國立交通大學

電子工程學系 電子研究所碩士班

碩 士 論 文

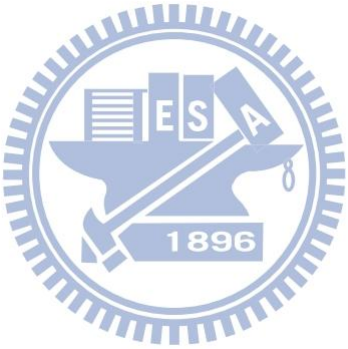
使用無響室錄音合成虛擬聆聽點

Virtual Listening Point Audio Synthesis using Anechoic Chamber Recording

研 究 生：簡士傑

指導教授：杭學鳴 教授

中 華 民 國 一O一 年 七 月


E S A
1896

使用無響室錄音合成虛擬聆聽點

Virtual Listening Point Audio Synthesis using Anechoic Chamber Recording

研 究 生： 簡士傑　　Student：Shih-Jie Chien

指導教授： 杭學鳴　博士　　Advisor：Dr. Hsueh-Ming Hang

國 立 交 通 大 學

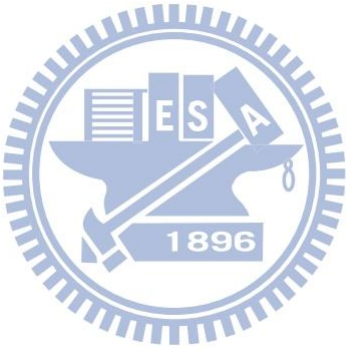
電子工程學系 電子研究所

碩 士 論 文

A Thesis
Submitted to Department of Electronics Engineering & Institute of Electronics
College of Electrical and Computer Engineering
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master of Science
in
Electronics Engineering

July 2012
Hsinchu, Taiwan, Republic of China

中華民國一 O 一年七月


E S A
1896

使用無響室錄音合成虛擬聆聽點

研究生:簡士傑　　指導教授: 杭學鳴 博士

國立交通大學

電子工程學系 電子研究所碩士班

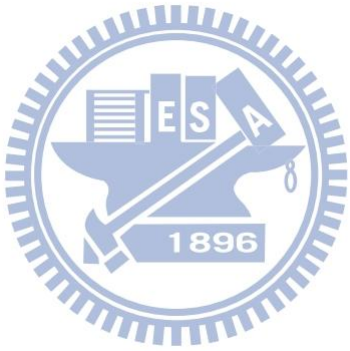
摘要

本論文的目的在於藉由無響室建造實體錄音環境，設計並且實現一個虛擬聆聽點系統。我們利用盲訊號分離(Blind Source Separation，BSS)、到達方向(Direction of Arrival，DOA)偵測、語音去雜訊等技術來建立虛擬聆聽位置的音訊，即*虛擬聆聽點語音合成*。為了達成這個目的，我們於自由聲場的無響室中佈置麥克風陣列，並且以揚聲器做為聲源錄製混合聲音訊號。

語音訊號合成主要可分成三個主要步驟，第一步驟是將錄製的混和訊號來估測原始各個音源訊號，此步驟一般使用盲訊號分離的技術來達成。第二步驟是為了估測原始各個音源訊號的來源方向，此步驟一般使用到達方向的技術來完成。第三步驟是為了在原始無麥克風錄音的虛擬位置上合成音訊，在我們系統中此步驟使用 SLAB 軟體來實現。

在實際的環境中，空氣中的雜訊干擾、音訊的失真是必然的。而這些因素會影響著訊號分離與到達方向的偵測。本篇論文中，我們提出許多主題來探討此兩

種技術，我們提供許多數據來進行驗證分析。我們將數據分成三大部分：CASE-A、CASE-B.1 和 CASE-B.2。CASE-A 為真實環境無響室中錄製的音源訊號。CASE-B.1 是利用 NASA 研究中心所研發的 SLAB 軟體來錄製音源訊號。CASE-B.2 為增加可加性白噪聲(Additive White Gaussian Noise，AWGN)於 CASE-B.1 而得的音源訊號。接著我們用語音去雜訊的技術，以改善人類主觀聽覺的品質，最後使用 SLAB 軟體來完成空間 3D 音訊處理程序。



Virtual Listening Point Audio Synthesis using Anechoic Chamber Recording

Student: Shih-Jie Chien Advisor: Dr. Hsueh-Ming Hang

Department of Electronic Engineering
Institute of Electronics
National Chiao Tung University

Abstract

The goal of this thesis is to design and implement a virtual listening point audio system by constructing a physical testing environment in an anechoic chamber. Several techniques are employed in implementing this system. They are blind source separation (BSS), direction of arrival (DOA) estimation and denoising filtering. The final outcome is constructing an audio signal at the desired virtual listening position, which is called *Virtual Listening Point Audio Synthesis*. In the Free Field Acoustic Room Chamber, each speaker represents a sound source and a microphone array records the received signals.

The audio synthesis procedure can be divided into three major steps. The first step is to separate each source signal from the recorded mixed signals. This step is usually accomplished by using the blind source separation (BSS) technique. The second step is to estimate the direction (angle) of a sound source. This step is usually accomplished by using the direction of arrival (DOA) technique. The third step is to synthesize an audio signal at a virtual point, where the original recording microphone does not exist. In our

system, this step is accomplished by using the SLAB software.

In a real acoustic environment, noise and distortion are inevitable. They disturb the BSS performance and the DOA estimation. In this project, we study the effects of several key parameters in the system. We conduct experiments, collect data, and analyze data to verify the proposed schemes. The experiments are classified into CASE-A, CASE-B.1 and CASE-B.2. CASE-A denotes the speech source recorded from the microphone arrays in the anechoic chamber. CASE-B.1 denotes the signals produced by using SLAB developed by the NASA Ames Research Center to simulate the recorded mixture signals in an ideal acoustic environment. CASE-B.2 denotes that we add the Additive White Gaussian Noise (AWGN) to CASE-B.1. We also adopt audio denoising technique to improve the subjective hearing quality. Finally, the 3-D audios are synthesized with the aid of the SLAB software.

誌謝

能完成這篇論文，我要誠摯地感謝我的指導教授 杭學鳴老師。在這兩年的碩士生涯中，讓我從老師身上學習到很多事情，讓我學習到做研究的方法，如何找問題、解決問題，還在研究上面給予專業的建議，並且在論文校稿方面以及口試投影片的製作上也給予我很好的建議。不僅如此，在老師身上也學會研究以外的事情，做人處事的道理、待人處事，老師也非常關心我們的日常生活，讓我在遭遇困難的時候，能適時給予幫助。在此向老師致上最高的感謝。

同時，我也要感謝張寶基及胡竹生兩位教授，特別抽空來擔任口試委員。還有胡竹生教授的實驗熱心贊助麥克風陣列，讓我能實際的錄製聲音。還有桑梓賢教授提供的無響室設備，我才能完成此論文。

我也要感謝羅偵源學長、張哲鳴同學，能夠在研究跟我一起討論，給予我很多意見，熱心地跟我討論研究內容，讓我可以更快進入狀況，讓我的 AUDIO 研究之路不會孤單。

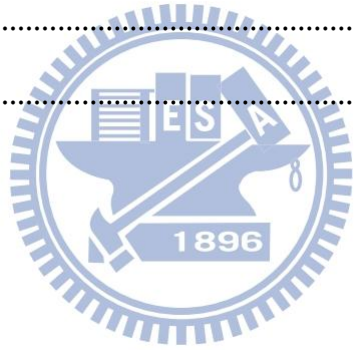
接下來我要感謝兩年碩士中一起努力的夥伴：維哲、讀修、義文、基峰、政憲、郁婷、凱翔、建志、建宏、峻利、朝雄學長和 Commlab 的各位們，陪我度過研究所的求學生活。有人可以一起討論修課，一起討論功課，因為大家，讓我的學習到很多新知識。也因為我們都有共同的目標，彼此相互扶持，相互鼓勵，讓我順利地走過這兩年的求學生活。

最後最重要的是，我要感謝我的家人，還有我的女朋友，你們永遠都是我心中的依靠。當我心情低落的時候，你們能給我與最大的安慰，讓我有信心能繼續往前走，不在徬徨、無助。感謝我的父母，因為你們的支持，讓我在求學之路可以無後顧之憂。有太多太多人需要感謝，謝謝你們。

誌於 2012 年 7 月

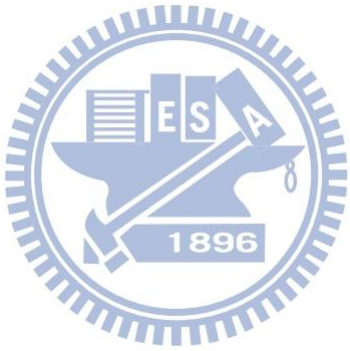
士傑

CONTENTS



LIST OF FIGURES

LIST OF TABLES



Chapter 1 Introduction

1.1 Motivation

Recently, the free-viewpoint TV (FTV) system concepts have been developed. It is well-known that the FTV system uses a sequence of images to synthesize and compress multiple images. The target of FTV is to synthesize an image at any viewpoint as we want. Thus, it allows the users to choose an arbitrary viewing angle.

Similarly, we can use multiple microphones to record multiple audio signals. At the receiver, an audio signal can be synthesized at a virtual listening point, which is the so-called *Virtual Listening Point Audio Synthesis*. However, the procedure of solving the audio problem is quite complex and is different from that of the video. Our goal is to reproduce a virtual audio from the recorded mixed signals by using the microphone array in an anechoic chamber.

1.2 Contributions and Organization of Thesis

In this thesis, our main target is to synthesize virtual listening-point audio in a real environment. The acoustic signal synthesis procedure can be divided into three major steps. The first step separates the source signals under the blind condition and the second step estimates the source directions (locations). The third step synthesizes the new listening-point audio.

For the first step, we use the blind source separation (BSS) technique to separate individual sound source from the mixed signals. We model and use the known mathematical tools [1], [2] to solve the separation problem. The subspace of interest is extracted by the principal component analysis (PCA) method [4]. For solving the permutation problem, there are many conventional methods are proposed in [5], [6], and

we adopt [7] and [8]. The scaling problem is solved by the minimum distortion principle (MDP) [10] or the subspace methods [9]. There are many well-known BSS methods and one of the most popular methods is the so-called independent vector analysis (IVA) [8]. The IVA method has different learning rules [12] and different properties from the conventional ICA methods.

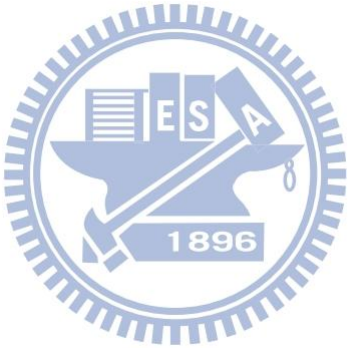
For the second step, we use the direction of arrival (DOA) technique to locate individual sound source from the mixed signals. The time difference of arrival (TDOA) [1] is a basic concept to explain the technique. It also has to satisfy some conditions in order to avoid the spatial aliasing [20]. The DOA technique can be solved under the invariant property assumption [21]. There are many proposed methods such as [22], [23], [24], [25], [26] and [27]. We adopt [28] to estimate DOA estimation for 3-D sources.

For the third step, we separate sources and identify their locations using the methods described in the first step and the second step. We adopt the software developed by the NASA Ames Research Center to synthesize the audio at a virtual listening point.

Because we record the audio signal in a real world environment, we need to consider the noise effect. There are many advanced applications require audio denoising techniques [13], [14], [15]. We adopt [16] to solve the denoising problem in our system.

This thesis contains six chapters. In Chapter 2, we describe the acoustic signals model and the adopted BSS method for sound separation. In Chapter 3, we describe the adopted denoising method for improving audio quality, and we also describe the adopted DOA method for the source localizations. In addition, we describe the overall system for virtual listening point audio synthesis. In Chapter 4, we describe the recording environments setting and show the BSS experimental results. In Chapter 5,

we show the DOA experimental results. In Chapter 6, based on the experimental results, we make a brief conclusion and suggest future research topics.



Chapter 2 Blind Source Separation

2.1 Model of Acoustic Signals

For a microphone array system, we can use signal processing techniques to describe and solve microphone array problems. We construct a signal model that includes acoustic signal and microphone array in the space, and then we model and use the mathematical tools to solve the target problems. A system is often modeled as a machine that takes in "inputs" and produces "outputs". In our research, the "Input" is the sound source signals and the "Output" is the received signals. There are four types of Input-Output systems: [1] single-input single-output (SISO), single-input multiple-output (SIMO), multiple-input single-output (MISO) and multiple-input multiple-output (MIMO). We assume the system model is linear and shift-invariant, and the channel impulse response is a Finite Impulse Response (FIR) filter, whose impulse response is of finite duration [2].

MIMO is an abbreviation of the Multiple-Input Multiple-Output system. In this study, we use the multiple-input multiple-output system to model the sound signals with microphone array. Assume the system model involves K input signal and N output signals as shown in Fig. 1, which can be modeled as:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

where

$$\mathbf{x}(t) = [x_1(t) \quad x_2(t) \quad \dots \quad x_N(t)]^T$$

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2 \quad \cdots \quad \mathbf{A}_K]$$

$$\mathbf{A}_k = \begin{bmatrix} a_{1k,0} & a_{1k,1} & \cdots & a_{1k,L-1} \\ a_{2k,0} & a_{2k,1} & \cdots & a_{2k,L-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{Nk,0} & a_{Nk,1} & \cdots & a_{Nk,L-1} \end{bmatrix}_{N\times L}, k = 1,2,...,K$$

$$\mathbf{n}(t) = [n_1(t) \quad n_2(t) \quad \cdots \quad n_N(t)]^T$$

where a_{nk} $(n = 1,2,...,N, k = 1,2,...,K)$ denotes the channel impulse response of the input of k-th signal and the output of n-th signal and L denotes the channel length. Then, we can represent the transfer function of the system in the z-domain.

$$\mathbf{X}(z) = \mathbf{A}(z)\mathbf{S}(z) + \mathbf{N}(z) \tag{2}$$

where

$$\mathbf{A}(z) = \begin{bmatrix} A_{11}(z) & A_{12}(z) & \cdots & A_{1K}(z) \\ A_{21}(z) & A_{22}(z) & \cdots & A_{2K}(z) \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1}(z) & A_{N2}(z) & \cdots & A_{NK}(z) \end{bmatrix}$$

$$A_{nk}(z) = \sum_{l=0}^{L-1} a_{nk,l} z^{-l}, n = 1,2,\ldots,N, k = 1,2,\ldots,K$$

and $\mathbf{X}(z), \mathbf{S}(z), \mathbf{N}(z)$ denote the $\mathbf{x}(t), \mathbf{s}(t), \mathbf{n}(t)$ in the z-domain.

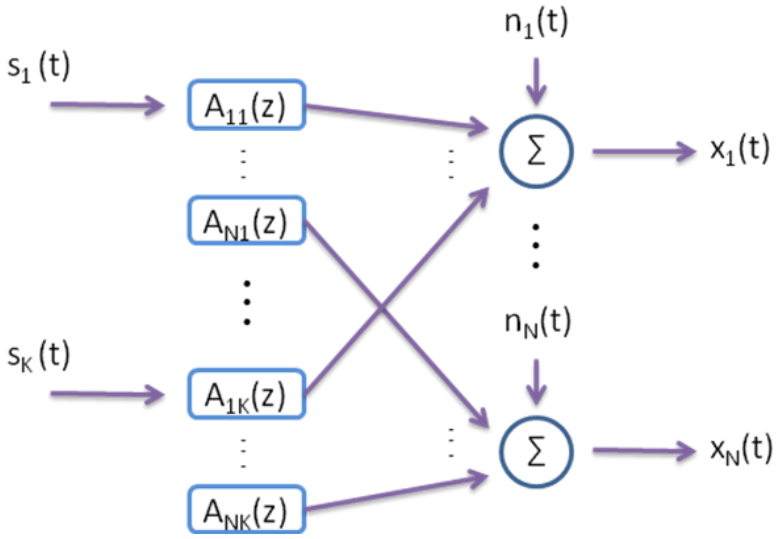


Fig. 1 MIMO System

2.2 Cocktail Party Problem

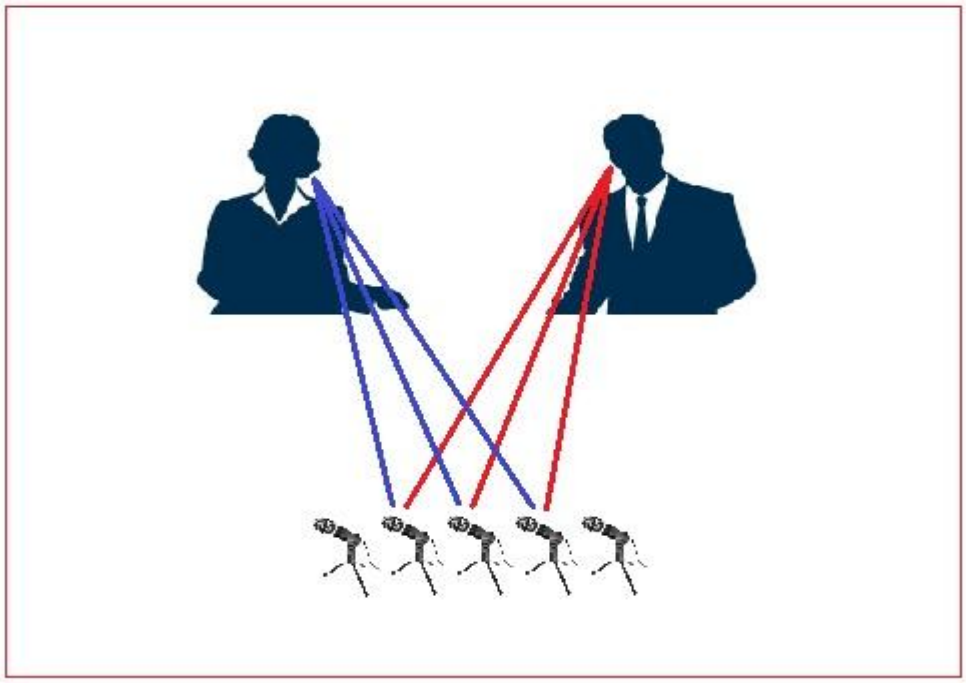


Fig. 2 Cocktail Party Problem

The “Cocktail Party Problem” in [3] is known as one of the most famous problems in the area of acoustic signal processing. Cocktail Party Problem refers to how to focus on a single speaker among a mixture of conversations and background noises. In this study, we view each speaker as a sound source and a microphone array placed in the Free Field Acoustic Room Chamber. How can we separate the sound sources based on the mixture signals recorded by a microphone array? It is difficult to solve this problem under condition of “blind”, which means that the mixture signals and mixing procedure are unknown. The goal of Blind Source Separation (BSS) is how to use signals processing techniques to recover the sound sources from the recorded mixture signals.

2.3 Independent Component Analysis (ICA)

2.3.1 Fundamental of ICA

Independent Component Analysis (ICA) is a popular BSS method to separate a mixture signals into components. In this study, we consider the frequency domain approach of ICA, which means that we transform signals into frequency domain, called FD-ICA. In the section 2.1, we use the MIMO system to model the signals. In many BSS methods, it is necessary to have the prior knowledge about the number of sound sources. However, we make an assumption that the number of microphones is greater than the number of source signals, which means that the exact number of sound sources is not critical but we need to have sufficient number of microphones to solve the problems. Let K be the number of source signal $\mathbf{s}$ and N be the number of recorded signal $\mathbf{x}$ with $N > K$. It can be modeled as:

$$\mathbf{x}(f,t) = \mathbf{A}(f)\mathbf{s}(f,t) + \mathbf{n}(f,t) \tag{3}$$

where

$$x_1(f,t) = a_{11}(f)s_1(f,t) + a_{12}(f)s_2(f,t) + \cdots + a_{1K}(f)s_K(f,t)$$

$$x_2(f,t) = a_{21}(f)s_1(f,t) + a_{22}(f)s_2(f,t) + \cdots + a_{2K}(f)s_K(f,t)$$

$$\vdots$$

$$x_N(f,t) = a_{N1}(f)s_1(f,t) + a_{N2}(f)s_2(f,t) + \cdots + a_{NK}(f)s_K(f,t)$$

where $\mathbf{x}(f,t) = \left[x_1(f,t),\ldots,x_N(f,t)\right]^T$ denotes the vector of mixture signals and $x_n(f,t)$ denotes the short-time Fourier transform (STFT) of the n-th microphone in the t-th time frame, $\mathbf{A}(f) = \left[\mathbf{a}_1(f),\ldots,\mathbf{a}_K(f,t)\right]$ denotes the mixing matrix and $\mathbf{a_k}(f) = \left[a_{1k}(f), a_{2k}(f) + \cdots + a_{Nk}(f)\right]^T$ denotes the STFT of the k-th channel impulse response in the t-th time frame, $\mathbf{n}(f,t)$ denotes the noise components in the recorded

signals. Therefore, $\mathbf{A}(f)\mathbf{s}(f,t)$ represents the principal components of $\mathbf{x}(f,t)$. For simplicity, we assume there are no room reflections and ambient noises, the received signals can be modeled as:

$$\mathbf{x}(f,t) = \mathbf{A}(f)\mathbf{s}(f,t) \tag{4}$$

The goal of Independent Component Analysis (ICA) is to estimate the demixing matrix $\mathbf{W}$, which can be written as:

$$\mathbf{y}(f,t) = \mathbf{W}(f)\mathbf{x}(f,t) \tag{5}$$

where $\mathbf{y}(f,t)$ denotes the separated signals.

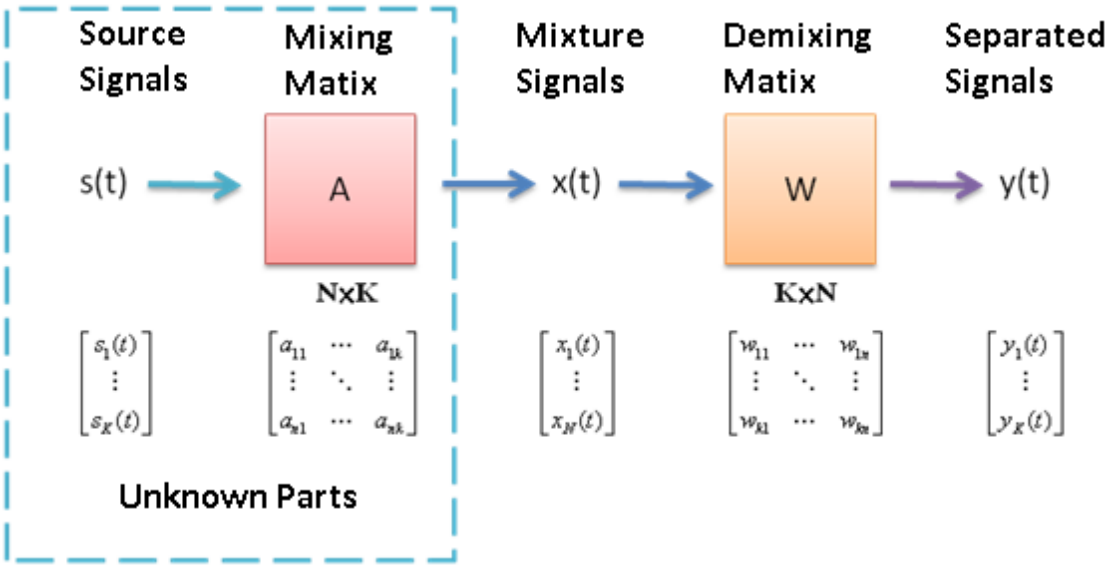


Fig. 3 BSS Filter Structure in Time Domain

In Fig.4, the recorded signals $\mathbf{x}(t)$ are transferred from time domain to frequency domain. We also apply the Principal Component Analysis (PCA) to pre-process the recorded signals $\mathbf{x}(f,t)$. Then, we use FD-ICA Algorithm to deal with the permutation and the scaling problems. At the end, after inverse DFT, we obtain the separated signals $\mathbf{y}(t)$ in time domain.

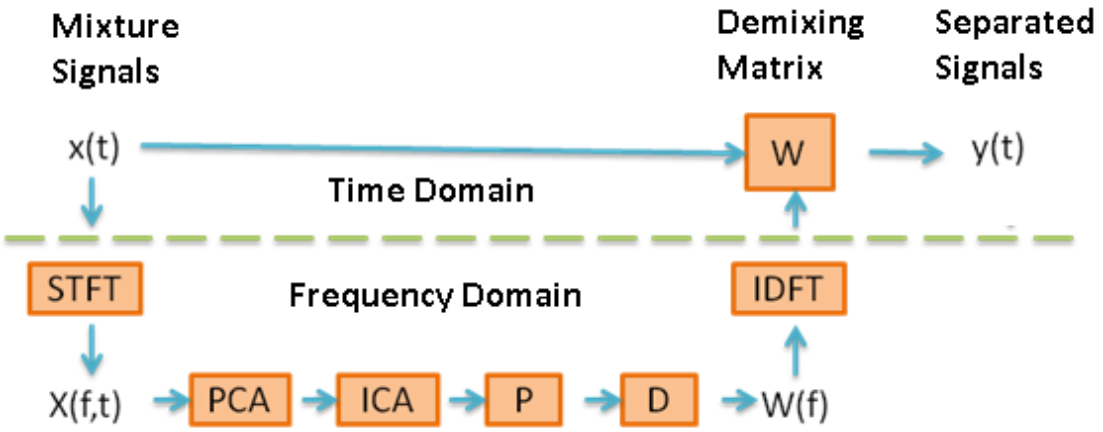


Fig. 4 Flow Chart of Demixing Matrix in Frequency Domain

2.3.2 Principal Component Analysis (PCA)

Since the number of mixture signals is greater than the number of source signals, by utilizing Principal Component Analysis (PCA) [4], we can obtain the subspace signals in which the room reflections and ambient noises are reduced. In other words, the subspace of principal components is reserved and the subspace of the reflection components is discarded. Thus, we simplify the recorded signals by using the PCA procedure, and then the estimation complexity of ICA can be reduced. The subspace signal $\mathbf{x}'$ is obtained by the following expression:

$$\mathbf{x}' = \mathbf{x} - E\{\mathbf{x}\} \tag{6}$$

It is mean that by centering signal $\mathbf{x}$, we obtain the zero-mean signal for decreasing the estimation complexity.

$$\mathbf{x}_p(f,t) = \mathbf{V}(f)\mathbf{x}(f,t) \tag{7}$$

where the subspace filter $\mathbf{V}(f) = \mathbf{D}(f)^{-1/2}\mathbf{S}(f)^T$ is derived by the Principal Component Analysis (PCA) method, $\mathbf{D}(f)$ denotes the eigenvalue diagonal matrix and $\mathbf{S}(f)$ denotes the eigenvector matrix corresponding to $\mathbf{D}(f)$.

Define the covariance matrix by $\mathbf{C}_{\mathbf{x}}(f) = E\left[\mathbf{x}(f,t)\mathbf{x}(f,t)^T\right]$. Then, $\mathbf{D}(f)$ and $\mathbf{V}(f)$ are obtained from the covariance matrix, $\mathbf{C}_{\mathbf{x}}(f)$ can be decomposed into $\mathbf{C}_{\mathbf{x}}(f) = \mathbf{S}(f)\mathbf{D}(f)\mathbf{S}(f)^{-1}$; $\mathbf{S}(f)$ is an orthogonal matrix that should satisfy the equation $\mathbf{S}(f)\mathbf{S}(f)^T = \mathbf{S}(f)^T\mathbf{S}(f) = \mathbf{I}$.

It is assumed that the significant eigenvalues are associated with the direct components from the source signals and the rest are associated with the noises. From the covariance matrix, we form an eigenvalue subspace matrix $\mathbf{D} = diag\left\{d_1,\ldots d_p, d_{p+1},\ldots,d_N\right\}$, where $d_1 > d_2 > \cdots > d_p \gg d_{p+1} > \cdots > d_N$. The eigenvectors corresponding to $d_1,\ldots,d_p$ represent the principal components of the signals, and the eigenvectors corresponding $d_{p+1},\ldots,d_N$ represent the reflections and ambient noises. Therefore, PCA can reduce the complexity of ICA process.

2.3.3 Characteristics of ICA

In the above PCA procedure in [4], we have to make some assumptions or restrictions in solving the BSS problem using ICA. First, the goal of ICA is to make the output signals $\mathbf{y}$ to be statistically independent. In other words, the joint probability distribution of the output signals $\mathbf{y}$ equals to the product of each marginal distribution, which can be shown by the following expression:

$$p(\mathbf{y}) = \prod_{i=1}^{K} p_i(y_i) \tag{8}$$

Second, we make the output signals to be non-Gaussian distribution. Assume that there exists a group of non-Gaussian distribution, and the sum of these independent signals y_i will to be close to Gaussian distribution, which is asserted by the so-called central limit theorem (CLT). On the other hand, if there exists a group of Gaussian

distribution, the sum of a group of signals is also Gaussian distribution. Then, the ICA method cannot separate these source signals.

2.4 Permutation Problem and Scaling Problem

2.4.1 Permutation Problem

The permutation problem is that we do not know the permutation of the independent components. In other words, the problem can make the signals be confused at transformation from frequency domain to time domain. We assumed that there exists a permutation matrix $\mathbf{P}$, which satisfies the following equation:

$$\mathbf{x} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s} \tag{9}$$

Assume $\mathbf{Ps}$ is kind of the independent components. If we re-arrange the order of the permutation matrix, it does not affect the independence of the internal signals in matrix $\mathbf{s}$. When we select another matrix $\mathbf{P}$, $\mathbf{Ps}$ corresponds to another permutation of the independent component, so the demixing matrix $\mathbf{W} = \mathbf{P}\mathbf{A}^{-1}$ is not unique. However, the ICA method is to find the demixing matrix at the each frequency bin. If the permutation of each frequency is not consistent, we transform the signals from frequency domain to time domain cannot be correct. For finding the solution of the permutation problem, there are many conventional methods in [5][6][7]. We adopt [8] to solve the permutation problem.

2.4.2 Scaling Problem

The scaling problem is that we do not known the energy of independent component, which means that the separated signals are multiplied by non-zero constants, that is, the separated signals are different from the original sources in magnitude. To solve the

scaling problem, there is a conventional method which filters individual output by the pseudo-inverse of the demixing matrix as proposed by [9].

In this study, we solve the scaling problem by using the minimal distortion principle [10]. Since we adopt the blind separation method, the mixing matrix $\mathbf{A}(f)$ is unknown. For simplicity, we assume the permutation matrix is that $\mathbf{P}(f)=\mathbf{I}$. The ideal scaling matrix $\mathbf{D}(f)$ should satisfy the following equation:

$$\mathbf{D}(f)\mathbf{W}(f)\mathbf{A}(f)=diag[\mathbf{A}(f)] \tag{10}$$

Once the separated signals are well-separated by the ICA method, there exists a diagonal matrix $\mathbf{Q}(f)$ such that $\mathbf{W}(f)\mathbf{A}(f)=\mathbf{Q}(f)$. Hence, the equation can be rewritten as $\mathbf{D}(f)\mathbf{Q}(f)=diag[\mathbf{A}(f)]$, and the mixing matrix $\mathbf{A}(f)$ can be estimated by $\mathbf{A}(f)=\mathbf{W}^{+}\mathbf{Q}(f)$, where $\mathbf{W}(f)^{+}$ is the Moore-Penrose pseudo-inverse of the separation matrix $W(f)$. Therefore, the estimation, $\mathbf{D}(f)=diag[\mathbf{W}(f)^{+}]$, is an approximation to the solution of the scaling problem in the FD-ICA.

2.5 Fast fixed-point Independent Vector Analysis

In this study, we adopt [7][8] as the ICA method, which is described in the preceding sections. The Independent Vector Analysis (IVA) method uses a different approach to solve the BSS problem by assuming that the source signals have certain dependency in the frequency domain. Under this hypothesis, the original sources are dependent together as a group by using the multidimensional prior. The model is a maximum likelihood approach to the multidimensional ICA (MICA), which is called independent vector analysis.

2.5.1 Independent Vector Analysis (IVA)

When the mixture signals are transformed into frequency domain, the mixing process can be modeled by the following mixing model:

$$\mathbf{x}(f,t) = \mathbf{A}(f)\mathbf{s}(f,t) \tag{11}$$

$$\mathbf{y}(f,t) = \mathbf{W}(f)\mathbf{x}(f,t) \tag{12}$$

$$f = 1, 2, \ldots, F$$

where F denotes the number of frequency bin. Since the ICA algorithm treats the source signals as independent and identically distribution (i.i.d), the IVA consists of a group of ICA, which is so-called multidimensional ICA (MICA) as shown in Fig. 5. Each 2×2 IVA mixture model denotes the ICA layer at a single frequency bin. For simplicity, we will drop the index t with only one variable f in the equations.

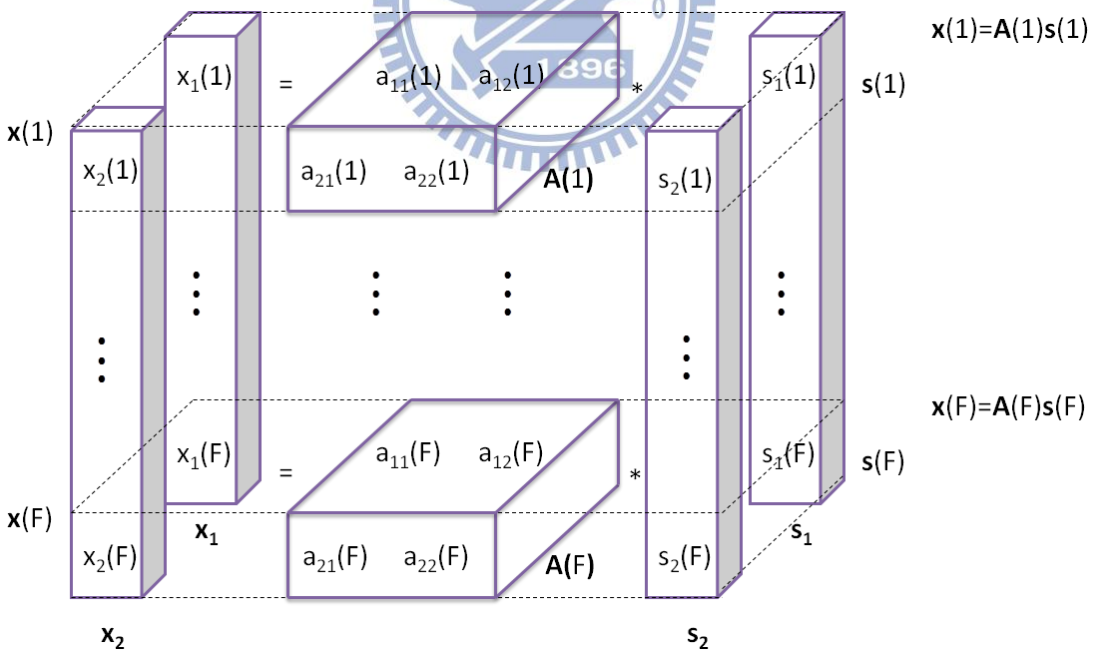


Fig. 5 IVA Structure in the frequency domain

In Fig. 5, $\mathbf{s}_1 = [s_1(1), \ldots, s_1(F)]^T$ and $\mathbf{s}_2 = [s_2(1), \ldots, s_2(F)]^T$ denote the multivariate

sources and $\mathbf{x_1} = [x_1(1), \ldots, x_1(F)]^T$ and $\mathbf{x_2} = [x_2(1), \ldots, x_2(F)]^T$ denote the mixture signals.

In the MICA algorithm, it also assumes the zero-mean and whitening input signals $\mathbf{x}$, so it can increase the learning speed. Therefore, the demixing matrix $\mathbf{W}$ needs to be orthogonal. These conditions are shown by the following functions:

$$E[\mathbf{x}(f)\mathbf{x}(f)^H]^T = \mathbf{I} \tag{13}$$

$$\mathbf{W}(f)\mathbf{W}(f)^H = \mathbf{I} \tag{14}$$

The IVA method can be represented by the mutual information of the output signals $\mathbf{y}$. Its contrast function is represented by the following equation:

$$D(P(\mathbf{y}) \| \prod_i P(\mathbf{y}_i)) \tag{15}$$

where $D(\cdot \| \cdot)$ denotes a contrast function of the Kullback-Leibler divergence, which measures the distance between two density distributions, and $P(\mathbf{y}_i)$ denotes the marginal distribution of $\mathbf{y}_i$. When the contrast is minimized to zero, $\mathbf{y}_i$ is expected to be independent of each other, which is represented as $P(\mathbf{y}) = \prod_i P(\mathbf{y}_i)$.

The ICA algorithm based on IVA consists of two steps. The first step is to find contrast function as the input learning function. The second step is to choose optimization method.

2.5.2 Contrast Function

In the above of the expression, the contrast function of IVA can be represented by the mutual information among multidimensional variable $\mathbf{y}_i$'s:

$$D(P(\mathbf{y}) \| \prod_i P(\mathbf{y}_i)) = \sum_i H(\mathbf{y}_i) - H(\mathbf{y}) \tag{16}$$

where $H(\mathbf{y}) = -\sum_i p(\mathbf{y}_i) \log p(\mathbf{y}_i)$ denotes the entropy function, and $\mathbf{y}_i = [y_i(1), y_i(2), \ldots, y_i(F)]^T$ is a vector of the i-th separated signal. However, the term $H(\mathbf{y})$ is constant since $\log\left|\det(\mathbf{W}(f))\right| = 0$ at any frequency bin f. The equation is equivalent to minimize the sum of the entropies of $\mathbf{y}_i$.

$$\arg \min_{\mathbf{W}(1),\ldots,\mathbf{W}(F)} \sum_i H(\mathbf{y}_i) \tag{17}$$

As in IVA method, negentropy can be employed for the derivation of the entropic contrast, which is defined as:

$$N(\mathbf{y}_i) = D(P(\mathbf{y}) \,\|\, P(\mathbf{y}^N)) \tag{18}$$

where $P(\mathbf{y}^N)$ denotes the information projection of y onto the Gaussian manifold. By Pythagorean relation, $N(\mathbf{y}_i) = H(\mathbf{y}_i^N) - H(\mathbf{y}_i)$ represents the relation in information geometry. Substituting (18) into (17), that is, another contrast function is obtained by the following expression:

$$\arg \min_{\mathbf{W}(1),\ldots,\mathbf{W}(F)} \sum_i H(\mathbf{y}_i) = \arg \max_{\mathbf{W}(1),\ldots,\mathbf{W}(F)} \sum_i N(\mathbf{y}_i) \tag{19}$$

$$\sum_i N(\mathbf{y}_i) = \sum_i E_{\mathbf{y}_i}[\log(P(\mathbf{y}_i))] + const \tag{20}$$

where $E_{\mathbf{y}_i}[\cdot]$ denotes the expectation of probability distribution $\mathbf{y}_i$. In spite that we have the entropy contrast, there exists the problem which is difficult to obtain the true distribution $p(\mathbf{y}_i)$ in finite data size. In order to solve the problem, it uses source prior to substitute the source distribution.

$$\sum_i E\left[\log(P_{\mathbf{s}_i}(\mathbf{y}_i))\right] = E\left[\log\left(P_{\mathbf{s}}(\mathbf{y})\right)\right] = E\left[\log\left(P_{\mathbf{W}^{-1}\mathbf{s}}(\mathbf{x})\right)\right] \tag{21}$$

where $P_{\mathbf{s}_i}$ denotes the estimation of source signal probability distribution, which is the so-called source prior. Here, we introduce a symmetric exponential norm distribution (SEND), which has the following expression:

$$\hat{P}_{\mathbf{s}_i}(\boldsymbol{s}_i) \propto \frac{\exp\left\{-\sqrt{\frac{2}{F}\sqrt{\sum_f |\boldsymbol{s}_i(f)|^2}}\right\}}{\sqrt{\sum_f |\boldsymbol{s}_i(f)|^2}^{2F-1}} \tag{22}$$

By replacing $P_{\mathbf{s}_i}(\cdot)$ in the contrast function $E\left[\log\left(P_{\mathbf{s}_i}(\mathbf{y}_i)\right)\right]$, the contrast function becomes:

$$G\left(\sum_f |\mathbf{y}_i(f)|^2\right) = -\log P_{\mathbf{s}_i}(\mathbf{y}_i) \tag{23}$$

where $G(z) = \sqrt{\frac{2z}{f}} + (F - \frac{1}{2})\log z$ with the constraint that $\mathbf{W}_i(f)$ are normalized. By using the Lagrange multiplier λ_i, we can include the normalization constraint in the contrast function:

$$\sum_i \left[E\left[G\left(\sum_f |\mathbf{W}_i(f)^H \mathbf{W}_i(f)|^2 \right)\right] - \sum_f \lambda_i(f)\left(\mathbf{W}_i(f)^H \mathbf{W}_i(f) - 1\right)\right] \tag{24}$$

2.5.3 Optimization Method

In order to apply the IVA algorithm to the BSS problem in frequency-domain, we have to deal with complex-valued variables. However, these complex variables can be expressed as circular symmetry around origin in most cases. Hence, the complex values should satisfy the following equations:

$$E\left[\mathbf{x}\mathbf{x}^T\right] = \mathbf{O}$$

$$E\left[\mathbf{x}\mathbf{x}^H\right] = \mathbf{I}$$

Once the contrast function is selected, we can derive the separating matrix by selecting the optimization method. Most ICA algorithms use the gradient descent technique as the learning rule [11]. However, [12] uses the Newton's method, which is

called FastICA algorithm [12]. Here, we assume $g(\cdot)$ is the input learning function from (23), which can be expressed as:

$$g\left(\mathbf{W}_i(f)\right) = E\left[G\left(\sum_f |\, \mathbf{W}_i(f)^H \mathbf{W}_i(f)\,|^2 \right) - \sum_f \lambda_i(f)\left(\mathbf{W}_i(f)^H \mathbf{W}_i(f) - 1\right) \right] \tag{25}$$

The function can be approximated by the quadratic Taylor polynomial:

$$\begin{aligned} g\left(\mathbf{W}_i(f)\right) \approx\; & g\left(\mathbf{W}_{i,o}(f)\right) + \frac{\partial g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^T}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right) \\ & + \frac{\partial g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^H}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^* \\ & + \frac{1}{2}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^T \frac{\partial^2 g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f) \partial \mathbf{W}_i(f)^T}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right) \\ & + \frac{1}{2}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^H \frac{\partial^2 g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^* \partial \mathbf{W}_i(f)^H}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^* \\ & + \left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^H \frac{\partial^2 g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^* \partial \mathbf{W}_i(f)^T}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right) \end{aligned} \tag{26}$$

The optimization of $g(\cdot)$ will set the gradient $\partial g(\cdot) / \partial$ to zero.

$$\begin{aligned} \frac{\partial g\left(\mathbf{W}_i(f)\right)}{\partial \mathbf{W}_i(f)^*} \approx\; & \frac{\partial g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^*} + \frac{\partial^2 g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^* \partial \mathbf{W}_i(f)^T}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right) \\ & + \frac{\partial^2 g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^* \partial \mathbf{W}_i(f)^H}\left(\mathbf{W}_i(f) - \mathbf{W}_{i,o}(f)\right)^* \equiv \mathbf{O} \end{aligned} \tag{27}$$

Note that, the equation is equivalent to Newton step equation, which can be reduced to

$$\mathbf{W_i}(f) - \mathbf{W}_{i,o}(f) = -\frac{1}{c\left(\mathbf{W}_{i,o}(f)\right)} \cdot \frac{\partial g\left(\mathbf{W}_{i,o}(f)\right)}{\partial \mathbf{W}_i(f)^*} \tag{28}$$

where $c\left(\mathbf{W}_{i,o}\right)$ is the constant multiplication term. By substitution, the iterative algorithm becomes as the following equation:

$$\mathbf{W}_i(f) \leftarrow \mathbf{W}_{i,o}(f) - \frac{E\left[y_{i,o}(f)^* G'\left(\sum_f |\, y_{i,o}(f)\,|^2 \right) \mathbf{x}(f) \right] - \lambda_i(f) \mathbf{W}_{i,o}(f)}{E\left[G'\left(\sum_f |\, y_{i,o}(f)\,|^2 \right) + |\, y_{i,o}(f)\,|^2 G''\left(\sum_f |\, y_{i,o}(f)\,|^2 \right) \right] - \lambda_i(f)} \tag{29}$$

where G' and G'' denotes the first order and second order differentials, and $\lambda_i(f) = E\left[|y_{i,o}(f)|^2 G'\left(\sum_f |y_{i,o}(f)|^2 \right) \right]$, which denotes the Lagrange multiplier. Also, instead of evaluating $\lambda_i(f)$, we can remove it by replacing the numerator and denominator of equation (29):

$$\mathbf{W}_i(f) = \hat{E}\left[G'(\sum_f |y_{i,o}(f)|^2) + |y_{i,o}(f)|^2 G''(\sum_f |y_{i,o}(f)|^2) \right] \mathbf{W}_{i,o}(f)$$
$$- \hat{E}\left[(y_{i,o}(f)^*) G'(\sum_f |y_{i,o}(f)|^2) \mathbf{x}(f) \right]$$

In addition to normalization, the rows of the demixing matrix $\mathbf{W}$ have to be decorrelated. The learning rules of $\mathbf{W}$ can be expressed as:

$$\mathbf{W}(f) \leftarrow \left(\mathbf{W}(f)\mathbf{W}(f)^H \right)^{-\frac{1}{2}} \mathbf{W}(f)$$

It should be calculated by above equation to make $\mathbf{W}(f)$ convergent at each frequency bin.

2.6 Evaluation of the BSS Performance

In evaluating the performance of BBS algorithms, one way is to measure the signal to interference ratio (SIR). The definition of SIR is described below:

$$SIR = \frac{10}{K} \sum_{i=1}^{K} \log_{10} \frac{\left\langle \max_n |y_{n,s_i}(t)|^2 \right\rangle}{\left\langle \sum_{j \neq i} |y_{n,s_j}(t)|^2 \right\rangle}, n = 1, 2, \ldots, N$$

where the signal $y_{n,s_i}(t)$ denotes the n-th output separated signal corresponding to the i-th source signal. For $n = i$, $|y_{n,s_i}(t)|^2$ denotes the power of the n-th desired separated signal with the same signal, and for $\sum_{j \neq i} |y_{n,s_j}(t)|^2$ denotes the sum of the interference power from the other source signal to the n-th separated signal.

Chapter 3 Acoustic Signal Processing and Synthesis

3.1 Audio Denoising Algorithm

3.1.1 Introduction

In acoustic signal processing, audio noise reduction is an important issue. In a real acoustic environment, the environment parameters including the air absorption, the surface reflection and microphone intrinsic distortion, and others, all generate audio noises. However, in many cases, it is assumed that there is no reverberation effect, which is called the single-path assumption. Many advanced applications require audio denoising techniques [13][14][15]. We adopt [16] to solve the denoising problem in our system.

3.1.2 Audio Noise Modeling and the Pointwise Wiener filter

Here, the noises are considered to be random variables, and they all have corrupted by the additive Gaussian noise. Considering the observation z_i of the recorded signals, it can be modeled as:

$$z_i = x_i + n_i \tag{30}$$

where n_i denotes a zero-mean white Gaussian random sequence, which means that $E[n_i] = 0$ and $E[n_i n_j] = \sigma_n^2 \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta function and $E[\cdot]$ denotes the expectation; and x_i denotes the audio signals. It is known that:

$$E[x_i] = E[z_i] \tag{31}$$

$$\overline{\sigma}_x^2 = E\left[\left(x_i - \overline{x}_i\right)^2\right] = E\left[\left(z_i - \overline{z}_i\right)^2\right] - \sigma_n^2 \tag{32}$$

where $\bar{x}_i$ denotes the sample mean and $\bar{\sigma}_x^2$ denotes the sample variance in a W-size window. Under this hypothesis, a local linear minimum mean square error filter (LLMMSE) is formed. It is a well-known technique to solve the denoising problem. It is proposed in [17] and given by:

$$x_i = \bar{x}_i + \frac{\bar{\sigma}_x^2}{\bar{\sigma}_x^2 + \sigma_n^2}(z_i - \bar{x}_i) \tag{33}$$

where x_i denotes an estimate (filter output) of the i-th sample point.

3.1.3 A Contextual Wiener Filter

An adaptive scheme designed based on minimization of the Fisher information metric is proposed by [16]. It is an approach of using the Markovian model. In other words, the audio is modeled as a Gaussian Markov Random Field (GMRF), which denotes a sample dependency on neighboring audio signals (non-causal filter). A MRF is a set of random variables that have the Markov property. In general, a window size is selected as $W = 3$ or $W = 5$ of the observed signals. Then, we use the expected Fisher Information as the cost function. The Fisher information can be defined as the variance of the score in [18]:

$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] \tag{34}$$

where $E_\theta[\cdot]$ denotes the expectation over θ and $\ell(\theta)$ denotes the logarithm of the pseudo-likelihood function. In the statistical theory, a pseudo-likehood is an approximation to the probability distribution, and this approximation provides a good calculation alternative to the Fisher Information on the Markove Random Field models [19]. Because it is difficult to calculate the expected Fisher information at real situation,

the observed Fisher Information can be approximated by the pseudo-likelihood equation:

$$I_{obs}(\theta) = \left[\frac{\partial}{\partial \theta} \log PL(\theta) \right]^2 \tag{35}$$

By the Law of Large Numbers, it can be estimated by the following equation:

$$\hat{I}^1_{obs}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\partial}{\partial \theta} \log p(x_i \mid \eta_i, \theta) \right]^2 \Bigg|_{\theta=\theta} \tag{36}$$

where $p(x_i \mid \eta_i, \theta)$ denotes the local conditional density function (LCDF) of the Markovian model and η_i denotes the neighboring element of the i-th sample. The LCDF is a density function depending on the neighboring elements. Therefore, $\hat{I}^1_{obs}(\theta)$ is an unbiased estimator of the observed Fisher Information, that is, $I_{obs}(\theta) = E\left[\hat{I}^1_{obs}(\theta) \right]$, making $\hat{I}^1_{obs}(\theta) \approx I_{obs}(\theta)$.

For the continuous acoustic signal, an isotropic GMRF is suitable for the analysis, which is characterized by a set of LCDF`s, each one is described by equation:

$$p(x_i \mid \eta_i, \mu, \sigma^2, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[x_i - \mu - \sum_{x_j \in \eta_i} \beta \left(x_j - \mu \right) \right] \right\}^2 \tag{37}$$

where μ, σ^2 denotes the mean and the variance, and β is a control parameter depending on neighboring samples. By $\hat{I}^1_{obs}(\theta) \approx I_{obs}(\theta)$, substituting (37) into (36), the observed Fisher Information with a closed expression in the GMRF model is given by:

$$\hat{I}^1_{obs}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{\sigma} \left[x_i - \mu - \sum_{x_j \in \eta_i} \beta (x_j - \mu) \right] \left[\sum_{x_j \in \eta_i} (x_j - \mu) \right] \right\}^2 \tag{38}$$

Let the first factor of (38) be zero, we derive the following equation:

$$x_i = \mu + \sum_{x_j \in \eta_j} \beta (x_j - \mu) \tag{39}$$

We assume a non-stationary model, and the parameters $\mu_i, \sigma_i^2, \beta_i$ are time variant. The parameter β_i, which is the gain of a filter, can be rewritten as:

$$\beta_i\left(\sigma_i^2, \sigma_n^2\right) = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_n^2} \tag{40}$$

where σ_n^2 denotes the noise variance. Substituting (40) into (39) to obtain the following expression:

$$x_i = \mu_i + \left(\frac{\sigma_i^2}{\sigma_i^2 + \sigma_n^2}\right) \sum_{x_j \in \eta_i} \left(x_j - \mu_i\right) \tag{41}$$

where the parameters μ_i, σ_i^2, β_i are estimated in a local adaptive way, which essentially calculate these parameters based on the observed samples in a window.

In section 3.1.2, we introduce the Pointwise Wiener Filter [17]. By renaming the variables using the same notations, we combine these two adaptive filters into a contextual adaptive Wiener filter, which is represented by

$$x_i = \bar{x}_i + \left(\frac{\bar{\sigma}_x^2}{\bar{\sigma}_x^2 + \sigma_n^2}\right)\left[\alpha(z_i - \bar{x}_i) + (1-\alpha)\sum_{z_j \in \eta_i} (z_j - \bar{x}_i)\right] \tag{42}$$

where $\alpha \in [0,1]$. For $\alpha = 1$, the formulation becomes the pointwise filter and for $\alpha = 0$, the formulation becomes the pure contextual filter. However, according to the Peak Signal-To-Noise Ratio (PSNR) measured in [16], $\alpha = 0.79$ is the optimum tradeoff value between two adaptive filters.

3.2 Direction of Arrival

3.2.1 Introduction to Direction of Arrival (DOA)

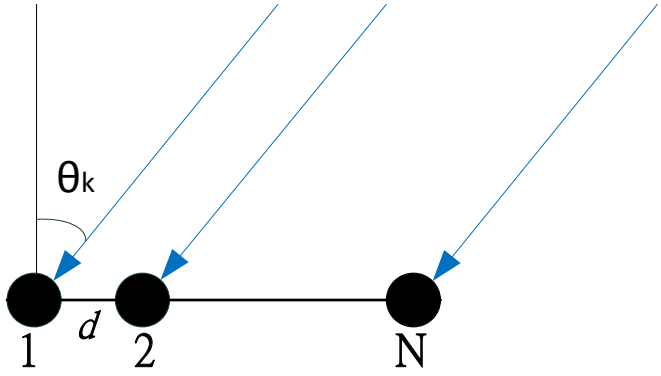


Fig. 6 Wave propagation and Microphone Array geometry

In acoustic signal processing, the source direction of interest is an important issue, which would be estimated by the Direction of Arrival (DOA) technique. In a real acoustic environment, the propagation of sound wave reaches walls, ground, and etc, which result in reflection. The phenomenon is called reverberation or multi-path effect. However, in many cases, it is assumed that there is no reverberant effect as we receive the source signals from the microphone array. Furthermore, we assume that the source signals are located far away from the microphone array, and it is the so-called far-field signals. Then, the propagation of source signal can be approximated as plane wave. As illustrated by in Fig. 6, different microphones receive signals along paths of different lengths, the phenomenon of Time Difference of Arrival (TDOA) [1]. Because the recording microphone array is linear and uniform placed, the time delays satisfy the following equation:

$$\tau_{nk} = \frac{(n-1)d\sin\theta_k}{c} \tag{43}$$

where τ_{nk} represents the delay time between the k-th source signal reaching the first sensor and the n-th sensor, c denotes the sound propagation velocity, which is about 340 m/s, d denotes the distance between two adjacent microphones, and it has to satisfy the condition $d \le (1/2)\lambda$ in order to avoid the spatial aliasing [20]. The DOA problem refers to how to estimate the angle of θ_k from the mixture signals. In this study, we view each speaker as a sound source and a microphone array is placed in the Free Field Acoustic Room Chamber.

3.2.2 DOA Estimation Based on ICA

We assume that there is no reverberant effect as we receive recorded signals from the microphone array, and all source signals reach the microphones at the same time. The mixing model x is obtained by the following expression:

$$x_n(t) = \sum_{k=1}^{K} a_{nk} s_k(t) \tag{44}$$

where a_{nk} is the attenuation generated by the k-th source to the n-th microphone. Considering the vector-matrix notation, the mixing model can be written as:

$$\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{a}_k \mathbf{s}_k(t)$$

and

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

where $\mathbf{a}_k$ is the k-th column of matrix $\mathbf{A}$, which is called mixing matrix in Chapter 2. Here, it is also called the steering vector matrix. This model is called the instantaneous or delay-less mixture model.

Considering the steering vector matrix, the matrix can be written as following expression:

$$\mathbf{A}(f)=[\mathbf{a}_1(f),\mathbf{a}_2(f),\ldots,\mathbf{a}_K(f)] \tag{45}$$

where $\mathbf{a}_k(f)$ denotes the k-th column vector of matrix $\mathbf{A}(f)$, which is made of the attenuation coefficients.

When the source signals are well-separated exactly at each frequency, we can obtain the steering matrix $\mathbf{H}(f)=\mathbf{W}(f)^{-1}\mathbf{D}(f)^{-1}$, where $\mathbf{W}(f)$ and $\mathbf{D}(f)$ denote the demixing matrix and scaling matrix in Chapter 2. However, the scaling matrix does not affect to the ratio of elements in the same column. This invariant property can be shown below [21].

$$\frac{\mathbf{A}_{ik}(f)}{\mathbf{A}_{mk}(f)}=\frac{\left[\mathbf{W}^{-1}(f)\mathbf{D}^{-1}(f)\right]_{ik}}{\left[\mathbf{W}^{-1}(f)\mathbf{D}^{-1}(f)\right]_{mk}}=\frac{\left[\mathbf{W}^{-1}(f)\right]_{ik}}{\left[\mathbf{W}^{-1}(f)\right]_{mk}} \tag{46}$$

Where $[\cdot]_{ik}$ denotes the i-th row and the k-th column element of the matrix and $i\neq m$. Then, we can use the invariant property to derive the relation as below in [21]

$$\frac{\mathbf{A}_{ik}(f)}{\mathbf{A}_{mk}(f)}=\frac{a_{ik}}{a_{mk}}\exp\left\{-j2\pi f(i-m)d\sin\theta_k c^{-1}\right\} \tag{47}$$

And then we can extract the angle θ_k:

$$\theta_k(f)=\sin^{-1}\frac{angle\left[\mathbf{A}_{ik}(f)/\mathbf{A}_{mk}(f)\right]}{2\pi f(i-m)dc^{-1}}$$

$$=\sin^{-1}\frac{angle\left[\mathbf{W}_{ik}^{-1}(f)/\mathbf{W}_{mk}^{-1}(f)\right]}{2\pi f(i-m)dc^{-1}} \tag{48}$$

3.2.3 DOA Estimation for 3-D Source Signals

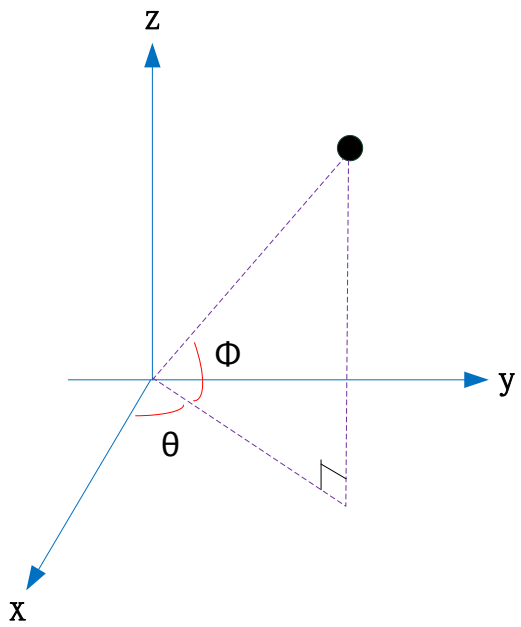


Fig. 7 Relationship between Source and Microphone Array

The 2-D DOA estimation is known as one of the popular methods in the area of acoustic signal processing. In general, the source signals and the microphone arrays are placed in the same plane as shown in Fig.6, and the estimated angle is the azimuth angle θ. In many cases, the problem is solved by using the uniform linear array (ULA), and there are many conventional method such as [21][22][23]. When we consider the source signals and the microphones in a real environment, the DOA is a three dimensional problem as shown in Fig. 7. Then, in addition to the azimuth angle θ, and we also need to find the elevation angle ϕ. In this approach, the microphone array is not necessary to be an ULA. However, the source direction has the higher accuracy of estimation when it is arranged as two ULA such as L-shape microphone array in [24]. There are many methods in [25][26][27], and we adopt [28] to estimate our DOA estimation of 3-D source signals.

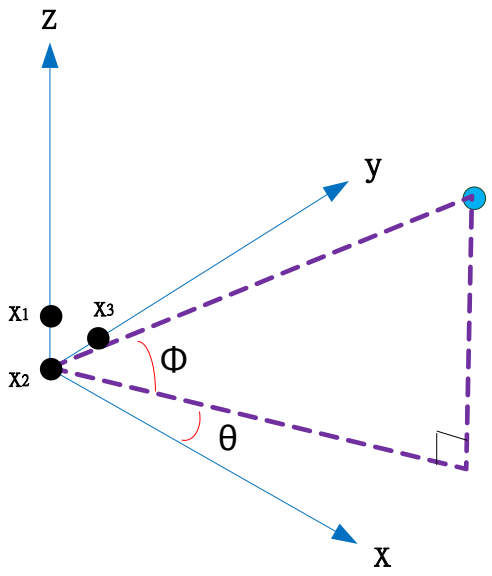


Fig. 8 Spatial Relationship of a Microphone Array and a Source Signal

We assemble three microphones as a microphone array for estimating the azimuth and elevation of the source signal as shown in Fig. 8. Considering the time delay mixture model, we convert the time-domain signals into frequency-domain by Short-Time Fourier Transform (STFT). Let K be the number of source signal $\mathbf{s}$ and N be the number of recorded signal $\mathbf{x}$ with $N > K$. It can be modeled as:

$$\mathbf{x}(f,t) = \mathbf{A}(f,t)\mathbf{s}(f,t) \tag{49}$$

and

$$\mathbf{A}(f) = \left[\mathbf{a}_1(f,\theta_1,\phi_1) \quad \mathbf{a}_2(f,\theta_2,\phi_2) \quad \cdots \quad \mathbf{a}_K(f,\theta_K,\phi_K)\right]$$

$$\mathbf{a}_k(f,\theta_k,\phi_k) = [a_{1k}(f,\theta_k,\phi_k) \quad a_{2k}(f,\theta_k,\phi_k) \quad \cdots \quad a_{Nk}(f,\theta_k,\phi_k)]^T$$

$$a_{nk}(f,\theta_k,\phi_k) = g_{nk} \exp\left\{ j\frac{2\pi f}{c} \vec{r}^{\,T} \cdot \vec{v}(\theta_k,\phi_k) \right\}$$

where $\mathbf{A}(f)$ is the $N \times K$ mixing matrix, whose k-th column vector represents the transfer function of the k-th source signal, which is the so-called steering matrix. The g_{nk} denotes the gain of a_{nk}, $\vec{r} = (x_n, y_n, z_n)^T$ denotes the coordinate vector of the n-th

microphone, and $\vec{v}(\theta_k, \phi_k) = (\cos\theta_k \cos\phi_k, \sin\theta_k \cos\phi_k, \sin\phi_k)$ represents the look direction vector of the n-th microphone as shown in Fig. 8. Then, we can obtain the equation by dividing two elements as shown in following equation:

$$\frac{a_{1k}}{a_{2k}} = \left|\frac{a_{1k}}{a_{2k}}\right| \exp\left\{ j\frac{2\pi f}{c}\left[(y_1 - y_2)\sin\theta_k \cos\phi_k + (z_1 - z_2)\sin\phi_k \right] \right\} \tag{50}$$

$$\frac{a_{1k}}{a_{3k}} = \left|\frac{a_{1k}}{a_{3k}}\right| \exp\left\{ j\frac{2\pi f}{c}\left[(y_1 - y3)\sin\theta_k \cos\phi_k + (z_1 - z_3)\sin\phi_k \right] \right\} \tag{51}$$

The elevation angle ϕ and the azimuth angle θ can be derived by the following equations:

$$\sin\theta_k \cos\phi_k = \frac{(y_1 - y_3)A - (y_1 - y_2)B}{(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)} \tag{52}$$

$$\sin\phi_k = \frac{(x_1 - x_2)B - (x_1 - x_3)A}{(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)} \tag{53}$$

and

$$A = \frac{angle(a_{1k} / a_{2k})}{2\pi fc^{-1}}$$

$$B = \frac{angle(a_{1k} / a_{3k})}{2\pi fc^{-1}}$$

Then, we extract the angles ϕ and θ:

$$\theta_k = \cos^{-1}\left\{ \frac{(y_1 - y_3)A - (y_1 - y_2)B}{\left[(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)\right]\cos\phi_k} \right\} \tag{54}$$

$$\phi_k = \sin^{-1}\left\{ \frac{(x_1 - x_2)B - (x_1 - x_3)A}{(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)} \right\} \tag{55}$$

3.3 Acoustic Signal Synthesis

3.3.1 Virtual Listening Point Audio Synthesis

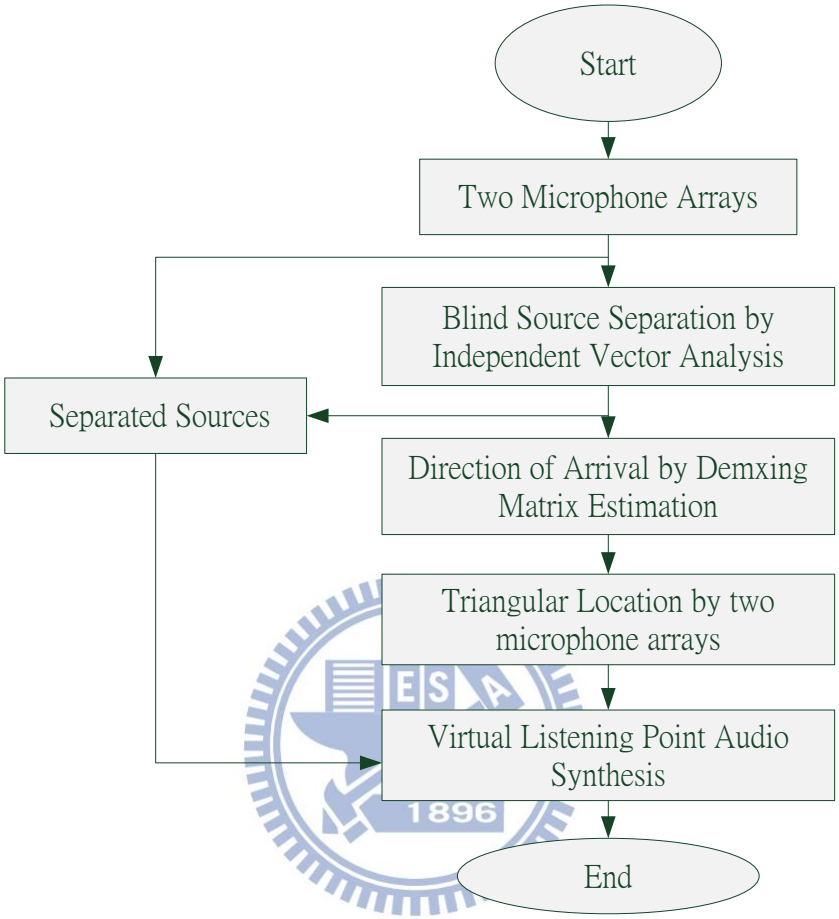


Fig. 9 Flow Diagram of 3D Acoustic Signal Synthesis

In our study, the main purpose is to synthesize a virtual listening point audio from the mixture signals. We record the signals using a microphone array in an anechoic chamber. Fig. 9 shows the acoustic signal synthesis flowchart. We are able to construct the acoustic signal at the desired virtual listening position, which is the so-called *Virtual Listening Point Audio Synthesis*. We assume that there are two source signals and two microphone arrays in our experiment. Each array contains three or seven microphones. The task includes three major steps. First, we adopt [8] to separate the mixture signals recorded by the microphone array. Second, by employing the IVA method, we can

obtain the demixing matrix $\mathbf{W}(f)$. Thus, we derive the steering matrix $\mathbf{A}(f) = \mathbf{W}(f)^{+}$, where $\mathbf{W}(f)^{+}$ denotes the pseudo-inverse of $\mathbf{W}(f)$. Then, we use the steering matrix $\mathbf{A}(f)$ and [28] to estimate the DOA of two source signals. Third, we select an arbitrary point to synthesis the virtual audio in the space. Fig. 10 shows three parts in main system.

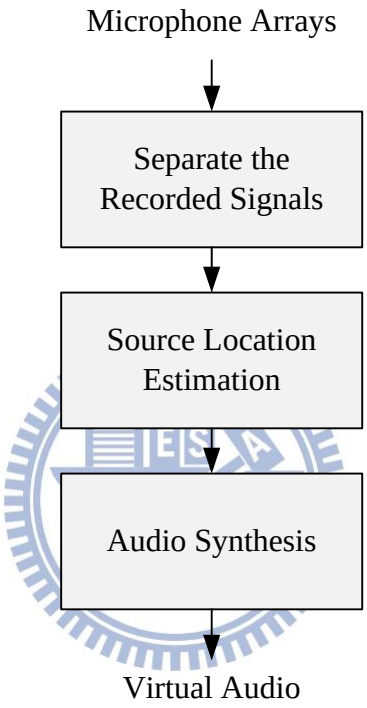


Fig. 10 Flow Diagram of Overall System

3.3.2 Recorded Signals Separation

In this section, the main purpose is to separate recorded signals by utilizing [7]. There are many different kinds of blind source separation methods. However, it is quite difficult to completely separate the source signals in general cases since the information about the source signals and the mixing system is not known. Fig. 11 shows our chosen BSS system flow diagram. First, we use the PCA method, which includes centering and whitening to pre-process the mixture signals. The purpose of the centering is to remove

the mean of the mixture signals. The whitening process decorrelates the mixture signals, and it converts the variance of mixture signals to be unitary. After the mixture signals are decorrelated, the separated signals are closer to be independent components in the ICA scheme. Then, we can use the Newton method to optimize the contrast function. When the iteration converges, we solve the scaling problem by using the method in [10]. Finally, we obtain the separated signals by multiplying the mixture signals by the demixing matrix, and then we convert the frequency–domain signals back to the time-domain by IDFT. Here, the separated signals $y(t)$ are recorded for the purpose of synthesizing the virtual audio.

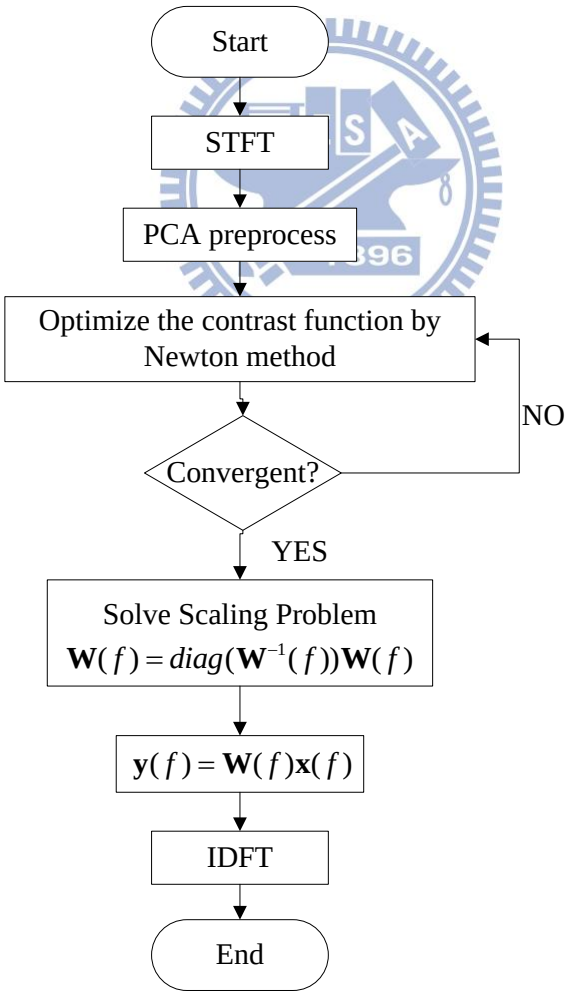


Fig. 11 Flow Diagram of ICA scheme

3.3.3 Source Localization

In a real acoustic environment, the source direction contain azimuth angle θ and elevation angle ϕ. Fig. 12 gives the DOA flow diagram of overall procedure. We use the pseudo-inverse to obtain the demixing matrix $\mathbf{A}(f)=\mathbf{H}(f)^{+}$, which is also called steering matrix. The i-th column of the steering matrix represents the transfer function of the i-th source. Then, the azimuth angle θ and the elevation angle ϕ are solved by dividing two elements within the i-th column as dicussed in section 3.2 [28].

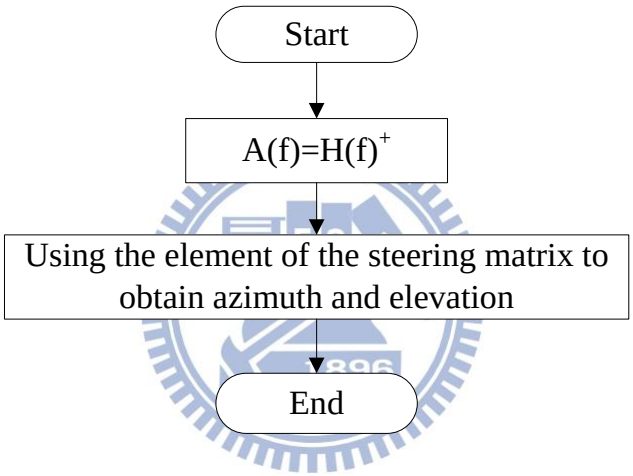


Fig. 12 The DOA Flow Diagram

For estimating the distance between the source and the microphone array, it is necessary to use two microphone arrays. Here, we use trigonometry to estimate the distance of the two source signals. When we obtain the steering matrices from two different microphone arrays, the column vectors correspond to the source signals but their orders may not match between two microphone arrays. Assume there exists steering vector matrix $\mathbf{A}_i=\left[\mathbf{a}_{ij}^{k}\right]$, where $\mathbf{A}_i$ denotes the steering matrix which is obtained from i-th microphone array; $\mathbf{a}_{ij}^{k}$ denotes the j-th steering vector, which represents the transfer function of the k-th source signal from the i-th microphone

array. If $\mathbf{A}_1 = \left[\mathbf{a}_{11}^1 \quad \mathbf{a}_{12}^2\right]$ and $\mathbf{A}_2 = \left[\mathbf{a}_{21}^2 \quad \mathbf{a}_{22}^1\right]$ represent the first and the second microphone arrays, respectively, we notice that the first column vector $\mathbf{a}_{11}^1$ and $\mathbf{a}_{21}^2$ represent the transfer function associated with different sources. However, $\mathbf{a}_{11}^1$ and $\mathbf{a}_{22}^1$ represent the transfer function with the same source. Let $\mathbf{Y}_1 = \left[\mathbf{y}_{11}{}^1 \quad \mathbf{y}_{12}{}^2\right]$ and $\mathbf{Y}_1 = \left[\mathbf{y}_{21}{}^2 \quad \mathbf{y}_{22}{}^1\right]$ represent the separated signals from two microphone arrays. We compute the correlation coefficient of two signals to be matched. Considering the time delay between two microphone arrays, the maximum time delay is caused by the distance between two microphone arrays. Therefore, we can select the maximum correlation coefficient to match signals.

$$\arg\max_i corrcoef\left(\mathbf{y}_{11}{}^p(1:L-n), \mathbf{y}_{2i}{}^g(n+1:L)\right), i = 1,2,...,K \qquad n = 0,1,...,T$$

$$T = \frac{r}{c} \times Fs$$

where K denotes the number of the steering vectors, L denotes the signal length, i and n denote the index of the sample point, r denotes the distance between two microphone arrays, c denotes the sound speed, Fs denotes the sampling rate, and T is the upper bound of the delay.

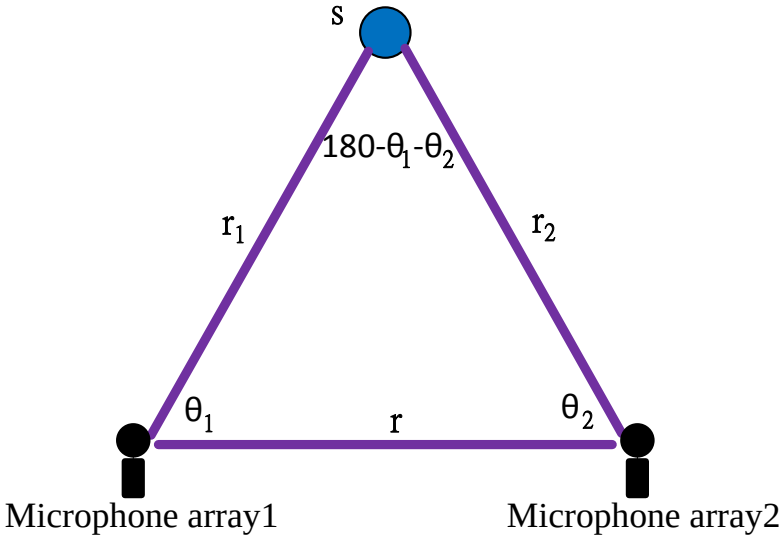


Fig. 13 Schematic Diagram of "Law of sines"

The distance r and the azimuth angle θ_i are shown in Fig. 13, and it can be shown that:

$$\frac{r}{\sin(180-\theta_1-\theta_2)} = \frac{r_1}{\sin\theta_2} = \frac{r_2}{\sin\theta_1} \tag{56}$$

When we obtain the distance r between two microphones and θ_1 and θ_2, we can derive r_1 and r_2, respectively. Finally, we can estimate the distance from the source to the microphone array by following equation:

$$R_i = r_i / \cos\phi_i \tag{57}$$

where ϕ_i denotes the elevation angle.

Because of the restrictions on the instrument, we do not use trigonometry to estimate the distance of the two source signals. Here, the distances, r_1 and r_2, are represented by true outcomes.

3.3.4 Audio Signal Synthesis

In our study, we adopt the software developed by the NASA Ames Research Center to synthesize the virtual listening point. The software implements the spatial 3D-sound processing procedure. It also supports placing the source signals in the space arbitrarily. We perform BSS to separate signals from the recorded mixture signals. Then, we take separated signals as inputs. Fig. 14 shows the arrangement of separated signals and the microphone array on the X-Y plane. S_1, S_2 and P_o respectively represent the first source, the second source and the position of the original microphone array. θ_1, θ_2 respectively represent the azimuth angles of the first source and the second source. d_1, d_2 respectively represent the distances of the first source and the second source from the microphone. Because we do not use trigonometry to estimate the distance, the distances here are true outcomes. We then synthesize the audio signals at P_1, P_2, P_3 and etc. Thus, we obtain the virtual listening point audio signals.

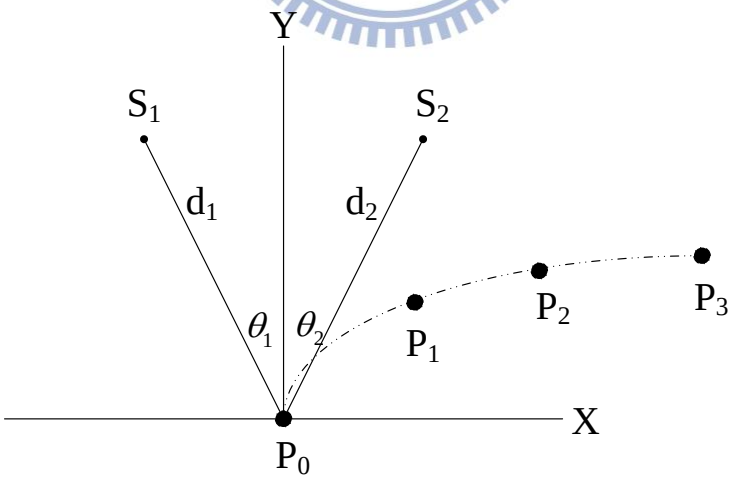


Fig. 14 Schematic Diagram of Audio Synthesis

Chapter 4 Experimental Results: Part A

4.1 Real and Virtual Acoustic Environment

4.1.1 Anechoic Chamber

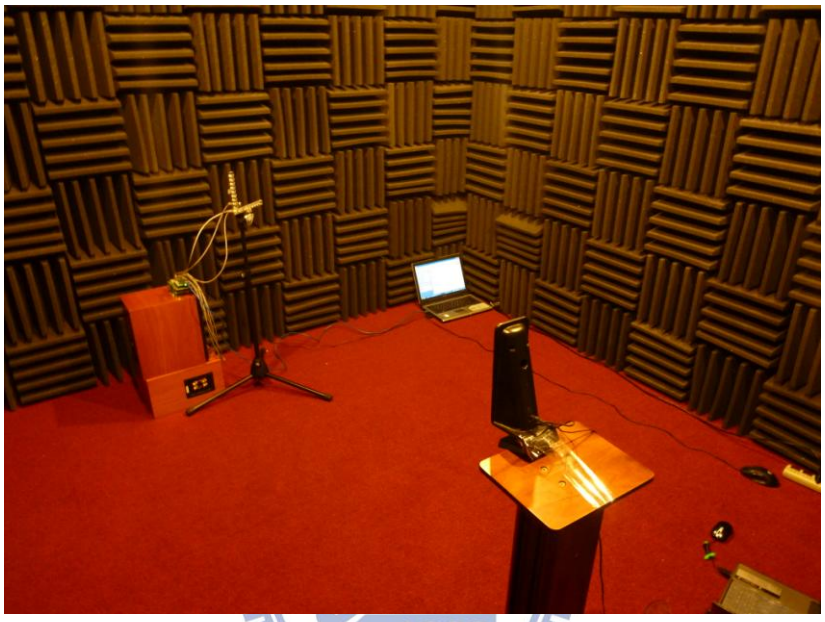


Fig. 15 Physical acoustic room in an anechoic chamber

An anechoic chamber is a room with special walls designed to prevent the sound reflection. It can also insulate the outside interference or noise. An anechoic chamber is commonly used to conduct experiments for simulating “free field” conditions or noise reduction. The material covering the walls of the anechoic chamber is wedge-shaped panels. The panel can dissipate the source energy before reflecting it away.

In our experiments, the microphones that we use are developed by Ario Company. These microphones have a diaphragm and backplate structure. It means that the voltage is changed by the distance between two plates, which is called Condenser Microphone or Capacitor Microphone. Because there is no coils and magnet, this kind of microphones have high sensitivity.

4.1.2 NASA Sound Lab (SLAB) Software

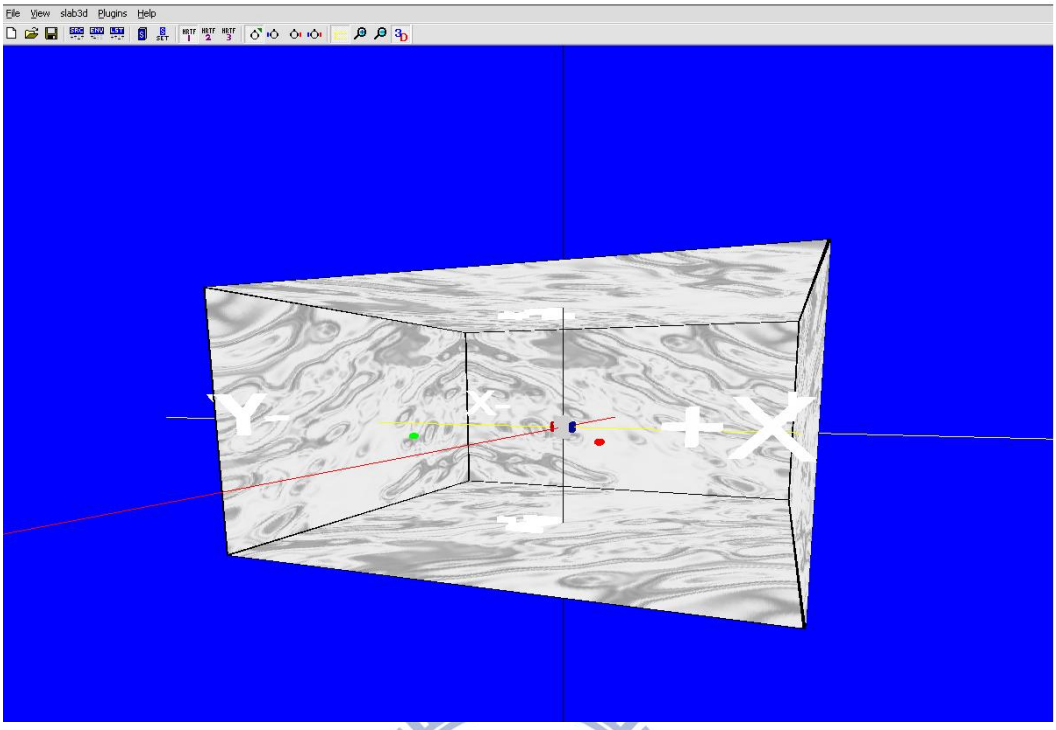


Fig. 16 Snapshot of the 3D virtual acoustic room in SLAB

SLAB is a virtual acoustic environment rendering system developed by the NASA Ames Research Center. The software simulates a virtual acoustic environment. It helps us to evaluate our algorithms in spatial hearing and psychoacoustics field. In other words, it is not necessary to construct a physical environment. There are three major categories of parameters: the source, the environment and the listener. The source parameters include the source locations and the source types. The environment parameters include the sound speed, the room size and the surface reflections, etc. The listener parameters include the listener location, the Head Related Transfer Function (HRTF) model and the interaural time difference (ITD), etc. There are some other specifications given in Table. 1, 2 and 3 [29].

Table. 1 Scenario Specifications [29]

Scenario	
Room	**Rectangular Room**
Reflections	**6 First-order Reflections**
Direct Path FIR Taps	**128**
Reflection FIR Taps	**32**
Material Filter	**First-order IIR Filter**

Table. 2 System Dynamics Specifications [29]

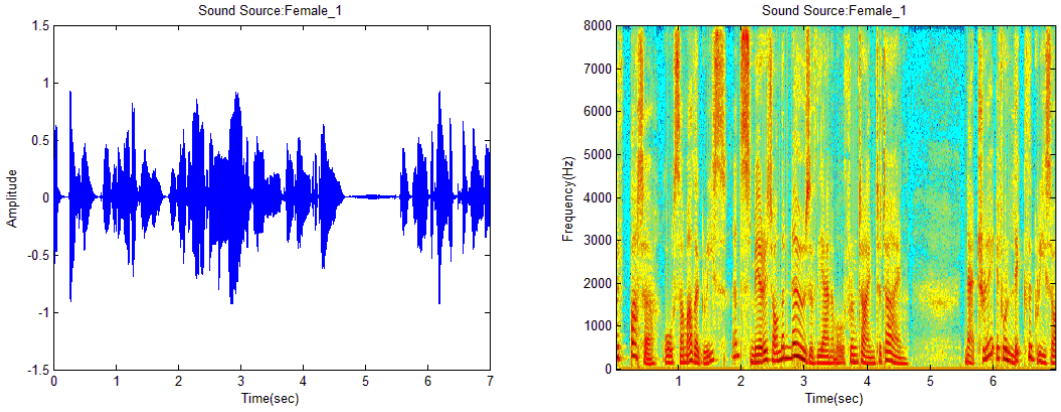
System Dynamics	
Sampling Rate	**44.1 kHz**
Update Rate	**120 Hz**
Internal Latency	**24 msec**
FIR Update	**Every 64 Samples (1.45 msec)**
Delay Line Update	**Every Sample (22.7 μsec)**

Table. 3 Numerical Precision Specifications [29]

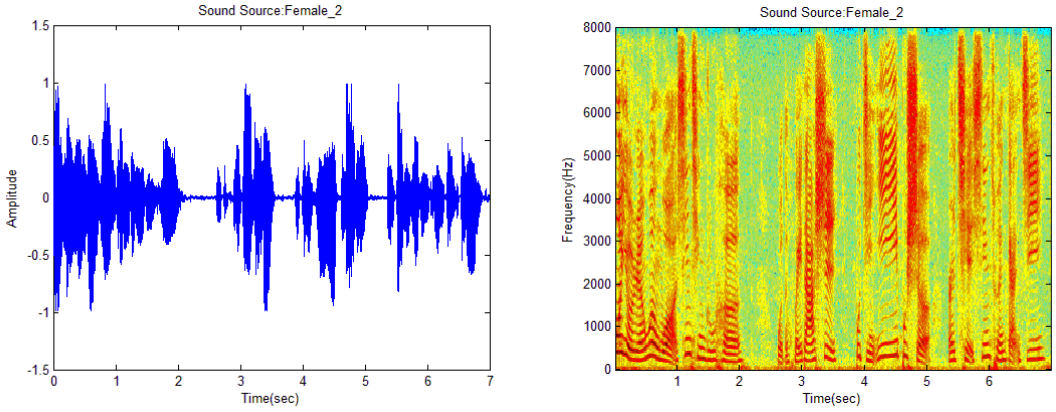
Numerical Precision	
Sound Input / Output	**16-bit Integer**
Scenario	**Double-precision Floating-point**
Signal Processing	**Single-precision Floating-point**

4.1.3 Experiment Setting

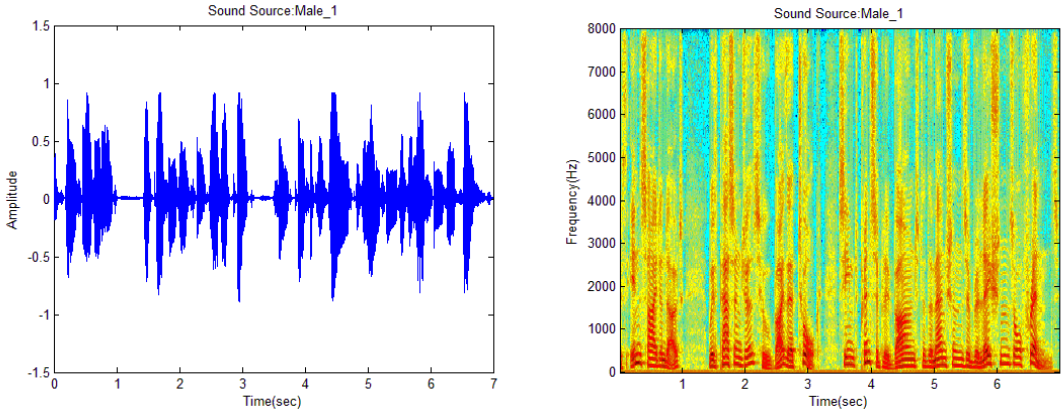
In our experiments, the testing source signals consist of four speeches: Female_1 (English-speaker), Female_2 (Chinese-speaker), Male_1 (English-speaker), Male_2 (Chinese-speaker). The power of the source signals is normalized. The waveforms and spectrograms of these signals are shown in Fig. 17 $(a) \sim (d)$.



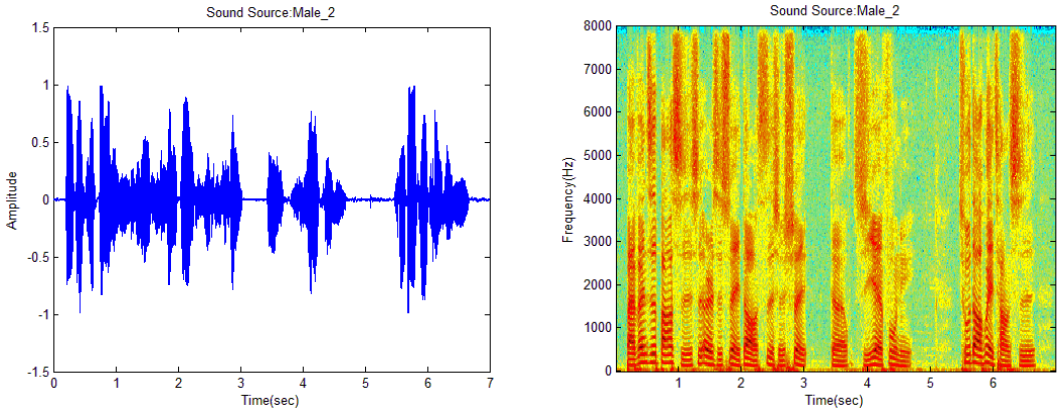
(a) Female_1



(b) Female_2



(*c*) Male_1



(d) Male_2

Fig. 17 The waveforms in time domain and the spectrograms in time-frequency domain

In theory, in the case of more than two sources, the estimation procedure would be similar to that in the two-source case. In our experiments, we only consider the two-source case without reverberant effect. As mentioned in Chapter 3, the microphone spacing d should be less than a half of wavelength to prevent spatial aliasing, that is, $d \le (1/2)\lambda$. We assume that the sound velocity $c = 345$ m/sec and the maximum frequency of source signal is 4 KHz (Sampling Rate: 8KHz). The microphone spacing should be less than 4.25 cm. Therefore, we consider $d = 3$ cm case.

In our experiments, we design twelve groups of the source signals, which are

shown in Table. 4. We classify the experiments into CASE-A, CASE-B.1 and CASE-B.2. The CASE-A denotes the speech source recorded from the microphone arrays in an anechoic chamber. In addition, each case includes three and seven microphone tests. They are labeled as the First-EXP and the Second-EXP. The CASE-B.1 denotes that we use the SLAB developed by the NASA Ames Research Center to simulate the recorded mixture signals in an ideal acoustic environment. The CASE-B.2 denotes that we add the Additive White Gaussian Noise (AWGN) to CASE-B.1. The noise is estimated with an approximation value in the CASE-A. Thus, we have two noise estimates from the First-EXP and the Second-EXP. We use the CASE-B.2 to simulate the situation in CASE-A.

Table. 4 Twelve Groups

Groups	**Source1**	**Source2**
1	Female_1	Female_2
2	Female_1	Male_1
3	Female_1	Male_2
4	Female_2	Female_1
5	Female_2	Male_1
6	Female_2	Male_2
7	Male_1	Female_1
8	Male_1	Female_2
9	Male_1	Male_2
10	Male_2	Female_1
11	Male_2	Female_2
12	Male_2	Male_1

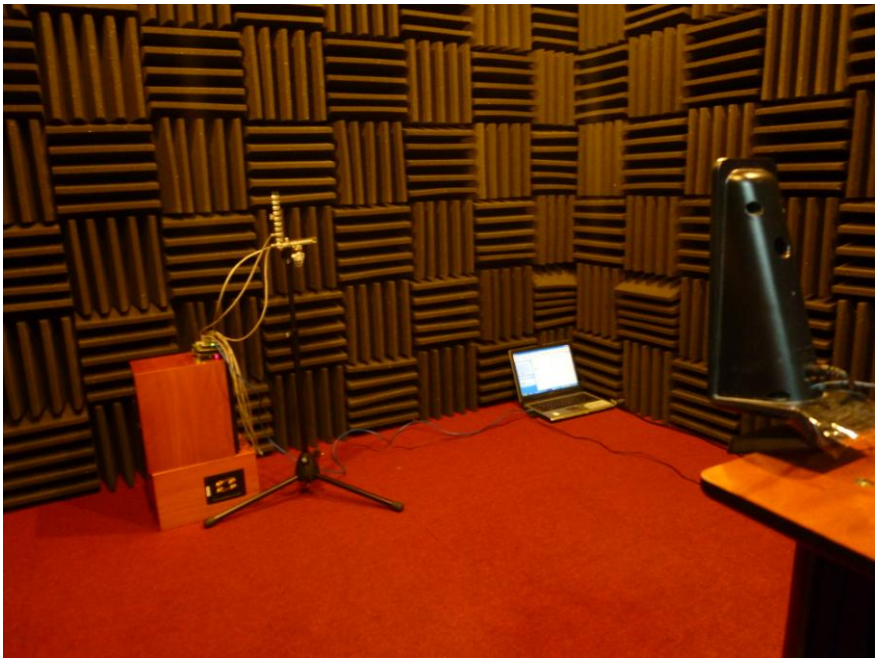
For convenience, we move the speech source instead of moving microphone array in recording. In addition, we set up a source and a sensor at the same horizontal plane in our experiments as shown in Fig. 18. The speech source varies from -30^o to 30^o with a 15^o step. It represents that the angles include $\pm 30^o$, $\pm 15^o$, 0^o with

different directions. The set-up of the microphone array and the sources is shown in Fig. 19. Finally, we adopt the Adobe Audition 3.0 software to combine two speeches into the mixture signals.

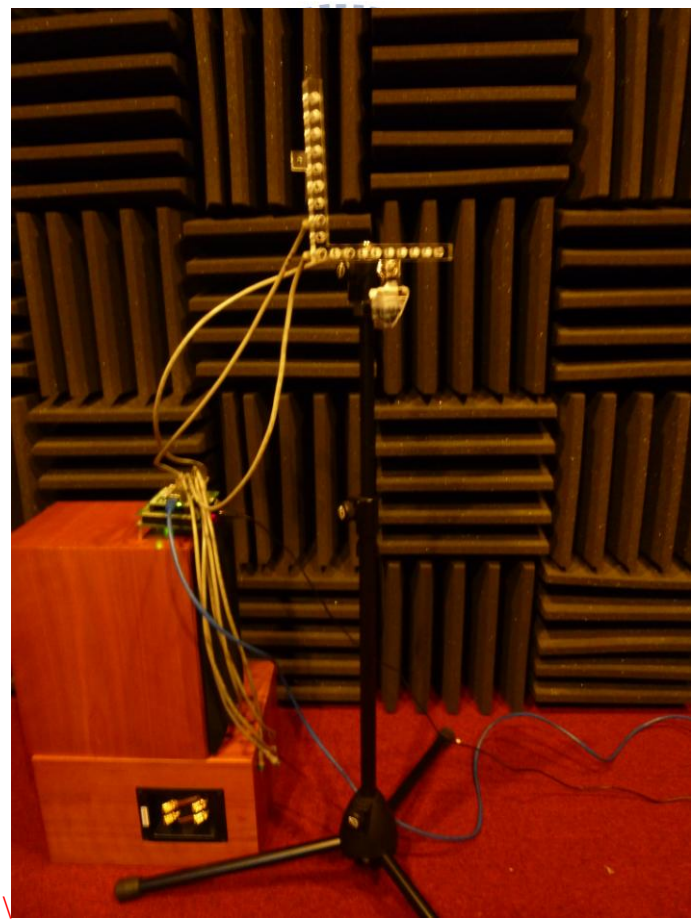
For CASE-B.2, we estimate an approximate Signal to Noise Ratio (SNR) by the following equation:

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{observe} - P_{noise}}{P_{noise}}\right) \tag{58}$$

where $P_{observe}$ denotes the power of the observation recorded from the microphone array, P_{noise} denotes the power of the noise. We choose the silent speech in about three-second period do estimate the noise. Furthermore, we calculate an average SNR from twenty segments to estimate SNR. The SNR with different sensors is shown in Fig. 20. The x-axis represents the order of the sensors, which is shown in Fig. 22. We notice that the average SNR of the First-EXP is higher than that of the Second-EXP. The seventh sensor x_7 has lower SNR in the Second-EXP. The test in CASE-B.2 demonstrates a mismatch between microphones. Fig. 21 shows the difference between CASE-A and CASE-B.2. Fig. 21(a) denotes the time domain signal of a sensor in the First-EXP of CASE-A. Fig. 21(b) denotes the simulation of the time domain signal of a sensor in the First-EXP of CASE-B.2. Fig. 21(c) denotes the time domain signal recorded from SLAB. Furthermore, in a real situation, we know that the channel impulse response has an impact on our experimental results. However, it is difficult to simulate a dynamic system. Thus, the simulation is only done with the additive AWGN.



(a) Placement of the Speech and the Sensor



(b) The Microphone Array

Fig. 18 Anechoic Chamber Scenario

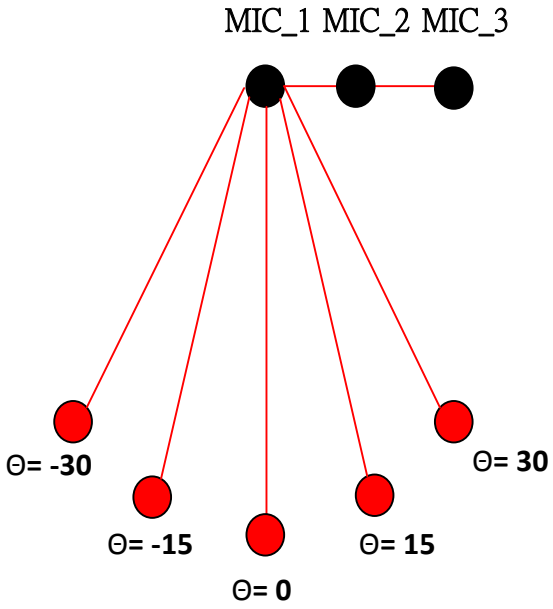


Fig. 19 The Location of the Source and the Microphone Array

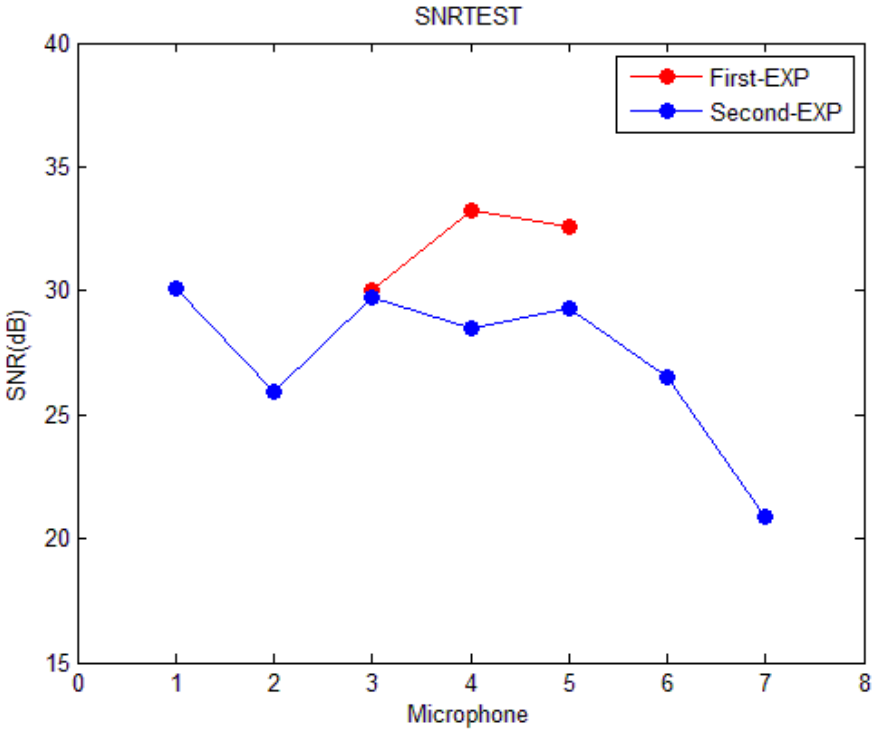
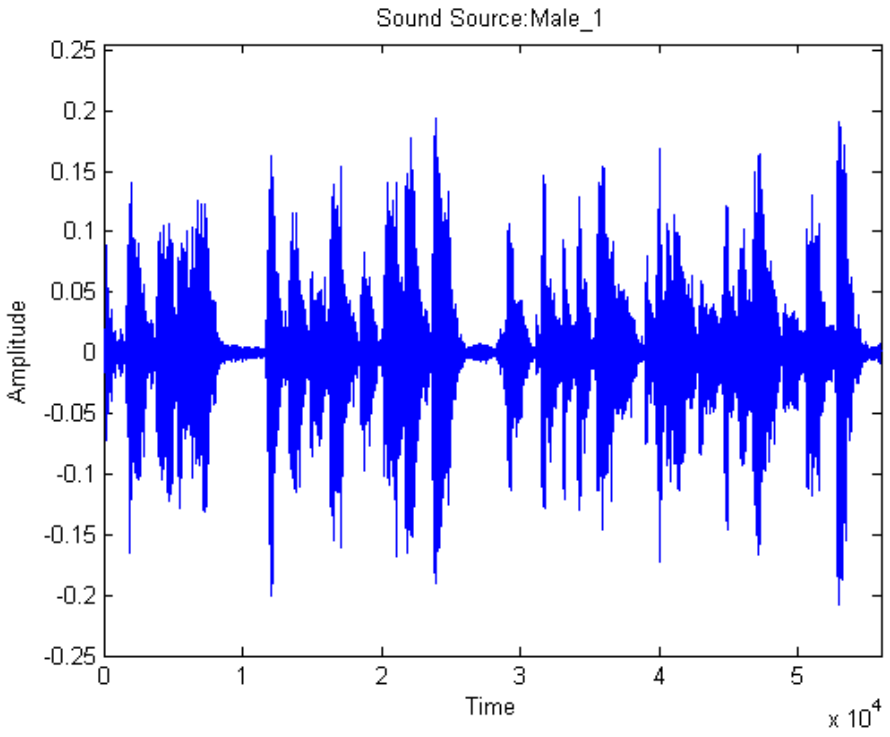
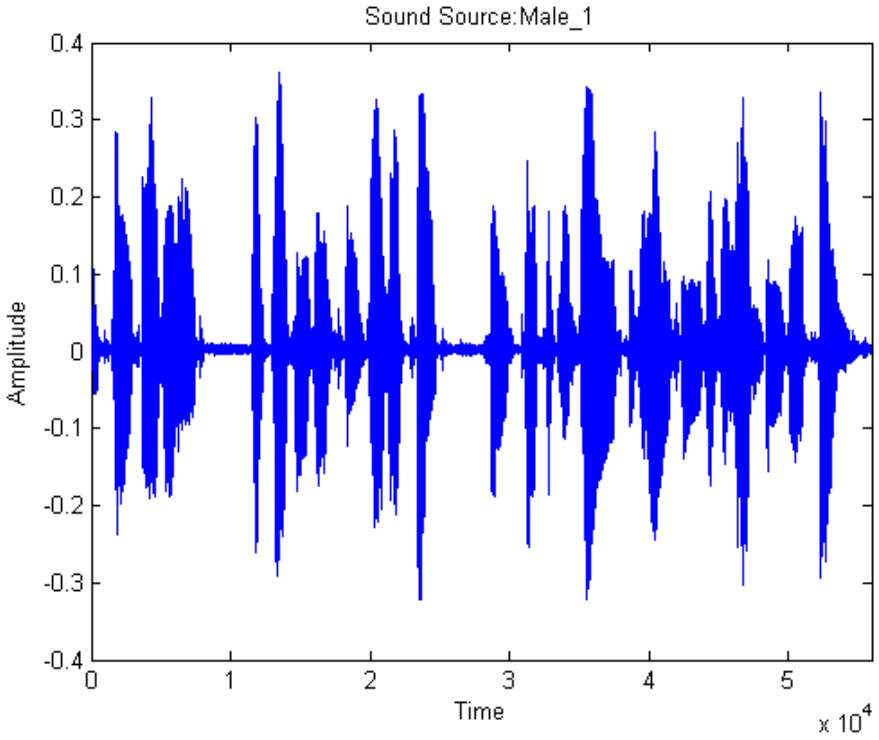


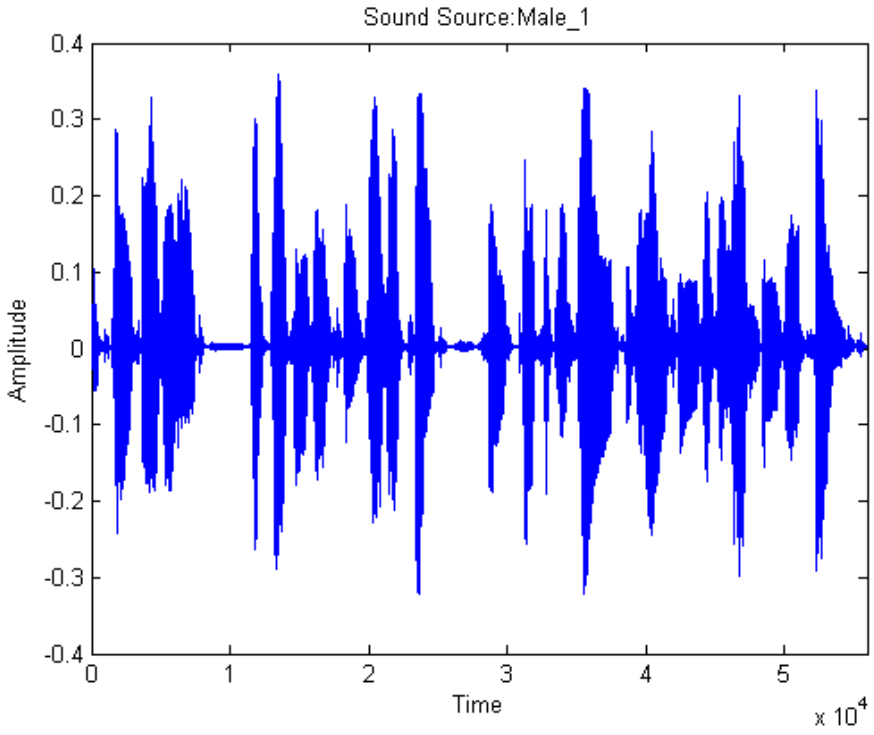
Fig. 20 First-EXP and Second-EXP SNR Test in CASE-A



(a) Male_1 in the First-EXP of CASE-A



(b) Male_1 in the First-EXP of CASE-B.2



(c) Male_1 in the First-EXP of CASE-B.1

Fig. 21 Compare the noise between CASE-A and CASE-B.2

4.2 Blind Source Separation Data Analysis

4.2.1 The Effect of Microphone Number

In this section, we focus on the effect of microphone numbers in the BSS algorithm. Thus, we select three sets of microphones: $x_3 \sim x_5$ (3 microphones), $x_2 \sim x_6$ (5 microphones) and $x_1 \sim x_7$ (7 microphones), which are shown in Fig. 22. There are many popular metrics of evaluating the BBS quality, and one way is to measure the Signal to Interference Ratio (SIR) as described in Chapter 2. The definition of SIR is described below:

$$SIR = \frac{10}{K}\sum_{i=1}^{K}\log_{10}\frac{\left\langle \max_{n} |\, y_{n,s_i}(t)\,|^2 \right\rangle}{\left\langle \sum_{j\neq i} |\, y_{n,s_j}(t)\,|^2 \right\rangle}, n = 1, 2, \ldots, N$$

Here, we set the input data duration four seconds (32000 sampling points). The window size is 512 sampling points and the source distance is 1.5M.

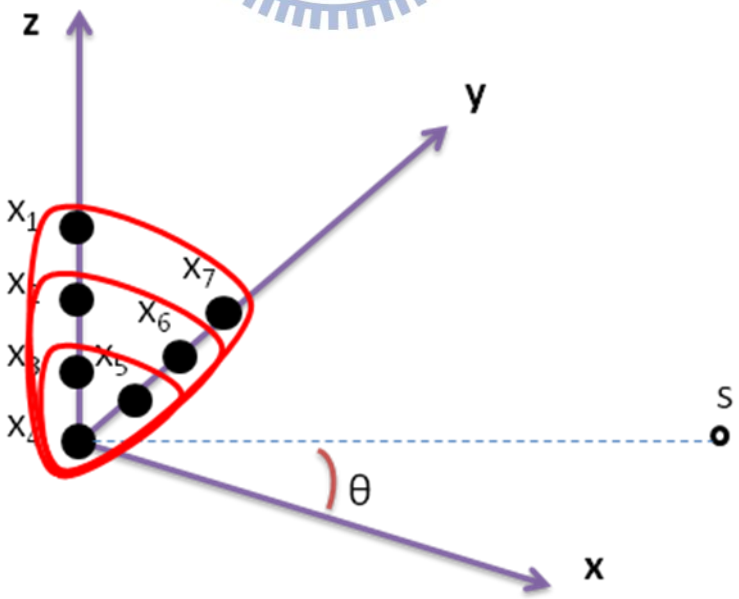
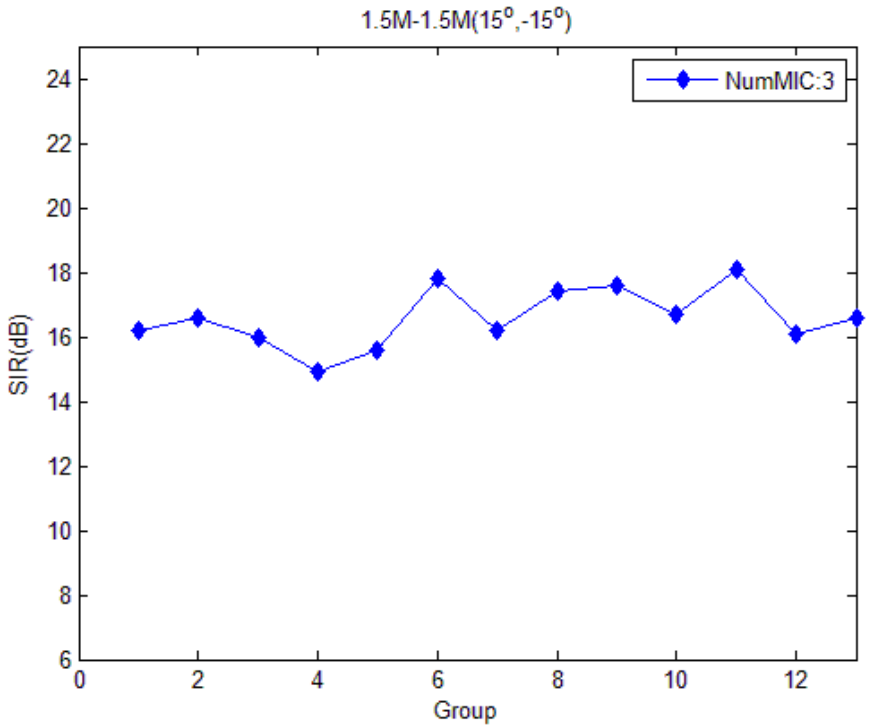
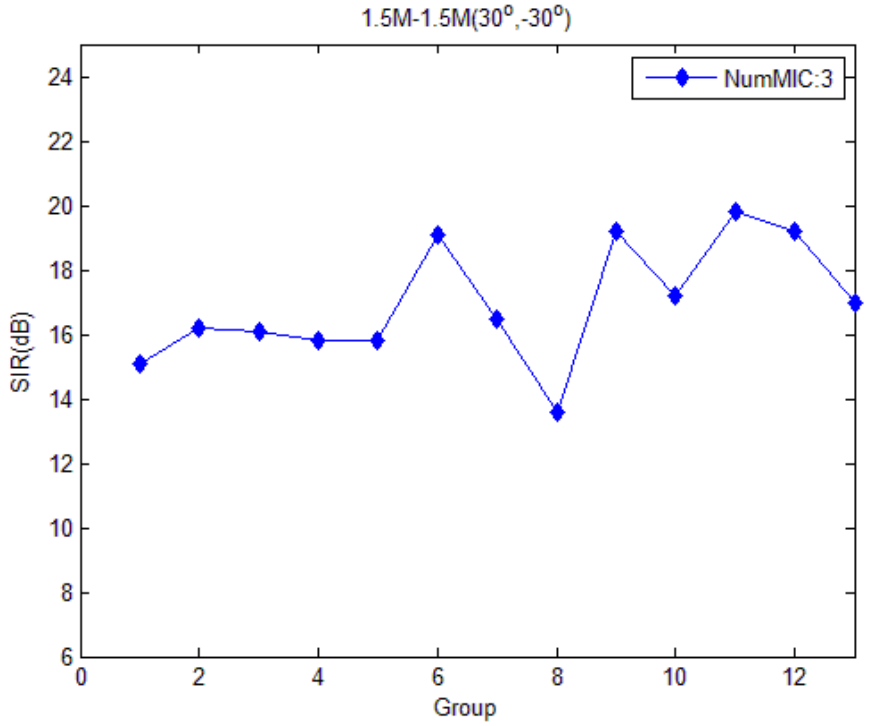


Fig. 22 The Placement of a Microphone Array

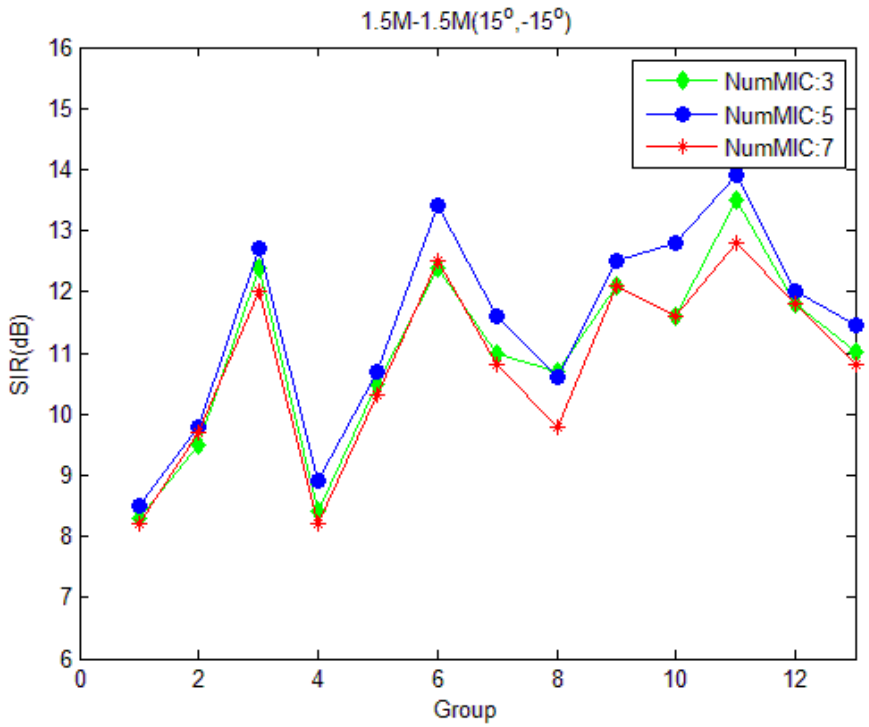


(a) SIR with angle 15^{o} and -15^{o}

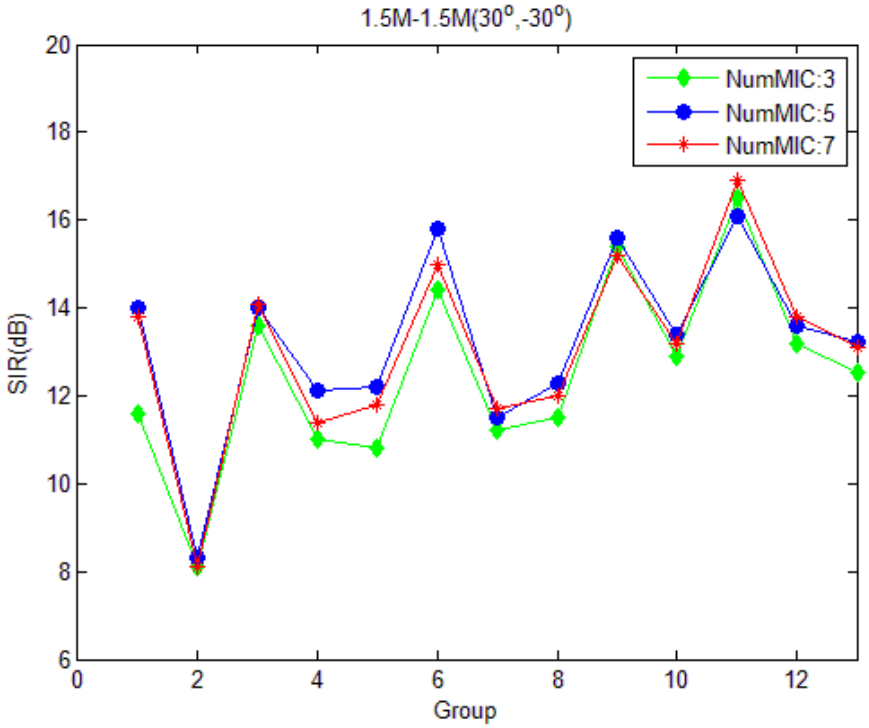


(b) SIR with angle 30^{o} and -30^{o}

Fig. 23 Microphone number test in the First-EXP of CASE-A (Real)



(a) SIR with angle 15^{o} and -15^{o}



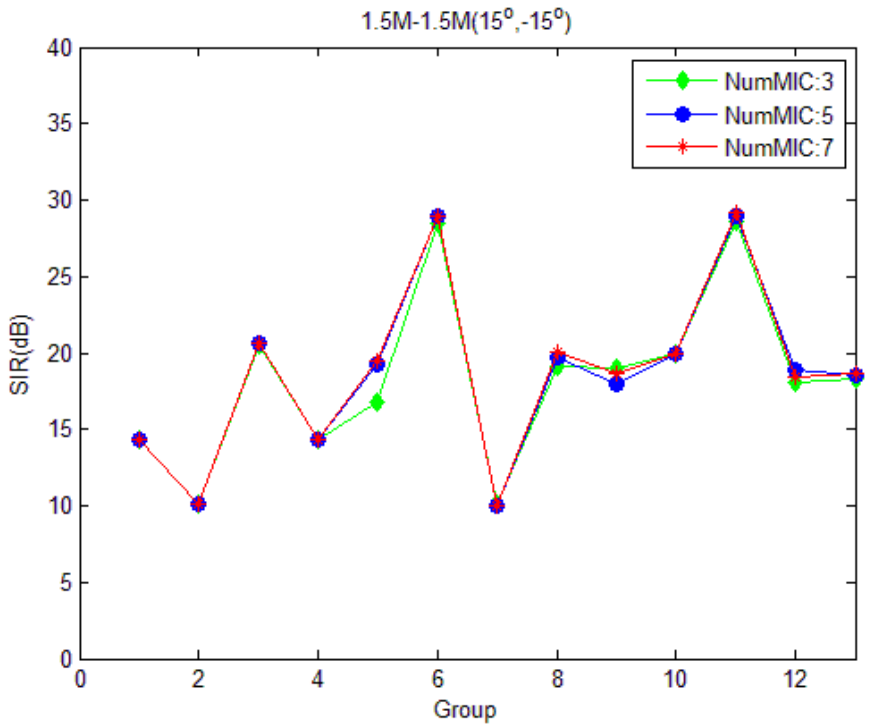
(b) SIR with angle 30^{o} and -30^{o}

Fig. 24 Microphone number test in the Second-EXP of CASE-A (Real)

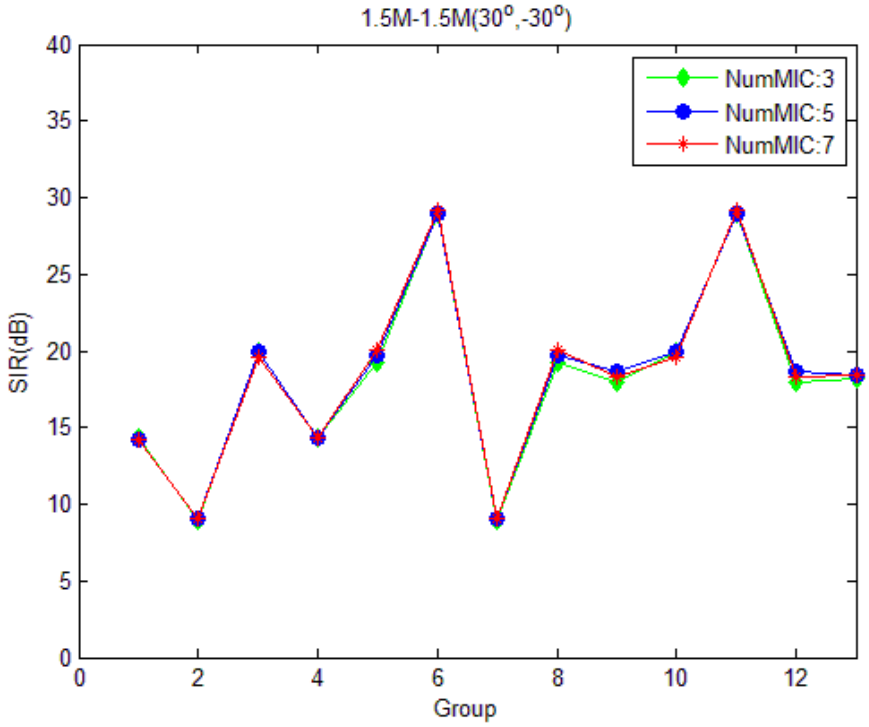
In CASE-A, we have the First-EXP and the Second-EXP experiments as shown in Fig. 23 and Fig. 24. Fig. 23(a) shows the SIR corresponding to different sources from angle 15^o and -15^o. Fig. 23(b) shows the SIR corresponding to different sources from angle 30^o and -30^o. The x-axis represents the test groups as shown in Table. 4. Group 13 represents the average SIR of all 12 groups. In a similar way, there are different sources from different angles in Fig. 24(a)~(b). In Fig. 23(a)~(b), we notice that the results provide good performance. The average SIR is approximately 16dB. We can observe that the overall results in Fig. 23 are better than that in Fig. 24. On the other hand, there are some other trends in Fig. 24. The results get better performance by adding more sensors as shown in Fig. 24(a)~(b). For example, using five and seven sensors outperforms than using three sensors. The average SIR is about 11~13dB. However, we also see that the results with five sensors show the better performance than that with seven sensors. In fact, the situation is reasonable. In the Second-EXP, the average SNR with five sensors provides a higher value than that with seven sensors. On the other hand, the average SNR of the First-EXP also provides the higher value than that of the Second-EXP as shown in Fig. 20. Therefore, we know that the sensors with higher SNR determine the BSS quality.

In CASE-B.1, it shows the test in a SLAB-based simulation. Fig. 25(a)~(b) show that the SIR values are almost the same for the tests with different sensors because the SLAB environment in CASE-B.1 is noise free and distortion free. In CASE-B.2, we see that more sensors have a higher SIR in this simulation as shown in Fig. 26 and Fig. 27.

Based on these figures, we can observe some trends from these data. First of all, BSS provides good quality when the average SNR of sensors is high. In other words, the sensors with high SNR determine the BSS quality. Second, the performance of BSS algorithm is higher by adding more sensors.

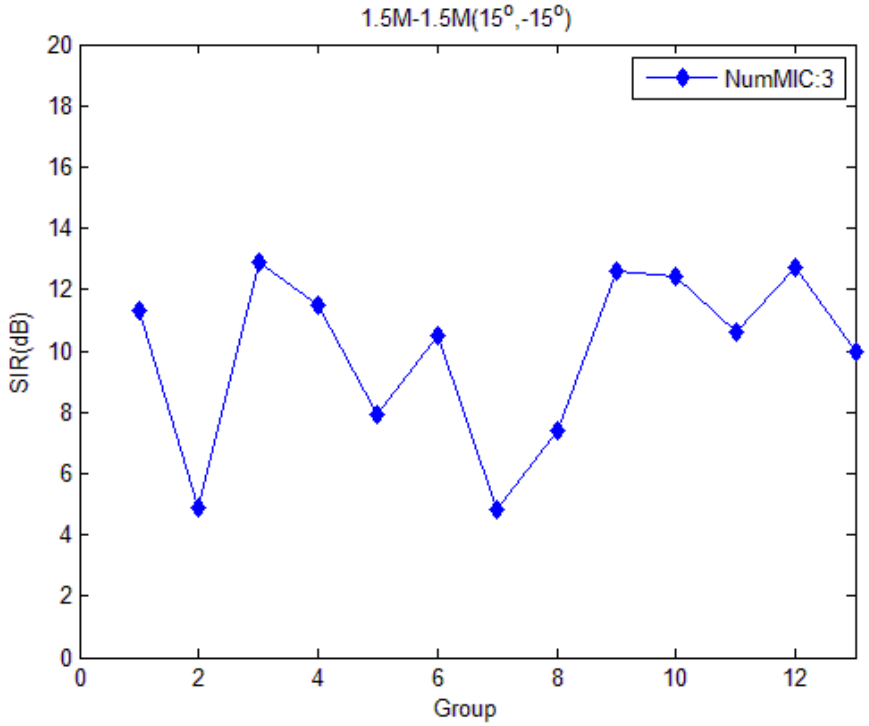


(a) SIR with angle 15^{o} and -15^{o}

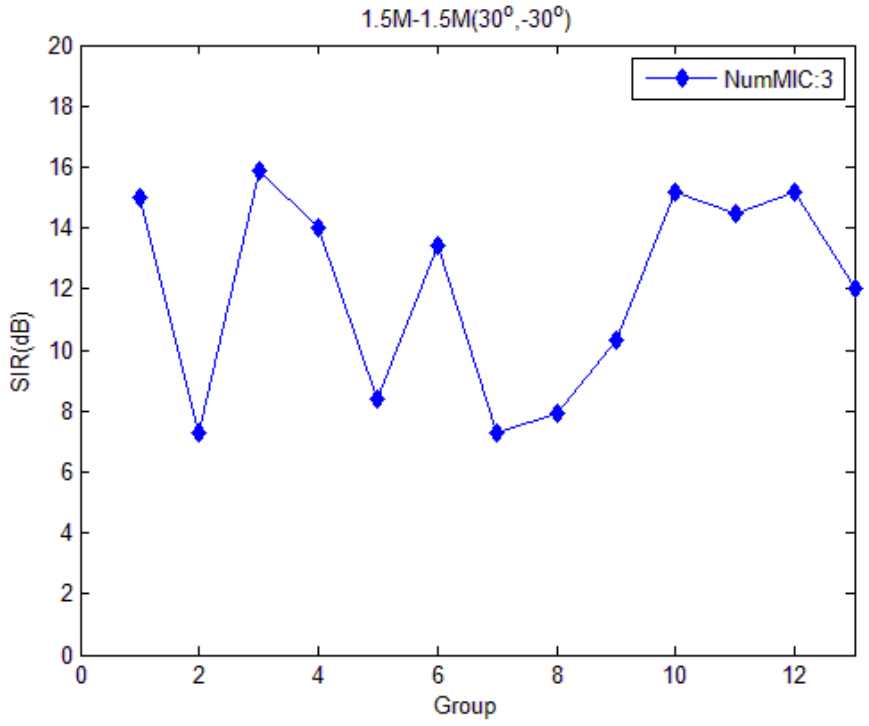


(b) SIR with angle 30^{o} and -30^{o}

Fig. 25 Microphone number test in CASE-B.1 (SLAB)

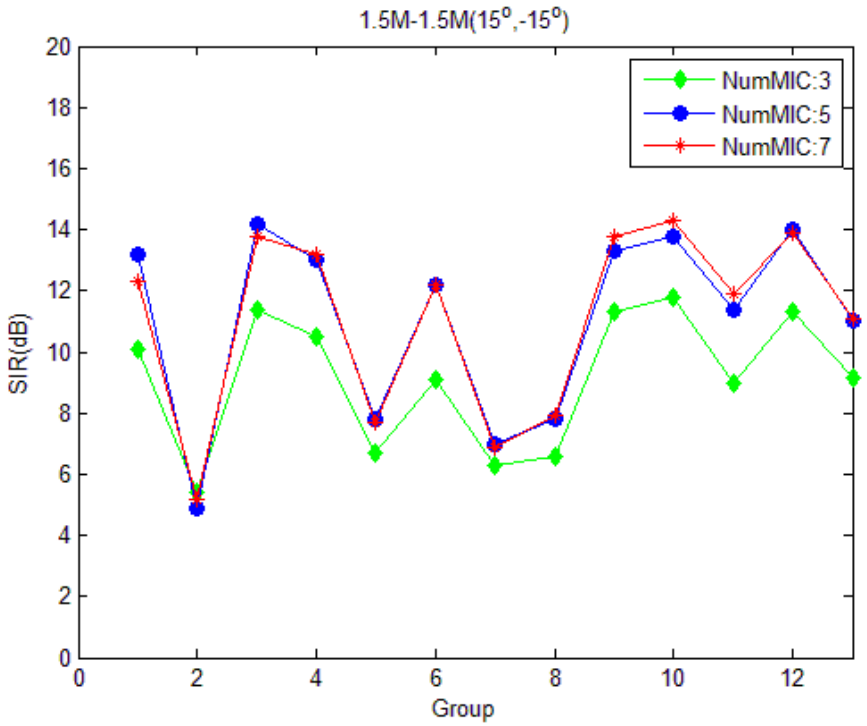


(a) SIR with angle 15^{o} and -15^{o}

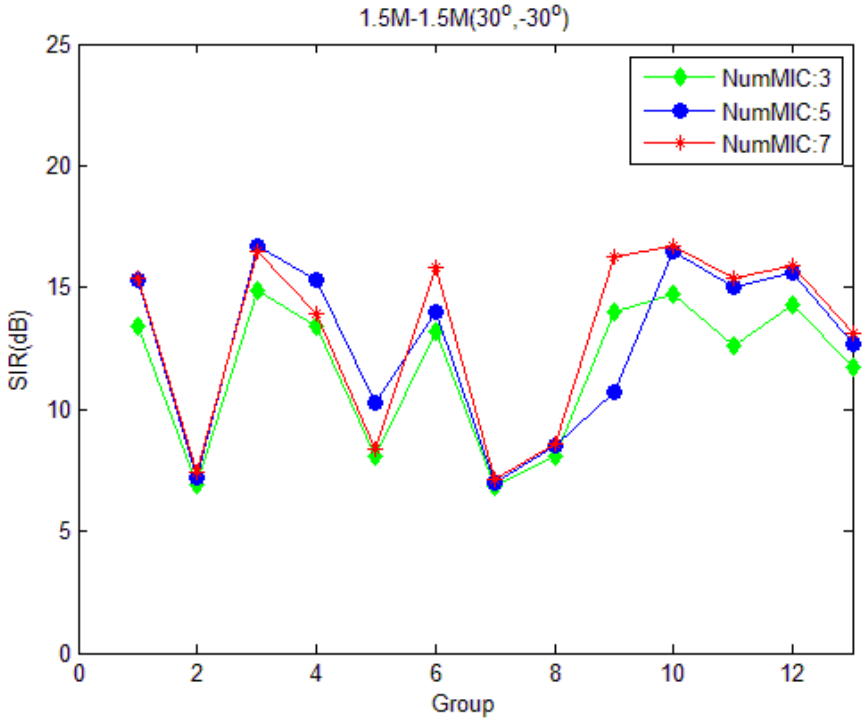


(b) SIR with angle 30^{o} and -30^{o}

Fig. 26 Microphone number test in the First-EXP of CASE-B.2 (AWGN)



(a) SIR with angle 15^{o} and -15^{o}

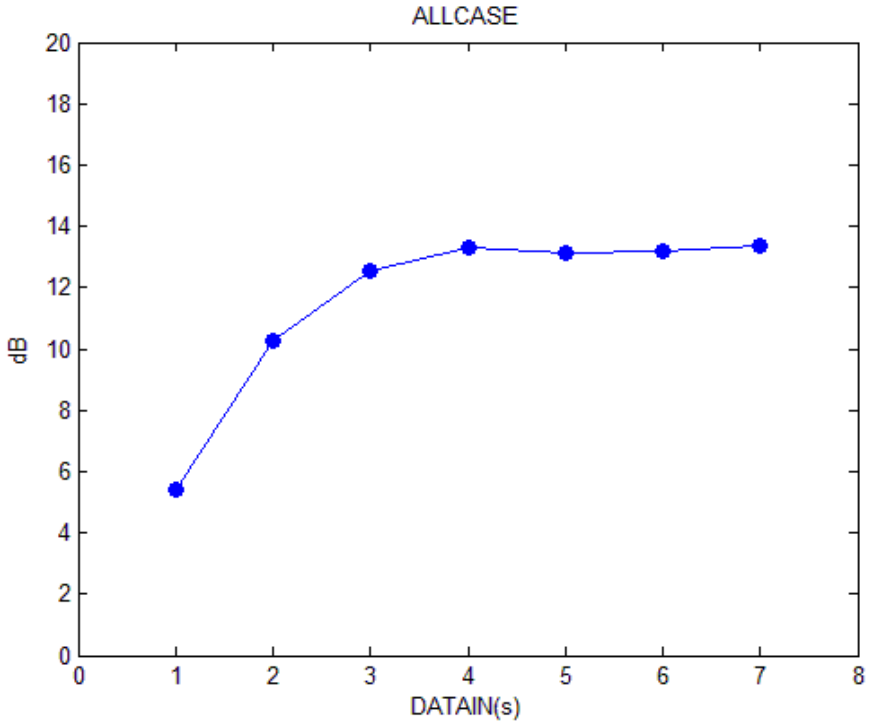


(b) SIR with angle 30^{o} and -30^{o}

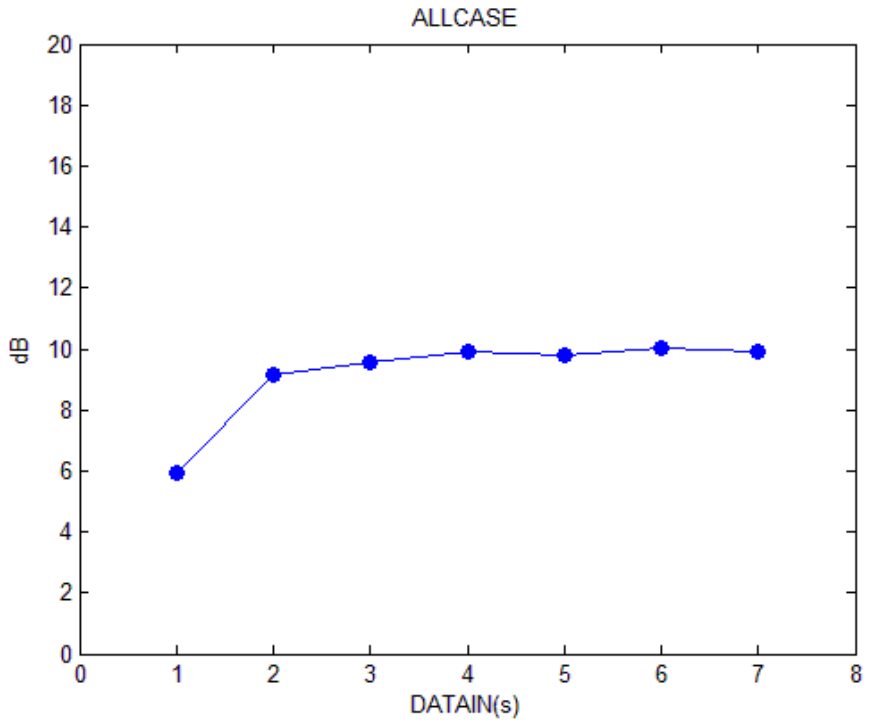
Fig. 27 Microphone number test in the Second-EXP of CASE-B.2 (AWGN)

4.2.2 The Effect of Data Size

In this section, we focus on the effect of input data size, that is, we choose different data length of mixture signals to test the BSS algorithm. Starting from a small size inputs, increasing data size can significantly improve the performance. When the input data reach a certain amount, we get less improvement. Therefore, we ought to limit data to a proper size to reduce delay and processing cost. In our experiments, we have one hundred and twenty test sequences. The sequences contain ten combinations of $\theta = 0^o, \pm 15^o, \pm 30^o$ as shown in Fig. 19. Each combination has twelve groups as shown in Table. 4. Here, we set the data window size to 512 samples and the source distance to 1.5M. For convenience, data length is abbreviated as DL.



(a) First-EXP



(b) Second-EXP

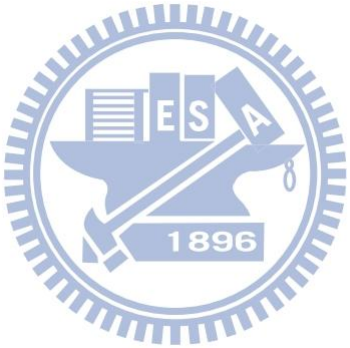
Fig. 28 Data Length test in CASE-A (Real)

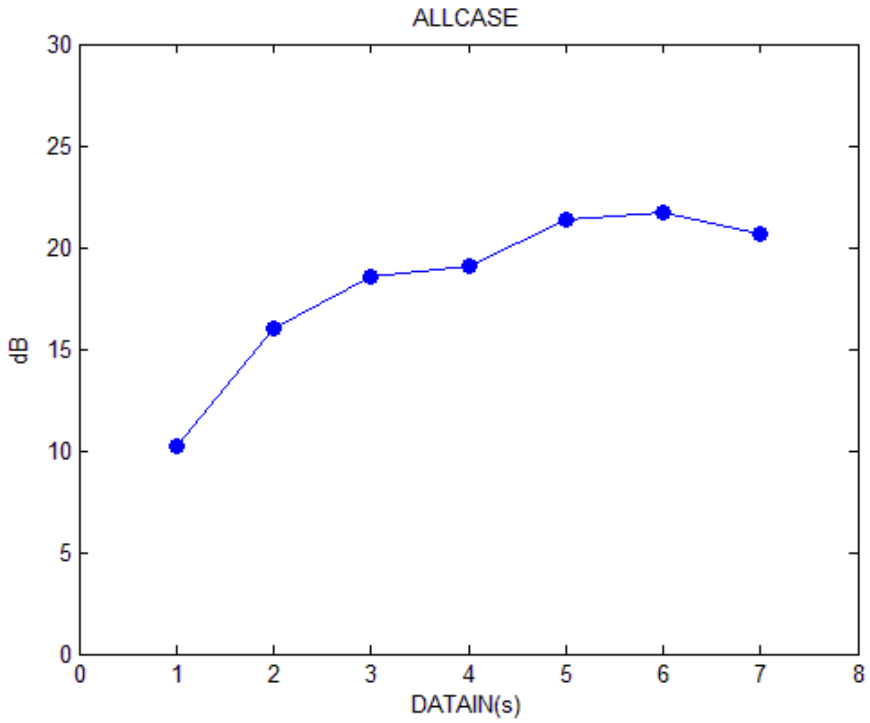
In CASE-A, we show the First-EXP and the Second-EXP results in Fig. 28. They indicate that different DL affects the BSS performance. The data points in Fig. 28 , Fig. 29 and Fig. 30 are collected from 120 test sequences. The x-axis represents the size of DL (sampling rate: 8KHz). According to Fig. 28(a), we observe that the performance of one-second DL, the shortest length, is the worst. When the DL increases to two seconds, the performance gets better obviously. However, we notice that the performance saturates at about four-second DL. In other words, when the DL gets beyond four seconds, the SIR does not gain much. A similar trend shows in Fig. 28(b), and the performance saturates at about three-second DL.

Fig. 29(a) and (b) show the same test in a SLAB-based simulation. They show that there is the similar trend as in the real data case. In Fig. 30, we add the AWGN into the

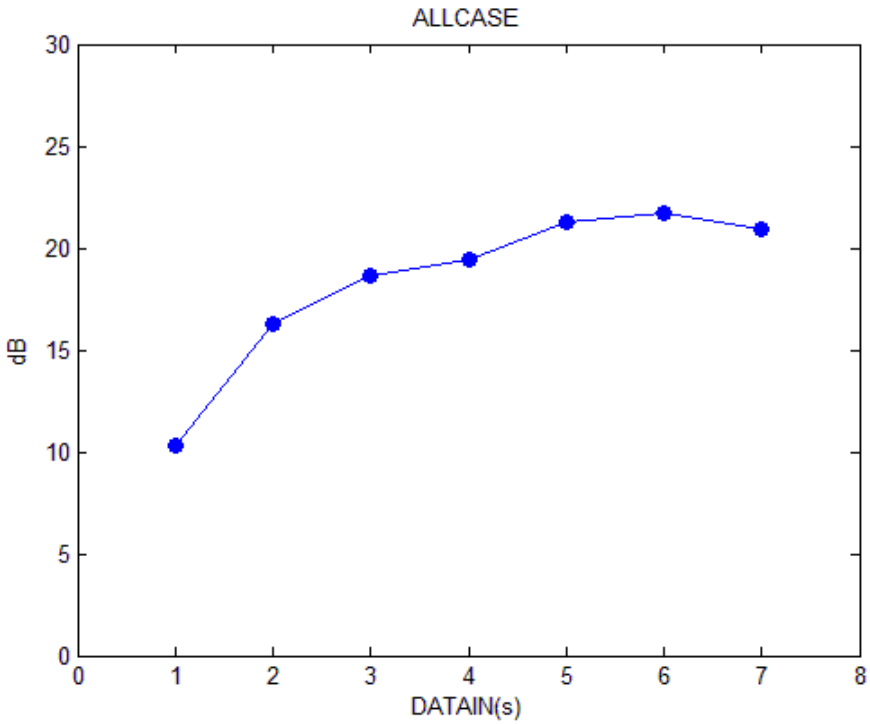
SLAB-based simulation. Fig. 30(a)~(b) also show that the curve is similar to the other cases.

Based on these results, we have the following conclusions. First, the larger data inputs provide better BSS performance. The improvement saturates at about four seconds for the real test data. Second, we should avoid having silent part in the BSS algorithm.



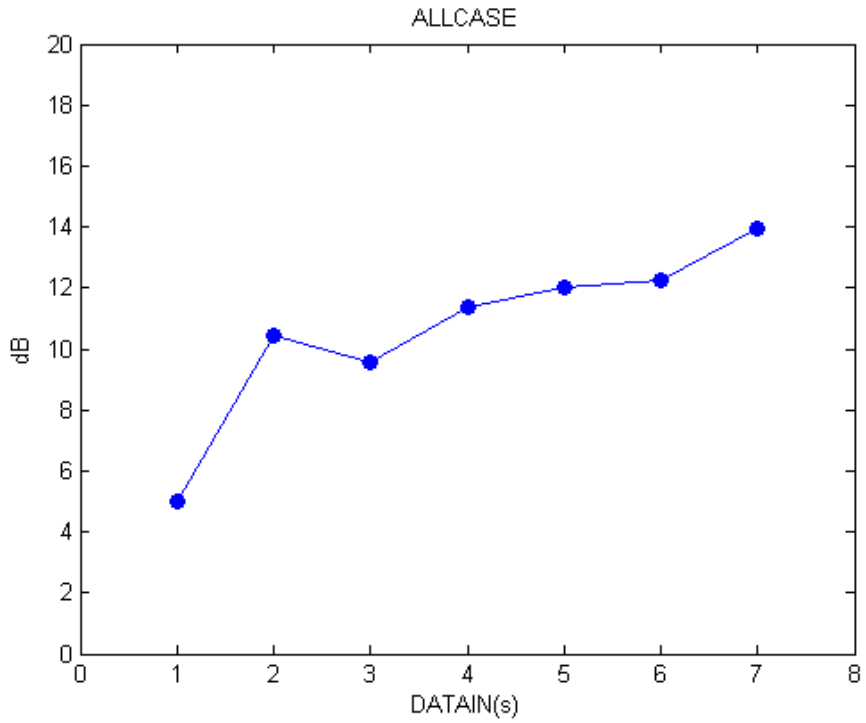


(a) Simulation with three Microphones

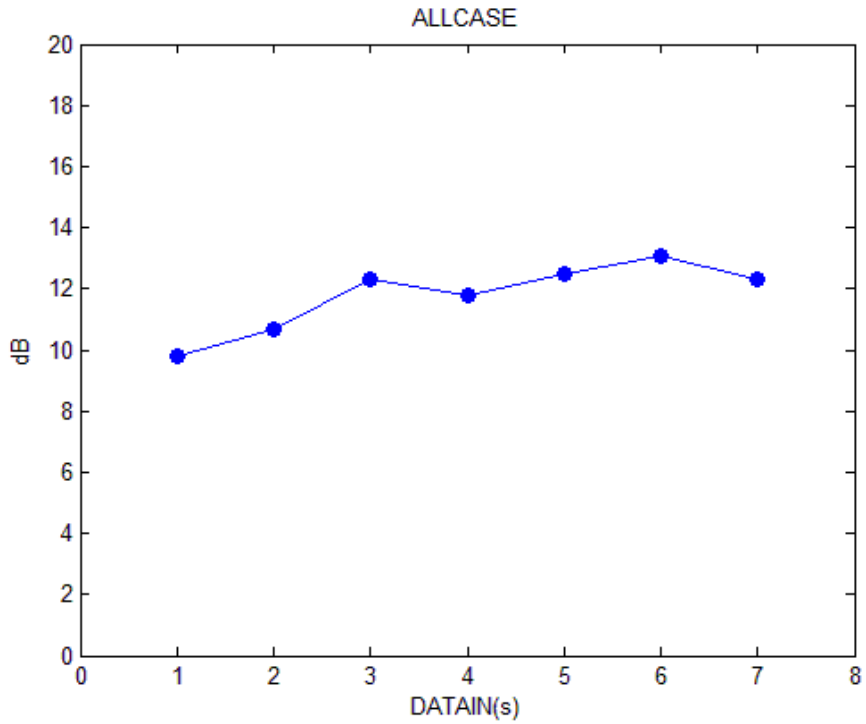


(b) Simulation with seven Microphones

Fig. 29 Data Length test in CASE-B.1 (SLAB)



(a) Simulation of the First-EXP



(b) Simulation of the Second-EXP

Fig. 30 Data Length test in CASE-B.2 (AWGN)

4.2.3 The Effect of source Distance

In this section, we focus on the effect of the distance between source and sensor. When we record the speech signals, in general, they come from different distances. Thus, we select speeches with different distances to create source signals. In our experiments, we fix the microphone array at a point and place the speech source with $\theta = 15^{o}$. We set the distances between the microphone array and the source to 1M, 1.5M and 2M. The set-up is shown in Fig. 31. Here, the microphone number is three, the data input is four seconds (32000 sampling points) and the window size is 512 samples.

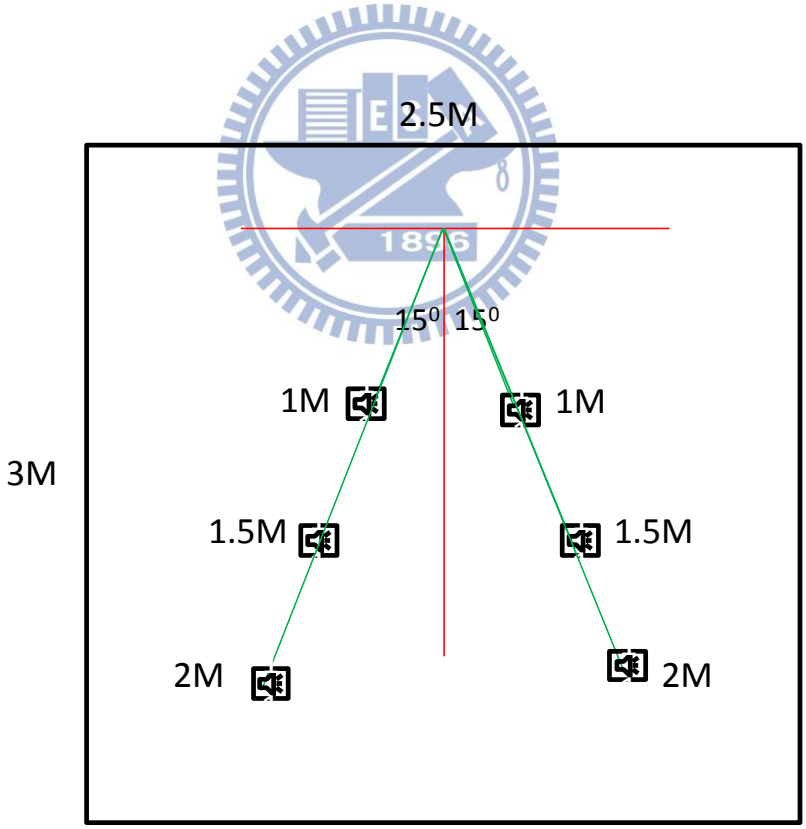
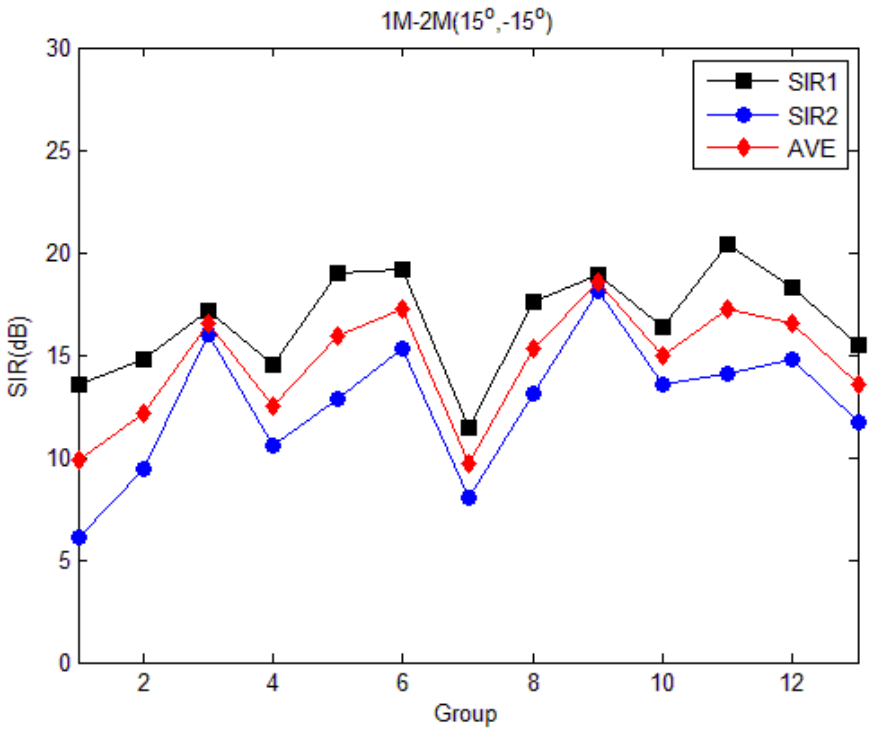
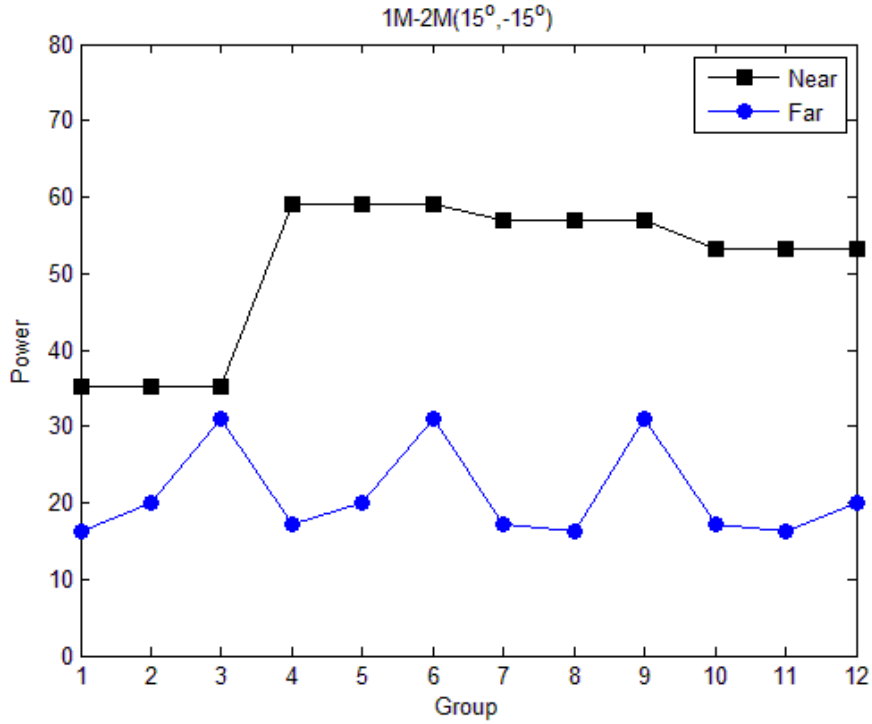


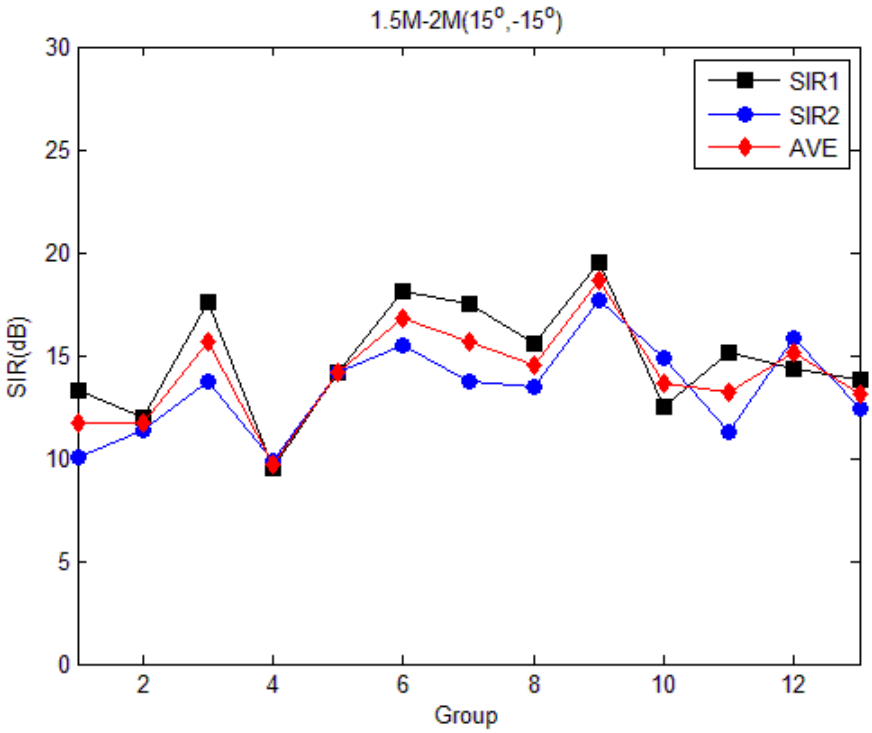
Fig. 31 The Placement of distance test in an Anechoic Chamber



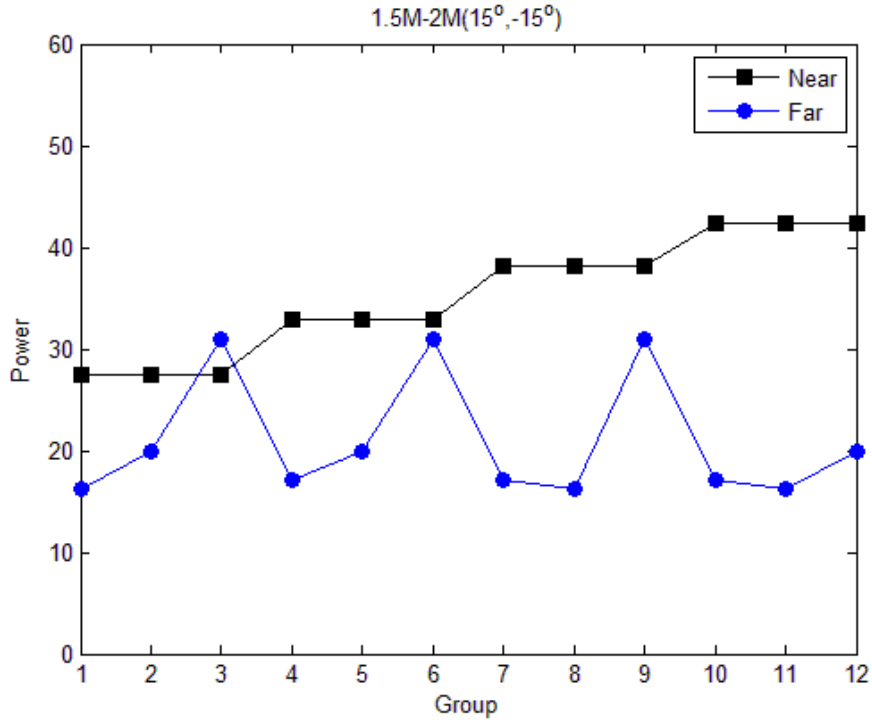
(a) SIR at Distances 1M and 2M



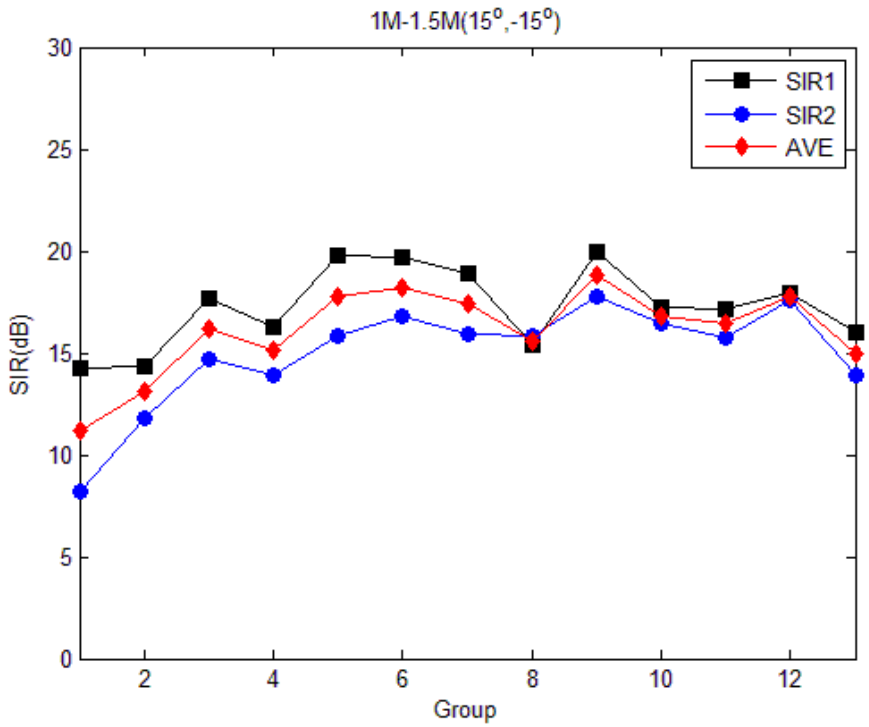
(b) Power at Distances 1M and 2M



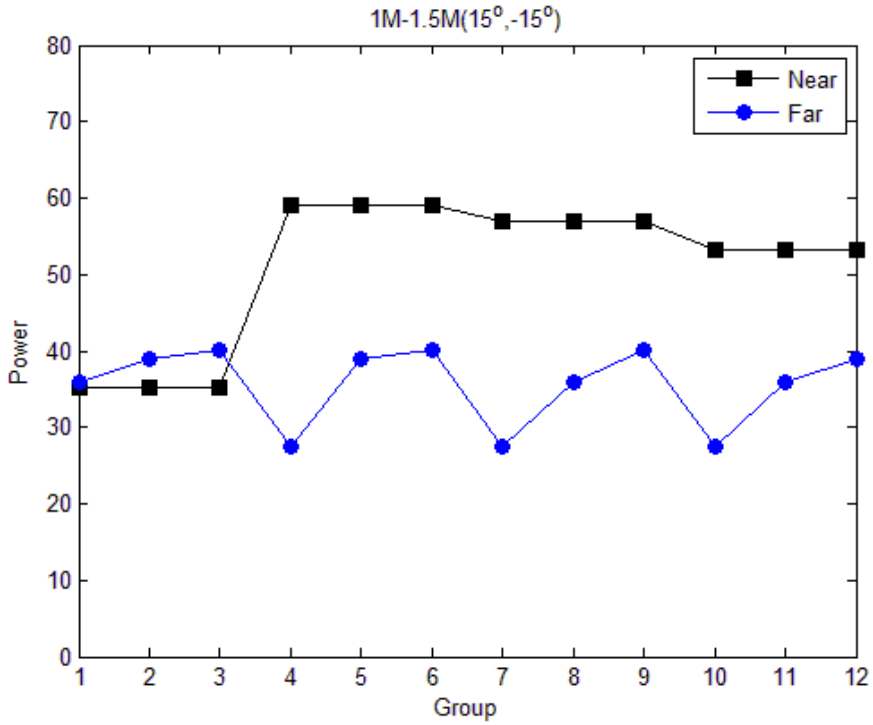
(c) SIR at Distances 1.5M and 2M



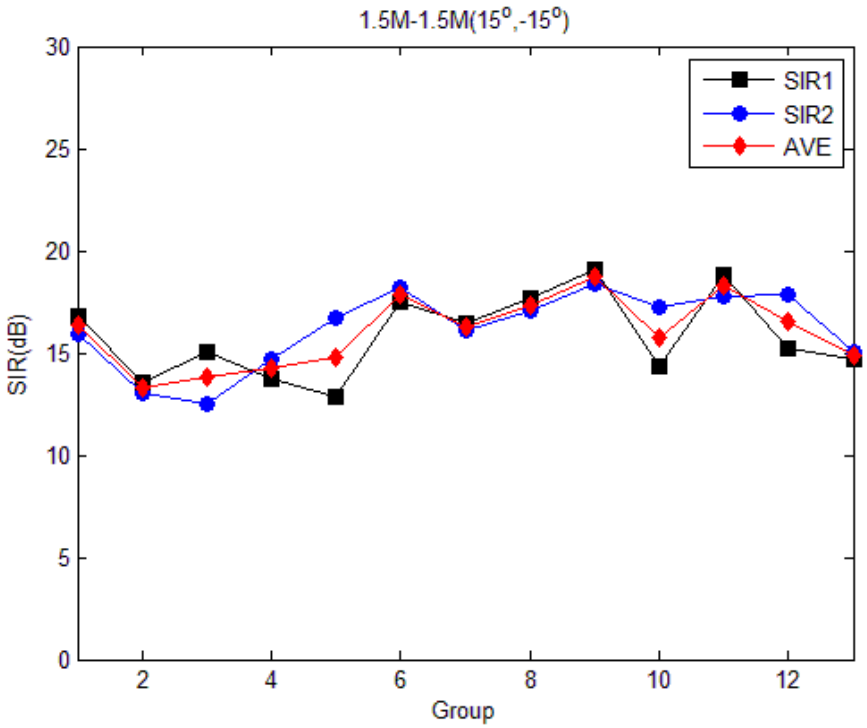
(d) Power at Distances 1.5M and 2M



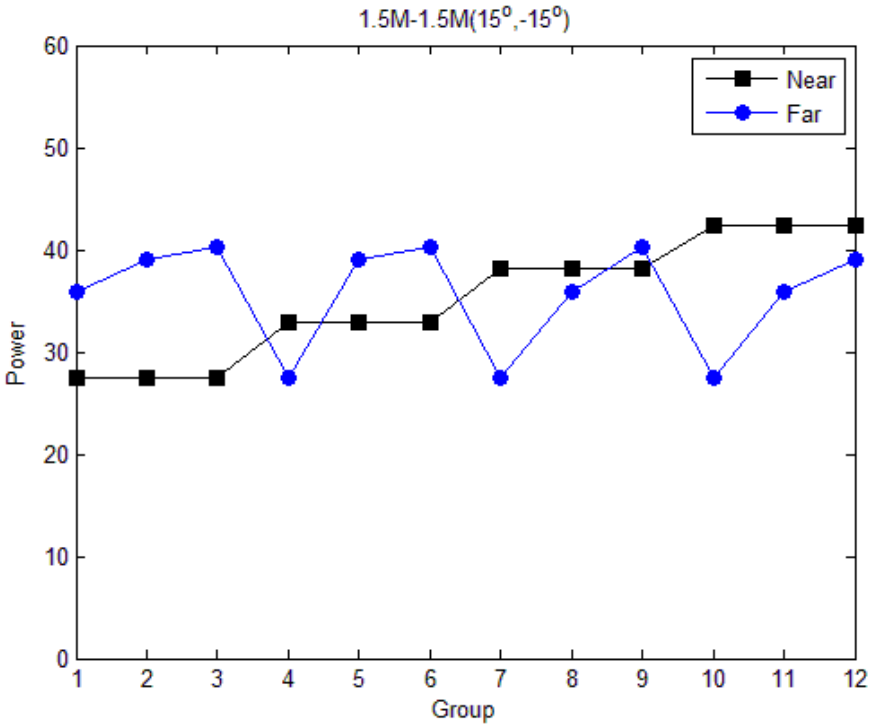
(e) SIR at Distances 1M and 1.5M



(f) Power at Distances 1M and 1.5M



(g) SIR at Distances 1.5M and 1.5M



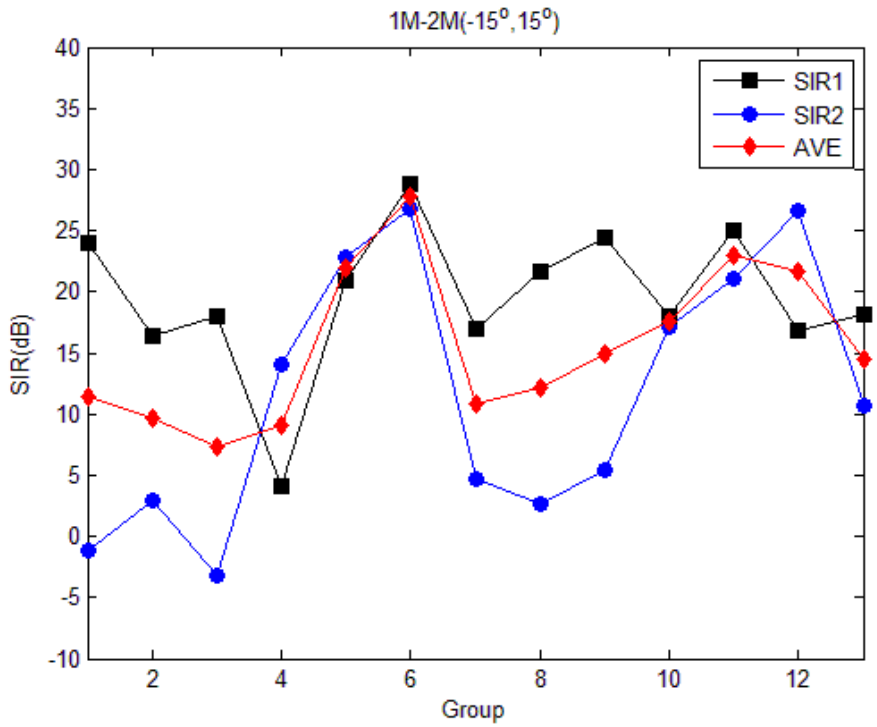
(h) Power at Distances 1.5M and 1.5M

Fig. 32 Different Distance test in the First-EXP of CASE-A (Real)

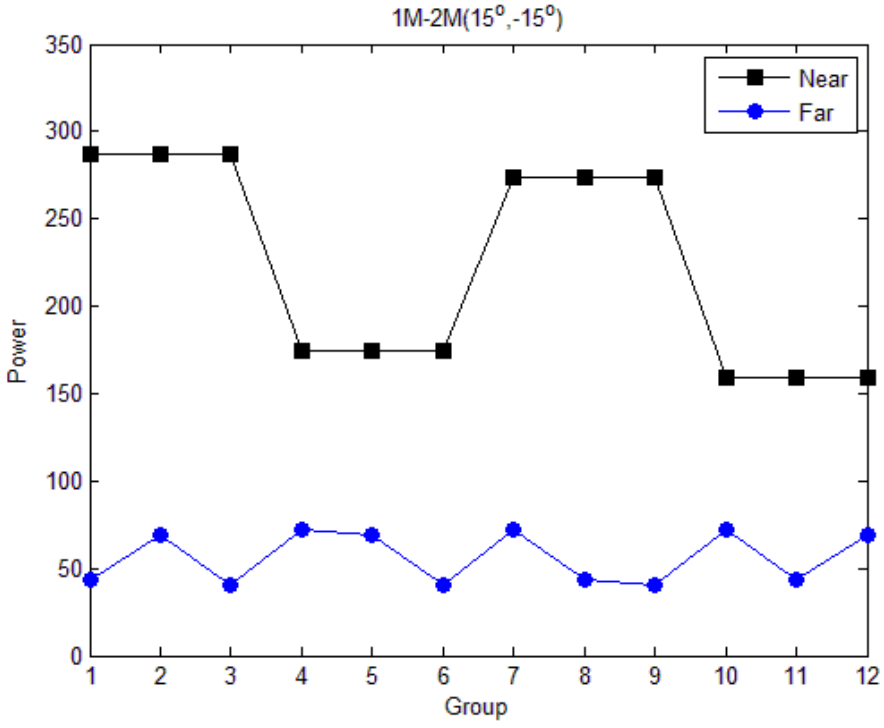
Here, we only show the First-EXP experiment of CASE-A in Fig. 32. Fig. 32(a) shows sources at 1M and 2M. Fig. 32(b) shows sources at 1.5M and 2M. Fig. 32(c) shows sources at 1M and 1.5M. Fig. 32(d) shows both sources at 1.5M and 1.5M. The x-axis represents the groups specified in Table. 4. In Fig. 32(a), we notice that the result is better when the source is close to the sensor. Also, the separated far source signal has lower SIR. In the meantime, we calculate the power of two received source signals, shown in Fig. 32(b). We can observe that the high SIR comes from the high power signal. In Fig. 32(c)~(f), we observe the same trend. When two speeches have the same distance, each SIR value is close to the other one as shown in Fig. 32(g) and (h). Typically, the average SIR for the cases with different distances is worse than those with the same distance.

Fig. 33 shows the test in the SLAB-based simulation. Here, we only show two test sequences. Fig. 33(a) shows sources at 1M and 2M. Fig. 33(c) shows both sources at 1.5M and 1.5M. We observe the same trends as discussed previously. The SIR of high power source has a better performance. In Fig. 34, we add AWGN into the SLAB-based simulation. The SIR has similar curves to the other simulations. For the most part, the results are better for the sources with the same distance.

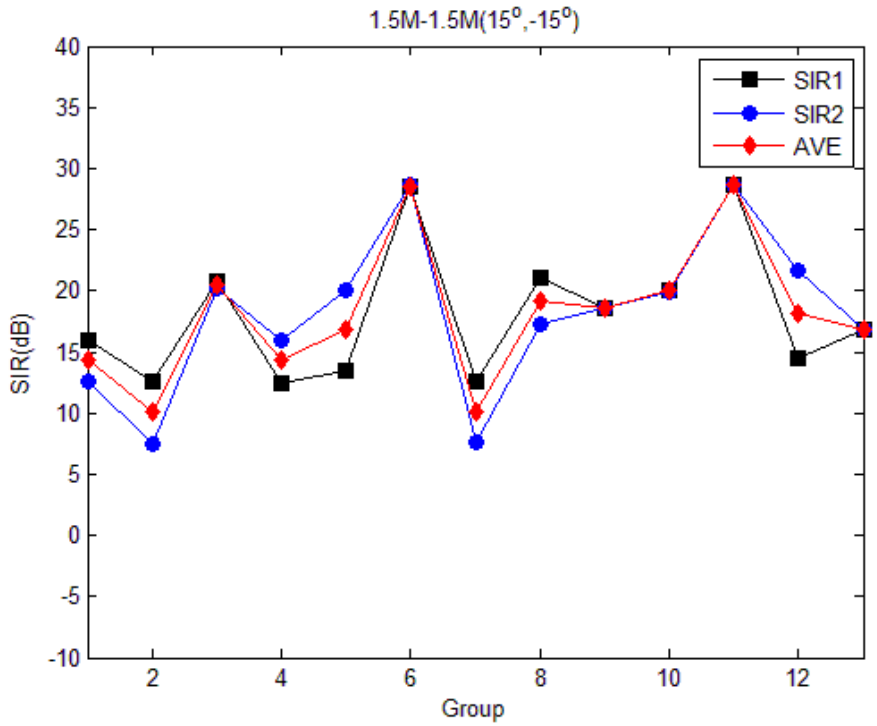
Based on these results, we can observe some trends from them. The result provides the better performance as two source signals has the same distance. In other words, the power of two signals is quite similar.



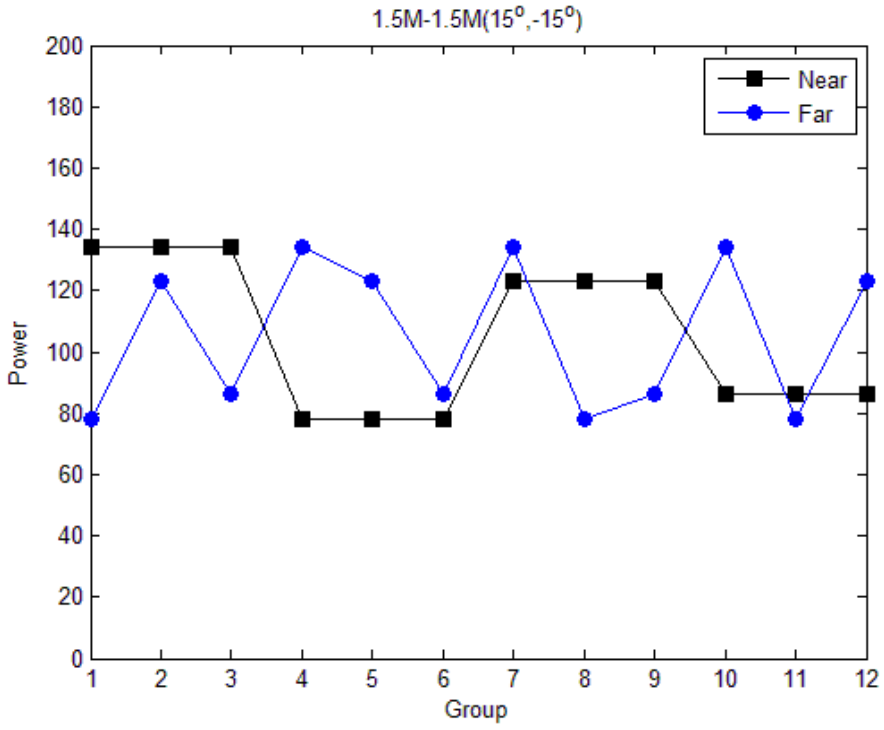
(a) SIR at Distance 1M and 2M



(b) Power at Distance 1M and 2M

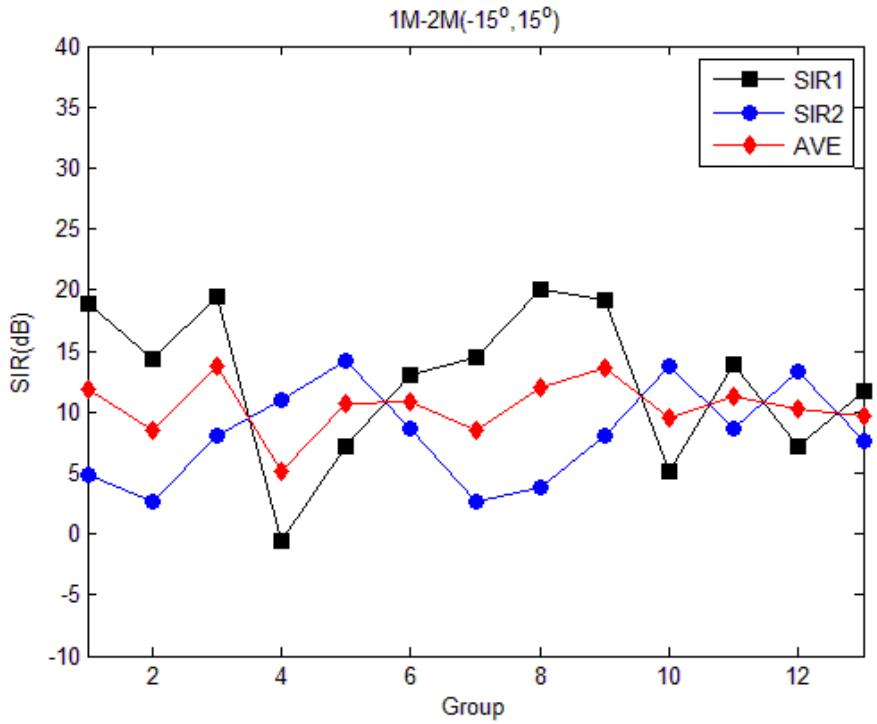


(c) SIR at Distance 1.5M and 1.5M

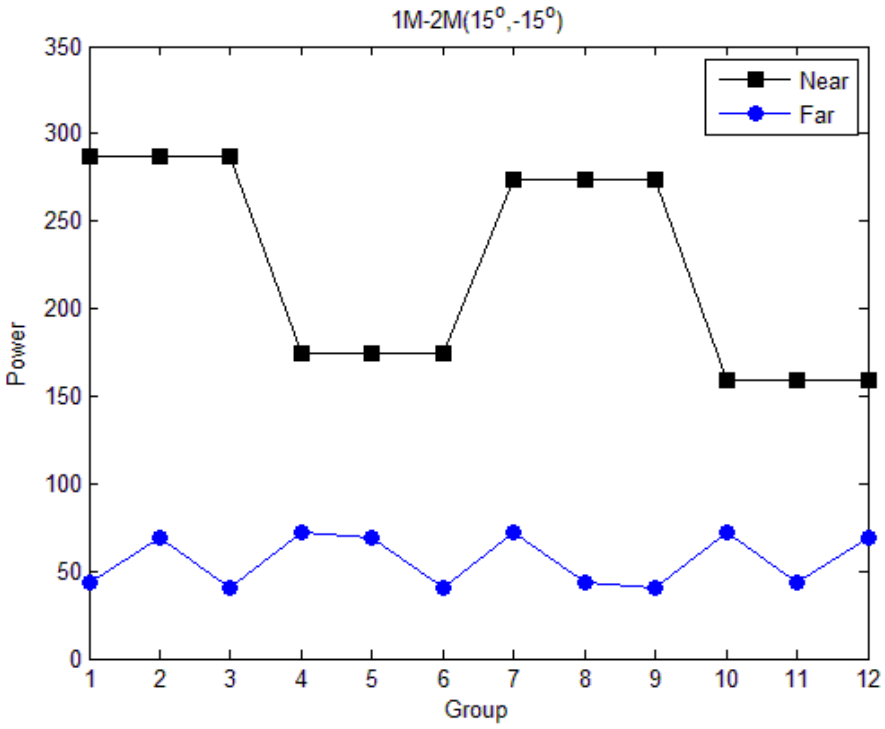


(d) Power at Distance 1.5M and 1.5M

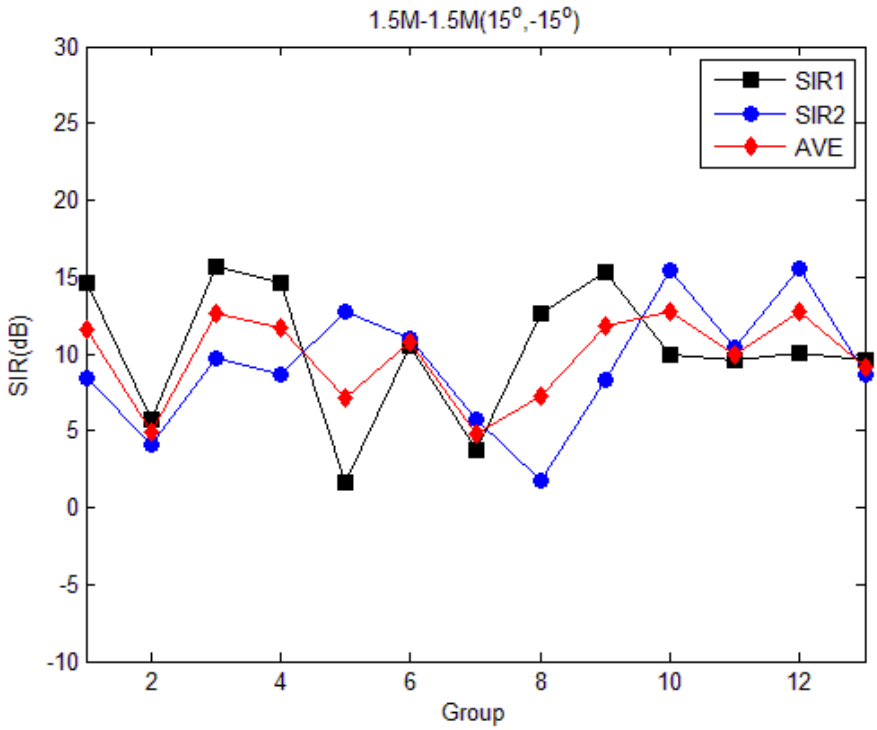
Fig. 33 Different Distance test with 3 Sensors in CASE-B.1 (SLAB)



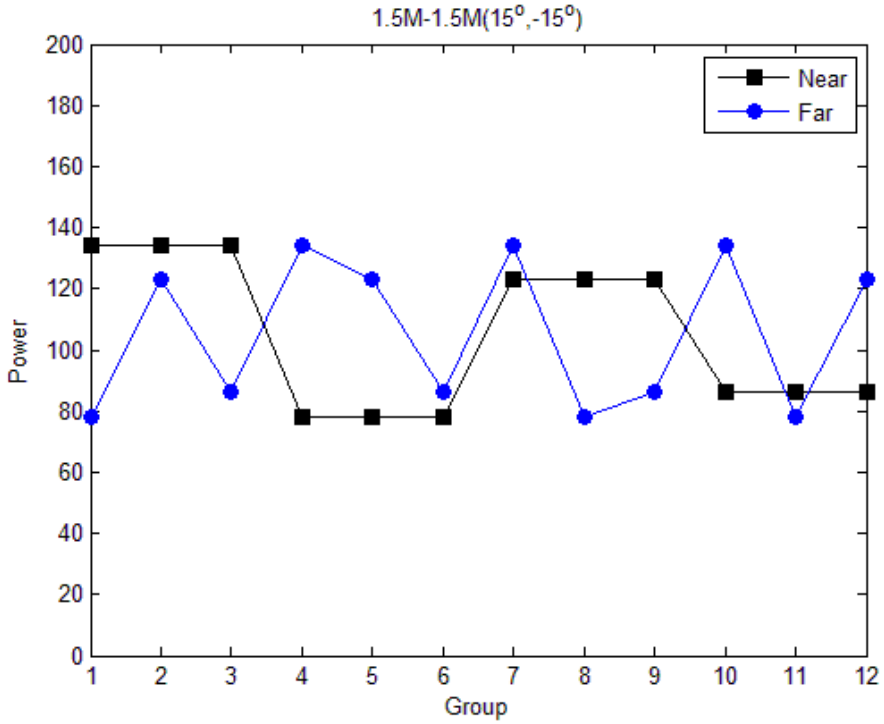
(a) SIR at Distance 1M and 2M



(b) Power at Distance 1M and 2M



(c) SIR at Distance 1.5M and 1.5M



(d) Power at Distance 1.5M and 1.5M

Fig. 34 Different Distance test in the First-EXP of CASE-B.2 (AWGN)

4.2.4 BSS Performance in Three Types

In this section, we focus on the effect of denoising. In a real acoustic environment, the environment parameters include the air absorption and microphone intrinsic distortion, and others. They generate audio noises and degradation. We adopt the denoising technique [4] described in Chapter 3 to reduce the noise. Here, we compare three cases BSS, which are shown in Fig. 35. In the Type_1 case, we use the IVA algorithm [13] in Chapter 2 to obtain the separated signals. In type_2, the denoising technique is applied the separated signals. In type_3, the denoising technique is applied to the inputs before they are processed by BSS.

Fig. 36~Fig. 39 show the results of Type_1, Type_2 and Type_3 in the First-EXP and the Second-EXP of CASE.A. Here, we, respectively, show two test sequences for the First-EXP and the Second-EXP. The first test sequence consists of Female_1 and Female_2. The second test sequence consists of Male_1 and Male_2. The angles of the sources in the sequences come from 15^o and -15^o, and the distance are all equal to 1.5M. The left-hand side of figures shows the time-domain signals we separate from the mixed signals. The right-hand side of figures shows the frequency-domain representations. Based on these figures, we observe some trends. In Type_1, we notice that there are still some signals at high frequency. In Type_2 and Type_3, the high frequency components are reduced and the low frequency components are reserved. It is quite obvious that the denoising technique removes the high frequency components. The filtered signals are rather similar in both Type_2 and Type_3. Thus, applying the denoising filter, before or after BSS, does not seem to be critical in our applications.

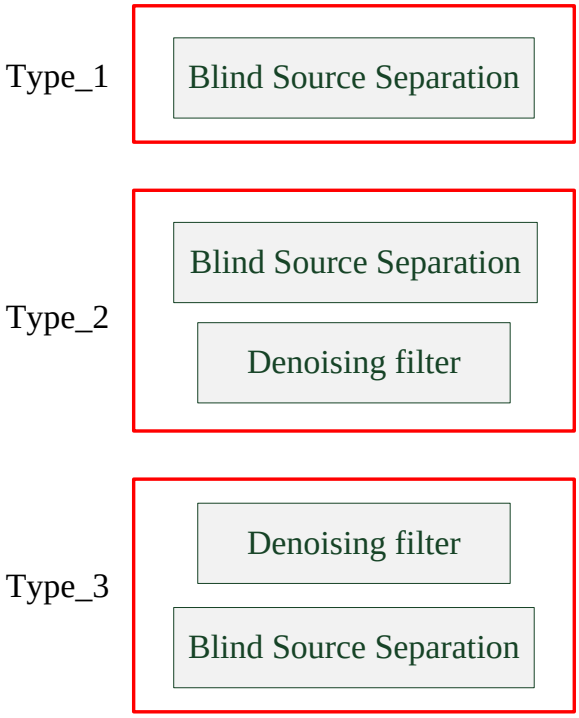
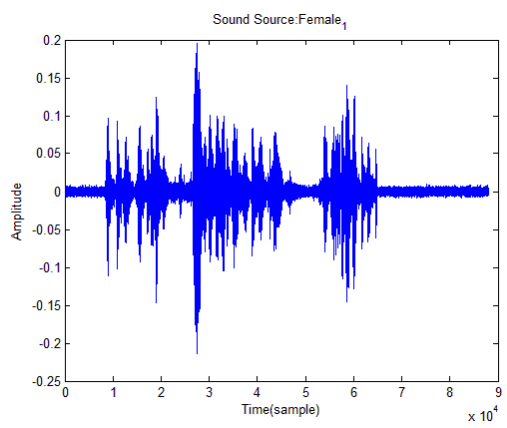
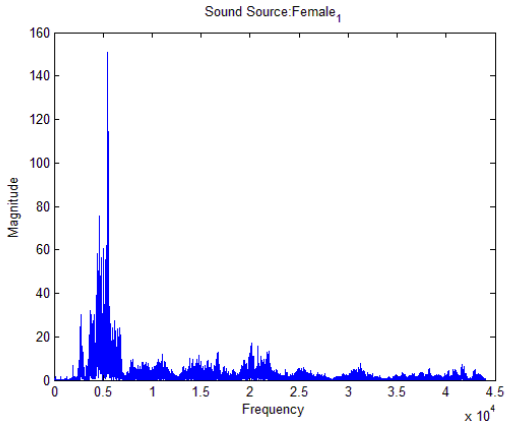
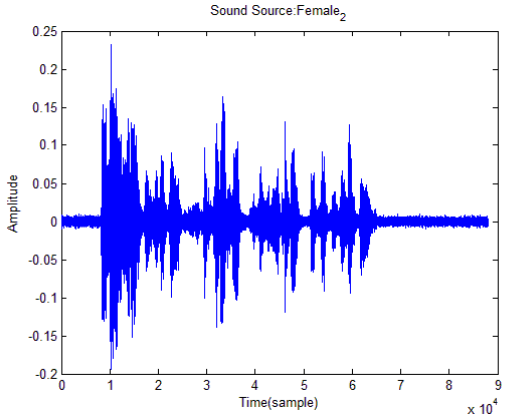


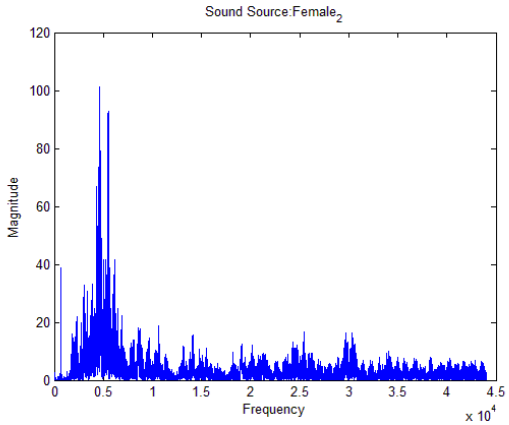
Fig. 35 BSS and inserting denoising filters

Type 1

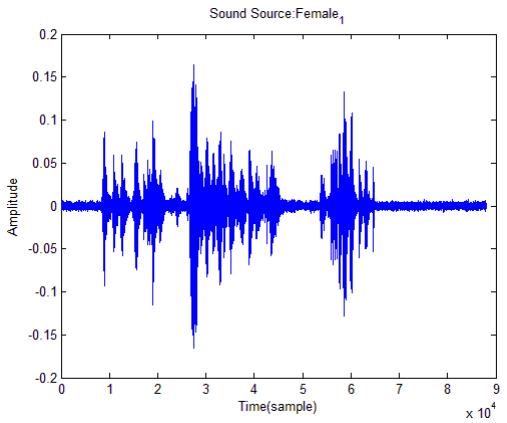

Sound Source:Female1
Amplitude
Time(sample)
x 10^4

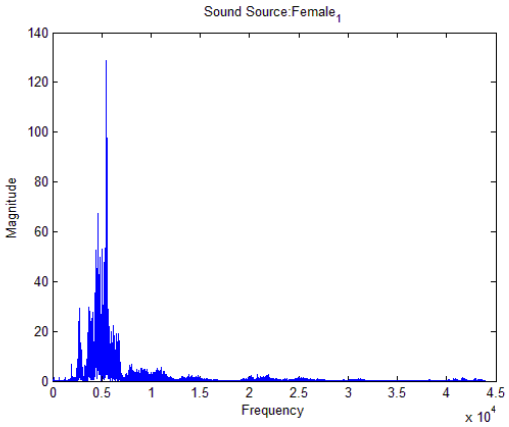

Sound Source:Female1
Magnitude
Frequency
x 10^4

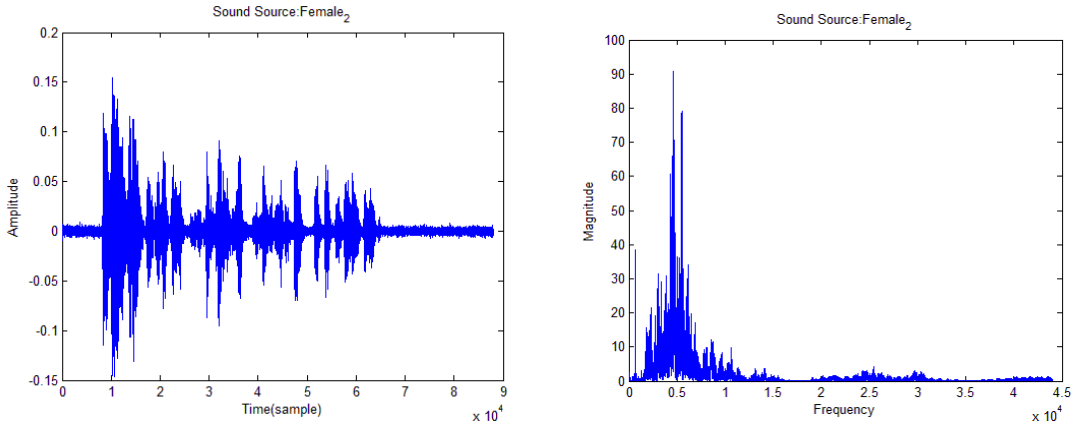

Sound Source:Female2
Amplitude
Time(sample)
x 10^4


Sound Source:Female2
Magnitude
Frequency
x 10^4

Type 2


Sound Source:Female1
Amplitude
Time(sample)
x 10^4


Sound Source:Female1
Magnitude
Frequency
x 10^4



Type 3

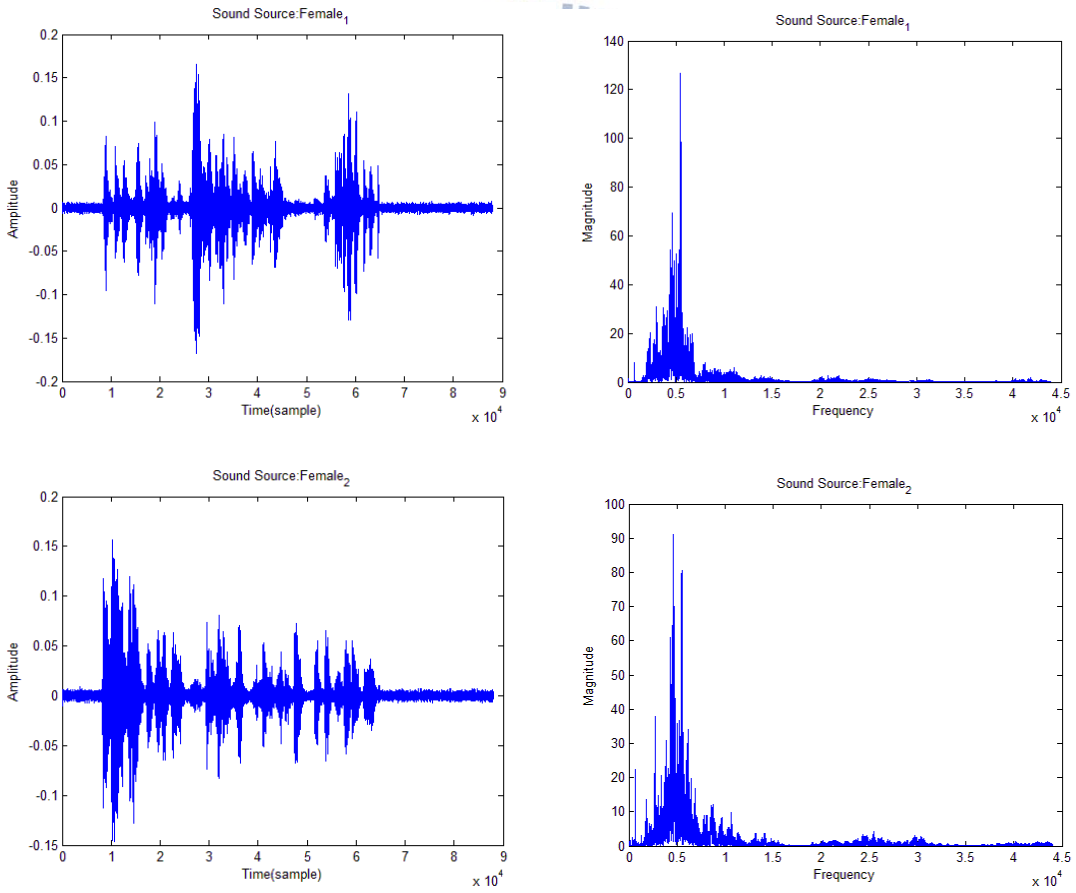
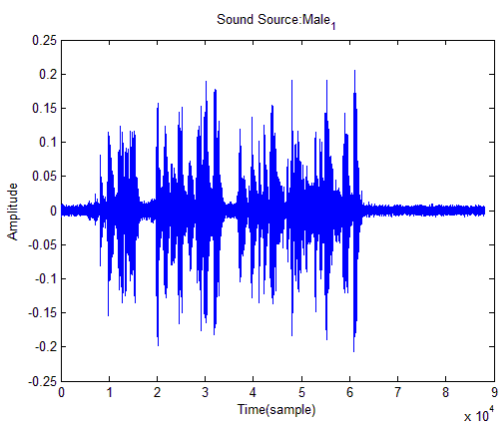
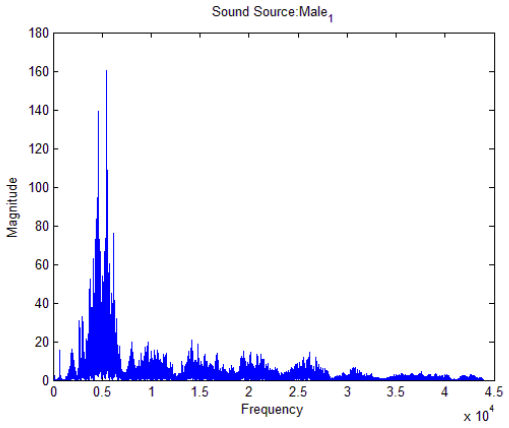


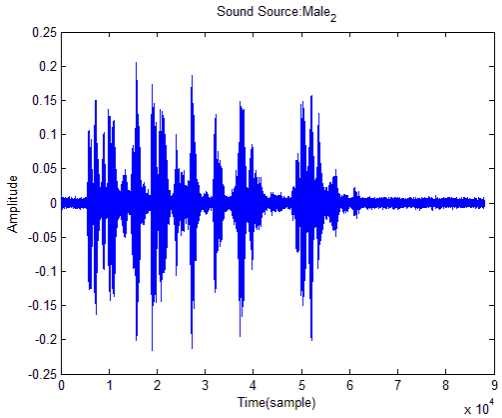
Fig. 36 Denoising effects in the First-EXP of CASE-A (Real)

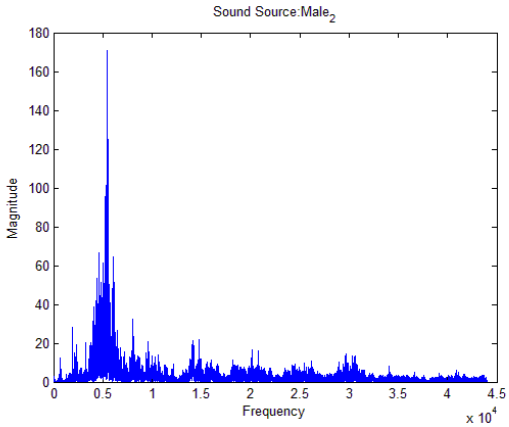
(Female_1 & Female_2)

Type 1

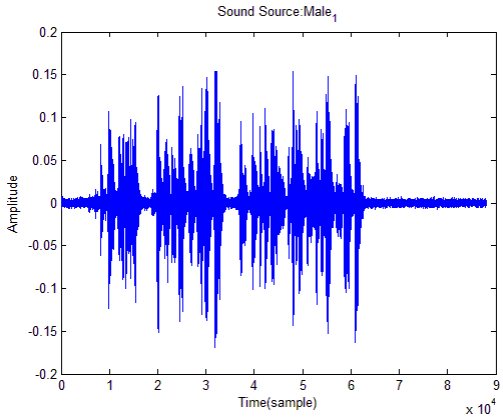

Sound Source:Male1
Amplitude
Time(sample)
x 10^4

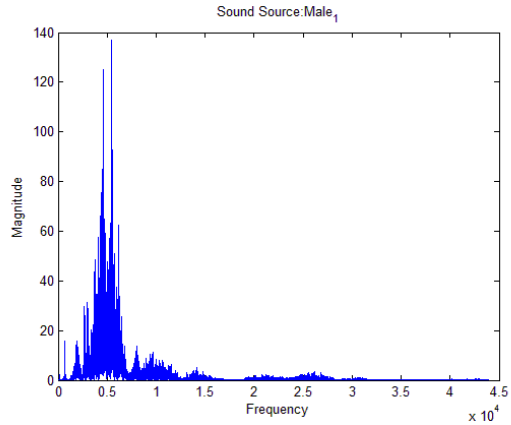

Sound Source:Male1
Magnitude
Frequency
x 10^4

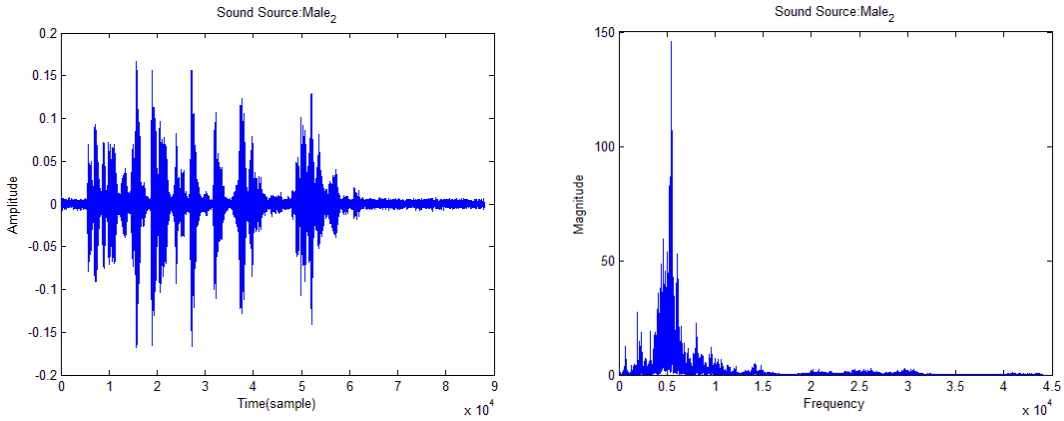

Sound Source:Male2
Amplitude
Time(sample)
x 10^4


Sound Source:Male2
Magnitude
Frequency
x 10^4

Type 2


Sound Source:Male1
Amplitude
Time(sample)
x 10^4


Sound Source:Male1
Magnitude
Frequency
x 10^4



Type 3

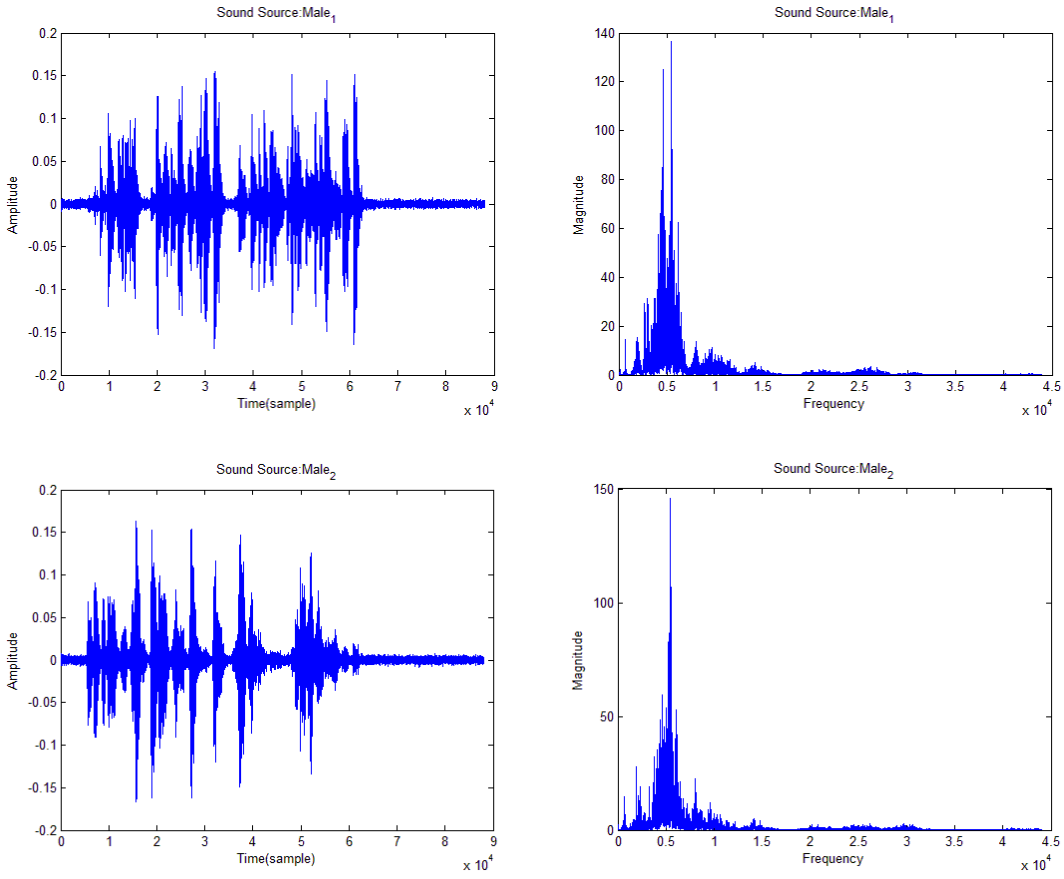
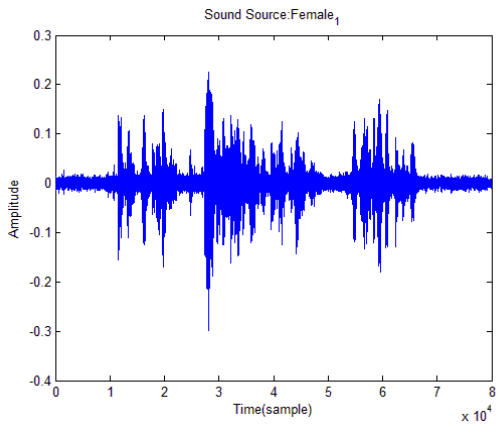
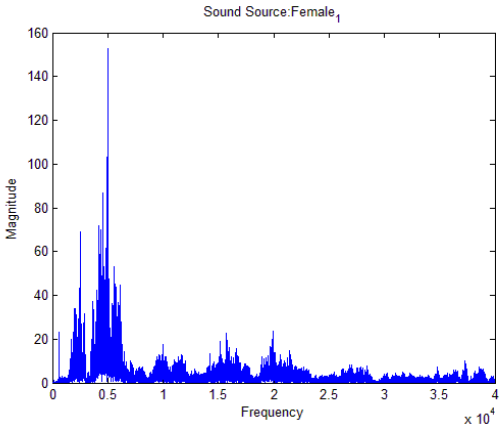


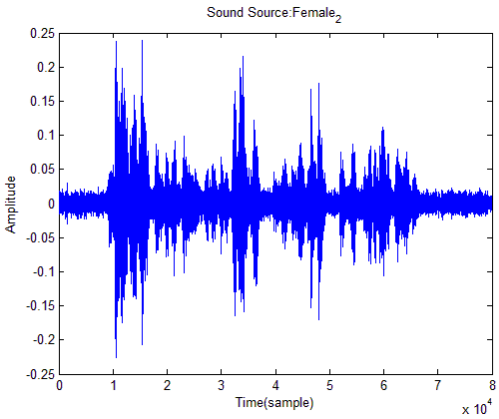
Fig. 37 Denoising effects in the First-EXP of CASE-A (Real)

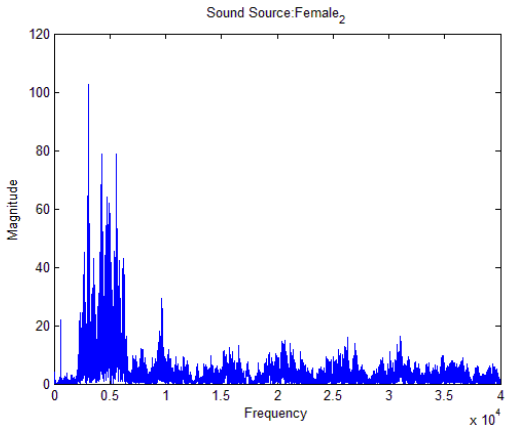
(Male_1 & Male_2)

Type 1

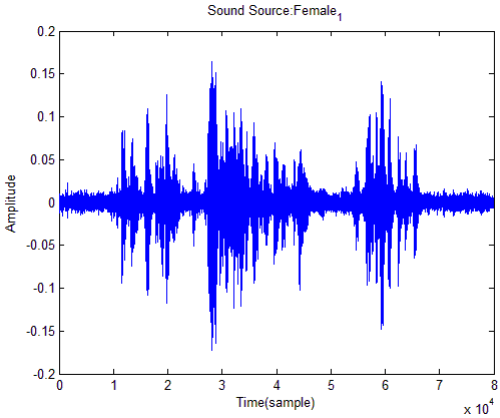

Sound Source:Female1
Amplitude
Time(sample)
x 10^4

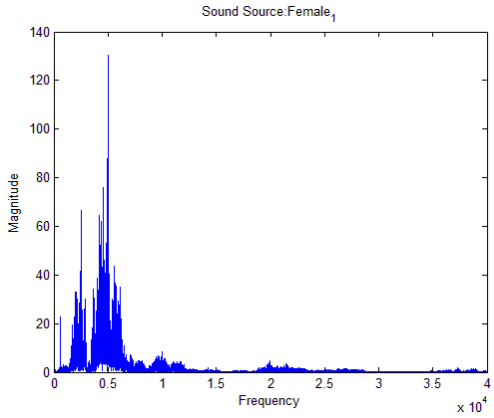

Sound Source:Female1
Magnitude
Frequency
x 10^4

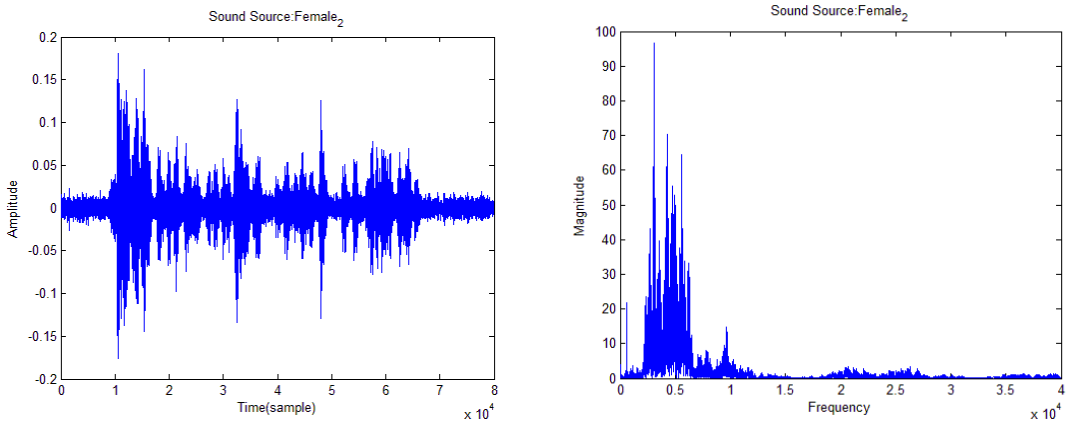

Sound Source:Female2
Amplitude
Time(sample)
x 10^4


Sound Source:Female2
Magnitude
Frequency
x 10^4

Type 2


Sound Source:Female1
Amplitude
Time(sample)
x 10^4


Sound Source:Female1
Magnitude
Frequency
x 10^4



Type 3

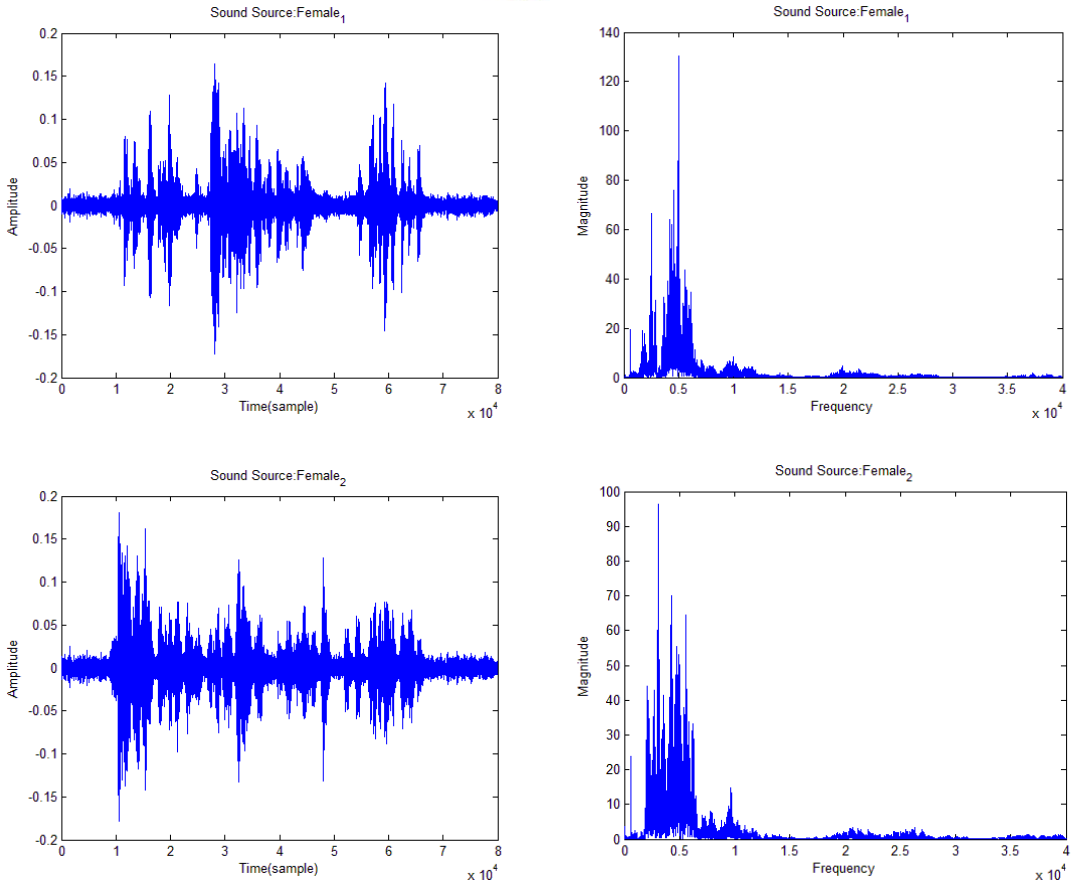
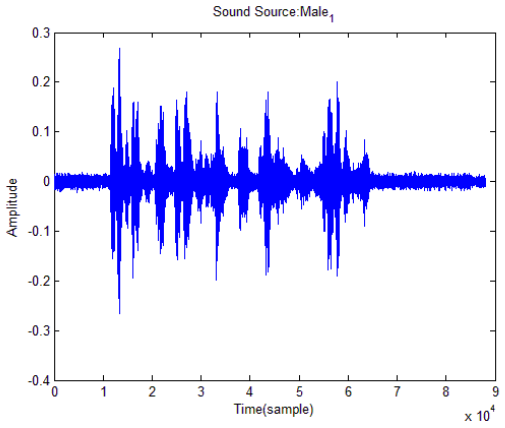
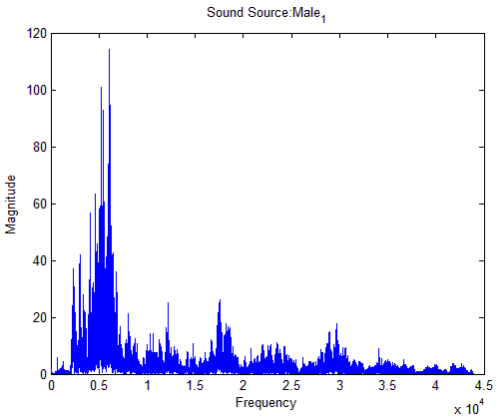


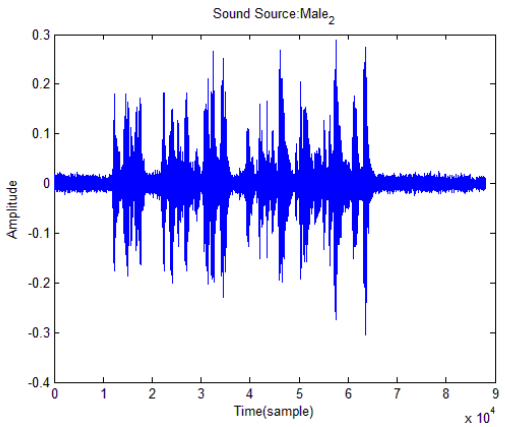
Fig. 38 Denoising effects in the Second-EXP of CASE-A (Real)

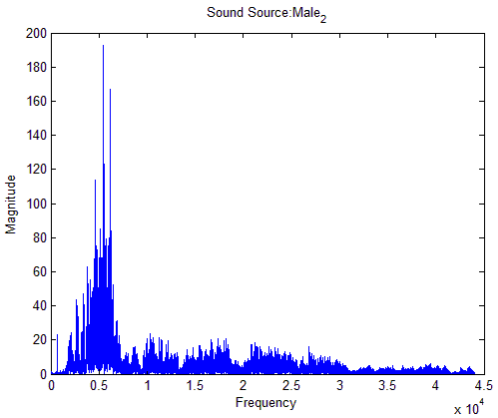
(Female_1 & Female_2)

Type 1

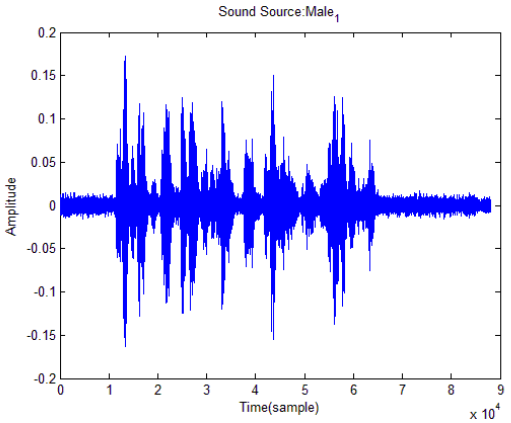

Sound Source:Male1
Amplitude
Time(sample)
x 10^4

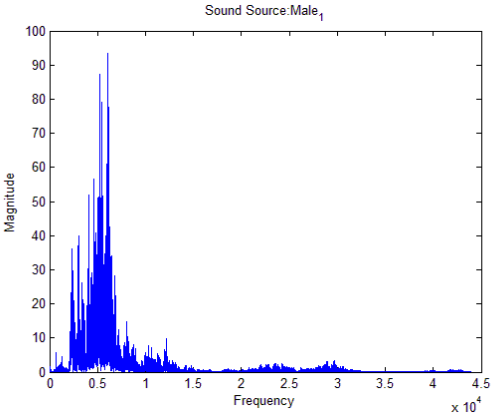

Sound Source:Male1
Magnitude
Frequency
x 10^4

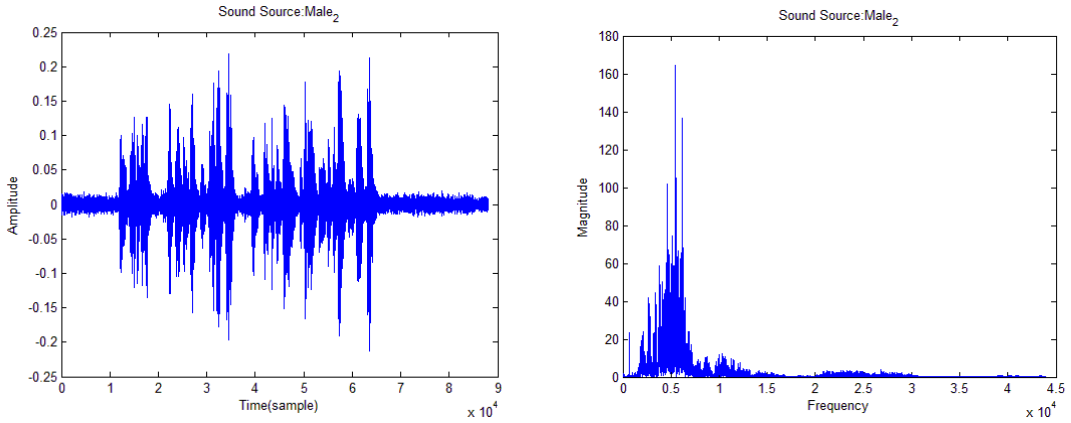

Sound Source:Male2
Amplitude
Time(sample)
x 10^4


Sound Source:Male2
Magnitude
Frequency
x 10^4

Type 2


Sound Source:Male1
Amplitude
Time(sample)
x 10^4


Sound Source:Male1
Magnitude
Frequency
x 10^4



Type 3

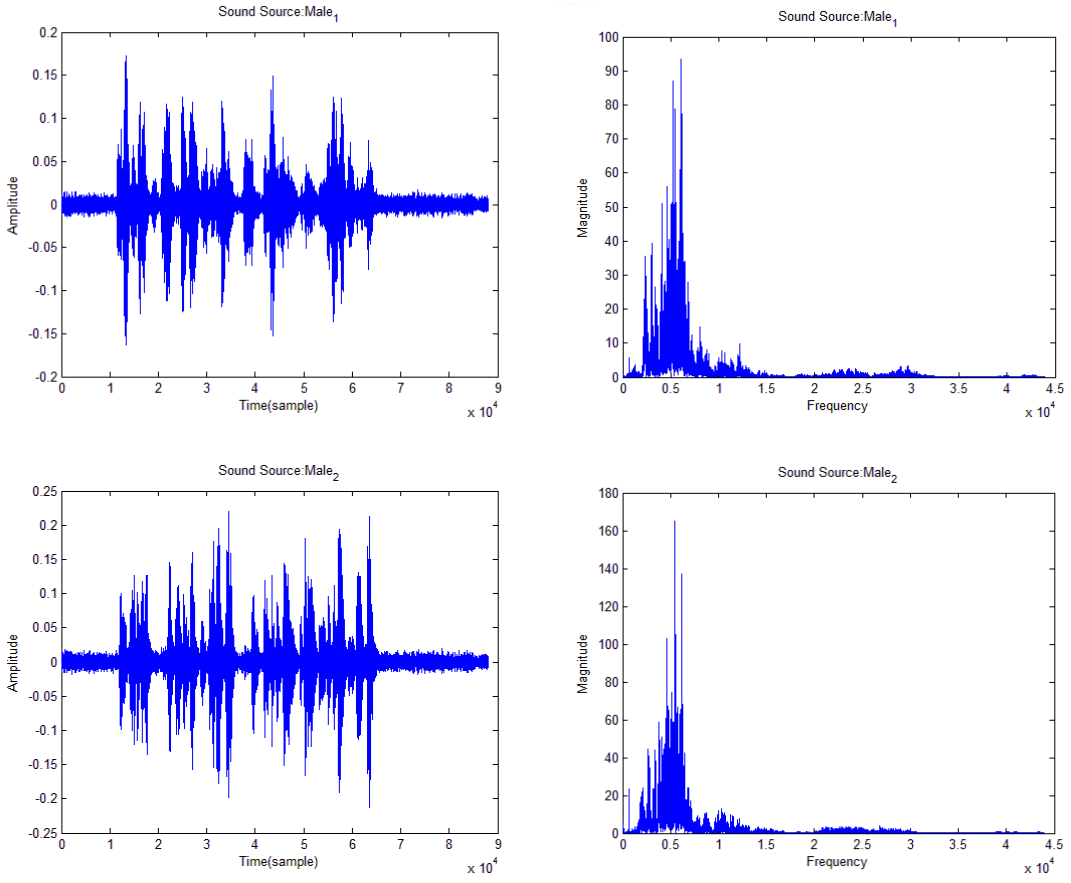


Fig. 39 Denoising effects in the Second-EXP of CASE-A (Real)

(Male_1 & Male_2)

Chapter 5 Experimental Results: Part B

5.1 Direction of Arrival Data Analysis

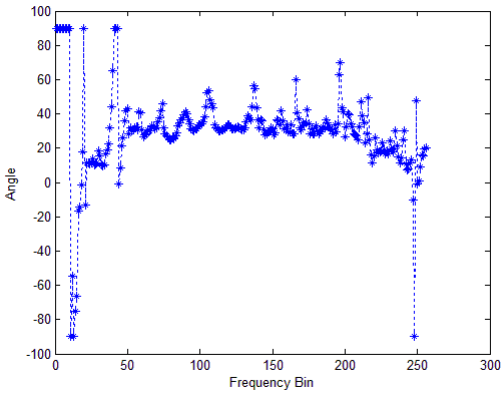
5.1.1 Frequency Bin Selection

In this section, we focus on the effect of frequency bins in the DOA estimation algorithm. In our experiments, we examine several cases to derive our final selections. We measure the estimation accuracy by using the mean absolute error (MAE). In statistics, the MAE is a formulation used to measure how close the predictions are to the true values. The definition of MAE is given below:

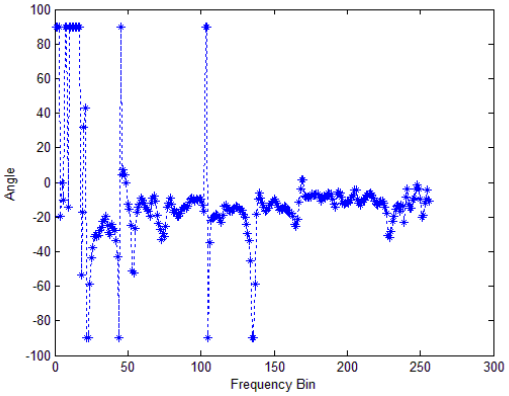
$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\theta_i - \theta|$$

where θ_i denotes the estimation of the i-th frequency bin and θ denotes the true value.

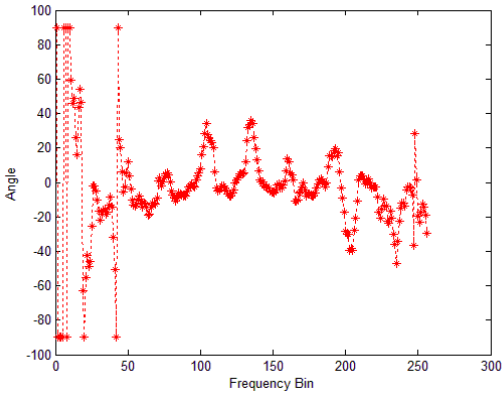
Here, we set the window size to 512 samples, the data input size to four seconds (32000 sampling points) and the distance between source and sensor to 1.5M. The window size, which is 512, leads to the same number of transform coefficients. The magnitudes of coefficients are symmetric at the middle point. Thus, we only use the first 256 bins to calculate the DOA estimates. The bin represents that the frequency components are placed at even intervals of f_{SAMPLE} / N_{WINDOW}. They are referred as frequency bins or FFT bins. We divide the 256 frequency bins into five intervals. Each interval contains 50 frequency bins, and we discard the final 16 bins since a typical speech signal contains less high frequency components.



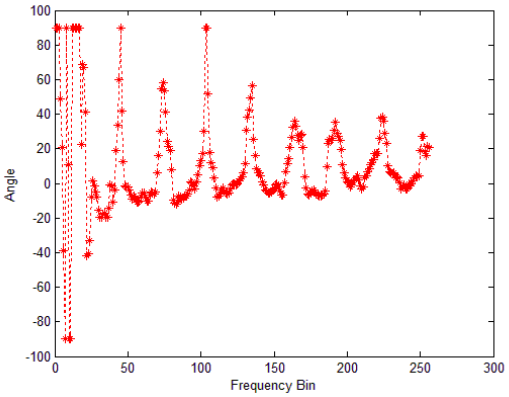
(a) $\theta = 15^o$



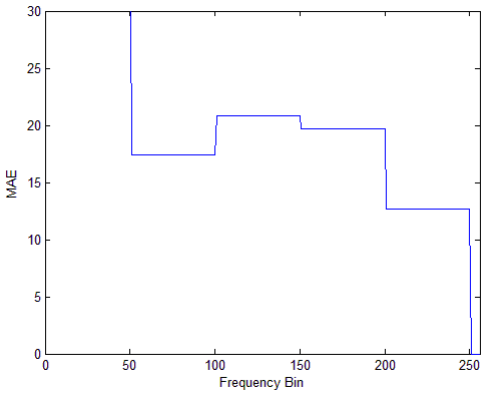
(b) $\theta = -15^o$



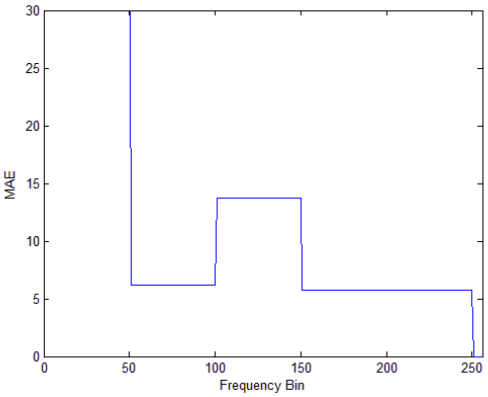
(e) $\phi = 0^o$



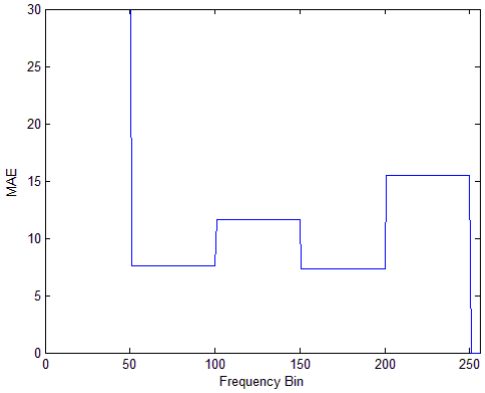
(f) $\phi = 0^o$



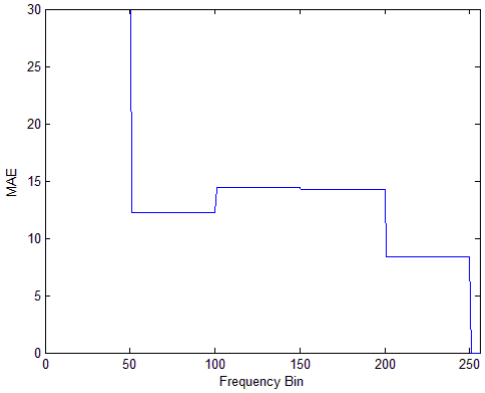
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$



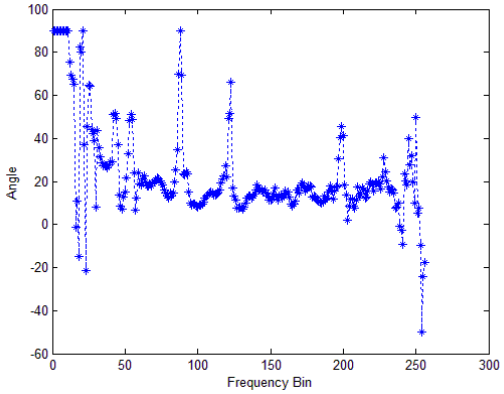
(g) MAE for $\phi = 0^o$



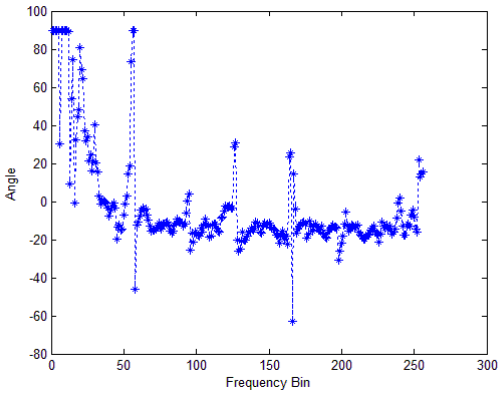
(h) MAE for $\phi = 0^o$

Fig. 40 DOA estimates (in various bins) in the First-EXP of CASE-A (Real)

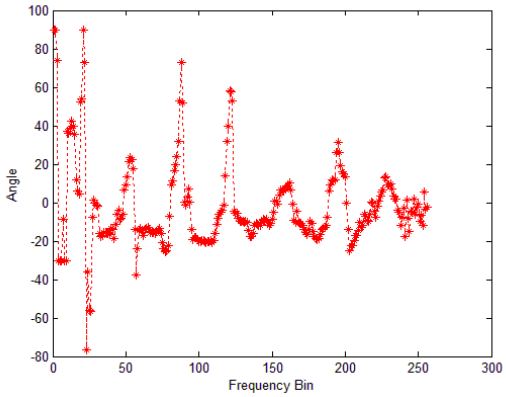
(Female_1 & Female_2)



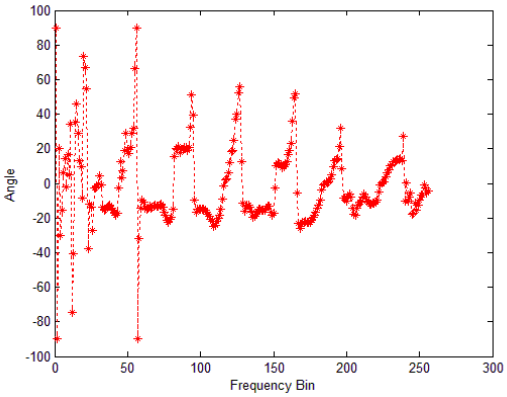
(a) $\theta = 15^o$



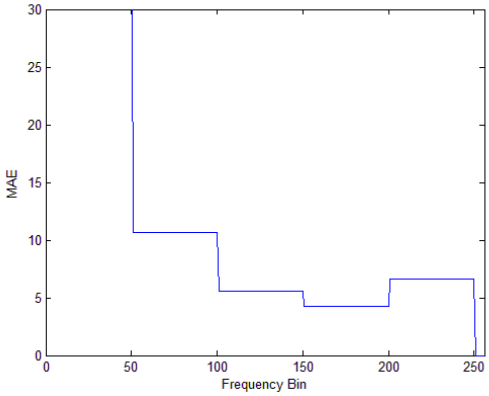
(b) $\theta = -15^o$



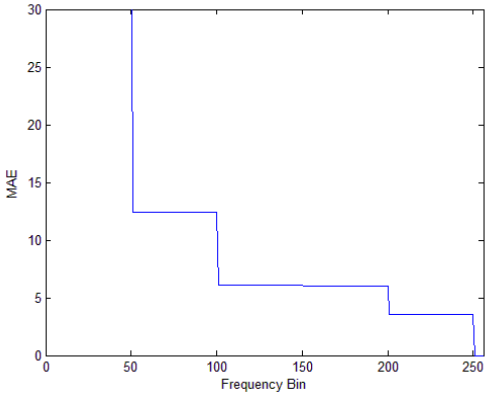
(e) $\phi = 0^o$



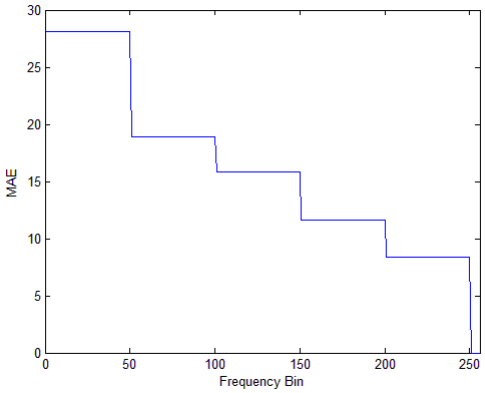
(f) $\phi = 0^o$



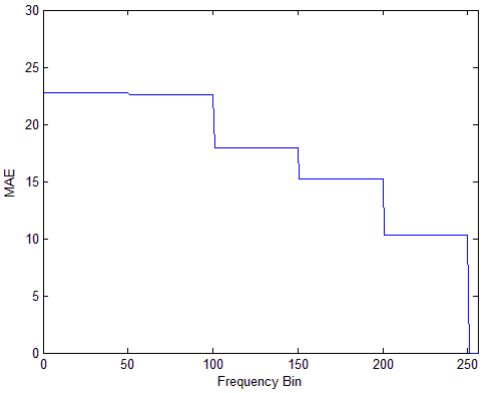
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$

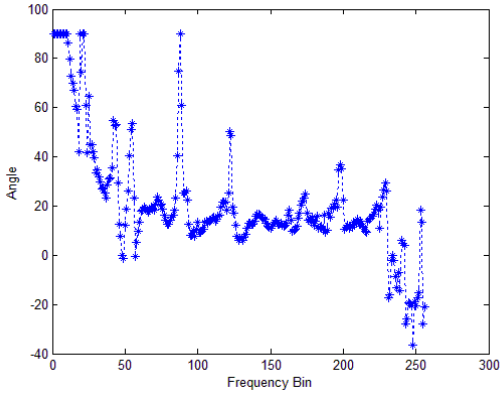


(g) MAE for $\phi = 0^o$

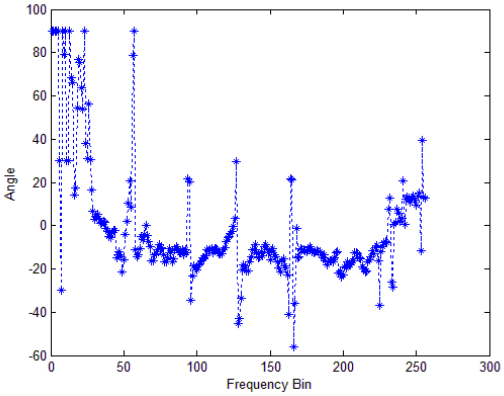


(h) MAE for $\phi = 0^o$

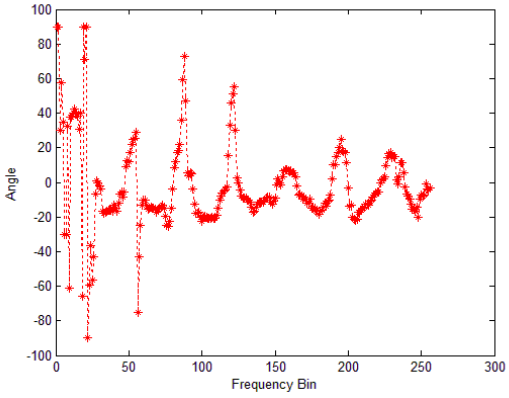
Fig. 41 DOA estimates (in various bins) in the Second-EXP of CASE-A (Real) (Female_1 & Female_2)



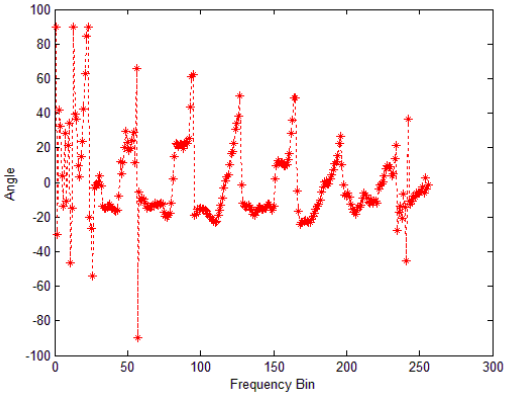
(a) $\theta = 15^o$



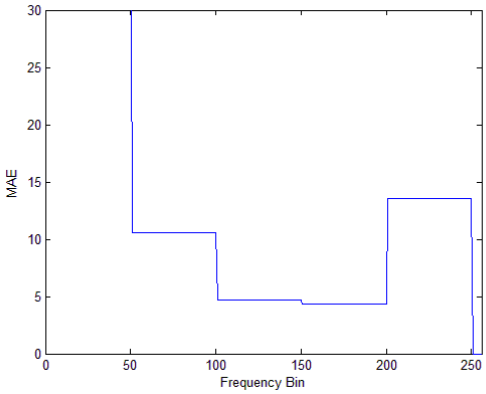
(b) $\theta = -15^o$



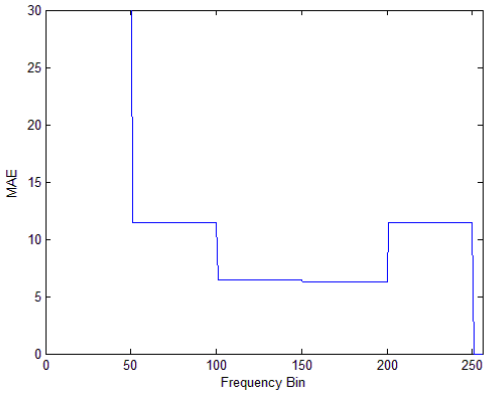
(c) $\phi = 0^o$



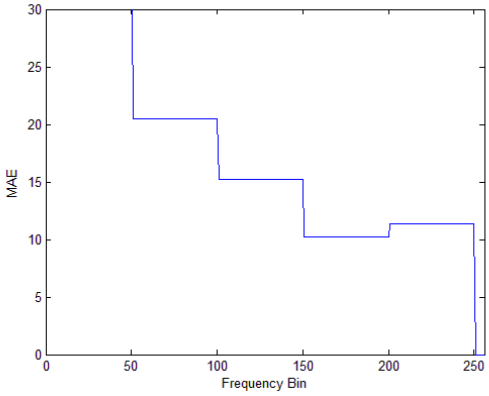
(d) $\phi = 0^o$



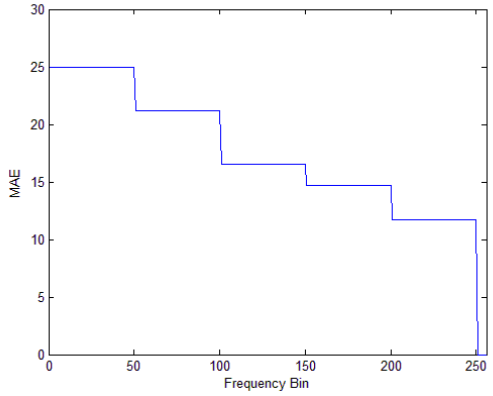
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



(g) MAE for $\phi = 0^o$



(h) MAE for $\phi = 0^o$

Fig. 42 DOA estimates (in various bins) in the Second-EXP of CASE-A (Real)

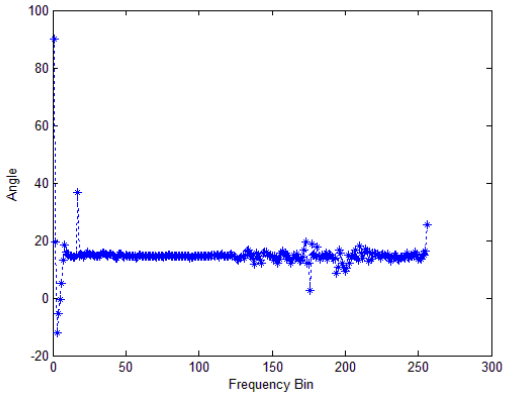
(Male_1 & Male_2)

In CASE-A, we show the First-EXP and the Second-EXP experiment results in Fig. 40~Fig. 42. Fig. 40(a)~(d) show the angle estimated at each frequency bin for Female_1 and Female_2 in the First-EXP. The x-axis represents 256 bins from low frequency to high frequency. The source speeches come from $\theta = 15^o$ and $\theta = -15^o$. For convenience, we always set the elevation angle $\phi = 0^o$. Fig. 40(e)~(h) show the MAE corresponding to Fig. 40(a)~(d). We notice that the estimation errors are high in the First-EXP. In principle, the DOA estimation would be more accurate when there are more sensors. In Fig. 41 and Fig. 42, we show two test sequences with seven sensors. The first test sequence consists of Female_1 and Female_2. The second test sequence consists of Male_1 and Male_2. Fig. 41(e)~(h) show the MAE corresponding to Fig. 41(a)~(d). In Fig. 41(a)~(b), we observe that the median frequency bins have better estimations. In fact, the situation is reasonable. The low frequency has large wavelength. Theoretically, the wavelength should be smaller than the distance between source and sensor; otherwise, the angle (phase shift) cannot be accurately estimated. In Fig. 42, we see a similar trend. Furthermore, there are also high estimation errors in high frequency bins. This is particularly true for the Male_1 and Male_2 test sequence. In general, the male voice seldom includes high frequency components. According to the above discussions, we should avoid using low frequencies and high frequencies in DOA estimation. Because of the high estimation errors on the elevation angles in CASE-A, we no longer consider the estimation of the elevation angles in the rest of this chapter.

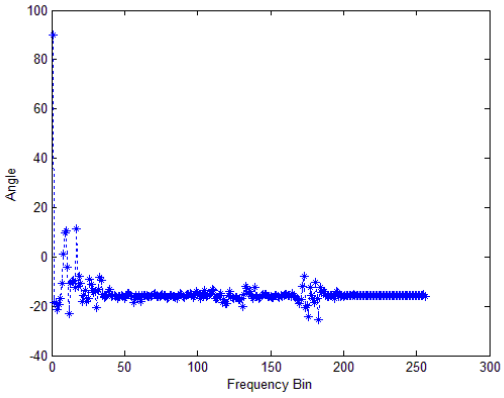
Fig. 43 and Fig. 44 show the test in the SLAB-based simulations with three and seven sensors. Fig. 43 shows the angle estimation at each frequency bin for Female_1 and Female_2. According to these data, the results are very good. In the ideal situations, the ICA-based scheme is quite accurate in DOA estimation. We also notice that the angle estimations have no difference between three and seven sensors. That is, the

recorded signals without noise interference provide accurate estimation. In Fig. 45 and Fig. 46, we observe that the angle estimations with seven sensors have better performance than that with three sensors.

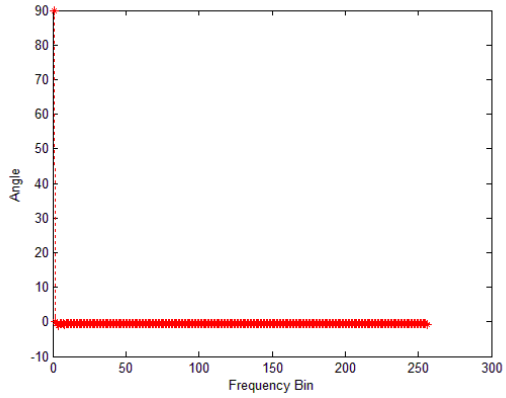
As a summary of the above discussions, firstly, the results of DOA estimation are improved when the number of sensors increases. Secondly, the low frequency bins and the high frequency bins are improper for the purpose of DOA estimation. Therefore, we choose the median frequency bins as our reliable intervals.



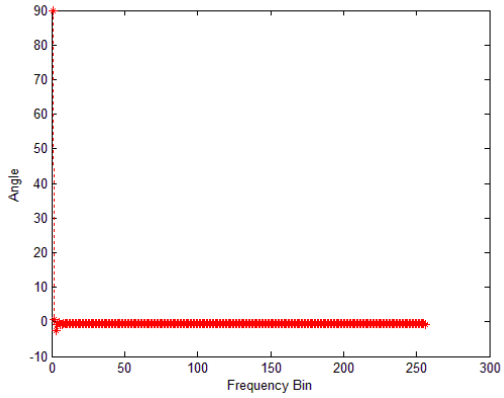
(a) $\theta = 15^o$



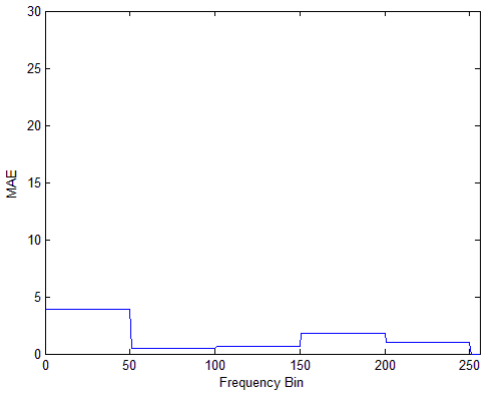
(b) $\theta = -15^o$



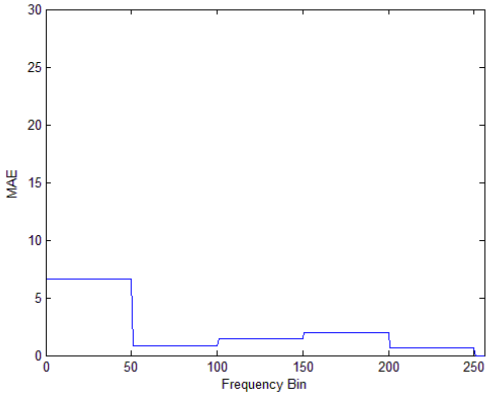
(e) $\phi = 0^o$



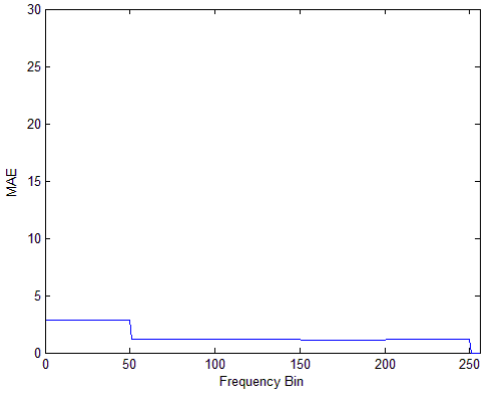
(f) $\phi = 0^o$



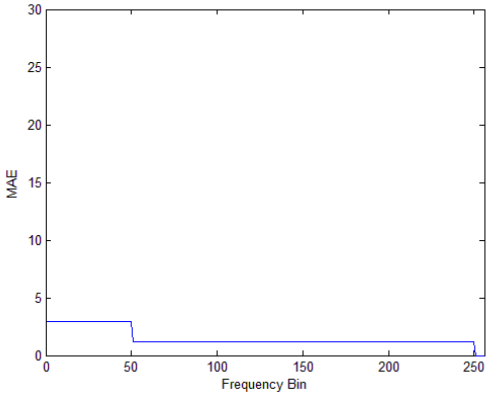
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$



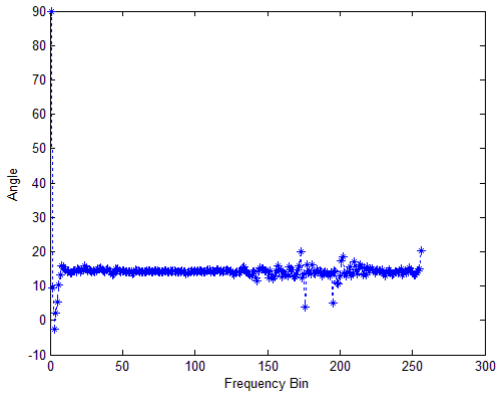
(g) MAE for $\phi = 0^o$



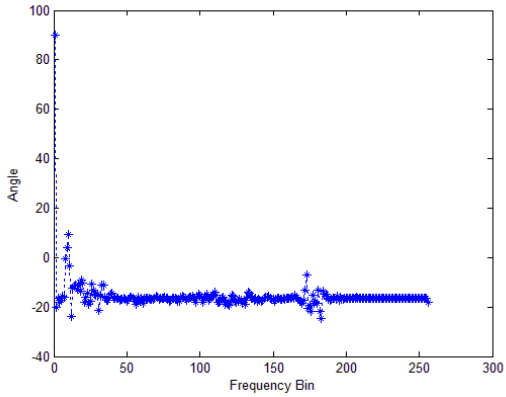
(h) MAE for $\phi = 0^o$

Fig. 43 DOA estimates (in various bins) with 3 MICs in CASE-B.1 (SLAB)

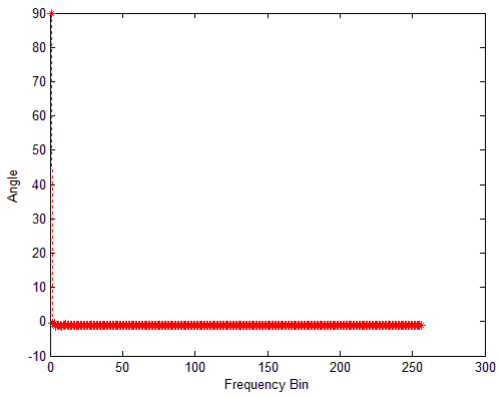
(Female_1 & Female _2)



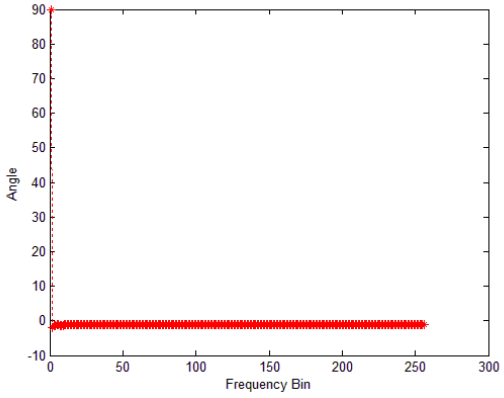
(a) $\theta = 15^o$



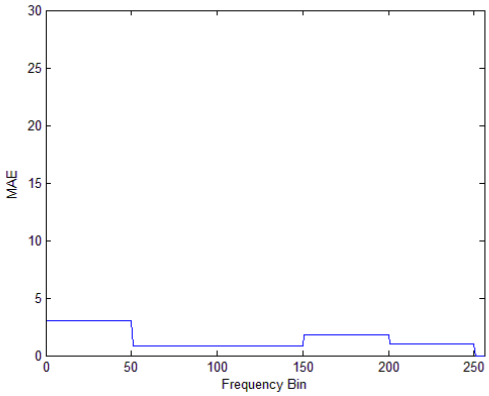
(b) $\theta = -15^o$



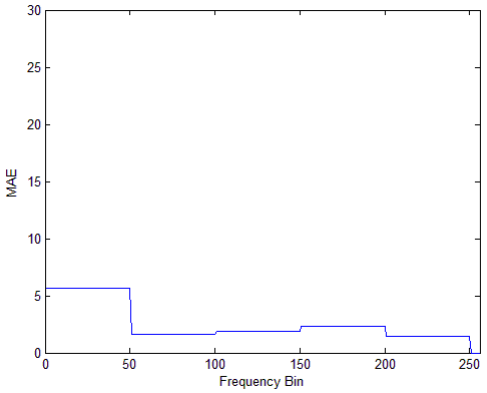
(e) $\phi = 0^o$



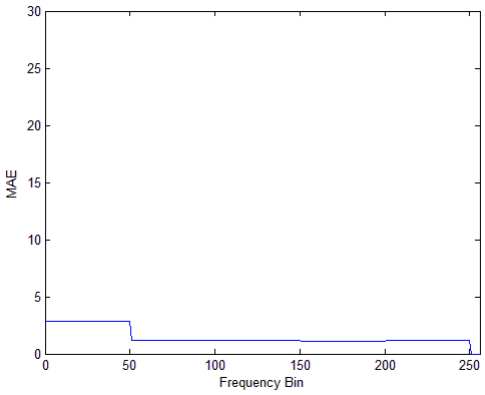
(f) $\phi = 0^o$



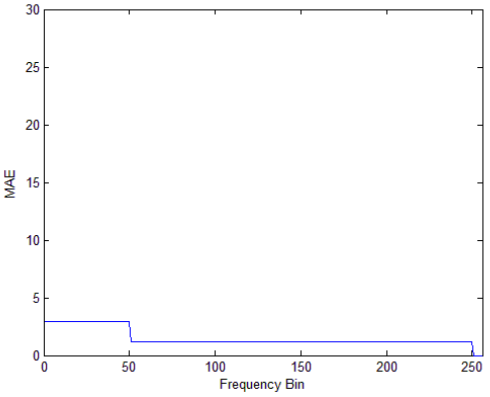
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$



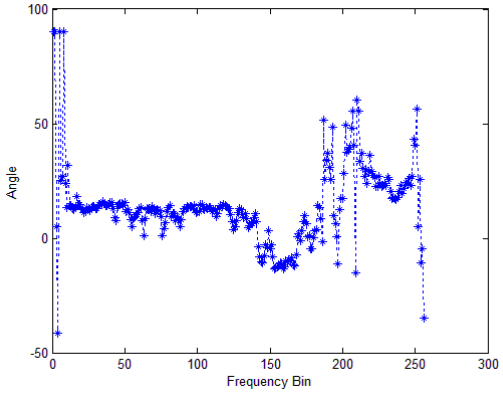
(g) MAE for $\phi = 0^o$



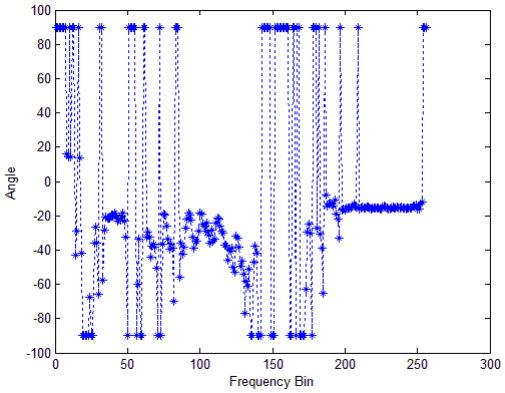
(h) MAE for $\phi = 0^o$

Fig. 44 DOA estimates (in various bins) with 7 MICs in CASE-B.1 (SLAB)

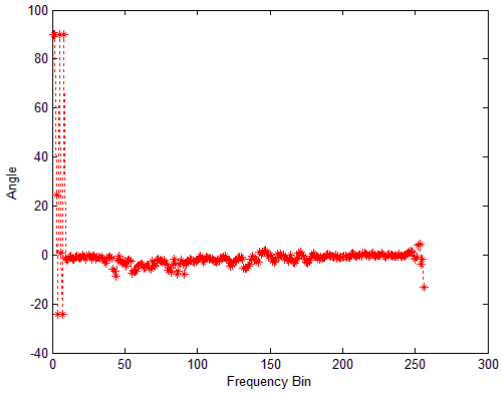
(Female_1 & Female _2)



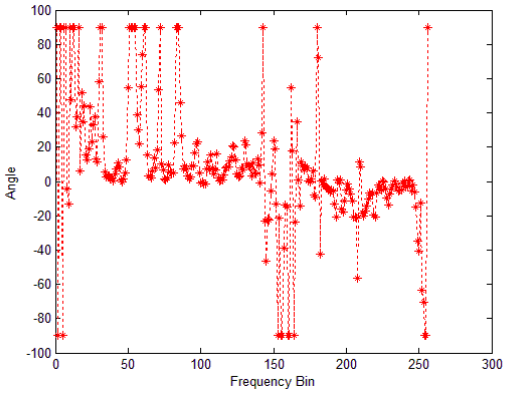
(a) $\theta = 15^o$



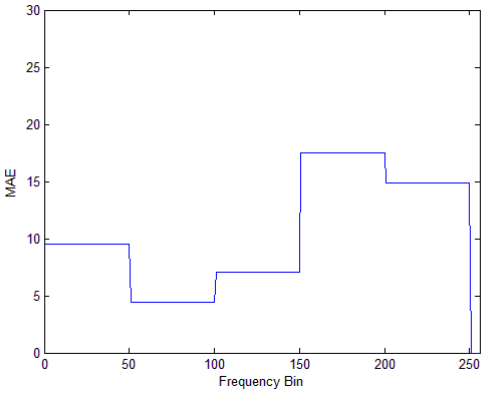
(b) $\theta = -15^o$



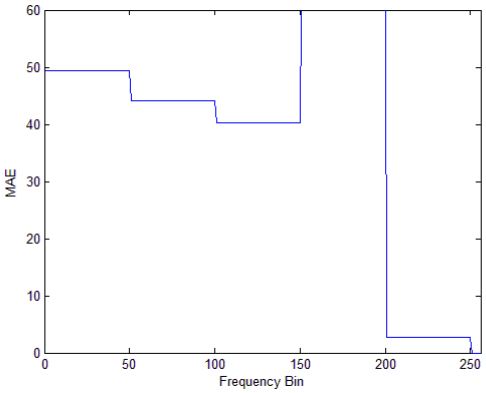
(e) $\phi = 0^o$



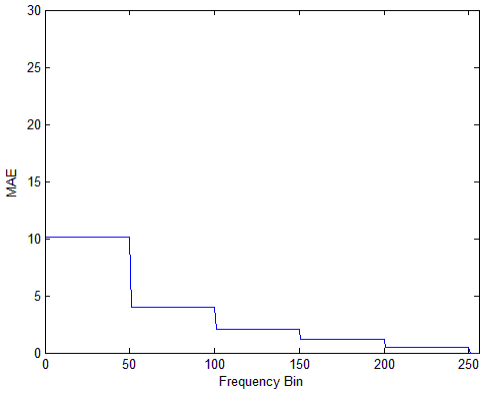
(f) $\phi = 0^o$



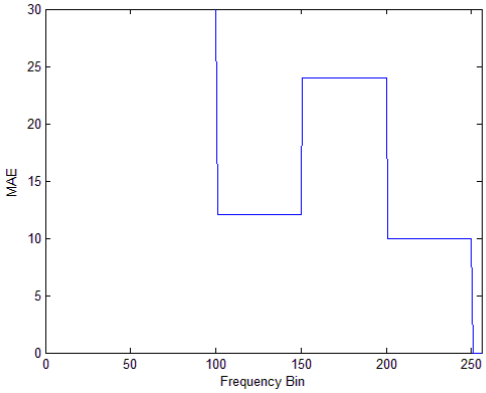
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$



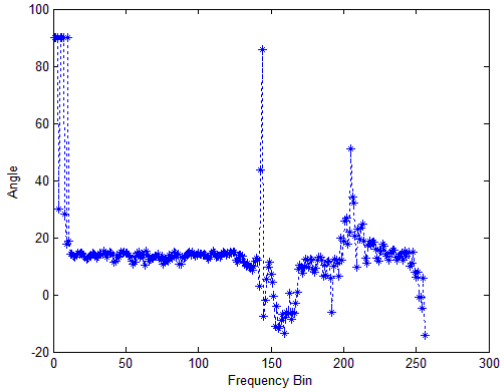
(g) MAE for $\phi = 0^o$



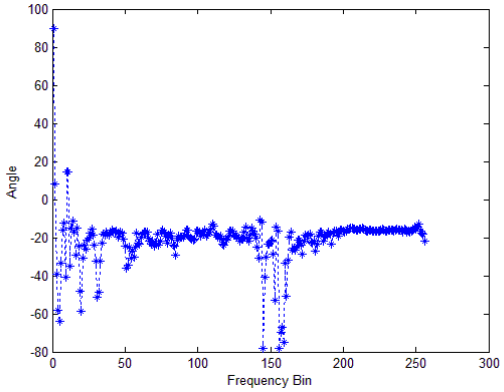
(h) MAE for $\phi = 0^o$

Fig. 45 DOA estimates (in various bins) with 3 MICs in CASE-B.2 (AWGN)

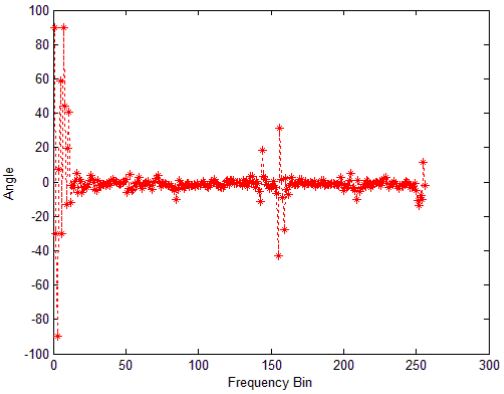
(Female_1 & Female _2)



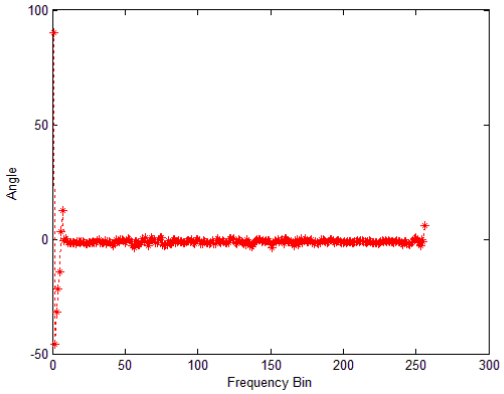
(a) $\theta = 15^o$



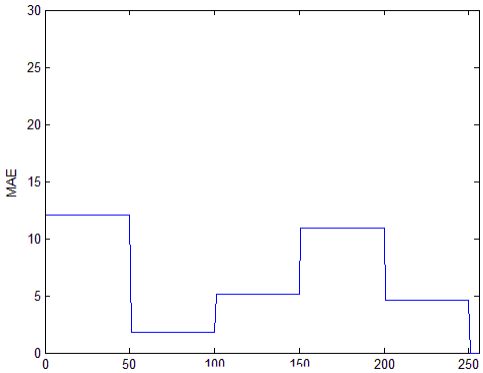
(b) $\theta = -15^o$



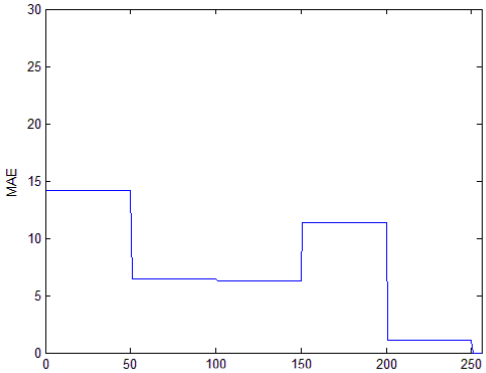
(e) $\phi = 0^o$



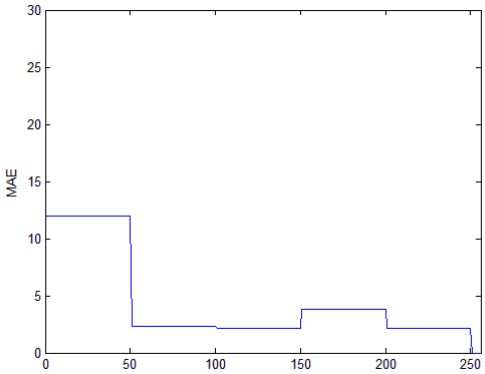
(f) $\phi = 0^o$



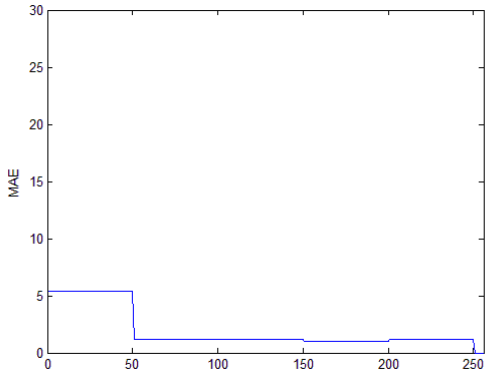
(c) MAE for $\theta = 15^o$



(d) MAE for $\theta = -15^o$



(g) MAE for $\phi = 0^o$



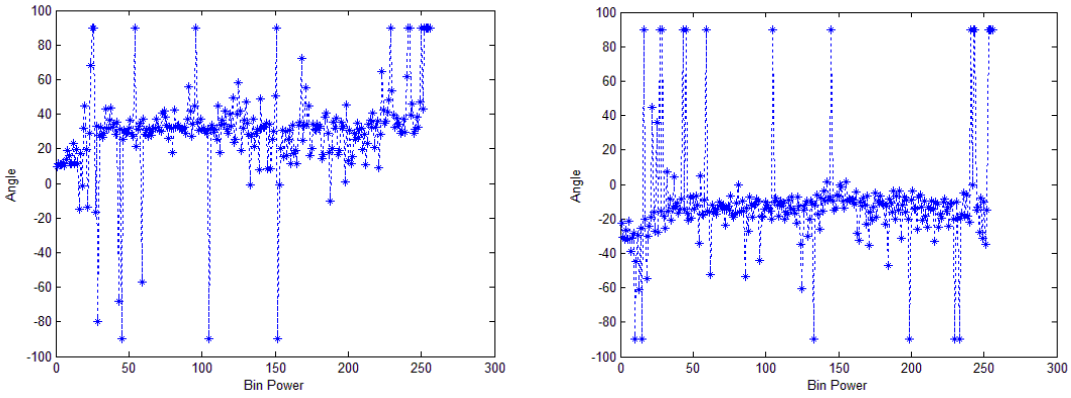
(h) MAE for $\phi = 0^o$

Fig. 46 DOA estimates (in various bins) with 7 MICs in CASE-B.2 (AWGN)

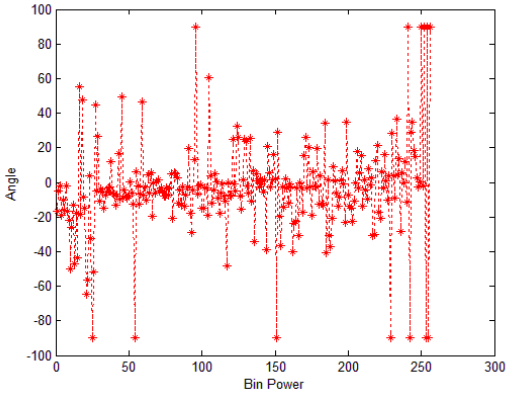
(Female_1 & Female _2)

5.1.2 Effect of Power in DOA Estimation

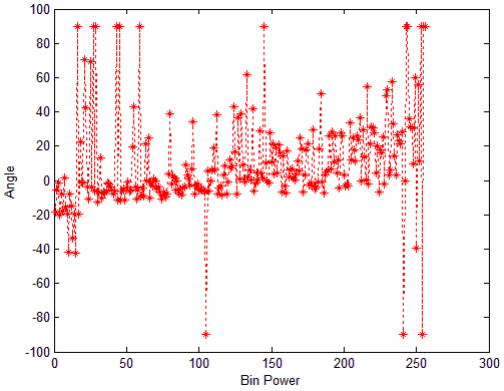
In this section, we focus on the effect of the power of bins in DOA estimation. We randomly pick up a mixture signal recorded from a microphone, and we convert the time-domain signal into frequency-domain by STFT. The power is estimated at the each frequency bin. Then, we sort the bins according to the power in each bin. That is, the power of bins is in the decreasing order. We measure the performance (accuracy) by MAE. We set the window size to 512 samples, the data input size to four seconds (32000 sampling points) and the distance between source and sensor to 1.5M. We also divide the power of frequency bins into five intervals. Each interval contains 50 indexes, and the last 16 indexes are discarded since their power is typically very low. For convenience, we abbreviate the power of a frequency bin as PFB.



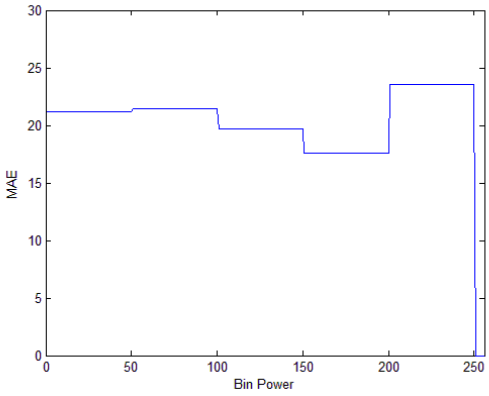
(a) $\theta = 15^o$ (b) $\theta = -15^o$



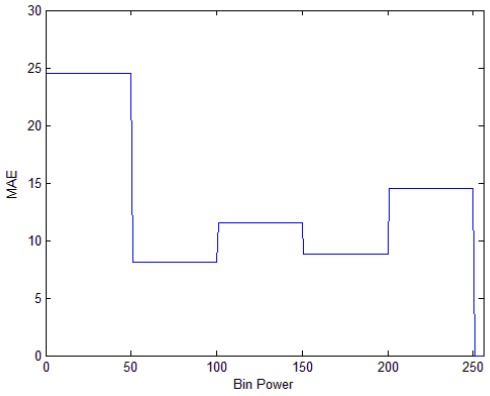
(c) $\phi = 0^o$



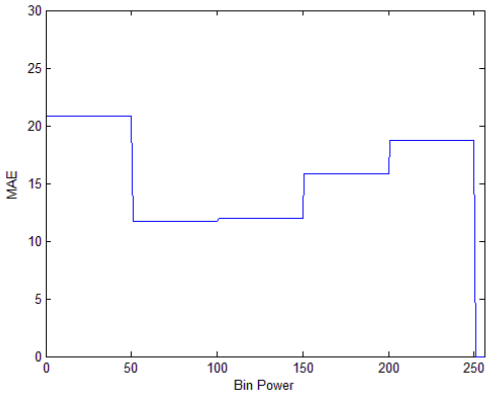
(d) $\phi = 0^o$



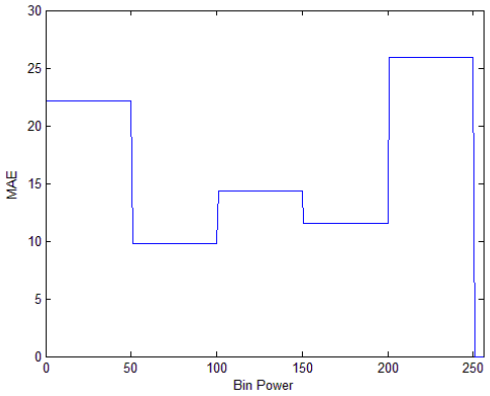
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



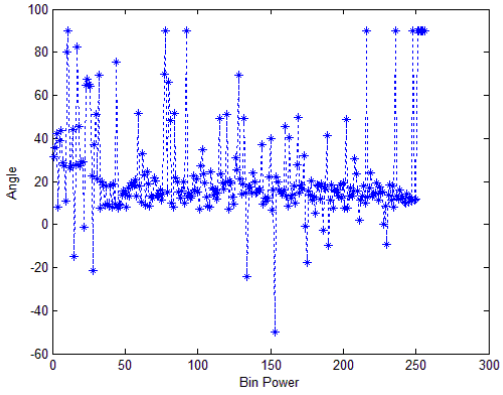
(g) MAE for $\phi = 0^o$



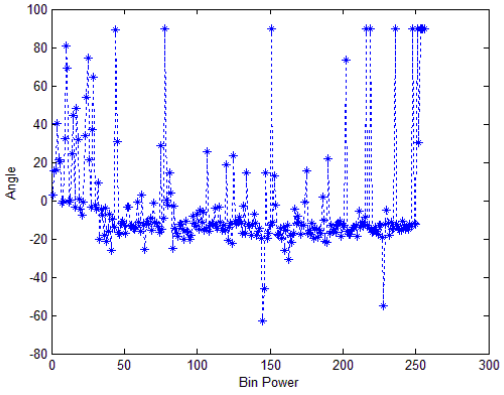
(h) MAE for $\phi = 0^o$

Fig. 47 DOA estimates (power sorted bins) in the First-EXP of CASE-A (Real)

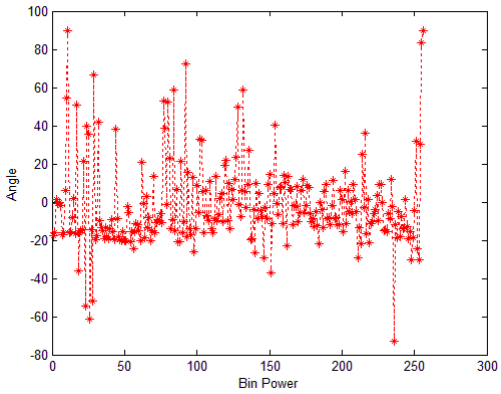
(Female_1 & Female_2)



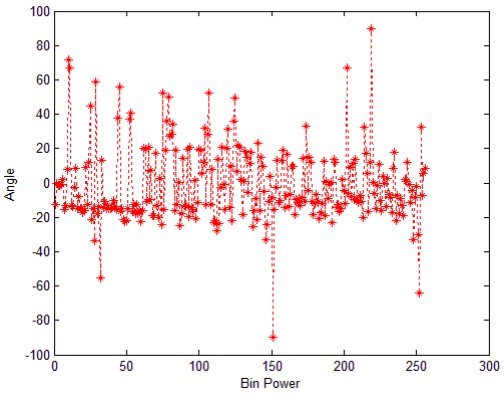
(a) $\theta = 15^o$



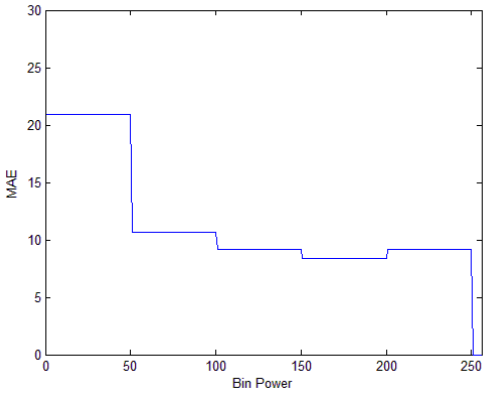
(b) $\theta = -15^o$



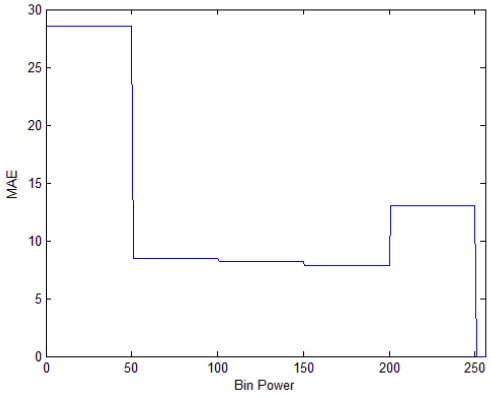
(c) $\phi = 0^o$



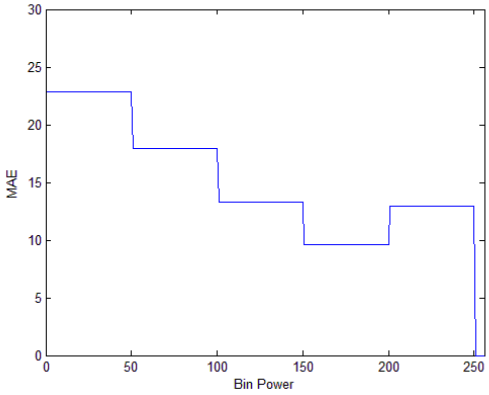
(d) $\phi = 0^o$



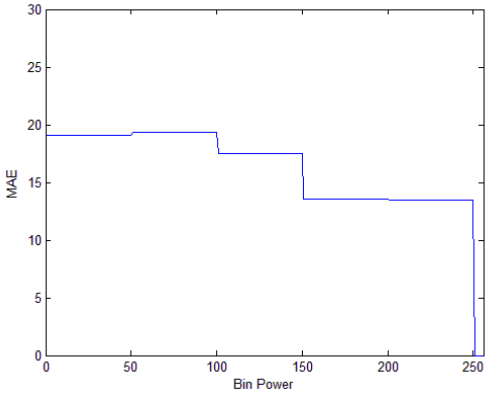
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$

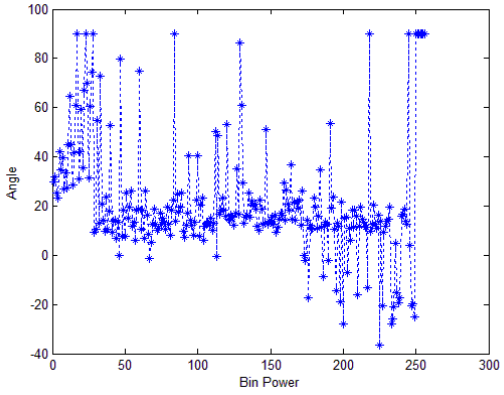


(g) MAE for $\phi = 0^o$

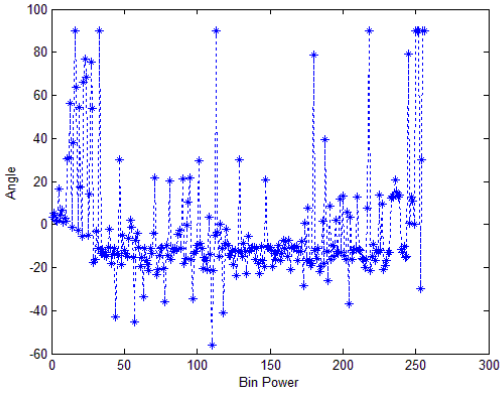


(h) MAE for $\phi = 0^o$

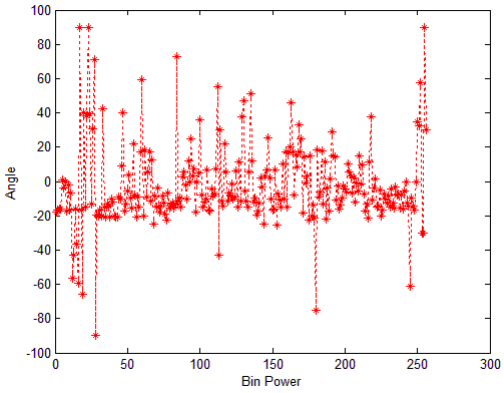
Fig. 48 DOA estimates (power sorted bins) in the Second-EXP of CASE-A (Real) (Female_1 & Female_2)



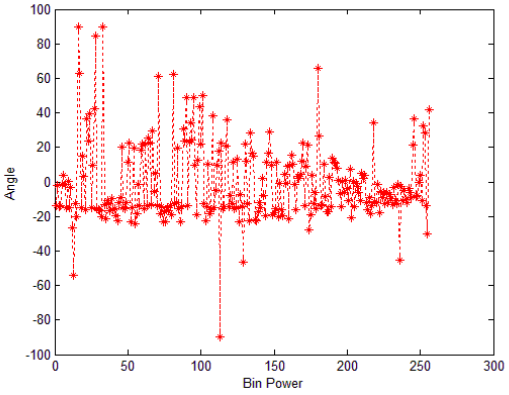
(a) $\theta = 15^o$



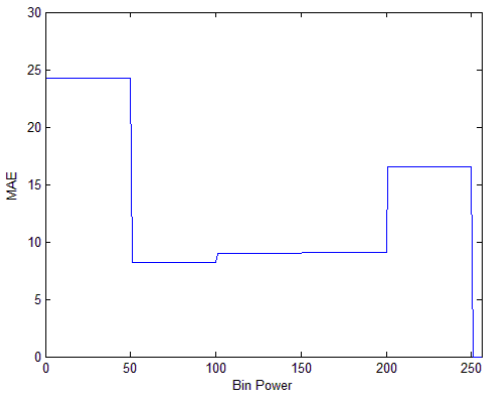
(b) $\theta = -15^o$



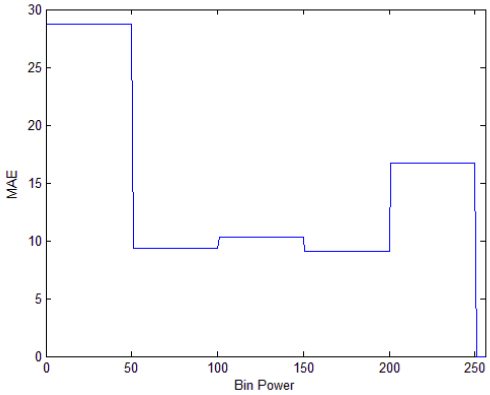
(c) $\phi = 0^o$



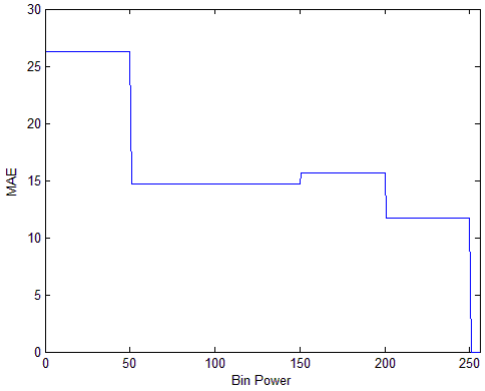
(d) $\phi = 0^o$



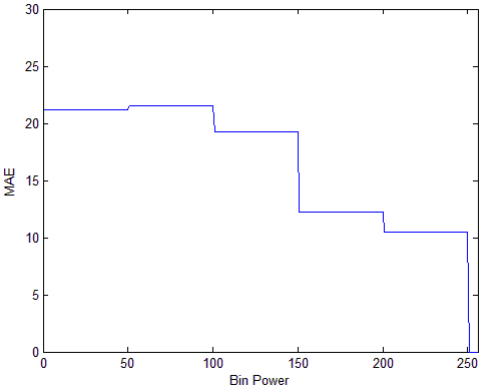
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



(g) MAE for $\phi = 0^o$



(h) MAE for $\phi = 0^o$

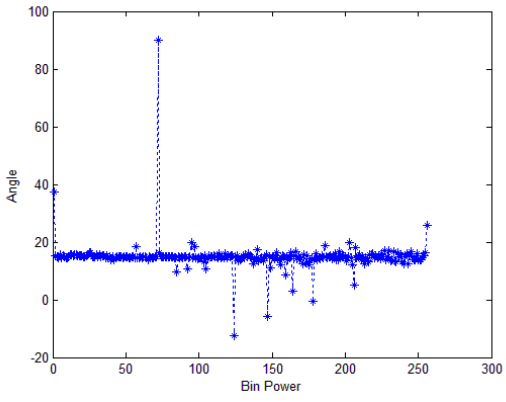
Fig. 49 DOA estimates (power sorted bins) in the Second-EXP of CASE-A (Real) (Male_1 & Male_2)

In CASE-A, we show the First-EXP and the Second-EXP experimental results in Fig. 47~Fig. 49. Fig. 47(a)~(d) show that the angle is estimated at each PFB for Female_1 and Female_2 in the First-EXP, and the sources are located at $\theta = 15^o$ and $\theta = -15^o$. The x-axis represents 256 indexes from the high PFB to the low PFB. Because of the poor performance in the First-EXP where three sensors were used, we will not discuss it in detail. In Fig. 48 and Fig. 49, we show two test sequences with seven sensors. The first test sequence consists of Female_1 and Female_2. The second

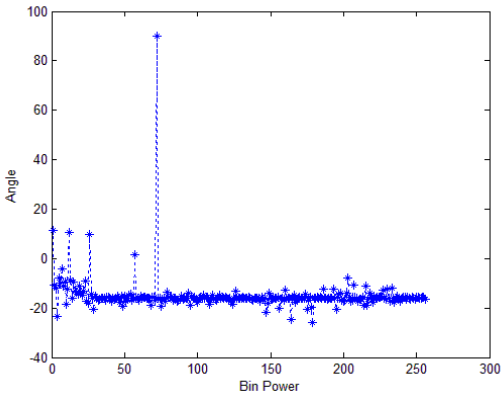
test sequence consists of Male_1 and Male_2. Fig. 48(e)~(h) show the MAE corresponding to Fig. 48(a)~(d). In Fig. 48(a)~(b), we find that the median PFB have better estimates. The PFB with the small indexes are often the low frequency components. In the preceding section, we discussed that low frequency components cannot provide accurate estimates. In Fig. 49, we notice a similar trend. We also find that the estimation errors of the high index PFB are high. It is quite well-known that the estimation errors increase when the SNR decreases. Clearly low power signals are unreliable in estimation. In conclusion, we should avoid using the lower index PFB and the high index PFB to estimate DOA.

Fig. 50 and Fig. 51 show the tests in the SLAB-based simulations with three and seven sensors. According to these figures, the tests have the good performance in almost all PFB. We also notice that the angle predictions have no difference between three and seven sensors. In Fig. 52 and Fig. 53, we observe that the angle estimations with seven sensors have the better performance than that with three sensors when there is a noise.

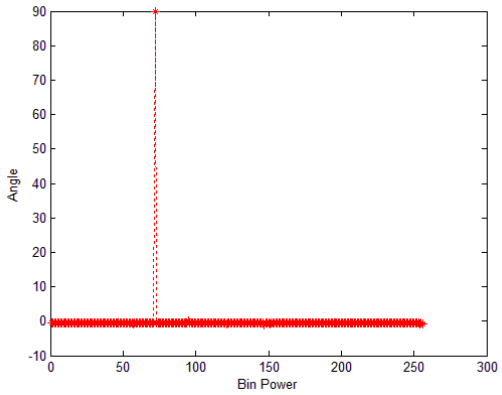
Based on these data, we conclude that the low power components are not reliable and more sensors provide better results for received signals containing noises (real cases).



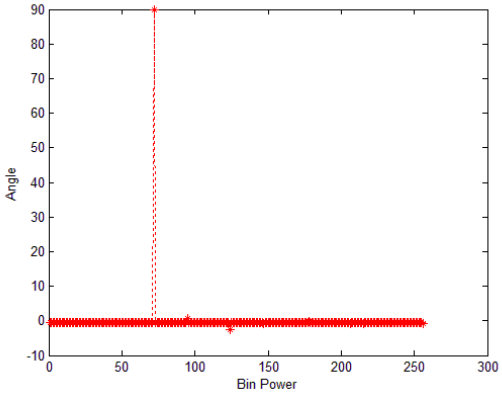
(a) $\theta = 15^o$



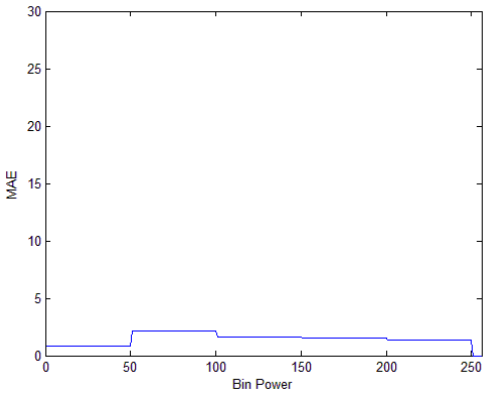
(b) $\theta = -15^o$



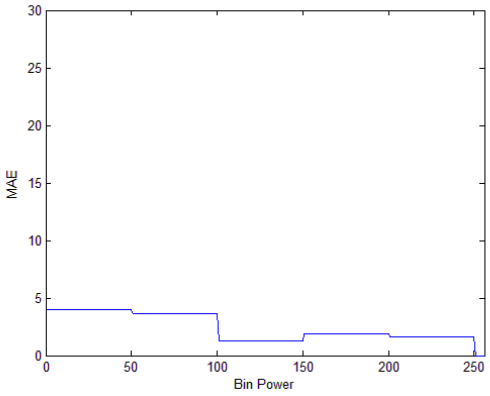
(c) $\phi = 0^o$



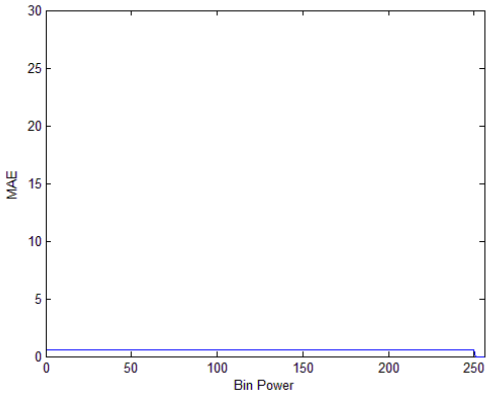
(d) $\phi = 0^o$



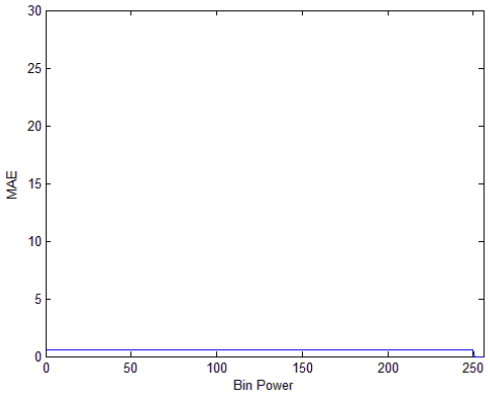
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



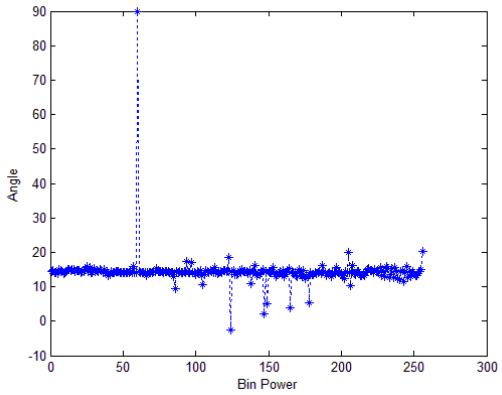
(g) MAE for $\phi = 0^o$



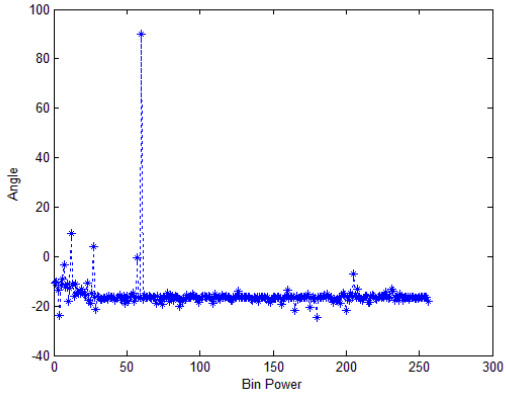
(h) MAE for $\phi = 0^o$

Fig. 50 DOA estimates (power sorted bins) with 3 MICs in CASE-B.1 (SLAB)

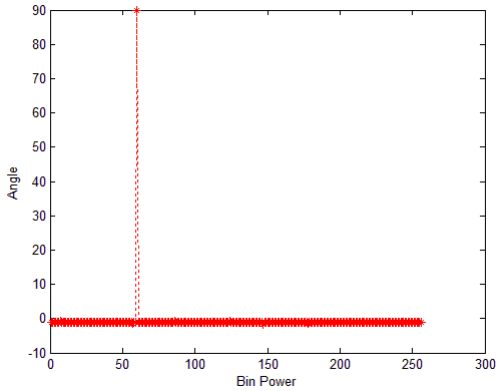
(Female_1 & Female _2)



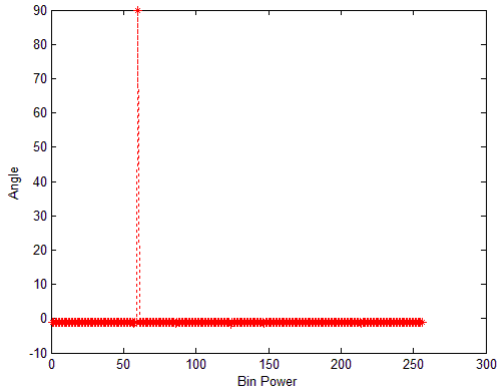
(a) $\theta = 15^o$



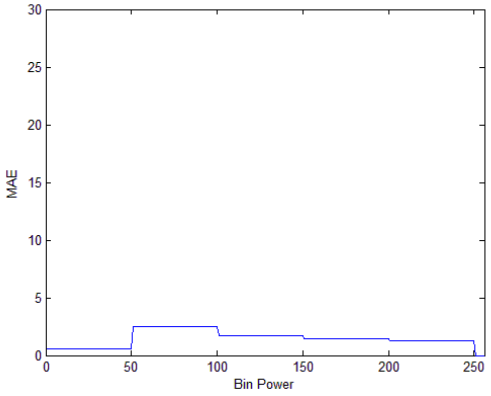
(b) $\theta = -15^o$



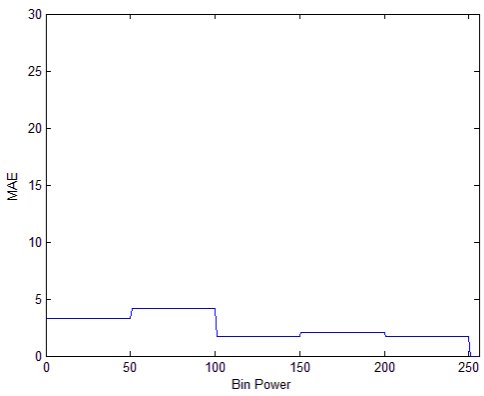
(c) $\phi = 0^o$



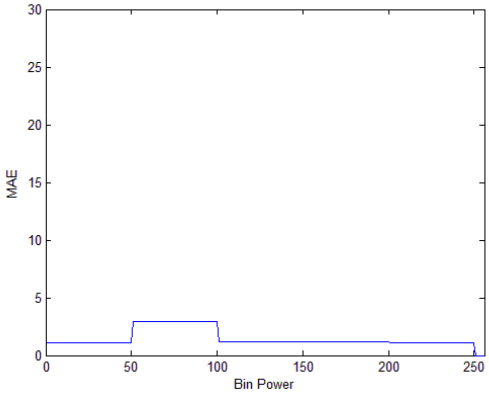
(d) $\phi = 0^o$



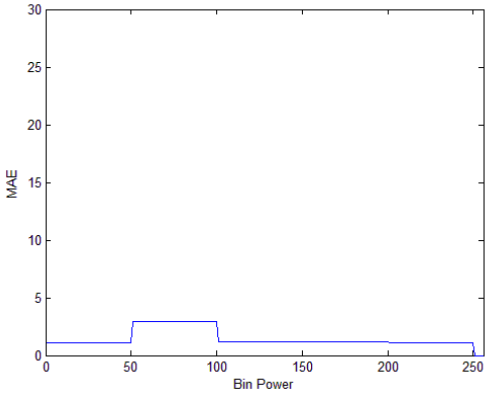
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$

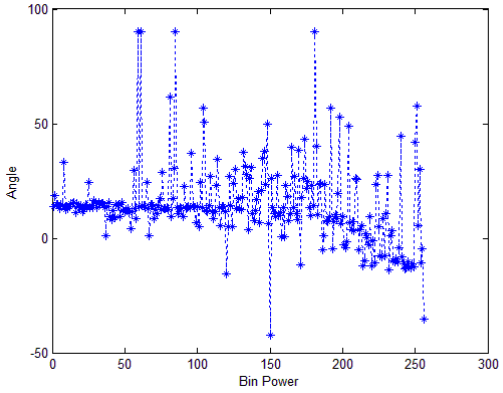


(g) MAE for $\phi = 0^o$

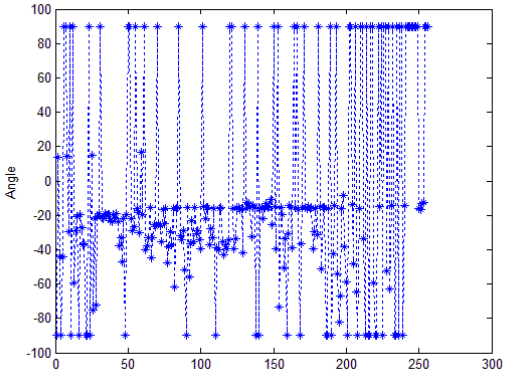


(h) MAE for $\phi = 0^o$

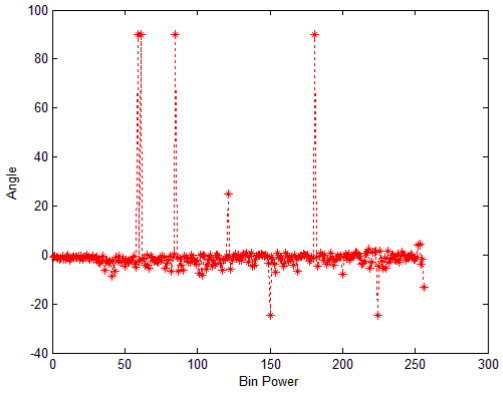
Fig. 51 DOA estimates (power sorted bins) with 7 MICs in CASE-B.1 (SLAB) (Female_1 & Female _2)



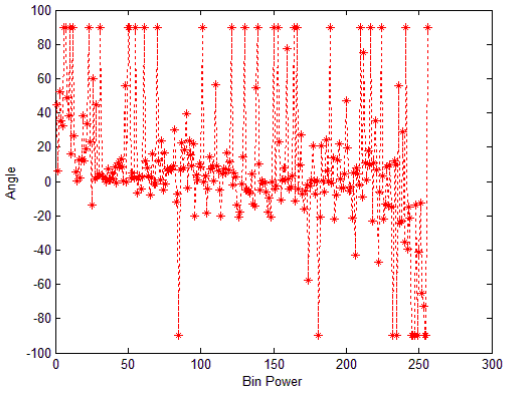
(a) $\theta = 15^o$



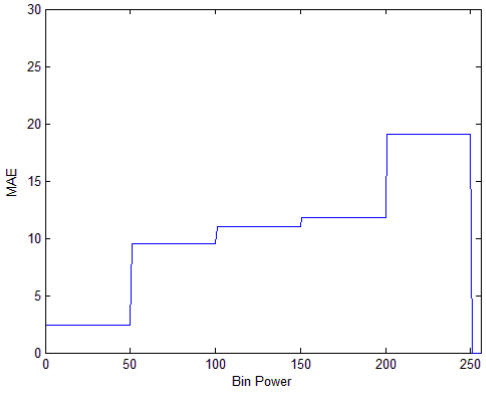
(b) $\theta = -15^o$



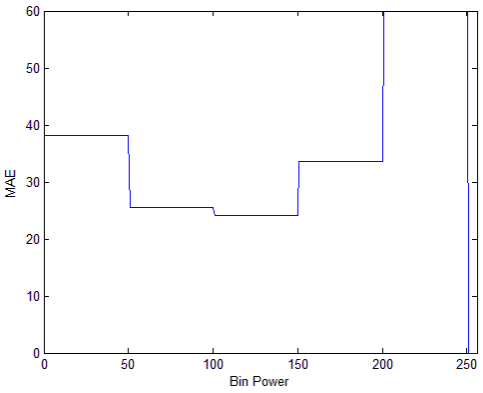
(c) $\phi = 0^o$



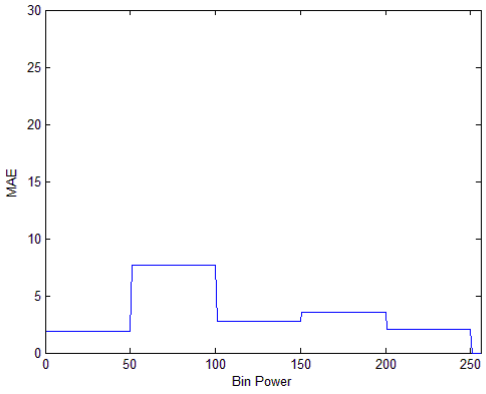
(d) $\phi = 0^o$



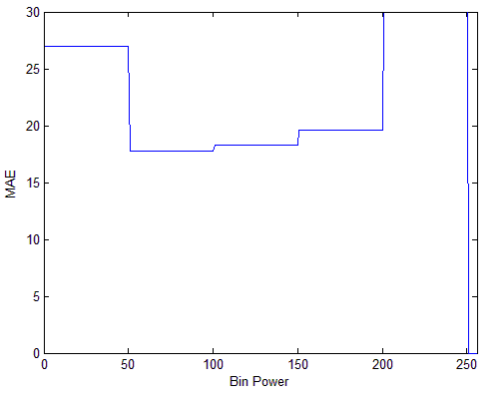
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



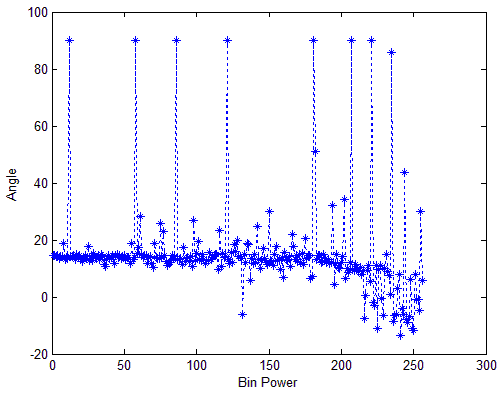
(g) MAE for $\phi = 0^o$



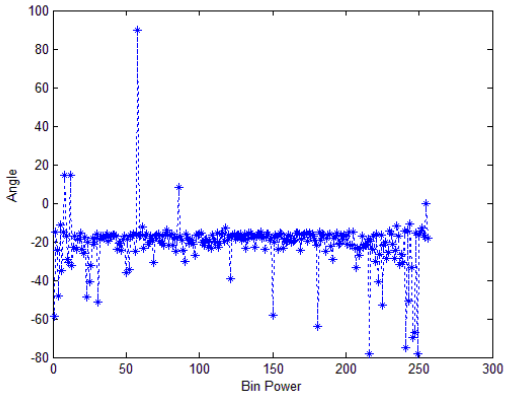
(h) MAE for $\phi = 0^o$

Fig. 52 DOA estimates (power sorted bins) with 3 MICs in CASE-B.2 (AWGN)

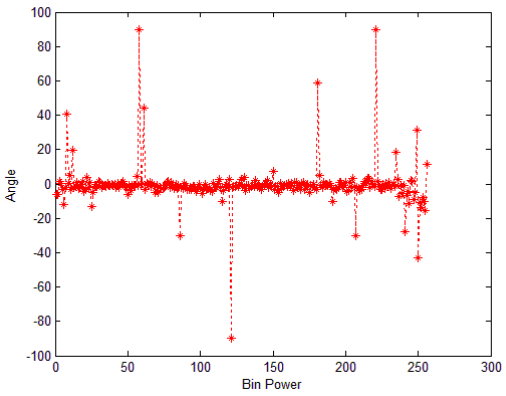
(Female_1 & Female _2)



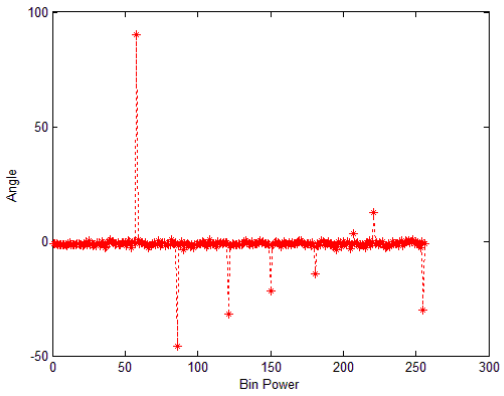
(a) $\theta = 15^o$



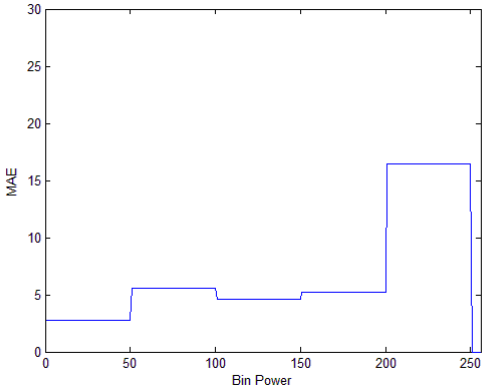
(b) $\theta = -15^o$



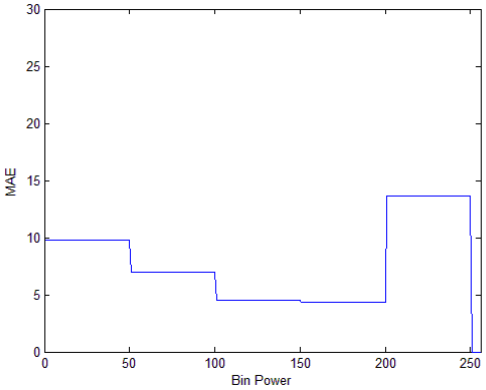
(c) $\phi = 0^o$



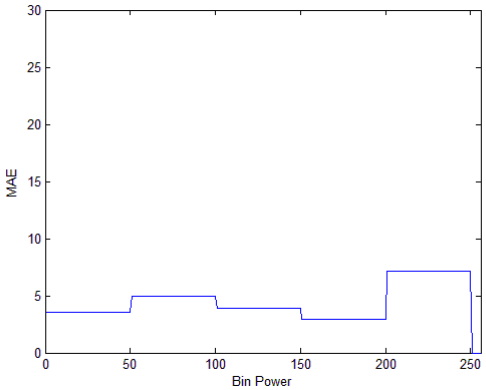
(d) $\phi = 0^o$



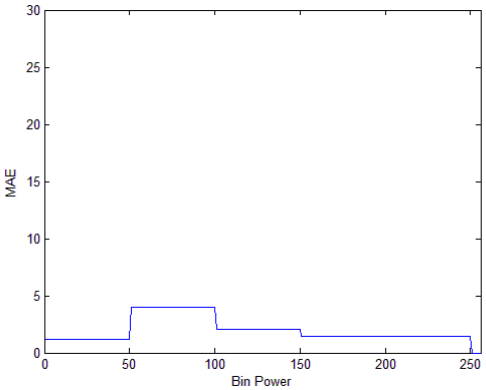
(e) MAE for $\theta = 15^o$



(f) MAE for $\theta = -15^o$



(g) MAE for $\phi = 0^o$



(h) MAE for $\phi = 0^o$

Fig. 53 DOA estimates (power sorted bins) with 7 MICs in CASE-B.2 (AWGN)

(Female_1 & Female _2)

5.1.3 Confidence Region

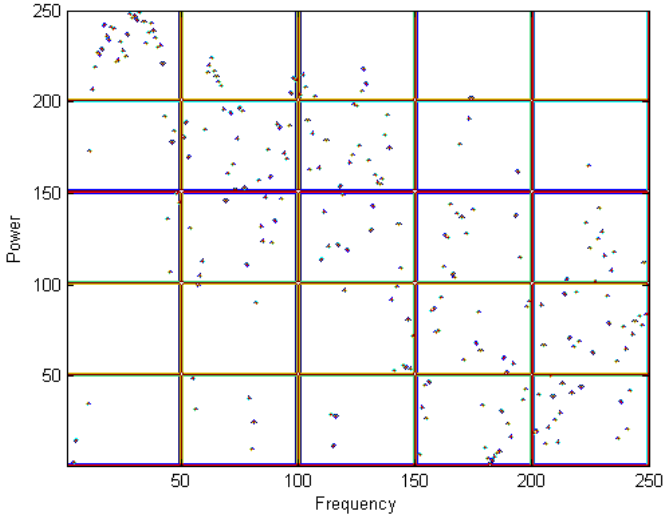
In this section, we discuss the find selections of reliable region; that is, the region provides good performance of DOA estimates. Again, we measure performance by the mean square error (MSE). The definition of MSE is described as below:

$$MSE(\theta) = E\left[\left(\theta - \theta\right)^2\right]$$

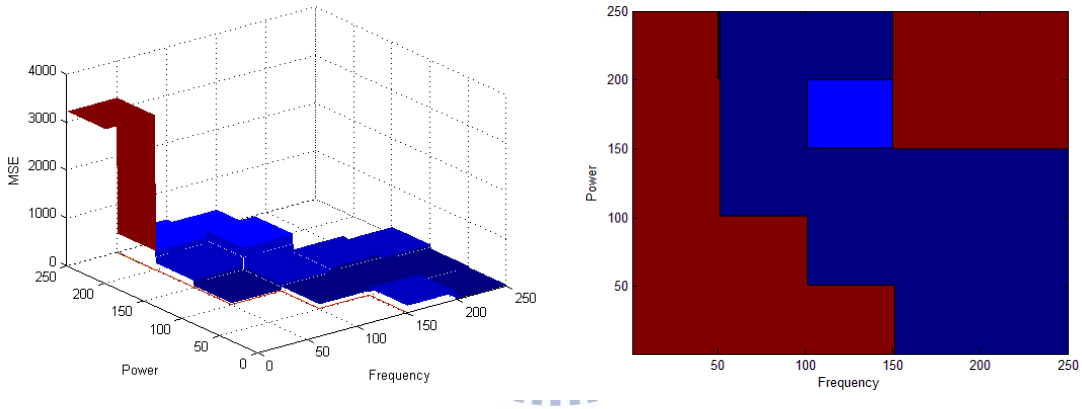
where θ denotes the angle predictions and θ denotes the true outcome.

Here, we integrate two observations we have concluded in the preceding two sections. For example, Fig. 54(a) shows the constellation with one axis in frequency and the other axis in PFB, and the entire domain is divided into 25 areas. We try to identify the confidence region in this plot. Fig. 54(b)~(c) show that the MSE of the DOA estimate corresponding to the constellation. The x-axis represents the frequency bins in increasing order. And, the y-axis represents the PFB in increasing order. Fig. 54 and Fig. 55 show two cases in the Second-EXP. We notice that the MSEs are high at high PFB and low frequency bins. On the other hand, the MSEs are high at low PFB and high frequency bins. We have discussed these trends already in preceding sections. Fig. 56 and Fig. 57 show that the low MSEs are low in the SLAB-based simulations because there is no noise. Fig. 58 and Fig. 59 show the confidence regions in the CASE-B.2.

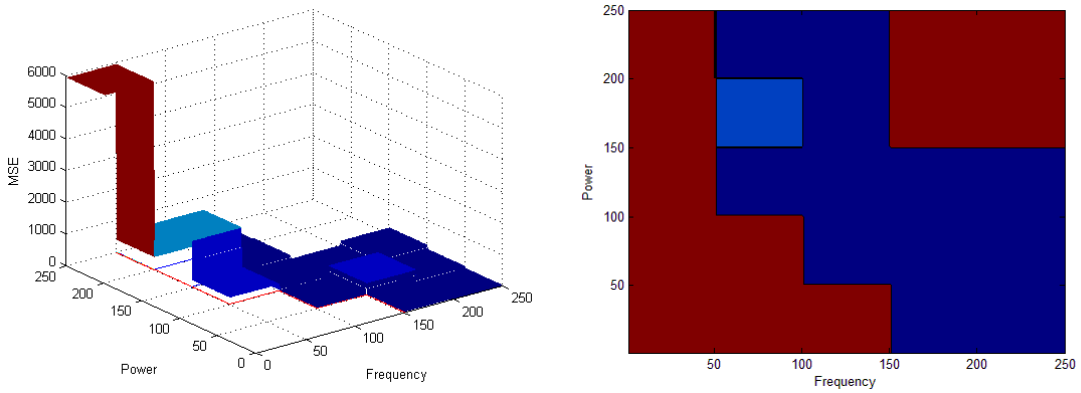
According to these data, we finally select the confidence regions that have the frequency bin ranging from 50 to 200 and the PFB ranging from 50 to 200 (512-size window).



(a) Constellation



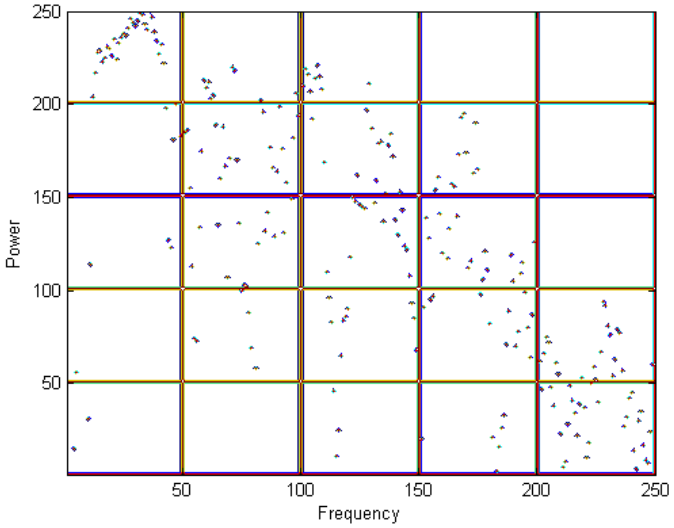
(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



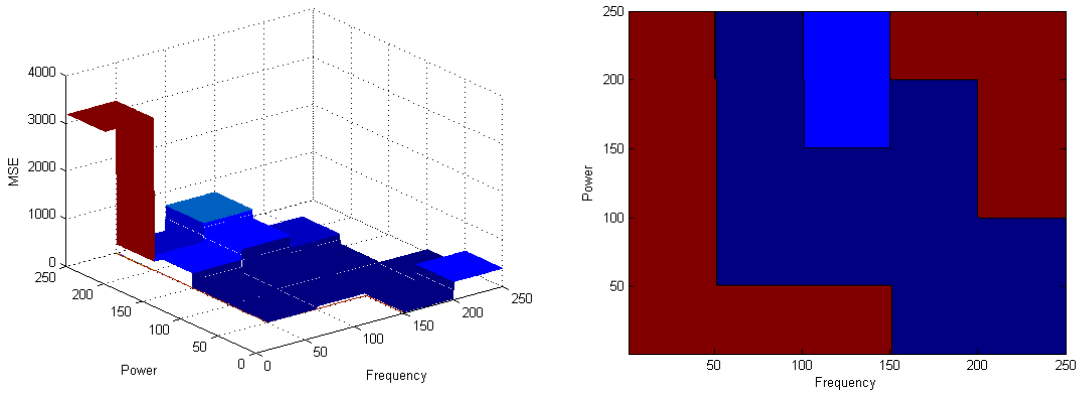
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 54 The Second-EXP in CASE-A (Real)

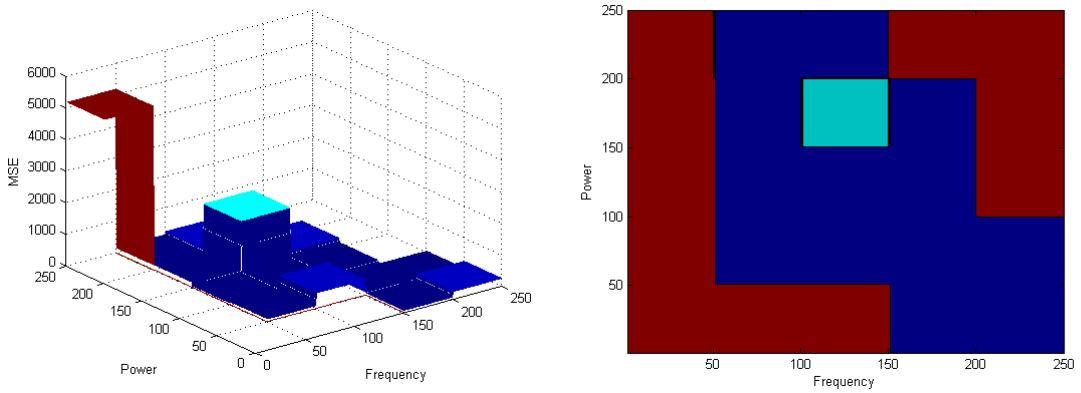
(Female_1 & Female_2)



(a) Constellation



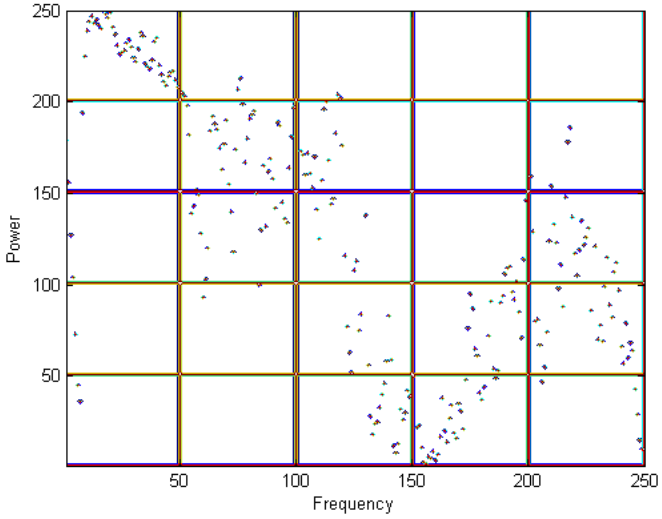
(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



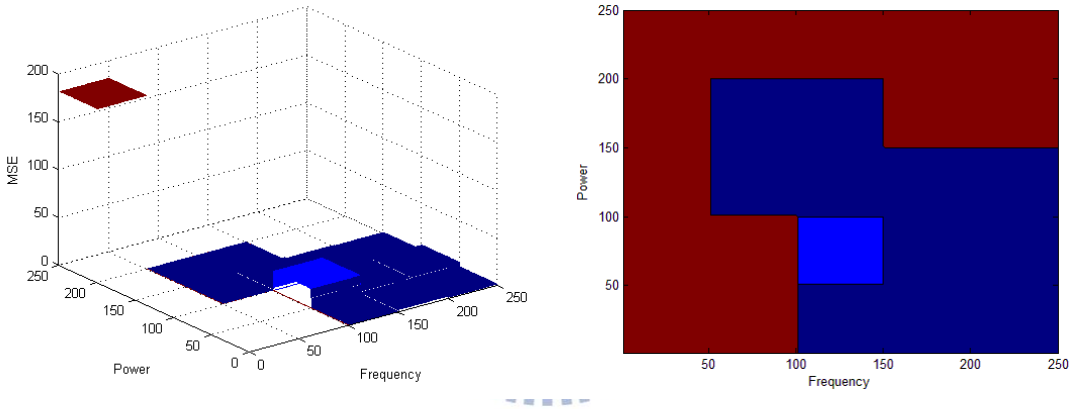
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 55 The Second-EXP in CASE-A (Real)

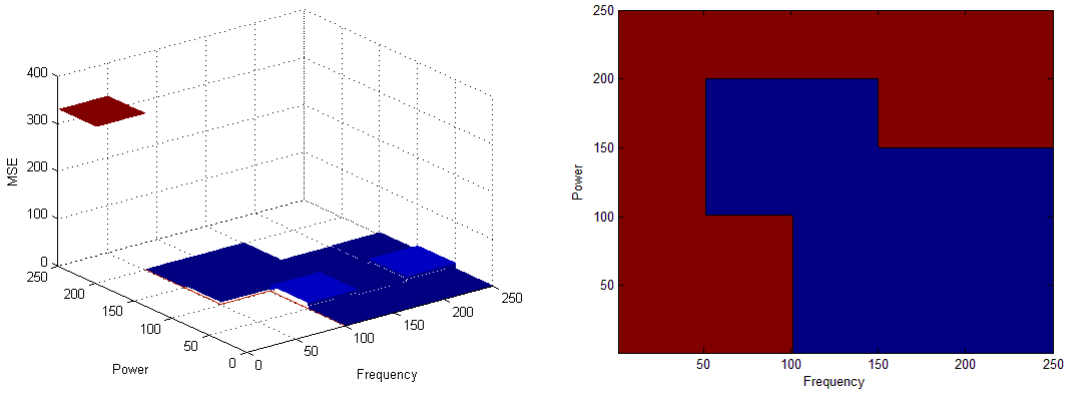
(Male_1 & Male_2)



(a) Constellation



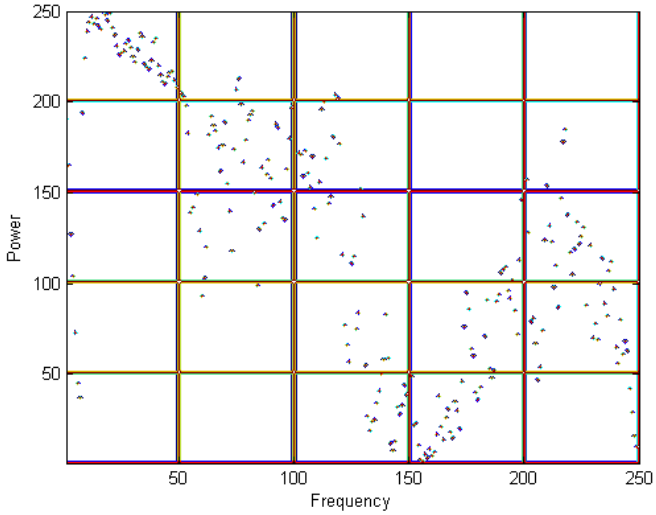
(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



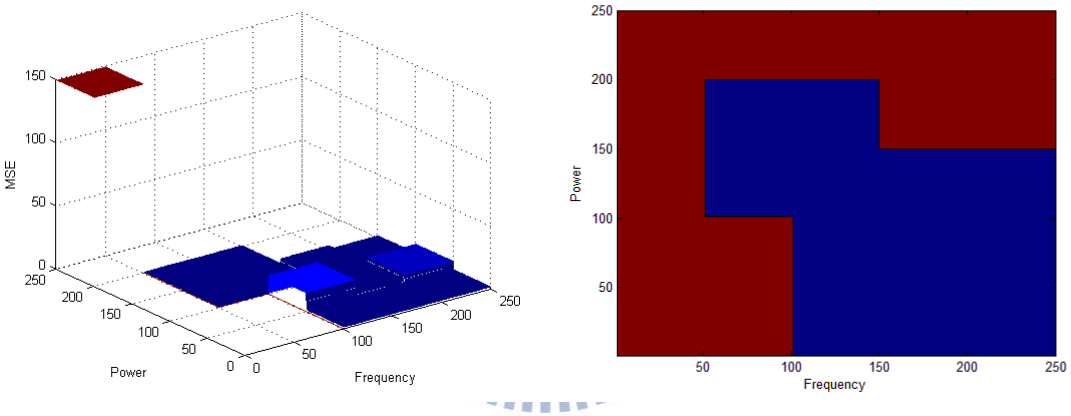
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 56 Three MICs in CASE-B.1 (SLAB)

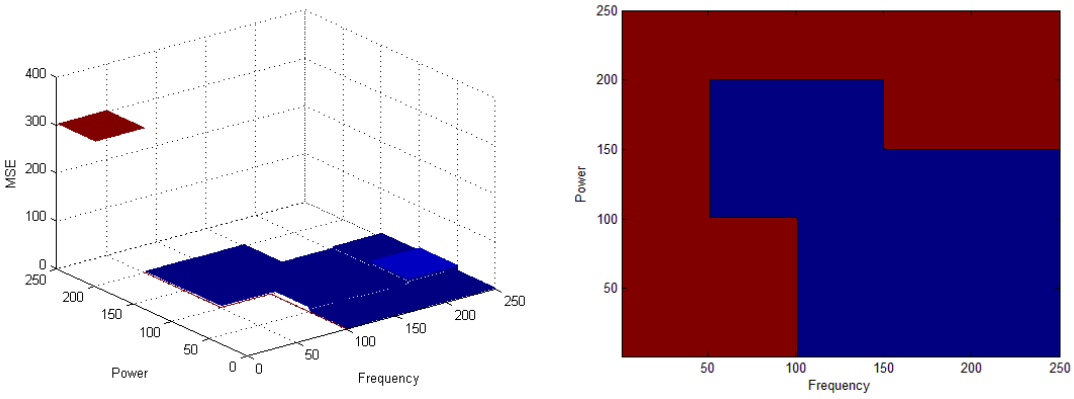
(Female_1 & Female _2)



(a) Constellation



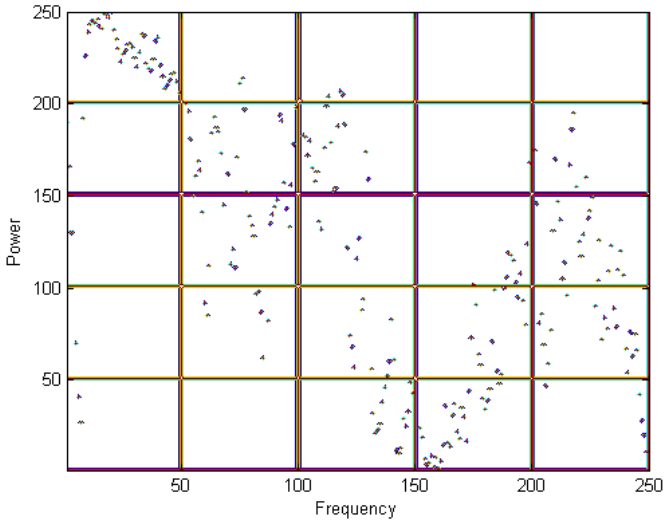
(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



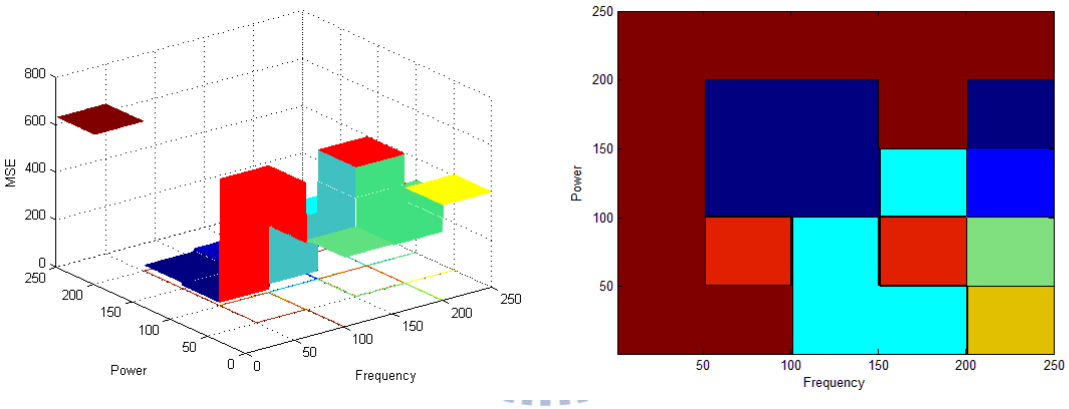
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 57 Seven MICs in CASE-B.1 (SLAB)

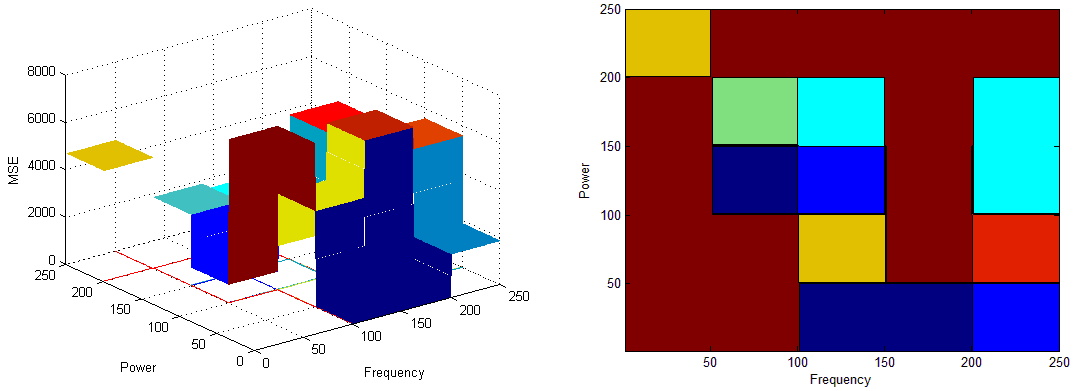
(Female_1 & Female _2)



(a) Constellation



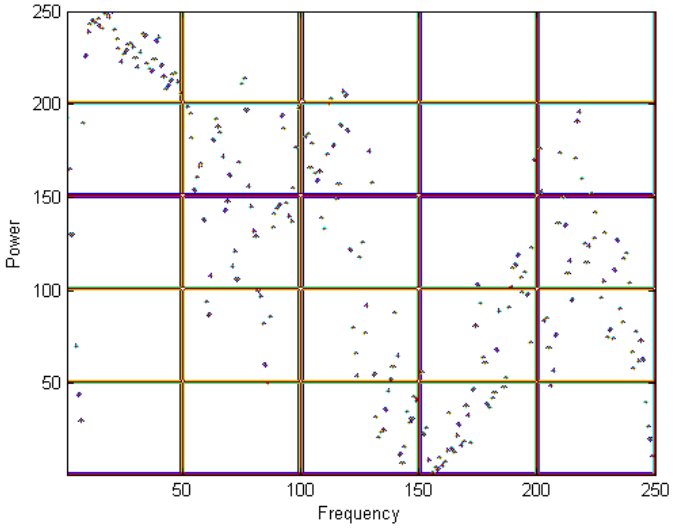
(b) MSE at $\theta = 15^o$ and MSE Contour at with $\theta = 15^o$



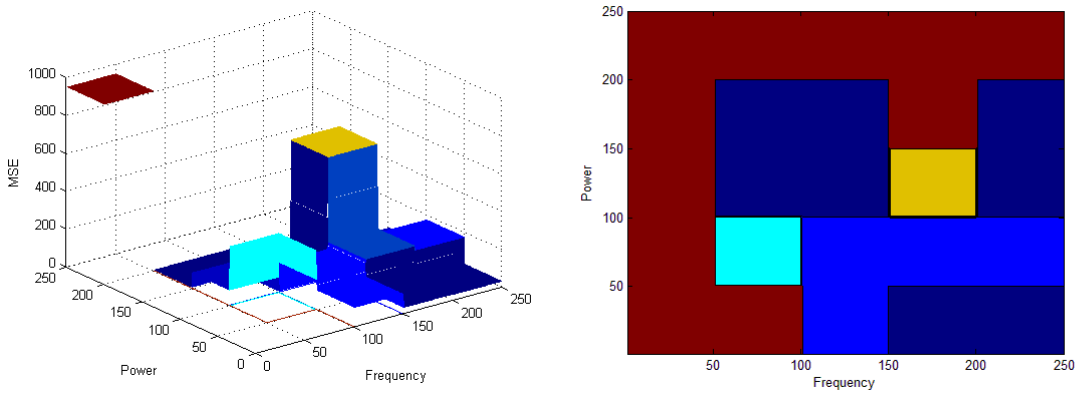
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 58 Simulation of the First-EXP in CASE-B.2 (AWGN)

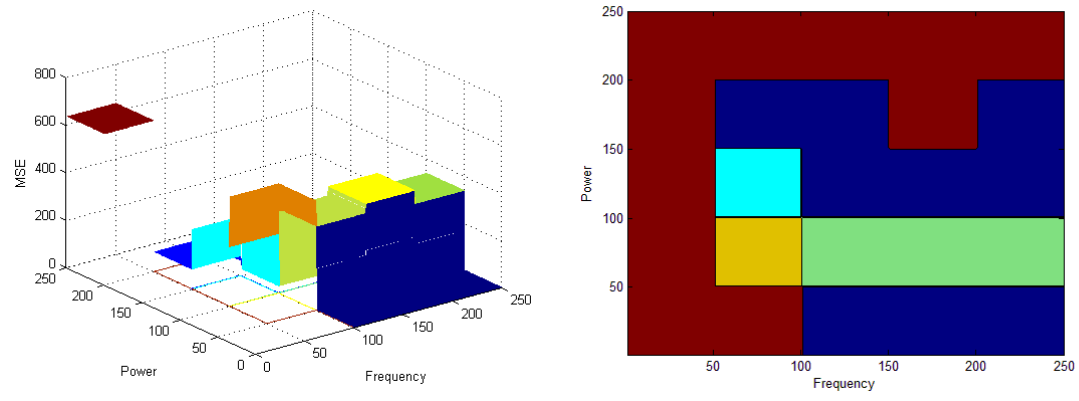
(Female_1 & Female _2)



(a) Constellation



(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



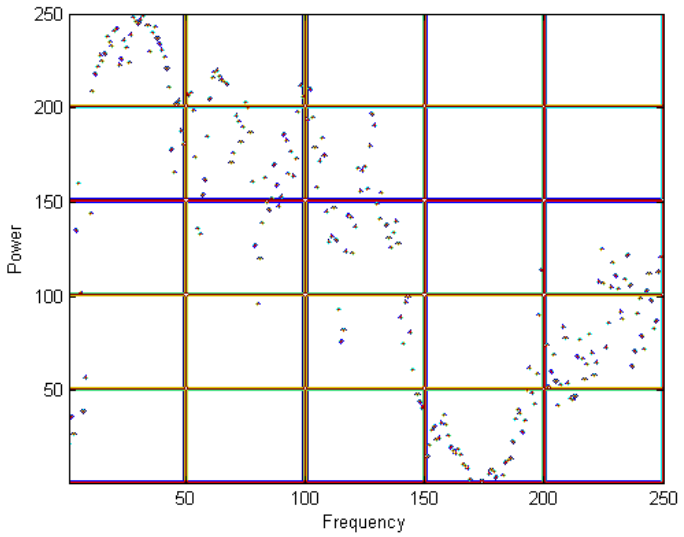
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$

Fig. 59 Simulation of the Second-EXP in CASE-B.2 (AWGN)

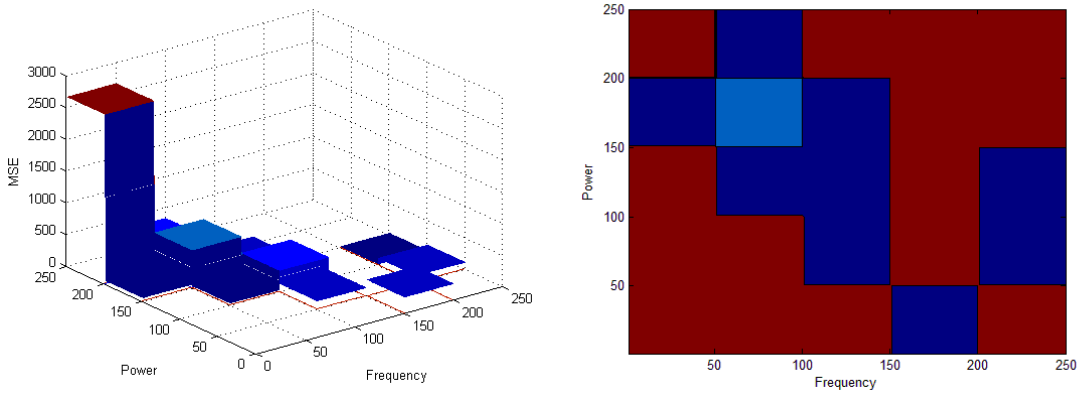
(Female_1 & Female _2)

5.1.4 Effect of Denoising on DOA Estimation

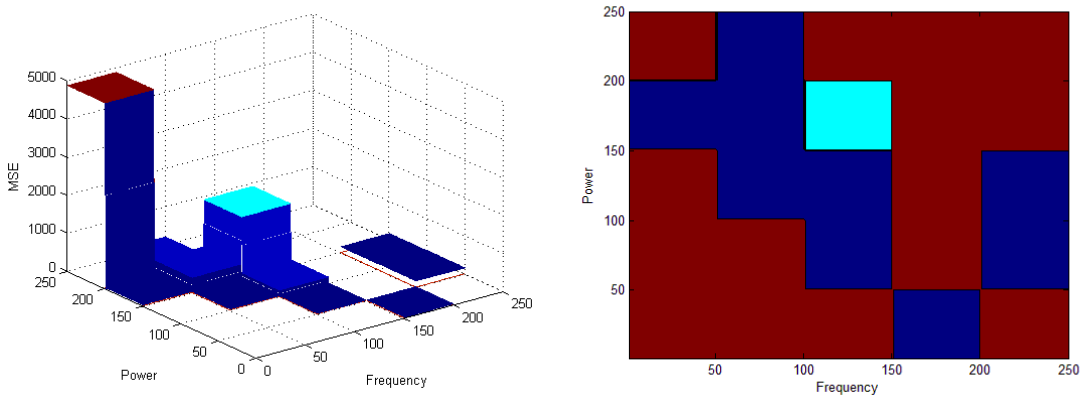
In this section, we look at the Type_3 set-up described earlier and its DOA performance. Here, we show four test sequences. Two test sequences consist of Female_1 and Female_2. Another two sequences consist of Male_1 and Male_2. Fig. 60 show that the test consists of Female_1 and Female_2. The some angles are 15^{o} and -15^{o}. In Fig. 60(a)~(c), we observe that the MSE is large in high PFBs and low frequency bins. Somewhat different from the previous conclusions, Fig. 60(d) shows the angle estimates at each frequency bin. Again, most reliable region is the median frequency bins. Thus, the audio denoising technique [4] has some effects on the DOA estimation but the difference is generally small. We can derive the same conclusions from Fig. 60~Fig. 63. It seems that the denoising technique may expand the confidence regions lightly to the higher frequency bins if their power is not too low. But finally, the middle frequency band is most reliable as discussed earlier.



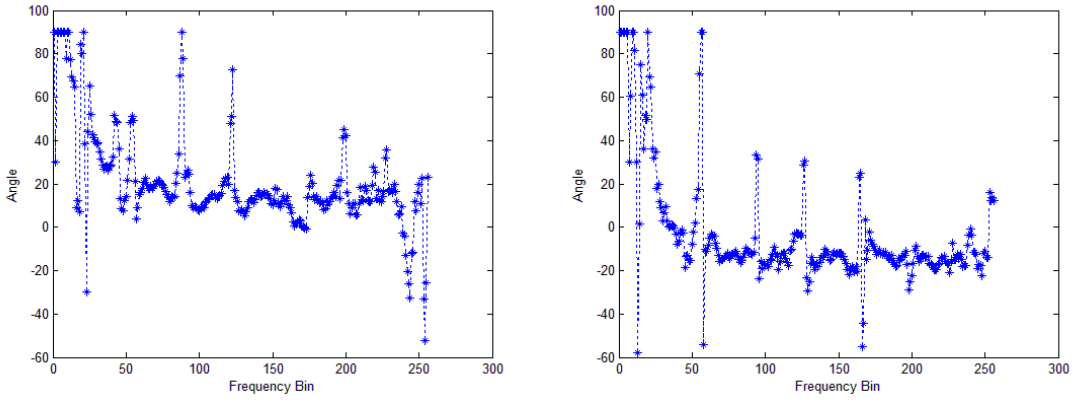
(a) Constellation



(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



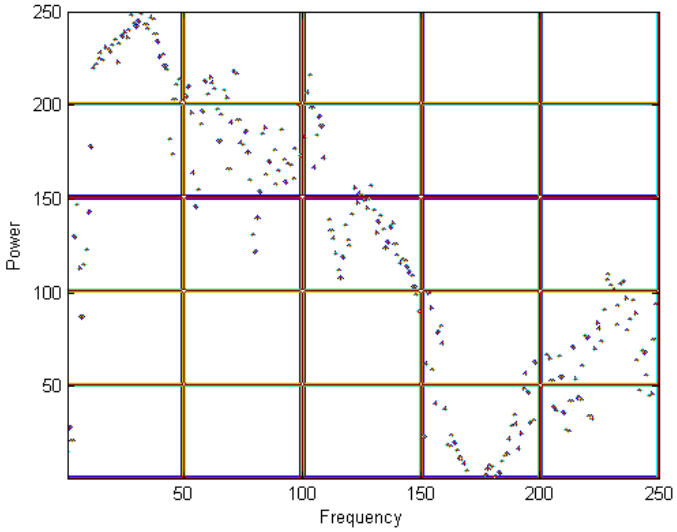
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$



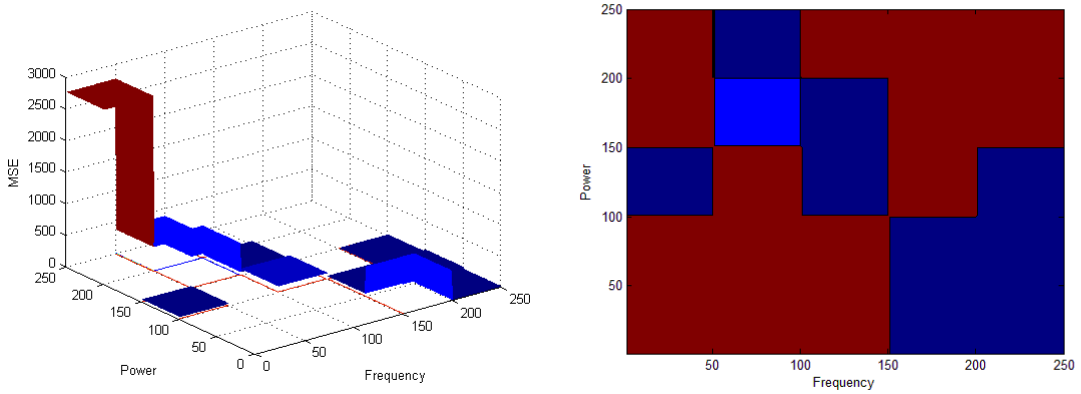
(d) $\theta = 15^o$ and $\theta = -15^o$

Fig. 60 Second-EXP in CASE-A (Real)

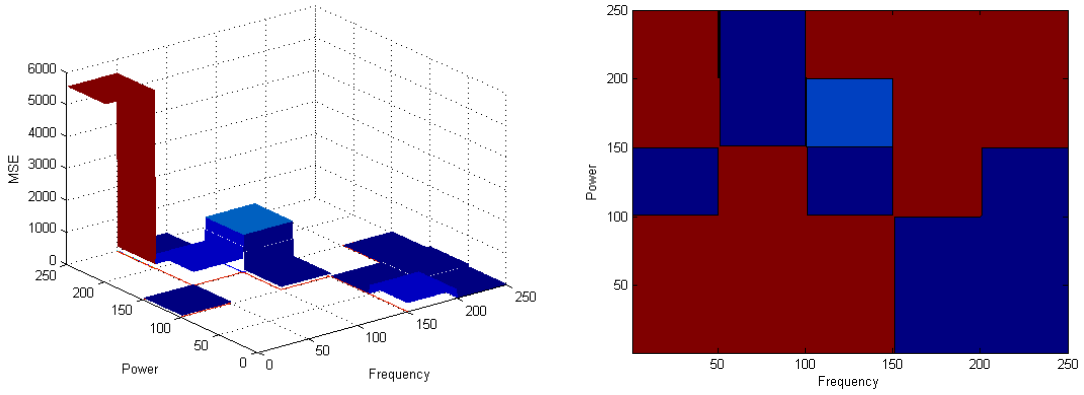
(Female_1 & Female_2)



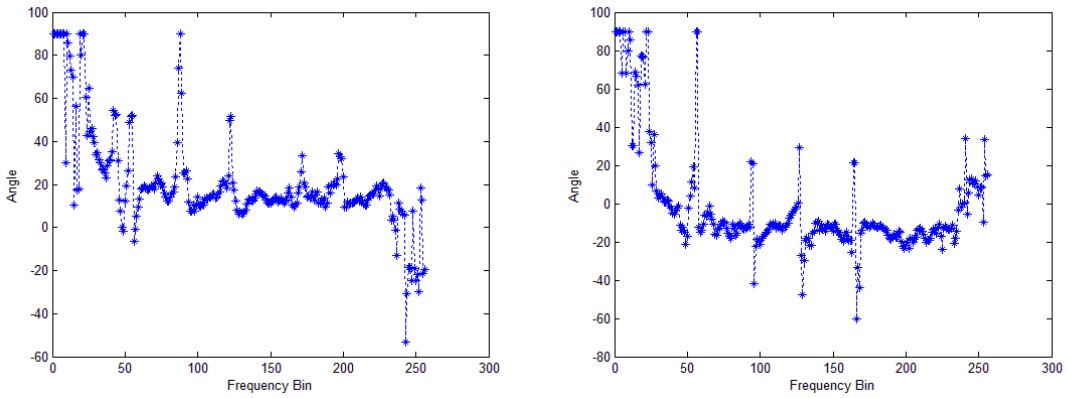
(a) Constellation



(b) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



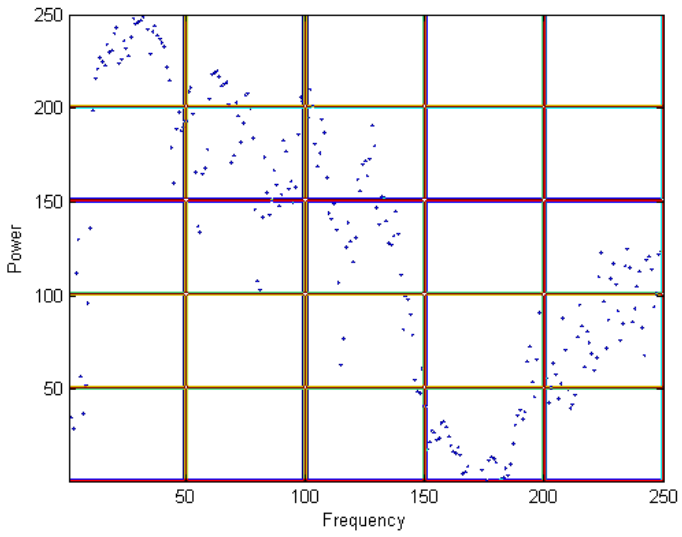
(c) MSE at $\theta = -15^o$ and MSE Contour at $\theta = -15^o$



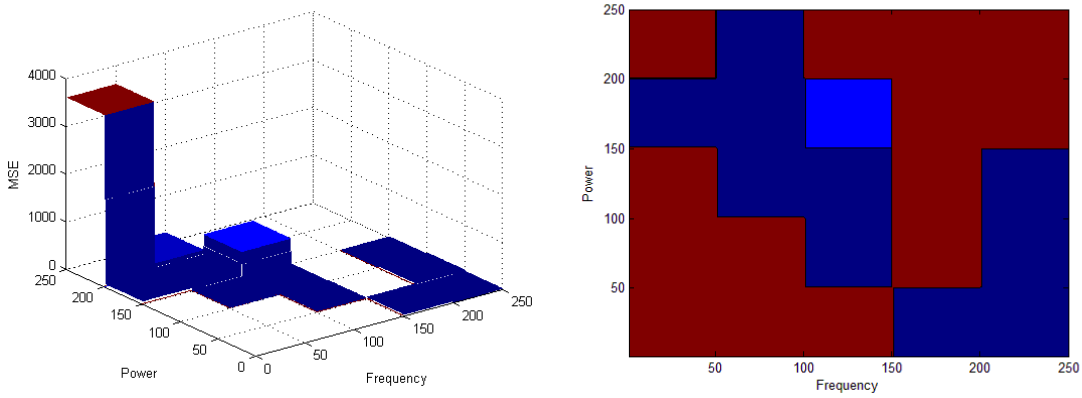
(d) $\theta = 15^o$ and $\theta = -15^o$

Fig. 61 Second-EXP in CASE-A (Real)

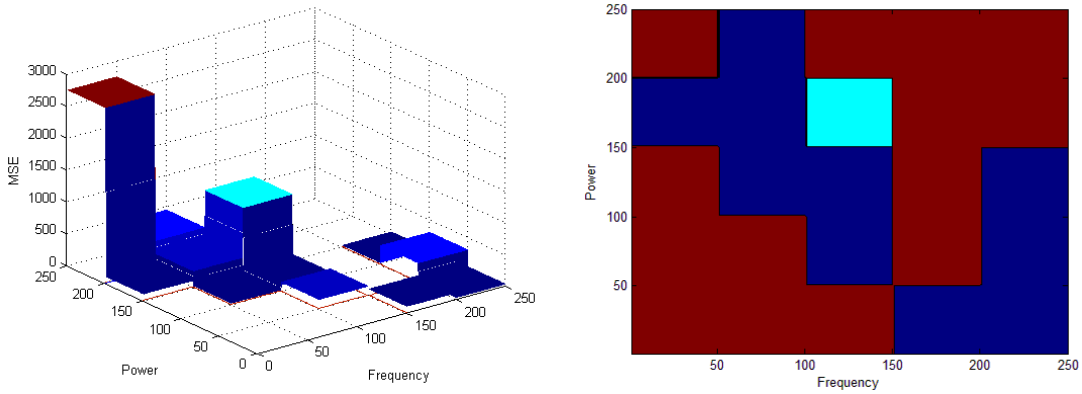
(Male_1 & Male_2)



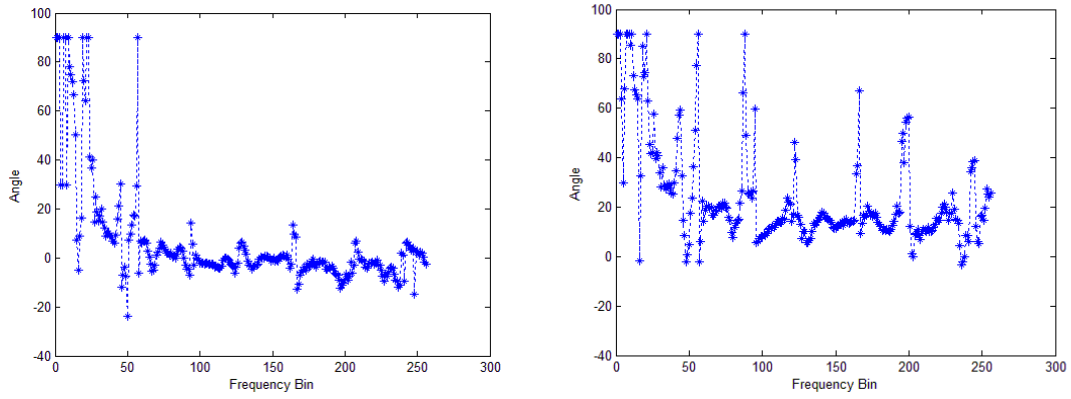
(a) Constellation



(b) MSE at $\theta = 0^o$ and MSE Contour at $\theta = 0^o$



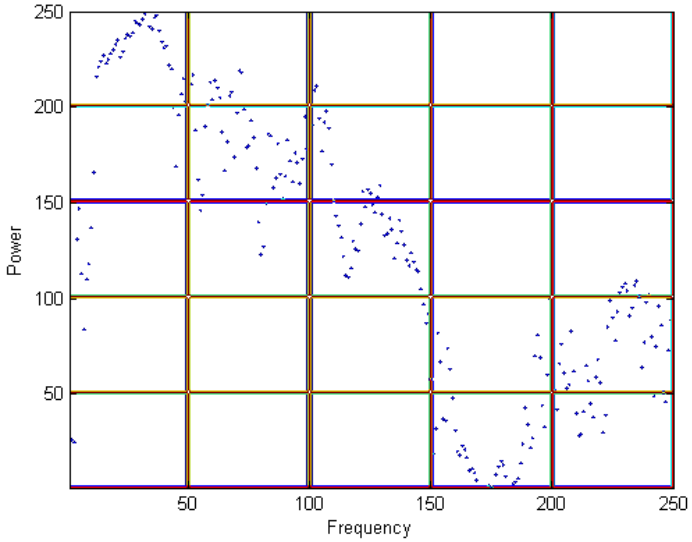
(c) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



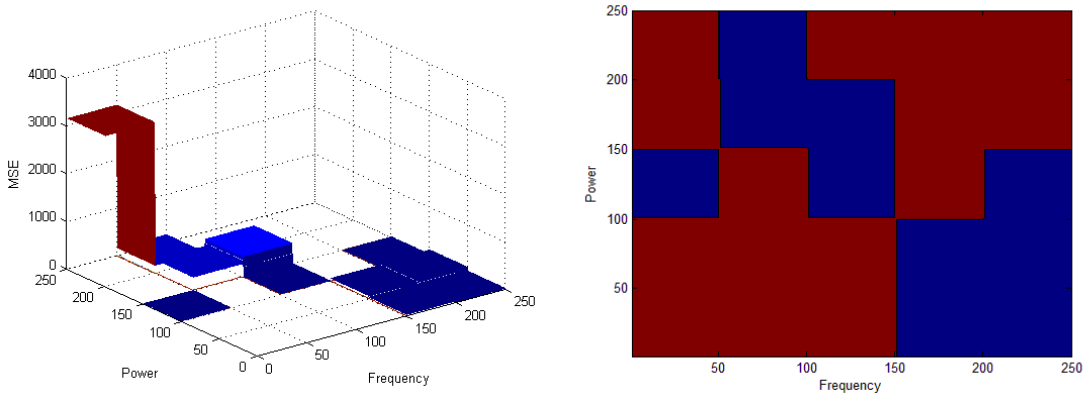
(d) $\theta = 0^o$ and $\theta = 15^o$

Fig. 62 Second-EXP in CASE-A (Real)

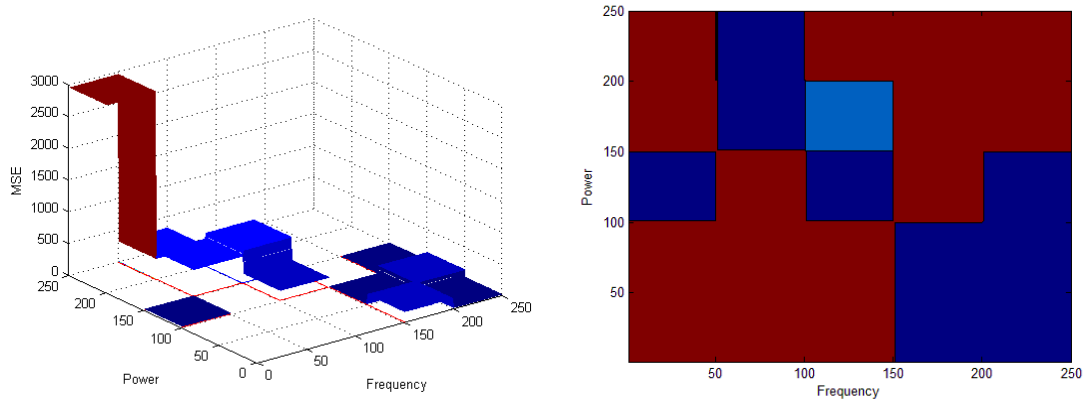
(Female_1 & Female_2)



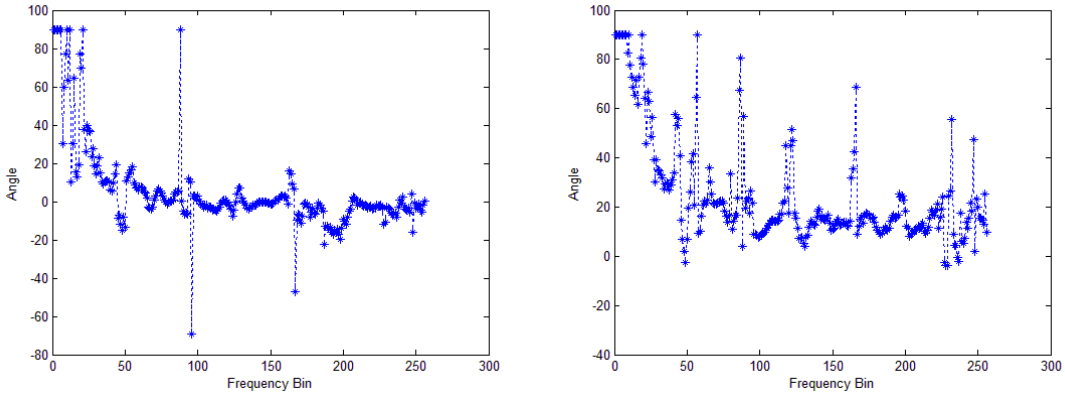
(a) Constellation



(b) MSE at $\theta = 0^o$ and MSE Contour at $\theta = 0^o$



(c) MSE at $\theta = 15^o$ and MSE Contour at $\theta = 15^o$



(d) $\theta = 0^o$ & $\theta = 15^o$

Fig. 63 Second-EXP in CASE-A (Real)

(Male_1 & Male_2)

5.2 Virtual Listening Point Audio Synthesis

In our study, we adopt the SLAB software developed by the NASA Ames Research Center to synthesize the audio at a virtual listening point. The software implements the spatial 3D-sound processing procedure.

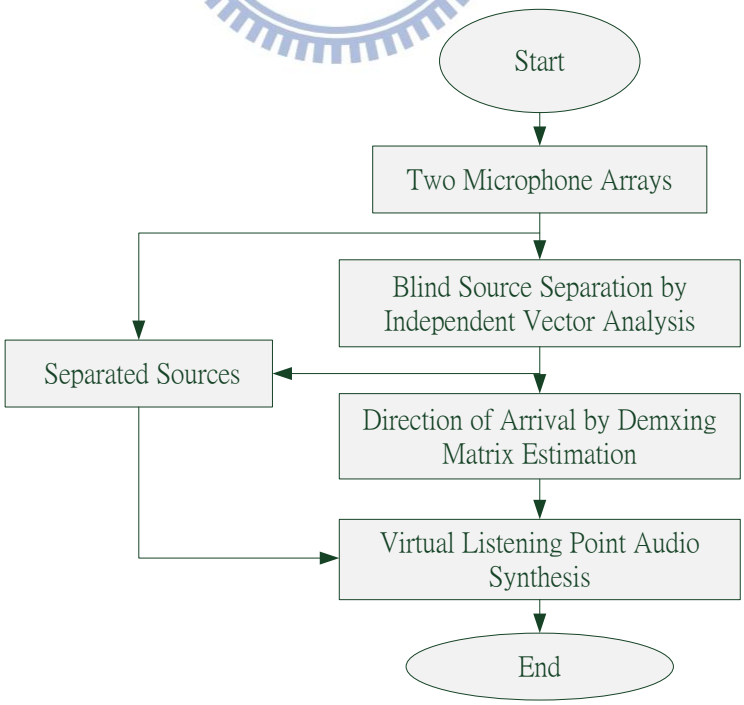


Fig. 64 Flow Diagram of 3D Acoustic Signal Synthesis

In our proposed audio synthesis system, we first perform BSS to separate signals from the recorded mixture signals. Then, we take separated signals as inputs to SLAB. Fig. 65 shows the arrangement of separated signals and the microphone array on the X-Y plane. $Female\_2$, $Male\_2$ and P_o represent the original recording layout. They are respectively the first separated source, the second separated source and the position of original microphone array. The azimuth angles of the first source and the second source, estimated by the DOA algorithm, are 14^o and 14^o respectively. Because of the restrictions on the instruments, we have only one set of microphone array. We did not use trigonometry to estimate the distance of the two sources. The distance of these sources from the microphone are the true values, 1.5M and 1.5M respectively. With all the above set-up, we can synthesize the virtual listening point audio using SLAB. Fig. 66(a)~(d) show the audio signals we synthesize at P_1, P_2, P_3, P_4. The X-Y coordinates in Table. 5 represent the exact positions in Fig. 65.

Table. 5 Spatial Location

	Female_2	Male_2	P_1	P_2	P_3	P_4
Coordinate	(-0.36, 1.46)	(0.36, 1.46)	(-1, 1.3)	(-0.4, 1.1)	(0.4, 1.1)	(1, 1.3)

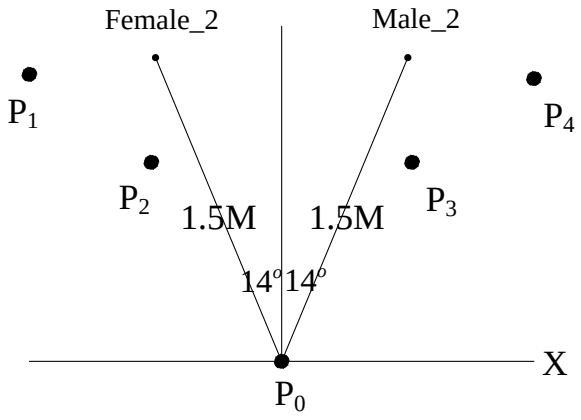
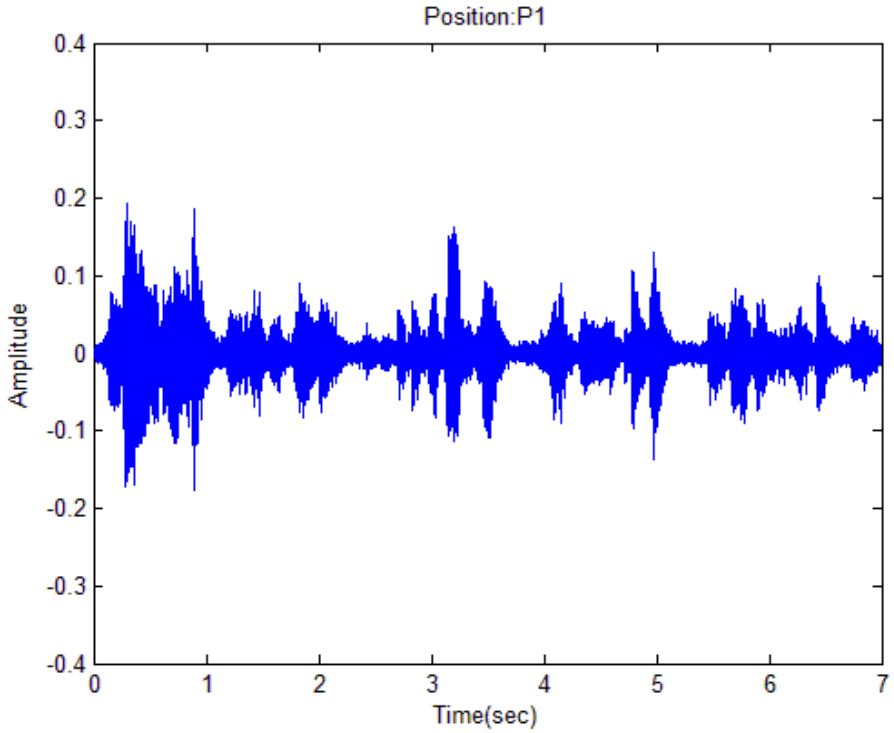
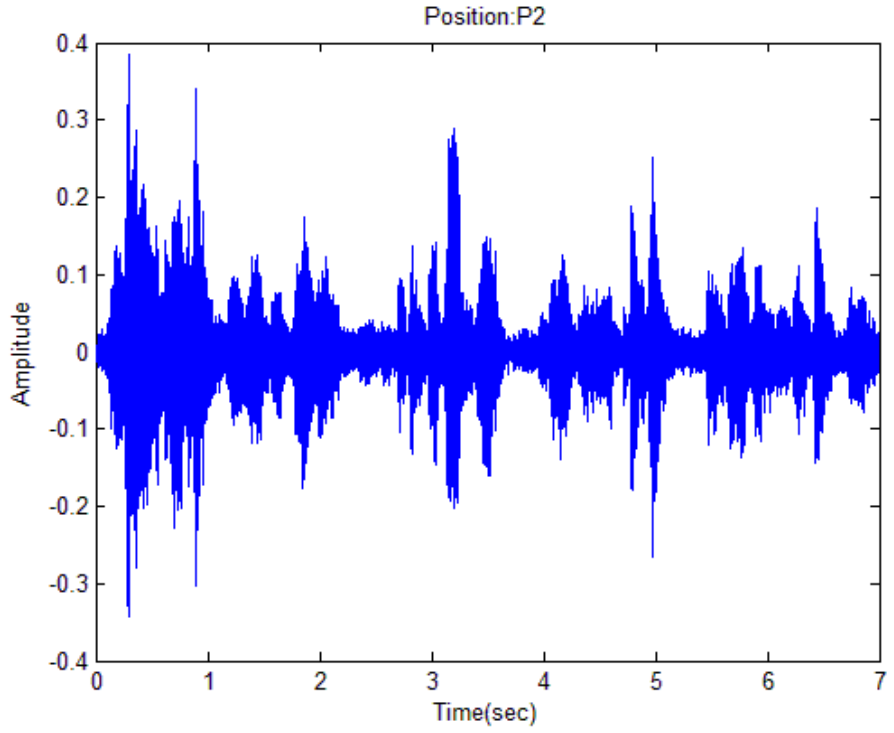


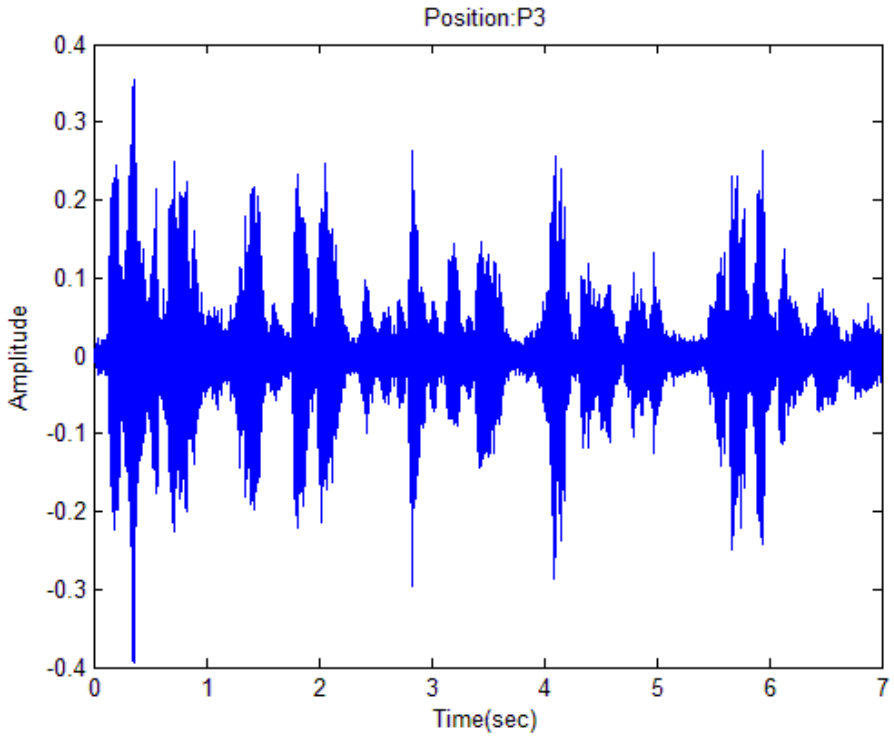
Fig. 65 Locations of Synthesized Audio



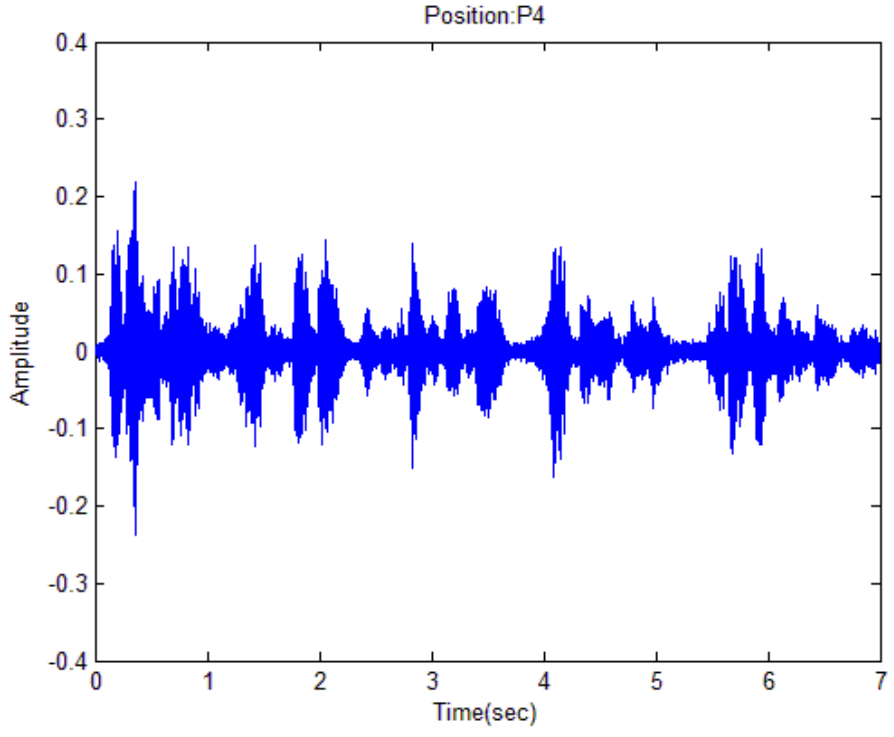
(a) Audio signal at P1



(b) Audio signal at P2



(c) Audio signal at P3



(d) Audio signal at P4

Fig. 66 Virtual Listening Point Audio

Chapter 6 Conclusion and Future Work

6.1 Conclusion

The main propose of this thesis is to synthesis virtual listening point audio from the recorded mixture signals in an anechoic chamber. We adopt the FastIVA method to separate the sound sources from the recorded microphone array audio. We use the DOA technique (in an ICA-based scheme) to estimate the source directions. We also adopt the audio denoising technique to improve the subjective hearing quality. Based on the experimental results, we have the following conclusions:

For BSS technique:

1. The microphones with high SNR determine the BSS quality.
2. The performance of BSS algorithm often improves by adding more sensors.
3. The BBS quality provides the better results when we input more data. Also, in a real environment, there is good performance with four-second data length.
4. Two source signals with similar power results in better BSS quality.

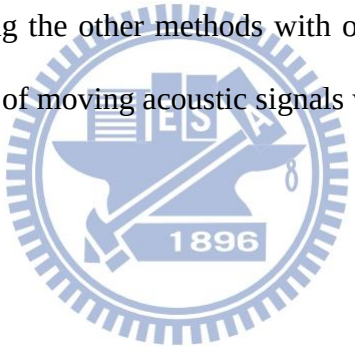
For DOA technique:

1. The results of DOA estimation are more accurate when there are more sensors.
2. Low frequency components are unreliable in DOA estimate because of their long period in time. However, the high frequency components often have high noise. To obtain reliable DOA estimates, the SNR of that signal needs to be sufficiently large. This requirement eliminates the high frequency components. Therefore, from our statistics, the median frequency bins have more reliable estimates. These reliable frequency bins are ranging from 50 to 200 (512-size window).

For denoising technique, we can find Type_2 or Type_3 shown in Fig. 35 are both usable. This technique can improve the subjective hearing quality in the BSS method but does not seem to help the source direction estimations in the DOA method.

6.2 Future Work

In our experiments, the elevation angle estimate is inaccurate. Although people cannot dearly distinguish the voice source vertical direction at different the elevation angle, it may still be worthwhile to improve the elevation angle estimation. Because of the restrictions on our instruments, the distance of the two source signals is not estimated with trigonometry. In the future, the problem can be solved by multiple microphone arrays or by using the other methods with one microphone array. Another possible topic is the synthesis of moving acoustic signals with the Doppler effects.



REFERENCES

[1] J. Benesty, et al, *Microphone Array Signal Processing*, Berlin, Germany: Springer-Verlag, 2008.

[2] Alan V. Oppenheim, et al., *discrete-time signal processing (Second Edition)*, Prentice Hall, 1999.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875-1902, 2005.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley Intrerscience, 2001.

[5] H. Sawada, et al., "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 5, Sep. 2004.

[6] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320-327, Mar. 2000.

[7] I. Lee, et al., Complex FastIVA: "A robust maximum likelihood approach of MICA for convolutive BSS," in *Lecture Notes in Computer Science*, pp. 601-608, 2006.

[8] I. Lee, et al., "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," in *Signal Process*, vol. 87, pp. 1859-1871, 2007.

[9] F. Asano, et al., "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, May 2003.

[10] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. ICA*, pp. 722-727, Dec. 2001.

[11] S. Amari, "Natural gradient works efficiently in learning," *Neural Computing*, vol. 10, no. 2, pp. 251-276, 1998.

[12] S.-I. Amari, A. Cichocki, H.H. Yang, A new learning algorithm for blind signal separation, in: Advances in Neural Information Processing Systems, vol. 8, pp. 757–763, 1996.

[13] G. Yu and S. Mallat, "Audio denoising by time-frequency block thresholding." *IEEE Trans. on Speech and Audio Processing*, vol. 56(5), pp. 1830–1839, 2008.

[14] B. E. M. S. Yu, G., "Audio signal denoising with complex wavlets and adaptive block attenuation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICAASP)*, 2007.

[15] D. L. G. J. T. B. Fvotte, C., "Sparce regression with structured priors: pplication to audio denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[16] A. L. M. Levada, D. C. Correa, "An Adaptive Approach for Contextual Audio Denoising using Local Fisher Information," *IEEE International Symposium on Circuits and Systems*, pp.125-128 , May. 2011.

[17] J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, no. 2, pp. 165–168, 1980.

[18] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. New York: Duxbury, 2002.

[19] A. L. M. Levada, N. D. A. Mascarenhas, and A. Tannus, "A novel map-mrf approach for multispectral image contextual classification using combination of suboptimal iterative algorithms," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 1795–1808, 2010.

[20] J, Dmochowski, J. Benesty, and S. Affes, "On Spatial Aliasing in Microphone Arrays," *IEEE Transactions on Signal Processing*, vol. 57, pp. 1383-1395, 2008.

[21] H. Sawada, et al., "Direction of arrival estimation for multiple source signals

using independent component analysis," in *Proc. International Symposium on Signal Processing and its Applications*, pp. 411–414, July 2003.

[22] R. Roy and T. Kailath, "ESPRIT-Estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, No.7, pp. 984-995, July 1989.

[23] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antenna and Propagation*, vol. AP-34, pp. 276-280, Mar. 1986.

[24] Y. Hua, et al., "An L-shaped array for estimating 2-D directions of wave arrival," *IEEE Trans. Antennas propagation*, 1991, 39(2): 143-146.

[25] C. Jian, et al., "2-D DOA Estimation by MEMP Based on L-shape Array," *IEEE 2006 8th international conference on signal processing*, vol. 1, 2006.

[26] W. M. G. Diab and H. M. Elkamchouchi, "A Novel Approach for 2D-DOA Estimation using Cross-Shaped Arrays," *Antennas and Propagation Society International Symposium, 2008. AP-S 2008. IEEE*, On page(s): 1-4, July 2008.

[27] M. G. Porozantzidou and M. T. Chryssomallis, "Azimuth and Elevation Angles Estimation using 2-D MUSIC Algorithm with an L-shape Antenna" *(APSURSI), 2010 IEEE* , On page(s):1-4, July 2010.

[28] H. Yuan, et al., "A DOA estimation method for 3D multiple source signals using independent component analysis," in *Proc. EUSIPCO2006*, Italy, Sept. 2006.

[29] J. D. Miller, "SLAB: a software-based real-time virtual acoustic environment rendering system," in *Proc. of the 2001 International Conference on Auditory Display*, Espoo, Finland, Jul. 2001.

自傳

簡士傑，1988 年 5 月 11 日出生於宜蘭縣。2010 年 6 月畢業於國立中正大學電機工程學系，之後進入國立交通大學電子研究所攻讀碩士學位，承蒙杭學鳴教授的指導，進入通訊電子與訊號處理實驗室(Commlab)，主要研究方向為多媒體訊號處理，論文題目為「使用無響室錄音合成虛擬聆聽點」。