A binaural sound source localization approach based on equalization-cancellation theory and its applications

By Chau Thanh Duc

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

September, 2010

A binaural sound source localization approach based on equalization-cancellation theory and its applications

By Chau Thanh Duc (0810205)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masato Akagi
Associate Professor Isao Tokuda
Associate Professor Kazunori Kotani

August, 2010 (Submitted)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sound source localization overview

Sound source localization (SSL) refers to the ability of human and animals to identify the location of sound source based on sounds received at their ears [35]. In signal processing, sound localization is the problem of determining the position of sound source or estimating the direction of arrival (DOA) of sound by analyzing signals recorded by microphones or sensors. The main approach in this field relies upon the auditory system of human and animals to obtain location cues of signals. Psychoacoustic research have discovered a number of cues for localization, including Interaural Time Difference (ITD) and Interaural Level Difference (ILD), also known as Interaural Intensity Differnce (IID), between both ears, spectral information, timing analysis, correlation analysis, and pattern matching.

SSL plays important role in many systems where location information of sound source is required as an indispensable factor. In humanoid robot, SSL is the key component to help the robot interact naturally with human. The information of sound sources is also used to specify the distribution of sound in 3D space which is the basic concept in sound visualization. In addition to these systems, SSL acts as an initial task in several other applications, such as speech enhancement and blind source separation. For multi-channel based speech enhancement, some algorithms assume DOA of target as a prior known information to perform beamforming enhancement [29, 30]. For blind source separation, in the case that the number of microphones is less than the number of sources, blind beamforming is considered as an alternative way to separate sounds replying on the detection of DOA. So far, a lot of algorithms, which were shown as effective in localization, have been introduced. A good review on this field can be found in [6].

1.2 Previous approaches on sound source localization

Existing SSL algorithms can be loosely classified into three general approaches: those based upon steered response power (SRP) of a beamformer, techniques adopting high-resolution spectral estimation concepts, and methods employing time-difference of arrival (TDOA) information [6]. The first approach involves estimation techniques rely on filtered, weighted, summed of signals received at microphones. The second approach refers to algorithms localize sound source through analysis of signal correlation matrix. And the last one includes all methods estimating sound location based on time delay between microphones.

1.2.1 Beamforming-based approach

The simplest and most common type of beamforming approach is delay-and-sum beamformer. This method applies time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. The signals are time-aligned and summed together to form a beamforming output. The location of sound source is determined as the position at which beamforming output is maximum. Advance beamformer methods apply filters to signals received from microphone array for time-alignment. The differences in filters derive different methods in these filter-and-sum beamformer.

The advantage of beamforming-based approach is the ability to be extended to the case of multiple signal sources [37]. Conventional procedures in this approach usually face with the problem of high computational expense in searching. The stochastic region contraction (SRC) technique was also applied to deal with this problem in talker localization [2]. However, this approach, in general, still not enough effectiveness for high accurate in real-time systems [6].

1.2.2 High-resolution spectral estimation approach

This approach includes methods adapted from the field of high-resolution spectral analysis. Typical techniques are autoregressive (AR) modeling, minimum variance (MV) spectral estimation, and eigen-analysis of which the multiple signal classification (MUSIC) is the most popular algorithm in this approach [23]. Although these methods have successfully applied in a variety of array processing applications, they all have their own restrictions for speech-source localization.

The high-resolution process is based upon the spatiospectral correlation matrix derived from signals received at the microphones. When exact knowledge of this matrix is unknown, it must be estimated from the observed data. To do this work, it is required assumptions that the sources of noise is statistically stationary and the location of sound source must be fixed. Another limitation is that these methods were developed in context of far-field plane waves projecting onto a linear array and few of them can be extendible to the case of general array geometries and near-field sources. Moreover, this approach has high computational complexity and tend to be less robust to source and sensor modeling errors than conventional beamforming methods [11].

1.2.3 TDOA-based approach

This approach is based on a two-step procedure. The first step is to estimate the time delay between two microphones in each pair. This time delay is along with the knowledge of microphone positions. Then, the second step employs this knowledge to generate hyperbolic curves, which intersect in the position at the source location.

There are some differences in derivation of these methods, for example, 2D vs. 3D, near source vs. distant source, etc. The most effectiveness of TDOA-based methods is the accurate and robust Time Delay Estimation (TDE) using cross-correlation. For two signals received at two microphones $x_1(t)$ and $x_2(t)$, the cross-correlation with the time delay τ is defined as

$$c_{12}(\tau) = \int_{-\infty}^{+\infty} x_1(t)x_2(t+\tau)dt. \tag{1.1}$$

Since the cross-correlation is sensitive and suffers from noise and reverberation, there have been many attempts to study for weighting function. The most basic method, which considered this function, is Generalized Cross-Correlation (GCC) [26]. However, once the room reverberations rise, this method begins to exhibit dramatic performance degradations and become unreliable. Hence, some methods have been study to make the GCC function more robust by deemphasizing the frequency-dependent weightings. Among them, the GCC with Phase Transformation (GCC-PHAT) is a method received most attention and is considered as the basis of speech source localization systems. The function of GCC-PHAT method is described as follow

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Psi(\omega)X_1(\omega)X_2(\omega)^* e^{j\omega\tau}d\omega \tag{1.2}$$

in which $X_1(\omega)$ and $X_2(\omega)$ are Fourier transforms of $x_1(t)$ and $x_2(t)$ respectively, * denotes the conjugate operator and $\Psi(\omega)$ is the weighting function

$$\Psi(\omega) = \frac{1}{|X_1(\omega)X_2(\omega)^*|} \tag{1.3}$$

The GCC-PHAT is then expanded for the case of multiple pairs of microphones by integrating the beamforming strategy leads to an effective sound source localization, namely SRP-PHAT [13]. Although these algorithms achieved considerable results on microphone array, they usually face with the problem of effect caused by recording system. In humanoid robot, if the recording system is the two sensors placed inside the pinnae of robot head, the recored sounds is affected by the head-related transfer function (HRTF). When such effects occur, the performance of these algorithms will be degraded significantly.

1.3 The challenges of binaural sound localization

Binaural sound localization, sometime known as binaural DOA estimation, has been extensively used in multi-channel signal processing systems, such as multi-channel signal separation, speech enhancement, binaural hearing aids and humanoid robot audition [8]. In multi-channel noise reduction, for example, the DOA information (steer vector) of the desired signal is normally needed to compensate for the differences among different microphones [6]. In robot audition, the sound source localization helps the robot to face and possibly come closer to the speakers that it is talking to [33]. The DOA information of sound sources in binaural hearing aid provide the users with the spatial cues of sound sources, improving the speech intelligibility in noise conditions and the perceptual impression of acoustic scene [4]. Such systems require a binaural DOA estimation method using two microphones, being robust under noisy conditions, and adapting to system effects (e.g. HRTF).

Although a huge number of studies have introduced a variety of sound localization algorithms with well performance, each algorithm has its own limitations when implemented in binaural applications. Among three approaches were described in section 1.2, the TDOA based methods have received extensive investigation due to its high performance and low computational expense. Several of them have been considered as state-of-the-art in this field, such as, GCC-PHAT and SRP-PHAT. However, these methods just deal with low noise environments and are not effective under HRTF effects. Techniques in high-resolution spectral estimation approach are limited to the far-field, statistically stationary source and noise, and especially, less robust to source and sensor modeling errors [11]. Even if the beamforming-based approach is potentially robust under noisy conditions and can deal with multiple sources, they also suffer from the requirement of large array of microphones and high complexity for real-time systems.

Regarding to the problem of HRTF, recently, F. Keyrouz *et al.* proposed Inverse-HRTF method, which was shown as effective in dealing with HRTF effect [25]. Although the Inverse-HRTF was reported with relatively highly-accurate estimation using two microphones, it is a kind of HRTF-dependent algorithm and limited specifically to artificial dummy head. Such algorithms are difficult to apply widely in SSL systems because the effect on sound completely depends on the systems' shape and the positions where microphones are placed. For example, if the shape of robot's head is cube the effect will not be HRTF anymore and constructing inversed filters becomes a big problem to concern.

In summary, a number of sound localization algorithms have been proposed. In some experimental conditions where the requirements are met, these algorithms are able to work effectively. However, considering binaural applications, localization methods are facing with following problems:

- *The requirement of using only two microphones:* Since computer cannot simulate human perception mechanism perfectly, large microphone arrays are employed to obtain more localization information of sound for improving accuracy. However, this will be the problem when applying to binaural-channel applications such as hearing aids in which there is only two microphones can be used.

- *The effect caused by system shape:* Sound localization is sometime implemented on special systems which may affect the recorded sound by their shape. For example, the robot head with a spherical shape causes the head-related-transfer function (HRTF).

- *The degradation factors:* for example, background noise and reverberation.

1.4 Equalization-Cancellation: A promising model for binaural sound localization

Human auditory system specify the location or direction of sound based on the differences between sounds at two ears. These differences are known as binaural cues of signals. In principle, binaural approach simulates human auditory mechanism to extract these cues for localization. Cancellation is a common strategy for exploiting binaural cues. In [30], Li *et al.* showed that the Cancellation process will be more effective if the Equalization is performed beforehand. These techniques are derived from a theory called Equalization-Cancellation.

Equalization-Cancellation (EC) model was originally developed by Durlach [17] and further improved by Culling and Summerfield [12]. In the original EC model, when subject is presented with a binaural-masking stimulus, the auditory system attempts to eliminate the masking components by transforming the signal arriving at one ear relative to the signal at the other ear to make the masker components equalized (the E process). Then part of the signal in each ear is canceled by subtracting the signal in the other ear (the C process) [17]. In theory, the cancellation process yields binaural cues of sound, such as ITD and IID, which are very important for binaural signal analysis, including sound localization. This model was then improved in [12] where the E and C processes were independently performed for the interfering signals in each channel.

In binaural signal processing, EC theory has been exploited to explain many psychoacoustic phenomena. Typically, the EC model is exploited to account for binaural masking level difference (BMLD) [12, 17], in which considerable benefits in understanding a signal in noise can be obtained when the DOA information of the signal is not the same as that of the maskers. In speech enhancement, the EC model is applied to estimate interference signals based on the dissimilarities between signals at two ears [29, 30]. So far, EC model has been used in signal detection [18], distance estimation [32]. Moreover, as it can extract binaural cues of sound, including ITD and IID, this model has a great potentiality to be applied in binaural sound source localization.

1.5 Thesis goal

Motivated by EC-theory, the first goal of this thesis is to develop an accurate and efficient binaural SSL method for DOA estimation by integrating EC model into beamforming strategy. Concerning the three challenges of binaural sound localization in section 1.3, the proposed method is expected to be able to localize accurately DOA of sound with two microphones, be robust under noisy environments, and has adaptability to overcome the problem of effect caused by system shape, e.g., HRTF.

The effectiveness of EC-BEAM is also expected to be verified in terms of binaural applications. The second goal of this thesis is to design binaural signal processing systems to show the advantages of EC-BEAM. Considering binaural applications, the binaural speech enhancement and blind sound separation are strongly related to DOA estimation. Correspondingly, this goal includes two sub-goals:

- *Proposing an intelligent speech enhancement based on EC-BEAM.* Speech enhancement is an important field in signal processing. In principle, speech enhancement systems, such as hearing aids, attempt to suppress the degradation factors of sound (e.g.,noise and reverberation) and preserve target signal. However, in daily communication, there are many important non-target signals need to be perceived by listener, for example, sound of telephone, sound of a call from someone. This issue has not been considered in the state-of-the-art of this field. In an effort to verify the applicability of EC-BEAM, this sub-goal is to construct an intelligent speech enhancement system which is able to extract meaningful signal as well as enhance target. Specifically, beside enhancing the target signal, this system detect and extract meaningful signal to present in final output and the effectiveness of EC-BEAM will be verified as the ability of this system to detect meaningful signal.

- *Developing a directional blind source separation based on EC-BEAM.* Blind source separation concerns the problem of recovery original sounds in a mixture of signals. There are many methods which can solve this problem successfully when the number of channels (microphones) is greater than the number of sources. Contrary, in case the number of microphone is less than the number of sources, this problem must be solved alternatively by searching the sources through a beamforming process, known as directional source separation. Since the performance of this method highly depends on effectiveness of sound source detection, the EC-BEAM is applied to develop an directional source separation as a second sub-goal.

Chapter 2

The proposed EC-BEAM algorithm for DOA estimation

Inspired by EC theory, we propose a two-microphone DOA estimation algorithm, namely EC-BEAM, by taking advantage of the steered beamformer based technique. This chapter firstly gives the basic concepts of EC-BEAM, then, describes the procedure of the proposed algorithm in detail, and finally evaluates its performance through experimental results.

2.1 Concepts of EC-BEAM

Essentially, the proposed EC-BEAM algorithm is a kind of filter-and-sum beamformer-based methods in which the EC model is applied to construct beamformer filters. These filters are obtained beforehand by Equalization process and preserved to be used in Cancellation process. The main difference between the proposed DOA estimation algorithm and the traditional beamformer-based DOA estimation approaches is that: in the traditional beamformer-based DOA estimation approaches, the main lobe of the acoustic beam points to the strongest sound source and the DOA estimate is decided by searching the direction with maximal power of beamformer output. By exploiting the EC model, in contrast, the proposed EC-BEAM algorithm generates and steers the main null beam towards the strongest source in which the signal from this direction will be removed via Cancellation process, which is much preferred for speech enhancement in multiple-noise-source conditions [30]. As a result, the output of beamformer contains only signals from other directions.

Following the strategy of the steered beamformer based DOA estimation, in principle, the EC-BEAM algorithm scans the space of interest with the acoustic beam at the pre-defined directions, and determines the DOA estimate based on the energy of the beamformer output. More specifically, the estimation of DOA is given by finding out the direction at which the power of null beamformer output reaches to minimum. Furthermore, the interpolation technique is exploited to reduce the computational cost, and increase the spatial resolution of DOA estimate even if only several pre-defined directions were scanned by the acoustic null beamformers.

Concerning the three challenges were pointed out in section 1.3, the EC-BEAM firstly uses only two microphones to estimate DOA of sound source. Secondly, the filters used in EC-BEAM are constructed beforehand in an Equalization process using signals recorded by the same system, therefore, theoretically, the EC-BEAM should be able to "learn" the effect caused by system's shape. Note that this effect is not limited to only HRTF. This ensures the EC-BEAM is adaptable to various systems. Finally, Since the EC model accounts for the BMLD, which is the phenomenon that human has ability to perceive signal in noise, the EC-BEAM is potentially robust under noisy environments.

2.2 EC-BEAM algorithm

The proposed EC-EAM DOA estimation algorithm is performed through three main stages: the rough DOA estimation by EC-model based beamformer technique; the interpolation for the directions between the rough DOA estimates; the final DOA estimation by seeking the direction with the minimal energy among all the interpolated directions and roughly estimated directions.

2.2.1 Beamforming with EC-model

The amount of acoustic beams used for scanning the space of interest at the pre-defined directions is directly dependent on the spatial resolution of DOA estimation. Obviously, the more the acoustic beams are, the higher spatial resolution is at the higher computational cost; and vice verse. For each predefined direction, an acoustic null beamformer that steers the null beam towards that direction is designed based on the EC model, which finally outputs the signals from the directions other than the pre-defined direction. Note that when a null beamformer is steered to the sound source with the largest energy, the energy of this null-beamformer output gets to minimal. Following sub-sections describe how to perform beamforming with EC model.

Signal model

In binaural scenarios, the characteristics of sound coming from a direction, for example θ, are involved to the differences in amplitude and phase of signals at the left and right ears. The observed signals, $X_L(k,\ell)$ and $X_R(k,\ell)$, in the kth frequency bin and the ℓth frame at the left and right ears, includes signal comes from θ and signals come from other directions (Fig. 2.1). Mathematically, these signals can be expressed as

$$X_i(k,\ell,\theta) \quad = \quad S_i(k,\ell,\theta) + N_i(k,\ell,\theta), \quad i = L, R, \tag{2.1}$$

where $S_i(k,\ell,\theta) = H_i(k,\ell,\theta)S(k,\ell,\theta)$ and $N_i(k,\ell,\theta)$ are respectively the spectra of the target and interfering signals; $H_i(k,\theta)$ represents the transfer functions between the target sound source to two ears, referred to as *head-related transfer function* (HRTF) in the context of binaural hearing. Note that the interfering signals, $N_i(k,\ell,\theta)$, might be a combination of multiple interfering signals and background noise.

The equalization and cancellation processes are implemented as the same way in [29], however, instead of estimating interference signals, these processes are applied to remove signal at a given direction, yielding remaining signals from other directions. More specifically, the E-C processes are performed as follows:

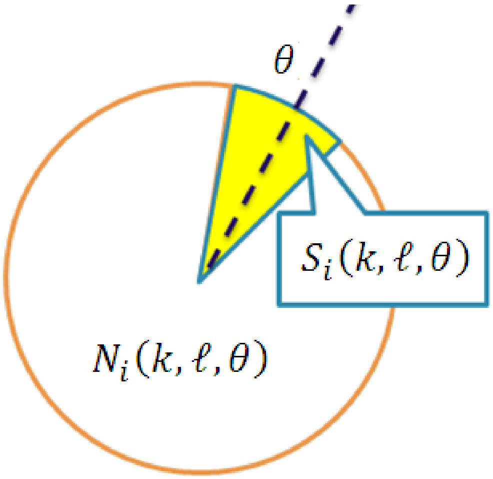


Figure 2.1: Signal model

Equalization process

This process aims to construct two equalizers which make the signal components in the given pre-defined direction at the left input and those at the right input to be equalized. After compensation for the differences in intensity and phase of the signal component at two ears, the received two equalizers and should satisfy

$$S_L(k, \ell, \theta) - W_R(k, \ell, \theta) S_R(k, \ell, \theta) = 0,$$
$$S_R(k, \ell, \theta) - W_L(k, \ell, \theta) S_L(k, \ell, \theta) = 0.$$

Specifically, these equalizers are obtained using the *normalized least mean square* (NLMS) algorithm, which is given as (θ is omitted for simplicity)

$$\mathbf{W}_L(\ell+1) = \mathbf{W}_L(\ell) + \mu \frac{\mathbf{X}_L(\ell)}{||\mathbf{X}_L(\ell)||^2} \left[\mathbf{X}_R(\ell) - \mathbf{W}_L^T(\ell)\mathbf{X}_L(\ell)\right], \tag{2.2}$$
$$\mathbf{W}_R(\ell+1) = \mathbf{W}_R(\ell) + \mu \frac{\mathbf{X}_R(\ell)}{||\mathbf{X}_R(\ell)||^2} \left[\mathbf{X}_L(\ell) - \mathbf{W}_R^T(\ell)\mathbf{X}_R(\ell)\right], \tag{2.3}$$

where $\mathbf{W}_i(\ell) = [W_i(1, \ell), W_i(2, \ell), \ldots, W_i(K, \ell)]^T$, $\mathbf{X}_i(\ell) = [X_i(1, \ell), X_i(2, \ell), \ldots, X_i(K, \ell)]^T$ ($i = L, R$), K is the STFT length, and the superscript ^{T} denotes the transposition operator; μ is the step size.

The Equalization is performed beforehand with clean signal for each direction in training at the off-line mode. The trained equalizers will be further used in the following cancellation processing at the online testing mode.

Cancellation process

The coefficients of two equalizers are fixed and applied to the observed mixture signals in the presence of interfering signals. Since the equalizers are calibrated in the scenarios without interfering signals, the signal components with the given direction at the null beamformer output should be approximately, if not exactly, equivalent to the signal components of the right (left) channel. As the result, the beamformer output signals are derived as

$$\begin{aligned} Z_L(k,\ell,\theta) &= X_L(k,\ell,\theta) - W_R(k,\ell,\theta)X_R(k,\ell,\theta) \\ &\approx N_L(k,\ell,\theta) - W_R(k,\ell,\theta)N_R(k,\ell,\theta), \quad (2.4) \\ Z_R(k,\ell,\theta) &= X_R(k,\ell,\theta) - W_L(k,\ell,\theta)X_L(k,\ell,\theta) \\ &\approx N_R(k,\ell,\theta) - W_L(k,\ell,\theta)N_L(k,\ell,\theta). \quad (2.5) \end{aligned}$$

From Eqs. (2.4) and (2.5), we observe that the signal from the given direction has been cancelled, yielding the estimate of the signals from all other directions. It is clear that the energy of this null beamformer output will be significantly reduced if the given direction is the same as that of highest energy source to be estimated. Consequently, the power of the beamformer output, which is actually the power of signals coming from directions other than θ, is given by

$$P(\theta) = \sum_{k,\ell} \Big[Z_L(k,\ell,\theta)^2 + Z_R(k,\ell,\theta)^2 \Big]. \tag{2.6}$$

2.2.2 Interpolation for non-beamformed directions

Since scanning all possible directions in the space of interest is really time-exhausting, the steered beamformer-based approaches normally focus the acoustic beams towards several per-defined directions, followed by the interpolation process to improve the spatial resolution of DOA estimation. Given the beamformer outputs in the neighboring directions, the interpolation is performed by yielding the output energies in the directions between the neighboring directions at the desired spatial resolution. The purpose of interpolation is to obtain the beamformer output at the directions other than the pre-defined ones, finally increasing the spatial resolution of DOA estimation. In this research, the cubic spline interpolation is exploited [5]. The interpolation process finally yields the energy of the outputs for the directions at the desired spatial resolution, which will be further utilized in the final DOA estimation stage.

2.2.3 DOA estimation

Since the problem of DOA estimation is considered as finding a direction with the strongest intensity in the space of interest, the DOA estimate is determined as the direction at which the power of the null beamformer output at the steered directions or the interpolated directions is minimal. Suppose the power of the remaining signal except for the energy at the direction θ be $P(\theta)$, the final DOA estimate should be

$$\hat{\theta} = \arg\min_{\theta} P(\theta) \tag{2.7}$$

2.3 Experiments and Results

The proposed EC-BEAM algorithm was examined under various conditions and compared with the well-known GCC-PHAT algorithm. To evaluate the performance of EC-BEAM and its robustness under noisy environments as well, a number of experiments have been carried out with clean and noisy conditions. The adaptability of the proposed algorithm is also verified by experimental results on signals under effects of in-ear recorded HRTFs and behind-the-ear recorded HRTFs.

2.3.1 Experiment on data under in-ear HRTF

The purpose of this section is to test EC-BEAM with signals affected by HRTF recorded from microphones placed in ear of artificial dummy head, and test its robustness under noisy conditions. In these experiments, the KEMAR HRTF database measured at 44.1 KHz of MIT [20] was applied to synthesize speech signals. Regarding DOA estimation, we just used the HRTF measurement in horizontal plane (0^o elevation) with 5^o-intervals in an azimuth range from -90^o to 90^o. For speech, 110 Japanese utterances by 11 speakers, in which each speaker has 10 utterances, were selected from ATR database [28]. For each sample, by convoluting with the HRIR, we created 37 signals for 37 directions from -90^o to 90^o. In total, 4070 signals were created, of which 370 signals were used for training to obtain 37 pairs of equalizers (left and right) corresponding to those directions. The 3700 remaining signals then were used to produce testing data. The acoustic environments tested in this experiment include clean condition and noisy conditions with one-source-noise, two-source-noise and real noise recorded at cafeteria.

Clean condition

In this condition, no interfering signal is present. To confirm whether the EC-BEAM can well estimate the DOA in cases the directions of observed signals have not been trained, we just used the trained equalizers at 10^o-intervals from [-90^o, -80^o,..., 90^o]. Suppose the desired spatial resolution of DOA estimation be 5^o, the interpolation process was further implemented to get 1^o-interval beamforming resolution. In testing, EC-BEAM was used to estimate DOAs of signals from 37 directions, each direction has 100 signals. In total, 3700 estimates were performed.

DOA estimation results are plotted in Fig. 2.2 in which the value at each direction is the *average estimation error* (AEE) over 100 estimates. It can be observed that although the equalizers at azimuths $-85^o, -75^o, ..., 85^o$ had not been applied, these directions were also correctly estimated. Overall, AEEs are bellow 4.5^o and the average error for all estimates at all directions is 1.29^o as shown in Table 2.1. Consider that the estimation is correct if the difference between estimated DOA and real DOA does not exceed 5°, the accuracy (in Table 2.1) is relatively high, 98.21%. Furthermore, the Standard Deviation (Std.) of AEEs for all direction is only 1.29, this means the error does not change so much among these directions. More detail results of estimation at each direction can be referenced in Table 2.2.

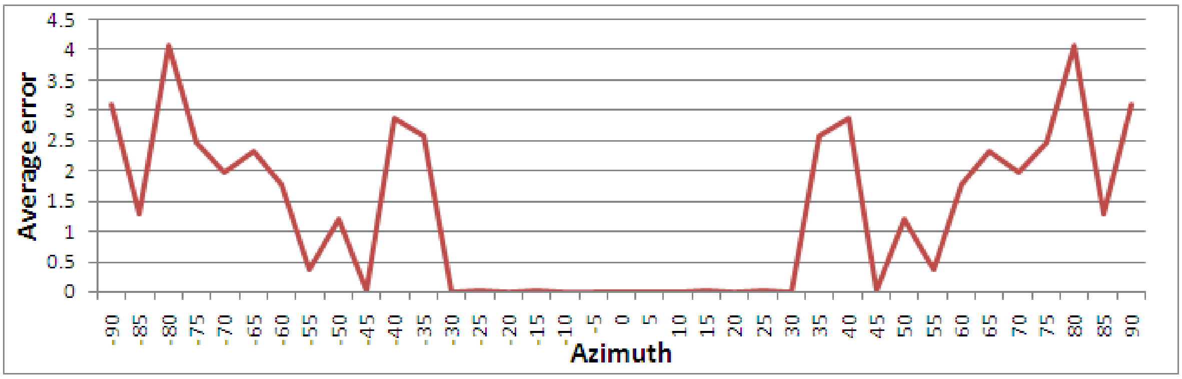


Figure 2.2: Average estimation errors of EC-BEAM with clean signals

Table 2.1: Summarized result on clean signals under in-ear HRTF effect

Average Error	Average Standard Deviation	Accuracy
1.29	1.29	98.21 %

Noisy conditions

In this condition, the EC-BEAM was evaluated with simulated noise and real recorded noise. For simulated noise, clean speeches were mixed together to obtain noisy data in which one signal was considered as target and the others are noise. For one-source-noise signal, the direction of noise-source was fixed at 60°, while the direction of the target varied from -90° to 90° (5°-intervals). When mixing these signals, the amplitude of noise was controlled to make the Signal-to-Noise Ratios (SNR) of 5dB, 10dB and 15dB. At each SNR level, a total 3700 signals were created (100 speech data×37 directions). For two-source-noise signal, mixing method was also performed in the same way of one-source-noise signal, but the directions of noise-sources were fixed at -30° and 60°.

The summary of results are shown in Table 2.3 and Table 2.4. These tables show that in the accuracy in all cases remains high while the average errors do not increase so much. For more detail, the DOA estimation results in terms of estimation error under the one- and two-noise-source conditions are plotted in Fig. 2.3. Fig. 2.4(a) shows that the DOA estimation errors increase as the increase of noise level in both noise conditions. It is important to note that the increase in DOA estimation error is very small in the conditions with SNRs larger than 5 dB. In the low noise conditions, the introduced DOA estimation

Table 2.2: Detail results on clean signals under in-ear HRTF effect

Azimuth (deg)	**AEE**	**Std.**	**Acc (%)**		**Azimuth (deg)**	**AEE**	**Std.**	**Acc (%)**
-90	3.09	3.44	100		*5*	0.02	0.10	100
-85	1.35	1.96	100		*10*	0.01	0.10	100
-80	4.19	4.43	93		*15*	0.02	0.14	100
-75	2.69	3.84	93		*20*	0.01	0.10	100
-70	2.01	6.02	96		*25*	0.03	0.14	100
-65	2.6	6.28	97		*30*	0.02	0.10	100
-60	1.81	3.47	98		*35*	2.57	11.27	95
-55	0.35	0.92	99		*40*	1.99	11.27	94
-50	1.23	1.48	100		*45*	0.04	0.14	100
-45	0.03	0.14	100		*50*	1.25	1.48	100
-40	2.39	11.27	94		*55*	0.44	0.92	99
-35	2.59	11.27	95		*60*	1.75	3.47	98
-30	0.02	0.10	100		*65*	2.4	6.28	97
-25	0.03	0.14	100		*70*	1.49	6.02	96
-20	0.01	0.10	100		*75*	2.48	3.84	93
-15	0.01	0.14	100		*80*	4.15	4.43	93
-10	0.01	0.10	100		*85*	1.33	1.96	100
-5	0.01	0.10	100		*90*	3.1	3.44	100
0	0	0.00	100					

errors are quite comparable to those in the clean conditions. These observations are quite similar in both one- and two-noise-source conditions.

Concerning the comparison between the DOA estimation performance of the proposed EC-BEAM in the two simulated noise conditions, no significant difference is observed. While, the estimation errors in two-noise-source condition are slightly smaller than those in the one-noise-source condition. The possible reason is that the individual sound source in the two-noise-source condition is much weaker than the source in the one-noise-source condition, even in the same SNR condition. The detail estimation result at each direction of these conditions can be observed at section A.1 and A.2 in Appendix A

Table 2.3: Summarized result on one-source-noise signals under in-ear HRTF effect

SNR	Average Error	Average Standard Deviation	Accuracy
5 dB	4.02	3.47	89.95 %
10 dB	1.55	1.48	97.27 %
15 dB	1.33	1.32	98.08 %

For real recorded noise, the EC-BEAM was experimented noise recorded in the cafeteria of Japan Advanced Institute of Science and Technology (JAIST). In mixing process, the noise amplitude was also controlled to get SNR of 5dB, 10 dB and 15 dB. The summarized result is shown in Table 2.5 and the details are illustrated in Fig. 2.5. In general, there

Table 2.4: Summarized result on two-source-noise signals under in-ear HRTF effect

SNR	Average Error	Average Standard Deviation	Accuracy
5 dB	2.84	2.94	93.19 %
10 dB	1.45	1.35	97.70 %
15 dB	1.31	1.31	98.14 %

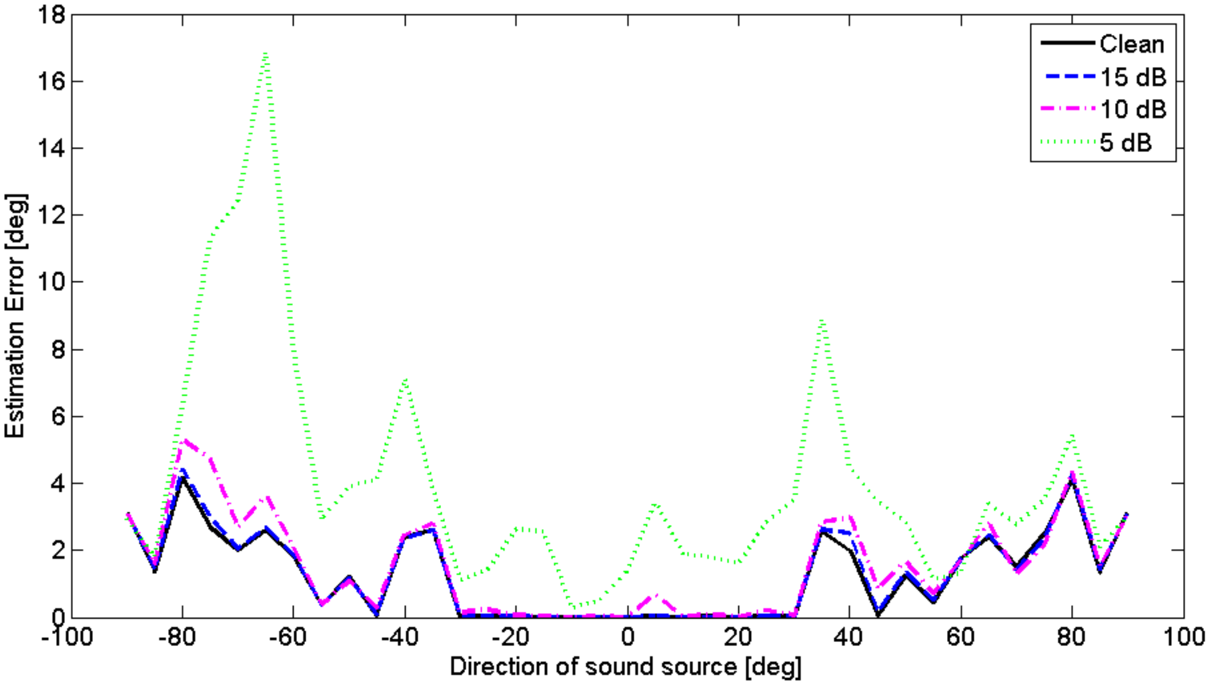


Figure 2.3: Average estimation errors of EC-BEAM with in-ear one-source noisy signals

is almost no difference between the results in this case and those of two-source noise condition. That means the performance of EC-BEAM is still stable even in real noise conditions.

Further observation in the noisy conditions, it is shown that the results on cafeteria-noise signals (Table 2.5) are quite similar to those on two-source-noise signals 2.4 and higher the results on one-source-noise signals (Table 2.3). This indicates that at the same SNR, EC-BEAM has higher accuracy in the condition where noise is diffused.

Table 2.5: Summarized results on cafeteria noisy signals under in-ear HRTF effect

SNR	Average Error	Average Standard Deviation	Accuracy
5 dB	3.05	3.41	92.11 %
10 dB	1.43	1.46	97.43 %
15 dB	1.31	1.31	98.11 %

In summary on both clean and noisy conditions, observation illustrates that when the strongest source lies in around the front (e.g., about $-30^o \sim 30^o$), the estimation errors

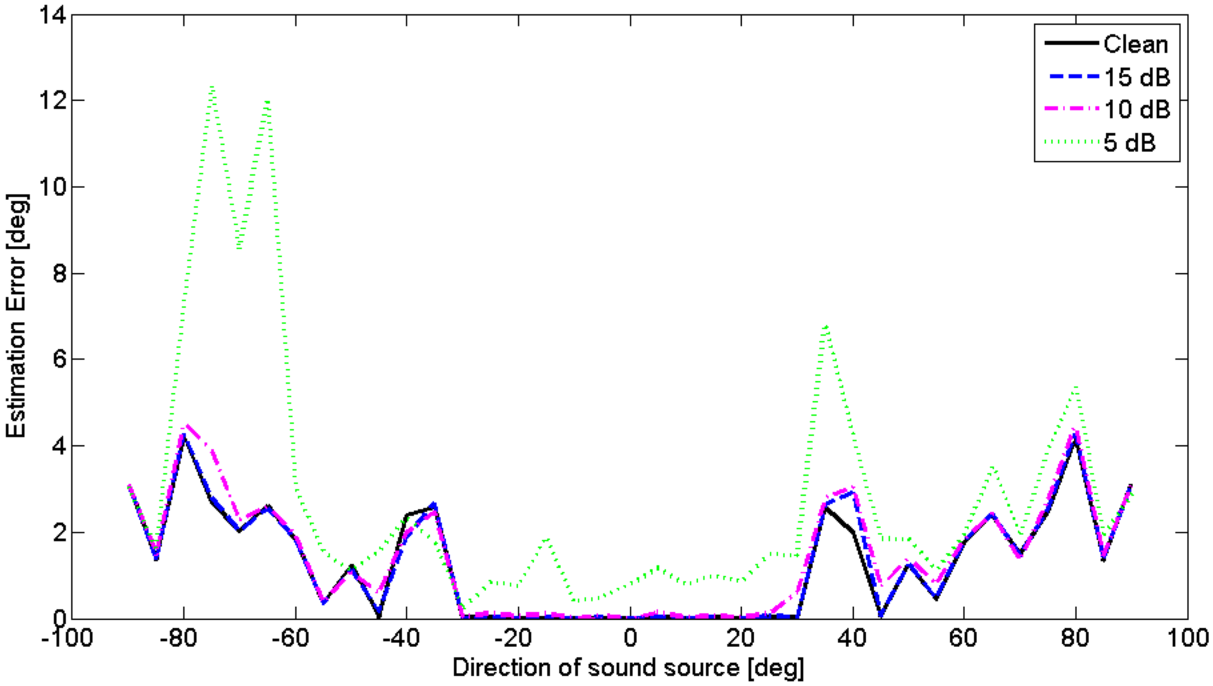


Figure 2.4: Average estimation errors of EC-BEAM with in-ear two-source noisy signals

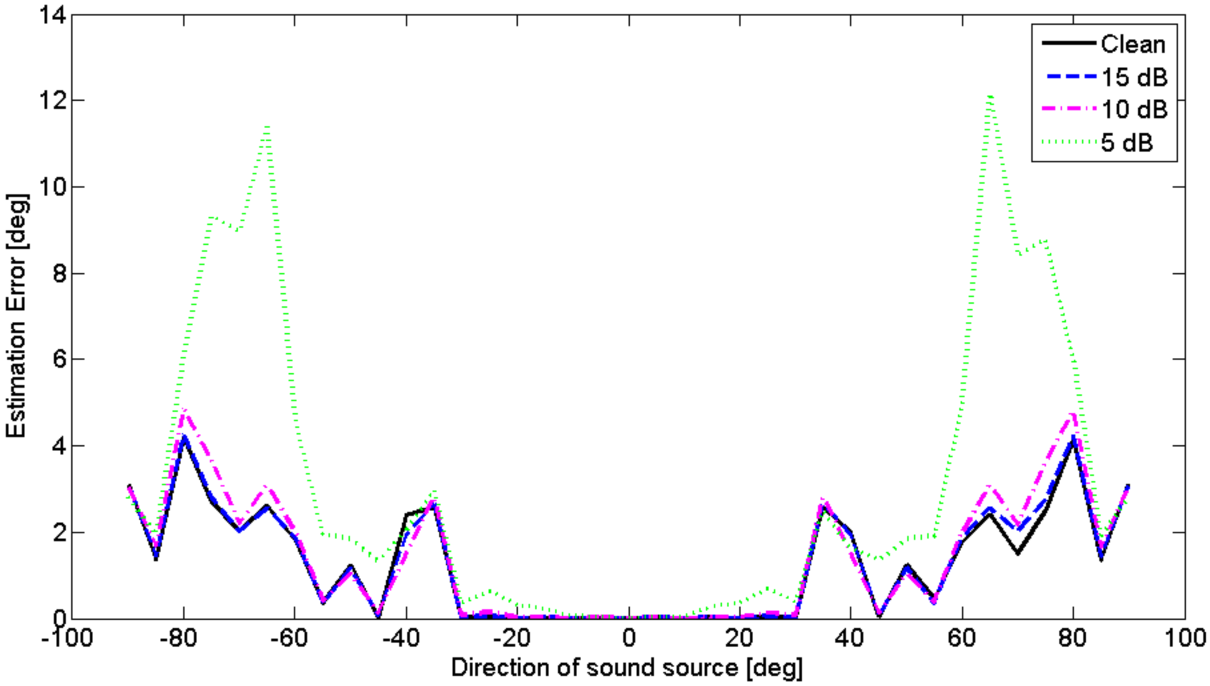


Figure 2.5: Average estimation errors of EC-BEAM with in-ear cafeteria noisy signals

of the proposed EC-BEAM algorithm are significantly small in the low noise conditions, and the errors are relatively larger in the lateral areas (e.g., close to -90^o or 90^o) under all tested conditions including the clean condition. This result is also consistent with the localization ability of human in practical environments [3].

2.3.2 Experiment on data under behind-the-ear HRTF

This experiment aims to evaluate the adaptability of EC-BEAM with different systems. The database used in this experiment is HRIR database from University of Oldenburg [24], in which recording system had 8 microphones, with 2 inside-ear microphones and 6 behind-the-ear mikes. In order to test EC-BEAM with signals under effects other than in-ear HRTF (like KEMAR Database), we used the recorded signals from the first 2 of 6 behind-the-ear microphones in Anechoic set. The dataset was created in the same way as in *Experiment on data under in-ear HRTF*. We also used 370 signals (of one speaker) for training, and the 3700 remaining signals for testing. Since the robustness of the proposed algorithm under noisy conditions was confirmed by experiments in 2.3.1, this experiment was carried out with only clean data.

The result shown in Table 2.7 indicates that although the accuracy decreased a little, the average error and its standard deviation remained low. Moreover, as shown in Fig. 2.6, the average errors of all directions are almost less than 4. Further observation on Fig. 2.7, we can see that the error in estimation under effect of behind-the-ear HRTF is quite stable compared to the strong fluctuation with that of under in-ear HRTF. Even though the average error in this case higher than the case of in-ear HRTF, its performance is quite comparable.

Table 2.6: Summarized result on clean signals under behind-the-ear HRTF effect

Average Error	Average Standard Deviation	Accuracy
1.66	1.19	92.27 %

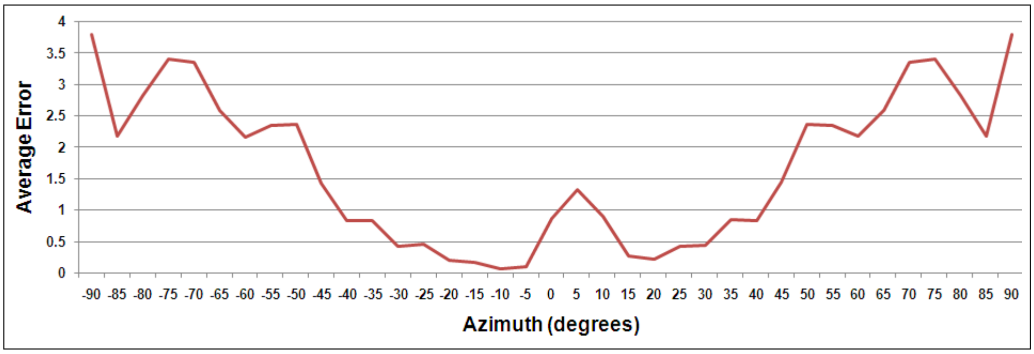


Figure 2.6: Average estimation errors of EC-BEAM with clean signals under behind-the-ear HRTF effect

Table 2.7: Detail result on clean signals under behind-the-ear HRTF effect

Azimuth (deg)	**AEE**	**Std.**	**Acc (%)**		**Azimuth (deg)**	**AE**	**Std.**	**Acc (%)**
-90	3.8	5.09	72		*5*	1.32	1.43	100
-85	2.17	3.21	90		*10*	0.89	1.02	100
-80	2.83	3.13	100		*15*	0.26	0.69	100
-75	3.4	4.58	72		*20*	0.21	0.66	100
-70	3.36	4.75	87		*25*	0.42	1.14	99
-65	2.59	5.72	90		*30*	0.44	1.31	99
-60	2.16	5.91	94		*35*	0.84	2.04	96
-55	2.35	6.26	90		*40*	0.82	2.66	96
-50	2.36	7.19	90		*45*	1.44	5.03	96
-45	1.43	5.03	96		*50*	2.36	7.19	90
-40	0.82	2.66	96		*55*	2.35	6.26	90
-35	0.83	2.00	96		*60*	2.17	5.91	94
-30	0.42	1.28	99		*65*	2.59	5.72	90
-25	0.45	1.17	99		*70*	3.36	4.75	87
-20	0.19	0.64	100		*75*	3.4	4.58	72
-15	0.16	0.62	100		*80*	2.83	3.13	100
-10	0.06	0.32	100		*85*	2.17	3.21	90
-5	0.1	0.37	100		*90*	3.8	5.09	72
0	0.86	0.93	100					

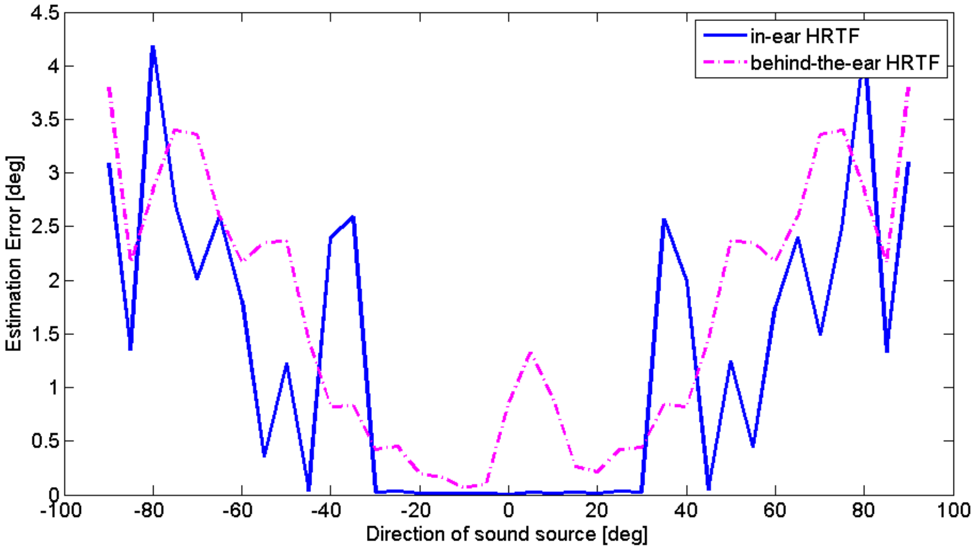


Figure 2.7: Estimation errors with clean signals under in-ear HRTF effect and behind-the-ear HRTF effect

2.3.3 Superiority of the proposed EC-BEAM algorithm

To demonstrate the superiority of the proposed EC-BEAM algorithm, its performance is further compared with that of the traditional GCC-PHAT algorithm in the clean condition. The results in DOA estimation error is plotted in Fig. 2.8. Fig. 2.8 shows that the estimation errors introduced by the EC-BEAM algorithm are much lower than those by the GCC-PHAT algorithm especially in the ranges of $[-90^o \sim -65^o]$ and $[65^o \sim 90^o]$. The possible reason is that no HRTF effect is considered in the current implementation of the GCC-PHAT algorithm. The effect of HRTF leads to the significant performance degradation of the GCC-PHAT algorithm even in the clean condition. In contrast, the effect of HRTF was fully learned through the training process in the proposed EC-BEAM algorithm.

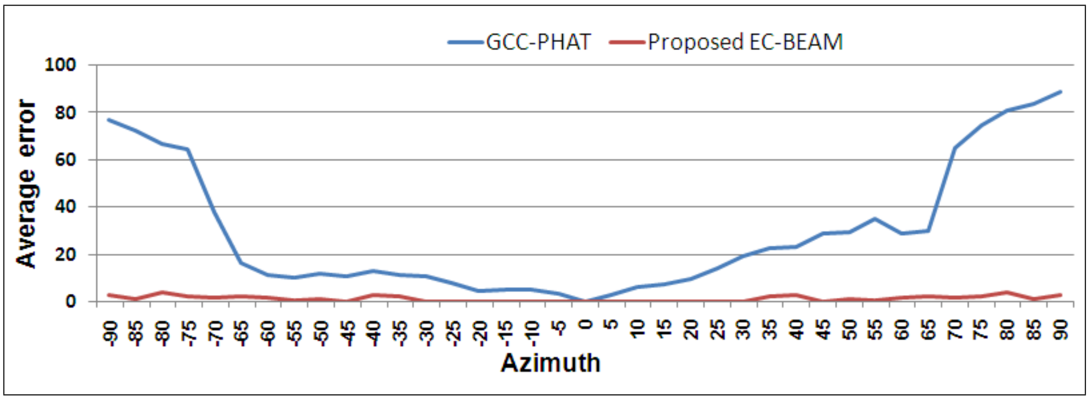


Figure 2.8: Comparison of EC-BEAM and GCC-PHAT

2.4 Conclusion

This chapter introduced the proposed EC-BEAM algorithm for DOA estimation. Regarding to the first goal of this thesis, the EC-BEAM was successfully carried with only two microphones and achieved a relatively high results with the average estimation error at 1.29^o and the accurate at 98.21 %. In noisy condition, the performance of the proposed algorithm degraded as the noise became high, however, there is almost no difference in estimation errors between the case of low-noise conditions and clean condition, especially at SNR from 10 dB or higher. This indicated that the EC-BEAM is strongly robust under low-noise conditions. Since the case in which SNR is not lower than 5 dB is the most common case in daily life, this algorithm promises to be implemented in practical system. Moreover, the experimental relatively high results in both kinds of HRTF effects, in-ear HRTF and behind-the-ear HRTF, verified that this algorithm is potentially adaptable to effects caused by system's shape. In summary, the proposed EC-BEAM fairly satisfies the three challenges of binaural sound localization, which were mentioned in section 1.3 of chapter 1.

Chapter 3

EC-BEAM in speech enhancement

In recent years, binaural speech enhancement has been extensively studied for many binaural applications, such as hearing aids. Approaches on this problem mainly attempt to enhance the target signal with preservation of its cues and suppress all signals other than target. In fact, in addition to the target signal, human beings pay attention to other important or meaningful sounds (e.g., the call from others) in the daily conversation. This attention mechanism to meaningful signals has not been considered in the state-of-the-art speech enhancement systems.

In this chapter, we firstly propose an intelligent speech enhancement system, which is able to not only enhance the target signal but also extract and present meaningful signals, based on EC-BEAM. Then, some experiments are conducted to confirm the ability of the proposed system.

3.1 Introduction

The main purpose of speech enhancement is to preserve only one signal which is considered as the target signal and reduce all undesired signals such as background noise, reverberation and non-target speech. However, in addition to the target speech, there may be other meaningful signals which usually provide important (at least useful) information. Such meaningful signals are quite popular in daily life, e.g., the ring of telephone and the call from someone probably behind the listener. In some urgent cases, furthermore, it is quite dangerous if some non-target (meaningful) signals e.g., the sound from car hooter and fire-alarm signal, are not perceived. However, state-of-the-art speech enhancement systems do not involve the function of extracting these meaningful signal, which may lead to inconvenient and/or dangerous for users [6]. Therefore, detecting and extracting meaningful signals should be indispensable for speech enhancement in speech communication and hearing assistant systems.

Due to the high performance in suppressing interfering signals, multi-channel speech enhancement technique has shown great superiority to single-channel technique. So far, many multi-channel speech enhancement systems have been proposed and widely researched, such as, delay-and-sum beamformer, generalized sidelobe canceller (GSC) beamformer [21] , transfer function GSC [19], GSC with post-filtering [10], multi-channel Wiener filter [14], and blind source separation (BSS) [1]. However, these systems normally require a large array of spatially distributed microphones to achieve higher spatial selectivity and yield single-channel monaural output, which suffers from the high complexity and loss of binaural cues at the output.

Consequently, binaural speech enhancement with two-input two-output has been studied for small physical size and low computational cost. Dorbecker *et al.* proposed a two-input two-output spectral subtraction approach [15]. Kollmeier et al. introduced a binaural noise reduction scheme based on interaural phase difference (IPD) and interaural level difference (ILD) in frequency domain [27] . Lotter et al. proposed a dual-channel speech enhancement approach based on superdirective beamforming [31]. These methods are usually based on some strict assumptions that might not be satisfied in practical environments. More recently, Li *et al.* proposed a two-stage binaural speech enhancement (TS-BASE) algorithm, which was confirmed effective in dealing with non-stationary multiple-source interference signals and preserving binaural cues [30]. However, in the original TS-BASE algorithm, no meaningful signals (other than the target signal) are taken into account and preserved at the outputs [30].

Motivated by the idea of preserving meaningful signal and taking advantage of TS-BASE and EC-BEAM, we propose an intelligent speech enhancement approach for hearing aids, namely intelligent TS-BASE (iTS-BASE), by enhancing both target and meaningful signal at the same time. To do that non-target meaningful signal is automatically detected, extracted concurrently with enhancing target signal. Specifically, the proposed model is performed through two parallel processes. The first process is to enhance target signal from a predefined direction by using the traditional TS-BASE. In the second process, non-target meaningful signal will detected by EC-BEAM, and extracted by employing TS-BASE again. Finally the enhanced target signal and the extracted signal are combined together to generate the final outputs.

3.2 The original TS-BASE model

Two-stage binaural speech enhancement (TS-BASE) was firstly proposed by Li [29] and consequently improved in [30]. Basically, the TS-BASE approach exploits Equalization-Cancellation (EC) model and Wiener Filter to enhance target signal through two stages, shown in Fig. 3.1. Specifically, it consists of the following two stages.

- *Estimation of interference signals*

  In this stage, the EC model is applied to estimate interference signals in which the equalization process is performed in training process to construct two equalizers (left and right), the cancellation process applies the two equalizers to cancel the target signal in each channel. A compensation process is further performed to make the remaining signal equivalent to interference signals based on Wiener theory. As a result, the remaining signal contains only interference signals received in each microphone.

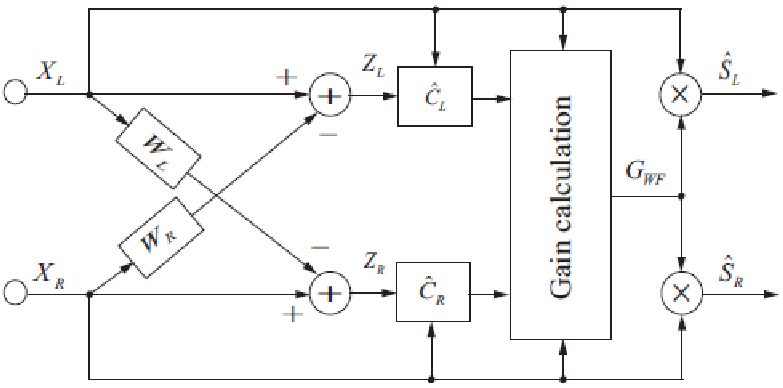


Figure 3.1: Block diagram of TS-BASE

- *Enhancement of target signal*

  The estimated interference signals in the first stage are used to construct the gain function of speech enhancer which is shared in both channels for binaural cues preservation. Finally, the gain function is applied to the original binaural input to get the enhanced signal.

3.3 The proposed intelligent TS-BASE

3.3.1 Principle of iTS-BASE

To construct an intelligent TS-BASE, a conceptual model is proposed as shown in Fig. 3.2, including two main parallel processes: (1) The first process implements the original TS-BASE to enhance target signal from a specific direction. The advantage of TS-BASE to be applied in this process is that it can extract signal from a priori known direction and can deal with multiple non-stationary noise sources. As expected, the result from this process is only the signal from target direction and the signals from other directions should be suppressed. (2) The second process attempts to detect and extract the meaningful signal which is considered as important to listener. It is strictly required that the this process must be concurrently performed and share the same input with the first process. Moreover, the meaningful signal from the non-target direction is also binaural signal with binaural cues, which are very important in some serious cases. One typical example is that when someone hears a sound from car hooter, he should be able to guess where the car is.

The key factors in this research are detecting and extracting the meaningful sounds which were never 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 car hooter, sound of fire alarm). In principle, however, 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 situations where human perceives sounds. Though meaningful signals have diverse

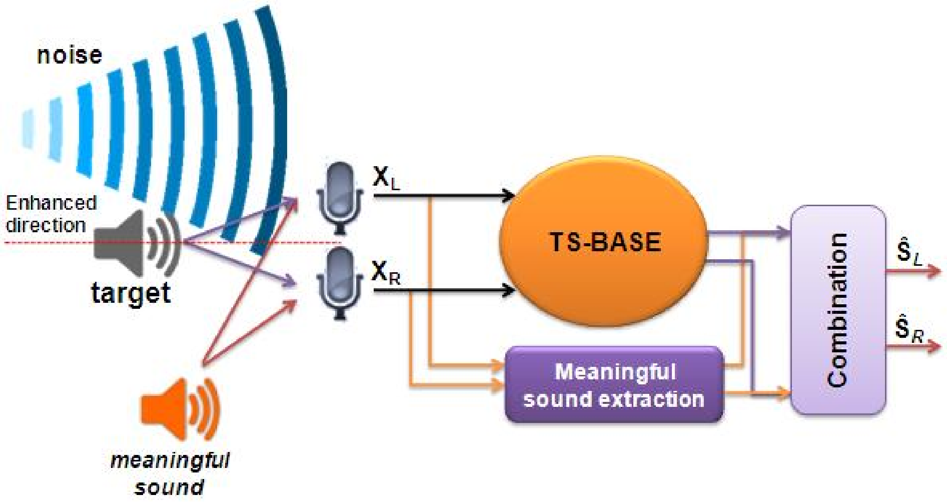


Figure 3.2: The conceptual model of the proposed intelligent TS-BASE (iTS-BASE)

characteristics that attract human's perceptual attention, in this research, the meaningful signals were limited to the sounds with the following physical characteristics for simplicity:

- *Strong energy:* The meaningful signals that human beings are interested in are normally strong enough in intensity. This is because that the weak sounds will be masked by other stronger sounds in practical environments.
- *Enough temporal duration:* The meaningful sounds are normally long enough for human to perceive. The too short sound in duration is difficult to be recognized by human.
- *Sudden occurrence:* Some meaningful sounds (e.g., telephone ring) occur with sudden increase in energy, which easily attracts the attention of human in daily-life conditions.

Actually, in addition to the above-mentioned basic characteristics, there are a lot of other characteristics for the diverse meaningful signals that generally depend on the perceptual attention of listeners in different environments. Though the dominant factors for determining meaningful signals are highly varying in different conditions, generally speaking, the above characteristics are common features for most meaningful signals in real-world conditions. In this research, only the first two characteristics are considered in the current implementation of our iTS-BASE model, detailed in the following section.

3.3.2 Implementation of iTS-BASE

Due to the the importance of meaningful sound in daily-life environments, in this research, we propose a system which can detect and extract the meaningful signals, such as, the ring of telephone, the call of someone. The proposed iTS-BASE approach consists of the original TS-BASE for enhancing the target signal, and the meaningful signal extraction which is detailed in this section.

To extract the meaningful signal, the TS-BASE is again employed in the second process since it can enhance signal from a specified direction. For the meaningful signal, we define in this research that a signal will be considered as meaningful 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). To extract the meaningful signal with TS-BASE, the direction of the meaningful signal must be determined. Specifically, the meaningful signal extraction is performed as follows: a sound source localization task is carried out to estimate the direction of arrival (DOA) of the meaningful signal, followed by the meaningful signal extraction by TS-BASE, eventually outputting the binaural target and meaningful signals by combining the output signals from two paths. The implementation flowchart of the proposed iTS-BASE is shown in Fig. 3.3.

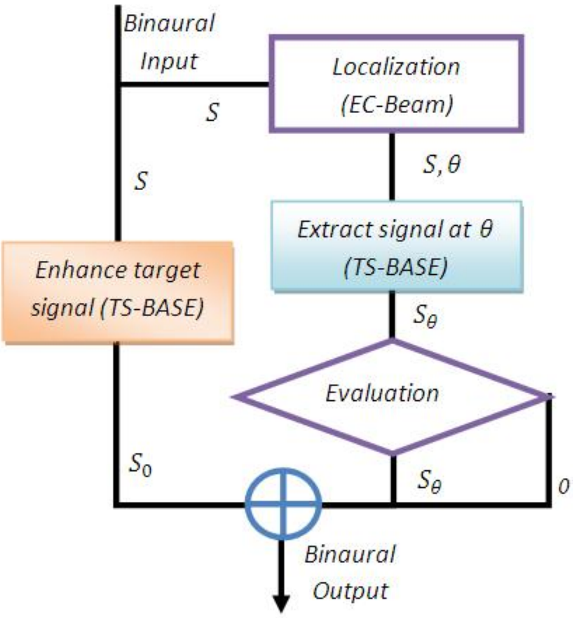


Figure 3.3: The implementation flowchart of the proposed intelligent TS-BASE

DOA estimation of the meaningful signal

In 2.3, the EC-BEAM was shown effective in high-accurately estimating the DOA of sound source under the presence of HRTF effects. In these model, the EC-BEAM is again applied for DOA estimation of the meaningful signal. Since the meaningful signal is different from the target signal, in the current implementation, the DOA of the meaningful signals is determined by scanning the non-target directions through EC-based beamforming.

Detection and extraction of meaningful signal

Since the EC-BEAM is able to estimate the DOA of the signal with the strongest energy, the possible meaningful signal can be extracted by using the TS-BASE algorithm [30]. One advantage of utilization of EC-BEAM and TS-BASE in this algorithm is that both of them based on EC-model, so they can share the same equalizers. According to the criteria for meaningful signal, it is necessary to further judge whether the extracted signal is meaningful or not. Specifically, this process checks whether the extracted signal is stronger than some threshold in intensity and longer than a certain duration. In the implementation, these thresholds were experimentally set: the threshold in intensity was 0.1 time the power of noisy signal, and that in duration was 0.2 second. The output of the meaningful signal extraction will be the extracted signal if it satisfied all criteria; otherwise, the output will be zero.

Enhancement of target and meaningful signals

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

3.4 Experiment and discussion

3.4.1 Experimental configuration

In the experiments, the situation in which the target speaker is located in the front of the listener and another guy call the listener from behind (i.e., the meaningful signal) was simulated. The target signal is the utterance selected from ATR database [28] and the meaningful signal is a recorded sound of speech "hello". To obtain the binaural sounds, the HRTF database from MIT Media lab [20] was used. The speech data were first up-sampled to 44.1 kHz and convolved with the HRTF, then down-sampled to 8 kHz. Binaural background noise was recorded at cafeteria using two microphones at the two ears of a dummy head. The target signal was assumed from the front of the listener (i.e., 0°), while the direction of the meaningful signal was set to 60°. The amplitude of the meaningful signals was controlled to make the ratio of the meaningful signal to the target signal (MTR) in average amplitude be 0.5 and 1.0, 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. In DOA estimation of the meaningful signal by EC-Beam, the direction from $[-10^{o}; 10^{o}]$ was considered as the target direction and was ignored for the meaningful signal.

3.4.2 Experimental results and discussions

The performance of the proposed iTS-BASE was evaluated in terms of two measures, namely, perceptual evaluation of speech quality (PESQ) score [22] and log-spectral distance (LSD). The evaluation results of PESQ are shown in Fig. 3.4. In general, the PESQ of the iTS-BASE algorithm is higher than that of the TS-BASE algorithm, which indicates the performance of the iTS-BASE algorithm is better than the original TS-BASE algorithm in improving speech quality. Both TS-BASE and iTS-BASE algorithms provide much higher PESQ improvements compared with the unprocessed noisy inputs. In the case MTR = 1.0, it can be observed that the PESQ of iTS-BASE is steady above the other PESQs. In this case, when SNR becomes high (or the noise becomes low), the performance of 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, however, the TS-BASE is just able to enhance signal from only one direction (target) and tends to reduce signal from other direction, including meaningful signal. When the noise becomes low, the energy of the non-target signal is mainly from meaningful signal. Since the TS-BASE algorithm removed the meaningful sound, its PESQ value becomes lower even compared to the un-proceeded 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 level.

In term of LSD measurement method, the results plotted in Fig. 3.5 show that the performance of the TS-BASE algorithm becomes worse compared to that of unprocessed when the SNR increases in both cases MTR = 0.5 and MTR = 1.0. This is due to the fact that TS-BASE removes all non-target signals including meaningful signals as explained above. Contrary, the iTS-BASE algorithm generally remains well performs stability. There is one notice that, in both cases, the LSD value of TS-BASE and iTS-BASE is the same when SNR = 0. It is because at this SNR, the noise is much bigger than meaningful sound, so that the extracted signal is not considerable compare to the remaining noise. However, in high SNR conditions, the iTS-BASE algorithm becomes better than the TS-BASE algorithm more and more. This confirms the effectiveness of the proposed iTS-BASE algorithm in extracting meaningful signals.

The performance of the original TS-BASE and proposed iTS-BASE algorithms were further evaluated through spectrograms, shown in Fig. 3.6. The spectrograms of the original and processed signals by the TS-BASE and iTS-BASE algorithm in the condition with MTR of 1 at SNR of 0 dB. It can be observed that noise in the signal enhanced by the TS-BASE algorithm (Fig. 3.6 e) was significant reduced. However, the meaningful signal also suppressed to a large degree, which leads to listeners ignoring such important sound. On the contrary, the enhanced signal by the iTS-BASE algorithm, as shown in (Fig 3.6 e), still maintains the good noise reduction performance from TS-BASE, and more importantly the meaningful signal was preserved.

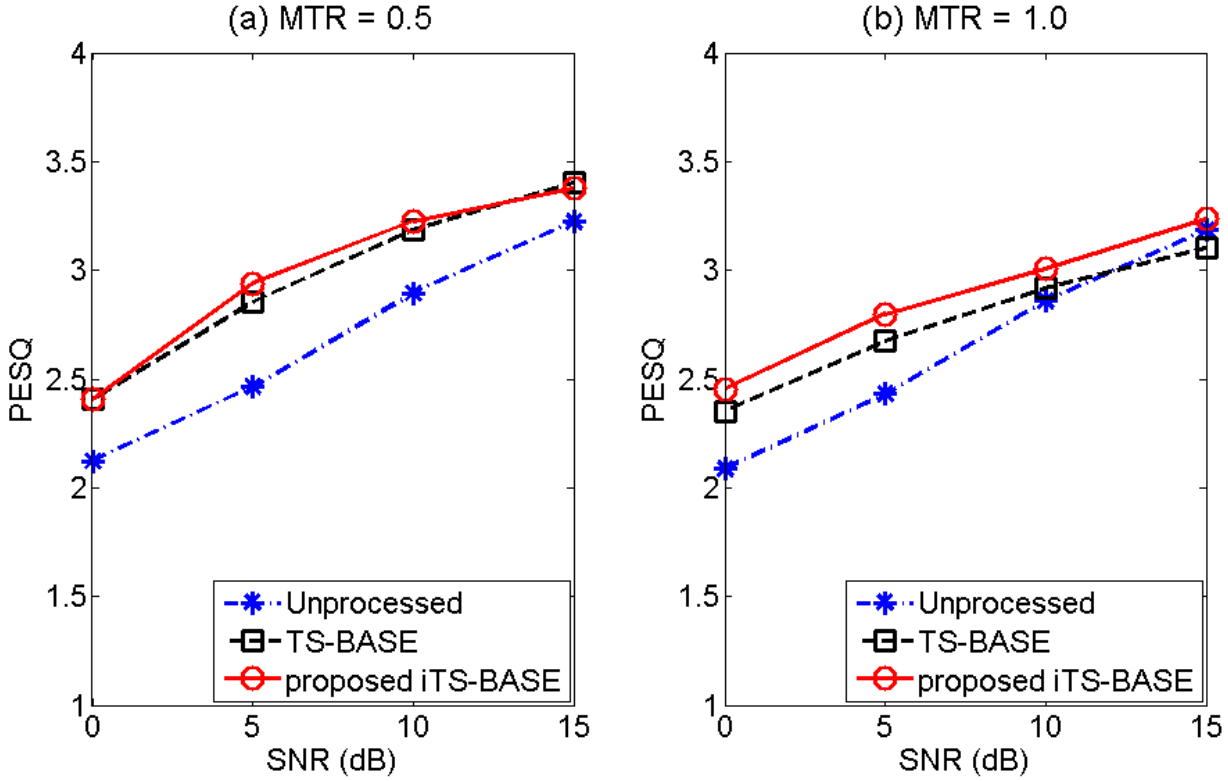


Figure 3.4: Experimental results in terms of perceptual evaluation of speech quality (PESQ) of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTS-BASE algorithm

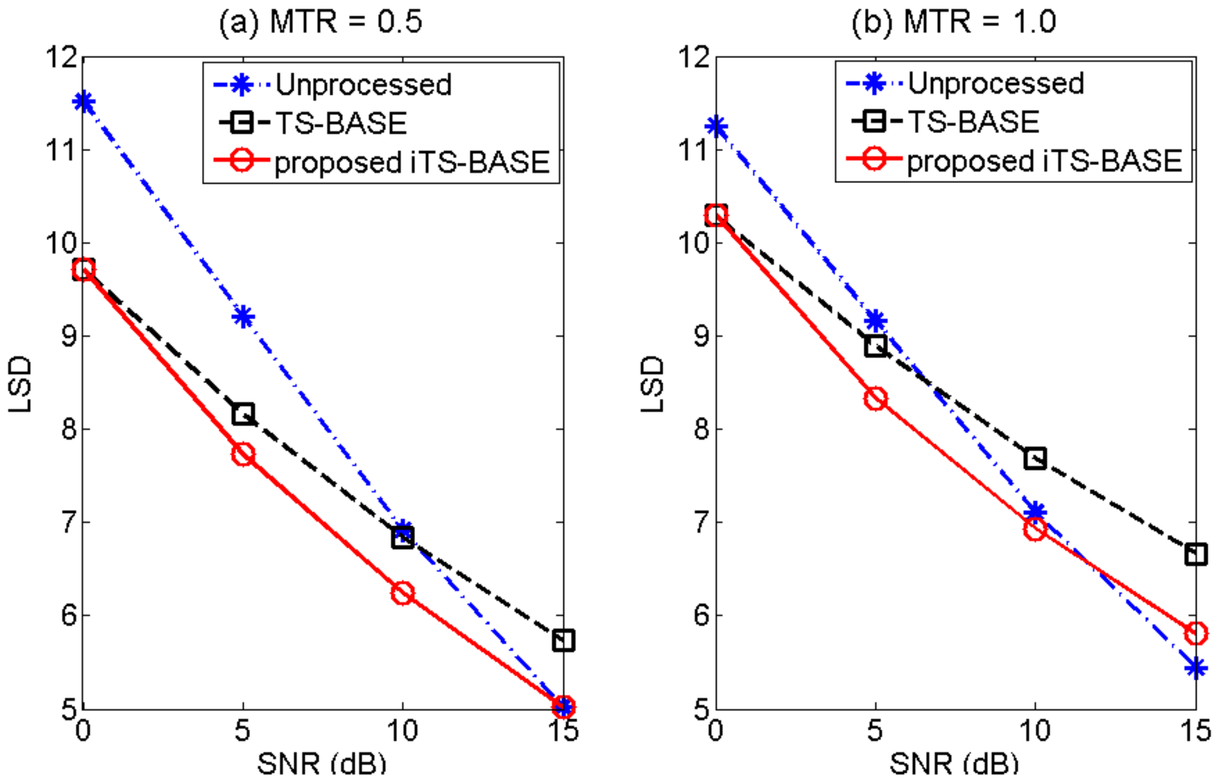


Figure 3.5: Experimental results in terms of log-spectral distance (LSD) of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTS-BASE algorithm.

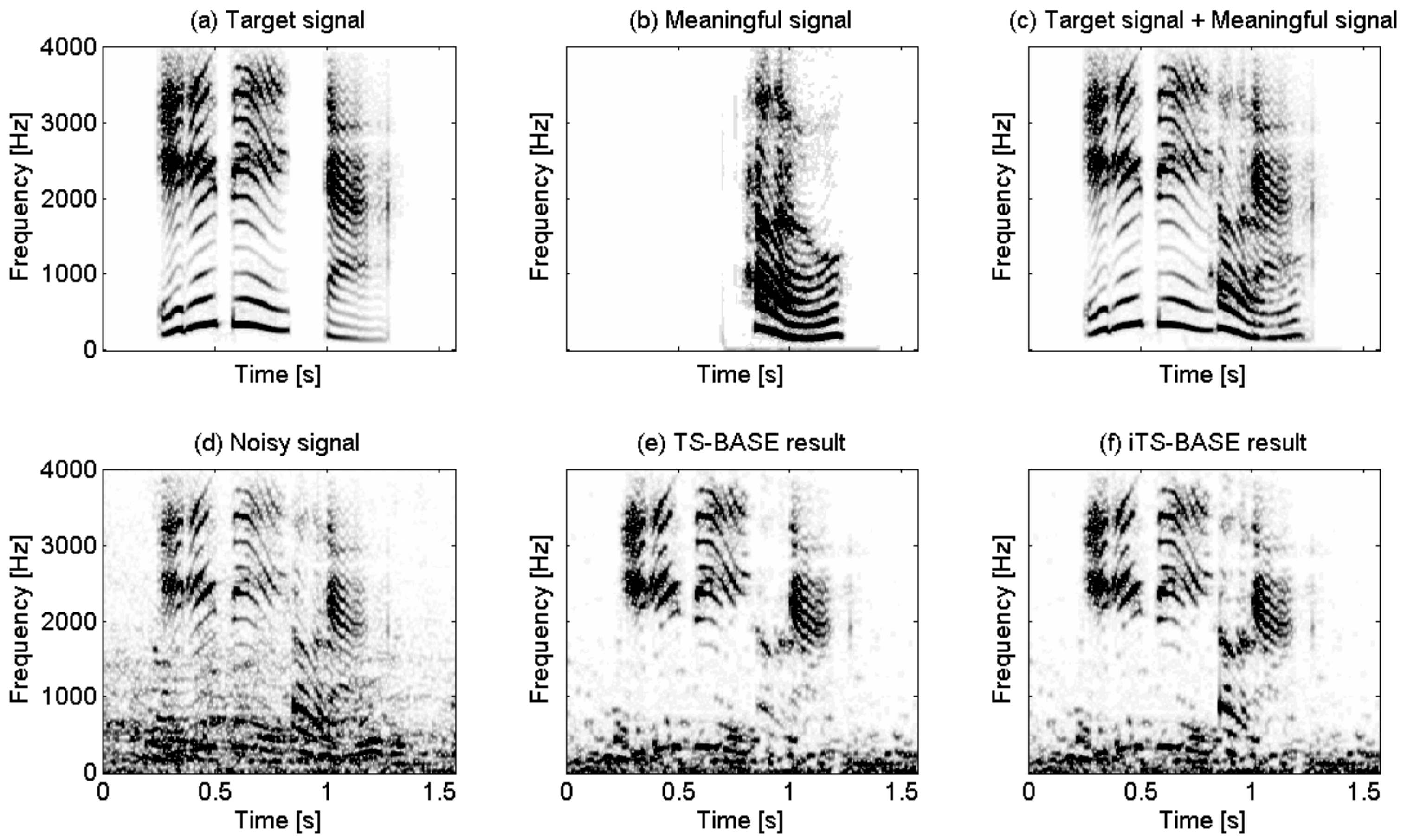


Figure 3.6: Spectrograms of the target signal, the meaningful signal, the target+meaningful signal, the noisy signal, the signals enhanced by the TS-BASE and the iTS-BASE algorithms.

3.5 Conclusion

In this chapter, an intelligent speech enhancement model, namely iTS-BASE, which is able to enhance not only target signal but also non-target meaningful signal, was proposed. The proposed iTS-BASE includes two parallel processes: target signal enhancement and meaningful signal extraction. The first process was performed by TS-BASE algorithm to enhance target signal. In the second process, we defined some criteria physically for meaningful sound, including strong energy and long temporal. The EC-BEAM was applied to detect candidate for meaningful signal by searching the biggest non-target signal as the first criterion. The detected signal was then extracted by using TS-BASE algorithm again and an evaluation task further determined whether the extracted signal was meaningful or not.

Experimental result showed that, with two simple physical criteria, the iTS-BASE can detect simple meaningful sound, for example, the speech "Hello". The evaluation in terms of PESQ, LSD and spectrogram analysis confirmed that the enhanced signal obtained by iTS-BASE is closer to the original signal than the enhanced signal obtained by TS-BASE and the unprocessed noisy signal. This result verified that this system can present two sounds (target sound and extracted meaningful sound) with sound source directions' information. This also indirectly confirms the effectiveness of EC-BEAM.

However, since determining the meaningful sound is highly depends on human perception and the listening situation, two considered criteria cannot involve all meaningful sounds in real world. Consequently, the iTS-BASE system needs more investigation on psychoacoustic cues of meaningful signal to be implemented in practice.

Chapter 4

EC-BEAM in blind source separation

The performance of EC-BEAM for localization of one source was verified in section 2.3 of chapter 2. In those experiments, the DOA of sound source was determined by finding the global minimal peak of the energy of beamformer outputs. Actually, there were several local minimal peaks and some of them corresponded to smaller sound sources. In this chapter, the EC-BEAM will be expanded to localize multiple sources and applied into blind source separation (BSS).

Conventional blind sound separation methods require information sound sources, or the number of microphones must be more than the number of sound sources. In case the number of microphones is greater than the number of sources, there are many methods can deal with this problem successfully. However, in the case that the number of microphones is less than the number of sound sources, blind beamforming technique is employed to extract individual sounds as an alternative method. Inspired by EC-BEAM and TS-BASE which was described in section 3.2 of chapter 3, in this chapter we propose a blind beamforming method using only two microphones by exploiting both algorithms for blind source separation.

4.1 Blind sound separation overview

4.1.1 The blind source separation problem

Blind source separation (BSS), also known as blind signal 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 capability of focusing his attention on his friend's speech [9].

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, such 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 speech intelligibility enhancement, noise reduction, hearing aids and cochlear implants.

Although human can effectively separate and focus on any sound in a mixture, most of BSS methods require a microphone array in which the number of microphones must more than the number of 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 state-of-the-art methods in this field.

4.1.2 Previous work

The crucial approach in BSS replies on microphone array and the problem was firstly considered as linear deconvolution. Mathematically, assume that there is a source signal vector

$$s(k) = [s_1(k)s_2(k)...s_m(k)]^T \tag{4.1}$$

where m is the number of sources and $s_i(k)$ is the ith signal source. These source signals pass through an $(m \times n)$ linear, time-invariant system with matrix impulse response $A_i (0 < i < \infty)$, the mixture of signals obtained can be expressed as follow [6]

$$x(k) = [x_1(k)...x_n(k)]^T = \sum_{i=0}^{\infty} A_i s(k-i) \tag{4.2}$$

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

$$y(k) = \sum_{l=0}^{\infty} B_l x(k-l) \tag{4.3}$$

in which $y(k) = [y_1(k)...y_m(k)]^T$ is the output vector sequence contains the estimates of individual source signals.

There have been many research investigated in this issue, which can be found in [7, 16]. BSS was then further considered as non-linear deconvolution. However, it is not clear whether human binaural processing is linear [38].

Currently, a number of approaches for BSS have been proposed. Follows are the typical methods:

- 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)

The above methods have a common requirement that the number of source m must be not greater than the number of microphones n, whereas it is often the case in human binaural processing that the number of sources m far outnumber the two ears used to collect acoustical information. An other approach to BSS is based upon the DOA of sound sources, namely directional BSS [34] and 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 [36]. Also in [36] binaural approach for BSS was mentioned (Fig. 4.1). However, the problem of preservation of binaural cues in BSS has not been considered in these methods. This will be a gap in this problem when BSS be carried out in some practical applications, especially for speech intelligibility enhancement, hearing aids.

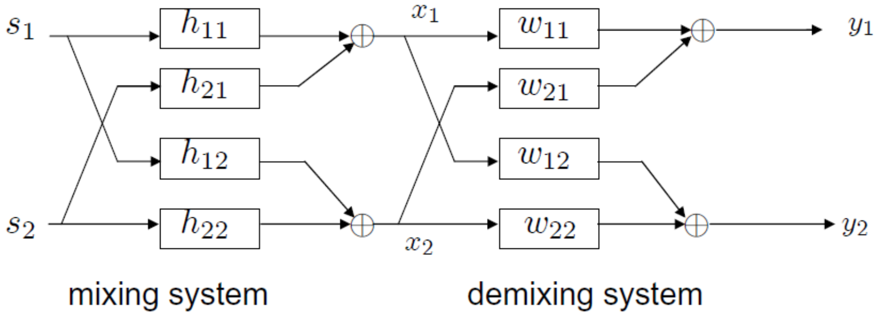


Figure 4.1: Basic two-channel BSS

4.2 Blind source separation by EC-BEAM and TS-BASE

Regarding human perception, psychoacoustic studies of binaural hearing show that the phenomenon in which human has ability to understand a signal in noise can be can explained by *binaural masking level difference* (BMLD) [3]. In signal processing, the BMLD is accounted for by EC-theory [17], which was successfully applied to speech enhancement (the TS-BASE model) [29, 30] as well as in sound localization shown in chapter 2. Motivated by EC-theory taking advantages of EC-BEAM and TS-BASE, we proposed a blind beamforming method for binaural directional BSS. Specifically, the proposed system includes two main stages: source detection by EC-BEAM and source extraction by TS-BASE.

4.2.1 Source detection

This stage is to detect the DOAs of all sound sources by applying EC-BEAM. In the beamforming process of EC-BEAM, though analysis of null-beamformer outputs, several local minimal peaks are specified, yielding the candidates for DOAs of sources. The candidate DOAs are selected by following selecting function

$$\Psi(\theta_i) = \begin{cases} 1 \quad if\ P(\theta_i) \leq P(\theta_{i-1})\ \ and\ \ P(\theta_i) \leq P(\theta_{i+1}) \\ \qquad\qquad\quad 0 \quad otherwise \end{cases} \tag{4.4}$$

where $P(\theta_i)$ this the power of beamformer output at direction θ, θ_{i-1} and θ_{i+1} are two neighbors of θ. These candidates are then checked and reduced to ensure that no candidate is close to other, in which the candidate at which the power of beamformer output is bigger will not be selected. Finally, to prevent the peaks from noise or correlation between sources, a candidate is considered as a true sound source if its beamforming value bellow a pre-defined threshold. Within this study, these thresholds are experimentally set.

4.2.2 Source extraction

In this stage, the TS-BASE is employed to extract (separate) signals at directions which were detected by Source detection stage. The extraction process is performed the same way as explained in iTS-BASE model. The equalizers are also shared for using in both EC-BEAM and TS-BASE. To execute this system, a training process is required to construct these equalizers beforehand.

4.3 Experiment and results

4.3.1 Experimental configuration

- *Data:* For speech, utterances of Japanese speakers were selected from ATR database [28]. To obtain directional sounds, the HRTF database from MIT Media lab [20] was used again. The speech data were first up-sampled to 44.1 kHz and convolved with the HRTF, then down-sampled to 20kHz.

- *Mixing:*The mixture of signals was simulated by mixing directional signals together. In this experiment, three directional speeches from three Japanese speakers were at directions -50°, 0°, 60° respectively. These directions are visualized in Fig. 4.2.

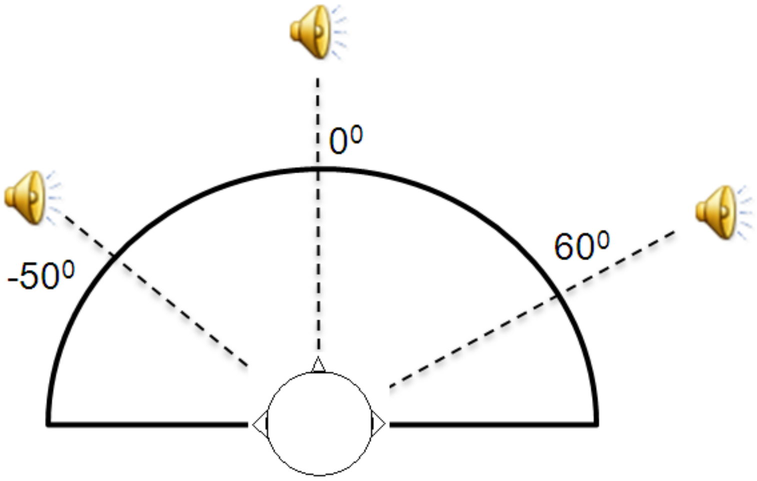


Figure 4.2: Directions of mixed signals

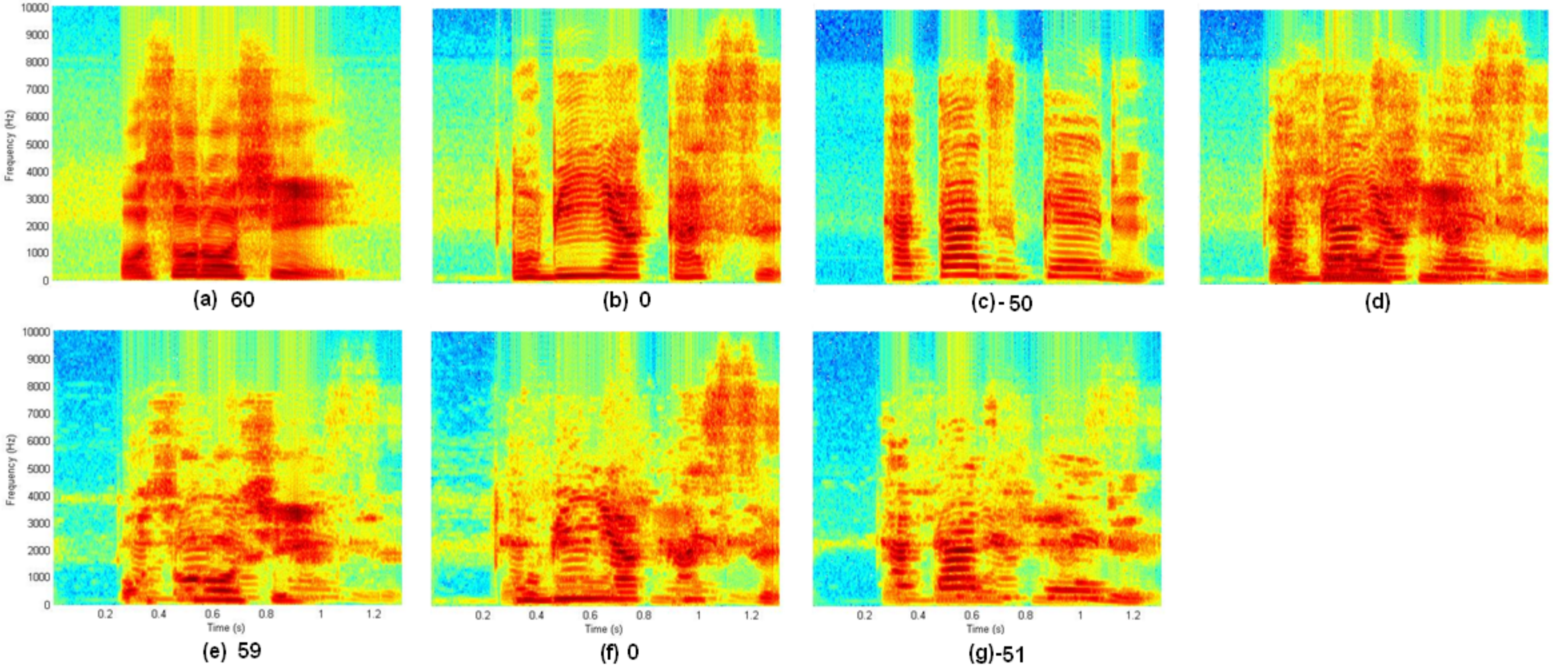
4.3.2 Experimental result

The goal in this experiment is to automatically detect and separate the three mixed signals with their binaural cues. To achieve this goal, the thresholds of *close candidates* and *beamforming power of candidate* are experimentally set for which, two candidates are close if their distance is less then 5° and the power threshold of candidate beamformer is set at 0.1 time the average power of processed signal. In result, there are three signals were detected at direction -51^o, 0^o and 60^o. There extracted signals are listened and compared with the original signals. Personally, we perceive that the extracted signals are quite similar to the original ones, and much better than mixture. However, official subjective evaluation has not been done yet.

The performance of the proposed BSS algorithms was further evaluated through spectrograms. In Fig. 4.3, the spectrogram (a), (b), (c) and (d) correspond to original signals coming from direction 60, 0, -50 and the mixture respectively. The spectrograms of signals extracted at detected DOAs (-51^o, 0^o and 59^o) are shown in Fig. 4.3 (e), (f) and (g). It can be observed that the separated signals have similar patterns as those of original ones, and much improved compared to the mixture. Moreover, the signals (e), (f) and (f) were extracted by using TS-BASE, their binaural cues should be preserved due to the ability of TS-BASE which was verified in [30].

4.4 Discussion

In this chapter, the EC-BEAM and the TS-BASE algorithms were applied to proposed a new method for blind beamforming separation technique using only two microphones. Specifically, the EC-BEAM was performed to detect multiple sound sources and TS-BASE was used to extract (separate) detected signals. In experiment, some parameters were set experimentally, for example, the power threshold for the output of candidate beamformer. Experimental result showed that, this system was able to effectively detect multiple sound sources, and in term of spectrogram analysis, the extracted signals have quite similar patterns compared with the original signals. Especially, the binaural cues of extracted signals are preserved.


(a) 60

(b) 0

(c)- 50

(d)

(e) 59

(f) 0

(g)-51

Figure 4.3: 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 this thesis, a new binaural sound localization algorithm for DOA estimation, namely EC-BEAM, has been proposed by integrating EC-model into beamforming strategy. Essentially, EC-BEAM is kind of filter-and-sum beamformer method, in which the Equalization process is applied to construct beamformer filters and the Cancellation process is applied into null-beamforming process. For each null-beamformer to certain direction, the Cancellation process applies the constructed filter to cancel the signal coming from current direction, yielding signals from other directions in beamformer output. The sound source is determined at the direction at which the energy of beamformer output gets to minimum. Several experiments have been conducted to evaluate the effectiveness of EC-BEAM, including:

- *Experinemt on signal under in-ear HRTF effect* In this part, the performance of EC-BEAM is verified in the presence of HRTF recorded by two microphones placed in ear of dummy head. Some experimental conditions have been considered, for example clean condition (target speech without any interference signal), noisy condition using simulated noise (one-source and two-source) and the real noise recorded at cafeteria.

- *Experinemt on signal under behind-ear HRTF effect* The EC-BEAM was carried out with data under effect of HRTF recored by microphones placed behind-the-ear. Only clean condition was considered to compare performance of EC-BEAM in this case and the case of clean condition under in-ear HRTF. This experiment aims to verify performance of the proposed method under different effects.

Experimental results show that the EC-BEAM with two microphone is able to localize accurately sound source, be robust under various normal noise conditions and be adaptable to different effects. This result also confirmed the satisfaction of EC-BEAM to the basic challenges in binaural sound localization. The superiority of EC-BEAM is further evaluated by comparing with the well-known SSL method, GCC-PHAT, in which the EC-BEAM achieved much higher performance since the original GCC-PHAT is not robust under the presence of HRTF.

Finally, the effectiveness of EC-BEAM is verified by applying into speech enhancement and blind source separation.

- For speech enhancement, we proposed an intelligent speech enhancement model (iTS-BASE) which is able not only enhance target signal but also detect and extract meaningful signal to present in output. Experiment was carried out with a simple meaningful sound (the speech "hello"). Experimental result which was evaluated in terms of PESQ and LSD measurement, and spectrogram analysis show that the enhanced signal obtained by iTS-BASE is the closest to the clean signal compared to noisy signal and enhanced signal obtained TS-BASE. This indicates that the iTS-BASE was successful in enhancing target signal and meaningful signal at the same time.
- For blind signal separation, a new model is proposed in which the EC-BEAM was applied to detect DOAs of multiple sound sources and TS-BASE was used to extract individual signals. Experimental result evaluated in term of spectrogram analysis show that the the spectrograms of extracted signals have similar pattern with the original individual signals. Especially, the binaural cues of extracted signals are preserved.

Through the experimental results in the proposed systems, the applicability of EC-BEAM is also confirmed.

5.2 Contributions

The main purpose of this thesis, which is to propose an approach in binaural sound source localization based on equalization-cancellation theory, has been achieved. A new binaural SSL method, EC-BEAM, was constructed by integrating EC-model into beamforming technique to estimate DOA of sound. Specifically, the proposed EC-BEAM was successfully carried out with only two microphones to estimate DOA of sound accurately in experiment of clean condition. It is also robust under various noisy environments, especially, it was shown as strongly robust under normal noise conditions. Moreover the proposed EC-BEAM was verified to be well adaptable to the effects caused by system, at least the in-ear and behind-the-ear HRTF effects. In summary, it can be said that the proposed EC-BEAM algorithm for DOA estimation well satisfies the three challenges of binaural sound localization, which was mentioned in section 1.3 of chapter 1.

Concerning EC-BEAM, two models have been proposed in speech enhancement and blind signal separation. For speech enhancement, an intelligent model, iTS-BASE, which is able to extract meaningful signal and enhance target signal, was presented. This model is an important and indispensable model to be study because in daily life, beside target sound, there are a lots of meaningful signals need to be perceived by listener but state-of-the-art methods in speech enhancement have not considered this problem. Although the iTS-BASE has just applied some typical physical criteria for meaningful detection, it was able to present two sound(target sound and extracted meaningful sound). This will be a promising system in the future if the criteria for meaningful detection are well-defined.

For blind sound separation, a new blind beamforming method was proposed to deal with this problem in case the number of microphones is less than the number of sound sources. Although with just two microphones this model was shown potential to separate sources with binaural cues preserved, which is very important in many binaural application.

5.3 Future work

In terms of EC-BEAM, the experimental result in section 4.3 of chapter 4 showed that this algorithm is potential to be expanded to multiple sources localization. There are two issues to be considered for this expansion: the first is technique to detect good minimal peaks which correspond to the candidates of sound sources; the second is how to determine the threshold to evaluate the peaks of true sound sources. Once the this expansion is successful, the source detection process in the proposed BSS method will be more effective and the proposed BSS system will be more reliable. On the other hand, the EC-BEAM also promises to be employed in microphone array for source locator. Regarding to sound localization based on GCC approach, SRP-PHAT [13] is a combination of GCC-PHAT and SRP strategy. In SRP-PHAT, all microphone pairs will steer to some candidate locations, the a position is the true sound source location, the total cross-correlation values of all pairs will reach the maximum. The EC-BEAM can follow this basic concept to localize sound source in which every pair of microphones steers the null-beamformer to candidate locations, the true sound source is the position that minimizes total energy of all beamformer outputs. An effective searching strategy may be need to be investigated.

For iTS-BASE, this thesis has just considered some simple physical criteria for meaningful signal. However, the meaningful signals in real world are more difficult to define since it highly depends on the human perception and hearing situation. As a result, such physical criteria can not cover all meaningful signal in daily life and the proposed iTS-BASE model has just been a first step to a real expected intelligent speech enhancement system. Consequently, there is a need to deeply study the characteristics of meaningful signal based upon psychoacoustic research for meaningful signal detection.

Appendix A

Supplemental Results of EC-BEAM

Following section gives detail results of EC-BEAM in several experiments. Each table summarizes results from 3700 estimates (100 utterances × 37 directions), including Average Estimation Error (AEE), Standard deviation of error (Std.), and Accuracy (Acc) in which an estimate is accurate if its error does not exceeds 5^o.

A.1 Experimental results of EC-BEAM on one-source-noise signals under in-ear HRTF effect

SNR	**5dB**			**10dB**			**15dB**		
Azimuth (deg)	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**
-90	2.93	3.25	100	3.07	3.37	100	3.09	3.41	100
-85	1.81	2.27	100	1.58	2.03	100	1.42	1.98	100
-80	6.22	6.47	52	5.30	5.60	84	4.44	4.83	91
-75	11.33	16.00	28	4.70	7.26	74	2.97	4.29	91
-70	12.40	21.46	62	2.67	6.65	93	2.04	6.03	96
-65	16.88	36.95	70	3.63	8.52	94	2.67	6.76	96
-60	7.86	21.32	88	2.05	3.65	96	1.85	3.52	98
-55	2.89	10.99	94	0.38	1.03	99	0.34	0.95	99
-50	3.94	16.07	95	1.10	1.36	100	1.19	1.46	100
-45	4.13	17.79	95	0.27	0.54	100	0.07	0.26	100
-40	7.12	21.70	88	2.45	10.35	95	2.38	10.29	95
-35	3.78	12.60	94	2.80	11.47	95	2.62	11.36	95
-30	1.13	5.01	99	0.14	0.40	100	0.04	0.20	100
-25	1.45	4.97	99	0.23	0.52	100	0.04	0.20	100
-20	2.61	14.87	98	0.07	0.27	100	0.02	0.14	100
-15	2.55	14.30	98	0.03	0.17	100	0.01	0.10	100
-10	0.28	0.53	100	0.00	0.00	100	0.01	0.10	100
-5	0.50	0.73	100	0.02	0.14	100	0.01	0.10	100
0	1.36	7.66	98	0.01	0.10	100	0.00	0.00	100
5	3.42	11.69	95	0.69	5.15	99	0.02	0.14	100
10	1.89	8.14	96	0.03	0.22	100	0.01	0.10	100
15	1.81	7.55	96	0.06	0.35	100	0.01	0.10	100
20	1.60	7.07	96	0.03	0.22	100	0.01	0.10	100
25	2.85	9.25	95	0.20	0.45	100	0.04	0.20	100
30	3.52	11.94	93	0.07	0.27	100	0.03	0.17	100
35	8.90	19.55	82	2.83	11.39	95	2.61	11.32	95
40	4.45	10.84	90	2.98	10.34	95	2.52	10.25	95
45	3.46	5.43	87	0.85	1.15	100	0.14	0.40	100
50	2.85	3.04	100	1.69	1.95	100	1.36	1.64	100
55	1.15	1.23	100	0.73	0.96	100	0.50	0.92	100
60	1.33	1.51	100	1.71	3.42	99	1.76	3.44	99
65	3.41	4.15	98	2.80	5.97	98	2.46	6.28	97
70	2.76	4.02	88	1.29	3.62	97	1.40	3.61	97
75	3.52	5.04	79	2.14	3.60	95	2.33	3.97	93
80	5.44	6.26	77	4.36	4.74	92	4.26	4.66	92
85	2.11	3.57	99	1.52	2.15	100	1.42	2.07	100
90	3.04	4.38	99	3.03	3.37	100	3.05	3.39	100

A.2 Experimental results of EC-BEAM on two-source-noise signals under in-ear HRTF effect

SNR	**5dB**			**10dB**			**15dB**		
Azimuth (deg)	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**
-90	3.03	3.34	100	3.09	3.40	100	3.08	3.41	100
-85	1.57	2.05	100	1.47	2.01	100	1.33	1.96	100
-80	7.28	11.23	69	4.51	4.84	90	4.27	4.67	92
-75	12.35	19.43	47	3.89	6.66	86	2.79	4.17	93
-70	8.54	16.05	74	2.27	6.02	96	2.02	6.03	96
-65	12.05	30.68	80	2.59	6.75	96	2.54	6.74	96
-60	3.04	6.28	94	1.92	3.64	95	1.83	3.52	98
-55	1.53	4.92	95	0.36	1.08	99	0.35	0.97	99
-50	1.12	2.00	98	1.06	1.32	100	1.12	1.39	100
-45	1.56	2.09	99	0.59	0.83	100	0.14	0.37	100
-40	2.34	8.10	96	2.00	9.21	96	1.90	9.20	96
-35	1.73	5.30	99	2.47	10.22	96	2.67	11.32	95
-30	0.22	0.51	100	0.05	0.22	100	0.02	0.14	100
-25	0.85	1.02	100	0.12	0.35	100	0.02	0.14	100
-20	0.75	0.99	100	0.05	0.22	100	0.01	0.10	100
-15	1.87	10.16	99	0.11	0.33	100	0.02	0.14	100
-10	0.38	0.69	100	0.01	0.10	100	0.01	0.10	100
-5	0.50	0.80	100	0.05	0.22	100	0.02	0.14	100
0	0.79	5.32	99	0.00	0.00	100	0.00	0.00	100
5	1.17	5.19	99	0.15	0.44	100	0.02	0.14	100
10	0.79	4.87	98	0.03	0.22	100	0.01	0.10	100
15	1.00	4.92	98	0.06	0.35	100	0.02	0.14	100
20	0.85	4.97	98	0.02	0.14	100	0.01	0.10	100
25	1.49	7.12	98	0.11	0.33	100	0.05	0.22	100
30	1.46	7.88	98	0.59	5.50	99	0.02	0.14	100
35	6.82	17.70	87	2.78	11.37	95	2.64	11.32	95
40	4.25	12.38	91	3.03	11.24	94	2.91	11.23	94
45	1.84	4.62	96	0.73	4.16	99	0.06	0.24	100
50	1.83	2.15	100	1.39	1.68	100	1.25	1.52	100
55	1.11	3.56	99	0.79	3.49	99	0.46	0.92	100
60	1.92	3.51	96	1.82	3.49	98	1.79	3.47	98
65	3.51	6.49	95	2.43	5.83	98	2.43	6.29	97
70	1.90	4.28	95	1.36	3.61	97	1.46	3.61	97
75	3.88	5.70	78	2.74	4.46	91	2.53	4.05	93
80	5.36	5.97	74	4.44	4.85	91	4.25	4.63	92
85	1.96	3.34	99	1.43	2.06	100	1.35	2.00	100
90	2.88	3.25	100	3.05	3.38	100	3.07	3.40	100

A.3 Experimental results of EC-BEAM on cafeteria-noise signals under in-ear HRTF effect

SNR	**5dB**			**10dB**			**15dB**		
Azimuth (deg)	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**	**AEE**	**Std.**	**Acc (%)**
-90	2.80	3.18	100	3.02	3.37	100	3.05	3.39	100
-85	1.96	2.49	100	1.68	2.21	100	1.44	2.05	100
-80	6.02	6.29	62	4.85	5.24	87	4.26	4.63	92
-75	9.32	14.93	59	3.66	6.49	87	2.79	4.23	93
-70	8.97	17.33	77	2.18	6.07	96	2.01	6.02	96
-65	11.38	21.40	75	3.08	8.41	95	2.54	6.74	96
-60	4.68	10.09	90	2.00	3.63	96	1.86	3.52	98
-55	1.96	6.34	96	0.39	1.11	99	0.34	0.96	99
-50	1.84	7.23	98	1.04	1.28	100	1.17	1.45	100
-45	1.32	4.90	99	0.13	0.36	100	0.06	0.24	100
-40	2.05	9.32	96	1.45	7.97	97	1.92	9.20	96
-35	2.97	10.11	96	2.80	11.42	95	2.65	11.32	95
-30	0.37	0.70	100	0.09	0.33	100	0.02	0.14	100
-25	0.61	0.89	100	0.14	0.40	100	0.05	0.22	100
-20	0.31	0.61	100	0.03	0.17	100	0.01	0.10	100
-15	0.23	0.50	100	0.04	0.20	100	0.02	0.14	100
-10	0.05	0.26	100	0.01	0.10	100	0.01	0.10	100
-5	0.02	0.14	100	0.02	0.14	100	0.01	0.10	100
0	0.00	0.00	100	0.00	0.00	100	0.00	0.00	100
5	0.02	0.14	100	0.02	0.14	100	0.02	0.14	100
10	0.04	0.24	100	0.01	0.10	100	0.01	0.10	100
15	0.28	0.58	100	0.04	0.24	100	0.02	0.14	100
20	0.35	0.64	100	0.04	0.24	100	0.01	0.10	100
25	0.68	0.94	100	0.13	0.39	100	0.06	0.24	100
30	0.39	0.71	100	0.09	0.33	100	0.02	0.14	100
35	2.48	8.74	97	2.81	11.42	95	2.67	11.32	95
40	1.61	8.10	97	1.44	7.97	97	1.92	9.20	96
45	1.35	4.90	99	0.12	0.35	100	0.06	0.24	100
50	1.84	7.24	98	1.04	1.29	100	1.18	1.46	100
55	1.91	6.33	97	0.38	1.09	99	0.34	0.96	99
60	5.03	10.87	89	1.99	3.63	96	1.86	3.52	98
65	12.17	22.30	73	3.08	8.41	95	2.55	6.74	96
70	8.41	16.96	80	2.17	6.06	96	2.01	6.02	96
75	8.77	14.65	65	3.62	6.47	88	2.74	4.18	93
80	5.94	6.20	65	4.81	5.19	87	4.23	4.61	92
85	1.93	2.45	100	1.64	2.20	100	1.43	2.05	100
90	2.82	3.20	100	3.02	3.37	100	3.05	3.39	100

References

[1] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann. Multi-channel source separation preserving spatial information. *in Proc. ICASSP2007*, pages I5–8, 2007.

[2] V. Alvarado. *Tolker localization and Optimal Placement of Microphones for a Linear Microphone Array using Stochastic Region Contraction.* PhD thesis, Brown University, Providence RI, USA, 1990.

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

[4] M. Bodden. Binaural hearing and future hearing-aids technology. *Journal de Physique III*, 4, 1994.

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

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

[7] J.-F. Cardoso. Blind signal separation: Statistical principles. *Proc. IEEE*, 90:2009–2026, 1998.

[8] S. Chandran. *Advances in direciton-of-arrival estimation.* Artech House Publisher, 2005.

[9] C. Cherry. Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Amer.*, 25:975–981, 1953.

[10] I. Cohen. Multichannel post-filtering in nonstationary noise environments. *IEEE Trans. Signal Process*, pages 1149–1160, 2004.

[11] R. Compton. *Adaptive Antennas.* Prentice Hall, 1988.

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

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

[14] S. Doclo, A. Spriet, J. Wouters, and M. Moonen. Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction. *Speech Communication*, 49(7-8):636–656, 2007.

[15] M. Dorbecker and S. Ernst. Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation. *EUSIPCO1996*, pages 995–998, 1996.

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

[17] N. Durlach. Equalization and cancellation theory of binaural masking level differences. *JASA*, 35(8):1206–1218, 1963.

[18] N. Durlach. Binaural signal detection: Equalization and cancellation theory. *In J. V. Tobias Editor, Foundations of Modern Auditory Theory*, 2:369–462, 1972.

[19] S. Gannot, D. Burshtein, and E. Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. On Signal Processing*, 49(8):1614–1626, 2001.

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

[21] J. Griffiths. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennnas Propagat.*, 30:27–34, 1982.

[22] 2000 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.

[23] D. Johnson and D. Dudgeon. *Array Signal Processing Concepts and Techniques.* Prentice Hall, 1993.

[24] H. Kayser, S. Ewert, J. Anemuller, 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*, 2009.

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

[26] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustic Speech Signal Processing*, ASSP-24:320–327, 1976.

[27] B. Kollmeier, J. Peissig, and V. Hohmann. Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain. *Scand. Audio. Suppl.*, 38:28–38, 1993.

[28] A. Kurematsu, K.Takeda, H.Kuwabara, K.Shikano, Y.Sagisaka, and S.Katagiri. Atr japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9(4):357–363, 1990.

[29] 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*, pages 263–268, 2007.

[30] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki. Two-stage binaural speech enhancement with wiener filter for high-quality speech communication. *In Press, Speech Communication*, 2010.

[31] T. Lotter, B. Sauert, and P. Vary. A stereo input-output superdirective beamformer for dual channel noise reduction. *Eurospeech2005*, pages 2285–2288, 2005.

[32] Y. Lu and M. Cooke. Auditory distance perception based on direct-to-reverberant energy ratio. *Int. Workshop Acoust. Echo Noise Contr*, 2008.

[33] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano. Active audition based humanoid system and ist evaluation: Localization, seperation and recognition of simultaneous. *Proceedings of IEEE/RSJ International Conference on Humanoids (Humanoids-2003), Springer-Verlag, IEEE*, 2003.

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

[35] A. Popper and R. Fay. *Sound source localization.* Springer -handbook for Auditory research, 2005.

[36] 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.

[37] M. Wax and T. Kailath. Optimum localization of multiple sources by passive arrays. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-31:1210–1217, 1983.

[38] W. Yost. *Fundamentals of Hearing.* Academic Press, 3 edition, 1994.

Publications

[1] Duc Thanh Chau, Junfeng Li, and Masato Akagi. A doa estimation algorithm based on equalization-cancellation theory. *Meeting of the Technical Committee of Psychological and Physiological Acoustics*, June 2010.

[2] Duc Thanh Chau, Junfeng Li, and Masato Akagi. A doa estimation algorithm based on equalization-cancellation theory. *To be appeared in Proc. Interspeech2010*, September 2010.