can efficiently exploit the internal effect from a binaural system. As mentioned in Section 1.2, the internal effect is caused by the obstacles of head, shoulder and pinna, which block the sound from directly reaching the ear drums. Although the internal effect corrupts the binaural cues, e.g. ITD and ILD, and makes them different from those in a normal microphone array, it emphasizes the cues, leading to the fact that the azimuths are more distinguishable. A general SSL method, such as GCC-PHAT, may be not effective on such kinds of systems because it is not able to adapt this effect. There has been research dealing with this issue by assuming that the HRTFs measures are available [8, 9]. In this way, the internal effect can be fully exploited. However, their applicability may be limited because accurately measuring HRTFs in an arbitrary system would not be an easy work. The SSL method that this thesis aims at is expected to make fully use of the internal effect, but simple enough for implementing in practical systems.

The second objective is related to the robustness against the external effects, in which the main focus is put on the background noise and reverberation. The background noise may consist of not only diffuse noise but also directional noise, which makes the problem of SSL more complicated, and is rarely considered in previous SSL research. Since these noises are quite popular in practice, both of them are considered within this thesis. We limit the background noise to stationary noise, or at least its energy changes much slower

than that of the target signal. For reverberation, we consider early reverberation as sounds observed from multiple sources, in which the energy of each source is lower that of the target source, while late reverberation can be modeled as diffuse random signals. With these assumptions, the targeted SSL method is expected to exploit the characteristics of the above components for robustness in noisy reverberant environments.

In short, this thesis is motivated by the directional hearing ability of the human auditory system with regard to the well explanation of the EC model on binaural hearing phenomena, and targeted at the issue of sound localization in binaural systems. The main goal is to develop a new SSL approach based on the EC model to deal with the existing problems of binaural SSL. To achieve this goal, the proposed approach must be applicable in practical systems, which can be evaluated by two objectives: effectiveness with a binaural system, and robustness again noise and reverberation.

1.4 Thesis outline

This dissertation consists of 5 chapters, in which the remaining ones are organized as follows.

Chapter 2 provides some important background knowledge related to the issues in the following chapters. First, several basic concepts underlying directional hearing are introduced. The EC model is presented in detail to understand the mechanism on which a new SSL approach is proposed in Chapter 3. An overview of binaural sound localization is then given. From this overview, position of the proposed localization approach can be recognized among the current trends of SSL. The last section of this chapter further discusses the challenges of binaural SSL as well as to reveal the promising direction that this study is investigating.

Chapter 3 is the main chapter in this dissertation, which presents the proposed SSL approach based on the EC model. This chapter begins with a general binaural SSL algorithm that integrates the EC mechanism into a beamforming technique (named as EC-BEAM), and the experimental results to evaluate its effectiveness regarding to the first objective of this thesis. Following that, the problems of EC-BEAM in practical environments are analyzed and two approaches are proposed to make the EC-BEAM meet the second objective, i.e robustness against noise and reverberation. The first approach is to make EC-BEAM robust against reverberation by adapting the EC model to the reverberation level in room (Adaptive EC-BEAM). The second approach is to make EC-BEAM robust against both noise and reverberation in an opposite way, in which the EC-model is kept unchanged while two weighting functions are applied to eliminate the

effects of noise and reverberation on EC procedures (Weighted EC-BEAM). Although both approaches improve the performance of EC-BEAM in noisy reverberant environments, only Weighted EC-BEAM is used in Chapter 4 because its assumptions are easier to be satisfied in practice. Experiments are carried out to verify the performance of each approach and an insightful discussion of each approach is also given.

Chapter 4 further verifies the SSL approach proposed in Chapter 3 via applications. The proposed Weighted EC-BEAM is selected to apply in binaural speech enhancement for hearing aids and blind source separation. In the application of speech enhancement, the proposed SSL method is applied to localize the important sounds other than the target sound, which are called meaningful signals. This consideration is to enable the system to extract and present the important sounds together with the target at the final output. For blind source separation, the proposed Weighted EC-BEAM is applied to detect the DOAs of all concurrent sources. The competing sound signals are then separated based on the detected DOAs without knowing the number of sources. Experimental results are also presented to verify the localization ability of the proposed method in these applications.

Finally, Chapter 5 summarizes this thesis, points out the contributions and suggests some directions for further research.

Chapter 2

Background

One can easily recognize that listening to sounds using two ears is much better than doing it with only one ear alone, especially in the presence of multiple concurrent speakers. This comes from the advantages of binaural hearing, which enable us to localize and separate sound sources, as well as to improve speech intelligibility. So far, much research has been investigated to study the mechanisms related to the binaural hearing abilities. The trend of this topic may be loosely remarked from the 'duplex theory' of Lord Rayleigh [21]. Later studies have presented a number of binaural auditory interaction models, e.g., coincidence-counting, equalization-cancellation (EC) and their derivations (see the reviews in [2, 3, 22, 23, 24]). These models have achieved some successes when applied in binaural signal processing systems. Although the coincidence counting mechanism proposed by Jeffress, realized as cross-correlation in recent research, has dominated the models of binaural interaction, the EC model developed by Durlach has retained its appeal for many years due to its conceptual simplicity and its ability to predict binaural data. Nowadays, it has still being applied to deal with a wide range of signal processing problems, including speech intelligibility prediction, target detection, speech enhancement and distance estimation [1, 4, 14, 25, 26, 27].

Along with the development of binaural model, binaural SSL have also been extensively researched. Significant attention has been paid on the series of binaural SSL based on the generalized cross-correlation (GCC) function, such as GCC, GCC-PHAT and their derivations [15, 28]. These methods localize the sound source by exploiting the ITD of the target signal. Later research improved the SSL performance by combining the ITD and ILD [10], or even the monaural cue [11] in a statistical model. Other research approached the problem indirectly via machine learning, such as artificial network, or based on the measures of HRTFs. All of these methods were successful in several experimental conditions but still remained some disadvantages that need to be further improved for

practical applications.

With the purpose to provide some background knowledge related to the problem of binaural sound localization, this chapter is organized with two main sections: binaural hearing and binaural SSL. The first section discusses the basic factors underlying binaural hearing, along with an appreciation of the binaural processing ability of the EC model in noisy acoustical environments, and in the presence of competing speakers. This section begins with the binaural cues and the BMLD since they are the most important concepts to understand the mechanism of the binaural phenomena. A brief overview of binaural interaction models is given, from which the position of the EC model can be realized. Following that is a subsection that clearly describes the procedure of the EC model for binaural processing in 'cocktail party' conditions. Attention will be paid mainly on the logical concepts of binaural hearing rather than the physiologic aspect. The second section of this chapter concentrates on the current development of binaural SSL. An overview of binaural SSL is provided to summarize methods investigated in the field. The limitation of these methods is also analyzed, which is one of the motivations that this study is investigated.

2.1 Binaural hearing

Binaural hearing, usually mentioned as directional hearing, provides a number of abilities in the real world. Among them, two important ones are localization and selectivity. Sound localization ability allows a subject to determine the directions of sound sources in the space which is benefit in communication or helps people to react properly in urgent cases. Humans can discriminate a difference in the azimuth (horizontal plane) of approximately 2^o and of approximately 7^o in the vertical plane [29, 30]. Selective listening enables us to pay attention to individual sound sources in a cluttered acoustical environment. For instance, it aids a subject in listening to one speaker in an environment where several speakers are speaking at the same time, which is usually mentioned as "the cocktail party effect" [20]. In addition, the binaural system plays a major role in improving speech intelligibility in noisy and reverberant environments. The advantages of binaural hearing involve a number of binaural interactions, in which the major factors are binaural cues and binaural masking-level differences.

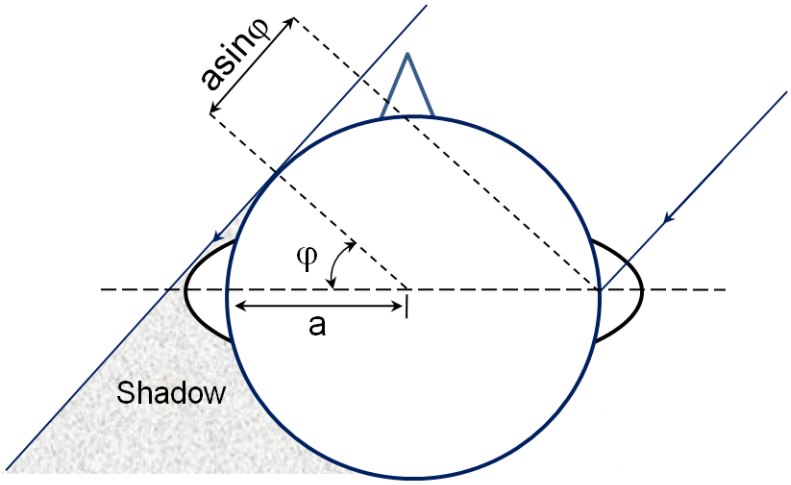

Figure 2.1: Binaural cues: The ITD comes from the fact that sounds arrive at the farther ear are slightly later than those at the closer one; the ILD appears because the head blocks a part of sound from reaching the farther ear.

2.1.1 Binaural cues

Many researches have been investigated to specify the factors affecting the spatial aspects of how a sound is perceived. It had not been remarked until Lord Rayleigh introduced the duplex theory [21] for a comprehensive analysis of the physics of binaural perception. Though this theory was too simple to completely explain the mechanism of directional hearing, it remains basically valid until now. According to the duplex theory, two physical cues dominate the perceived location of an incoming sound source due to the differences in time and intensity of the sounds received at the left and right ears, unless the sound source is located directly in front of or behind the head. As illustrated in Fig. 2.1, sounds arrives slightly earlier in time with higher intensity at the ear that is closer to the source.

The difference in arrival time of the sound at two ears is usually referred as interaural time difference (ITD) or interaural phase difference (IPD) when measured as the phase shift of a sinusoidal tone. This originates from the fact that it takes longer for the sound to arrive at the ear that is farther from the source. Classical studies model the head as a solid sphere, such as Woodworth model [31], and result in a ITD range from 0 μsec (sound source at 0^o, i.e., in front of the listener) to 660 μsec (sound source at 90^o). These findings are consistent with the study of Feddersen *et al.* [32]. However, recent research with well-established model found that the ITDs below about 500 Hz are on the order of 800-820 μsec in stead of 660 μsec as in classical model, and the ITDs for the higher

frequencies came closer to the 660 μsec value of the classical study [33]. These results are also supported by the research of Roth *et al.* [34] and confirmed by the data of Bronkhorst and Plomp [35].

The interaural intensity difference (IID), also referred as the interaural level difference (ILD) when measured in dB, is produced because of the shadowing effect of the head that prevents a part of the incoming sound energy from reaching the ears, in which the farther ear receives a slightly lower energy than the closer one. Similarly to ITD, the IID approaches the maximum value when the sound source is at 90^o and decreases to 0 as the direction of the sound source moves to the front of the listener. IID is also a function of frequency, in which the IIDs are negligible at 200 Hz and increase with frequency, reaching about 20 dB at 6000 Hz.

In terms of sound localization, the ITD and IID cues operate in complementary ranges of frequencies. Low frequencies that have wavelengths larger than the path around the head tend to bend over the head to the far ear, while higher frequencies have wavelengths smaller than the head, so they are blocked in the path to that ear. Therefore, the ITDs are useful localization cues in low frequencies (below about 1.5 kHz) and IIDs are pronounced at higher frequencies. Moreover, as both ITD and IID vary in frequencies, the binaural cues information in individual frequency bands are likely more useful than those in overall.

2.1.2 Binaural masking-level differences

In a condition that more than one sound are concurrently present, the auditory system may fail to detect the presence of a sound source due to its low level in comparison with other sound sources. The sound source is said to be masked by other sound sources. Masking is a phenomenon that the reception of a specified set of the acoustic stimuli ("targets") is degraded by the presence of other stimuli ("maskers"). Consider a scenario of target detection with a target signal S and a masker (noise) N, the subscript 'm' denotes the monaural condition, 'o' denotes the binaural condition in which the inputs at the left and right ears are identical and 'τ' denotes the binaural condition in which the phase difference of the inputs at the left and right ears is τ. Thus, S_mN_m indicates that the signal and noise are presented to one ear, and S_oN_o means that the same signal and the same noise are simultaneously presented to both ears. These conditions can be considered as the starting points. The target signal is more audible in the case the phase difference of either target or noise is not zero, i.e., $S_\tau N_o$ or $S_o N_\tau$. This is the phenomenon illustrating the masking level difference in binaural hearing.

Binaural masking-level difference (BMLD) is defined as the difference in detection

threshold between monaural condition and a binaural condition. Mathematically,

$$BMLD = Threshold(S_m N_m) - Threshold(S_\tau N_o) \tag{2.1}$$

BMLD is a popular concept describing the benefit of binaural hearing in the presence of multiple concurrent sound or 'cocktail party effect' [20]. It is traditionally used to explain the mechanism of binaural sound (target) detection and extensively used in some other related issues, such as speech intelligibility prediction (see the reviews in [1, 2, 22, 36]).

The value of BMLD varies depending on the relative difference between IPDs of the target signal and the masker component. Reported in the study of Green *et al.*, the BMLD gets the largest value at about 15 dB when either the signal or the noise is opposite at two ears ($S_\pi N_o$ or $S_o N_\pi$) [37]. Its value decreases to 0 dB as the phases of both signal components are identical ($S_o N_o$) . In a room condition, the amount of masking strongly depends on the position of both sound sources. BMLD generally tends to increase when the difference in azimuth of the sources gets larger, especially with white noise masker [38]. As a result, the BMLD gets the maximum as the difference in azimuth of the two sources are 180^o and gets minimum as both sources locate at the same direction. At the same condition concerning IPD of both the target and masker, the BMLD is larger as the spectrum level of the masking noise is increased, especially when the noise is identical at both ears [39, 40].

The BMLD is also dependent on the stimulus frequency in the summary of Durlach [1], as shown in Fig. 2.2. Its value is largest for low frequencies with the size of about 15 dB at 250 Hz, and decreases as the frequency gets higher until a stable value at 3 dB is obtained at the frequencies about 1500-2000 Hz. This may be because the BMLD is highly dependent on the relative IPDs of both signal components. Moreover, as mentioned in Section 2.1.1, the role of IPD is reduced at high frequencies. This finding has been supported by a number of studies, e.g., [14, 41, 42, 43].

The BMLD may be obtained differently with various kinds of maskers. Robinson and Jeffress used both correlated (same at both ears) and uncorrelated noise in their experiments to specify how noise correlation affects the value of BMLD [44]. They found that the BMLD from uncorrelated noise is on the order of 3-4 dB and BMLD becomes larger as the degree of correlation decreases. From this result, an interesting hypothesis that the human ability to overcome the effect of reverberation because of the uncorrelation was introduced. This issue is further supported by Koenig *et al.* [45], who found that room reverberation decorrelates the noise at two ears, resulting in an BMLD of about 3 dB.

In summary, the BMLD comes from the differences in binaural cues (e.g., ITD and

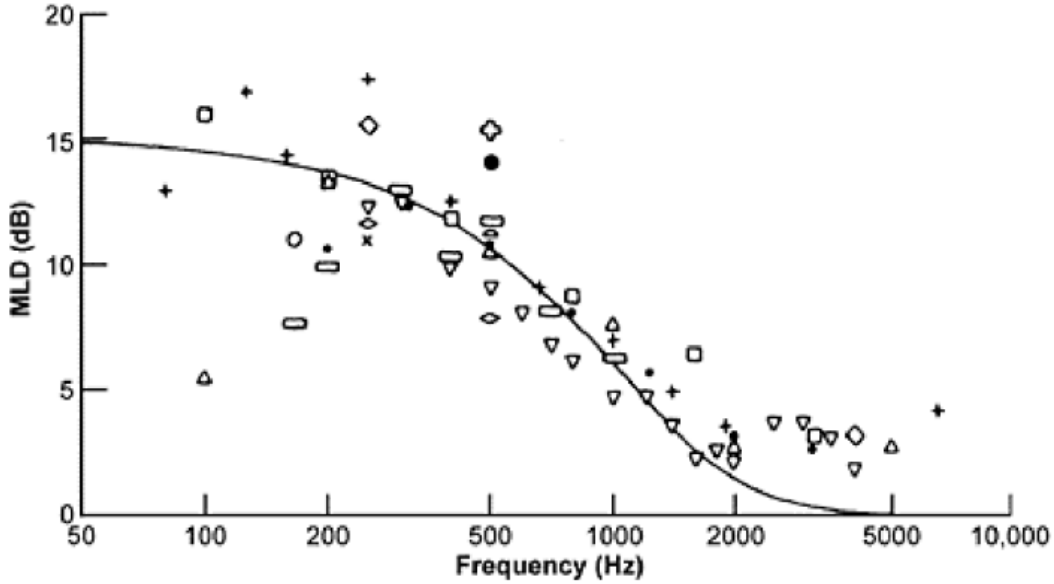

Figure 2.2: BMLD as a function of frequency for many studies, summarized by Durlach [1].

ILD) of the target and the masker. Its value is dependent on the relative IPDs of the two signal components, frequency and the correlation level of the masker at two ears. BMLD is the important concept to describe the mechanism of binaural separation of the human auditory system in the 'cocktail party' scenario.

2.1.3 Binaural interaction models

Inspired by the impressive abilities of human auditory system in processing and perceiving sound, many auditory theorists have been investigated into the field to model the processing mechanism behind these abilities. After the duplex theory of Rayleigh [21], modern modeling researches mostly came from the suggestion of 'coincidence theory' of Jeffress [13] and the concept of binaural masking-level differences introduced by Hirsh and Licklider [46, 47]. There were a number of studies that verified and explored these works in a variety of subjective experiments, which were summarized in the review of Hafter and Trahiotis [48]. In addition, significant efforts have been made to describe these data in terms of quantitative models, e.g., the cross-correlation (CC) model formulated by Sayers and Cherry [49]. Some other researches highlighted the interaural time delay produced by a vectorial combination of the target and masker components, namely 'vector model', in a series of papers and connected the concept with a number of experiments in binaural detections [42, 50, 51, 52]. Among the variety of proposed models, the equalization-cancellation model of Durlach [1, 14] and the model based on auditory-nerve activity of Colburn [53, 54] are two of the most noticed models during this period. Although it

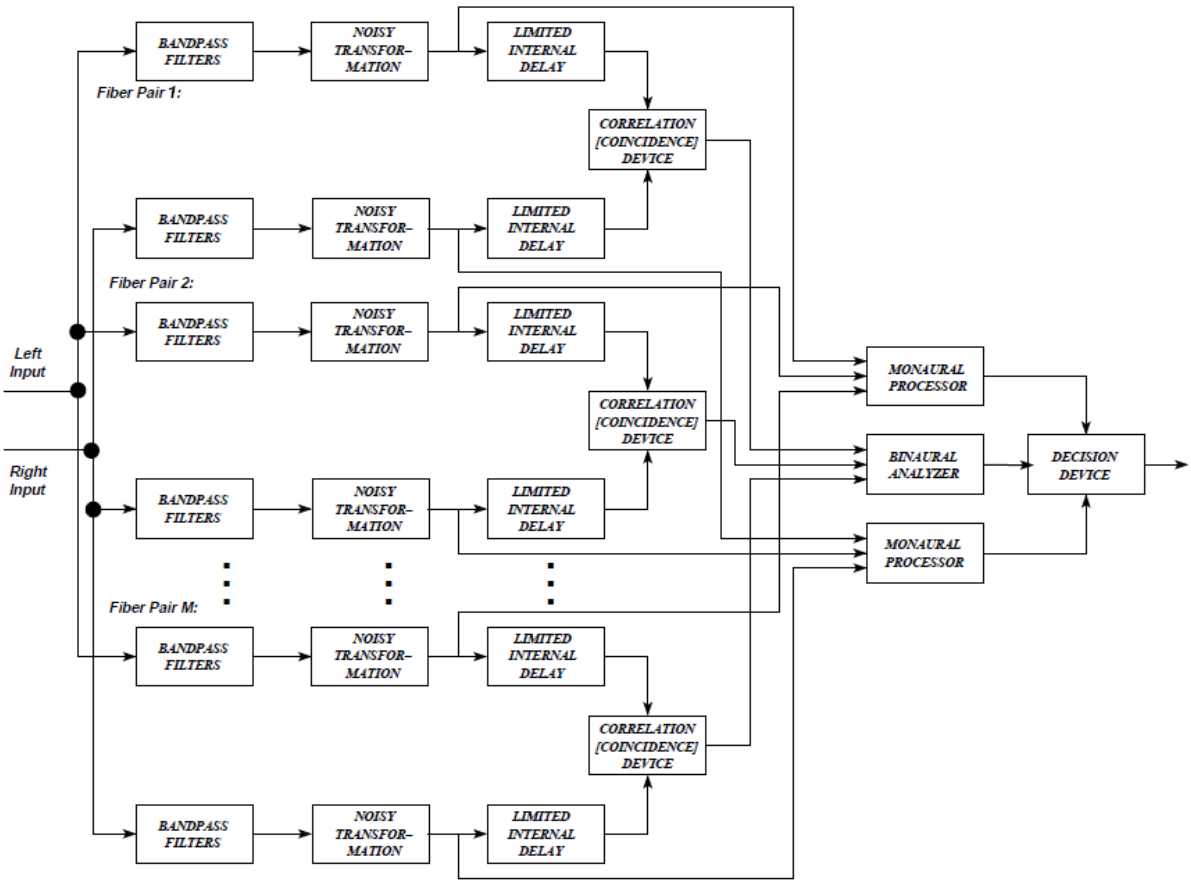

Figure 2.3: Generic binaural interaction model suggested by Colburn and Durlach [2].

has been argued that these two models are linear functions of each other, Breebaart *et al.* showed that they are not equivalent as their predictions in the case of finite-length stimuli are different [55]. Excellent review can be found in the work of Colburn and Durlach [2], in which most of the major theoretical models and experimental results are summarized.

Also in the summary of Colburn and Durlach [2], a generic model of binaural interaction was suggested (Fig. 2.3) and it is believed that all the current models are its realizations. This generic structure includes a series of peripheral processing steps (bandpass filtering, rectification and stochastic neural representation of the signals), the binaural analysis (with interaural timing information, internal delays using a correlation or coincidence mechanism and interaural intensity differences), and a decision-making component to select the best processing between monaural and binaural paths at final outputs. Most of the recent models can be traced to be a derivation of this generic model. Two typical example models, which are dominant in the field and are still popularly used nowadays are the EC model and the cross-correlation model. The cross-correlation model can be considered as an implementation of the coincident model of Jeffress. Essentially, the CC

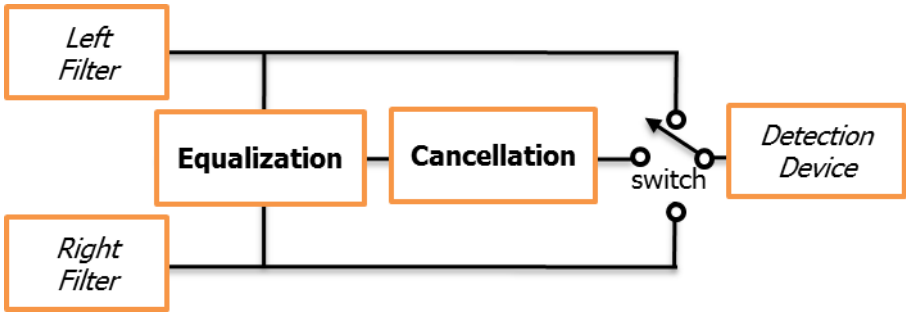

Figure 2.4: Conceptual model of equalization-cancellation theory.

model analyzes the binaural inputs by exploiting their similarity; the EC model, on the other hand, exploiting their dissimilarity. In recent years, the findings on the human auditory system have been widely applied in both simulation and reality. Binaural interaction models have become the central processing in a huge number of applications such as speech recognition systems [56, 57], virtual environments and displays [7, 23], speech intelligibility prediction [25, 26, 58], speech enhancement [4, 59], sound localization [11, 15, 16, 19, 60], source distance estimation [27, 61], etc.

2.1.4 Equalization-Cancellation model

The EC model was originally developed by Durlach [14] for binaural signal detection. This model is one of the bases on which the generic model described in Section 2.1.3 is developed. The structure of the model is illustrated in Fig. 2.4, including the preprocessing, binaural computation and decision device to select the best processing between monaural and binaural paths. Within this thesis, only the binaural path is concentrated on since we aim to exploit the binaural advantages.

Consider a scenario that the target speech $x(t)$ is competed or masked by a binaural stimulus $m(t)$ (masker).

$$y_i(t) = x_i(t) + m_i(t),\ i = L, R \tag{2.2}$$

where $y_i(t)$ is the total observed signal. In the traditional EC model, it is assumed that when subject is presented with the masker, the auditory system is able to calibrate the optimal ITD τ and optimal ILD α to make the masker components at two ears 'equalized'(the E process), i.e. $m_L(t) = \alpha m_R(t - \tau)$. Detection of the target signal is obtained by subtracting the 'equalized' signal at one ear from that of the other ear (the

C process).

$$EC(\tau, \alpha) = y_L(t) - \alpha y_R(t - \tau) = x_L(t) - \alpha x_R(t - \tau). \tag{2.3}$$

It can be observed that in the final output of Eq. (2.3), the masker component was completely eliminated, remaining only the residual of target signal. Moreover, the residue of target is dependent on the compensating parameters τ and α. If the difference in azimuth between target source and noise source is small, the target signal components at the two ears may also be equalized. As a result, the residue of target signal at the output is low. In contrary, if the difference of the two sources get larger, the target signal components at the two ears are much different and the output is higher. This observation shows that the output of cancellation is consistent with the amount of binaural masking discussed in section 2.1.2 Therefore, the EC model has extensively used to describe BMLD phenomena and to predict the masking amount in various of binaural data in various studies. This model was further improved by Culling and Summerfiled [62] in which the EC parameters are selected independently in each frequency band.

Among the variety of research based on EC model, there is an interesting branch that the EC model is employed to eliminate the target signal instead of the masking components. Typically, in binaural speech enhancement, Li *et al.* employed the EC model to remove the target signal component in the stage of noise estimation; then the estimated noise is used to construct a gain function to enhance the target speech using Wiener filter theory [4]. A similar mechanism was applied in the distance estimation study of Lu *et al.*, in which the EC model is used to eliminate the direct-path component of reverberant speech, remaining only reverberation component; following that, DRR is obtained based on the estimated reverberation and distance of sound source is specified by matching the DRR to a prior trained model [27]. The success of these research makes the EC model more powerful. It can be generalized that either the target and the masker can be individually eliminated from a 'cocktail party' mixture signal, depending on the curtain purpose. This ability is quite similar to the selective hearing ability, which enables human to pay attention to individual conversation in the condition that multiple speakers are concurrent active. Furthermore, this suggests that the EC model is promising for the issue of sound localization in the presence of multiple sound sources.

2.2 Binaural SSL

General SSL methods can be classified into three groups based on the available input: monaural SSL, binaural SSL, and microphone-array SSL. The monaural SSL group includes methods using only one microphone integrated in a robot ear. Most of the monaural SSL studies aim to estimate the DOA of sound in terms of elevation since it is well-known that elevation of the sound source is specified by the sound spectrum [2], which is a monaural cue. Recent research have extended the monaural methods to azimuth localization and provided evidences showing that such kinds of SSL is possible, such as the study of Wightman and Kistler [7]. Due to the limitation of spatial information, such as ITD and ILD, these results are still far from real applications. Binaural SSL related to the methods that determine the source location in a binaural system, normally a dummy head with HRTF effects. This approach has received a lot of attention in the field because it is the natural approach that mimics the localization ability of human beings. Microphone-array approach is related to the SSL methods using multiple microphones without shadow effects. This approach achieves the highest performance among the three approaches due to the advantage spatial information. However, this advantage normally comes with a high computational complexity and so is not suitable with real-time systems, especially the binaural applications. As this thesis concentrates on binaural SSL, only binaural approach is discussed within this section.

Most of the binaural SSL methods exclusively focus on localization in azimuth rather than in elevation and distance due to the popular need of azimuthal DOA in binaural applications. A huge attention was paid on the cross-correlation (CC) technique, which is a realization of the coincidence model developed by Jeffress. The trend of the CC based SSL was considerably remarked since Knapp and Carter [15] introduced the generalized cross-correlation framework. Methods based on this mechanism produce point estimates of the time delay between two microphones by including a weighting function in a cross-correlation calculation. A popular technique to improve this framework is the Phase Transform (PHAT), which whitens the two signals before cross-correlating to provide a more sharply peaked correlation. The general formula of GCC-PHAT is given by:

$$GCC(\tau) = \int_{-\infty}^{\infty} \frac{X_L(\omega)X_R(\omega)^* e^{j\omega\tau}}{|X_L(\omega)X_R(\omega)^*|} d\omega. \tag{2.4}$$

Although PHAT was originally used as an ad-hoc method for robustness against noise with the assumptions of uncorrelated, stationary, Gaussian sources, this method has showed an

impressive performance in many conditions. In a later research, by using those assumptions, Zhang *et al.* showed that PHAT is a kind of maximized likelihood function [63]. The GCC-PHAT method was further extended to multiple microphones systems in the study of Dibiase, giving birth to a more general SSL algorithm, named as SRP-PHAT [28]. The formula of SRP-PHAT is described as follows:

$$SRP(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{X_i(\omega)X_j^*(\omega)}{|X_i(\omega)X_j^*(\omega)|}e^{j\omega(\tau_i-\tau_j)}d\omega = \int_{-\infty}^{\infty}\left|\sum_{i=1}^{N}\frac{X_i(\omega)}{|X_i(\omega)|}e^{j\omega\tau_i}\right|^2 d\omega. \quad (2.5)$$

This derives another form of GCC-PHAT by considering two-microphone version of SRP-PHAT,

$$GCC(\theta) = \int_{-\infty}^{\infty}\left|\frac{X_L(\omega)}{|X_L(\omega)|}e^{j\omega\tau_L} + \frac{X_R(\omega)}{|X_R(\omega)|}e^{j\omega\tau_R}\right|^2 d\omega. \quad (2.6)$$

From this form, it is easily realized that SRP-PHAT is a kind of whitened delay-and-sum beamforming method. There should be no doubt that GCC-PHAT and SRP-PHAT are effective methods in the uncorrelated noise conditions. However, these methods are less robust in reverberant conditions because of the increase of bias, variance, and spurious detections [64]. Moreover, since this method is based on ITD only, there has been an analysis showing that it suffers from the binaural setups [65].

In order to effectively localize a sound source in a binaural system, various azimuth-dependent models have been built. Andersson *et al.* [66] and Raspaud *et al.* [10] investigated to explicitly join the estimations of ITD and ILD. Other research exploited these cues in an implicit way, such as machine learning. Berglund *et al.* [67] extracted binaural cues in a feature vector and mapped it into the location factors using an artificial neural network (ANN). In this way, the problem of SSL is transformed into the problem of machine learning, which is quite powerful and rapidly developed recent years. Although these methods have shown relatively good performance in the ideal conditions, there has been lack of evidence that such kinds of methods can perform in a practical environment where noise and reverberation may significantly corrupt the sound attributes. More recently, in the localization model of Woodruff *et al.* [11], interference signals were taken into consideration in a joint ITD-ILD statistical model, in which the undesired factors are characterized by the direct-to-residual energy ratio. They achieved significant improvement in comparison with previous methods in the experimental conditions. However, the applicability of this method is still remained as a difficult problem as such kinds of information is normally not available in practice.

Concerning SSL under the effect of HRTF, one of the effective approaches to tackle the problem is using HRTFs directly. In a binaural system with a dummy head, the sound observed at an ear differs from that at the other ear because of the differences in the propagation path, which are characterized by HRTFs. Therefore, by exploiting the propagation informtion in HRTFs, the direction of a sound source can be obtained. McDonald *et al.* [8] introduced a binaural SSL based on the observation that the travel path of the signals at two ears would be the same if they are convoluted with the HRTF of the opposite ear, which can be described as follows:

$$x_L(\phi, t) \circledast h_R(\phi) = h_L(\phi) \circledast s(t) \circledast h_R(\phi) = h_L(\phi) \circledast x_R(\phi, t), \tag{2.7}$$

where $\circledast$ denotes the convolution operator, $x_i(\phi, t)$ and $h_i(\phi)$ are respectively the observed signal and the HRTF (in the time domain) at the receiver i ($i = L, R$), provided that the source $s(t)$ locates at the direction ϕ. As a result, the DOA is specified via looking for the pair of HRTFs minimizing the dissimilarity of the cross HRTFs, that is

$$\begin{aligned} &\widehat{\phi}(t) = \underset{\theta}{\operatorname{argmin}}\, e(\theta, t) \quad \text{where} \\ &e(\theta, t) = \sum_t \left[x_L(t) \circledast h_R(\theta) - x_R(t) \circledast h_L(\theta)\right]^2 \end{aligned} \tag{2.8}$$

Keyrouz *et al.* [9] proposed a binaural SSL based on a similar mechanism, in which the inverse of HRTF at each ear is used as a filter to recover the original sound emitted at the source:

$$\begin{aligned} s(t) &= x_L(t, \phi) \circledast h_L^{-1}(\phi) \\ &= x_R(t, \phi) \circledast h_R^{-1}(\phi) \end{aligned} \tag{2.9}$$

where $h_L^{-1}(\phi)$ and $h_R^{-1}(\phi)$ are respectively the inverses of the left and right HRTFs. In this way, the pair of inverse HRTFs corresponding to the direction of the sound source should provide the most identical 'recovered' signals. They further improved this idea by reducing the number of points of HRTF, leading to the reduction of the computational cost for real-time systems. The effectiveness of these methods was verified in the experimental conditions with uncorrelated noise and/or low reverberations. However, in adverse conditions where noise can be directional or reverberation is high, these method may fail to detect the true sound sources. In addition, as this approach strictly relies on the HRTF information, their applicability may be limited in practice because accurately measuring HRTFs for an arbitrary system is a time-consuming work. An efficient and applicable

method should not only be able to exploit HRTF information but also be robust against the undesired factors, such as noise and reverberation, and be simple enough to be carried out in common systems.

2.3 Summary

Human auditory system has inspired a huge number of research into the field to understand the mechanism of binaural hearing, especially the localization and selective hearing abilities. Research revealed that these abilities are related to the processing of binaural cues and the binaural masking level differences. Two important models that mimic the human hearing mechanism are the cross-correlation and the equalization-cancellation. The cross-correlation model was employed in various sound localization methods with relatively good performance, in which the well-known method is GCC-PHAT. However, as the CC-based SSL methods exploited only ITD information, they are not effective on binaural systems. The EC model has successfully explained the BMLD, which is an important phenomenon to understand selective hearing in cocktail party effect. This revealed the potential of the EC model for binaural processing tasks, particularly for sound localization.

The main difficulties of binaural SSL come from the effects of HRTF, noise and reverberation. On the development of binaural SSL, the GCC method and its derivations have achieved relatively good performance. However, they seem not effective with HRTF effects since only the ITD was exploited. In fact, HRTF may become an advantage if its information is prior known because it emphasizes the binaural cues, leading to the fact that the spatial sources are more distinguished. Various azimuth dependent models for SSL have been studied to adapt this effect, such as joint ITD-ILD, ANN based SSL. So far, there is still lack of an effective method that is robust against the external effects, such as noise and revereberation. The approach based on HRTF, e.g. cross-channels HRTF and inverse HRTF, also faces the same problem although in this approach, HRTF information is more effectively exploited. For practical applications, binaural SSL methods should be able to fully make use of HRTF information and robust against noise and reverberation.

Chapter 3

Proposed binaural SSL approach based on EC Theory

Through out the series of binaural hearing research, the EC model has shown its effectiveness in explaining binaural phenomena, especially the BMLD. This inspires an idea to apply the EC model to the problem of binaural SSL. It should be noticed that in the original EC model, the target signal is detected by suppressing the noise masker. In practical environments, since noise may consist of diffuse noise and multi-source noise, it is difficult to suppress all the noise to look for the sound source. Instead, the target source can be eliminated, remaining the sounds from other directions. The DOA of the target source is specified via looking for the direction with minimal remained energy, which was mentioned as "steering the null" technique in previous research [3].

The goal of this chapter is to develop a new binaural sound localization method based on the EC model, in which it will focus on the engineering aspect with regard to the two objectives of this thesis, i.e. effectiveness with a binaural system and robustness in noisy reverberant environments. In the first section, a general algorithm related to the first objective is designed by integrating the EC model into a beamforming technique, namely EC-BEAM. The following section addresses the problems of the EC-BEAM in the presence of noise and reverberation. Two approaches are suggested to independently deal with these problems to make the proposed method meet the second objective, resulting in two separate derivative algorithms of the EC-BEAM. The first approach considers reverberation as the main factor degrading the SSL performance, and attempts to overcome it by adapting the EC model to the reverberation condition (Adaptive EC-BEAM). This approach is partly motivated by the finding of Shin-Cunningham [68], which showed that the human auditory system is able to 'learn' and adapt a room condition to improve the perception ability. The second approach is based on a hypothesis that the human

auditory system may improve the perception ability by some additional processings based on the different characteristics between the target signal (such as speech) and undesired factors (background noise and reverberation). Following this hypothesis, the perception model (the EC model) is kept unchanged, while two weighting functions are suggested to eliminate the effects of noise and reverberation on EC procedures (Weighted EC-BEAM). The main difference between these two approaches is that one adapts the localization model to the observed data while the other processes the observed data to fit the current model. Discussions on the advantages and disadvantages of each approach are given after the experimental results.

3.1 General principle: EC-BEAM

The general principle of the proposed approach is to integrate the EC model into a beamforming technique, namely EC-BEAM. A null is steered to each interest direction by using the EC operations to eliminate the signal component from that direction. Once the null is steered to the true sound source, the residual energy should be minimal. This principle is applied in this section in two scenarios: localization of one source and localization of multiple sources.

3.1.1 Localization of a single source

Given a set D of interest directions on the horizontal plane, for each direction $\theta \in D$, an equalizer is constructed so that if the source is located at θ, the signals observed from the source can be (or at least approximately) equalized,

$$S_L(\omega, t) - W(\omega) S_R(\omega, t) \approx 0, \text{ or } W(\omega) = \frac{S_L(\omega, t)}{S_R(\omega, t)}, \tag{3.1}$$

where ω and t denote the frequency bin index and the frame index respectively, $X_L(\omega, \theta, t)$ and $X_R(\omega, \theta, t)$ are the *short-time Fourier transforms* (STFTs) of the signals observed from the source at the left and right receivers, and $W(\omega, \theta)$ is the equalizer at the direction θ. Essentially, $W(\omega, \theta)$ represents the IPD and ILD of the sound components observed from the source. In the frequency domain, the signal observed at each receiver is related to its transfer function by (see Appendix A):

$$X_i(\omega, \theta, t) = H_i(\omega, \theta) S(\omega, t), \quad i = L, R, \tag{3.2}$$

where $S(\omega, t)$ and $H_i(\omega, \theta)$ are the sound signal emitted at the source and the transfer function representing the propagation of sound from the source to each receiver respectively. Therefore, the equalizer in Eq. (3.1) can be rewritten as:

$$W(\omega, \theta) = \frac{H_L(\omega, \theta)}{H_R(\omega, \theta)}. \tag{3.3}$$

From Eq. (3.3), $W(\omega, \theta)$ is specified by only the transfer functions at θ and independent of the sound signals. In the case the receivers are integrated in a dummy head, $H_i(\omega, \theta)$ becomes the *head-related* transfer function and $W(\omega, \theta)$ is equivalent to the concept of *interaural transfer function* in binaural hearing studies [3].

The equalizers are obtained by pre-training in anechoic condition where only one sound source is present at each interest direction. Each equalizer is constructed independently in frequency bands, which is consistent with the modified EC model suggested by Culling and Summerfield [62]. The signal-independence property of equalizer guarantees that an equalizer trained with some sound signal is able to perform with other unknown sound signals. In training, the equalizers are calibrated using *normalized least mean square* (NLMS) method, which is given by:

$$W_{t+1} = W_t + \mu \frac{X_R^*(t)}{|X_R(t)|^2} \left[X_L(t) - W_t X_R(t) \right], \tag{3.4}$$

where ω and θ are omitted for easy reading, the superscript * denotes the conjugate operator and μ is a scalar value specifying the step size. The training stage will be further described in Section 3.1.3.

In the stage of DOA estimation, in order to localize the sound source at an unknown direction ϕ, a null is steered to each interest direction by using EC operations to eliminate the target signal components. After a cancellation process through all interest directions, DOA of the target sound is specified as the direction at which the residual energy of the null is minimal, as shown in Fig. 3.1. That is:

$$\widehat{\phi}(t) = \underset{\theta \in D}{\operatorname{argmin}}\, C_X(\theta, t), \quad \text{where}$$

$$C_X(\theta, t) = \int_{-\infty}^{\infty} |X_L(\omega, \phi, t) - W(\omega, \theta) X_R(\omega, \phi, t)|^2 d\omega. \tag{3.5}$$

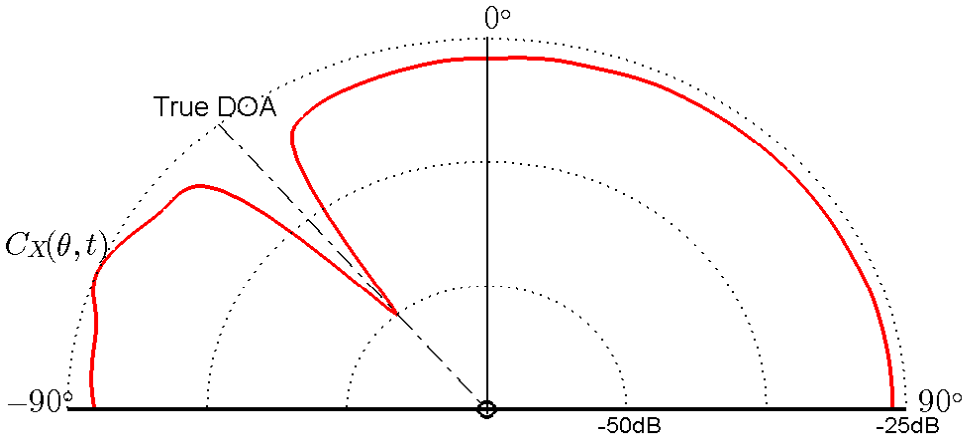

Figure 3.1: Illustration of null steering in the EC-BEAM. The target source is located at $-40°$ in clean-anechoic condition.

3.1.2 Localization of multiple sources

The EC-BEAM algorithm described in Section 3.1.1 localizes only one source with the hypothesis that the direction of source corresponds to the minimum of the residual energy of the null steering process. By observation, this hypothesis is also valid in the case multiple sources are present simultaneously. In this section, the EC-BEAM algorithm is extended to multiple sources localization, namely mEC-BEAM, without the requirement of knowing the number of sources. In order to achieve this goal, the DOA estimates must satisfy following criteria:

- Local minimum of null energy.

  As the EC model is able to eliminate sound from a given direction, if the null is steered to a true sound source, its energy should be lower than that of the neighbour null. Therefore, a candidate angle must be the angle with a local minimum of null energy but not necessary a global minimum as in the original EC-BEAM. Mathematically, the set of candidate angles C ($C \subset D$) of mEC-BEAM are specified as follows:

  $$C = \{\theta_i | C_X(\theta_i) \leq C_X(\theta_{i-1}) \text{ and } C_X(\theta_i) \leq C_X(\theta_{i+1})\}. \tag{3.6}$$

- Separate from other candidates.

  Under the effects of noise and reverberation, there may appear a number of local minima of null energy which are very close to each other. However, research on *minimal audible angle* (MAA) [29, 30, 69] revealed that in order to separate two

sources, the angle between them must be large enough. To exclude these unreliable candidates, we set up a threshold T_θ so that:

$$\forall \theta_i, \theta_j \in C, |\theta_i - \theta_j| \geq T_\theta. \tag{3.7}$$

- Low null's energy.

  The trend of null's energy through a beamforming process may produce many pseudo local minima; however, they do not correspond to any actual sources. In the case the number of sound sources N is known, these local minima are not important because only N best candidates are considered. However, in the case of blind source detection, these local minima must be excluded. A threshold T_E is set up to avoid these minima by

$$\forall \theta_i \in C, C_X(\theta_i) \leq T_E \tag{3.8}$$

The above criteria are to make mEC-BEAM be able to localize multiple concurrent sound sources. However, depending on the situation that the mEC-BEAM is carried out, some of those criteria or maybe all of them will be used.

3.1.3 Performance evaluation

Configuration

The proposed EC-BEAM algorithm was evaluated under simulated conditions, including clean-anechoic, noisy-anechoic, clean-reverberant and noisy-reverberant conditions. To generate directional sounds, sound utterances were convoluted with head-related impulse responses (HRIRs). The HRIRs were simulated using the ROOMSIM package [70] with a room size of $10 \times 10 \times 3\ m^3$ and an identical absorption coefficient for all the room surfaces. The ROOMSIM software utilizes the HRTF measurements obtained by KEMAR dummy head [71] to generate the HRIRs, using the image method [72]. HRIRs produced in this way are expected to represent reliable simulations since the software uses real measured HRTFs. Reverberation time (T_{60}) was set from 0 to 0.6 s (step of 0.2 s) while source distance was fixed at 3 m. The direction of sound source varied in horizontal plane from -90^o (left side of dummy) to 90^o (right side of dummy) with a step size of 10^o.

Training was executed in a clean-anechoic condition, where only one source was present at each interest direction respectively. Since the equalizer is independent of the training data, white noise was used as the sound utterance for training as its energy distributes

over all frequency bins. Testing was performed with five speech sentences selected from the ATR Japanese Database [73], which have an average length of 10 seconds. Background noise was simulated with directional noise and diffuse noise, where the directional noise is the sound of a fan at 40^o while the diffuse noise was created by convolving the sound recorded from an air-conditioner with HRIRs at all directions and then summing them all together. Energies of the two noises were controlled to be about the same in the mixture. Each DOA estimate was obtained using a frame length of 0.5 s.

DOA estimation performance was evaluated via *average estimation error* (AEE) and *error rate*. The error rate is the ratio of wrong estimates over the total estimates,

$$\text{Error rate} = \frac{\text{Number of wrong estimates}}{\text{Total estimates}}, \tag{3.9}$$

where an estimate is defined as a wrong estimate if its error is over 10^o threshold, i.e, $|DOA_{estimated} - DOA_{real}| > 10^o$.

Evaluation of training

The equalizer at each interest direction was trained following Eq. (3.4), in which μ was set to 0.01 while $W(\omega)$ was initialized by 1 and updated along the frame index t. Note that to obtain the equalizer $W(\omega)$ satisfying Eq. (3.1), a relatively long training data may be required. Instead, a short observed signal was used to train the equalizer repeatedly for a certain number of iterations until it is converged. Specifically, each equalizer was obtained using a 3-second white noise. Fig. 3.2 shows the error rates of the DOA estimation on the clean speech along the number of iterations for training. It can be observed that when the number of iterations is increased, the error rate is reduced because the equalizers are converged and satisfy Eq. (3.1). From Fig. 3.2, the equalizers are totally converged when the number of iterations is higher than 80. Therefore, in the following sections, the number of iterations for training is empirically set to 100.

Evaluation of performance

Fig. 3.3 shows the average estimation error of the EC-BEAM at each direction in anechoic conditions with/without noise. It is observed that in low noise conditions (SNR>0 dB), the estimation error at frontal directions (i.e. $[-50^o,50^o]$) is quite low while that at the side directions (especially the directions that are close to -90^o or 90^o) is relatively higher. These results are consistent with the results of binaural hearing research [29, 30, 69, 74] since the azimuth change at the side area results in a smaller change in binaural cues, e.g., IPD and ILD, than that at the middle. Therefore, the effect of noise is more emphasized

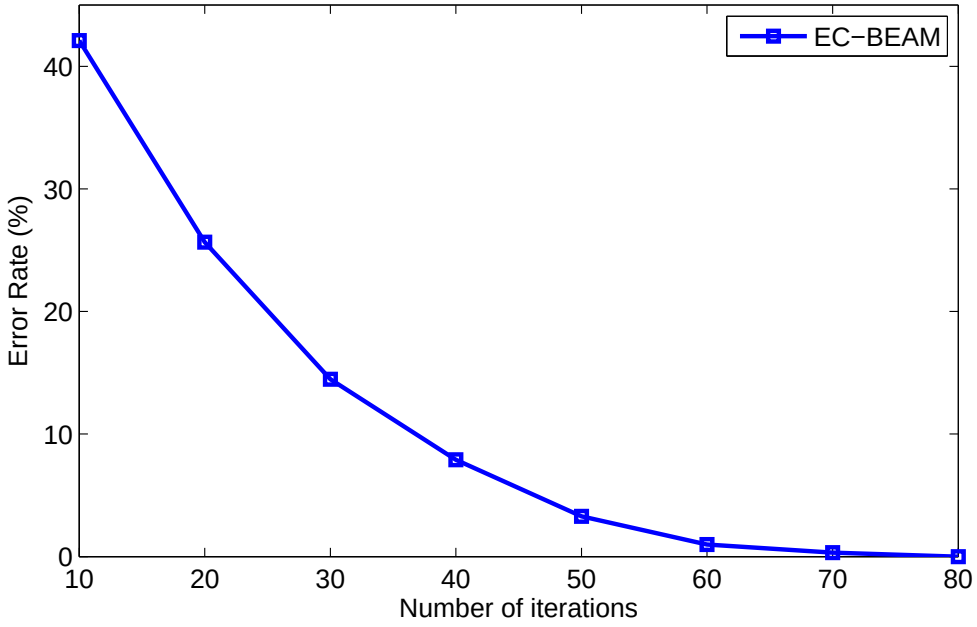

Figure 3.2: Convergent of equalizers

in this area. When the level of noise is increased, the estimation error also increases; however, the increment of estimation error is very small in the conditions that SNR is equal or larger than 5 dB. Especially, the estimation performance in the condition of 10 dB SNR is quite comparable to that in the clean condition. This indicates that the EC-BEAM can perform well in low noise conditions, but it suffers from high noise conditions.

Table 3.1: Summary of estimates performed by EC-BEAM in reverberant conditions.

T_{60} (s)	Average Error (°)	Standard Deviation	Error rate (%)
0 s	0.2	1.4	0.0
0.2 s	3.8	6.8	9.9
0.4 s	8.9	11.5	20.4
0.6 s	12.9	16.3	29.9

Similar results were obtained in the presence of reverberation, as summarized in Table 3.1. In anechoic and low reverberation conditions, the EC-BEAM can estimate the direction of the sound source quite well. When reverberation time gets higher, the estimation error is also increased as the effect of reflection is more serious.

The overall error rate of the EC-BEAM the conditions that both noise and reverberation are concurrently present is plotted in Fig. 3.4, and the detail values are shown in Table 3.2. It can be observed that the error rate increases in the case that either noise or reverberation is high. In noisy reverberant conditions, the error rate is higher than that

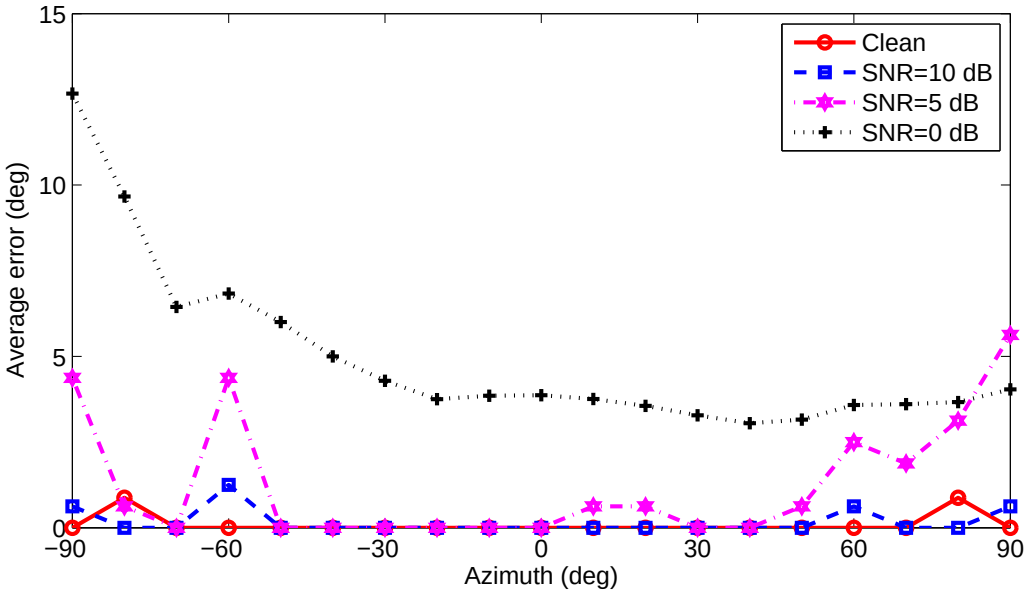

Figure 3.3: Average estimation errors of EC-BEAM in anechoic with/without noise conditions.

in the cases noise or reverberation is present alone. In summary, the proposed EC-BEAM method is able to work with binaural data in low noise and low reverberation conditions. In high noise and/or high reverberation conditions, the performance of the EC-BEAM is degraded and seems not reliable.

Table 3.2: Average error rate (%) of the EC-BEAM in noisy reverberant conditions.

T60/SNR	0 s	0.2 s	0.4 s	0.6 s
15 dB	0.0	9.2	20.4	30.9
10 dB	0.0	11.2	22.0	34.5
5 dB	1.3	17.8	29.9	39.8
0 dB	13.5	33.2	47.0	56.3

3.1.4 Discussion

The EC-BEAM algorithm has been proposed to take the EC model into the problem of binaural SSL. The experimental results showed that this method performed quite well in low noise and low reverberation conditions. This is because the equalizer trained in advance is a representation of the interaural transfer function, which fully accounts for the ITD and ILD of the target signal. From these results, it is believed that the proposed EC-BEAM can work effectively with binaural systems, which implies that the first goal

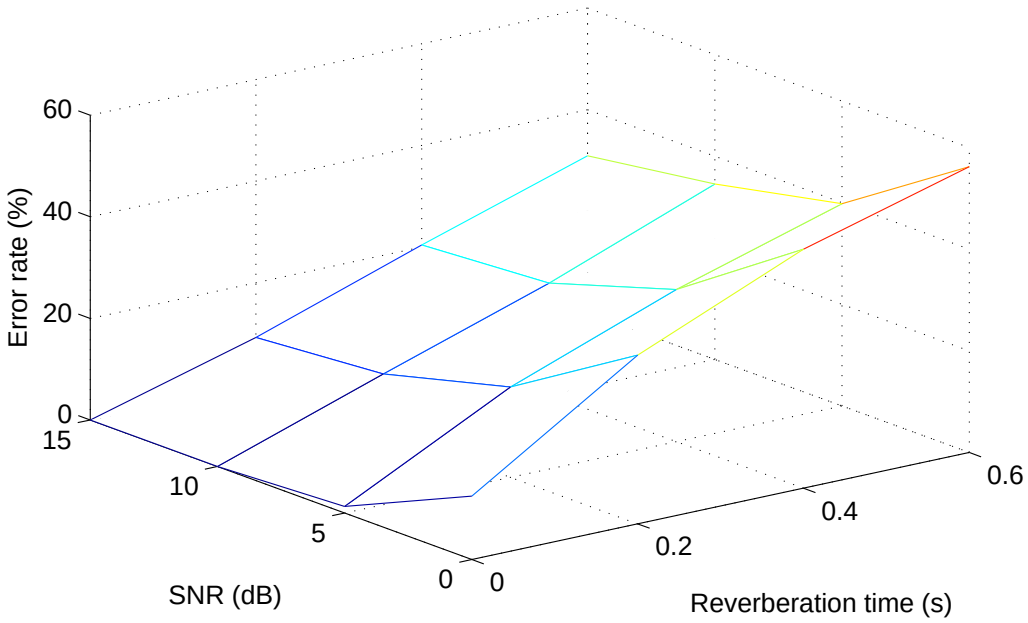

Figure 3.4: Overall error rate of the EC-BEAM in noisy reverberant conditions.

of this thesis has been achieved. However, it can be observed that the performance of the EC-BEAM is degraded quickly when noise or reverberation gets higher. This indicates that the EC-BEAM does not meet the second objective of this thesis, i.e., the robustness in adverse noisy reverberant environments. Therefore, further investigation is needed to improve its performance in these conditions and to achieve the goals of this thesis.

3.2 Problems of EC-BEAM in practical conditions

Binaural SSL in practice generally have difficulties of internal effect and external effect (see Section 1.2). Internal effect is mainly because head and ears block the sound from directly reaching the receivers, which is characterized by the HRTFs. The equalizer of EC-BEAM can be considered as a kind of filter adapting to the binaural system. Therefore, the EC-BEAM algorithm should account well for this effect as long as the equalizer is trained properly. However, the EC-BEAM is not able to account for the external effects, i.e. noise and reverberation, because they are unknown and unpredictable.

In enclose spaces, reverberation decorrelates the target source, leading to the degradation of directional perception. The mechanism is that sounds emitted from the source propagate to arbitrary directions and reflected by walls, floor, ceiling and other objects in room. The reflected sounds differ from the direct-path sound in both amplitude and time

delay. Theoretically, these sounds can be expressed as the delayed and decayed versions of direct sound. The total reverberation can be represented as follows (see Appendix A):

$$R(\omega) = \int_{0^+}^{\infty} X(\omega)\alpha(\tau)e^{-j\omega\tau}d\tau \tag{3.10}$$

where τ and $\alpha_i(\tau)$ are respectively the time delay after the direct-path and the decay coefficient due to the absorption of air, walls and other objects in room. Since reflected sounds decay in time, their effects are different depending on how long they are remained in room. In signal modeling, total reverberation is divided into two main components: early reverberation and late reverberation.

$$R(\omega) = R^E(\omega) + R^L(\omega) \tag{3.11}$$

Early reverberation arrives the ear within a time threshold t_0 after the direct-path sound (t_0 is commonly selected from $50ms$ to $80ms$). Although these reflections are weaker that the direct-path sound, their effect is still comparable and can be considered as sounds obtained from other sources, which has lower energy than that of the current source. Late reverberation can be modeled as uncorrelated random signal [75], whose impulse response is presented by $h(t) = b(t)e^{-\sigma t},\ t \geq 0$, where $b(t)$ is a zero-mean Gaussian stationary noise and σ is related to the reverberation time T_R by:

$$\sigma = \frac{3\ln 10}{T_R} \tag{3.12}$$

In the modeling, late reverberation is not correlated with direct signal and early reverberation, and together uncorrelated at both ears. With both early and late reverberation, the EC mechanism may fail to cancel completely the target signal component, leading to the degradation of performance.

In addition to reverberation, background noise reduces performance of EC-BEAM though it is not correlated with target signal. It is because noise increases sound energy at non-source directions, especially directional noise. Specifically, suppose that EC-BEAM succeeds in elimination of target source, if the energy of noise is higher than that of the target signal, the residual energy may remains high and direction of source may not be localized correctly.

In summary, observed signal in practical conditions may consist of not only target signal (direct path) but also its reverberation and background noise.

$$Y(\omega) = X(\omega) + R(\omega) + N(\omega) \tag{3.13}$$

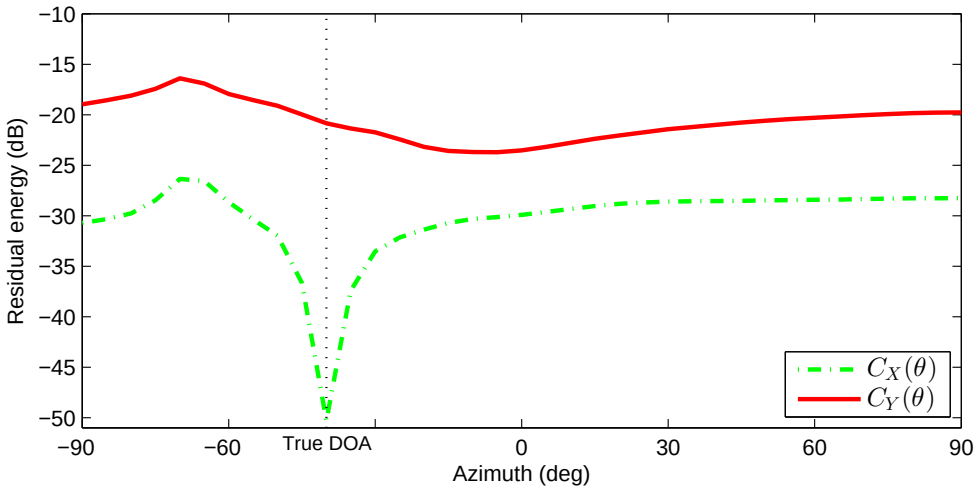

Figure 3.5: Residual energy of cancellation performed in clean-anechoic condition and noisy reverberant condition ($T_{60} = 0.5$s, SNR = 10 dB). Target source is fixed at $-40°$ in both conditions.

Early reverberation is correlated with target signal while late reverberation is not correlated with it, and also not correlated at both channels. Noise is not correlated with target signal, however, it may be correlated at both channels in case of directional noise. Cancellation of target signal at each interest direction $\theta \in D$ is performed as follows:

$$\begin{aligned} C_Y(\theta) &= \int_{-\infty}^{\infty} |Y_L(\omega) - W(\omega,\theta) Y_R(\omega)|^2 \, d\omega \\ &= \int_{-\infty}^{\infty} |[X_L(\omega) + R_L(\omega) + N_L(\omega)] - W(\omega,\theta)\,[X_R(\omega) + R_R(\omega) + N_R(\omega)]|^2 \, d\omega. \end{aligned} \tag{3.14}$$

As a result, cancellation operation is not applied to only target signal but also to reverberation and noise. The residues of cancellation in clean-anechoic and noisy revebrerant conditions are shown in Fig. 3.5. It can be observed that the residual energy of the steered null to true direction does not drop to minimum in noisy reverberant condition, which indicates that EC-BEAM fail to detect true sound source.

3.3 Approach 1: Adapting EC-BEAM to reverberant condition

The performance of the EC-BEAM has been verified in several noisy conditions in Section 3.1.3. However, it failed to localize the sound source in noisy reverberant environments, as showed in Section 3.2. One of the big problems can be recognized is the effect of reverberation. Since the equalizer was trained in anechoic data while the applied condition is reverberant room, there may be model mismatch, which leads to the result that the target source cannot be eliminated completely. In the series of research on hearing, it is showed that the perception of humans at later estimates is better than that of the initials [68, 76, 77]. This phenomenon occurs in various experimental conditions, especially in reverberant rooms, and is popularly referred to as 'room-learning'. Essentially, learning makes the auditory system familiar with the conditions which is equivalent to an adjustment of the model's parameters to fit the trial data. In order to make the EC-BEAM work in the certain conditions, the EC model must adapt to reverberation so that it can cancel the target successfully. Motivated by these studies, in this section, the EC model is modified to adapt the effect of reverberation. This model is then applied to the EC-BEAM, resulting in an improved algorithm, namely Adaptive EC-BEAM. Finally, the experimental results for evaluation are presented.

3.3.1 Adaptive EC model

The general structure of the improved EC model is shown in Fig. 3.6. In this scope, at-

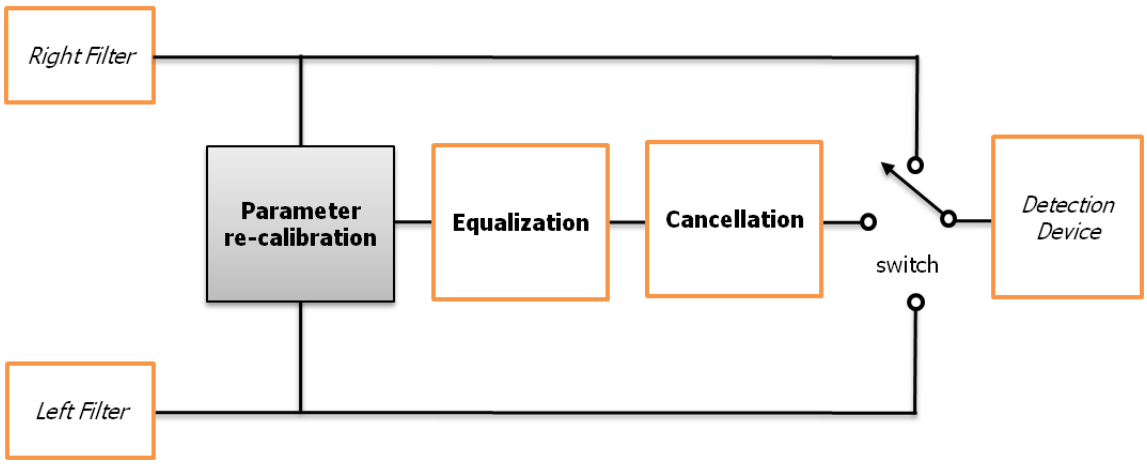

Figure 3.6: Block diagram of the adaptive EC model: Beside the main components of original EC model, the proposed model has an additional 'parameter re-calibration' step to accommodate the model's parameters corresponding to the reverberation level.

tention will be paid on the binaural processing path of the proposed adaptive EC model to enhance the binaural processing ability rather than to analyze the full model as in the application of signal detection of Durlach [14]. With the respect to the suggestion of Culling and Summerfield [62], which showed that binaural cues are frequency-dependent, the 'Equalization' and 'Cancellation' processes in the proposed model are performed independently in each frequency band. The procedure of adaptive EC model is as follows: the equalizer is trained in anechoic condition in advance to successfully compensate for the ILD and IPD of target signals at two ears as presented in Eq. (3.1). This training approach is one of the previous realizations of EC model applied in sound separation [59] and speech enhancement [4]; In the step of target elimination, the equalizer is adjusted before performing the E and C processes to match it with the reverberation level as showed in Fig. 3.6. In order to adapt the equalizer to reverberation level, we first represent the effect of reverberation on equalizer, then quantify the effect and modify the equalizer correspondingly.

Representation of reverberation effect on equalizer

The purpose of adaptive EC model is to eliminate completely sound energy from the target source. That means in reverberant condition, not only direct-path component but also reverberation generated by the source should be canceled. In order to do that, the corresponding equalizer (namely adaptive equalizer) must be able to equalize both direct-path component and reverberation. The adaptive equalizer is specified as follows:

$$W_{Adaptive}(\omega) = \frac{X_L(\omega,t) + R_L(\omega,t)}{X_R(\omega,t) + R_R(\omega,t)} \tag{3.15}$$

Let $\Delta(\omega,t)$ and $\delta(\omega,t)$ be the difference between the direct components and the difference between the reverberant components of sounds observed at the left and right receivers respectively, $\Delta(\omega,t) = X_L(\omega,t) - X_R(\omega,t)$, $\delta(\omega,t) = R_L(\omega,t) - R_R(\omega,t)$. Mathematically, the adaptive equalizer in Eq. (3.15) can be rewritten as:

$$W_{Adaptive}(\omega) = \frac{X_L(\omega,t) + Q(\omega,t)}{X_R(\omega,t) + Q(\omega,t)} \tag{3.16}$$

$$\text{where} \quad Q(\omega,t) = R_R(\omega,t) - \frac{\delta(\omega,t)\ [X_R(\omega,t) + R_R(\omega,t)]}{\Delta(\omega,t) + \delta(\omega,t)} \tag{3.17}$$

In Eq. (3.16), the effect of reverberation on the equalizer is presented by only $Q(\omega,t)$. It describes the masking effect of reverberation, which reduces the binaural differences of the target signals. This explanation is consistent with the results of the researches on the

effect of reverberation in which it is shown that reverberation degrades the directional sensitivity of auditory neurons [78, 79]. Since $X_i(\omega, t)$ $(i = L, R)$ and $Q(\omega, t)$ are not available in practice, we investigate to present the adaptive equalizer is as function Γ of the original equalizer $W(\omega)$ and the reverberation level, denoted by ℓ.

$$W_{Adaptive}(\omega) = \Gamma[W(\omega), \ell] \tag{3.18}$$

It is observed in Eq. (3.16) that the component $Q(\omega, t)$ shifts the value of $W_{Adaptive}(\omega)$ closer to 1 as its energy gets larger. Therefore, the function Γ is empirically selected as following form:

$$W_{Adaptive}(\omega) = W(\omega)^{\beta} \tag{3.19}$$

where β is specified based on the level of reverberation. Eq. (3.19) is the representation of the effect of reverberation on the equalizer in reverberant conditions.

Quantification of reverberation effect

As $Q(\omega, t)$ accounts for the effect of reverberation in Eq. (3.16), its energy should get larger when the reverberation increases and vice versa. Hence, $Q(\omega, t)$ is assumed to be described based on the right reverberation component as follows

$$Q(\omega, t) = \nu(\omega) R_R(\omega, t), \tag{3.20}$$

in which $\nu(\omega)$ depends only on room condition and is time-invariant. From Eq. (3.10), let

$$\psi(\omega) = \int_{0^+}^{\infty} \alpha(\tau) e^{-j\omega\tau} d\tau, \tag{3.21}$$

The reverberation component can be represented by (see Appendix A):

$$R_R(\omega, t) = \psi(\omega) X_R(\omega, t) \tag{3.22}$$

The adaptive equalizer in Eq. (3.16) is rewritten as follows:

$$\begin{aligned} W_{Adaptive}(\omega) &= \frac{X_R(\omega, t) + \Delta(\omega, t) + [\nu(\omega)\psi(\omega)]\ X_R(\omega, t)}{X_R(\omega, t) + [\nu(\omega)\psi(\omega)]\, X_R(\omega, t)} \\ &= 1 + \frac{1}{1 + \nu(\omega)\psi(\omega)} \frac{\Delta(\omega, t)}{X_R(\omega, t)} \end{aligned} \tag{3.23}$$

In a binaural system, the distance between two receivers is normally much smaller than that between the system and the source, hence $|\Delta(\omega,t)| \ll |X_R(\omega,t)|$. By applying Taylor expansion, $W(\omega)^\beta$ can be approximated as follows

$$W(\omega)^\beta = \left[1 + \frac{\Delta(\omega,t)}{X_R(\omega,t)}\right]^\beta \approx 1 + \beta\frac{\Delta(\omega,t)}{X_R(\omega,t)} \tag{3.24}$$

From Eq. (3.23) and Eq. (3.24), the equalizer in reverberant condition can be described using the equalizer in anechoic condition as follows

$$W_{Adaptive}(\omega) = W(\omega)^\beta \quad \text{with} \quad \beta = \frac{1}{1+\nu(\omega)\psi(\omega)} \tag{3.25}$$

It is difficult to estimate β in practice since $\nu(\omega)$ and $\psi(\omega)$ are unknown. As the reflected components may come from any direction with various time delays, the phase of reverberation may fluctuate like a random variable and may be unpredictable. Therefore, estimating the phase components of $\nu(\omega)$ and $\psi(\omega)$ to quantify the reverberation component as in Eq. (3.20) and Eq. (3.22) seems not practical. We empirically omit their phase components and the exponent β in Eq. (3.25) is approximated by $\widehat{\beta}$ as follows:

$$\widehat{\beta} = \frac{1}{1+|\nu(\omega)|.|\psi(\omega)|} \tag{3.26}$$

From Eq. (3.22),

$$|\psi(\omega)| = \frac{|R_R(\omega,t)|}{|X_R(\omega,t)|} = 10^{-\frac{DRR}{20}} \tag{3.27}$$

where DRR is the direct-to-reverberant energy ratio (in dB) defined as

$$DRR = 10\log\left(\frac{|X_R(\omega)|^2}{|R_R(\omega)|^2}\right) \tag{3.28}$$

In a relatively highly reverberant condition, the reverberation should diffuse to all spaces in room. Because of diffusion, the energies of reverberation observed at the left and right receivers should be insignificantly different. We heuristically set $|Q(\omega,t)| = |R_R(\omega,t)|$ or $|\nu(\omega)| = 1$. Finally, given the equalizer $W(\omega)$ trained in anechoic condition, the corresponding equalizer in reverberant condition, $W_{Adaptive}(\omega)$, can be obtained as follows:

$$W_{Adaptive}(\omega) = W(\omega)^{\widehat{\beta}} \quad \text{where} \quad \widehat{\beta} = \frac{1}{1+10^{-\frac{DRR}{20}}} \tag{3.29}$$

Intuitively, the Eq. (3.29) is consistent with the variation of the equalizer corresponding to room condition. Given a fixed source location, when $DRR \rightarrow +\infty$ (anechoic condition), $\widehat{\beta} \rightarrow 1$ and $W_{Adaptive}(\omega) \rightarrow W(\omega)$ according to Eq. (3.29). Similarly, in the case of extremely high reverberation or $DRR \rightarrow -\infty$, $\widehat{\beta} \rightarrow 0$ and $W_{Adaptive}(\omega) \rightarrow 1$ as a result. This is reasonable because in such condition the binaural differences between the left and right signals are completely destroyed due to the overlap masking caused by the late reverberation component.

The Eq. (3.29) can be further generalized in the case where training and testing are performed in arbitrary conditions specified by DRR_1 and DRR_2 respectively as follows:

$$W_2(\omega) = W_1(\omega)^{\widehat{\beta}} \quad \text{with} \quad \widehat{\beta} = \frac{1 + 10^{-\frac{DRR_1}{20}}}{1 + 10^{-\frac{DRR_2}{20}}} \tag{3.30}$$

The anechoic condition can be considered as the reverberant condition with $DRR = +\infty$. In that case Eq. (3.30) becomes Eq. (3.29).

3.3.2 Proposed Adaptive EC-BEAM

The concept of the adaptive EC model is applied into the EC-BEAM, resulting in an improved algorithm, named as Adaptive EC-BEAM. The procedure of the Adaptive EC-BEAM is as follows:

- Training: This step is executed same as that of the original EC-BEAM. The equalizer at each interest direction is obtained in advance using a clean anechoic sound signal. The training method is NLMS, which was suggested in previous research of Gannot *et al.* [80] and Li *et al.* [4].

- DOA estimation: The trained equalizers are first modified using DRR to match the current level of reverberation, using Eq. (3.29). In this manner, the reverberation is characterized by DRR information, which is assumed to be estimated prior the localization process. Recently, various research on DRR estimation have been investigated, such as the studies of Lu *et al.* [27] and Hioka *et al.* [61]. However, since the proposed approach in this section focuses only on the theory aspect of how the EC-BEAM can improved, estimation of DRR is ignore and this information is assumed to be known in advance.

Cancellation is performed using the modified equalizers at the interest directions,

$$C_{Adaptive}(\theta) = \int_{-\infty}^{\infty} |X_L(\omega) - W_{Adaptive}(\omega, \theta) X_R(\omega)|^2 d\omega. \tag{3.31}$$

Finally, DOA of target source is specified by looking for the minimum of residual energy, similarly to the original EC-BEAM algorithm,

$$\widehat{\phi} = \underset{\theta \in D}{\operatorname{argmin}}\, C_{Adaptive}(\theta). \tag{3.32}$$

3.3.3 Experiments and results

The experiments are divided into two parts. The first part is to verify the effectiveness of the Adaptive EC-BEAM algorithm in the presence of reverberation. This experiment is carried out with real recorded sounds in a reverberant room using two microphones. The second one is to evaluate the applicability of the Adaptive EC-BEAM regarding to the binaural SSL in noisy reverberant conditions, and to further compare it with the traditional GCC-PHAT algorithm [15].

Effectiveness of Adaptive EC-BEAM with real reverberation

This experiment evaluates whether the proposed EC model and the Adaptive EC-BEAM are able to adapt a reverberation condition; therefore, noise was not considered. Directional sound signals were speech recorded in a anechoic room and a reverberant room with the room size of $8 \times 5 \times 3.5\ m^3$ ($T_{60} \approx 0.4s$), using two microphones spacing at 0.34 m. Five speech utterances were selected from the ATR Japanese database [73] with an average length of 4 seconds. Source location varied from 0^o (the front) to 90^o (the right side) with a 10^o step at the distances of 1m, 2m and 3m. The training process was performed with anechoic sounds while the testing process was executed with reverberant ones. DRR was obtained using distance perception model proposed by Bronkhorst and Houtgast [81]. A signal length of 0.5 s was used for each estimation.

Fig. 3.7 shows the *average estimation errors* (AEEs) and the *standard errors* (SEs) along the azimuths of the EC-BEAM and Adaptive EC-BEAM algorithms in the condition sound source located at a distance of 3m. It can be observed that the estimates of EC-BEAM in the frontal area (from 0^o to 50^o) are relatively good while those in the side area (from 60^o to 90^o) are quite poor. These results are consistent with the results of binaural hearing research [29, 30, 69] since the azimuth change at the side area results in a smaller

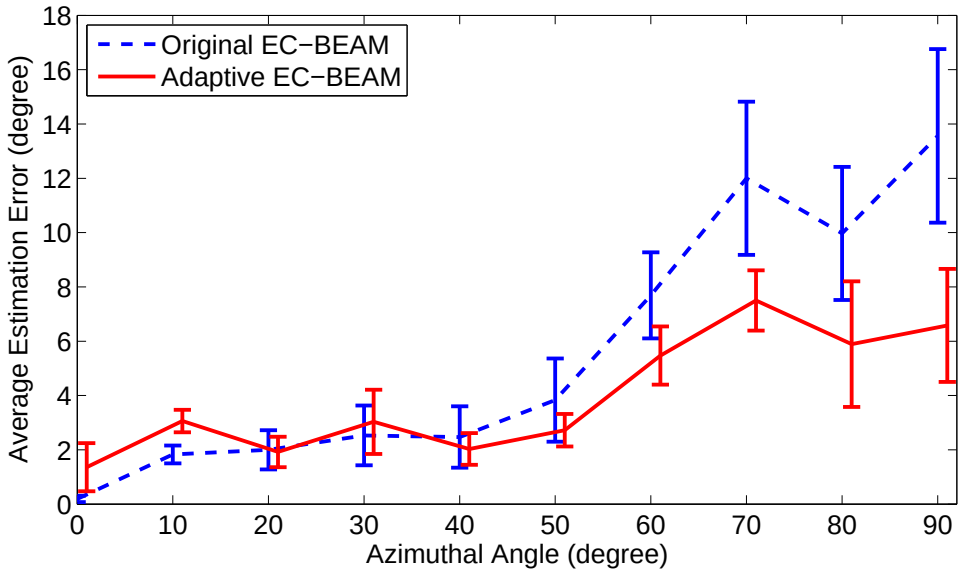

Figure 3.7: Average estimation errors and standard errors of EC-BEAM and Adaptive EC-BEAM along the azimuths, source distance at 3 m.

change in binaural cues, e.g. IPD and ILD, than that at the middle. Therefore, the effect of reverberation is more emphasised in this area. This effect is significantly reduced in Adaptive EC-BEAM since its equalizers were adjusted appropriately corresponding to the change of interaural differences along the reverberant level. It is noticed that the AEEs of Adaptive EC-BEAM at the frontal area slightly increase comparing with those of EC-BEAM. This may be because in experiment the alternative shifting factor $\widehat{\beta}$ was obtained from β by heuristically removing the phase components of $\kappa(\omega)$ and $\psi(\omega)$ in Eq. (3.25), which may cause some loss of information.

The effect of reverberation depends upon its energy in the whole signal, which is characterized by the DRR factor. Consequently, the analysis of reverberant signal based on DRR is reasonable. However, since DRR is a kind of distance-dependent factor [81], if the improvement of Adaptive EC-BEAM is strictly relied on DRR, its applicability would be relatively limited. Therefore, the feasibility of the adaptation was further investigated. Revealed from Eq. (3.29), the value of $\widehat{\beta}$ decreases very slowly when value of DRR is low enough, i.e. the distance is relatively large, as shown in Fig. 3.8. This leads to a possibility of using only one appropriate value of $\widehat{\beta}$ for each room condition in the scope of far-field localization. Table 3.3 shows the overall AEEs of the Adaptive EC-BEAM in comparison with EC-BEAM along various distances using the DRR computed at fixed 2m. From Table 3.3, the best improvement of Adaptive EC-BEAM is remarked at the matched DRR value (2m). Its improvements at 1m (over-estimated) and 3m (under-

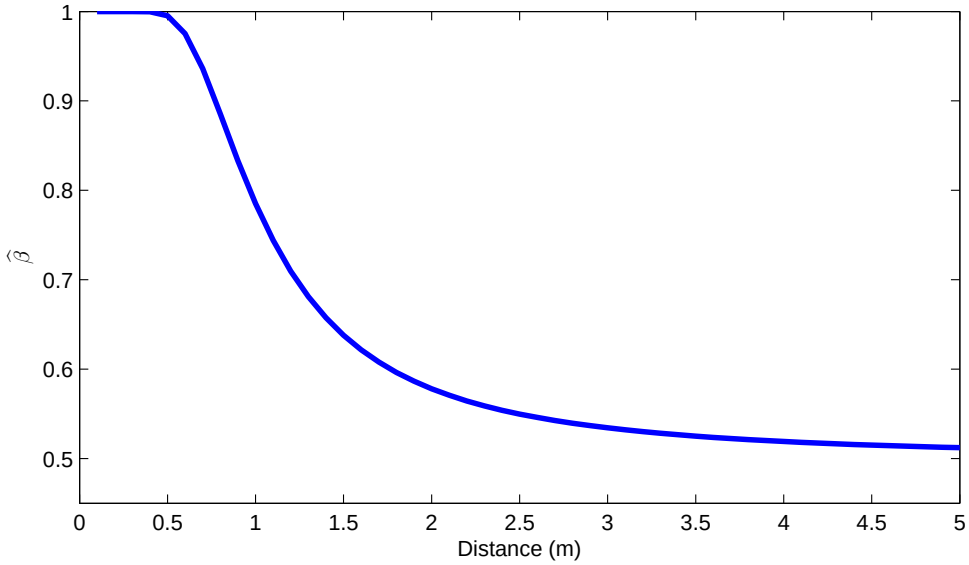

Figure 3.8: Values of $\widehat{\beta}$ along the distances calculated using distance perception model in the room $8m \times 5m \times 3.5m$, $T_{60} \approx 0.4s$.

Table 3.3: Overall average estimation errors (in degrees) of Adaptive EC-BEAM and EC-BEAM, modification using DRR computed at fixed 2m.

Distance	1 m	2 m	3 m
EC-BEAM	1.9	4.1	5.4
Adaptive EC-BEAM	1.9	2.0	4.0

estimated) are not as good as that at 2m, however, in general the Adaptive EC-BEAM method outperforms the original EC-BEAM algorithm. These results supported that the proposed adaptation is effective in the presence of reverberation.

Adaptive EC-BEAM in noisy reverberant conditions

This experiment is designed to extensively examine the proposed Adaptive EC-BEAM method with binaural sound signals in noisy reverberant environments. Directional sounds were created by convolving *impulse responses* from Head-related impulse response (HRIR) database of Odenburg University [82] with speech utterances, which were selected in ATR Japanese database as in the first experiment. The impulse responses were measured by two microphones placed at the *rear of dummy ear* in the room condition named *"Office I"*. Babble noise was added at an SNR of 10 dB. The training process was performed in anechoic condition, while the testing process was performed in noisy reverberant conditions. The DRR was also calculated in the same way of the first experiment. Original

EC-BEAM and GCC-PHAT were performed in the same condition for comparison. In the implementation of GCC-PHAT, since it is difficult to derive DOA correctly from time delay via mathematical geometry calculation due to the HRTF effect, we designed a prior "training step" to create a mapping from time delay to direction. Estimate was obtained by firstly calculating the time delay using GCC-PHAT, then matching to the mapping. All the algorithms used an equal signal duration of 200 ms for each estimation. The metric of evaluation is ratio of incorrect estimates to total estimates (error rate), in which an estimate is correct if its error is not over a certain threshold.

Fig. 3.9 shows the error rates along various error thresholds of estimates performed by the three algorithms. Since the EC-BEAM algorithm relies on the compensation of equalizer regarding to ILD and IPD, its performance is degraded when mismatching occurs. This is the common limitation of methods based on learning approach. The performance of GCC-PHAT method is relatively better than the EC-BEAM since there is no mismatching in ITD calibration. By modifying equalizer properly, the Adaptive EC-BEAM algorithm is more robust against reverberation and outperforms the GCC-PHAT method. It should be noticed that in the current system, the microphones were located at the rear of the dummy ears (not in ear). This leads to the fact that using ITD cannot distinguish the azimuths. As shown in Fig. 3.10, the time delays at several directions are identical and the time delay at 90^o is even lower than that at 80^o. In this case, ILD should be a valuable additional cue to improve localization ability and thereby the Adaptive EC-BEAM outperformed GCC-PHAT method. These results verified that the Adaptive EC-BEAM is adequate for binaural SSL and robust against reverberation.

3.3.4 Discussion

A mechanism has been provided to adapt the EC model to reverberation conditions by modifying its equalizer corresponding to the DRR. The experimental results demonstrated that the Adaptive EC-BEAM, which is based on the adaptive EC model concepts, outperformed the original EC-BEAM and the GCC-PHAT algorithms. This implies that the EC model was successful in eliminating the target signal. Moreover, the results in the second experiment showed that the Adaptive EC-BEAM is able to work effectively on a binaural system in noisy reverberant conditions, which supports that both objectives pointed out in this thesis are satisfied. Essentially, the modification of the equalizer is a re-calibration corresponding to the degradation of the binaural cues of the target signal under the effect of reverberation. The exponent $\hat{\beta}$ tends to reduce the compensation for the IPD and ILD when the reverberation level increases. In a reverberant room, as the reflection propagates

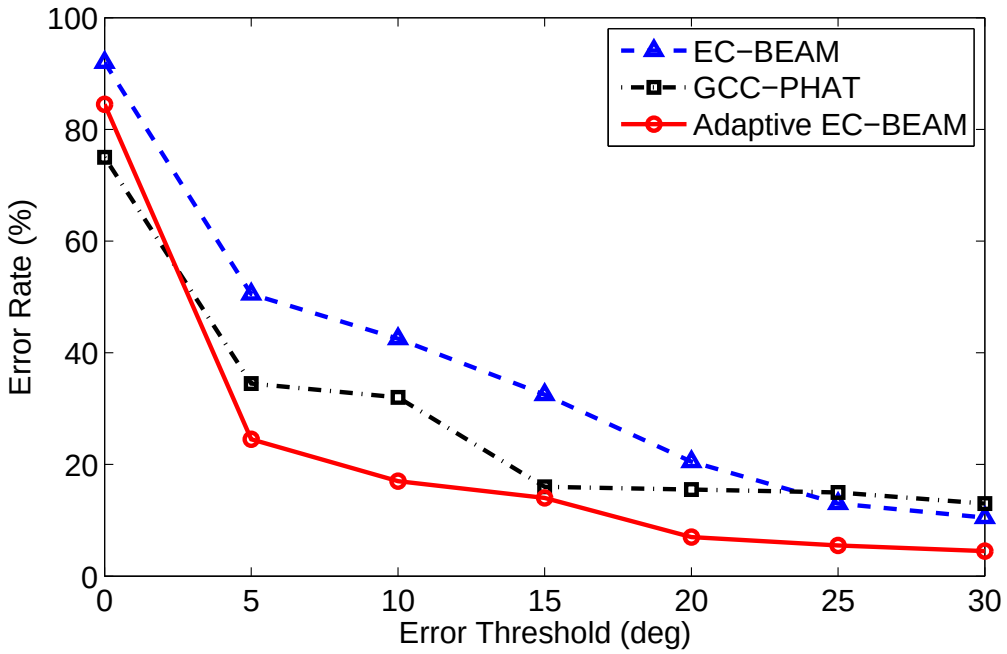

Figure 3.9: Error rate along the thresholds of estimates performed by EC-BEAM, GCC-PHAT and Adaptive EC-BEAM in noisy *'Office I'* condition.

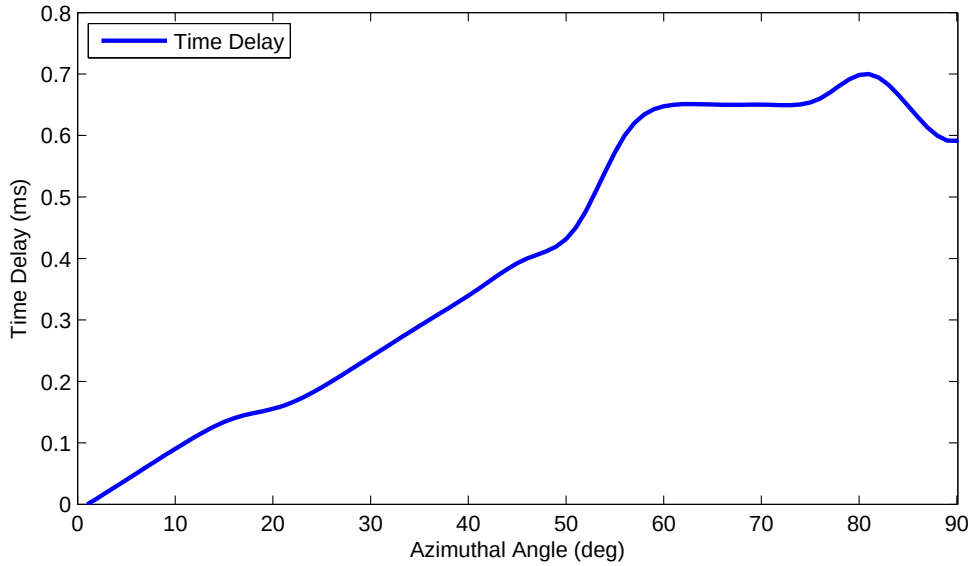

Figure 3.10: ITD is an ambiguous cue for DOA estimation under the shadow effect when microphones were placed at the rear of dummy ears.

to all spaces, its energy should be diffused and equivalent at both ears. Therefore, the adjustment of the equalizer is consistent with the reverberation effects, especially on the ILD.

In the study of Shinn-Cunningham [68], the human auditory system is able to adapt to a room condition naturally; thus, an ideal adaptive binaural model should be able to specify the acoustical condition and accommodate to itself automatically. However, this should be a complicated work because the process of hearing is affected by a large number of acoustical factors and the sound processing mechanism of human brain has not been fully understood. The adaptive EC-model proposed in this section is a step toward the ideal model, which concentrates on understanding the effect of reverberation and provides a theoretical way to account for this effect. Therefore, the method to obtain the DRR information has not been provided. So far, there have been a number of studies tackling the problem of DRR estimation and achieved some successes, such as Lu *et al.* [27] and Hioka *et al.* [61]. It would be interesting to link these works to the work presented in this thesis for a further development of the proposed model. Although the Adaptive EC-BEAM using the concept of adaptive EC model has achieved relatively good DOA estimation in some experimental conditions, there would be still many difficulties to realize it in practical applications.

3.4 Approach 2: Reducing the effects of noise and reverberation

In noisy conditions, the target signal can be detected easier when the characteristics of noise and the target signal are different. This inspires an idea to exploit these differences to improve performance of localization of the target signal (such as speech) against undesired factors (such as background noise and reverberation). In this section, the effects of noise and reverberation on the EC operations are analyzed and two weighting functions are suggested to suppress these effects based on their characteristics, while the perception model (the EC model) is kept unchanged. Since noise and reverberation together form a complicated effect, we consider two scenarios in which noise and reverberation are treated separately.

3.4.1 Eliminating the effect of background noise

Noise may enhance sound energy at the directions other than the target direction. It is more difficult if noise at one ear is correlated with that of the other ear, such as directional noise or localized noise. This is because if the energy of noise source is higher than that of the target source, DOA of noise will be detected instead of DOA of target sound. Normally, the background noise is stable in time regarding to energy and location

of sources, which can be assumed as wide sense stationary (WSS), while target sound (such as speech) normally does not continue at all the time. Therefore, it is possible to extract noise only periods, for example using voice activity detection (VAD), to analyze and improve SSL performance.

For simplicity, in this scenario reverberation component is omitted and observed signal consists of only target signal and noise, which can be represented as follows:

$$Y_i(\omega) = X_i(\omega) + N_i(\omega), \quad i = L, R \tag{3.33}$$

where $N_i(\omega)$ present the background noise, including diffuse noise and directional noise. With the assumption of uncorrelation between target signal and noise, the C operation in Eq. (3.14) can be rewritten as follows (see Appendix C for further explanations):

$$\begin{aligned} C_Y(\theta) &= \int_{-\infty}^{\infty} |X_L(\omega) - W(\omega,\theta)X_R(\omega)|^2 \, d\omega + \int_{-\infty}^{\infty} |N_L(\omega) - W(\omega,\theta)N_R(\omega)|^2 \, d\omega \\ &\triangleq C_X(\theta) + C_N(\theta). \end{aligned} \tag{3.34}$$

In Eq. (3.34), the C operation is applied to not only target signal but also noise. Due to the compensation of $W(\omega, \theta)$, the residual noise $C_N(\theta)$ varies along the steering direction θ and affects the final output of the C process. Fig. 3.11 is an example of this effect. Because of the considerable variation of $C_N(\theta)$, although the residual of target, $C_X(\theta)$, yields a quite good minimum to specify the DOA, the total output of C operation on both target and noise is not minimal at the direction of sound source.

Note that because $W(\omega)$ is time-invariant, if $N_L(\omega)$ and $N_L(\omega)$ are WSS, $Z_\theta(\omega) = N_L(\omega) - W(\omega,\theta)N_R(\omega)$ is also WSS. Considering $C_N(\theta)$ on N_ω discrete finite frequency bins, its value can be described as follows:

$$C_N(\theta) = \sum_{\omega} |Z_\theta(\omega)|^2 = N_\omega E[Z_\theta(\omega) Z_\theta^*(\omega)] = N_\omega \Phi_{Z_\theta Z_\theta}(0), \tag{3.35}$$

where $E[.]$ denotes the mean operator and $\Phi_{Z_\theta Z_\theta}(0)$ is the auto correlation of $Z_\theta(\omega)$ at the time shift 0, which is also time-invariant as $Z_\theta(\omega)$ is WSS. Therefore, $C_N(\theta)$ depends on only the steering direction θ and is independent of the observed time t. In order to adapt to the effect of noise on the C operation, we additionally use a noise compensation coefficient $\kappa(\theta)$ for each steering direction $\theta \in D$ so as the cancellation outputs of noise

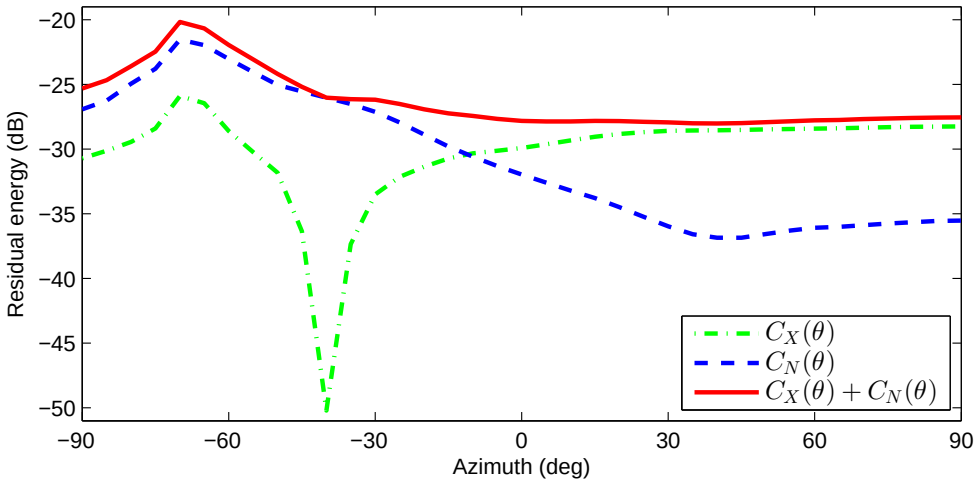

Figure 3.11: Outputs of C operation with target, noise, and noisy signal. Target source locates at $-40°$; noise consists of diffuse noise and directional noise (at $40°$) with equivalent energies; SNR of noisy signal is 0 dB.

at all steering directions are equalized, that is

$$\kappa(\theta_m)C_N(\theta_m) = \kappa(\theta_n)C_N(\theta_n), \quad \forall\theta_m, \theta_n \in D. \tag{3.36}$$

In this manner, $\kappa(\theta)$ is a scalar value characterizing for the distribution of noise energy at each direction. In implementation, $\kappa(\theta)$ is calibrated before DOA estimation by first performing the C operation with non-target signal period (where only noise is present) to obtain $C_N(\theta)$ at all directions $\theta \in D$, then by setting $\kappa(\theta)$ to 1 at $0°$, i.e. $\kappa(0°) = 1$, $\kappa(\theta)$ at other directions are specified by

$$\kappa(\theta) = \frac{C_N(0°)}{C_N(\theta)}, \quad \forall\theta \in D. \tag{3.37}$$

The C operation of EC-BEAM with consideration of noise, named as EC-BEAM/N, is suggested as follows:

$$\begin{aligned} C_Y^N(\theta) &= \kappa(\theta) \int_{-\infty}^{\infty} |Y_L(\omega) - W(\omega,\theta)Y_R(\omega)|^2 \, d\omega \\ &= \kappa(\theta)C_X(\theta) + \kappa(\theta)C_N(\theta) \end{aligned} \tag{3.38}$$

If the equalizer is well trained, the output of target cancellation should drops to approximately zero when the steering direction matches the direction of target source. As a

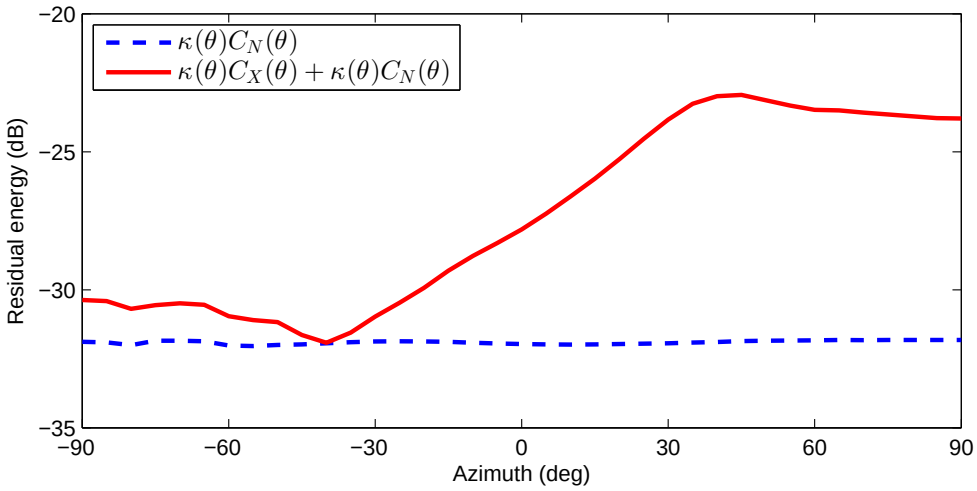

Figure 3.12: Outputs of the C operation incorporated with noise compensating coefficient $\kappa(\theta)$ performed on noise and noisy signal. Configuration is same as in Fig. 3.11.

result, the residual energy of the null at the direction of target source consists of only the energy of (equalized) residual noise, i.e. $C_Y^N(\phi) = \kappa(\phi)C_N(\phi)$, and the minimum is yielded at the direction of sound source, as illustrated in Fig. 3.12.

It is clear that EC-BEAM/N can work well with noise as long as the coefficients $\kappa(\theta)$ are properly obtained. In fact, $\kappa(\theta)$ can be constructed WSS noise, such as background noise, even in the case it contains directional noise. Such kind of noise is popular in normal room conditions, for example the noise from fans and air-conditioners. However, this strategy is not able to deal with reverberation because its effect cannot be learned in the absence of target signal.

3.4.2 Eliminating the effect of late reverberation

Similarly to the first strategy, in this scenario, noise component is omitted for simplicity. Observed signal now consists of target signal (direct-path) and reverberation, in which reverberation can be divided into early and late reverberation components with different characteristics. The early reverberation can be considered as sound observed from multiple sources where the energy of each source is lower than that of the target source. As EC-BEAM is able to work with multiple interference signals (3.1.3), it performance should be fine with early reverberation. We conceptually divide the observed signal into two

components: early response and late response:

$$\begin{aligned} Y_i(\omega) &= X_i(\omega) + R_i^E(\omega) + R_i^L(\omega) \\ &= X_i^E(\omega) + X_i^L(\omega), \quad i = L, R \end{aligned} \tag{3.39}$$

in which the early response $X_i^E(\omega)$ includes the direct-path sound component $X_i(\omega)$ and the early reverberation component $R_i^E(\omega)$, while the late response is equivalent with the late reverberation component. Because the component $X_i^L(\omega)$ at each receiver can be assumed as uncorrelated with $X_i^E(\omega)$ and together uncorrelated at both channels, the C operation in Eq. (3.14) can be rewritten as follows (see Appendix C):

$$C_Y(\theta) = \int_{-\infty}^{\infty} \left| X_L^E(\omega) - W(\omega,\theta) X_R^E(\omega) \right|^2 d\omega + \int_{-\infty}^{\infty} \left[|X_L^L(\omega)|^2 + |W(\omega,\theta)|^2 |X_R^L(\omega)|^2 \right] d\omega \tag{3.40}$$

Eq. (3.40) shows that cancellation output of late response component varies along the steering directions only because of the amplitude of the equalizers, which corresponds to the ILD of target signals. Therefore, the C operation of suggested EC-BEAM against late reverberation, named as EC-BEAM/R, is executed without ILD compensation as follows:

$$\begin{aligned} C_Y^R(\theta) &= \int_{-\infty}^{\infty} \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} - \frac{W(\omega,\theta)}{|W(\omega,\theta)|} \frac{Y_R(\omega)}{|Y_R(\omega)|} \right|^2 d\omega \\ &= \int_{-\infty}^{\infty} \left| \frac{X_L^E(\omega)}{|Y_L(\omega)|} - \frac{W(\omega,\theta)}{|W(\omega,\theta)|} \frac{X_R^E(\omega)}{|Y_R(\omega)|} \right|^2 d\omega + \int_{-\infty}^{\infty} \left[\left| \frac{X_L^L(\omega)}{Y_L(\omega)} \right|^2 + \left| \frac{X_R^L(\omega)}{Y_R(\omega)} \right|^2 \right] d\omega. \end{aligned} \tag{3.41}$$

Since the residual of late response component in Eq. (3.41) is independent of steering directions, it does not affect the minimum of cancellation output. As a results, DOA estimate is specified based on the minimum of cancellation of the early response component only. In this manner, the E operation just compensates for IPD to the early response components as the ILD of equalizers was excluded. Note that although EC-BEAM/R is proposed to deal with late reverberation, it is also effective with other interference having similar characteristic with this component, such as diffuse noise.

3.4.3 Proposed Weighted EC-BEAM

In order to estimate DOA in the presence of noise and reverberation simultaneously, we propose to integrate both strategies into the EC-BEAM algorithm to improve its performance, namely Weighted EC-BEAM. Final DOA estimate is decided based on the residual energies of both EC-BEAM/N and EC-BEAM/R. Because in EC-BEAM/R, the amplitude of the observed signals is eliminate before performing cancellation, its residual energy, $C_Y^R(\theta)$, is much lower than the residual energy $C_Y^N(\theta)$ from EC-BEAM/N. To keep the two outputs balance and make them contribute equivalently to the final estimation, $C_Y^R(\theta)$ and $C_Y^N(\theta)$ are normalized to [0,1] before summing. The normalization is suggested as follows:

$$NORM[a(\theta)] = \frac{a(\theta) - \min\limits_{\varphi \in D} a(\varphi)}{\max\limits_{\varphi \in D} a(\varphi) - \min\limits_{\varphi \in D} a(\varphi)}, \tag{3.42}$$

in which the symbol a stands for C_Y^R and C_Y^N. The final estimated DOA is specified by:

$$\begin{aligned} &\widehat{\phi} = \underset{\theta \in D}{\operatorname{argmin}}\, C_{Weighted}(\theta), \\ &\quad \text{where} \quad C_{Weighted}(\theta) = NORM[C_Y^N(\theta)] + NORM[C_Y^R(\theta)]. \end{aligned} \tag{3.43}$$

3.4.4 Experiments and Results

Experiments on Weighted EC-BEAM are divided into two parts. The first part is to verify the effectiveness EC-BEAM/N and EC-BEAM/R in noise and in reverberation separately, and to show how the EC-BEAM is improved by combining both two strategies. The second part is to evaluate the applicability of the proposed Weighted EC-BEAM algorithm in noisy reverberant environments and further compare it with the SRP-PHAT and cross-channel HRTF algorithms discussed in section 2.2.

Materials and configurations

In evaluation, we simulate various reverberant conditions by using the ROOMSIM package [70]. The ROOMSIM software utilizes the HRTF measurements obtained by KEMAR dummy head [71] to generate simulated reverberant binaural room impulse responses (BRIRs) based on the image method [72]. BRIRs generated in this way are expected to represent reliable simulations since the software uses real measured HRTFs. Reverberant BRIRs are generated in a $10 \times 10 \times 3$ (m^3) room with reverberation times (T_{60}) from 0 to 0.8 s, depending on experiments. Anechoic BRIRs are selected directly from HRTF

measurements without simulation. Source location varies from $-90°$ to $90°$ with the step of $10°$, at the distances from 1 to 4 m with the step of 1 m.

Speech data are selected from ATR Japanese database [73]. Six speech sentences with the average length of 10 seconds are chosen, in which three are uttered by males and the others are uttered by females. These speech sentences are convolved with the simulated BRIRs to generate directional sound signals. Simulated background noise is added into directional signals to produce noisy reverberant data. The background noise consists of diffuse noise and directional noise, in which diffuse noise is generated by first filtering sounds recorded from an air-conditioner by BRIRs at all directions then summing, while directional noise is created by filtering a sound from a fan with the BRIRs at $40°$. As these recorded noises have some different characteristics, the source of directional noise can still be perceived after mixing. The energies of two kinds of noise in the mixture are kept to be approximately equal. When adding the total noise to reverberant signals, the power of noise is controlled to obtained signal-to-noise energy ratio (SNR) from -5 dB to 15 dB (step of 5 dB).

The equalizer at each direction is trained using clean-anechoic signal generated from a speech sentence and used to estimate DOA of signals generated from other five sentences in all the experiment conditions. Training process is conducted using NLMS method as specified in Eq. (3.4). The noise compensation coefficient $\kappa(\theta)$ is calibrated using one-second period of noise. As $\kappa(\theta)$ is independent of noise level, it is calibrated only one time and applied to all SNR conditions. In test, we use a window length of 0.1 s with 50% overlapping and integrate the response energy over 0.5 s for each estimation. We evaluate the performance of SSL via the ratio of incorrect estimates to all estimates (*error rate*), where an incorrect estimate is defined as one having absolute error over $10°$.

Experiment part 1

Four algorithms are examined, including the original EC-BEAM, the EC-BEAMs using individual strategies to deal with noise (EC-BEAM/N) and reverberation (EC-BEAM/R) and the EC-BEAM combining both strategies (Weighted EC-BEAM). Test data are generated with source distance at 3 m.

Fig. 3.13 and Fig. 3.14 respectively demonstrate the improvements of EC-BEAM/N and EC-BEAM/R in comparison with the original EC-BEAM. Fig. 3.13 shows the error rates of the original EC-BEAM and EC-BEAM/N along SNRs in anechoic condition. The original EC-BEAM can localize sound source relatively well at low noise conditions. However, its error rate rapidly increases as noise level gets higher. This is due to the effect of noise as analyzed in Section 3.4.1. By learning and adapting to the effect of noise,

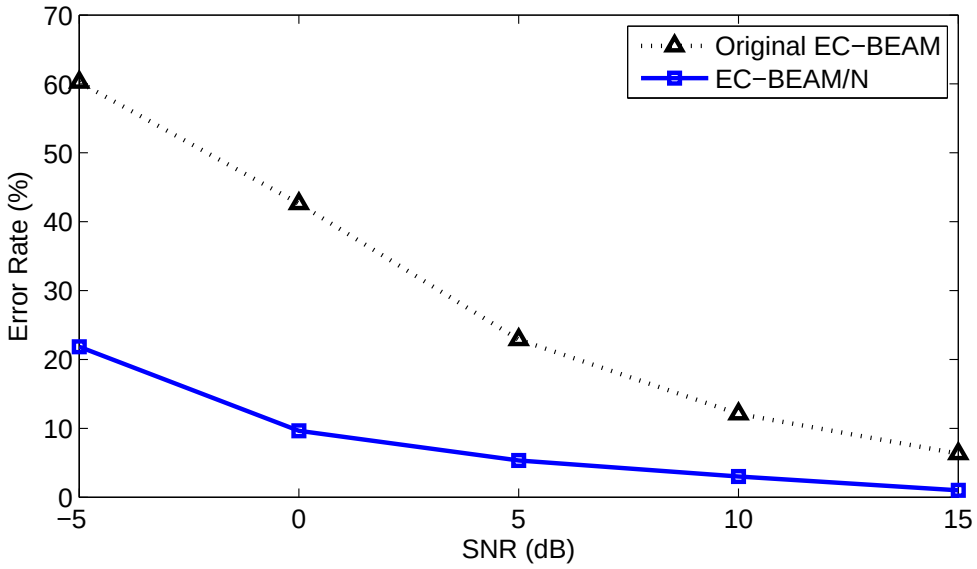

Figure 3.13: Performance of original EC-BEAM and EC-BEAM/N in noisy anechoic condition. Sound source locates at the distance of 3 m.

the error rate of EC-BEAM/N is dramatically reduced. Similar results are observed in Fig. 3.14, which represents the error rates of original EC-BEAM and EC-BEAM/R along reverberation time in the absence of noise. We can see that the original EC-BEAM also suffers from the effect of reverberation while the EC-BEAM/R is quite robust against this effect. These results indicate that each strategy works efficiently in the condition that it is designed for, i.e. noise or reverberation is present individually.

We further examine both strategies and their combination in the conditions where both noise and reverberation are concurrently present. Fig. 3.15 shows the error rates of the four algorithms in the case $T_{60} = 0.5$ s and SNR varies from -5 dB to 15 dB. It can be observed that the performances of the original EC-BEAM, EC-BEAM/N and EC-BEAM/R are relatively reduced in comparison with their performances in the conditions where noise or reverberation is present alone. However, EC-BEAM/N and EC-BEAM/R still work considerately better the original EC-BEAM. This indicates that the strategy to deal with an interference component does not so suffers from other interference component. EC-BEAM/N performs better than EC-BEAM/R in high noise conditions as it can well adapt to noise. When the SNR is higher than 5 dB, EC-BEAM/R outperforms EC-BEAM/N as it can account for reverberation. The Weighted EC-BEAM algorithm makes fully use of the advantages of both strategies and achieves the best performance through all SNR conditions. This supports that integrating the two strategies into EC-BEAM is reasonable and the Weighted EC-BEAM algorithm is able to deal with noise and reverberation

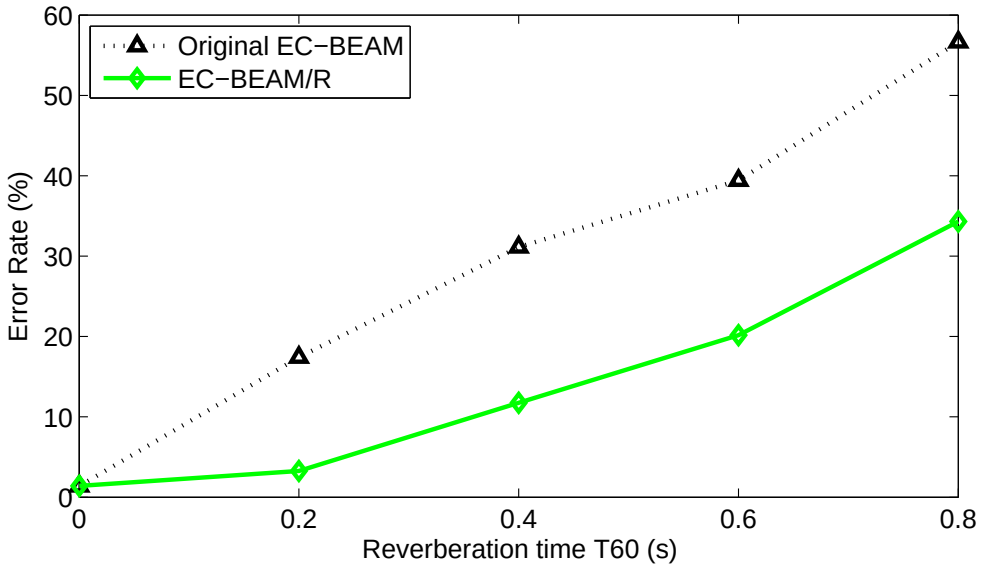

Figure 3.14: Performance of original EC-BEAM and EC-BEAM/R in reverberant conditions. Source distance is at 3 m and no noise is present.

simultaneously.

Experiment part 2

Reverberation time is set to 0.5 s while distance of sound source varies from 1 to 4 m (step of 1 m). In implementation, the SRP-PHAT method uses IPDs calculated in anechoic condition at each direction to perform beamforming. The Cross HRTF method is executed using directly the HRTFs measured in anechoic condition.

Fig. 3.16 shows the average error rates of Cross HRTF, SRP-PHAT and Weighted EC-BEAM respectively across all distances. It can be observed that the Cross HRTF method yields highest error rates among the three algorithms. The Cross HRTF has the advantage that it possesses the HRTFs, which provide the propagation information at all interest directions. Therefore, it was reported with relatively accurate estimation in low interference (noise and reverberation) conditions [8]. However, as it does not account for interference effects, its performance is dramatically degraded in the presence of either high noise or high reverberation. Moreover, since this method is strictly dependent on HRTFs, its applicability to practice may be limited because accurately measuring these information in an arbitrary binaural system is a time-consuming work.

The error rate of SRP-PHAT is consistently lower than that of Cross HRTF in both high and low noise conditions. This is partly because SRP-PHAT is quite robust against reverberation since it was shown as an approximation of maximum likelihood in low

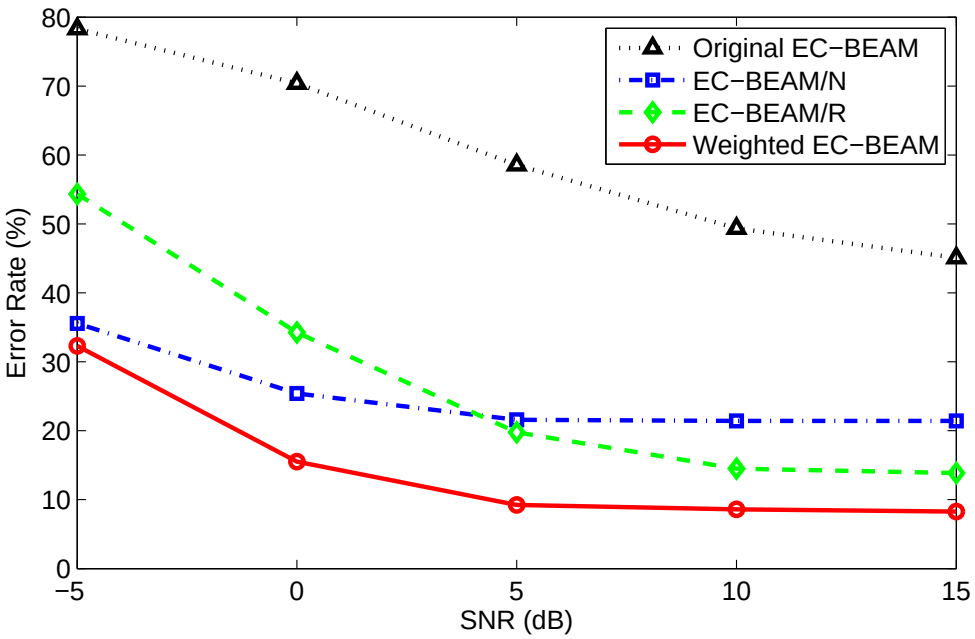

Figure 3.15: Performance of the four algorithms in noisy reverberant conditions. Source distance is at 3 m and $T_{60} = 0.5s$.

noise condition [63]. However, it still suffers from high noise condition as its error rate dramatically increases at low SNRs. This implies that SSL methods based on ITD (or IPD) only may not be fully adequate to binaural systems because the ILD, which varies largely through the azimuths due to the effect of HRTFs, is also a very important cue to specify source direction.

The Weighted EC-BEAM algorithm outperforms both Cross HRTF and SRP-PHAT, especially in high noise conditions. At the SNR of -5 dB, the proposed method improves roughly 20% and 25% error rate comparing to SRP-PHAT and Cross HRTF, respectively. In relatively low noise conditions when SNR is higher 10 dB, our method performs equivalently to SRP-PHAT but still improves about 8% error rate in comparison with Cross HRTF. This is because both noise and reverberation are taken into consideration in the proposed method. The noise adapting strategy makes Weighted EC-BEAM more robust against background noise as well as fit to the binaural setup, while the strategy for reverberation significantly reduces the effect of this factor, especially the late reflection component. On the other hand, our method is more flexible than the Cross HRTF method since training the equalizers with observed signals should be easier than meticulously measuring HRTFs.

Fig. 3.17 shows the error rates of the three algorithms along the error thresholds in a typical case where SNR is fixed at 5 dB. It can be observed that the error rate of Weighted

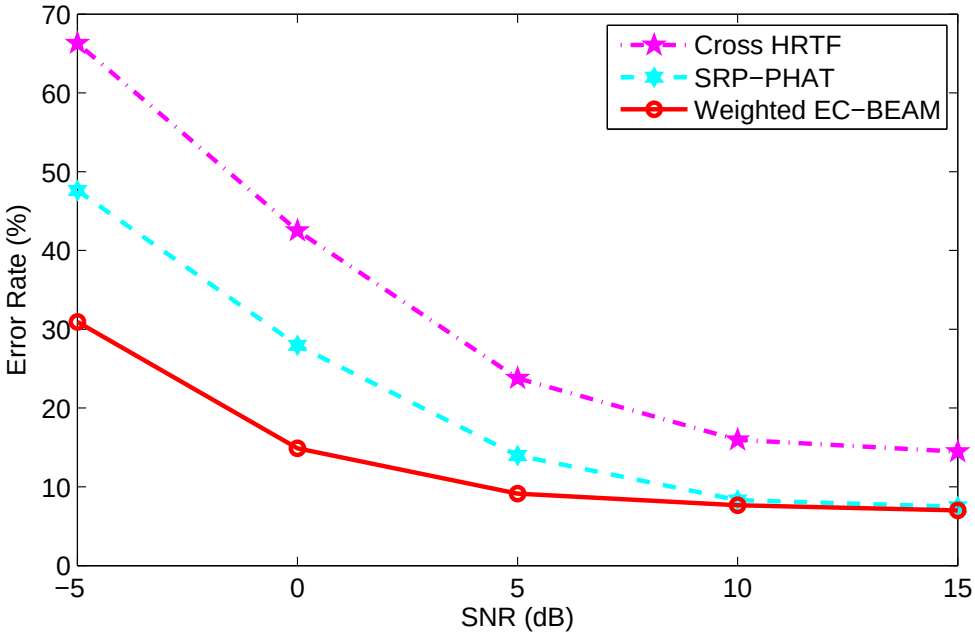

Figure 3.16: Average error rates of Cross HRTF, SRP-PHAT and Weighted EC-BEAM along SNRs. Error rate of each algorithm is calculated by taking mean of error rates through distances from 1 m to 4 m ($T_{60} = 0.5s$).

EC-BEAM is lower than those of Cross HRFT and SRP-PHAT at not only the defined 10^o-threshold but also at all of other thresholds. This implies that our proposed method is still better than the others in the case higher accuracy is required, and the angular error of our wrong estimates (whose error is higher than 10^o) is also smaller than those of the other algorithms. Fig. 3.18 further compares the robustness of the three algorithms against reverberation. In the same room condition, the energy of reverberation increases relatively to that of direct component when the distance of sound source increases [81]. As a result, the effect of reverberation on signals received from longer-distance source is also higher. Observation from Fig. 3.18, the error rate of Cross HRTF rises quickly through distances, indicating that this method quite suffers from reverberation. Both error rates of SRP-PHAT and Weighted EC-BEAM increase slower than that of Cross HRTF, in which the error rate of Weighted EC-BEAM is always below that of SRP-PHAT because it is able to adapt to noise (including directional noise) in the room.

3.4.5 Discussion

This section has introduced two strategies to suppress the effects of background noise and late reverberation on the EC-BEAM algorithm based on their characteristics. The

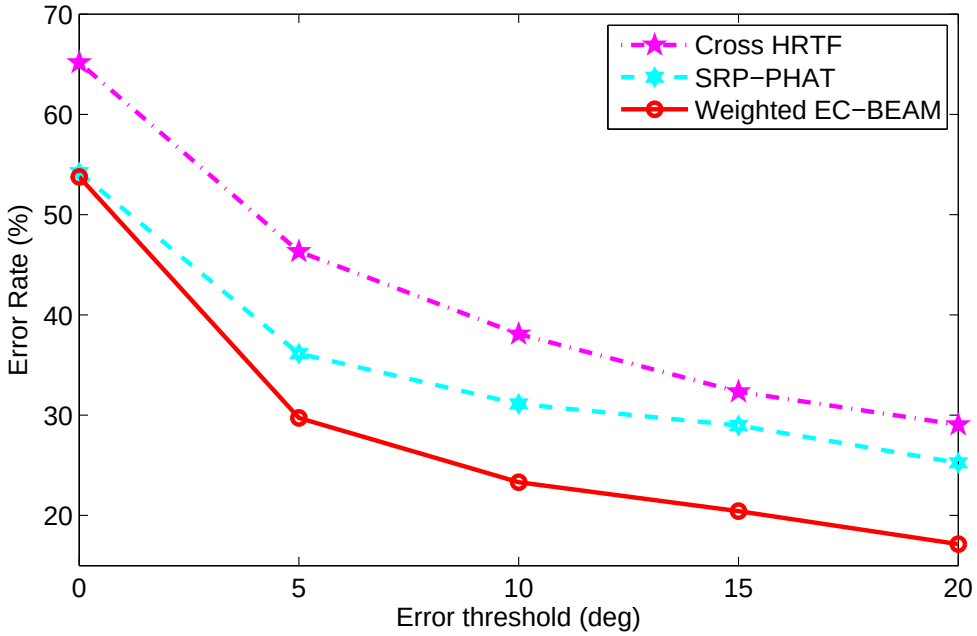

Figure 3.17: Error rates of Cross HRTF, SRP-PHAT and Weighted EC-BEAM along error thresholds at the fixed 5-dB SNR. Error rate at each threshold is mean of error rates through distance from 1 m to 4 m.

EC-BEAM/N equalizes the residual energies of background noise at all directions, while the EC-BEAM/R makes the residual energies late reverberation invariant of the steering directions. Experimental results showed that these processings can effectively reduce the effects of noise and reverberation on the EC procedures. As a result, the Weighted EC-BEAM significantly reduced the estimation error and outperformed the Cross-HRTF and SRP-PHAT methods, which indicates that two objectives of this thesis are satisfied.

On a mathematical view point, the Cross-HRTF method and the original EC-BEAM method may have some equivalences. The cost function of Cross HRTF method [8] discussed in Section 2.2 can be rewritten in the frequency domain as follows:

$$\begin{aligned} E(\theta) &= \int_{-\infty}^{\infty} |X_L(\omega)H_R(\omega,\theta) - X_R(\omega)H_L(\omega,\theta)|^2 d\omega \\ &= \int_{-\infty}^{\infty} \left| H_R(\omega,\theta)\left[X_L(\omega) - \frac{H_L(\omega,\theta)}{H_R(\omega,\theta)}X_R(\omega)\right]\right|^2 d\omega. \end{aligned} \tag{3.44}$$

From Eqs. (3.5), (3.16), and (3.44), it can be realized that Cross HRTF is a filtered version of the original EC-BEAM, where $H_R(\omega,\theta)$ is the filter in this manner. Because of the

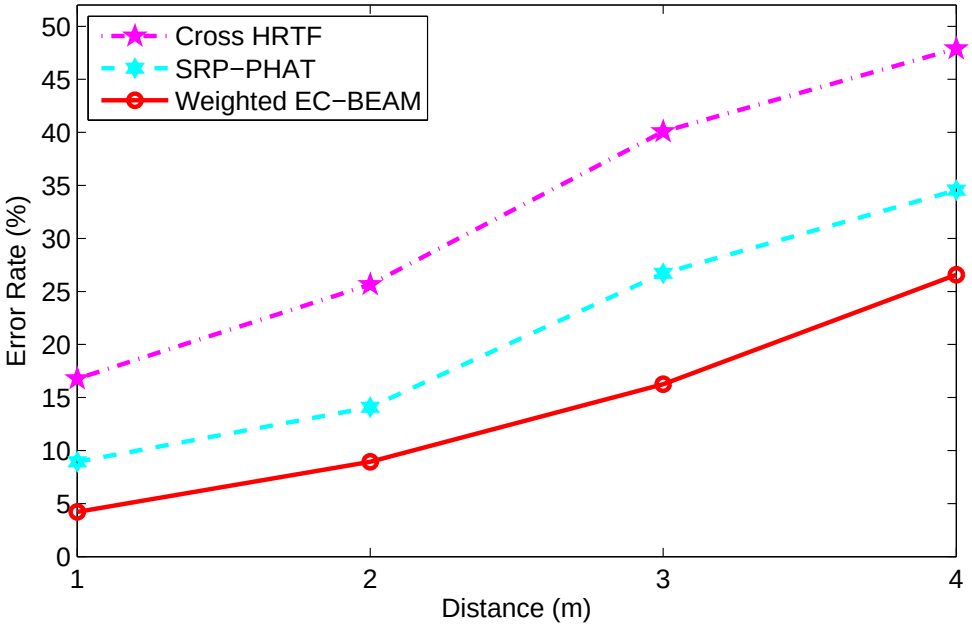

Figure 3.18: Average error rates of Cross HRTF, SRP-PHAT and weigted EC-BEAM along distances. Error rate of each algorithm is calculated by taking mean of error rates through SNRs from -5 dB to 15 dB ($T_{60} = 0.5s$).

equivalence of the two methods, the Cross HRTF method would face similar problems with that of the original EC-BEAM discussed in Section 3.2. This is the reason why the Cross HRTF performs poorly in high noise and high reverberant conditions. Since the Cross HRTF method strictly relies on HRTFs, it is hard to understand the effect of undesired factors on this method and it has less chance to be improved. From this point of view, it would be interesting to investigate whether the proposed strategies can improve the Cross HRTF in noisy reverberant environments.

In terms of sound localization using ITD, the strategy to make EC-BEAM robust against reverberation in Section 3.4.2 (EC-BEAM/R) has a close relation with SRP-PHAT. The cost function of SRP-PHAT with two microphones in Eq. (2.6) can be rewritten as follows:

$$\begin{aligned} P(\theta) &= \int_{-\infty}^{\infty} \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} e^{j\omega\tau_L} + \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega\tau_R} \right|^2 d\omega \\ &= \int_{-\infty}^{\infty} \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} + \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega(\tau_R - \tau_L)} \right|^2 d\omega. \end{aligned} \tag{3.45}$$

On the other hand, from Eq. (3.1), the phase component of the equalizer is the IPD of the target signal. Omitting the fact that the time delay τ_i may be frequency-dependent, we have:

$$\frac{W(\omega)}{|W(\omega)|} = e^{j\omega(\tau_R - \tau_L)}. \tag{3.46}$$

The EC-BEAM/R in Eq. (3.41) can be rewritten as follows:

$$C_Y^R(\theta) = \int_{-\infty}^{\infty} \left| \frac{Y_L(\omega)}{|Y_L(\omega)|} - \frac{Y_R(\omega)}{|Y_R(\omega)|} e^{j\omega(\tau_R - \tau_L)} \right|^2 d\omega. \tag{3.47}$$

From Eq. (3.45) and Eq. (3.47), we can see that the SRP-PHAT and EC-BEAM/R are quite similar. The only difference between these two methods is that one is based on the similarity of the observed signals at two channels by maximizing the beamformer, the other is based on their dissimilarity by minimizing the null. These two methodologies are mentioned as the equivalent approaches in DOA estimation [3] and should provide similar results. The advantage of the proposed Weighted EC-BEAM method in comparison with SRP-PHAT is the strategy accounting for the effect of noise in rooms, i.e. the EC-BEAM/N. However, the strategy of noise adaptation is mainly effective in high noise conditions. This explains why the proposed Weighted EC-BEAM algorithm localizes well at low SNRs while its performance remains as good as that of SRP-PHAT in the presence of reverberation at high SNRs.

3.5 General discussion

The main purpose of this chapter is to develop a new binaural SSL approach based on the EC model for practical binaural applications. In fact, in the review of Durlach [1], an SSL idea was briefly mentioned by steering the null to the sound source. However, there has been lack of a research to verify the feasibility of this idea as well as whether it is applicable in practice, though this question has been raised for about four decades. The original EC-BEAM method can be considered as a simple realization of Durlach's idea. It is based on a principle that the EC model is applied to cancel the target signal; as a result, the direction of sound source is specified by the null whose residual energy is minimal. The EC-BEAM is designed with a training stage to pre-calibrate EC parameters, i.e., the equalizers, to make the EC model work effectively on a given binaural system. Experimental results showed that the EC-BEAM can work excellently in a clean-anechoic

condition, and relatively well in low noise and low reverberation conditions. However, since this algorithm does not have a mechanism to account for interference, it suffers from high noise and/or high reverberant environments.

The proposed Adaptive EC-BEAM method is an approach to enable EC-BEAM to work in reverberant environments. It is based on the observation that when the reverberation is present, a model mismatch between training and estimation occurs, which may reduce the ability of the EC model in eliminating the reverberant target component. An adaption is performed by modifying the EC parameters accordingly to a certain reverberation level so as to successfully cancel all the sound signals produced by the target source. The reverberation level in this consideration is characterized by the DRR, which is assumed to be known in advance. This approach is consistent with the research of Shinn-Cunningham [68] about 'room learning' concept, in which the human auditory system adapts to the room condition to improve the perception. Through the experimental results, the Adaptive EC-BEAM has shown its effectiveness with real recorded reverberant sounds and simulated noisy reveberant data. Although these results are relatively promising, there remains two difficulties regarding to the applicability of this method in practice. First, the EC adaptation is empirically based on only DRR while the perception of reverberation may also depend on other factors, such as the reverberation time. The second difficulty concerns the problem of estimation of DRR in practice. Though various researchers have tackled the problem, estimating correctly DRR in arbitrary conditions is still a difficult task, especially when noise is present. Because of the assumption of prior known DRR, the applicability of Adaptive EC-BEAM would be limited. Therefore, the proposed Adaptive EC-BEAM approach seems to provide further understanding on the EC-BEAM and the effect of reverberation rather than to realize it in real applications.

The Weighed EC-BEAM approach is to make EC-BEAM robust against both reverberation and background noise with limited prior known information. This approach is opposite to the approach of Adaptive EC-BEAM, in which weighting functions were applied to the observed sound signals to reduce the affect of undesired factors rather than to modify the localization model to adapt certain data. Two strategies were suggested to deal with noise (EC-BEAM/N) and reverberation (EC-BEAM/R) separately. The EC-BEAM/N strategy relies on an assumption that the background noise is stable in time (WSS) and requires prior knowledge of a short noise-only period to construct the noise distribution weight at all interest directions. This idea comes from the observation in practice that the noise sources (regarding to background noise such as fans and air-conditioners) are always fixed and their energy is almost unchanged. The target sound signal (e.g. speech signal in robot controlling and human-robot communication)

may stop for some short periods and capturing a noise-only period is possible (e.g. using voice activity detection). On some point of view, the methodology of EC-BEAM/N is similar to the mechanism of 'room learning' in the study of Shinn-Cullingham [68], in which the adaptation is applied to noise instead of reverberation. The EC-BEAM/R strategy employs a weighting function to eliminate the effect of late reverberation based on the uncorrelation characteristic, without any prior known information. Although this strategy is to deal with reverberation, it is also effective with uncorrelated (diffuse) noise. The proposed Weighted EC-BEAM combines the EC-BEAM/N and EC-BEAM/R in an empirical way with regarding to the fact that their outputs are both residual sound energies. It is difficult to explain theoretically that the way of combining is the best way, as well as to prove such kind of combination is effective with both noise and reverberation. However, from the empirical results in various experimental conditions, the performance of Weighted EC-BEAM was verified.

In summary, this chapter proposed a new binaural SSL approach based on the EC model for localization in practical conditions with two objectives: the effectiveness on a binaural system and the robustness against noise and reverberation. The EC-BEAM algorithm is a simple realization of this idea and has achieved the first objective. However, it suffered from high noise and/or high reverberation conditions. Two suggested approaches, Adaptive EC-BEAM and Weighted EC-BEAM, are the solutions to make the EC-BEAM meet the second objective. Through the experimental results, the effectiveness of the two proposed approaches are verified and both objectives are achieved. Among these methods, the Weighted EC-BEAM seems more applicable as its assumption is easier to be satisfied in practice, and would be a typical realization of Durlach's idea. The requirements and advantages of these methods are summarized in Table 3.4.

Table 3.4: Summary of the proposed EC-based SSL methods.

Method	*Environment*	*Assumption*	*Requirement*
EC-BEAM	-Low noise -Low reverberation	-No	-No
Adaptive EC-BEAM	-Reverberation	-Diffuse reverberation ($\|R_L\| \approx \|R_R\|$)	-DRR
Weighted EC-BEAM	-Background noise -Late reverberation	-WSS noise -Uncorrelated reverberation ($R_L \perp R_R$)	-Noise-only period

Chapter 4

Applications of Weighted EC-BEAM

The effectiveness of the proposed binaural SSL approach based on EC model has been verified by various experiments in Chapter 3. In this chapter, it is further evaluated through two applications. Although both Adaptive EC-BEAM and Weighted EC-BEAM perform quite well in experimental conditions, the Adaptive EC-BEAM algorithm still remains some difficulties, such as prior knowledge of DRR, as discussed in Section 3.5. Therefore, only Weighted EC-BEAM is selected to use in this chapter. This algorithm will be applied in two applications: intelligent speech enhancement and blind source separation.

4.1 Application of Weighted EC-BEAM in speech enhancement

Speech enhancement is the problem of suppressing noise and other interference out of the observed sound, remaining only the target signal. This is an important task in signal processing field, of which one of the most popular applications is hearing aid. Among the variety of currently proposed methods, the two-stage binaural speech enhancement (TS-BASE) algorithm proposed by Li *et al.* [4] has been underlined because it can reduce noise with preservation of binaural cues, which is very important for speech perception and understanding. However, the TS-BASE as well as traditional speech enhancement methods focus on enhancing only the target signal and suppressing all other signals, while besides the target, humans also pay attention to other important or meaningful sounds (e.g., a call from someone) in daily conversations. This attention-mechanism to meaningful signals is seldom considered in the state-of-the-art signal processing systems. With the purpose to enhance the target and preserve the meaningful signal, we propose to apply

Weighted EC-BEAM algorithm to detect the direction of meaningful signal, extract and present it together with the target signal. This idea is implemented using the TS-BASE method, named as intelligent TS-BASE or iTS-BASE.

4.1.1 TS-BASE Algorithm

Two-stage binaural speech enhancement (TS-BASE) was firstly proposed by Li *et al.* [83] and consequently improved in [4]. Basically, the TS-BASE approach exploits the Equalization-Cancellation (EC) model and Wiener filter to enhance the target signal, as shown in Fig. 4.1. This algorithm consists of the following two stages:

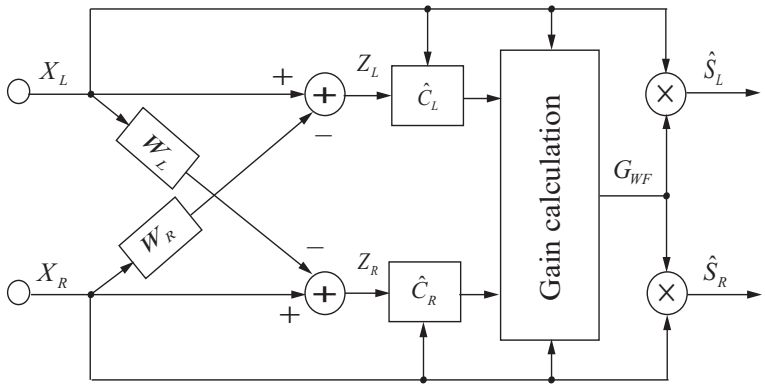

Figure 4.1: Block diagram of TS-BASE.

1. *Estimation of interference signals.*
   In this stage, the EC operations are applied to estimate the interference signals from the observed sound by eliminating the target signal component. This process is performed similarly to the null-steering procedure of the EC-BEAM at one interest direction. Specifically, suppose that the observed sound signal consists of a target and noise as follows:

$$Y_i(\omega, t) = X_i(\omega, t) + N_i(\omega, t), \quad i = L, R \tag{4.1}$$

   where ω denotes the frequency bin, t is the frame index, $X_i(\omega, t)$ and $N_i(\omega, t)$ are the target signal and noise at each channel respectively. In training, two equalizers, $W_L(\omega)$ and $W_R(\omega)$, are constructed using NLMS method to satisfy the following equations:

$$X_L(\omega, t) - W_R(\omega) X_R(\omega, t) \approx 0 \tag{4.2}$$

$$X_R(\omega, t) - W_L(\omega) X_L(\omega, t) \approx 0 \tag{4.3}$$

In order to estimate the noise component, the TS-BASE algorithm performs the cancellation operation by the following equations:

$$Z_L(\omega,t) = Y_L(\omega,t) - W_R(\omega)Y_R(\omega,t) \approx N_L(\omega,t) - W_R(\omega)N_R(\omega,t) \tag{4.4}$$

$$Z_R(\omega,t) = Y_R(\omega,t) - W_L(\omega)Y_L(\omega,t) \approx N_R(\omega,t) - W_L(\omega)N_L(\omega,t) \tag{4.5}$$

It can be observed that in Eq. (4.4) and Eq. (4.5), the target signal components $X_i(\omega,t)$ at both channels are completely removed, remaining only noise signal components. In this manner, the noise estimation factors, $Z_L(\omega,t)$ and $Z_R(\omega,t)$, are equivalent to the residual energy of null steering in the EC-BEAM algorithm. A time-variant frequency-dependent compensation factor, $B_i(\omega,t)$, $(i = L, R)$, is constructed to map the remained signal to the interfering components in the input mixture signals. $B_i(\omega,t)$ is derived by minimizing the mean square error between the target-canceled signal and the input mixture signal under the assumption of zero correlation between the target signal and the interfering signals, formulated as

$$\hat{B}_i(\omega,t) = \arg\min_{B_i} E[Y_i(\omega,t) - Z_i(\omega,t)B_i(\omega,t)], \quad i = L, R \tag{4.6}$$

where E is the expectation operator. Based on Wiener theory, $B_i^{opt}(\omega,t)$ is given by:

$$B_i^{opt}(\omega,t) = \frac{\Phi_{Y_iZ_i}(\omega,t)}{\Phi_{Z_iZ_i}(\omega,t)}, \quad i = L, R \tag{4.7}$$

where $\Phi_{Y_iZ_i}(\omega,t)$ denotes the cross-correlation spectrum of $Y_i(\omega,t)$ and $Z_i(\omega,t)$; and $\Phi_{Z_iZ_i}(\omega,t)$ is the auto-correlation spectrum of $Z_i(\omega,t)$.

2. *Enhancement of the target signal.*
The time-variant frequency-dependent compensation factor obtained in the first stage is used to construct a gain function to enhance the observed sound signal. The gain function is constructed by adopting the Wiener filter:

$$G(\omega,t) = \frac{\xi(\omega,t)}{1+\xi(\omega,t)}, \tag{4.8}$$

where $\xi(\omega,t)$ is *a priori* SNR calculated as:

$$\xi = \frac{E[X_LX_L^* + X_RX_R^*]}{E[(B_LZ_L)(B_LZ_L)^* + (B_RZ_R)(B_RZ_R)^*]} \tag{4.9}$$

in which ω and t are omitted for simple writing, the superscript * is the conjugation operator. The estimate of the *a priori* SNR, $\xi(\omega, t)$, is updated in a decision-directed scheme that significantly decreases the residual musical noise.

4.1.2 Proposed intelligent speech enhancement system

To construct an intelligent TS-BASE, a conceptual model is proposed as shown in Fig. 4.2, including two main parallel processes: (1) The first process implements the original TS-BASE to enhance the target signal from a specific direction. The advantage of the TS-BASE applied in this process is that it can extract a signal from a prior known direction and can deal with multiple non-stationary noise sources. As expected, the result from this process is only the signal from the target direction while the signals from other directions should be suppressed; (2) The second process attempts to detect and extract the meaningful signal from the non-target direction, which is considered as important to the listener. It is strictly required that this process must be concurrently performed and share the same input with the first process. Moreover, the binaural cues of the meaningful signal must be preserved because they may be important in some certain cases. One typical example is when someone hears a sound from a car hooter, he/she should be able to guess where the car is.

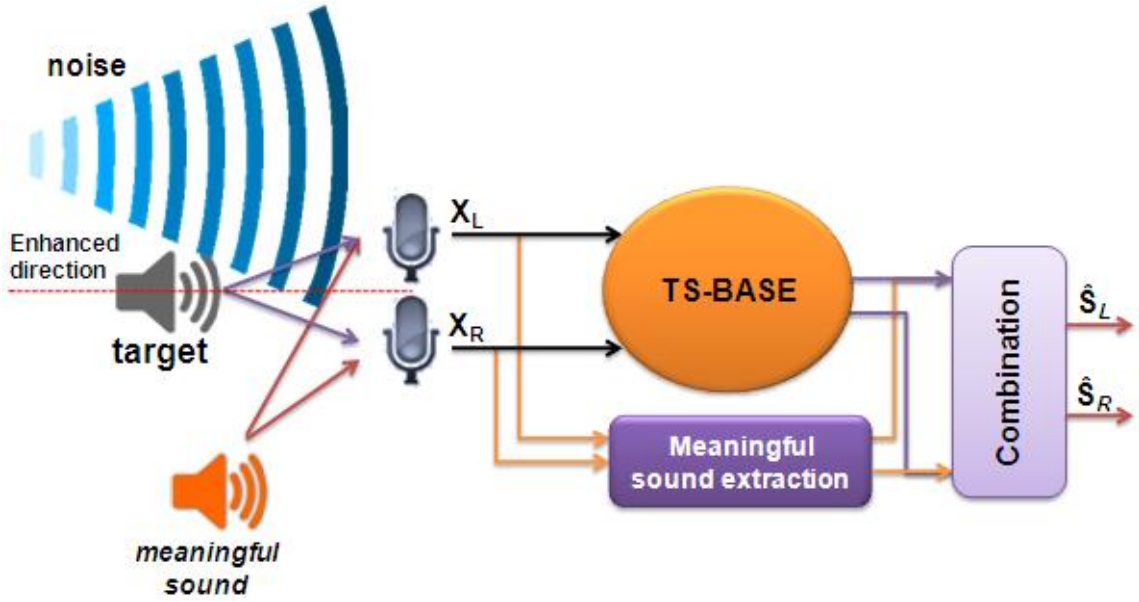

Figure 4.2: The conceptual model of the proposed intelligent TS-BASE system.

The key factors in this study are detecting and extracting the meaningful sounds which have never been considered by the state-of-the-art speech enhancement systems. In real-

world environments, there are a huge number of meaningful sounds including speech (e.g., a call from someone) and non-speech (e.g., telephone ring, sound of a car hooter, sound of a fire alarm, etc.). In principle, it is an extremely difficult problem to determine which sound is meaningful among a vast of mixture sounds because it is highly dependent on the situation that the subject perceives the sounds. Although for general sounds, there are a lot of properties that attract the human's perceptual attention, a meaningful signal normally has following physical characteristics:

- *Strong energy:* The meaningful signals that humans are interested in are usually strong in intensity. This is because the weak sounds will be masked by other stronger sounds in practical environments.

- *Enough temporal duration:* The meaningful sounds are normally long enough for humans to perceive. Too short sounds are difficult to be recognized by humans.

- *Sudden occurrence:* Some meaningful sounds (e.g., telephone ring) occur with sudden increase in energy that easily attracts the attention of a human in practice.

Within this scope, only the first two characteristics are considered.

4.1.3 Implementation of iTS-BASE

The proposed iTS-BASE approach consists two components: the original TS-BASE for enhancing the target signal, and the meaningful signal extraction, as in Fig. 4.3. This section will focus on how a meaningful signal is extracted.

We define a non-target signal as a meaningful signal if it satisfies the following two characteristics: (1) its energy is strong enough (e.g., larger than some threshold); (2) its duration is long enough (e.g., for a certain duration). The meaningful signal extraction process is performed through three steps:

- *DOA estimation of the meaningful signal:*
  The Weighted EC-BEAM algorithm is employed to localize all candidates of meaningful signals. Besides the good performance of Weighted EC-BEAM and TS-BASE, another reason that they are combined here is that both algorithms are based on EC model; therefore, they can share the same equalizers. As the meaningful signals are different from the target signal, the DOA of the meaningful signals is determined by scanning the non-target directions through an EC-based beamforming.

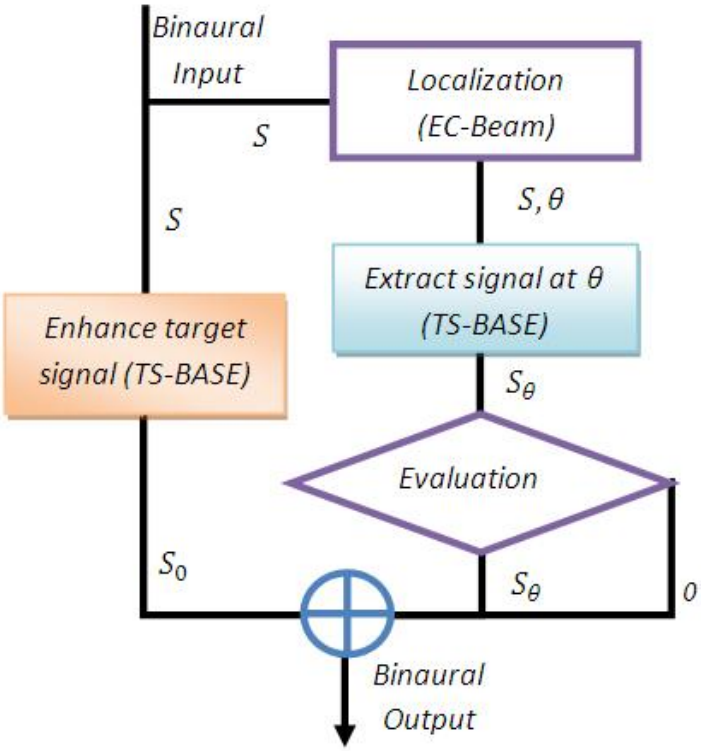

Figure 4.3: Flowchart of the proposed iTS-BASE.

- *Detection and extraction of the meaningful signal:*
  After being localized by Weighted EC-BEAM, the candidates of meaningful signals are extracted by using the TS-BASE algorithm. According to the criteria for meaningful signals, it is necessary to further judge whether the extracted signals are meaningful or not. Specifically, this process checks whether the extracted signals are larger than a predefined threshold in intensity and longer than a certain duration. In the current implementation, we consider only one meaningful signal at a time and the threshold in intensity was set to 1/10 power of the whole noisy signal, which means that the power ratio of the extracted signal to the whole signal is -10 dB, and the threshold in duration is to 0.1 second. The output of the meaningful signal extraction will be the extracted signal if it satisfies all criteria; otherwise, the output will be zero.

- *Enhancement of the target and meaningful signals:*
  The output of the proposed iTS-BASE algorithm is finally generated by combining the output of the original TS-BASE algorithm (the enhanced target signal), and the output of the meaningful signal extraction.

4.1.4 Experimental Evaluation

The experiments in this section were conducted in a simulated situation in which the target speaker was located in front of a listener and another person called him/her from behind (i.e., the meaningful signal). The target signal was the utterance selected from the ATR database [73] and the meaningful signal was a recorded sound of the speech "hello". The sound signals above were convolved with the HRTF measurement from the KEMAR database [92] to produce binaural sounds. Binaural background noise was recorded at a cafeteria using two microphones at two ears of a dummy head. The target signal was set to 0° (in front of the listener), while the direction of the meaningful signal was set to 60° . The amplitude of the meaningful signals was controlled to make the energy ratio of the meaningful signal to the target signal (MTR) at -3 dB and 0 dB, respectively. The mixture of the target and meaningful signals was then considered as the clean signal to be estimated. The noisy signal was generated by adding the recorded cafeteria noise into the mixture of the target and meaningful signal at SNRs of 0; 5; 10 and 15 dB.

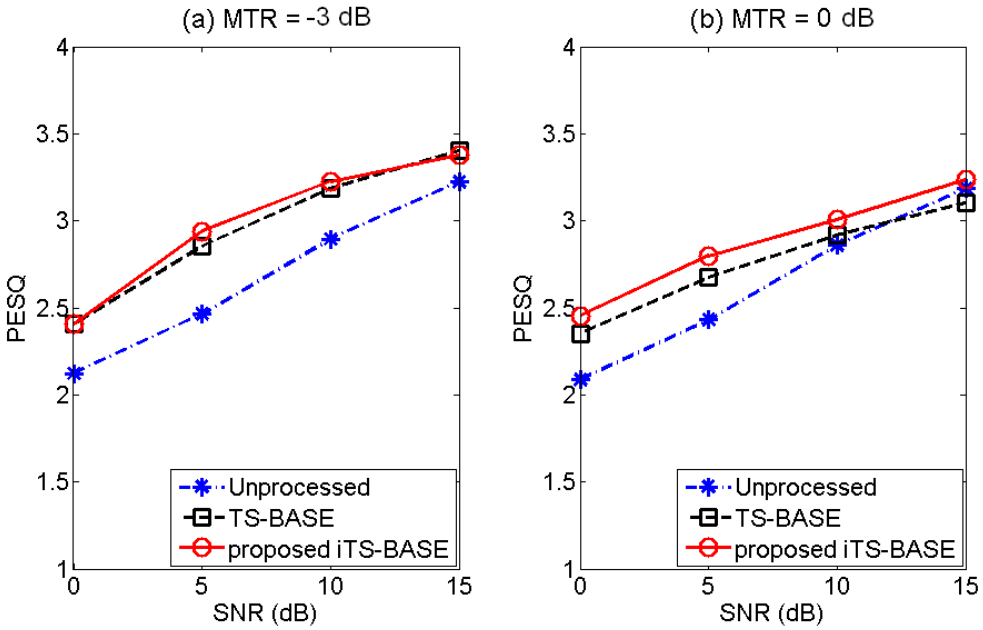

Figure 4.4: Experimental results in terms of perceptual evaluation of speech quality (PESQ) of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTS-BASE algorithm.

In the results, the Weighted EC-BEAM detected the meaningful signal correctly at 60° in all cases. The performance in terms of meaningful signal extraction and target enhancement of the iTS-BASE algorithm was evaluated by two measures: the perceptual evaluation of speech quality (PESQ) [84] and the log-spectral distance (LSD), in which

high PESQ and low LSD are the indications of a good performance. Fig. 4.4 shows the evaluation of PESQ. This value of the iTS-BASE algorithm is generally higher than that of the TS-BASE algorithm, which means that the performance of the iTS-BASE algorithm is better than the original TS-BASE algorithm in improving the speech quality. Both the TS-BASE and iTS-BASE algorithms provide relatively higher PESQ improvements comparing with the unprocessed noisy inputs. It can be observed that the PESQ of iTS-BASE in the case MTR = 0 dB is consistently above the other PESQs. This is because the energy of the meaningful signal in this case is relatively strong, so the improvement is quite clear. In both MTR condiitons, when the SNR becomes high (or the noise becomes low), the performance of the TS-BASE gets worse. The reason is that the clean signal contains signals from two separate directions: the target signal is from 0^o and the meaningful signal from 60^o, but the TS-BASE can enhance the signal from only one direction (the target) and tends to reduce the signal from other directions, including the meaningful signal. When the noise is low, the suppressed non-target signal is mainly from the meaningful signal. In contrast, by enhancing the target signal and extracting the meaningful signal at the same time, the iTS-BASE performs well and stable for almost all SNR levels.

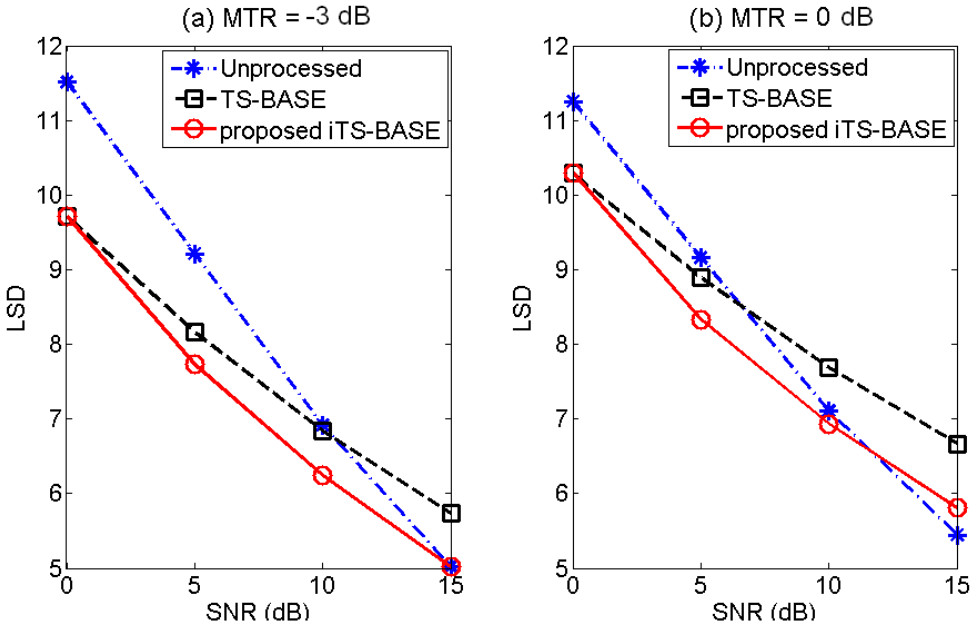

Figure 4.5: Experimental results in terms of log-spectral distance (LSD) of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTSBASE algorithm.

Similar results can be observed in the evaluation in term of LSD measurement, as plotted in Fig. 4.5. The TS-BASE algorithm is only able to improve the LSD in low SNR conditions and becomes worse in high SNR conditions due to the fact that it suppressed

not only the noise but also the meaningful signal. On the other hand, the iTS-BASE algorithm generally remains good performance at all SNR levels. It is noticed that in both cases of MTR, the LSD values of TS-BASE and iTS-BASE are nearly the same when SNR = 0 dB. The possible reason is that at this SNR the noise is much larger than the meaningful sound, so the extracted signal is not significant in comparing with the residual noise. However, when SNR is increased, the iTS-BASE algorithm performs gradually better than the TS-BASE algorithm.

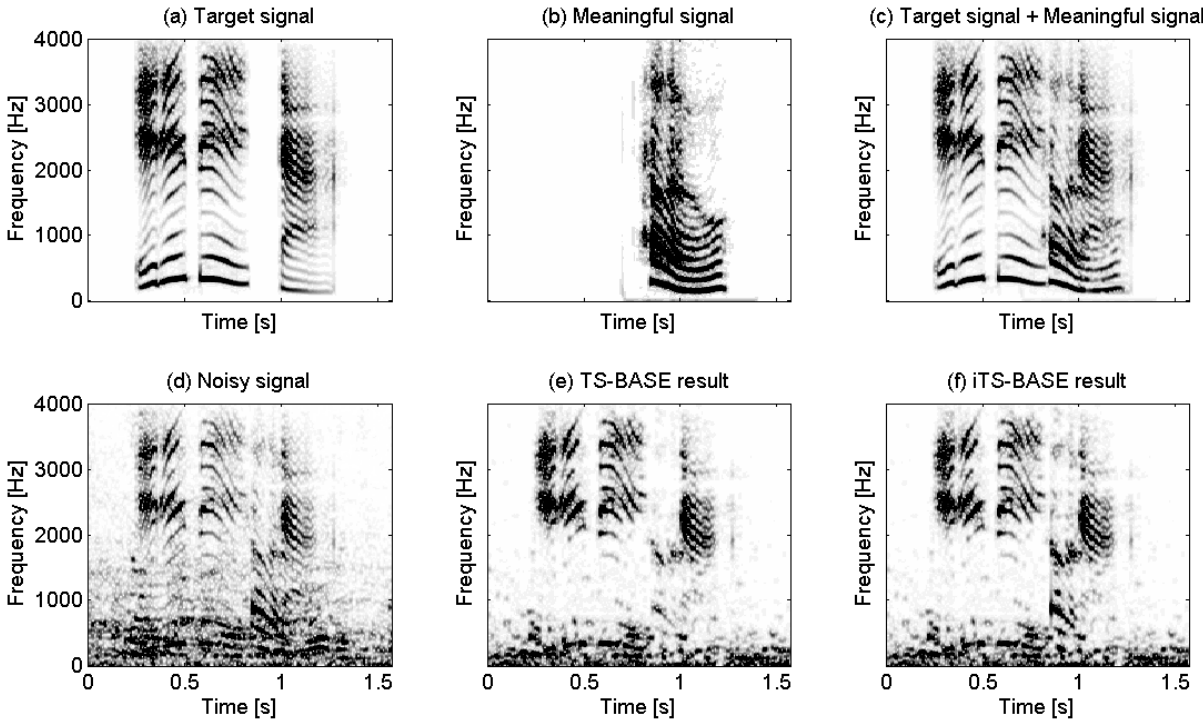

Figure 4.6: Spectrograms of the target signal, the meaningful signal, the target+meaningful signal, the noisy signal, the signals enhanced by the TS-BASE and the iTS-BASE algorithms.

Fig. 4.6 further visualizes the performance the TS-BASE and the proposed iTS-BASE algorithms via spectrograms in the case MTR = 0 dB. It can be observed that noise is significantly reduced in the signal enhanced by the TS-BASE algorithm (Fig. 4.6e). However, the meaningful signal is also suppressed to a large degree, which may lead to the fact that the listener cannot perceive the important sound. On the other hand, the output spectrum from the iTS-BASE algorithm shows that the good noise reduction performance is still maintained while the meaningful signal is also preserved (Fig 4.6f). From these results, the effectiveness of the proposed iTS-BASE algorithm as well as the applicability of the Weighted EC-BEAM algorithm are confirmed.

4.2 Application of Weighted EC-BEAM in blind source separation

Sound separation is related to the problem of separating sounds coming from multiple sources in a mixture signal. Conventional sound separation methods usually require information of the number of sound sources, and/or the number of microphones must be more than the number of sound sources. On psychoacoustic aspect, the human ability to perceive sound in noise can be explained by the BMLD [74]. The BMLD was successfully simulated by the EC Theory [14], on which the EC-BEAM algorithms were proposed in Chapter 3. With the purpose to verify the applicability of the Weighted EC-BEAM in the problem of sound separation, this section introduces a new blind source separation method by using the Weighted EC-BEAM for DOA estimation the TS-BASE algorithm for signal extraction.

4.2.1 Blind sound separation overview

Blind sound separation (BSS), also known as blind source separation, is the problem of finding out the original signals in a set of mixed signals without the aid of information (or very little information) about the source signals or the mixing process. A typical example is the "cocktail party problem", where one is talking with his friend and numerous conversations are occurring at the same time around him, he has the ability to focus his attention on his friend's speech [20].

BSS relies on the assumption that the source signals do not correlate with each other. For example, the signals may be statistically independent or decorrelated. BSS thus separates a set of signals into a set of other signals, so that the regularity of each resulting signal is maximized, and the regularity between the signals is minimized (i.e. statistical independence is maximized). In speech processing, BSS has been widely applied in speech intelligibility enhancement, noise reduction, hearing aids and cochlear implants.

Although humans can effectively separate and focus on any sound in a mixture, most of the BSS methods require a microphone array in which the number of microphones must be larger than the number of the sources to be separated. On the other hand, the binaural cues, which are very important in speech intelligibility enhancement systems and hearing aids, have not been considered in the state-of-the-art methods in this field.

The crucial approach in BSS replies on a microphone array. Mathematically, assume

that there is a source signal vector:

$$s(k) = [s_1(k)s_2(k)...s_m(k)]^T \tag{4.10}$$

where m is the number of sources and $s_i(k)$ is the i-th signal source. These source signals are passed through an $(m \times n)$ linear, time-invariant system with a matrix impulse response $A_i(0 < i < \infty)$, the mixture of the obtained signals can be expressed as follows [85]:

$$y(k) = [y_1(k)...y_n(k)]^T = \sum_{i=0}^{\infty} A_i s(k-i) \tag{4.11}$$

The goal of the method in this approach is to find out the sequence of the $(m \times n)$ matrices B_l so that each source $s_i(k)$ can be uniquely extracted from the mixture:

$$\widehat{s}(k) = \sum_{l=0}^{\infty} B_l y(k-l) \tag{4.12}$$

in which $\widehat{s}(k) = [\widehat{s}_1(k)...\widehat{s}_m(k)]^T$ is the output vector sequence containing the estimates of individual sources.

So far, there have been many research investigating this issue, which can be found in [86, 87]. Currently, preferred methods in blind signal separation include:

- Principal components analysis (PCA)
- Singular value decomposition (SVD)
- Independent component analysis (ICA)
- Dependent component analysis (DCA)
- Short-time Fourier transform (STFT)
- Degenerate unmixing estimation technique (DUET)
- W-disjoint orthogonality
- Joint approximate diagonalization eigen-matrices (JADE)
- Computational auditory scene analysis (CASA)
- Constant modulus algorithm (CMA)

In these research, the problem was considered in both linear and non-linear deconvolution. However, it is not clear whether human binaural processing is linear [88]. Furthermore, the number of sources m in the convolutive BSS must not be greater than the number of microphones n, whereas it is often the case in human binaural processing that the number of sources m far outnumbers the two ears used to collect acoustical information.

An other approach to BSS is based upon the DOA of sound sources, namely directional BSS [89], and is also called multiple signals extraction, in which BSS was regarded as a set of beamformers whose response is constrained to a set of angles $\theta = [\theta_1, ..., \theta_M]$ for recovering all m sources from the mixture. Blind beamforming technique was also investigated as an alternative method in case the number of microphones is less than the number of sound sources [90]. Also in [90], a binaural approach for BSS was mentioned. However, the problem of preservation of the binaural cues in BSS has not been considered in these methods. This will be a gap in this problem when BSS is carried out in some practical applications, especially in speech intelligibility enhancement and hearing aids.

4.2.2 BSS using Weighted EC-BEAM and TS-BASE

The proposed system includes two main stages: source detection by Weighted EC-BEAM and source extraction by TS-BASE.

Sources detection

The Weighted EC-BEAM algorithm is extended to localize multiple sources as described in Section 3.1.2. Specifically, the null-steering process is performed in the same way to that of the Weighted EC-BEAM in Section 3.4. After that, spline interpolation [91] is further applied to acquire the residual energy at non-steering directions. This step is to increase the resolution of searching in case the equalizers could not be trained at all directions due to the limit of the training data, and to reduce the computational complexity. Finally, the set C of DOAs of sound sources is specified by three criteria:

- Local minimum of null energy:

$$C = \{\theta_i | C_Y(\theta_i) \leq C_Y(\theta_{i-1}) \text{ and } C_Y(\theta_i) \leq C_Y(\theta_{i+1})\} \tag{4.13}$$

- Separate from other candidates:

$$\forall \theta_i, \theta_j \in C, |\theta_i - \theta_j| \geq T_\theta \tag{4.14}$$

- Low residual energy:

$$\forall \theta_i \in C, C_Y(\theta_i) \leq T_E \tag{4.15}$$

In the following experiment, T_θ is set to 5^o and $T_E = E[C_Y(\theta_i)]/2$, where $E[.]$ denotes the mean operator.

Sources extraction

The TS-BASE is employed to extract (separate) signals at the DOAs detected in *sources detection* stage. The extraction process is performed in the same way as that described in the iTS-BASE algorithm. Both the Weighted EC-BEAM and the TS-BASE algorithms used the same set of equalizers as they are based on the EC mechanism. However, in extraction, the TS-BASE used the nearest equalizer to the detected DOA if the equalizer at that DOA is not available.

4.2.3 Experiments and results

Experimental configuration

- *Data:* For speech, utterances of Japanese speakers were selected from ATR database [73]. To obtain directional sounds, the HRTF database from MIT Media lab [92] was used again. The speech data were first up-sampled to 44.1 kHz and convolved with the HRTF, then down-sampled to 20kHz.

- *Simulation:* 'Cocktail party' signals were simulated by mixing the directional signals together. The mixture consists of three directional speech sentences from three Japanese speakers at the directions -50^0, 0^0, 60^0 respectively. These scenario of simulation is demonstrated in Fig. 4.7.

- *Implementation:* The equalizers were trained at the directions from -90^o to 90^o (step of 5^o). In the stage of DOA estimation, interpolation was applied to get a searching resolution of 1^o.

Results and discussion

In the results, there were three sources detected at direction -51^0, 0^0 and 60^0, which are almost exactly the true DOAs. These sources were then extracted using the TS-BASE algorithm. Fig. 4.8 (a), (b), (c) and (d) show the spectra of the original signals at the

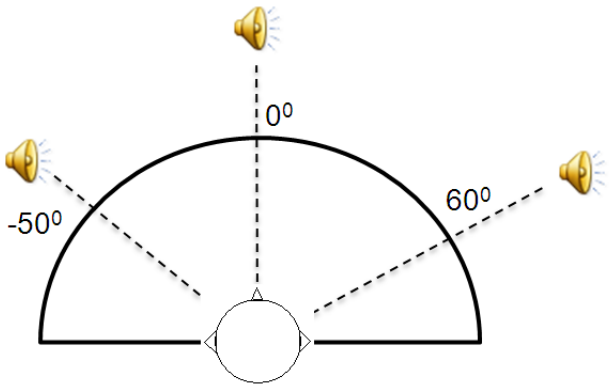

Figure 4.7: Directions of the mixed signals

directions 60^o, 0^o, -50^o and the spectrum of the mixture, respectively. The spectrograms of the extracted signals are shown in Fig. 4.8 (e), (f) and (g). It can be observed that the separated signals have similar patterns as those of the original ones, and are much improved compared with the mixture. Moreover, the binaural cues of the extracted signals were also preserved due to this ability of the TS-BASE algorithm [4].

4.3 Summary

The main purpose of this chapter is to verify the performance of the proposed EC-based binaural SSL approach via applications in noisy reverberant conditions. Two binaural applications were selected, including binaural speech enhancement for hearing aids and binaural blind source separation, as they are quite related to the issue of sound localization. The Weighted EC-BEAM algorithm was chosen to be used in the applications because its implementation is easier than that of the Adaptive EC-BEAM.

In terms of binaural speech enhancement for hearing aids, most of the methods tackle this problem by enhancing the target signal and suppressing all of the other signals. However, in daily life, humans listen to not only the target signal but also the non-target but useful signals to react properly in each conditions, for example avoiding a coming car. These signals, namely meaningful signals, have not been considered in previous speech enhancement systems. In the first application, an intelligent binaural speech enhancement system (iTS-BASE) was suggested to deal with this issue by enhancing both the target signal and the meaningful signals, with the preservation of binaural cues. In order to realize this idea, besides the target enhancement process, an additional process is designed to extract and present the meaningful signal at the final output. The extraction process was performed by first applying the Weighted EC-BEAM to localize the meaningful

source, then using the TS-BASE algorithm [4] to extract the signal at the detected DOA, and finally evaluating whether the extracted signal is meaningful to present at the final output.

In the second application, the Weighted EC-BEAM algorithm was applied in binaural sound separation. Traditional sound separation methods normally require that the number of microphones is greater than the number of sources. The method of Parra *et al.* [89] can separate sources without the constraint of the number of microphones; however, it requires the prior known DOAs of sources. Following this approach, we suggest a source separation method using the Weighted EC-BEAM and TS-BASE algorithms, in which the Weighted EC-BEAM was used to localize all the sources and the TS-BASE algorithm was applied to extract the sound signals at the detected DOAs. This results in a new blind source separation, which can separate sources using only two microphones without any requirement of prior known information.

In the experimental results, the Weighted EC-BEAM has shown its good performance in terms of localization. Specifically, it is able to localize correctly the direction of meaningful signal in noisy reverberant environment. This contributes to the results that the meaningful signal was extracted and presented at the final output while the performance of the target enhancement as well as the noise suppression was maintained. In the sources separation process, the Weighted EC-BEAM can almost localize three sources at the same time accurately. This is important because DOAs of sources is the crucial factor deciding the performance of a blind source separation based on beamforming. From this result, the Weighted EC-BEAM algorithm may be able to improve the method of Parra *et al.* [89]. More importantly, through out the results of the two applications, the effectiveness of the Weighted EC-BEAM in noisy reverberant conditions is once again verified.

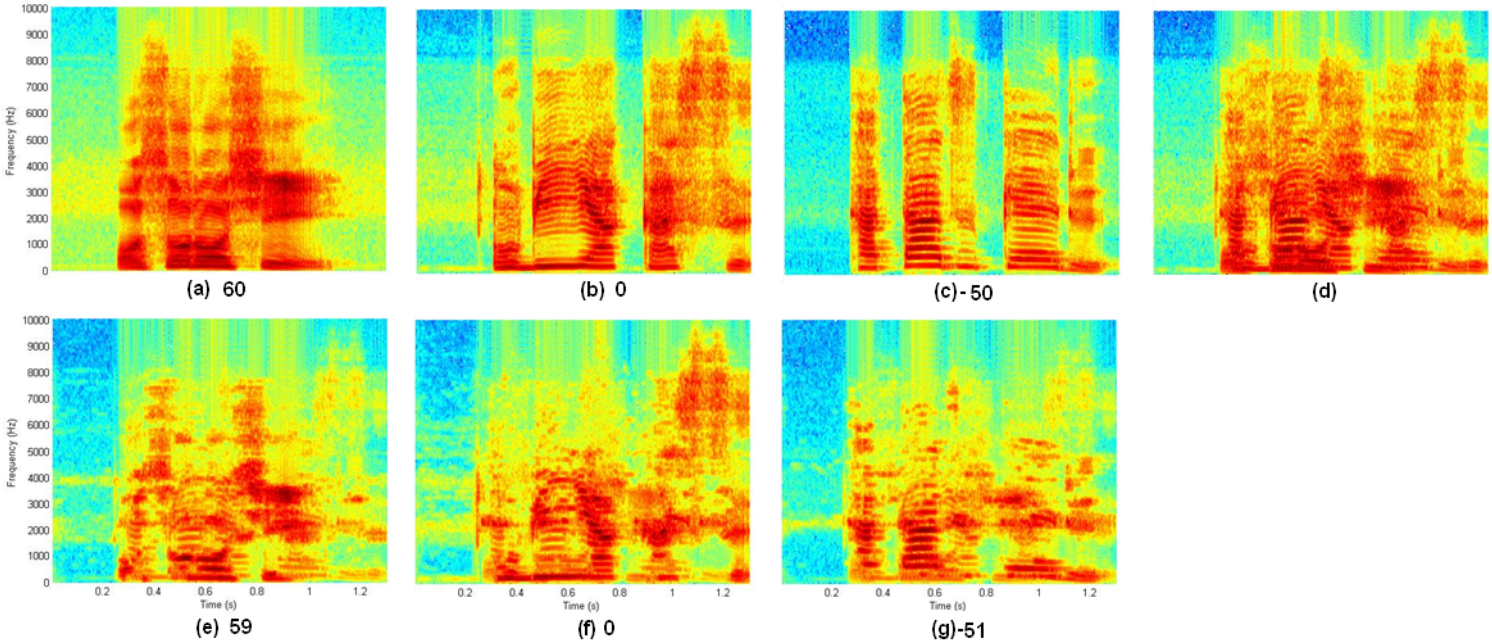

Figure 4.8: Spectrograms of the individual signals (a,b,c), the mixture (d), and the separated signals (e,f,g).

Chapter 5

Conclusion

5.1 Summary

In the past decades, much attention has been paid on the issues of how humans process and perceive sounds, and how this mechanism can be applied in practice. This thesis aims to link the findings from binaural hearing modeling with the applications in signal processing. Specifically, we tackle the problem of binaural sound source localization (SSL) in noisy reverberant environment based on the equalization-cancellation (EC) model.

Binaural SSL in practical environments is a challenging task because the observed sound is corrupted by noise and reverberation, while the input is limited to only two receivers under the effect of HRTFs. In the psychoacoustic research field, one of the binaural interaction theories to explain the mechanism of human perception in such conditions is the EC model of Durlach [1, 14]. The explanation of the EC model in the cocktail party conditions is consistent with the perception of human, which reveals the potential of EC model for the issue of sound localization in the presence of interference. This motivated our study to investigate a binaural SSL approach based on the EC mechanism with two objectives: (1) effectiveness on binaural systems, and (2) robustness against noise and reverberation.

The general idea, namely EC-BEAM, is conceptually a combination of an EC model and a beamforming technique. The EC-BEAM steers a null to each interest direction by utilizing the EC procedures to eliminate the target sound signal component. The direction of sound source is determined via looking for the null with minimal residual energy. The EC-BEAM is designed with a training processing to construct the equalizers beforehand. This guarantees that the EC-BEAM is able to work effectively with the internal effect. Experimental results showed that this method can work well in low noise and low

reverberant conditions, by which the first objective is achieved. The proposed method was further considered to localize a sound source under the effects of noise and reverberation regarding to the second objective. Two solutions have been suggested to make the EC-BEAM robust against these effects, namely Adaptive EC-BEAM and Weighted EC-BEAM.

The Adaptive EC-BEAM approach improves the SSL performance in reverberant conditions by adapting the EC model to the reverberation level in rooms. The hypothesis is that the equalizers, which were trained in an anechoic condition, are not able to equalizer the observed signals in an unknown reverberant condition, leading to the fact that the target signal cannot be eliminated. In the Adaptive EC-BEAM, for each applied condition, the equalizers are adjusted corresponding to the direct-to-reverberant energy ratio (DRR). This modification is to make the equalizers fully compensate for both the target and its reverberation at the two receivers, so that these components can be completely suppressed. Experimental results with real recorded data and various simulated room conditions showed that the Adaptive EC-BEAM performed significantly better than the original EC-BEAM algorithm. Although these results are quite promising and satisfy a part of the second objective, this approach still remains the difficulty of DRR estimation before performing SSL. Therefore, the proposed Adaptive EC-BEAM approach might provide more understanding of the EC-BEAM and the effect of reverberation rather than to be ready for an application in practice.

The Weighted EC-BEAM approach enables the proposed method to work in noisy reverberant conditions by using two weighting functions (corresponding to two strategies) to deal with background noise and reverberation separately. The first strategy (EC-BEAM/N) to deal with background noise is based on the assumption that the background noise is stationary over time and it is possible to extract a short noise-only period, such as using voice activity detection. A weighting function was used to equalize the residual noise energies from the null-steering process at all directions. This weighting function can be understood as a set of coefficients characterizing the distribution of noise in the searching space. This strategy also makes the SSL method fit to a binaural system better as the internal effect is indirectly included in the weighting function. The second strategy focuses on eliminating the effect of late reverberation (EC-BEAM/R) based on the assumption that the late reverberation components are uncorrelated with the target and together uncorrelated at both receivers. In this strategy, the C process is performed without ILD compensation. This strategy may also be effective with diffuse noise as it also satisfies this assumption. The proposed Weighted EC-BEAM is the empirical combination of the two strategies by summing their 'normalized' residual energies. Although it is hard to

explain theoretically how both strategies can work well when both noise and reverberation are present as well as whether the suggested combining method is the best way gathering them into one algorithm, the Weighted EC-BEAM has expressed an excellent performance in various noisy reverberant conditions and satisfied both objectives of this thesis.

The proposed EC-based SSL approach was further evaluated via applications. The Weighted EC-BEAM algorithm was selected to be used in two binaural applications, speech enhancement and sound separation, because of its simple requirement. In the first application, the purpose is to develop a binaural speech enhancement method for hearing aids, which not only enhances the target sound but also preserves the meaningful signal. The Weighted EC-BEAM was used to localize the candidate of meaningful signal, whose energy is the largest among the non-target signals. This candidate was then extracted using the TS-BASE algorithm [4] and evaluated whether it is really a meaningful sounds. Finally, the verified meaningful signal and the target signal, which had been enhanced by the TS-BASE in another parallel process, was combined and presented in the final output. In the application of sound separation, the Weighted EC-BEAM algorithm was used to estimate the DOAs of all sound sources. Then these sources were extracted (separated) by employing the TS-BASE algorithm again. This resulted in a blind source separation with neither the requirement that the number of microphones must be larger than the number of sources nor the DOAs of sources are known in advance. Experimental results showed that the Weighted EC-BEAM algorithm is able to estimate the DOAs accurately in both cases: the meaningful signal in the first application and the target sources in the second application. As a result, the TS-BASE algorithm was successful in extracting these signals with preservation of their binaural cues. More importantly, through out these results, the applicability of the proposed EC-based binaural SSL approach was further confirmed.

5.2 Contributions

This study is originally inspired by the binaural processing abilities of the human auditory system, particularly the selective hearing ability in the cocktail party conditions, with the main focus on sound localization and the EC model. The overall contribution of this thesis is a realization the EC model in a binaural SSL approach, which can be applied in practical noisy reverberant environments. This is a connection between the field of binaural hearing modeling and the field of sound signal processing in the series of research on the human perception and its applications. Inside the whole work, three subsequent contributions were derived. The first subsequent contribution is the general EC-BEAM algorithm, which gathers the EC model and a beamforming technique into

an SSL method. This is an important starting point to answer the question how the EC model can be applied in localization. In fact, the mechanism of the EC-BEAM is quite similar to the procedure of the SRP-PHAT [28], in which the CC model is replaced by the EC model and we look for the minimized residual energy of a steered null instead of the maximized beamforming energy. The EC-BEAM was further extended to multiple sources localization by defining several criteria for candidates of sound sources. Through experimental results, the feasibility of the proposed EC-BEAM algorithm was verified in ideal and low noise/reverberant conditions.

The general EC-BEAM algorithm was extensively considered in terms of implementation in practical noisy reverberant environments. This consideration yielded the second subsequent contribution of this thesis, which provided two approaches to make the EC-BEAM robust against environmental factors: Adaptive EC-BEAM and Weighted EC-BEAM. The Adaptive EC-BEAM approach introduced a new concept of an adaptive EC model to deal with the effect of reverberation by using DRR. In the proposed adaptive EC model, the equalizers are adjusted for each applied condition to overcome the mismatch problem of the training-based EC models suggested by Roman *et al.* [59] and Li *et al.* [4]. Although the Adaptive EC-BEAM is still not ready for applying in practice because the estimation of DRR has not been given, it provides a theoretical way to account for reverberation in the EC model as well as in the EC-BEAM. Moreover, as the relationship of the ideal equalizer and DRR was formulated in the adaptive EC model, it reveals a way to estimate the DRR by reversing the formulation, that is the DRR value that maximizes the performance of the EC operation is the true DRR of observed signal. In contrary to Adaptive EC-BEAM, the Weighted EC-BEAM approach is a practical way for a robust EC-BEAM algorithm. Weighted EC-BEAM provided two strategies to deal with noise (EC-BEAM/N) and reverberation (EC-BEAM/R) respectively. As the equivalence between EC-BEAM and Cross-channel HRTF method [8] or EC-BEAM/R and SRP-PHAT discussed in Section 3.4.5, it is interesting to investigate whether the proposed strategies can improve the previous algorithms.

The last subsequent contribution of this thesis is the applications of the proposed Weighted EC-BEAM to solve existing problems in binaural signal processing. The application of intelligent speech enhancement system introduced a new concept of meaningful signal, which has never been considered in previous speech enhancement systems. By applying the Weighted EC-BEAM, this system is able to not only enhance the target signal but also preserve the meaningful sound signal. This should be very benefit for practical systems, such as hearing aids. The application of the Weighted EC-BEAM in source separation provided a blind method to detect the DOAs of the target sources.

From this step, the sources can be blindly extracted and separated. Comparing with previous sound separation studies, the proposed BSS method uses only two channels without any requirement of prior knowledge about the sound sources. Moreover, although only speech enhancement and sound separation have been considered in this thesis, the proposed approach should be applicable to various binaural applications, especially ones that are related the sound localization issue.

5.3 Suggestions for further research

Within this thesis, the EC model has been taken into the problem of binaural sound localization. A general SSL method based on EC model has been proposed and two approaches to improve it in noisy reverberant environments have been suggested. Although the proposed methods performed well in experimental conditions and the Weighted EC-BEAM algorithm has been successfully applied in speech enhancement and sound separation, there are still rooms for further investigation related to this topic. The first possible issue to be investigated is the estimation of DRR, concerning the applicability of Adaptive EC-BEAM. Once the DRR is estimated, the adaptive EC model and adaptive EC-BEAM can be executed in reverberant conditions. Moreover, the adaptive EC model may not be limited to only the EC-BEAM algorithm but should be applicable in other binaural processing applications. In fact, some attentions have been paid on DRR estimation recent years, such as the work of Lu *et al.* [27] and Hioka *et al.* [61]. It should be interesting to link these research to the adaptive EC model as well as the adaptive EC-BEAM algorithm. Concerning this issue, the relation between the ideal equalizer, which maximizes cancellation performance, and the DRR was formulated in Eq. (3.29). By reversing this relation, the DRR may be estimated. That is, for a fixed direction, given the equalizer obtained in anechoic condition and an observed signal in reverberant condition, the DRR value that maximizes the cancellation performance (i.e. minimizes residual energy) should be the true DRR of the observed signal.

The second issue may be further tackled is the problem of estimation noise-only segments, regarding to the applicability of the Weighted EC-BEAM. In practical applications, such as mobile robot, the localizing system may be movable, leading to the fact that the DOAs and the energy of the noise sources may change in time. Fortunately, these changes are normally slow and it is possible to online update these noise information to recalibrate the hearing models, particularly the localization model in this manner. In order to do that, a noise-only duration must be estimated, provided that the target sound, such as speech, are not always continue over time. This problem can be solved via applying

voice activity detection (VAD) methods. VAD is an important front-end task in a lot of signal processing and understanding applications, e.g. automatic speech recognition (ASR) [93, 94], real-time speech transmission on the Internet [95], etc. This issue has been extensively researched and various methods have been proposed, such as VADs based on energy thresholds [96], pitch detection [97], spectrum analysis [98], zero-crossing rate [99], periodicity measure [100], and methods combining these features. These methods achieved quite good results in various conditions, which are believed to be good enough for the Weighted EC-BEAM. It should be noticed that the VAD is to discriminate speech/non-speech frames while the requirement here is noise-only frame, which may be simpler than the VAD problem.

The third issue may be further researched is to investigate whether applying the mechanism of EC-BEAM/N and EC-BEAM/R improves other SSL methods, such as the cross-channel HRTF [8] and the SRP-PHAT [28]. As discussed in Section 3.4.5, the original EC-BEAM algorithm is quite equivalent to the cross-channel HRTF method. The cross-channel HRTF has an advantage in comparison with the EC-BEAM that it possesses accurate measured HRTFs. Theoretically, the HRTF describes the propagation of sound to each channels better than the equalizer, which is essentially a presentation of the interaural transfer function. Moreover, the HRFT at each channel contains the spectral information, which is the main cue for localization in the vertical plane (elevation). Because of the equivalence between two SSL methods, the cross-channel HRTF may also face the similar problems of the EC-BEAM in noisy reverberant environments. Therefore, the EC-BEAM/N and the EC-BEAM/R may be useful for it to be robust in such conditions. Another method can be reconsidered is the SRP-PHAT, which was specified as an equivalent method of the EC-BEAM/R. Based on the analysis in Section 3.4.5, it is possible to explain why the SRP-PHAT works extremely well in reverberant conditions, besides the explanation of Zhang *et al.* in [63]. However, the performance of SRP-PHAT in high noise conditions, particularly the directional noise, is still low because it does not have a mechanism to adapt to this effect. From this point of view, the EC-BEAM/N should be a promising solution for the SRP-PHAT to be further improved in practical noisy reverberant environments.

Appendix A

Signal model

In the time domain, when a sound signal $s(t)$ is emitted at the target source, the signal observed at each receiver (ear or microphone) can be presented as follows:

$$x_i(t) = a_i s(t - t_i),\ i = L, R, \tag{A.1}$$

where a_i and t_i are the attenuation coefficient and the time delay describing the propagation of sound from the source to the left receiver ($i = L$) or the right receiver ($i = R$). Hereafter, the receiver index i will be omitted for simplicity. In signal modeling, the propagation of sound can be presented by a transfer function $h(t)$, by which the observed sound is expressed by the convolution of $h(t)$ and $s(t)$,

$$x(t) = h(t) \circledast s(t), \tag{A.2}$$

where $\circledast$ denotes the convolution operation.

In enclose space, the observed signal additionally includes the reverberation instead of the direct signal component $x(t)$ only. Reverberation is built up from multiple reflections, which are the sounds reflected from walls, floor or other objects in the room. When the sound $x(t)$ arrives at the receiver, the reflections of previous sound signals also arrive later with some delays. The reflections are weaker than the direct signals due to the absorption of air and reflecting surfaces. A reflection with a time delay τ can be represented based on the direct observed signal by $\alpha(\tau)x(t - \tau)$, where $\alpha(\tau)$ is a coefficient describing the attenuation of the reflection in comparison with the direct signal ($\alpha(\tau) < 1$). As a result,

the total reverberation component can be presented as follows:

$$r(t) = \int_{0^+}^{\infty} \alpha(\tau) x(t-\tau) d\tau. \tag{A.3}$$

In the time-frequency domain, Eq. (A.2) can be rewritten by:

$$X(\omega, t) = H(\omega) S(\omega, t), \tag{A.4}$$

where ω denotes the frequency bin index, $S(\omega, t)$, $H_0(\omega)$ and $X(\omega, t)$ respectively represent the signal at the source, the transfer function and the observed signal in the time-frequency domain respectively. These components can be obtained by applying the short-term Fourier transform to the corresponding components in the time-domain. Similarly, the reverberation component specified by Eq. (A.3) is represented as follows:

$$R(\omega) = \int_{0^+}^{\infty} \alpha(\tau) X(\omega) e^{-j\omega\tau} d\tau = \left[\int_{0^+}^{\infty} \alpha(\tau) e^{-j\omega\tau} d\tau \right] X(\omega). \tag{A.5}$$

Note that $\alpha(\tau)$ depends on only the room condition and is independent of the observed time t. By denoting

$$\psi(\omega) = \int_{0^+}^{\infty} \alpha(\tau) e^{-j\omega\tau} d\tau, \tag{A.6}$$

the reverberation component can be expressed by:

$$R(\omega) = \psi(\omega) X(\omega). \tag{A.7}$$

In Eq. (A.7), $\psi(\omega)$ plays the role of an impulse response to describe the reverberation. The difference between the representative form of reverberation in this study and those in previous research is that reverberation is described based on the direct observed component, $x(t)$, instead of the signal emitted at the source $s(t)$. The total observed signal consists of the direct component and the reverberation as follows:

$$Y(\omega) = X(\omega) + R(\omega). \tag{A.8}$$

Due to the absorption of the air and the reflecting surfaces, the reflections get weaker

along the time that they are present in the rooms. The reflections with short time delays have strong energy and are comparable with the direct component, while the reflections with long time delays are much weaker and tend to be diffuse. From these characteristics, the reverberation component can be divided into two parts: early reverberation and late reverberation.

$$R(\omega) = R^E(\omega) + R^L(\omega). \tag{A.9}$$

The early reverberation includes the reflections within a time delay threshold t_0, when their energy is still strong. In most of the research, t_0 is selected from 50 ms to 80 ms. The early reverberation can be considered as the signal observed from multiple weaker sources, but it is strongly correlated to the direct component. The late reverberation is diffuse to all directions and uncorrelated with either the direct component and the early reverberation, denoted by $R^L \perp X$, $R^L \perp R^E$. Furthermore, considering the binaural case, late reverberations at both channels are assumed as uncorrelated, $R^L_L \perp R^L_R$.

Besides the reverberation, there may be noise in normal room conditions, denoted by $N(\omega)$. $N(\omega)$ may consist of directional noise (or localized noise) where the sound comes from a specific source, and diffuse noise where the noise source is unclear. Within this thesis, $N(\omega)$ is assumed as uncorrelated with the sound signals generated from the target source, including $X(\omega)$ and $R(\omega)$.

In summary, sound observed in general conditions consists of direct component (target), early and late reverberation, and noise:

$$Y(\omega) = X(\omega) + R^E(\omega) + R^L(\omega) + N(\omega), \tag{A.10}$$

where $X \perp R^L$, $X \perp N$, $R^E \perp R^L$, $R^E \perp N$ and $R^L \perp N$. In addition, $R^L_L \perp R^L_R$ in the scenario of binaural processing.

Appendix B

Approximation using Taylor expansion

For any complex numbers a, b satisfying $|a| << 1$ and $|b| << 1$,

$$(1+a)^b \approx 1 + ab \tag{B.1}$$

Proof:
Let $f(x) = (1+x)^b$ be a complex-valued function. According to Taylor theorem, the value of $f(x)$ can be expressed as follows:

$$f(x)^b = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f^{(3)}(x_0)}{3!}(x - x_0)^3 + ... \tag{B.2}$$

for any complex value x_0. Replace x by a and x_0 by 0, Eq. (B.2) becomes:

$$(1+a)^b = 1 + \frac{b}{1!}a + \frac{b(b-1)}{2!}a^2 + \frac{b(b-1)(b-2)}{3!}a^3 + ... \tag{B.3}$$

Since $|a| << 1$ and $|b| << 1$, Eq. (B.3) can be approximated by:

$$(1+a)^b \approx 1 + ab \tag{B.4}$$

Appendix C

Energy independence of uncorrelated signals

Given two signals $x(t)$ and $y(t)$ in the time domain and their spectra $X(\omega)$ and $Y(\omega)$ in the frequency domain, if $x(t)$ and $y(t)$ are uncorrelated, then

$$\int_{-\infty}^{\infty} |X(\omega) + Y(\omega)|^2 d\omega = \int_{-\infty}^{\infty} \left[|X(\omega)|^2 + |Y(\omega)|^2\right] d\omega \tag{C.1}$$

Proof:

Since $x(t)$ and $y(t)$ are uncorrelated, we have:

$$\int_{0}^{\infty} x(t)y(t)dt = 0 \tag{C.2}$$

This derives a result that they have zero cross-correlation at the time lag $\tau = 0$:

$$R_{xy}(\tau)|_{\tau=0} = \int_{0}^{\infty} x(t)y(t+\tau)dt\Bigg|_{\tau=0} = 0. \tag{C.3}$$

On the other hand, the cross correlation between $x(t)$ and $y(t)$ is related to the cross power spectral density function by the well-known Fourier transform relationship

$$R_{xy}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega)Y^*(\omega)e^{j\omega\tau}d\omega. \tag{C.4}$$

where the superscript * denotes the conjugate operator. By replacing $\tau = 0$ into Eq. (C.4), the following equation is derived:

$$\int_{-\infty}^{\infty} X(\omega)Y^*(\omega)d\omega = 0. \tag{C.5}$$

Similarly,

$$\int_{-\infty}^{\infty} X^*(\omega)Y(\omega)d\omega = 0. \tag{C.6}$$

As a result,

$$\begin{aligned}
\int_{-\infty}^{\infty} |X(\omega) + Y(\omega)|^2 d\omega &= \int_{-\infty}^{\infty} [X(\omega) + Y(\omega)][X(\omega) + Y(\omega)]^* d\omega \\
&= \int_{-\infty}^{\infty} [X(\omega) + Y(\omega)][X^*(\omega) + Y^*(\omega)] d\omega \\
&= \int_{-\infty}^{\infty} [X(\omega)X^*(\omega) + Y(\omega)Y^*(\omega)] d\omega \\
&= \int_{-\infty}^{\infty} \left[|X(\omega)|^2 + |Y(\omega)|^2\right] d\omega.
\end{aligned} \tag{C.7}$$

References

[1] N. Durlach, Foundations of Modern Auditory Theory, ch. Binaural signal detection: equalization and cancellation theory, New York: Academic Press, 1972.

[2] H. Colburn and N. Durlach, Handbook of Perception: Hearing, ch. Models of binaural interaction, pp.467–518, Academic Press, New York, 1978.

[3] H. Colburn and A. Kulkarni, Sound Source Localization, ch. Model of Sound Localization, pp.276–316, Springer, 2005.

[4] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," Speech Communication, vol.53, no.5, pp.677–689, 2011.

[5] Y. Oh, J. Yoon, J. Park, M. Kim, and H. Kim, "A name recognition based call-and-come service for home robots," IEEE Transactions on Consumer Electronics, vol.54, no.2, pp.247–253, 2014.

[6] J. DiBiase, H. Silverman, and M. Brandstein, Microphone Arrays, ch. Robust Localization in Reverberant Rooms, Springer, 2001.

[7] F. Wightman and D. Kistler, Binaural and Spatial Hearing in Real and Virtual Environments, ch. Factors affecting the relative salience of sound localization cues, p.123, Lawrence Erlbaum Associates, 1997.

[8] J. McDonald, "A localization algorithm based on head-related transfer functions," Journal of Acoustical Society of America, vol.123, no.6, pp.290–4296, 2008.

[9] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3d localization based on hrtfs," ICASSP, vol.5, pp.341–344, 2006.

[10] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.1, pp.68–77, 2010.

[11] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," IEEE Transactions on Audio, Speech, and Language, vol.20, no.5, pp.1503–1512, 2012.

[12] N. Durlach, Source Separation, Localization, and Comperhension in Human, Machines, and Human-machine Systems, ch. 15, pp.221–243, Kluwer Academic, 2005.

[13] L. Jeffress, "A place theory of sound localization," Journal of comparative and physiological psychology, vol.41, pp.35–39, 1948.

[14] N. Durlach, "Equalization and cancellation theory of binaural masking level differences," JASA, vol.35, no.8, pp.1206–1218, 1963.

[15] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on In Acoustics, Speech, and Signal Processing, vol.24, no.4, pp.320–327, 1976.

[16] J. Dibiase, A High-Accurate, Low-Latency Technique for Talker Localization in Reverberation Environments Using Microphone Array, Ph.D. thesis, Brown University, Providence RI, USA, 2000.

[17] X. Lv and M. Zhang, "Sound source localization based on robot hearing and vision," Proceedings of the International Conference on Computer Science and Information Technology, pp.942–946, 2008.

[18] D. Li and S. Levinson, "A linear phase unwrapping method for binaural sound source localiation on a robot," Proceedings of the 2002 IEEE International Conference on Robotics and Automation, pp.19–23, 2002.

[19] J. Murray, H. Erwin, and S. Wermter, "Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks," Neural Networks, vol.22, no.2, pp.172–189, 2009.

[20] C. Cherry, "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Amer., vol.25, pp.975–981, 1953.

[21] J. Strutt, "On our perception of sound direction," Philosophical Magazine, vol.13, pp.214–232, 1907.

[22] H. Colburn, Auditory Computation, ch. Computational models of binaural processing, pp.332–400, Springer-Verlag, New York, 1995.

[23] E. Shaw, Binaural and Spatial Hearing in Real and Virtual Environments, ch. Acoustical features of the human external ear, pp.25–47, Lawrence Erlbaum Associates, 1997.

[24] G. Kuhn, Directional Hearing, ch. Physical acoustics and measurements pertaining to directional hearing, Springer-Verlag, 1987.

[25] P. Zurek and R.F.U. Balakrishnan, "Auditory target detection in reverberation," Journal of the Acoustical Society of America, vol.115, no.4, pp.1609–1620, 2004.

[26] R. Wan, N. Durlach, and S. Colburn, "Application of extended equalization-cancellation model to speech intelligibility with spatial distributed maskers," Journal of the Acoustical Society of America, vol.128, no.6, pp.3678–3690, 2010.

[27] Y. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.7, pp.1793 – 1805, 2010.

[28] J. Dibiase, A High-Accurate, Low-Latency Technique for Talker Localization in Reverberation Environments Using Microphone Array, Ph.D. thesis, Brown University, Providence RI, USA, 2000.

[29] A. Mills, "On the minimum audible angle," Journal of the Acoustical Society of America, vol.30, pp.237–246, 1958.

[30] D. Chandler and D. Grantham, "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity," Journal of the Acoustical Society of America, vol.91, no.3, pp.1624–36, 1992.

[31] R. Woodworth, Experimental Psychology, New York: Holt, Rhinehart, and Winston, 1938.

[32] W. Feddersen, T. Sandel, D. Teas, and L. Jeffress, "Localization of high-frequency tones," Journal of Acoustical Society of America, vol.29, pp.988–991, 1957.

[33] G. Kuhn, "Model for the interaural time differences in the azimuthal plane.," Journal of the Acoustical Society of America, vol.62, pp.157–167, 1977.

[34] R. Roth, R. Kochhar, and J. Hind, "Interaural time differences: Implications regarding the neurophysiology of sound localization," Journal of Acoustical Society of America, vol.68, pp.1643–1651, 1980.

[35] A. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," Journal of Acoustical Society of America, vol.83, pp.1508–1516, 1988.

[36] C. Plack, The Sense of Hearing, ch. A Journey through the auditory system, Taylor & Francis Group, 2005.

[37] D. Green and G. Henning, "Audition," Annual review of Psychology, vol.20, pp.105–128, 1969.

[38] N. Kuroda, J. Li, Y. Iwaya, M. Unoki, and M. Akagi, "Effects of spatial cues on detectability of alarm signals in noisy environments," International workshop on the principles and applications of binaural hearing, Miyagi, Japan, 2009.

[39] T. Dolan, "Effects of masker spectrum level on masking-level differences at low signal frequencies," Journal of Acoustical Society of America, vol.44, pp.1507–1512, 2013.

[40] D. McFadden, "Masking-level differences determined with and without interaural disparities in masking intensity," Journal of Acoustical Society of America, vol.44, pp.212–223, 1968.

[41] I. Hirsh, "Binaural effects in remote masking," Journal of Acoustical Society of America, vol.30, pp.827–832, 1958.

[42] F. Webster, "The influence of interaural phase on masked thresholds. 1. the role of interaural time-duration," Journal of Acoustical Society of America, vol.23, pp.452–462, 2013.

[43] D. Green, "Interaural phase effects in masking of signals of different durations," Journal of Acoustical Society of America, vol.39, pp.720–724, 1966.

[44] D. Robinson and L. Jeffress, "Effect of varying the interaural noise correlation on the detectability of tone signals," Journal of Acoustical Society of America, vol.35, pp.147–152, 2013.

[45] A. Koenig, J. Allen, D. Berkley, and T. Curtis, "Determination of masking-level differences in a reverberant environment," Journal of Acoustical Society of America, vol.61, pp.1374–1376, 1977.

[46] I. Hirsh, "The influence of interaural phase on interaural summation and inhibition," Journal of Acoustical Society of America, vol.29, pp.536–544, 1948.

[47] J. Licklider, "The influence of interaural phase relations upon the masking of speech by white noise," Journal of Acoustical Society of America, vol.20, pp.150–159, 1948.

[48] E. Hafter and C. Trahiotis, Handbook of Acoustics, ch. Functions of the binaural system, New York: John Wiley and Sons, 1994.

[49] B. Sayer and E. Cherry, "Mechanism of binaural fusion in the hearing of speech," Journal of Acoustical Society of America, vol.29, pp.973–978, 1957.

[50] L. Jeffress, H. Blodgett, T. Sandel, and C. Wood, "Masking of tonal signals," Journal of Acoustical Society of America, vol.28, pp.416–426, 1956.

[51] E. Hafter, "Quantitative evaluation of a lateralization model of masking-level differences," Journal of Acoustical Society of America, vol.50, pp.1116–1122, 1971.

[52] W. Yost, "Tone-on-tone masking for three listening conditions," Journal of the Acoustical Society of America, vol.52, pp.1234–1237, 1972.

[53] H. Colburn, "Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination," Journal of the Acoustical Society of America, vol.54, pp.1458–1470, 1973.

[54] H. Colburn, "Theory of binaural interaction based on auditory-nerve data. ii. detection of tones in noise," Journal of Acoustical Society of America, vol.61, pp.525–533, 1977.

[55] J. Breebaart, S. van de Par, and A. Kohlrausch, "On the difference between cross-correlation and ec-based binaural models," Proceedings of the Forum Acusticum, Sevilla, Spain, 2002.

[56] M. DeSimio and T. Anderson, "Phoneme recognition with binaural cochlear models and the stereoausis representation," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.521–524, 1993.

[57] T. Sullivan and R. Stern, "Multi-microphone correlation-based processing for robust speech recognition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.91–94, 1993.

[58] P. Zurek, Acoustical Factors affecting Hearing Aid Performance, 2nd ed., ch. Binaural advantages and directional effects in speech, Allyn and Bacon, Boston, 1992.

[59] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," Journal of the Acoustical Society of America, vol.120, no.6, pp.4040–4050, 2006.

[60] D. Chau, J. Li, and M. Akagi, "A DOA estimation algorithm based on Equalization Cancellation Theory," Interspeech, pp.2770–2773, 2010.

[61] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," IEEE Transactions on Audio, Speech, and Language Processing, vol.8, pp.2374 – 2384, 2012.

[62] J. Culling and Q. Summerfield, "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," JASA, vol.98, pp.785–797, 1995.

[63] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in low noise, reverberative environments?," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2565–2568, 2008.

[64] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," IEEE Transactions on Speech and Audio Processing, vol.4, no.2, pp.148–152, 1996.

[65] R.M. Stern, G.J. Brown, and D.L. Wang, Computational Auditory Scene Analysis: Principles, Algorithms and Applications, ch. Binaural sound localization, pp.147–185, New York: Wiley, 2006.

[66] S. Andersson, A. Handzel, V.S. P.S., and Krishnaprasad, "Robot phonotaxis with dynamic sound-source localization.," Proceedings of the 2004 IEEE International Conference on Robotics and Automation, pp.4833–4838, 2004.

[67] E. Berglund, J. Sitte, and G. Wyeth, "Active audition using the parameter-less self-organising map," Autonomous Robots, vol.24, no.4, pp.401–417, 2008.

[68] B. Shinn-Cunningham, "Learning reverberation: considerations for spatial auditory displays," International Conference on Auditory Displays, pp.126–134, 2000.

[69] J. Harris, "A florilegium of experiments on directional hearing," Acta Otolaryngologia, vol.Suppl. 298, pp.1–26, 1972.

[70] D. Campbell, The ROOMSIM User Guide (v3.4), 2004. Available: http://media.paisley.ac.uk/ campbell/Roomsim/. Accessed in December, 2013.

[71] W. Gardner and K. Martin, "Hrtf measurements of a kemar," Journal of the Acoustical Society of America, vol.97, no.6, pp.3907–3908, 1995.

[72] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," Journal of the Acoustical Society of America, vol.65, no.4, pp.943–950, 1979.

[73] A. Kurematsu, K. Takeda, H. Kuwabara, K. Shikanoand, Y. Sagisaka, and S. Katagiri, "ATR japanese speech database as atool of speech recognition and synthesis," Speech Communication, vol.9, no.4, pp.357–363, 1990.

[74] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, revised ed., MIT Press, Canbridge, Massachusetts, USA, 1997.

[75] J. Polack, La transmission de lenergie sonore dans les salles, Ph.D. thesis, Univ. Maine, Le Mans, France, 1988.

[76] D. Brungart and N. Durlach, "Auditory localization of nearby sources ii: Localization of a broadband source in the near field," Journal of Acoustical Society of America, vol.106, pp.1956–1968, 1999.

[77] B. Shinn-Cunningham and K. Kawakyu, "Neural representation of source direction in reverberant space," Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustic. New York: IEEE Press, 2003.

[78] S. Devore, A. Ihlefeld, K. Hancock, B. Shinn-Cunningham, and B. Delgutte, "Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain," Neuron, vol.62, no.1, pp.123–134, 2009.

[79] S. Devore and B. Delgutte, "Effects of reverberation on the directional sensitivity of auditory neurons across the tonotopic axis: influences of interaural time and level differences," Journal of Neuroscience, vol.30, no.23, pp.7826–7837, 2010.

[80] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," IEEE Trans. On Signal Processing, vol.49, no.8, pp.1614–1626, 2001.

[81] A. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," Nature, vol.397, no.6719, pp.517–520, 1999.

[82] H. Kayser, S. Ewert, J. Anemller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," EURASIP Journal on Advances in Signal Processing, no.6, 2009.

[83] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "A speech enhancement approach for binaural hearing aids," in Proc. the 22nd Signal Processing Symposium, pp.263–268, 2007.

[84] . ITU-T P.862, "Perceptual evaluation of speech quality (pesq), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation, 2000. ITU-T Recommendation.

[85] M. Brandstein and D. Ward, Microphone Arrays, Digital Signal Processing, Springer, 2001.

[86] J.F. Cardoso, "Blind signal separation: Statistical principles," Proc. IEEE, vol.90, pp.2009–2026, 1998.

[87] S. Douglas, Blind Signal Separation and Blind Deconvolution, in Handbook of Neural Networks for Signal Processing, CRC Press, 2001.

[88] W. Yost, Fundamentals of Hearing, 3 ed., Academic Press, 1994.

[89] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," IEEE Transaction on Speech and Audio Processing, vol.10, pp.352–362, 2002.

[90] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in Proc. 4th International Symposium on Communications, Control, and Signal Processing (ISCCSP), 2010.

[91] C. Boor, A Practical Guide to Splines, Springer-Verlag, 1978.

[92] B. Gardner and K. Martin, "Hrtf measurements of a kemar dummy head microphone," At http://sound.media.mit.edu/KEMAR.html, Accessed in April, 2010, 2010.

[93] L. Karray and A. Martin, "Toward improving speech detection robustness for speech recognition in adverse environments,," Speech Communication, no.3, pp.261–276, 2003.

[94] J. Ramrez, J. Segura, C. Bentez, A. Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," EUROSPEECH, Geneva, Switzerland, pp.3041–3044, 2003.

[95] A. Sangwan, M.C. andH. Jamadagni, R. Sah, R. Prasad, and V. Gaurav, "VAD techniques for real-time speech transmission on the internet," IEEE International Conference on High-Speed Networks and Multimedia Communications, pp.46–50, 2002.

[96] K. Woo, T. Yang, K. Park, and C. Lee, "bust voice activity detection algorithm for estimating noise spectrum," Electronics Letters, vol.36, no.2, pp.180–181, 2000.

[97] R. Chengalvarayan, "Roust energy normalization using speech/non-speech discriminator for german connected digit recognition," EUROSPEECH, Budapest, Hungary, pp.61–64, 2014.

[98] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Processing, vol.10, no.6, pp.341–351, 2002.

[99] I.T.R.G.A. B, A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70, 2014.

[100] R. Tucker, "Voice activity detection using a periodicity measure," Inst. Elect. Eng, pp.377–380, 1992.

Publications

JOURNAL

[1] D. Chau, J. Li and M. Akagi, "Towards intelligent binaural speech enhancement by meaningful signal extraction." Journal of signal processing, vol. 15, no. 4, pp. 291-294, 2011.

[2] D. Chau, J. Li and M. Akagi, "Binaural sound source localization in noisy reverberant environments based on Equalization-Cancellation Theory." To be published on IEICE Transaction on Fundamentals, Vol. E97-A, No. 10, Oct. 2014.

INTERNATIONAL CONFERENCE

[3] D. Chau, J. Li and M. Akagi, "A DOA estimation algorithm based on Equalization - Cancellation Theory." Interspeech, pp. 2770-2773, 2010.

[4] J. Li, L. Yang, Y. Yan, D. Chau and M. Akagi. "Intelligibility investigation of single-channel noise reduction algorithms for Chinese and Japanese." Chinese Spoken Language Processing–ISCSLP, pp. 7-11, 2010.

[5] D. Chau, J. Li and M. Akagi, "Towards an intelligent binaural speech enhancement system by integrating meaningful signal extraction." International Workshop on Nonlinear Circuits, Communication and Signal Processing (NCSP) pp. 344-347, 2011 (Student paper award).

[6] D. Chau, J. Li and M. Akagi, "Improve Equalization-Cancellation based sound localization in noisy reverberant conditions using direct-to-reverberant energy ratio," Proc. IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), pp. 322–326, Beijing, China, 2013.

DOMESTIC CONFERENCE

[7] D. Chau, J. Li and M. Akagi, "A DOA estimation algorithm based on Equalization-Cancellation Theory." IEICE Technical Report, vol. 110, no. 71, EA2010-28, pp. 37-40, 2010.

[8] D. Chau, J. Li and M. Akagi, "Binaural multi-source localization in noisy reverberant condition based on Equalization-Cancellation model." ASJ meeting, Spring 2013.

[9] D. Chau, J. Li and M. Akagi, "Adaptive equalization-cancellation model and its application to sound localization in noisy reverberant environments." ASJ meeting on Engineer Acoustic (EA), May 2013.

[10] D. Chau, J. Li and M. Akagi, "Binaural sound source localization in noisy reverberant environments based on Equalization-Cancellation Theory." ASJ meeting on Engineer Acoustic (EA), March 2014.