國 立 交 通 大 學

電子工程學系 電子研究所碩士班

碩 士 論 文

由麥克風陣列訊號合成出虛擬聆聽點的

3D音訊

3D Acoustic Signal Synthesis at

Virtual Listening Point Using

Microphone Array Signals

研 究 生：張欽淵

指導教授：杭學鳴 博士

中 華 民 國 九 十 八 年 七 月


E
S
A
1896

由麥克風陣列訊號合成出虛擬聆聽點的３Ｄ音訊

3D Acoustic Signal Synthesis at Virtual Listening Point Using Microphone Array Signals

研究生：張欽淵　　Student: Chin-Yuan Chang

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

國 立 交 通 大 學

電 子 工 程 學 系　電 子 研 究 所 碩 士 班

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
in
Electronics Engineering

July 2009

HsinChu, Taiwan, Republic of China

中華民國九十八年七月


E
S
A
1896

由麥克風陣列訊號合成出虛擬聆聽點的３Ｄ音訊

研究生：張欽淵　　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

本論文的目標是為了在無原始麥克風錄音訊號的虛擬聆聽點上合成出 3D 音訊。為了達到這個目標，我們在空間中佈置麥克風陣列用以進行音源訊號的錄製工作。３Ｄ音訊合成可分為兩個主要步驟，第一個步驟是由混合的錄製訊號去估測各個音源訊號，此步驟通常是以盲訊號源分離 (blind source separation, BSS) 的技術來達成。第二個步驟則是在選定的回響空間內某一個虛擬聆聽點上合成出該點的３Ｄ音訊。此音訊的３Ｄ空間感可藉由頭部相關轉移函數 (head-related transfer function, HRTF) 與代表該點房間回響感覺的聽覺轉移函數 (acoustic transfer function, ATF) 對已分離訊號進行濾波而得到。

在本論文內，我們採用頻率域獨立成份分析 (frequency domain independent component analysis, FD-ICA) 和最小平方誤差近似解 (least squares optimization approach) 將混合訊號分離。我們以訊號干擾比 (signal to interference ratio, SIR) 來評估分離矩陣的效果。在重建３Ｄ音訊的過程中，我們會先計算出該回響空間

的聽覺轉移函數總集 (ATF-pool)，接著從 ATF-pool 當中選取對應的 ATF 來對已分離訊號濾波，然後再以適當的 HRTF 合成出 3D 雙聲道音訊。對於不在 HRTF 和 ATF 測量點上的虛擬聆聽點，其對應的 HRTF 和 ATF 分別以現有的 HRTF 和 ATF 總集用內差的方式求得。最後，在任意位置的虛擬聆聽點和所選的空間回響環境內展示出具有３Ｄ效果的合成音訊。

關鍵詞：麥克風陣列、３Ｄ音訊合成、盲訊號源分離、頭部相關轉移函數、聽覺轉移函數、虛擬聆聽點

3D Acoustic Signal Synthesis at Virtual Listening Point Using Microphone Array Signals

Student: Chin-Yuan Chang Advisor: Dr. Hsueh-Ming Hang

Department of Electronic Engineering
Institute of Electronics
National Chiao Tung University

Abstract

The target of 3D virtual listening point audio synthesis is to reconstruct 3D audio at a virtual point where the original recording microphone does not exist. To facilitate this idea, the source music is recorded by a microphone array that consists of more than a few recording microphones arranged in a designed spatial pattern. The 3D acoustic signal synthesis can be divided into two key steps. The first step is to estimate the individual source signal from the mixed, recorded signals. This step is usually accomplished by using the blind source separation (BSS) technique. The second step is to synthesize a 3D acoustic signal at a virtual listening point in a chosen reverberant room environment. The 3D feeling of an acoustic signal can be enhanced by filtering the separated signals in step one by the head-related transfer function (HRTF) and the acoustic transfer function (ATF), which represents the room acoustic effect.

In this study, we adopt the frequency domain independent component analysis (FD-ICA) and a least-square optimization approach to separate the mixture signals. We investigate the effectiveness of the BSS methods by evaluating their demixing matrices using the signal to interference ratio (SIR) metric. In the reconstruction

process, we first calculate the ATFs of the reverberant room to form an ATF-pool. Then, the separated signals are mixed using the adequate ATFs drawn from the ATF-pool. Finally, the 3D two-channel audio is synthesized with the help of appropriately chosen HRTFs. A few problems have to be solved in the aforementioned procedure. For example, for an off-grid virtual listening point, its HRTF and ATF are interpolated using the existing HRTF library and the ATF-pool, respectively. At the end, the synthesized 3D acoustic signals are demonstrated with arbitrary virtual listening point and selected room reverberation environments.

Keywords: microphone array, 3D acoustic signal synthesis, blind source separation (BSS), head-related transfer function (HRTF), acoustic transfer function (ATF), virtual listening point

誌謝

能夠完成這篇論文，我首先要感謝的是指導教授杭學鳴老師，自從我大二開始的專題研究一直到碩士班的這五年之間，在老師的指導之下學習到做研究的方法和應有的態度。老師除了在研究上給予專業的意見之外，亦鼓勵我們繼續努力進步。老師也時常關心我們的生活，並提供許多相當實用的建議與協助。在此向老師致上我最誠摯的感謝。

我也要感謝峰誠學長，在我碩一時做數位典藏計劃的時候教我許多 Java 程式的技巧，家揚學長在我大學專題研究時期就熱心地從 Fourier 轉換開始帶我入門，以及繼大、大師和雄哥這些學長們在研究上給予許多資訊與幫助。諸位 Commlab 的同學、學弟們，我們在修課時一起打拼、在平時一起打球、在空閒時一起打 CS，有你們在的日子，在 Commlab 裡的生活就是多采多姿！

此外，Commlab 提供的研究設備與環境使得研究能持續進行，在熬夜趕進度之後，能睡在實驗室沙發上真是一大享受，大家都在忙的時候還要排隊的說！

最後要感謝我的家人和朋友，你們的支持是我前進的動力！

誌於 2009 年 7 月

欽淵


E
S
A
1896

Contents

List of Tables


E
S
A
1896

List of Figures

Chapter 1

Introduction

As we human beings live in a three dimensional space, our 3D acoustic feeling of the two ears is well-trained by every received audio signal so that it is easy for us to distinguish several different sound sources from a convolutive mixture signal such as a microphone signal in a room. However, the 3D acoustic feeling is lost in the transition of multiple natural source signals to the microphone signal. Our goal is to reproduce an audio signal with a reconstructed 3D acoustic feeling from the omni-directional microphone array signals. With the 3D acoustic signal, one can have the feeling of the direction, distance and elevation of each sound source and the reverberation of the room, which would be much impressive rather than a single channel mixture signal. Another application of the 3D acoustic signal synthesis is to match up with the 3D view point camera array, which can make the overall sequence vivid and lively.

Therefore, our main propose is to synthesize a 3D acoustic signal from the omni-directional microphone array signals. This task can be intuitively divided into two major steps. The first step separates the source signals blindly and the second step adds in the 3D acoustic feeling. The former is usually achieved by the blind source separation (BSS) method and the latter is realized by filtering with the acoustic transfer function (ATF) and the head-related transfer function (HRTF).

For the first step, there are many BSS methods [1] and one of the most popular methods is called independent component analysis (ICA). The concept of the ICA methods is to make the separated signals as statistically independent as possible. Different kinds of implementations of the ICA method [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] have different adaptive learning rules and different properties. For over-determined BSS methods,

the subspace of interest would be extracted by the principle component analysis (PCA) method [12] or the other subspace methods [1], [13], [14]. Some subspace method operates in the frequency domain. Thus the frequency domain ICA (FD-ICA) method is applied to the subsequent separation procedure with the permutation problem and the scaling problem for each individual frequency bins. The permutation problem can be solved by the hybrid method of direction of arrival (DOA) and the correlation method [15] and the scaling problem is solved under the minimum distortion principle (MDP) [16]. For the convolutive mixture signals, a least squares optimization based on the cross-power-spectrum approach is adopted [17].

For the second step, the HRTF is adopted in this thesis as we present the 3D feeling through the headphone [13], [18]. Due to the effect of HRTF, the degradation of the ill-separated signals can be reduced in a subjective way. We also include the room impulse response (RIR) by filtering the output signals with ATF [19]. The ATF characterizes the reflection effect of a reverberant room. The ATF is changed along with the room sizes and the wall materials, which will be discussed in this thesis.

With HRTF and ATF, we can synthesize audio signals for different scenarios based on the separated signals dynamically and thus the use of HRTF and ATF can bring the spatial impression of a virtual listening point in a specific room. We demonstrate the effect of these two methods for different scenarios.

This thesis is organized in the order of the processing flow from the captured microphone array signals to the 3D acoustic signal. In chapter 2, we describe the adopted BSS method for the sound separation. In chapter 3, the synthesis method of the 3D acoustic signal using the separated signals is presented. The HRTF and ATF of a virtual listening point can be interpolated by the adjacent measured points [19], [20]. In chapter 4, we describe the simulation setups, discuss the effectiveness of the adopted BSS methods, and demonstrate the overall system performance. Then we conclude the research results and

discuss the future work in chapter 5.


E
S
A
1896

Chapter 2

Blind Source Separation

2.1 Introduction to Blind Source Separation (BSS)

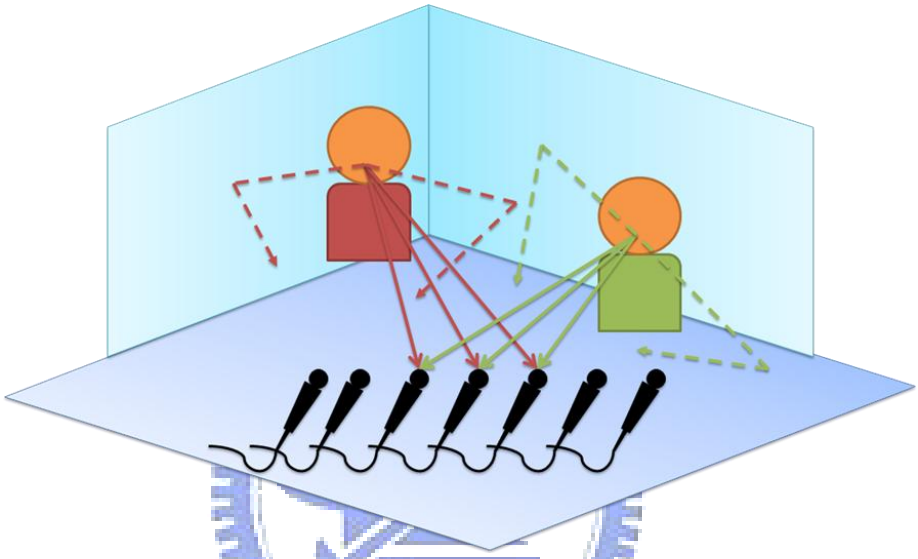


Fig. 2.1 Cocktail Party Problem

The “Cocktail Party Problem” is known as one of the most famous problems in the area of acoustic signal processing. It is described by the following sentence, “how to focus one's listening attention on a single talker among a mixture of conversations and background noises, e.g. cocktail party, and ignoring other conversations?” In the words we used in the rest of the article, considering that each talker as a sound source and an array of microphones placed in the room, how can we separate the sound sources or segregate a particular one by processing on the mixture signals we received from the microphone array? It is a difficult problem especially under the condition of “blind”, which means that the source signals and the mixing process are unknown and only the recordings of the mixtures are available. The goal of BSS is to recover all sound sources or a particular one from the

recorded mixtures under the condition of “blind”.

In many cases of using BSS methods, it is necessary for the BSS system to have the priori 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 number of sound sources is not essential as long as we have sufficient number of microphones.

The BSS system as shown in Fig. 2.1 is often quite effective. We adopt this filtering network described below to obtain the approximation of the source signal vector $\mathbf{s}$.

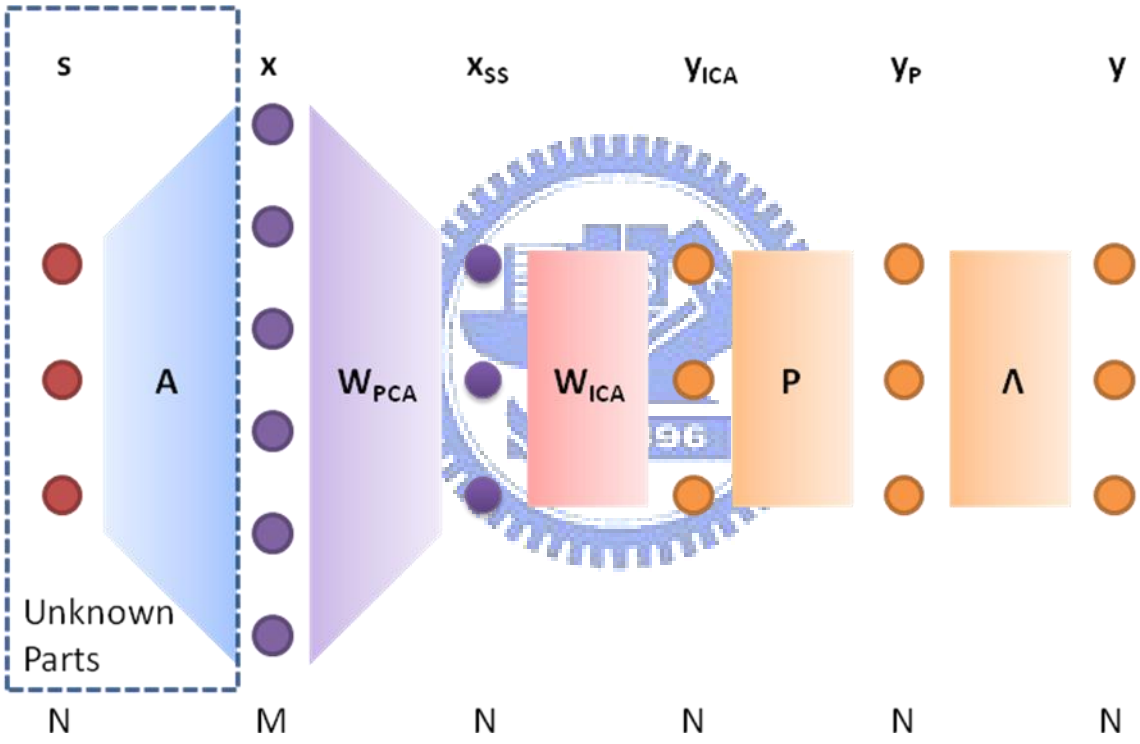


Fig. 2.2 BSS System Concept

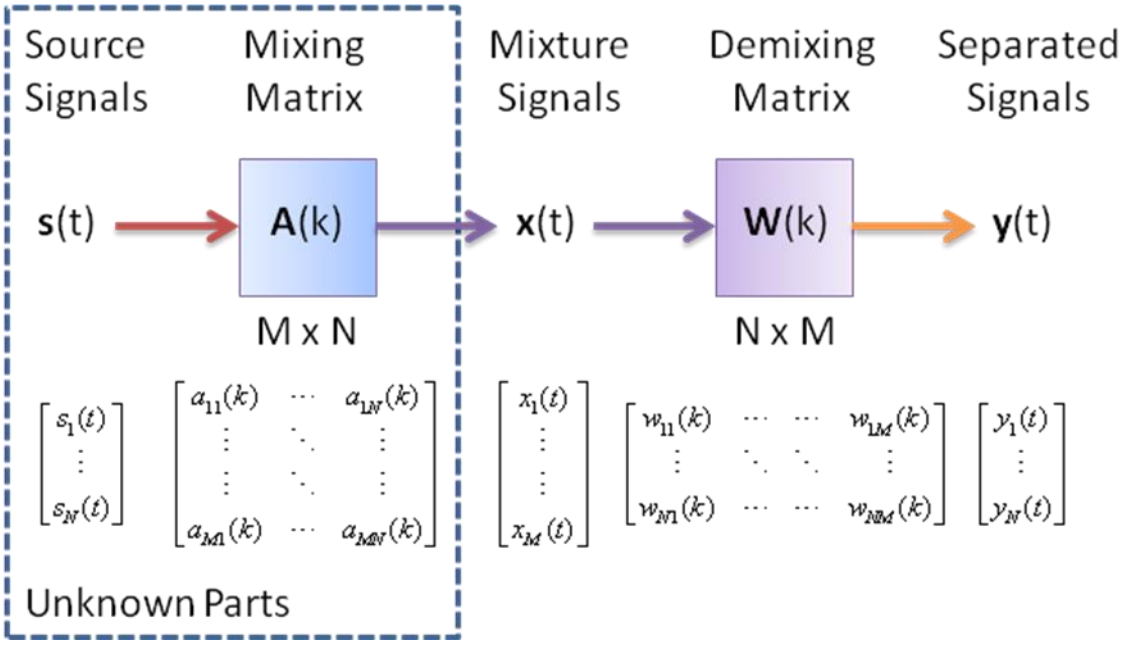


Fig. 2.3 BSS Filters Work in the Time Domain

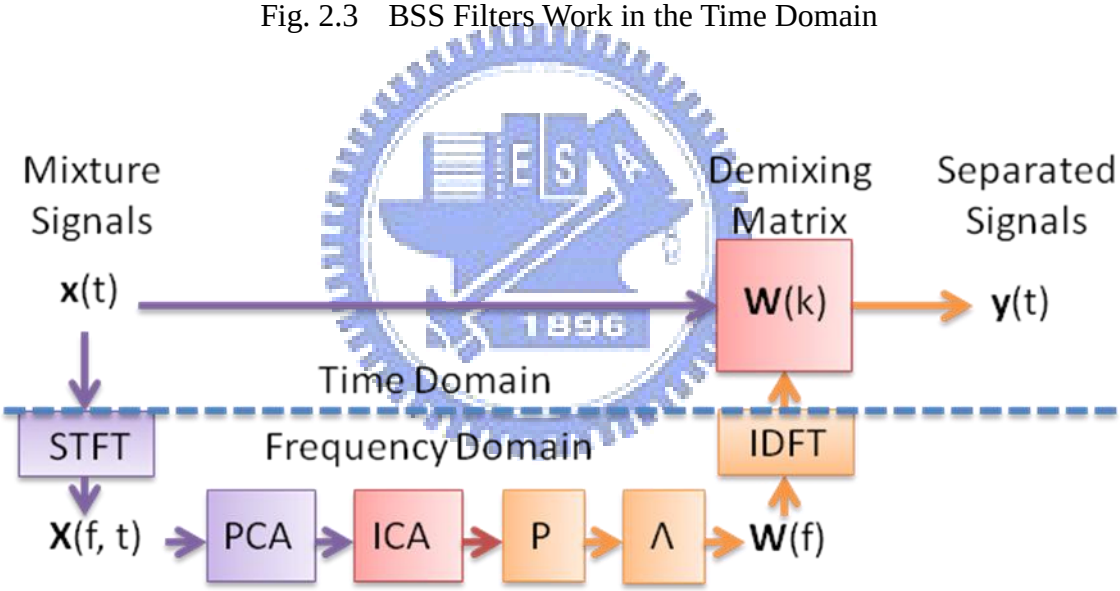


Fig. 2.4 Flow Chart of Obtaining Demixing Matrix in Frequency Domain

2.2 Model of Acoustic Signals

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

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

where $\mathbf{x}(f,t)=[X_1(f,t),\ldots,X_M(f,t)]^T$ denotes the vector of mixture signals, M denotes the number of microphones and $X_m(f,t)$ denotes the short-term Fourier transform (STFT) of the m-th microphone in the t-th time frame, $\mathbf{A}(f)$ denotes the mixing matrix, $\mathbf{s}(f,t)=[S_1(f,t),\ldots,S_N(f,t)]^T$ denotes the vector of source signals, N denotes the number of sound sources, $S_n(f,t)$ denotes the STFT of the n-th source in the t-th time frame, and $\mathbf{n}(f,t)$ denotes the mixture of less-directional components which includes room reflections and ambient noise. Therefore, $\mathbf{A}(f)\mathbf{s}(f,t)$ represents the directional components in $\mathbf{x}(f,t)$.

In addition, the (m,n) element of the mixing matrix $\mathbf{A}(f)$ can be considered as the transfer function from the n-th source to the m-th microphone, which is modeled as:

$$A_{m,n}(f)=|A_{m,n}(f)|e^{-j2\pi f\tau_{m,n}},$$

where $A_{m,n}(f)$ denotes the magnitude of the transfer function and $\tau_{m,n}$ denotes the propagation time from the n-th source to the m-th microphone.

2.3 Subspace Method

Since the number of mixture signals is greater than the number of source signals, by utilizing the subspace method [2], we can obtain the filtered subspace signals in which the room reflections and ambient noises are reduced. In other words, the subspace of direct components is selected and in the meanwhile the subspace of reflection components is discarded.

The subspace signals $\mathbf{x}_{ss}(f,t)$ is obtained by the following expression as shown in Fig. 2.2:

$$\mathbf{x}_{ss}(f,t)=\mathbf{W}_{PCA}(f)\mathbf{x}(f,t),$$

where the subspace filter $\mathbf{W}_{PCA}(f)=\mathbf{\Lambda}_{ss}^{-\frac{1}{2}}(f)\mathbf{E}_{ss}^{H}(f)$ is a special case of principal

component analysis (PCA) method with $M >> N$, $\mathbf{\Lambda}_{ss}(f)$ denotes the subspace eigenvalue matrix and $\mathbf{E}_{ss}(f)$ denotes the eigenvector matrix corresponding to $\mathbf{\Lambda}_{ss}(f)$.

$\mathbf{\Lambda}_{ss}(f)$ and $\mathbf{E}_{ss}(f)$ are obtained from the spatial correlation matrix $\mathbf{R}(f) = \langle \mathbf{x}(f,t)\mathbf{x}^H(f,t) \rangle_t$ and $\mathbf{R}(f)$ can be decomposited into $\mathbf{R}(f) = \mathbf{E}(f)\mathbf{\Lambda}(f)\mathbf{E}^{-1}(f)$ where $\mathbf{\Lambda}(f)$ denotes the eigenvalue matrix and $\mathbf{E}(f)$ denotes the eigenvector matrix. It is assumed that the significant eigenvalues are occupied by the direct components from the sound sources and the rest are full with the energy of room reflections and ambient noises. Therefore, we pick N significant eigenvalues to form a subspace eigenvalue matrix $\mathbf{\Lambda}_{ss}(f) = diag(\lambda_1(f), \ldots, \lambda_N(f))$ where $\lambda_k(f)$ is the k-th significant eigenvalue at frequency f.

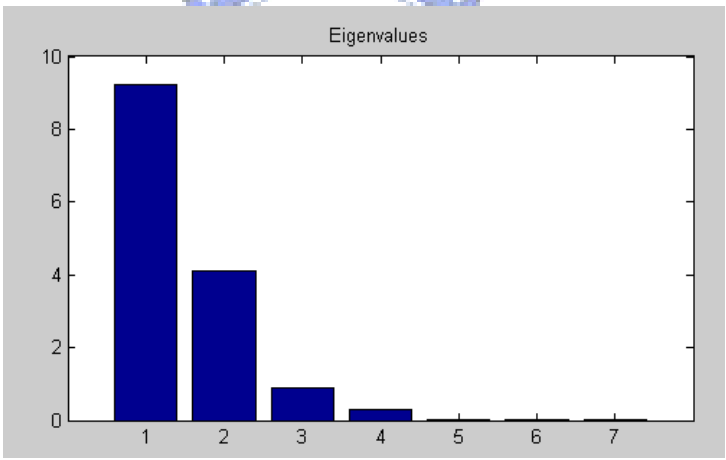


Fig. 2.5 A Typical Example of Eigenvalues for M = 7 and N = 2

2.4 Independent Component Analysis (ICA)

The goal of ICA is to make the output signals $\mathbf{y}$ be statistically independent. In other words, the joint probability distribution of output signals $\mathbf{y}$ equals to the product of each marginal distribution, which can be shown as the following expression:

$$f(\mathbf{y}) = \prod_{i=1}^{N} f_i(y_i).$$

The cost function to minimize the redundancy between each output signal y_i can be shown as the following expression:

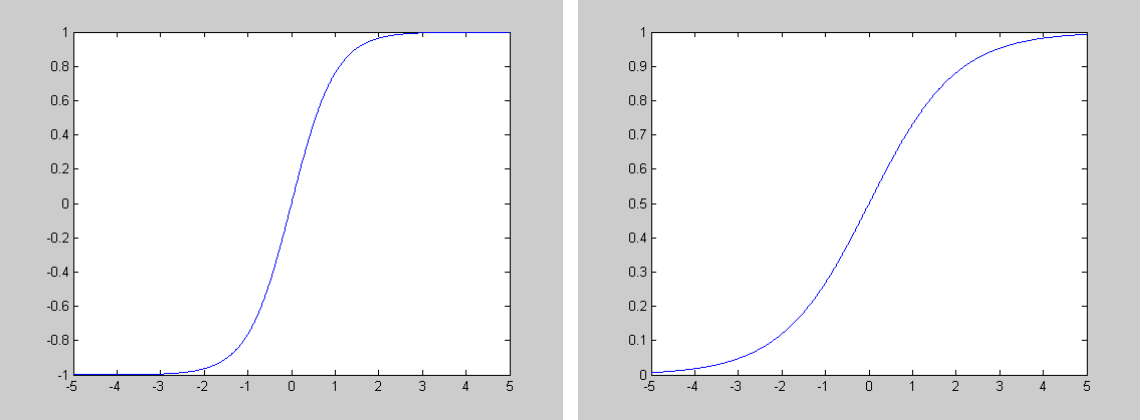
$$I(\mathbf{y}) = E\left[\log \frac{f(\mathbf{y})}{\prod_{i=1}^{N} f_i(y_i)}\right] = \int f(\mathbf{y}) \log \frac{f(\mathbf{y})}{\prod_{i=1}^{N} f_i(y_i)} d\mathbf{y}.$$

When $f(\mathbf{y}) = \prod_{i=1}^{N} f_i(y_i)$, the value of the cost function $I(\mathbf{y})$ equals to 0.

2.4.1 Information Maximization Method and Natural Gradient Method

One of the popular methods to approach the aim of having the statistically independent output signals is the information maximization method, as known as the Infomax method. The Infomax method maximizes the mutual information $I(\mathbf{y};\mathbf{x})$. Our purpose is to obtain the demixing matrix $\mathbf{W}$ through an adaptive learning algorithm. We are interested only in $\Delta\mathbf{W}$, so we take the differentiation with respect to $\mathbf{W}$. Thus, the maximization of the mutual information: $I(\mathbf{y};\mathbf{x}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x})$ is equal to the maximization of $H(\mathbf{y})$, because $H(\mathbf{y}|\mathbf{x})$ would be eliminated by the differentiation with respect to $\mathbf{W}$.

Let $\mathbf{y} = g(\mathbf{u})$ where $\mathbf{u} = \mathbf{W}\mathbf{x}$ and $g(\mathbf{u})$ is an invertible bounded nonlinear vector function such as $g(\mathbf{u}) = \frac{1}{1+e^{-\mathbf{u}}}$ or $g(\mathbf{u}) = \tanh(\mathbf{u})$.



(a) (b)

Fig. 2.6 (a) $g(\mathbf{u}) = \tanh(\mathbf{u})$ (b) $g(\mathbf{u}) = \dfrac{1}{1+e^{-\mathbf{u}}}$

When $g(\mathbf{u})$ has a unique inverse, $f(\mathbf{y}) = \dfrac{f(\mathbf{x})}{|J|}$, where $|J| = \det\left(\left[\dfrac{\partial y_i}{\partial x_j}\right]_{ij}\right)$.

Therefore, $H(\mathbf{y}) = -E[\log f(\mathbf{y})] = E[\log|J|] - E[\log f(\mathbf{x})] = E[\log|J|] + H(\mathbf{x})$. Since $J = (\det \mathbf{W})\prod_{i=1}^{N}\dfrac{\partial y_i}{\partial u_i}$ results in $\log|J| = \log\det\mathbf{W} + \sum_{i=1}^{N}\log\left|\dfrac{\partial y_i}{\partial u_i}\right|$, as we differentiate $I(\mathbf{y};\mathbf{x})$ with respect to $\mathbf{W}$, the learning rules of $\mathbf{W}$ can be derived as [4]:

$$\Delta\mathbf{W} \propto \frac{\partial I(\mathbf{y};\mathbf{x})}{\partial\mathbf{W}} = \frac{\partial\log|J|}{\partial\mathbf{W}} = \mathbf{W}^{-T} + \Phi(\mathbf{u})\mathbf{x}^T,$$

where $\Phi(\mathbf{u}) = [\phi_1(u_1), \ldots, \phi_i(u_i), \ldots, \phi_N(u_N)]^T$ and $\phi_i(u_i) = \dfrac{\partial}{\partial u_i}\log\left|\dfrac{\partial y_i}{\partial u_i}\right|$.

The natural gradient method multiplies the previous result by $\mathbf{W}^T\mathbf{W}$, which leads to an elegant form of the learning rule [11]:

$$\Delta\mathbf{W} \propto \frac{\partial\log|J|}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{W} = (\mathbf{I} + \Phi(\mathbf{u})\mathbf{u}^T)\mathbf{W}.$$

2.4.2 Frequency Domain ICA (FD-ICA)

Since the subspace method operates in the frequency domain, the FD-ICA is adopted in this BSS system.

The subspace signals $\mathbf{x}_{ss}(f,t)$ are used to obtain the adaptive ICA filter $\mathbf{W}_{ICA}(f)$,

where the learning rule of $\mathbf{W}_{ICA}(f)$ is shown by the following expression [13]:

$$\Delta\mathbf{W}_{ICA,\,i+1}(f) = \mu \cdot off - diag\left\{\left\langle \Phi(\mathbf{x}_{ss}(f,t))\mathbf{x}_{ss}{}^{H}(f,t)\right\rangle_{t}\right\}\mathbf{W}_{ICA,\,i}(f),$$

where μ denotes the learning rate, $\Phi(\mathbf{x}_{ss}(f,t)) = \left[\varphi(x_{ss,1}(f,t)),\ldots,\varphi(x_{ss,N}(f,t))\right]^{T}$ denotes the score function which is applied to each element $x_{ss,n}(f,t)$ in the vector $\mathbf{x}_{ss}(f,t)$ such that $\varphi(x_{ss,n}(f,t)) = \tanh\left(G\cdot \mathrm{Re}\left\{x_{ss,n}(f,t)\right\}\right) + j\tanh\left(G\cdot \mathrm{Im}\left\{x_{ss,n}(f,t)\right\}\right)$, where G is a gain constant.

2.5 Permutation Problem and Scaling Problem

The main goal of BSS is to obtain the separated signal vector $\mathbf{y}_{ICA}(f,t)$, which is achieved by the procedures described in the preceding sections. But for further applications of the separated signals, there are other problems need to be solved, which are the scaling problem and the permutation problem.

Since the original BSS system simply "separates" the mixture signals, there exists some magnitude distortions and incorrect permutation of the separated signals. The former may cause serious problem once these separated signals are used in the subsequent signal processing tasks, which is the so-called scaling problem, and the latter can mess up the signals at transformation from frequency domain back to time domain, which is the so-called permutation problem. Therefore, these problems should be studied for our further usage.

2.5.1 Permutation Problem

For finding the solutions of the permutation problem, there are two conventional methods: the one based on the direction of arrival (DOA) and the other based on the information among the adjacent frequencies. As shown in [15], a hybrid method combining

these two can solve the problem more confidently.

In order to solve the permutation problem in the frequency domain, let us look into the structure of the mixing matrix elements $A_{m,n}(f) = |A_{m,n}(f)| e^{-j2\pi f \tau_{m,n}}$. In the beamforming theory [21], $|A_{m,n}(f)|$ is set to 1 and $\tau_{m,n}$ is modeled as:

$$\tau_{m,n} = -\frac{d_m}{c} \cos\theta_n ,$$

where d_m means the position of the m-th microphone, θ_n is the angle of direction of the n-th source and c represents the sound speed. In order to fit the current situation, $A_{m,n}(f)$ is remodeled as:

$$A_{m,n}(f) = |A_{m,n}(f)| e^{j\left(2\pi f \frac{d_m}{c} \cos\theta_n + \phi_n\right)},$$

where ϕ_n is the phase modulation of n-th source.

The permutation matrix $\mathbf{P}(f)$ of each frequency f is the goal we want to achieve in producing the separated signals. Since $\mathbf{P}(f)$ can be realized as a row permutation of identity matrix $\mathbf{I}$ at each frequency f, $\mathbf{P}(f)$ can be transmuted into a function $\Pi_f(k)$ where k represents the k-th row of $\mathbf{P}(f)$ and the function returns the column index of the one-and-only nonzero element in this row. Therefore, the identification of the permutation function Π_f is equivalent to the one of $\mathbf{P}(f)$.

When the ICA method does separate the source signals at each frequency f, there exists a permutation matrix $\mathbf{P}(f)$ and a diagonal scaling matrix $\mathbf{\Lambda}(f)$ such that:

$$\mathbf{\Lambda}(f)\mathbf{P}(f)\mathbf{W}(f)\mathbf{A}(f) = \mathbf{I} .$$

In this case, $\mathbf{A}(f)$ can be approximated as $\mathbf{W}^{+}(f)\mathbf{P}^{-1}(f)\mathbf{\Lambda}^{-1}(f)$, where $\mathbf{W}^{+}(f)$ is the Moore-Penrose pseudoinverse of $\mathbf{W}^{+}(f)$. Arbitrarily choose two elements in the n-th column of $\mathbf{A}(f)$ with different row index m and m' and then we can remove the effect of the unknown $\mathbf{P}(f)$ and $\mathbf{\Lambda}(f)$ to identify the angle of the n-th source θ_n by the

following derivation [15]:

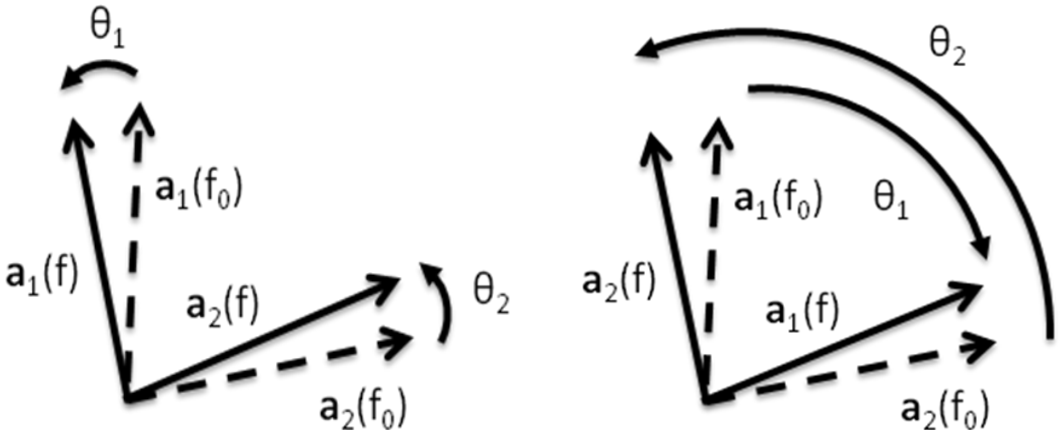
$$\frac{A_{m,n}}{A_{m',n}} = \frac{\left[\mathbf{W}^{+}\mathbf{P}^{-1}\mathbf{\Lambda}^{-1}\right]_{m,n}}{\left[\mathbf{W}^{+}\mathbf{P}^{-1}\mathbf{\Lambda}^{-1}\right]_{m',n}} = \frac{\left[\mathbf{W}^{+}\right]_{m,\Pi_{\omega}^{-1}(n)}}{\left[\mathbf{W}^{+}\right]_{m',\Pi_{\omega}^{-1}(n)}} = \frac{\left|A_{m,n}\right|}{\left|A_{m',n}\right|} e^{j2\pi f \frac{d_m - d_{m'}}{c}\cos\theta_n},$$

$$\theta_n = \arccos \frac{\arg\left(\frac{\left[\mathbf{W}^{+}(f)\right]_{m,\Pi_f^{-1}(n)}}{\left[\mathbf{W}^{+}(f)\right]_{m',\Pi_f^{-1}(n)}}\right)}{2\pi f \frac{d_m - d_{m'}}{c}}.$$

The above method is developed based on the direction of arrival, which is unreliable at low frequencies where the phase difference due to the small interval of linear-arranged microphones and the high frequencies, where the spatial aliasing appears. The hybrid method uses the information among the adjacent frequencies in all frequencies to make the solution more reliable.

Considering the frequency resolution of STFT Δf, for the current processing frequency f, its adjacent frequencies are the reference frequencies $f_0 = f - k \cdot \Delta f$, $k = 1, \ldots, K$, where K is an adjustable constant for confidence measurement. Let $\mathbf{a}_n(f)$ denote the n th column vector of $\mathbf{A}(f)$ at f, assuming $\left|A_{m,n}(f)\right| = 1$ for simplicity [2]:

$$\mathbf{a}_n(f) = \begin{pmatrix} e^{-j2\pi f \tau_{1,n}} \\ \vdots \\ e^{-j2\pi f \tau_{M,n}} \end{pmatrix}, \quad \mathbf{a}_n(f_0) = \begin{pmatrix} e^{-j2\pi (f-\Delta f) \tau_{1,n}} \\ \vdots \\ e^{-j2\pi (f-\Delta f) \tau_{M,n}} \end{pmatrix}.$$



(a) (b)

Fig. 2.7 (a) With Correct Permutation (b) With Incorrect Permutation

As Fig. 2.7 shows, we can say that "$\mathbf{a}_n(f)$ is the result of $\mathbf{a}_n(f_0)$ rotated by the rotation angle θ_n." Incorrect permutations usually results in a larger magnitude of rotation angle θ_n for each column vector $\mathbf{a}_n(f)$. Therefore, θ_n is expected to be the smallest when the permutation matrix is correct.

Let $\mathbf{W}^+(f)$ denote an estimation of the mixing matrix $\mathbf{A}(f)$, and $\mathbf{P}(f)$ denotes an arbitrary permutation matrix, which exchanges the row vectors of the transposed estimation matrix $(\mathbf{W}^+)^T$. The arbitrary permuted matrix $\bar{\mathbf{A}}(f)$ is calculated by the following expression:

$$\left(\bar{\mathbf{A}}(f)\right)^T = \mathbf{P}(f)\left(\mathbf{W}^+(f)\right)^T,$$

where $\bar{\mathbf{A}}(f) = \left[\bar{\mathbf{a}}_1(f), \ldots, \bar{\mathbf{a}}_N(f)\right]$ and $\bar{\mathbf{a}}_n(f)$ denotes the n-th permuted column vector. Then, the cosine of the angle θ_n between $\bar{\mathbf{a}}_n(f)$ and $\bar{\mathbf{a}}_n(f_0)$ is calculated by the following expression:

$$\cos\theta_n = \frac{\bar{\mathbf{a}}_n^H(f)\bar{\mathbf{a}}_n(f_0)}{\left\|\bar{\mathbf{a}}_n(f)\right\| \cdot \left\|\bar{\mathbf{a}}_n^H(f_0)\right\|}.$$

When the permutation matrix is a correct one, $\cos\theta_n$ is expected to be largest. Thus, the cost function $F(\mathbf{P}, k)$ can be written as [2]:

$$F(\mathbf{P},k)=\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\,,$$

where k denotes the index of the reference frequency f_0. When $\max_{\mathbf{P}} F(\mathbf{P},k)$ is close to $F(\mathbf{P},k)$ with other permutations, it may be difficult to determine which permutation is correct. Therefore, the confidence measure $C(k)$ is defined to represent how reliable the reference frequency f_0 is. The value of $C(k)$ is calculated by the following expression:

$$C(k)=\max_{\mathbf{P}\in\Omega}F(\mathbf{P},k)-\max_{\mathbf{P}\in\Omega'}F(\mathbf{P},k)\,,$$

where Ω denotes the set of all possible $\mathbf{P}$ and Ω' denotes the set of all possible $\mathbf{P}$ without $\hat{\mathbf{P}}_k=\arg\max_{\mathbf{P}\in\Omega}F(\mathbf{P},k)$. The approximate permutation matrix $\hat{\mathbf{P}}$ is obtained by the following expression:

$$\hat{\mathbf{P}}=\arg\max_{\mathbf{P}}F(\mathbf{P},\hat{k})\,,$$

where $\hat{k}=\max_{k}C(k)$.

The above method basically relies on the information associated with the adjacent frequencies, and there is a similar kind of method based on the interfrequency correlations. According to the observations of adjacent frequency spectrum envelopes, it can be found that the correlations among these adjacent frequencies are relatively higher than the others. Therefore, the interfrequency correlations are informative to determine the nearby frequency permutations.

The envelope of a separated signal $Y_i(f,t)$ is $|Y_i(f,t)|$, and the correlation between two signals $x(t)$ and $y(t)$ is defined as:

$$cor(x,y)=\frac{\mu_{x\cdot y}-\mu_x\cdot\mu_y}{\sigma_x\cdot\sigma_y}\,,$$

where μ_x, μ_y, and $\mu_{x\cdot y}$ are the means of $x(t)$, $y(t)$, and $x(t)\cdot y(t)$, respectively and σ_x and σ_y are the standard deviations of $x(t)$ and $y(t)$, respectively. Note that

$cor(x, y) = 0$ for the uncorrelated $x(t)$ and $y(t)$ and $cor(x, x) = 1$ and $cor(y, y) = 1$ for all $x(t)$ and $y(t)$.

By assuming that the adjacent frequencies are highly correlated, for each frequency f, the sum of correlations with the adjacent frequencies within a small range δ is maximized if the permutation function Π_f is correct. The priori condition of this maximization process is that the adjacent frequencies are fixed to the right permutation. Let F be the set of fixed frequencies, and then we can get the permutation function Π_f by exhausting all the possible permutation to maximize the sum of adjacent frequency correlations by the following expression [15]:

$$\Pi_f = \arg\max_{\Pi} \sum_{|g-f|<\delta,\, g\in F} \sum_{k=1}^{N} cor\left(\left|Y_{\Pi(k)}(f,t)\right|, \left|Y_{\Pi_g(k)}(g,t)\right|\right).$$

$$\Pi_f = \arg\max_{\Pi} \sum_{g\in Ha\cap F} \sum_{k=1}^{N} cor\left(\left|Y_{\Pi(k)}(f,t)\right|, \left|Y_{\Pi_g(k)}(g,t)\right|\right)$$

In the above maximization process, only the adjacent frequencies in the set of fixed frequencies are added into the summation. Therefore, we can use the DOA method to identify some permutation-predetermined frequencies in advance.

The permutation problem is solved by the following methods in the order of: the DOA approach as shown in Fig. 2.8 (a), the high interfrequency correlation method as shown in Fig. 2.8 (b), the harmonic frequency correlation method as shown in Fig. 2.9 (a), and finally the low interfrequency correlation as shown in Fig. 2.9 (b).

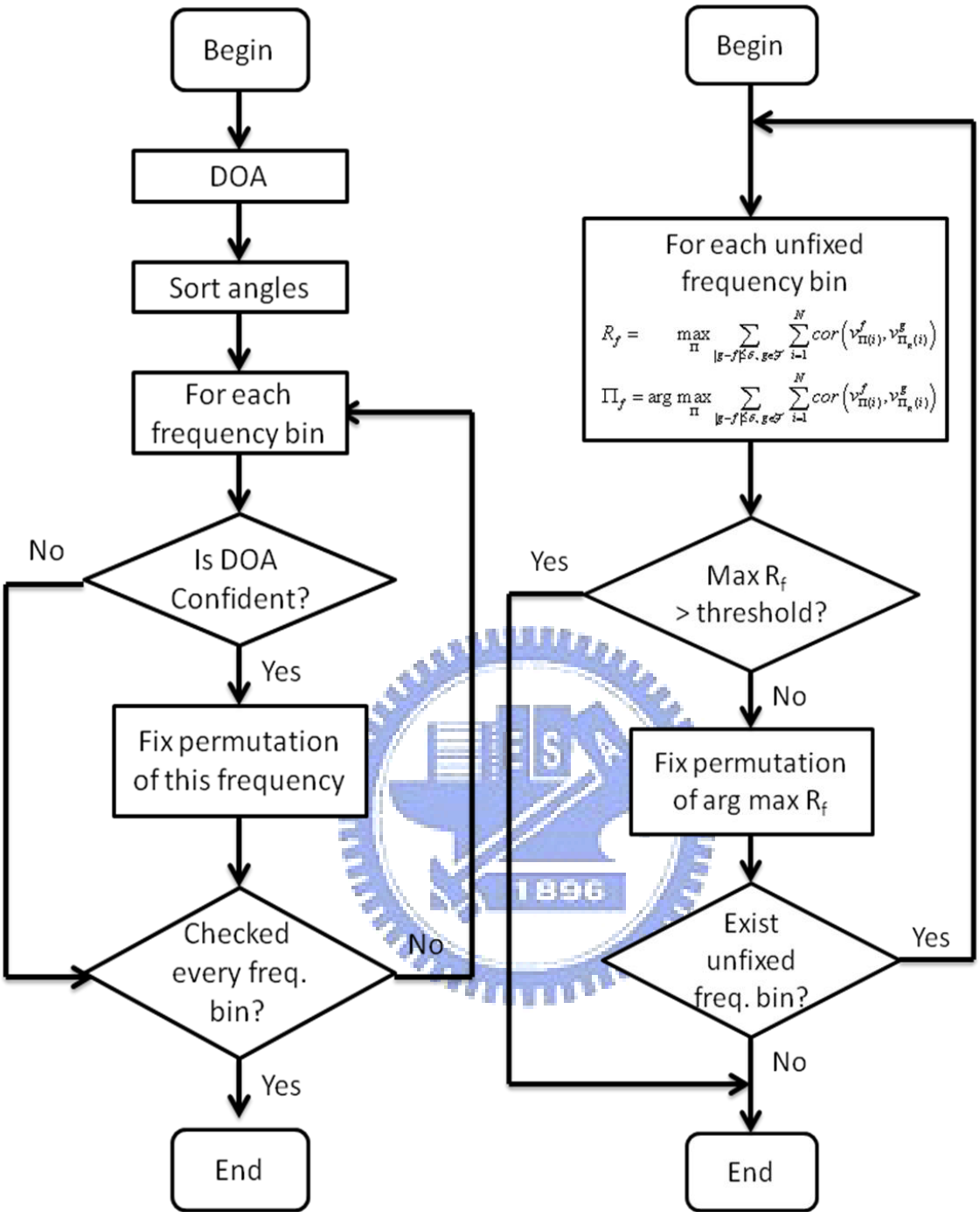


Fig. 2.8 (a) DOA Approach (b) High Interfrequency Correlation

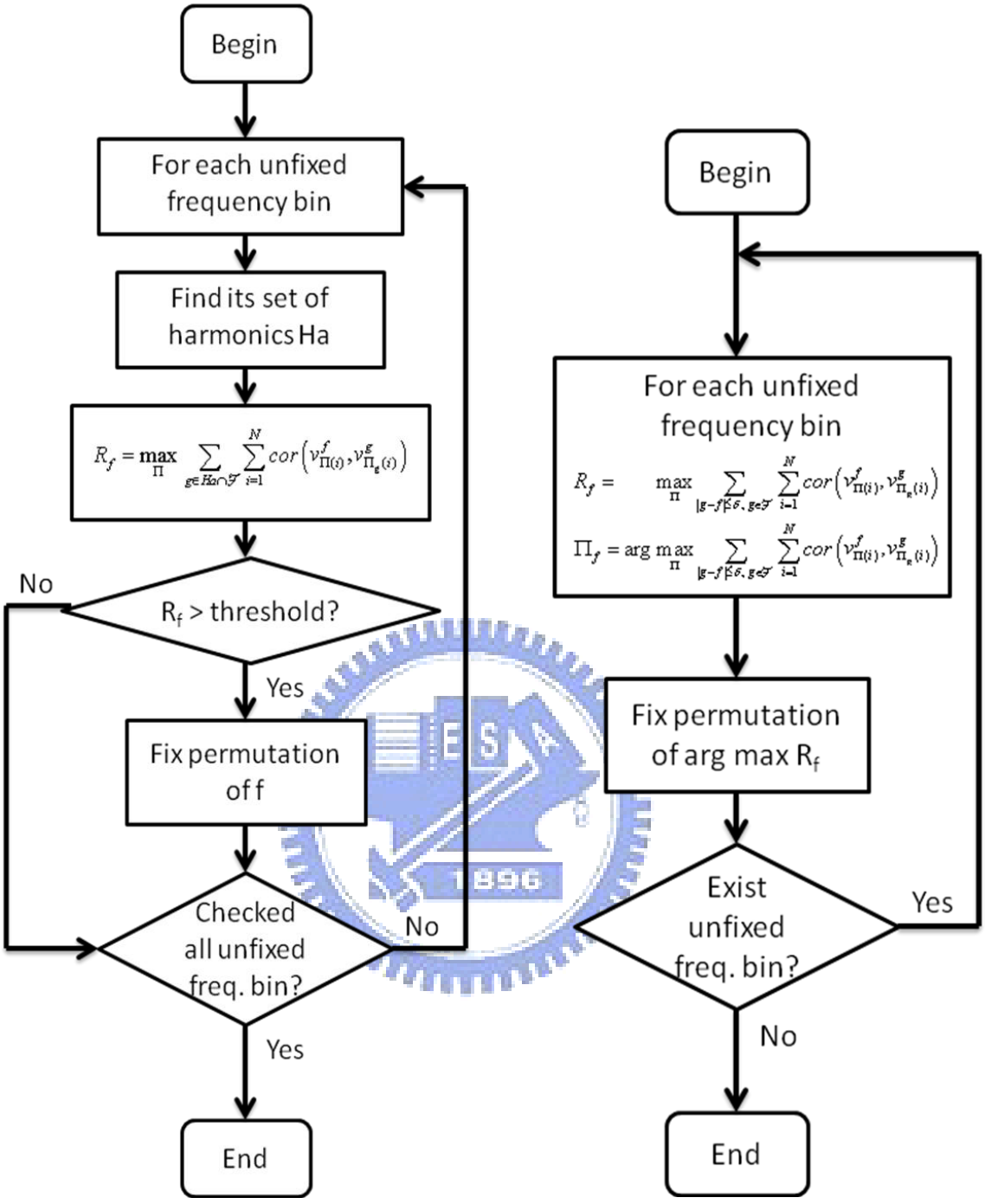


Fig. 2.9 (a) Harmonic Frequency Correlation (b) Low Interfrequency Correlation

2.5.2 Scaling Problem

Scaling problem can be solved by filtering individual output of the separation filter by

the pseudoinverse $\mathbf{B}(f)^+$ of the demixing matrix $\mathbf{B}(f)=\mathbf{P}(f)\mathbf{W}_{ICA}(f)\mathbf{W}_{PCA}(f)$. As $y_{p,n}(f,t)$, the n-th component of $\mathbf{y}_p(f,t)$, is filtered by $\mathbf{B}(f)^+$ with the following expression [2]:

$$\mathbf{y}_n(f,t)=\mathbf{B}(f)^+\left[0,\ldots,0,\,y_n(f,t),\,0,\ldots,0\right]^T,$$

where $\mathbf{y}_n(f,t)=\left[y_{1,n}(f,t),\ldots,y_{M,n}(f,t)\right]^T$ denotes the signal vector recovered from $y_{p,n}(f,t)$, and $y_{m,n}(f,t)$ denotes the recovered signal of the n-th source observed at the m-th microphone.

Therefore, by setting an arbitrary microphone number $\tilde{m}$ as the reference one, the magnitude of the recovered element $y_{\tilde{m},n}(f,t)=B^+_{\tilde{m},n}(f)y_{p,n}(f,t)$ is normalized to the $\tilde{m}$-th microphone, where $B^+_{\tilde{m},n}(f)$ denotes the $(\tilde{m},n)$-th element of $\mathbf{B}(f)^+$.

The scale recovered signal vector $\mathbf{y}(f,t)=\left[y_{\tilde{m},1}(f,t),\ldots,y_{\tilde{m},N}(f,t)\right]^T$ can be obtained by the following equation [2]:

$$\mathbf{y}(f,t)=\mathbf{\Lambda}(f)\mathbf{y}_p(f,t),$$

where $\mathbf{\Lambda}(f)=\tilde{\mathbf{B}}^+_{\tilde{m}}(f)=diag\left[B^+_{\tilde{m},1}(f),\ldots,B^+_{\tilde{m},N}(f)\right]$ is an $N\times N$ diagonal matrix.

Another way of solving the scaling problem is using the minimal distortion principle [16]. Since we conduct the separation method under the blind condition, the mixing matrix $\mathbf{A}(f)$ is unknown. For simplicity, we assume the permutation matrix $\mathbf{P}(f)=\mathbf{I}$ in the following derivation, the ideal scaling matrix $\mathbf{\Lambda}(f)$ should satisfy the equation that:

$$\mathbf{\Lambda}(f)\mathbf{W}_{ICA}(f)\mathbf{W}_{PCA}(f)\mathbf{A}(f)=diag\left[\mathbf{A}(f)\right].$$

Once the signals are well-separated by the preceding ICA method, there exists another diagonal matrix $\mathbf{D}(f)$ such that $\mathbf{W}_{ICA}(f)\mathbf{W}_{PCA}(f)\mathbf{A}(f)=\mathbf{D}(f)$. Hence, the unknown mixing matrix $\mathbf{A}(f)$ can be estimated as $\mathbf{W}^+(f)\mathbf{D}(f)$ where $\mathbf{W}^+(f)$ is the

Moore-Penrose pseudoinverse of the separation matrix $\mathbf{W}_{ICA}(f)\mathbf{W}_{PCA}(f)$. Therefore, the estimation of $\mathbf{\Lambda}(f)$ equals to $diag\left[\mathbf{W}^{+}(f)\right]$, which is an approximation to the solution of the scaling problem in the FD-ICA.

2.6 Convolutive BSS

The above ICA method is effective for the cases with no reflections, but for the convolutive mixture microphone array signals, the separation quality is severely degraded by the room reflection. Hence, we adopt another BSS method, which is developed based on a multiple decorrelation approach and a least squares optimization to estimate the mixing matrix $\mathbf{A}$ and the demixing matrix $\mathbf{W}$ [17].

The spatial correlation matrix $\mathbf{R}(f)=\left\langle \mathbf{x}(f,t)\mathbf{x}^{H}(f,t)\right\rangle_{t}$ can be written as :

$$\mathbf{R}(f)=\mathbf{A}(f)\mathbf{\Lambda}_{s}(f)\mathbf{A}^{H}(f)+\mathbf{\Lambda}_{n}(f),$$

where $\mathbf{\Lambda}_{s}(f)$ and $\mathbf{\Lambda}_{n}(f)$ are diagonal due to the independence assumption of the source signals [17]. The cross-power-spectrum average $\bar{\mathbf{R}}(f,t)$ can then be written as:

$$\bar{\mathbf{R}}(f,t)=\mathbf{A}(f)\mathbf{\Lambda}_{s}(f,t)\mathbf{A}^{H}(f)+\mathbf{\Lambda}_{n}(f,t).$$

Therefore, we want to find $\mathbf{W}(f)$, $\mathbf{\Lambda}_{s}(f,t)$, and $\mathbf{\Lambda}_{n}(f,t)$ which satisfy the following equation:

$$\mathbf{\Lambda}_{s}(f,t)=\mathbf{W}(f)\left(\bar{\mathbf{R}}(f,t)-\mathbf{\Lambda}_{n}(f,t)\right)\mathbf{W}^{H}(f).$$

A least squares optimization can be used to find the estimations $\hat{\mathbf{W}}(f)$, $\hat{\mathbf{\Lambda}}_{s}(f,t)$, and $\hat{\mathbf{\Lambda}}_{n}(f,t)$ which can minimize:

$$J=\sum_{f=1}^{T}\sum_{t=1}^{K}\left\|\mathbf{W}(f)\left(\bar{\mathbf{R}}(f,t)-\mathbf{\Lambda}_{n}(f,t)\right)\mathbf{W}^{H}(f)-\mathbf{\Lambda}_{s}(f,t)\right\|^{2},$$

where T denotes the frame size of the STFT and K denotes the range of optimization process along the time axis. The solutions can be obtained by using the gradient descent

algorithm [17].

2.7 Evaluation of the BSS Performance

One way to evaluate the BSS performance is to measure the signal to interference ratio (SIR). The definition of SIR is described below:

$$SIR = \frac{10}{N}\sum_{i=1}^{N}\log_{10}\frac{\left\langle \left| y_{i,S_i}(t) \right|^2 \right\rangle_t}{\left\langle \sum_{j\neq i} \left| y_{i,S_j}(t) \right|^2 \right\rangle_t}.$$

The overall separation matrix $\mathbf{W}$ is trained by the microphone array signals, which record the simultaneously played source signals. In order to test the effect of suppressing other source signals for each output separated signal, we only play one source signals at a time. The signal $y_{i,S_j}(t)$ represents the i-th output separated signal with only the j-th source signal being active. For $j = i$, $\left\langle \left| y_{i,S_i}(t) \right|^2 \right\rangle_t$ denotes the power of the i-th desired separated signal averaging over the time axis, and for all $j \neq i$, $\sum_{j\neq i}\left\langle \left| y_{i,S_j}(t) \right|^2 \right\rangle_t$ denotes the sum of other interference power from other source signals of the i-th separated signal.

Chapter 3

3D Acoustic Signal Synthesis

3.1 Acoustic Transfer Function Pool (ATF-Pool)

3.1.1 Measurement of ATFs

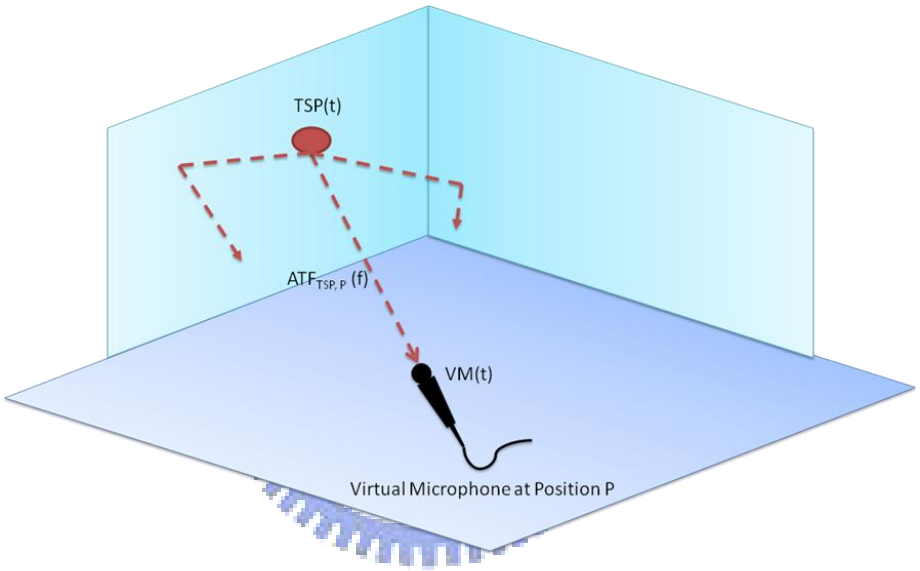


Fig. 3.1 Estimation of ATF by Using the TSP Signal

By transmitting a time-stretched pulse (TSP) signal from each source location to each virtual microphone position, the acoustic transfer functions (ATF) are estimated in the frequency domain with the following function:

$$ATF_{ij}(f) = \frac{VM_j(f)}{TSP_i(f)},$$

where $ATF_{ij}(f)$ denotes the acoustic transfer function from the i-th source location to the j-th virtual microphone location, $VM_j(f)$ denotes the j-th virtual microphone signal and $TSP_i(f)$ denotes the TSP signal from the i-th source location.

The frequency response of the time-stretched pulse signal we adopted in this thesis is shown as the following function [22], [23]:

$$TSP(k) = \begin{cases} e^{j\pi\frac{4Mk^2}{N^2}}, 0 \le k \le \frac{N}{2} \\ TSP(N-k), \frac{N}{2} < k < N \end{cases},$$

where N is the length of TSP signal and M is the TSP stretch parameter. For a TSP signal with $N = 2048$ and $M = 64$, the time domain response is shown as Fig. 3.2 below.

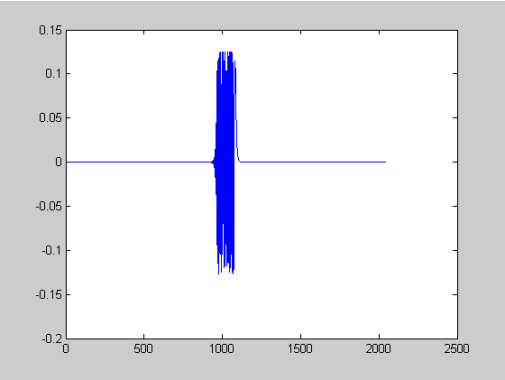


Fig. 3.2 TSP with N = 2048 and M = 64 in the Time Domain

The length N of the TSP signal determine the phase resolution and the TSP stretch parameter M has a trade-off between the signal to noise ratio (SNR) and the convergence. However, the averaged error level of the TSP signal is less than -100 dB [23], which is insignificant in our measurement of ATF; hence, the value of M can be assigned arbitrarily in our case.

3.1.2 ATF Interpolation

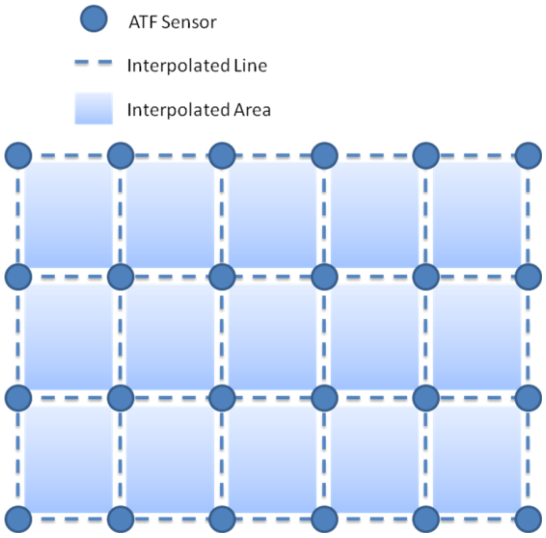


Fig. 3.3 Weighted Linear Interpolation of ATF

As revealed in the section 3.1.1, it is physically impossible to measure all the ATFs for each source-microphone pair, so it is reasonable to obtain the unmeasured ATF using the weighted linear interpolation method [19]. The ATF-Pool is consisted of the recorded ATF measurement. The ATFs from the source locations to the virtual listening point are synthesized only at the time they are needed, which can lower the amount of memory space requirement of the ATF-Pool.

3.2 Head-Related Transfer Function (HRTF)

It is necessary that having a head-related transfer function (HRTF) database in order to present the 3D spatial feeling through the headphone. There are several open-source HRTF database freely, such as [24]. The head-related transfer functions are measured with the dummy head recording in the anechoic chamber to simulate the reflection and diffraction characteristics of the torso, head and pinnae. For the same source signal, the head-related transfer function varies with different source elevation angles, azimuth angles and distances between the source and the head. Therefore, HRTF is actually a function of the elevation

angle ϕ, the azimuth angle θ and the distance r. The head-related transfer function at an arbitrary position is interpolated from the nearby captured HRTFs.

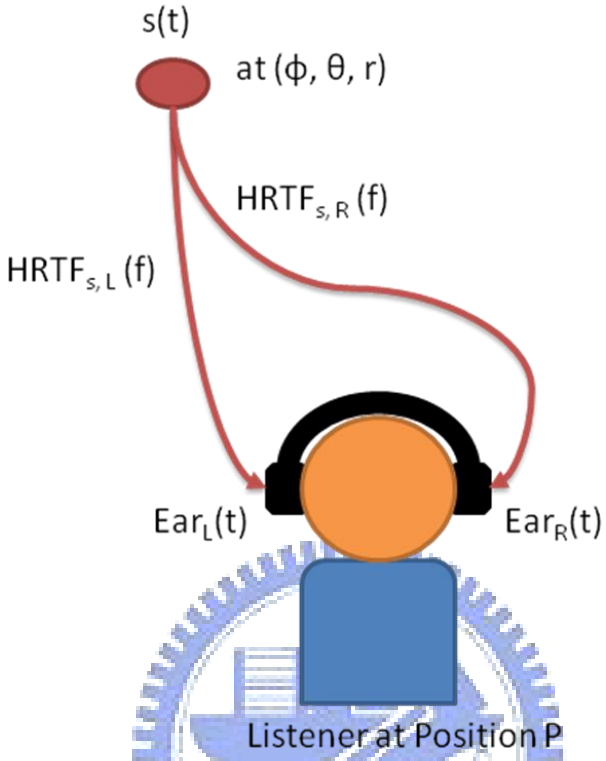
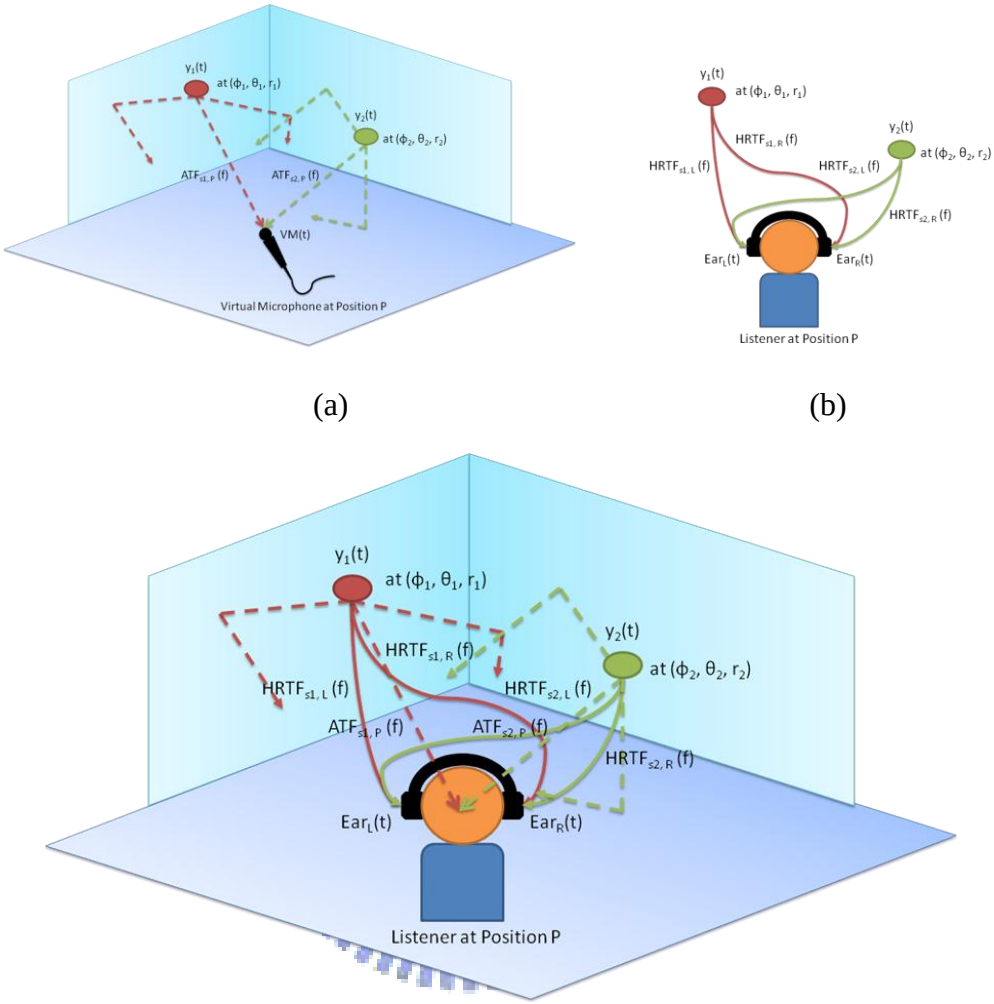


Fig. 3.4 HRTFs from $s(t)$ to $Ear_L(t)$ and $Ear_R(t)$

There are two channels for each HRTF measurement, which are left ear impulse response $HRIR_L(t)$ and right ear impulse response $HRIR_R(t)$. The frequency responses of $HRIR_L(t)$ and $HRIR_R(t)$ are $HRTF_L(f)$ and $HRTF_R(f)$ respectively. The calculations of $HRTF_L(f)$ and $HRTF_R(f)$ at a certain source position (ϕ, θ, r) related to the listener are shown as the following expression:

$$HRTF_L(f) = \frac{Ear_L(f)}{S(f)}, \quad HRTF_R(f) = \frac{Ear_R(f)}{S(f)}.$$

3.3 Combining HRTF and ATF



(a) (b)

(c)

Fig. 3.5 Combining ATF and HRTF

(a) ATF for Each Separated Signal (b) HRTF for Each Separated Signal

(c) 3D Acoustic Signal Synthesis

There are many different kinds of blind source separation methods, but 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 fully given. The performance of the separation results may degrade owing to the channel noise, room reflections and some

violations of the source signal stochastic model assumptions, which are usually different for speech signals and instrument signals. However, the interferences which are introduced by other source signals can be less significant as our main purpose of separating these source signals is to synthesize them back together.

With the HRTF database and the ATF-pool, the audience is allowed to choose the arrangement of the source signals and listening position arbitrarily. In other words, the audience can have one source signal at the left side and another at the right side, which are unrelated to the original geometric spots of these source signals in the room. The spatial impression is presented with the headphone by utilizing the HRTF database and the ATF-pool to simulate the user-customized listening scenarios. Therefore, the audience can hear the synthesized 3D feeling audio signals at their own sweet spots.

For the point which does not have an ATF measurement, the estimation of its ATF is calculated by a weighted linear interpolation from the nearby measured ATFs. The weighted linear interpolation method also appears in the calculation of HRTF when the desired spatial position of HRTF cannot be found from the HRTF database.

Let $y_i(t)$ be the separated signal corresponding to the source signal $s_i(t)$, head-related impulse response (HRIR) be the time domain filter of HRTF where $HRIR_{y_i,L}(t)$ and $HRIR_{y_i,R}(t)$ are the left and right ear HRIR from the position of $y_i(t)$ to the position of the head, and $AIR_{y_i}(P,t)$ be the acoustic impulse response (AIR) as known as the time domain filter of ATF from the position of $y_i(t)$ to the position P. The convolutive results of the two ear signals $Ear_L(t)$ and $Ear_R(t)$ can be derived as the following expressions:

$$Ear_L(t) = \sum_{i=1}^{N} AIR_{y_i}(P,t) * \left(HRIR_{y_i,L}(t) * y_i(t) \right),$$

$$Ear_R(t) = \sum_{i=1}^{N} AIR_{y_i}(P,t) * \left(HRIR_{y_i,R}(t) * y_i(t) \right).$$

The above expression can also be represented in the frequency domain as $Ear_L(f)$ and $Ear_R(f)$:

$$Ear_L(f) = \sum_{i=1}^{N} ATF_{Y_i}(P, f) HRTF_{Y_i,L}(f) Y_i(f),$$

$$Ear_R(f) = \sum_{i=1}^{N} ATF_{Y_i}(P, f) HRTF_{Y_i,R}(f) Y_i(f).$$

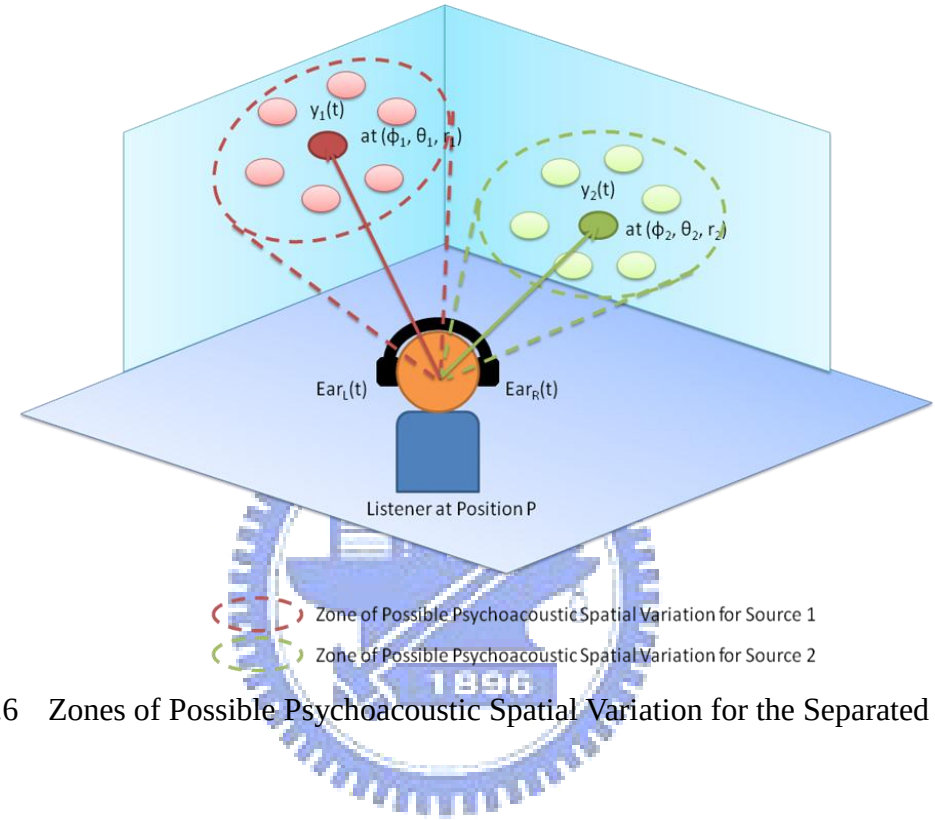


Fig. 3.6 Zones of Possible Psychoacoustic Spatial Variation for the Separated Signals

Owing to the interference in the separated signals, the psychoacoustic spatial impression may be degraded by the interaural time difference (ITD) and interaural level difference (ILD). The zone of possible psychoacoustic spatial variation for each source alters based on the SIR of each separated signal. The remaining interference for the i-th separated signal affects the j-th separated signal for all $j \neq i$. The subject performance degradation for such interferences depends on the human psychoacoustic resolutions of the azimuth angles, the elevation angles and the distance. For a far-field virtual listening point, the distance resolution would be less significant due to the human psychoacoustic characteristics, and the azimuth angles and the elevation angles dominate the main 3D acoustic feeling.


E
S
A
1896

Chapter 4

Experiment Results

4.1 Descriptions of the Adopted BSS System

We adopt the frequency domain independent component analysis (FD-ICA) in this paper with principle component analysis (PCA) as a preprocessing dimension reduction method. We choose the Infomax method combined with the natural gradient method due to the popularity and simplicity of these two methods. The signals are separated in the time-frequency domain and each frequency band is separated individually so that the permutation and scaling problems should be fixed after the ICA process. We solve the permutation problem by the combination of the DOA approach, the neighboring correlation approach and the harmonic frequency approach. The scaling problem is solved by using the minimum distortion principal method. For the convolutive BSS method, we adopt a least squares optimization technique based on the cross-power-spectrum approach with the gradient descent algorithm. The flow diagram for the overall BSS system is shown as Fig. 4.1 below.

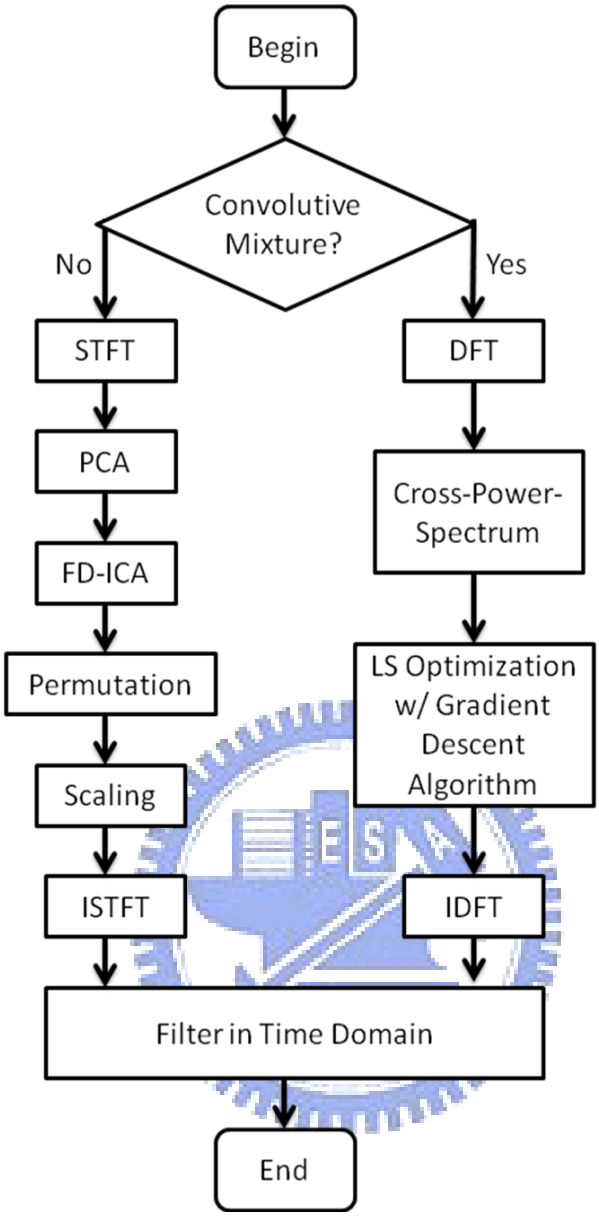


Fig. 4.1 Flow Diagram of the Adopted BSS System

Fig. 4.2 shows the arrangement of source signals and the microphone array on the X-Y plane. Two source signals are located 3.00 (m) away from each other and the interval length of the microphone array is equal to 0.50 (m). The middle point of the two source signals is 3.00 (m) away from the center of the seven microphone signals.

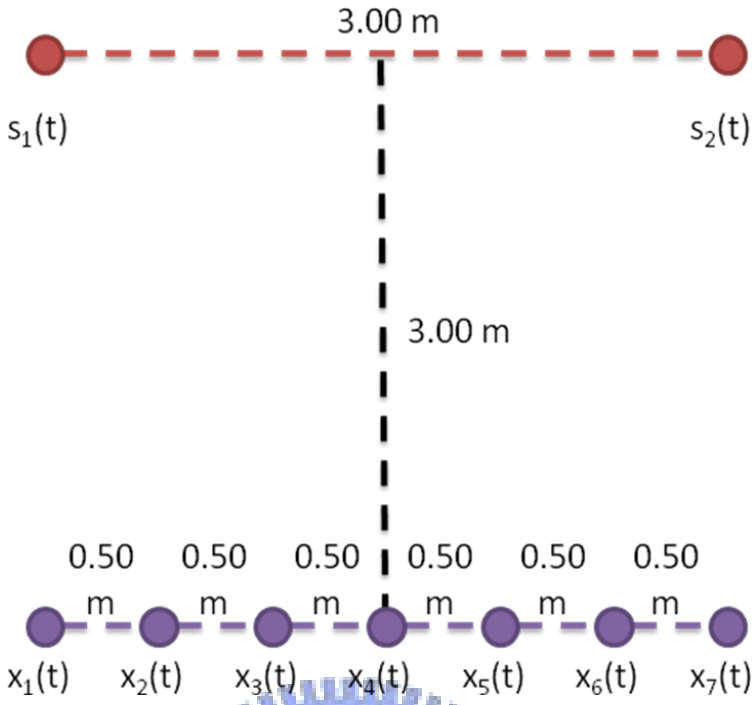


Fig. 4.2 Arrangement of the Source Signals and the Microphone Array

The settings of detailed parameters about the BSS system are shown in the Table 4.1. The thresholds th_θ and th_U are assigned to make sure the DOA calculation is confident, and the threshold th_{Ha} is adjusted based on the number of sources and the size of the harmonic set. The range K affects the convergence speed of the convolutive BSS method. For a larger K value, it takes more computational time to search for the valid demixing matrix $\mathbf{W}$.

Table 4.1 Settings of the BSS System Parameters

Parameters of the BSS System	Values
Sampling Frequency	44.1 kHz
Number of Microphones, M	7
Number of Sources, N	2
Length of STFT, T	8196 pt
Frame Shift of STFT	128 pt

Window Function	Hamming
Thresholds of confident DOA	$th_\theta = 1.5\sigma_\theta$, $th_U = 10\,\text{dB}$
Distance for Interfrequency Correlations, δ	$3 \cdot \Delta f$
Set of Harmonic Frequencies	$\{2f, 2f \pm \Delta f, 3f, 3f \pm \Delta f\}$
Threshold of Harmonic Correlations, th_{Ha}	1.2
Learning Rate, μ	1.0
Number of Iterations	1000
Nonlinear Function, $g(\mathbf{u})$	$\tanh(G \cdot \text{Re}\{\mathbf{u}\}) + j\tanh(G \cdot \text{Im}\{\mathbf{u}\})$
Gain of Score Function, G	100
Range of LS Optimization, K	5

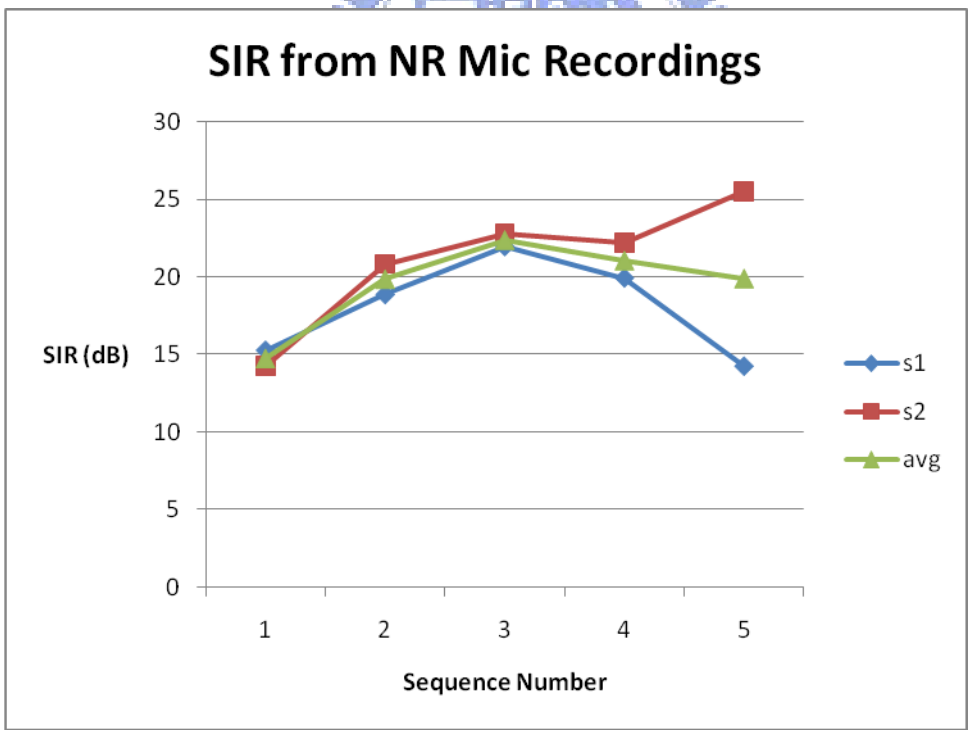


Fig. 4.3 SIR of the Demixing Matrix from No Reflection (NR) Microphone Recordings

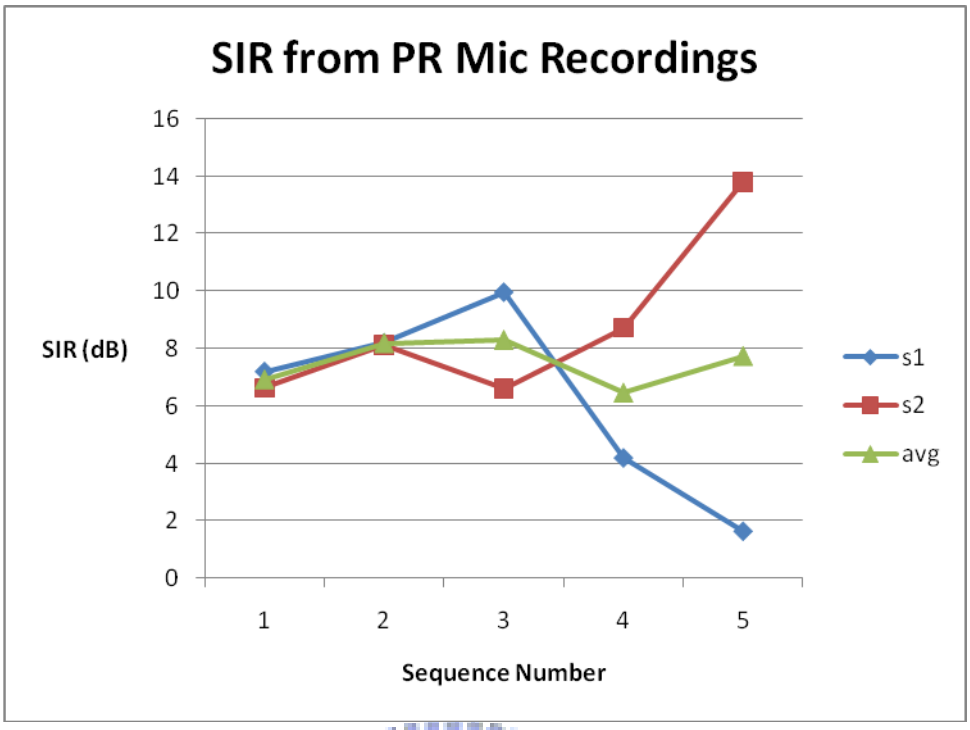


Fig. 4.4 SIR of the Demixing Matrix from Perfect Reflector (PR) Microphone Recordings

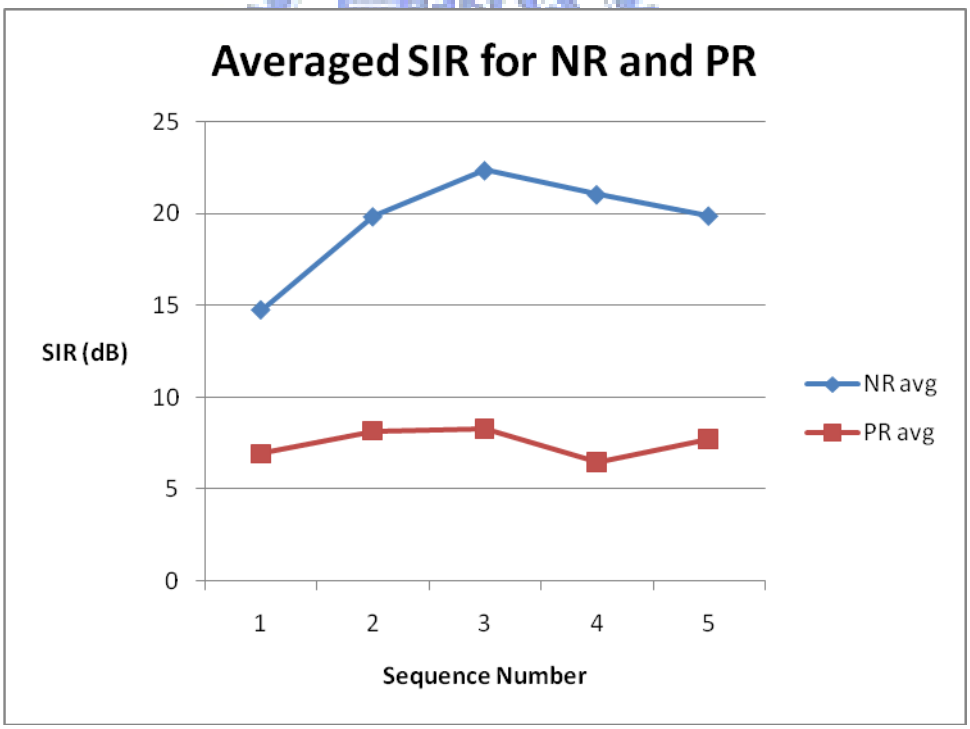


Fig. 4.5 Averaged SIR of NR and PR

Table 4.2 Source Types in Sequence Numbers

Sequence Number	Sequence Abbreviation	Source 1	Source 2
1	f01m01	Chinese speech, female	Chinese speech, male
2	instru	instrument, string 1	instrument, string 2
3	speech	Japanese speech, female	Japanese speech, male
4	winter	instrument, drums	instrument, piano
5	wistru	instrument, string 1	instrument, piano

There are five sets of data being processed from top to toe, which are “f01m01”, “instru”, “speech”, “winter”, and “wistru”. The “f01m01” sequences are two Chinese speech signals of a man and a woman; the “instru” sequences are two string instrument signals; the “speech” sequences are two Japanese speech signals of a man and a woman; the “winter” sequences are instrument signals of drums and a piano; the “wistru” sequences are a string in “instru” and the piano in “winter”. The lengths of all these wave files are about 6.8 second.

The effectiveness of the demixing matrix $\mathbf{W}$ can be measured as the SIR values of the microphone array signals. In Fig. 4.3, the SIR of the demixing matrix from no reflection (NR) recordings shows good performance in average. The sequence number corresponds to different test sequences which are shown in Table 4.2. When the wall material changes to the perfect reflectors (PR) in Fig. 4.4, the SIR values drop to around 7dB. In Fig. 4.5, the averaged SIRs of NR are higher than the ones of PR for all input sequences. The reason for this phenomenon can be easily understood since the reflections make the purely time-delayed BSS problem into a convolutive one. Thus, the independence of the source signals is disturbed.

For the fifth sequence “wistru”, the SIR difference of source 1 and source 2 is the

largest among the five sequences in both the NR and PR conditions. The explanation comes from the waveforms in Fig. 4.22 (c), (d) and Fig. 4.23 (c), (d). Note that the graphs of waveforms and spectrograms were normalized to the interval [-1, 1] for observation. Thus, the true amplitude cannot be observed from the waveforms of the source signals, but we can easily find that the mixture signals are dominated by the source 2 in the "wistru" sequence. Owing to the larger true magnitude of the source 2 (piano), the interference from source 2 to the separated signal 1 is still significant. On the other hand, the interference from source 1 to the separated signal 2 is insignificant in terms of the relative power ratio. However, in the two source case, the relative power ratio would be eliminated in the averaged SIR. Recall that the separated signals can be modeled as:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \mathbf{W} * \mathbf{A} * \mathbf{s} = \mathbf{U} * \mathbf{s} = \begin{bmatrix} u_{11}(t) * s_1(t) + u_{12}(t) * s_2(t) \\ u_{21}(t) * s_1(t) + u_{22}(t) * s_2(t) \end{bmatrix} = \begin{bmatrix} y_{1,S_1}(t) + y_{1,S_2}(t) \\ y_{2,S_1}(t) + y_{2,S_2}(t) \end{bmatrix},$$

where $\mathbf{U}$ denotes the overall filter of $\mathbf{s}$ to $\mathbf{y}$ and the averaged SIR is calculated as:

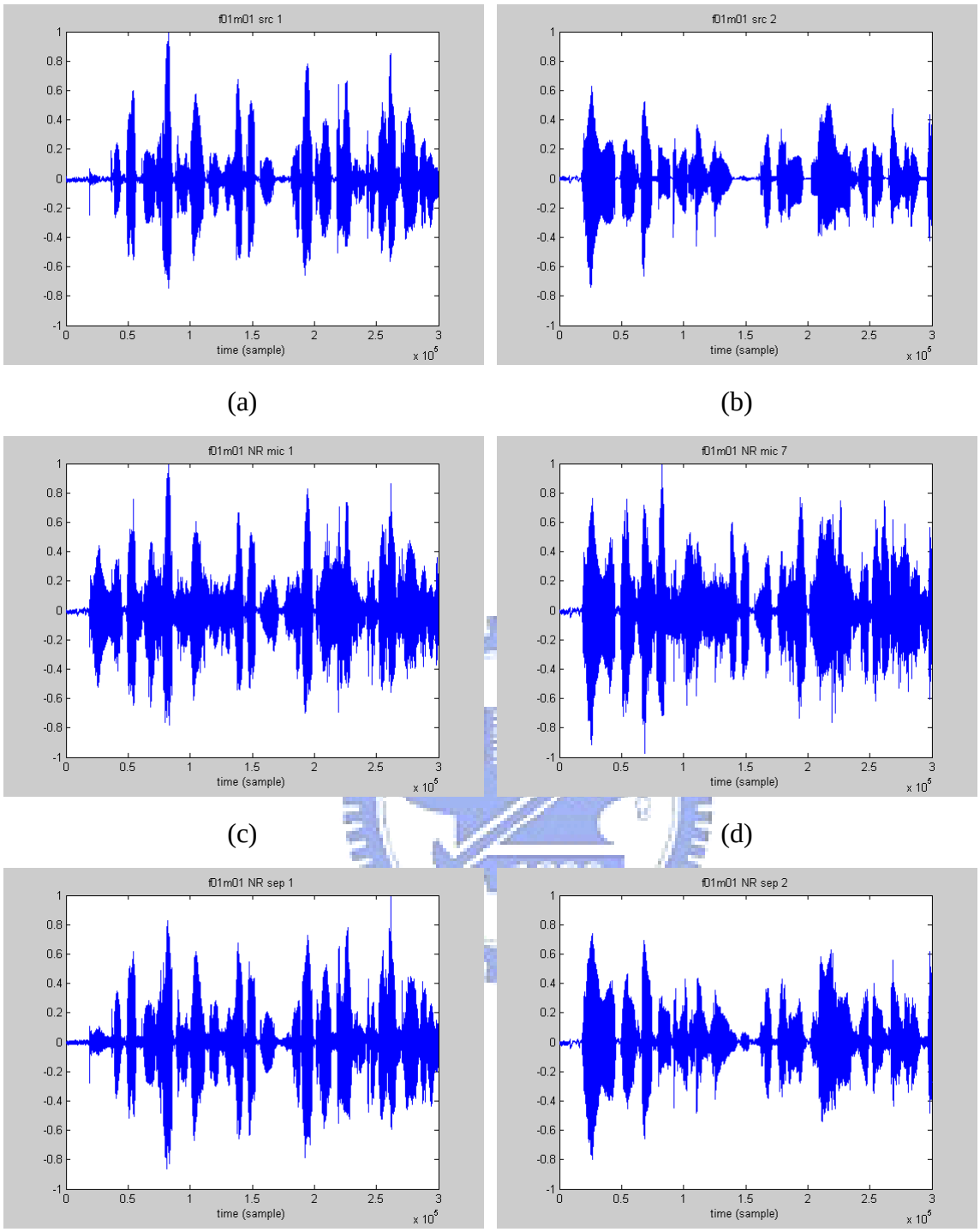
$$Averaged\ SIR = \frac{1}{2} \cdot 10 \log_{10} \left(\frac{\left\langle \left| y_{1,S_1}(t) \right|^2 \right\rangle_t}{\left\langle \left| y_{1,S_2}(t) \right|^2 \right\rangle_t} \cdot \frac{\left\langle \left| y_{2,S_2}(t) \right|^2 \right\rangle_t}{\left\langle \left| y_{2,S_1}(t) \right|^2 \right\rangle_t} \right).$$

Since $y_{i,S_j}(t) = u_{ij}(t) * s_j(t) \Rightarrow Y_{i,S_j}(f) = U_{ij}(f) S_j(f)$, it can be derived that the averaged SIR equals to:

$$Averaged\ SIR = \frac{1}{2} \cdot 10 \log_{10} \left(\frac{\left\langle \left| U_{11}(f) S_1(f) \right|^2 \right\rangle_f}{\left\langle \left| U_{12}(f) S_2(f) \right|^2 \right\rangle_f} \cdot \frac{\left\langle \left| U_{22}(f) S_2(f) \right|^2 \right\rangle_f}{\left\langle \left| U_{21}(f) S_1(f) \right|^2 \right\rangle_f} \right)$$

$$= \frac{1}{2} \cdot 10 \log_{10} \left(\frac{\left\langle \left| U_{11}(f) \right|^2 \right\rangle_f}{\left\langle \left| U_{12}(f) \right|^2 \right\rangle_f} \cdot \frac{\left\langle \left| U_{22}(f) \right|^2 \right\rangle_f}{\left\langle \left| U_{21}(f) \right|^2 \right\rangle_f} \right).$$

Therefore, the averaged SIR of "wistru" goes back to the normal range of the sequences.



(a) (b)

(c) (d)

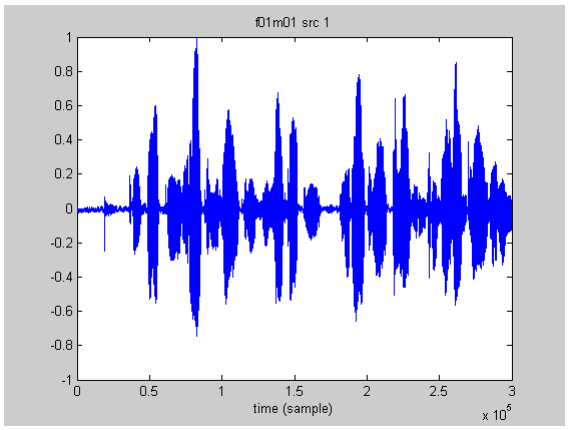
(e) (f)

Fig. 4.6 Sequence “f01m01” Waveforms in Time Domain

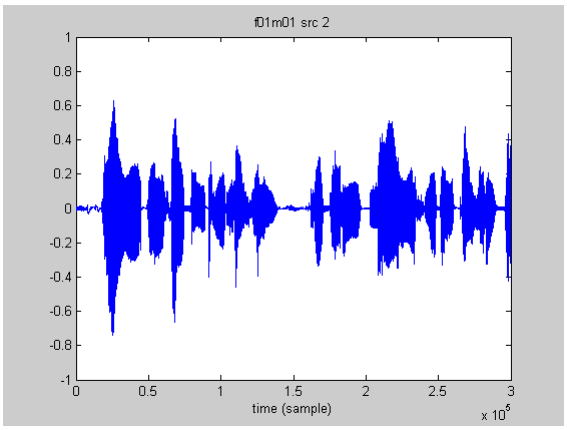
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

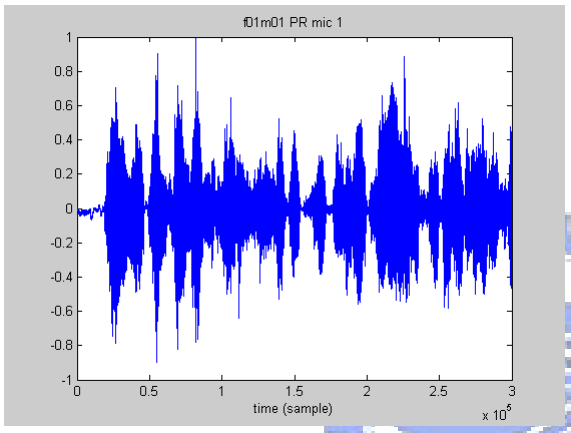
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



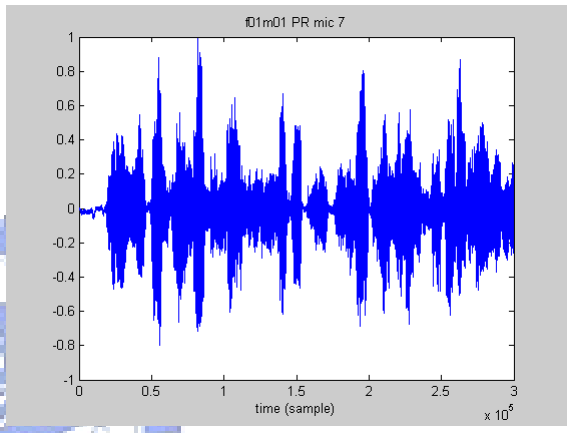
(a)



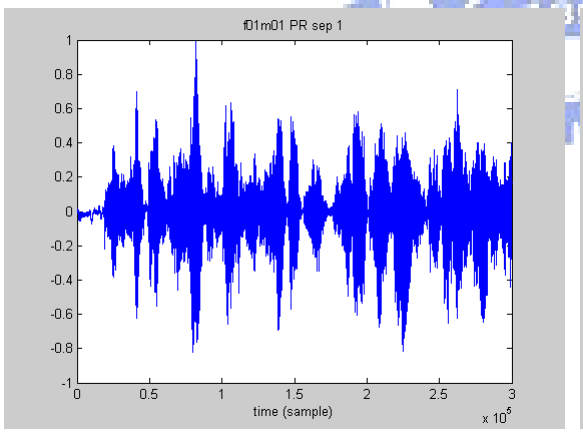
(b)



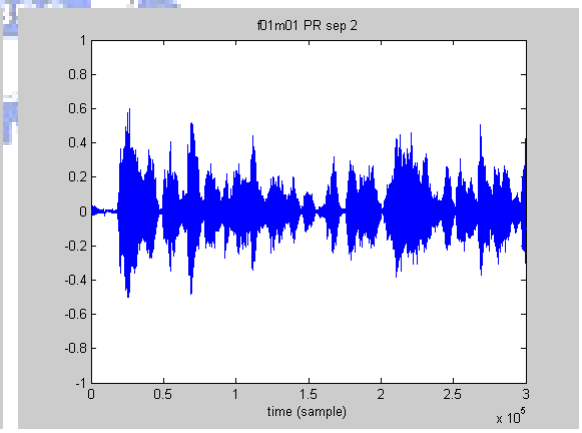
(c)



(d)



(e)



(f)

Fig. 4.7 Sequence “f01m01” Waveforms in Time Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR

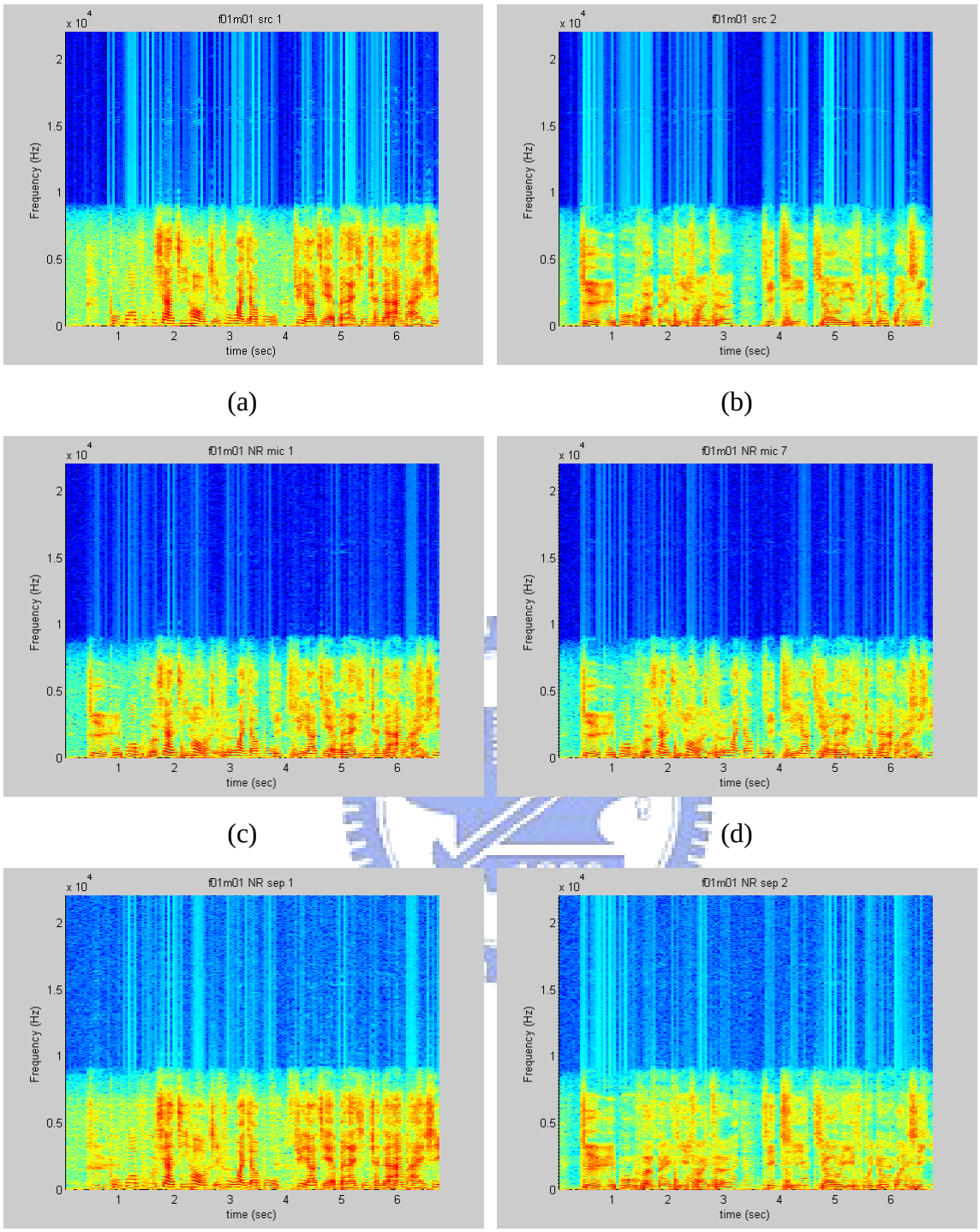
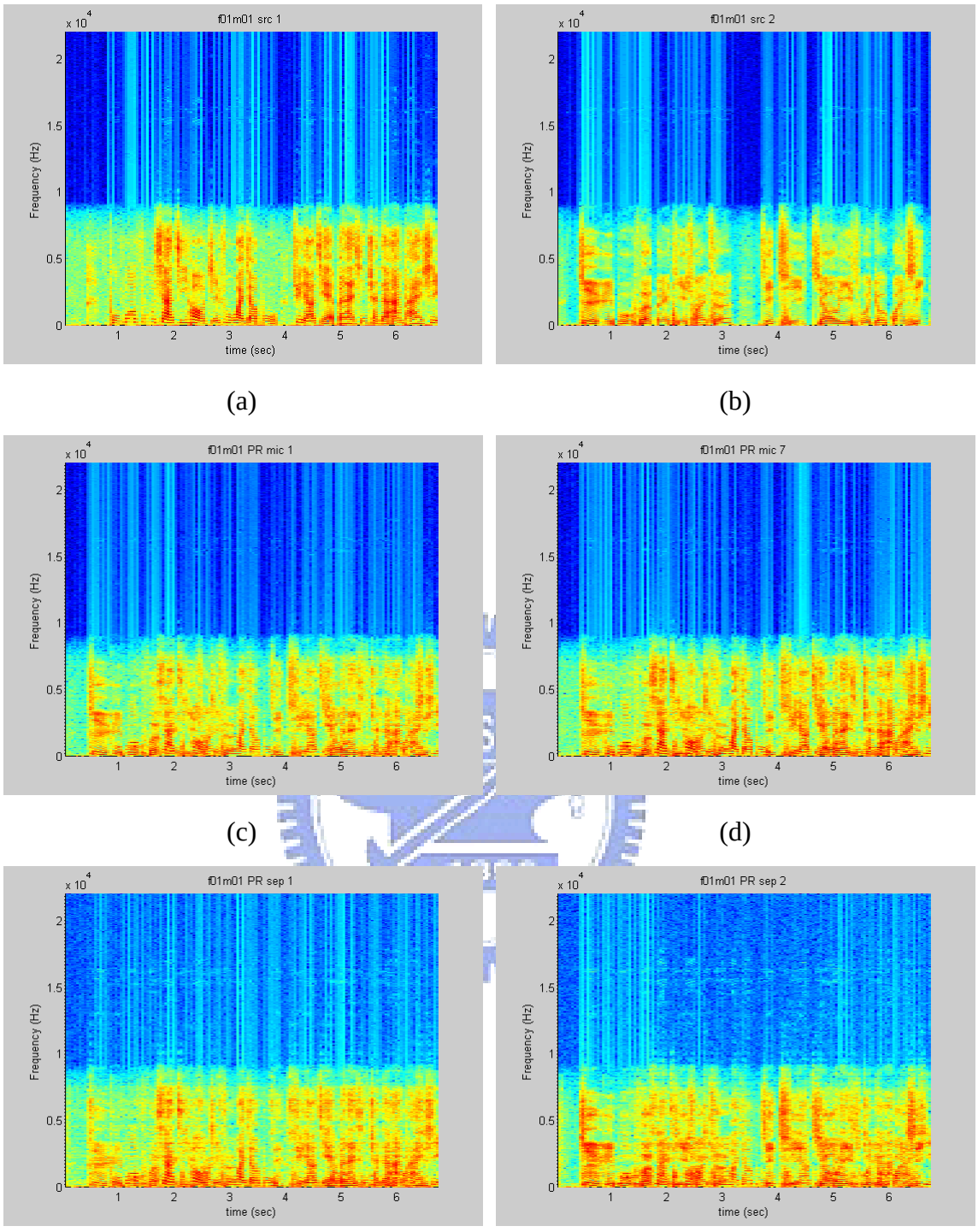


Fig. 4.8 Sequence “f01m01” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



(a) (b)

(c) (d)

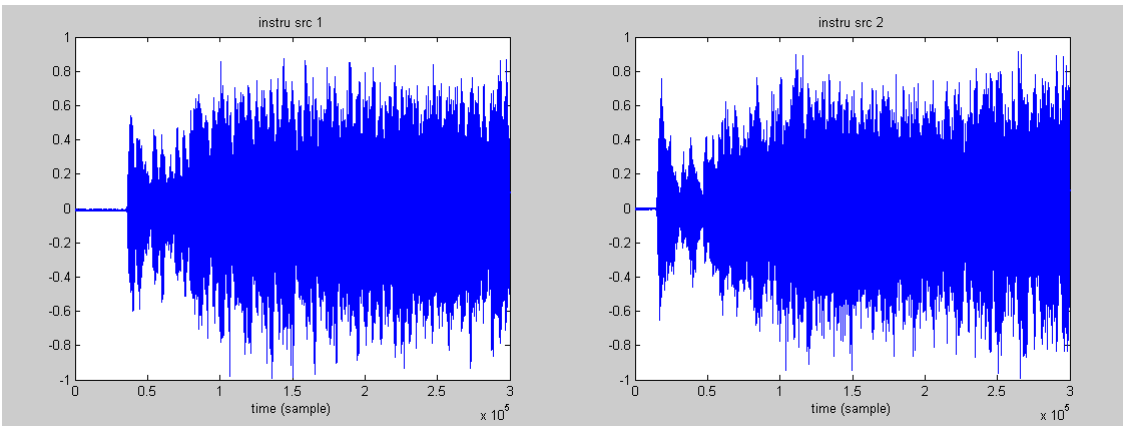
(e) (f)

Fig. 4.9 Sequence “f01m01” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

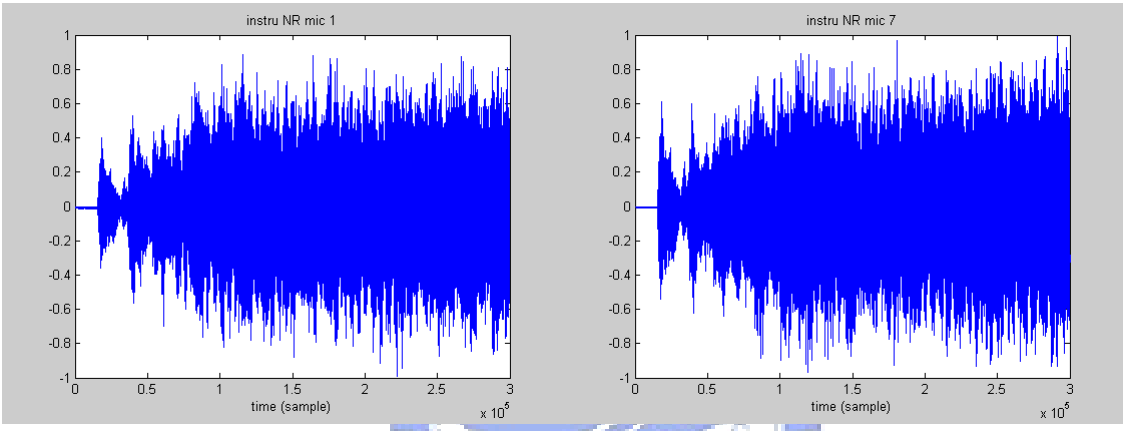
(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



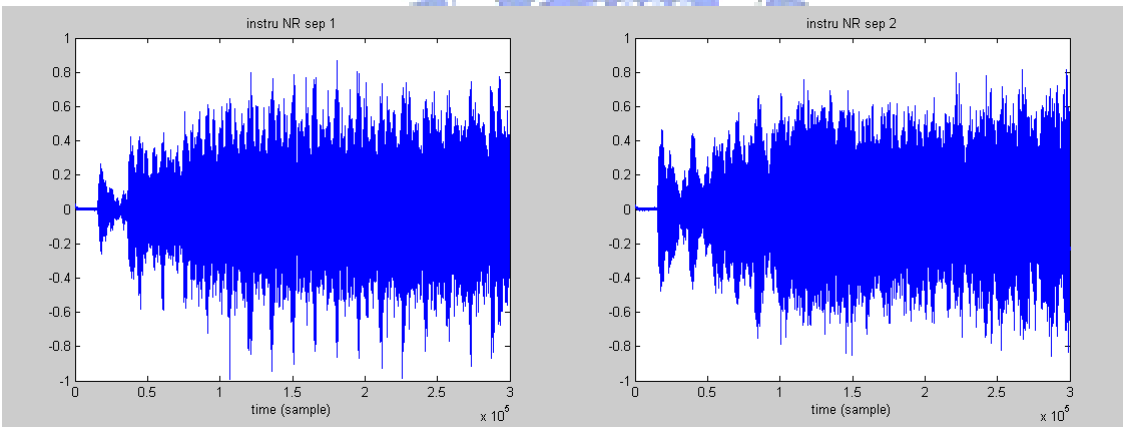
(a)

(b)



(c)

(d)



(e)

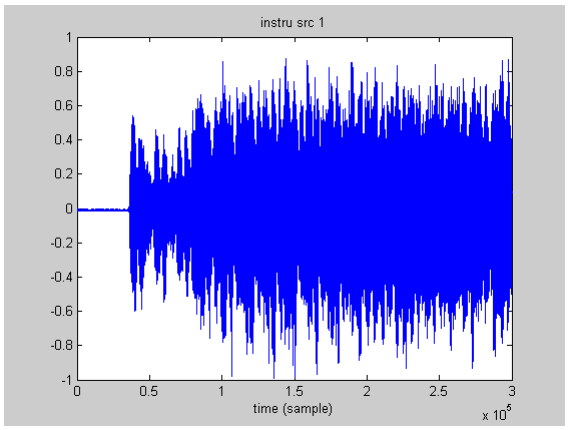
(f)

Fig. 4.10 Sequence “instru” Waveforms in Time Domain

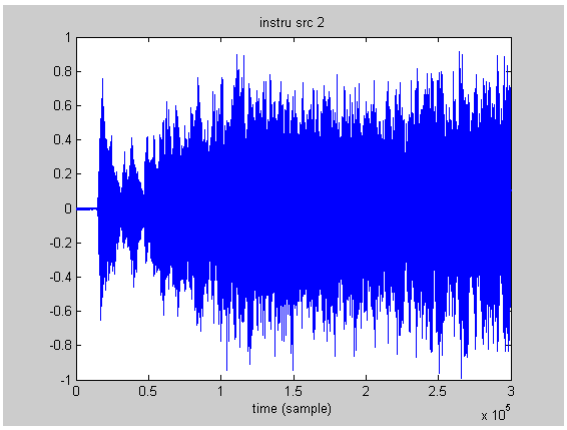
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

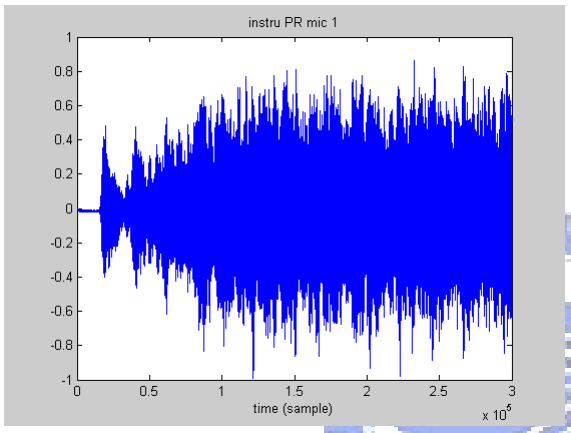
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



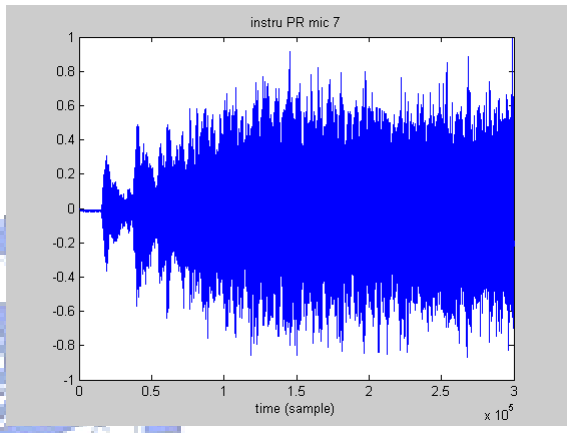
(a)



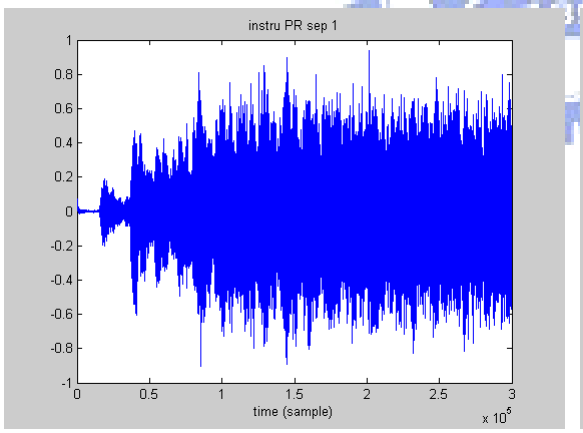
(b)



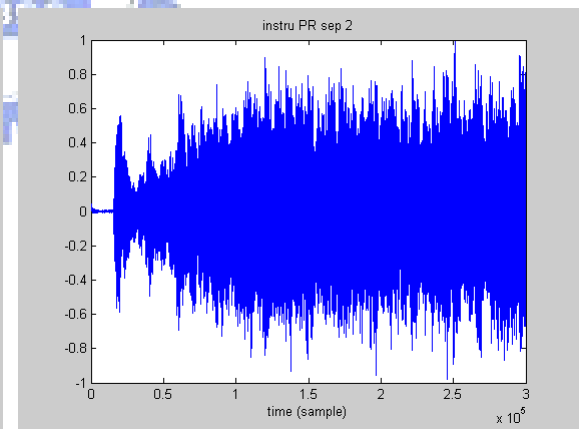
(c)



(d)



(e)



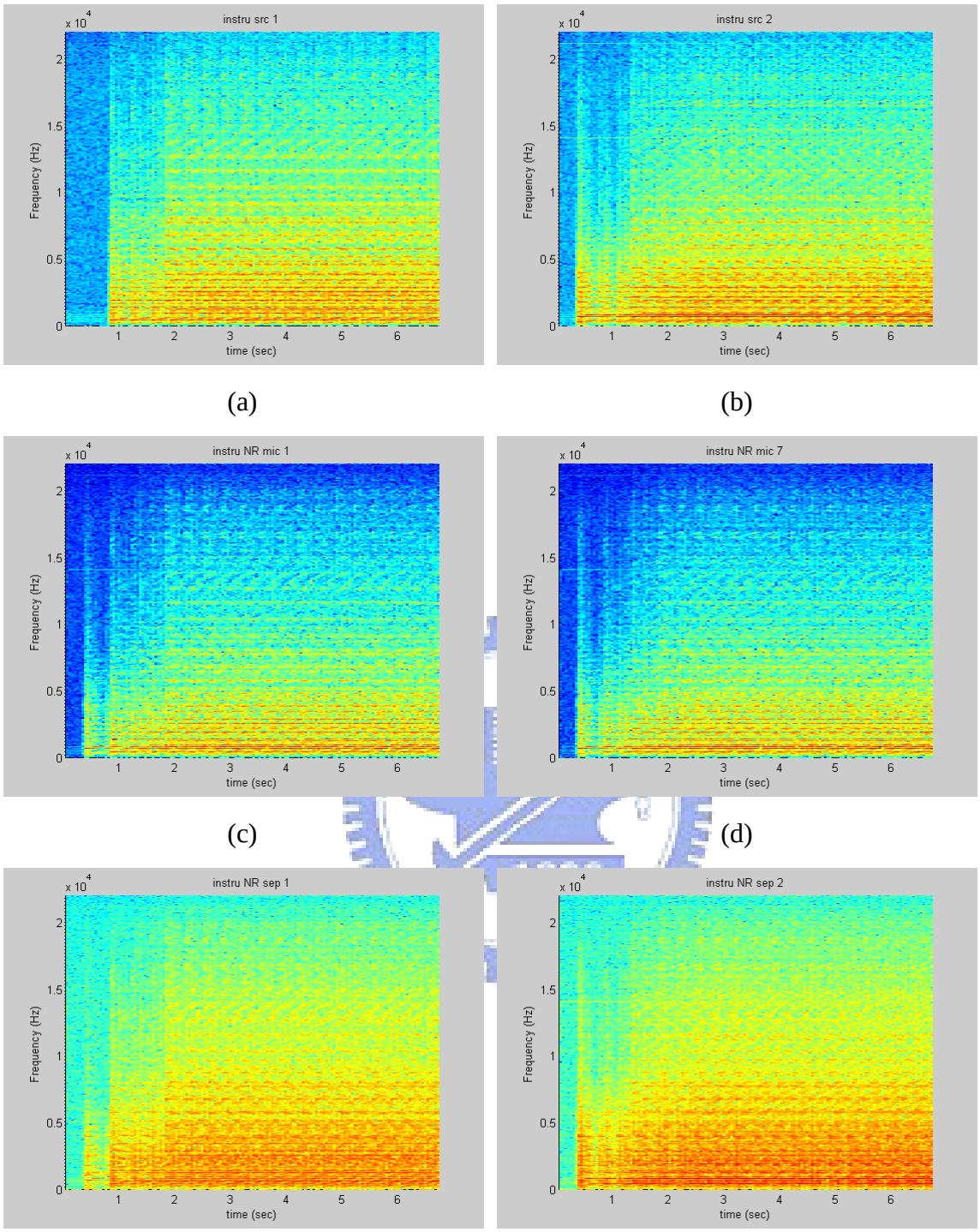
(f)

Fig. 4.11 Sequence “instru” Waveforms in Time Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



(a) (b)

(c) (d)

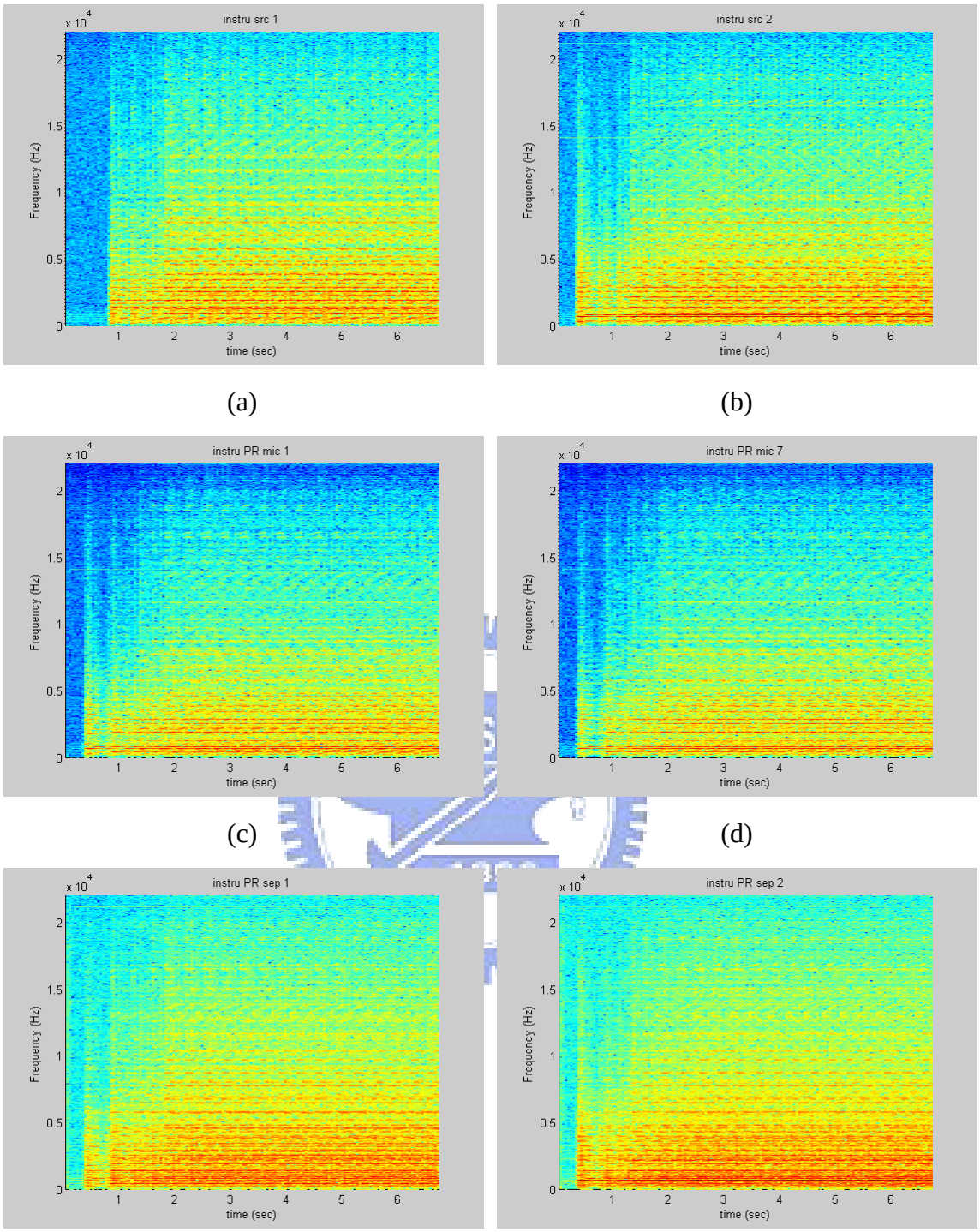
(e) (f)

Fig. 4.12 Sequence “instru” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



(a) (b)

(c) (d)

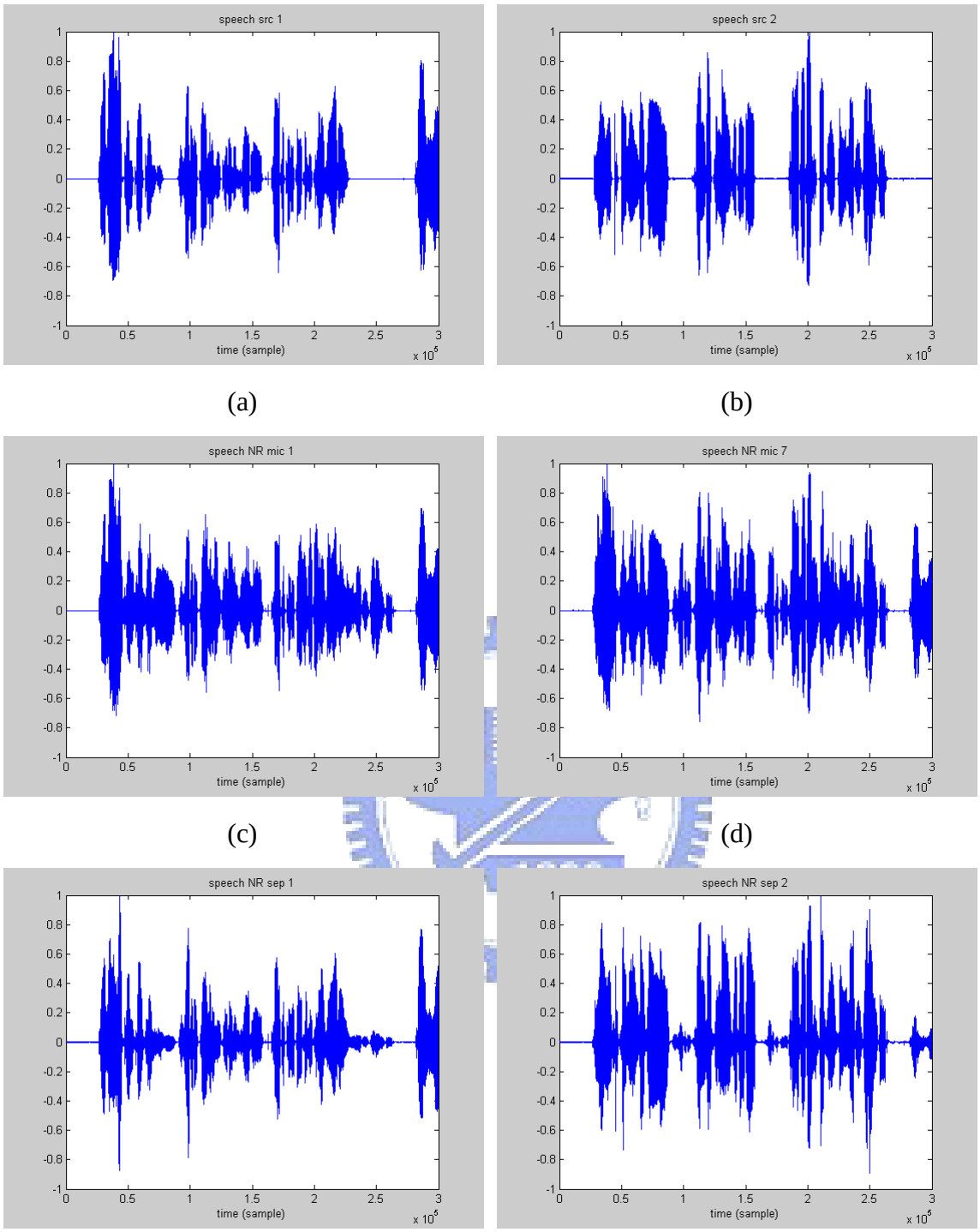
(e) (f)

Fig. 4.13 Sequence “instru” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



(a) (b)

(c) (d)

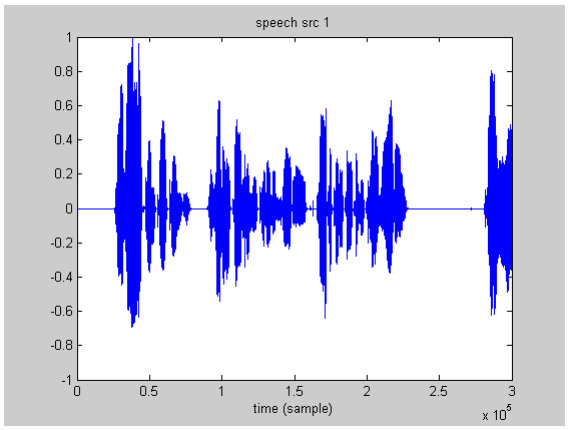
(e) (f)

Fig. 4.14 Sequence “speech” Waveforms in Time Domain

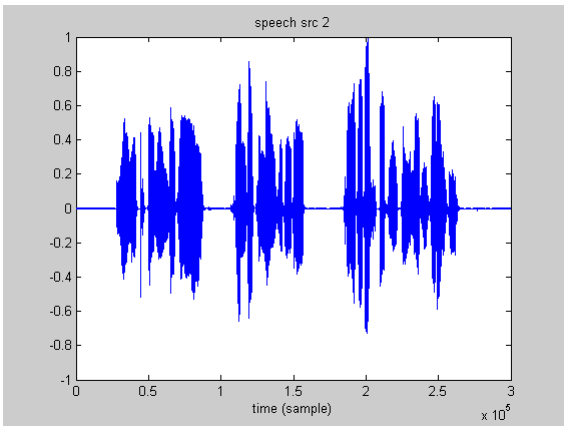
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

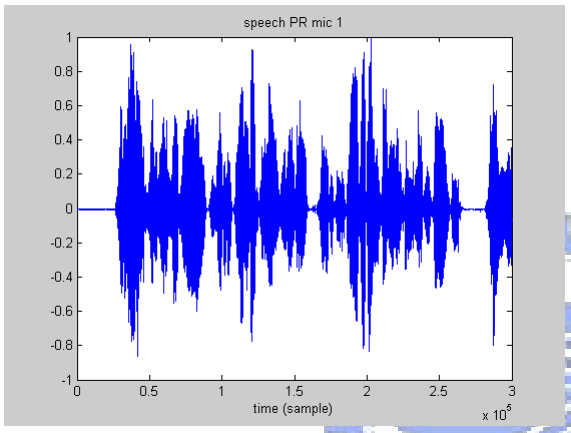
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



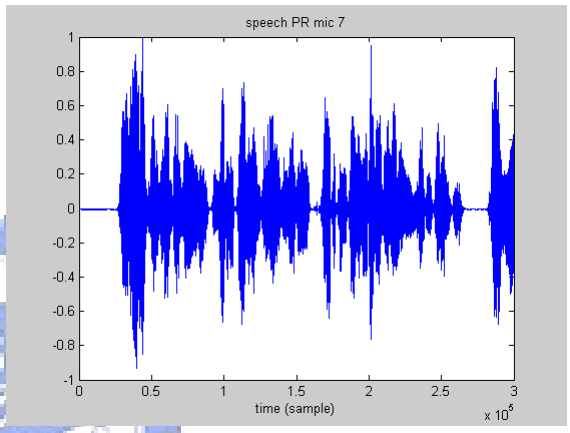
(a)



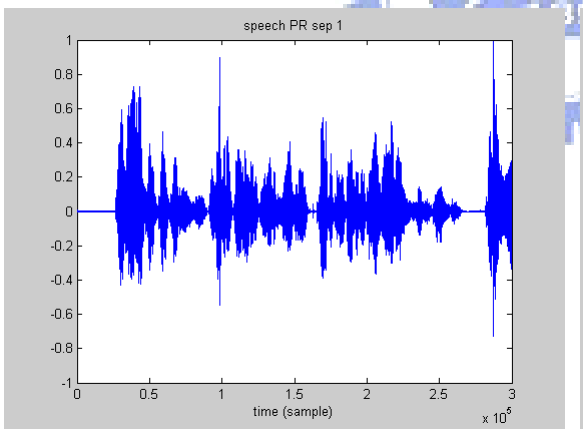
(b)



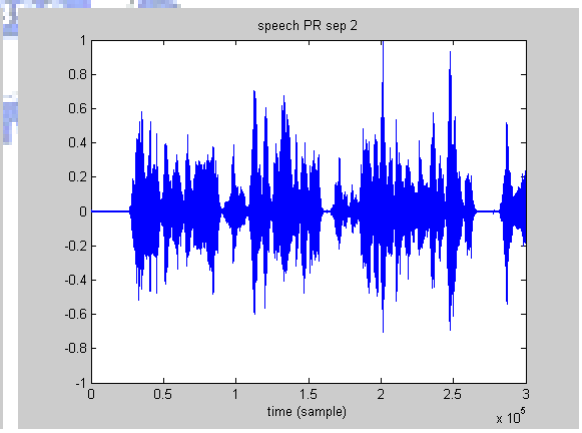
(c)



(d)



(e)



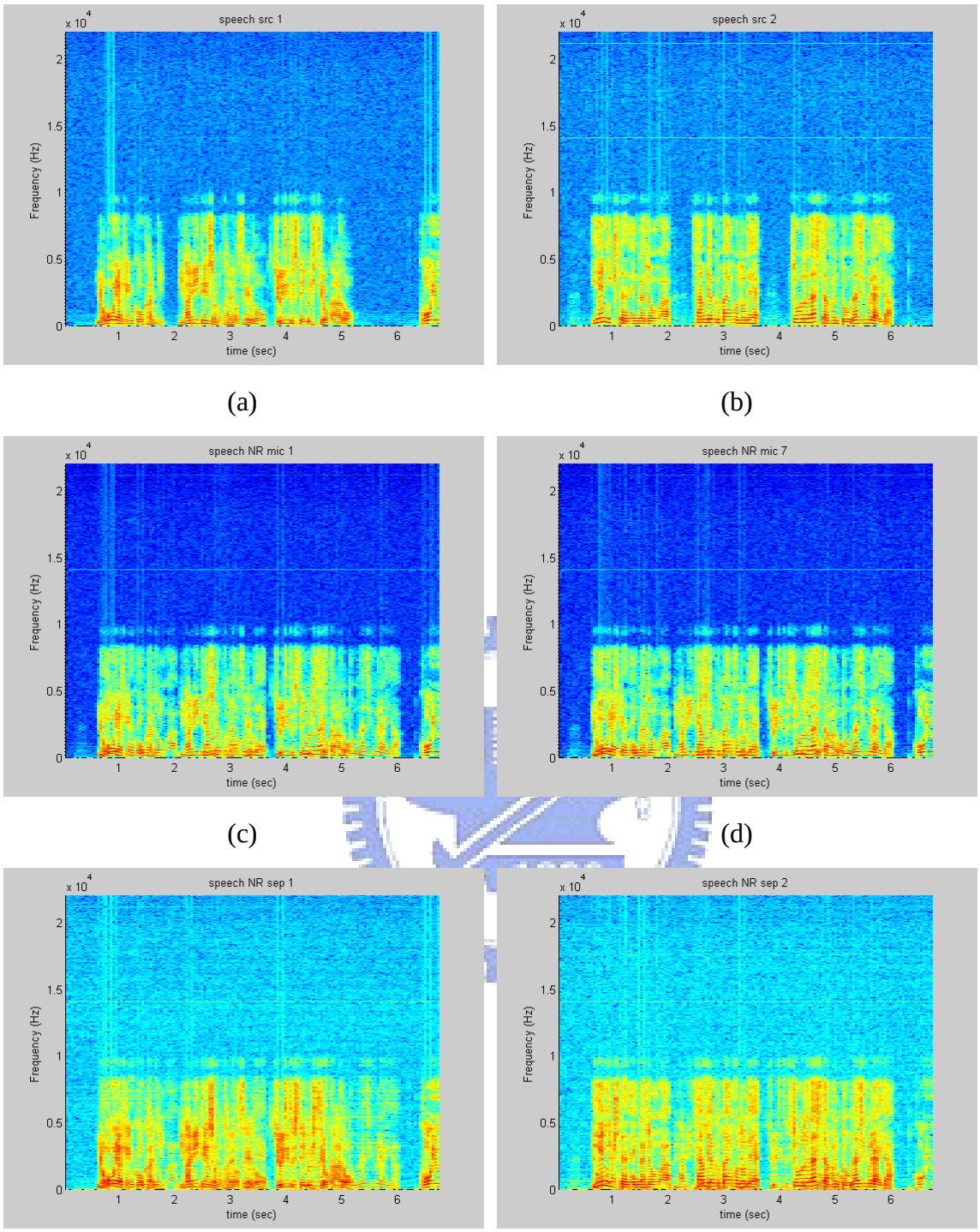
(f)

Fig. 4.15 Sequence “speech” Waveforms in Time Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



(a) (b)

(c) (d)

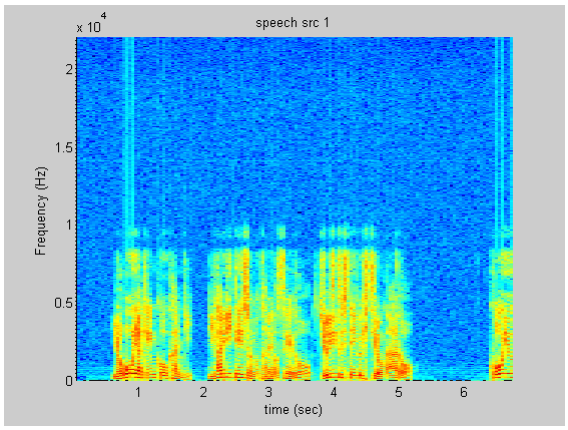
(e) (f)

Fig. 4.16 Sequence “speech” Spectrograms in Time-Frequency Domain

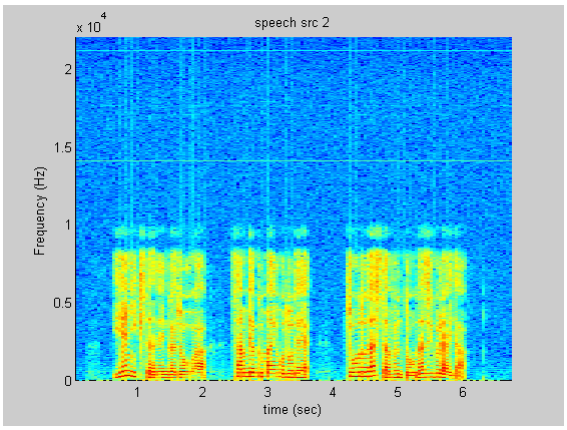
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

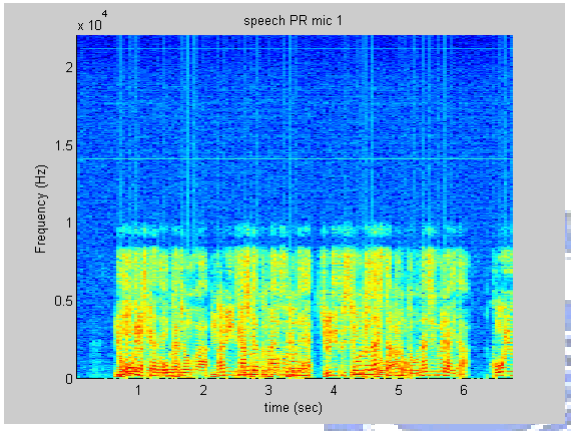
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



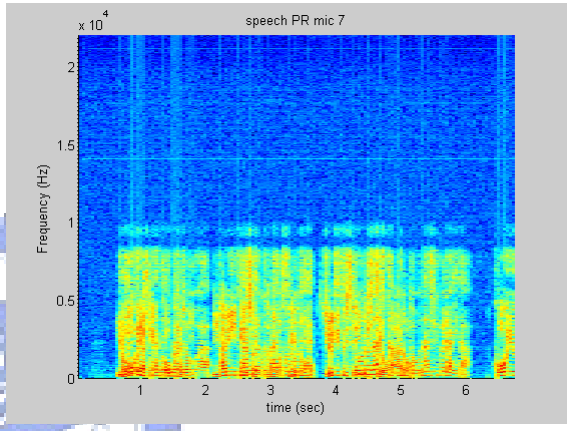
(a)



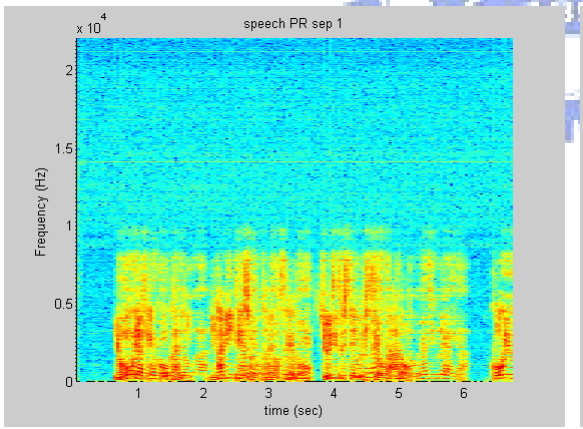
(b)



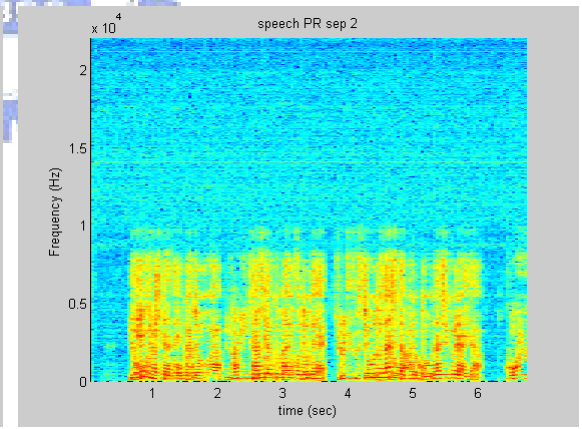
(c)



(d)



(e)



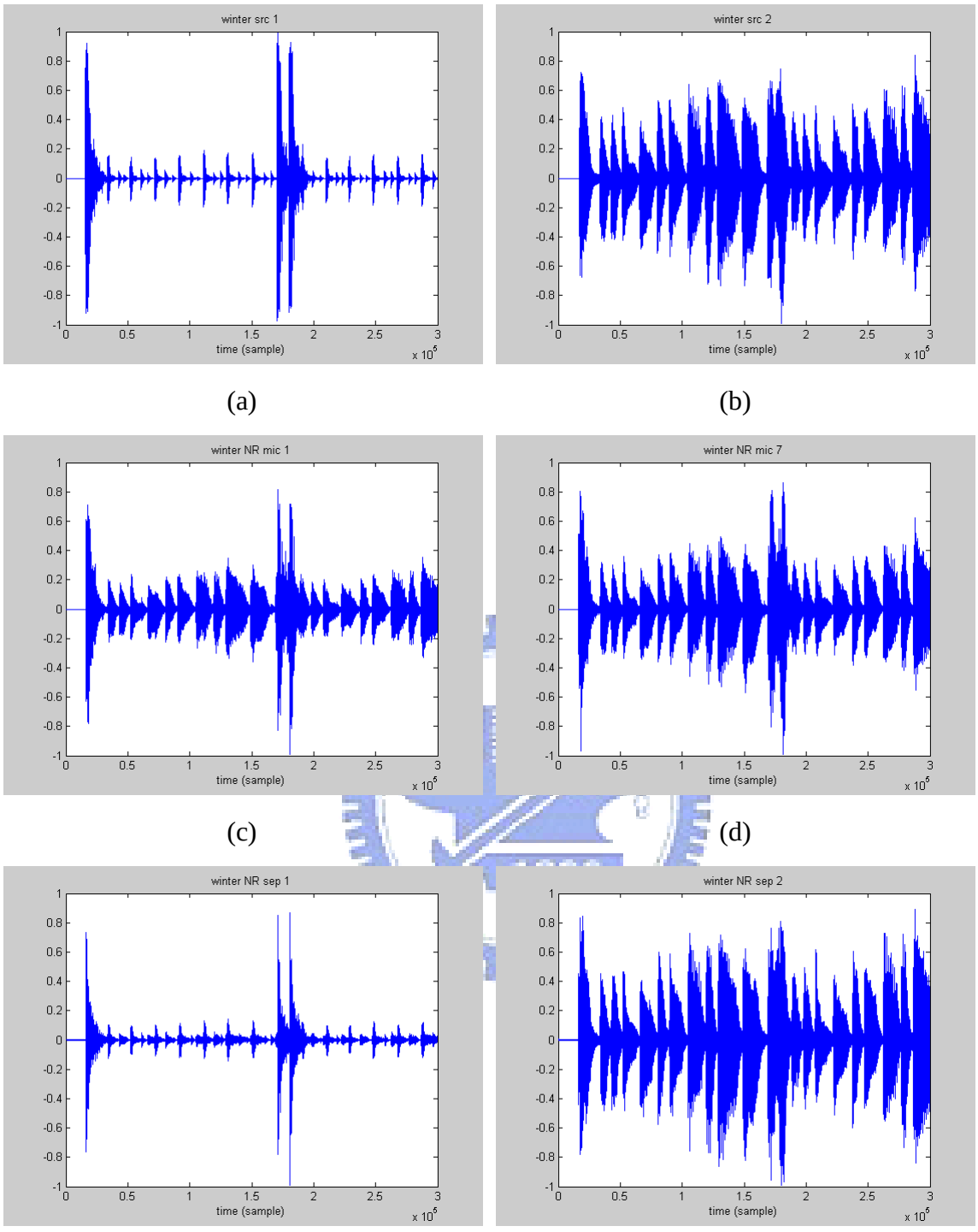
(f)

Fig. 4.17 Sequence “speech” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



(a) (b)

(c) (d)

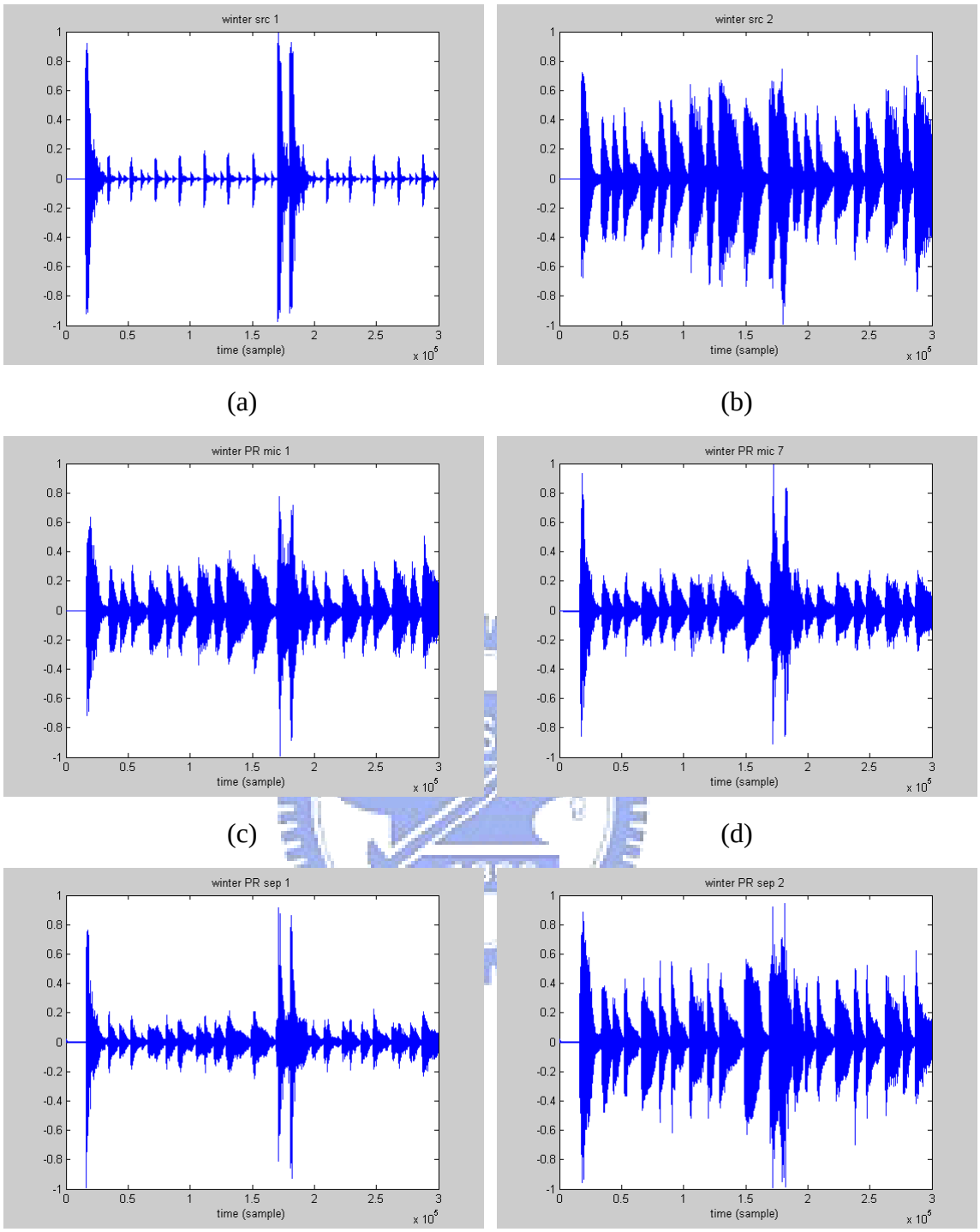
(e) (f)

Fig. 4.18 Sequence “winter” Waveforms in Time Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



(a) (b)

(c) (d)

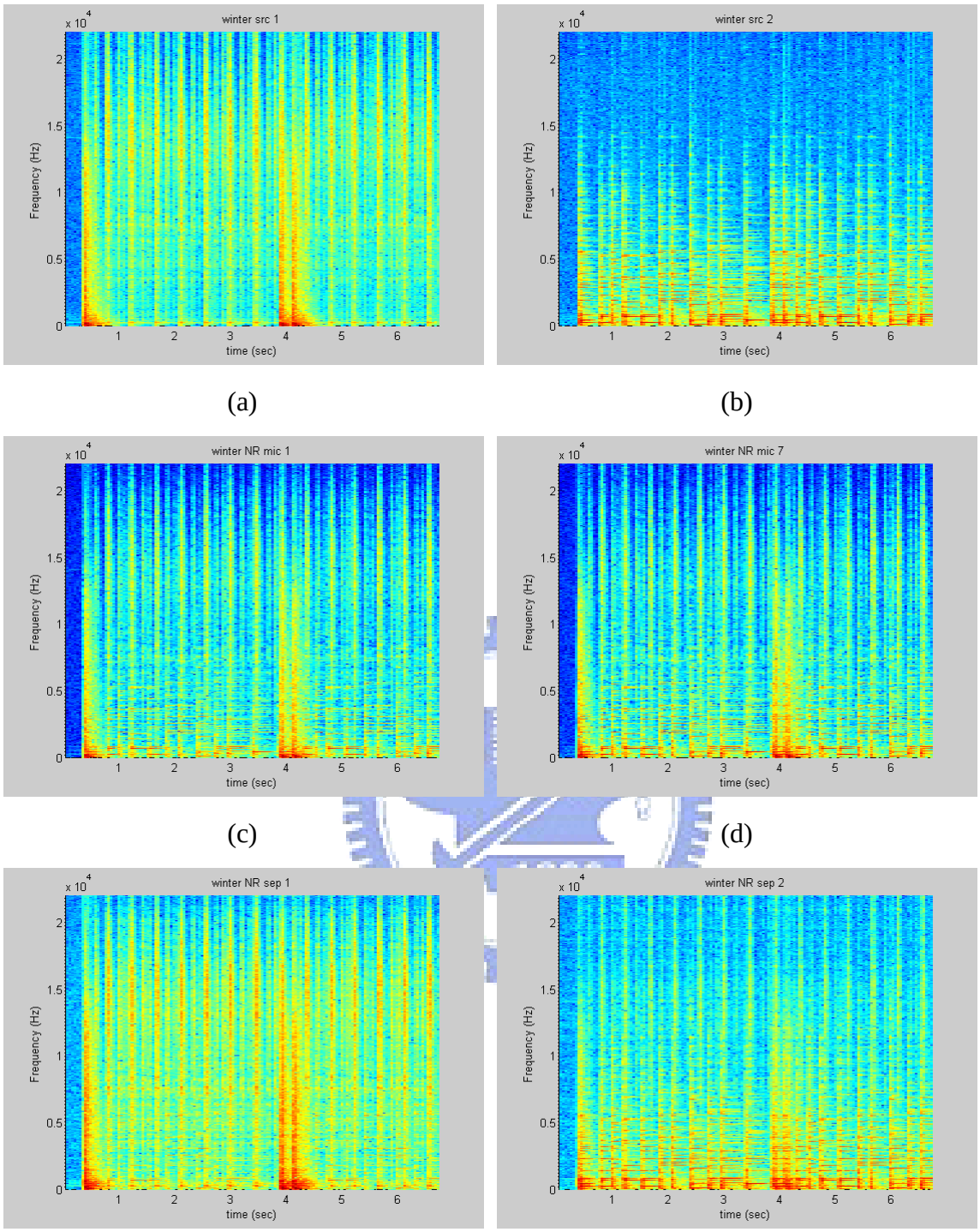
(e) (f)

Fig. 4.19 Sequence “winter” Waveforms in Time Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



(a) (b)

(c) (d)

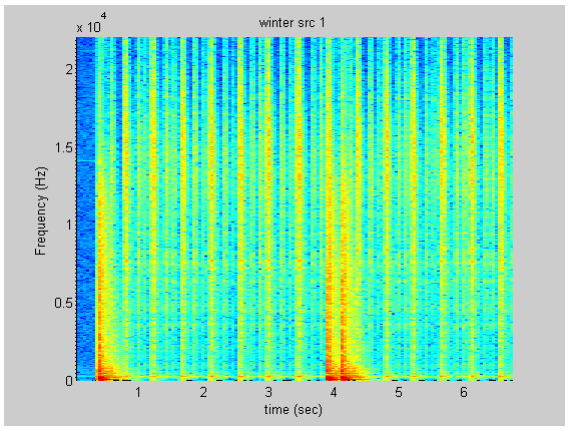
(e) (f)

Fig. 4.20 Sequence “winter” Spectrograms in Time-Frequency Domain

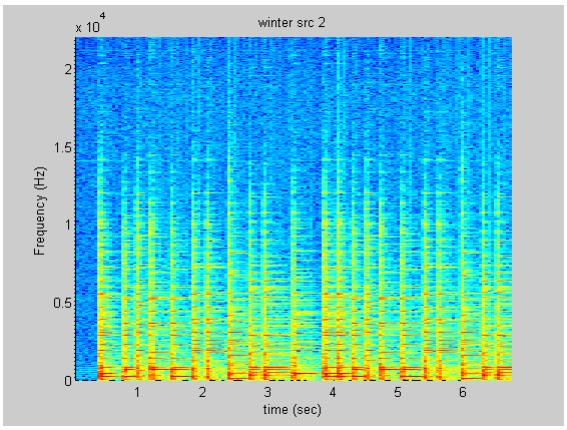
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

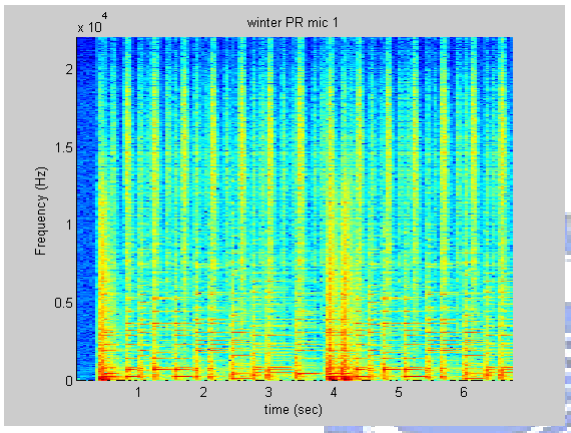
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



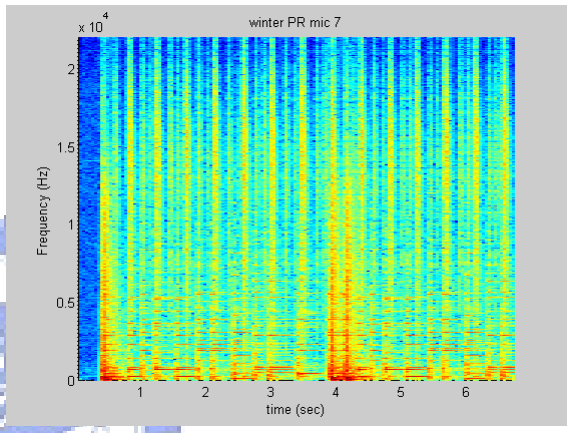
(a)



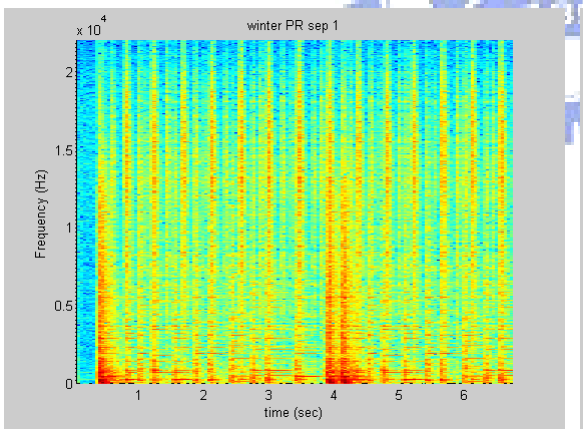
(b)



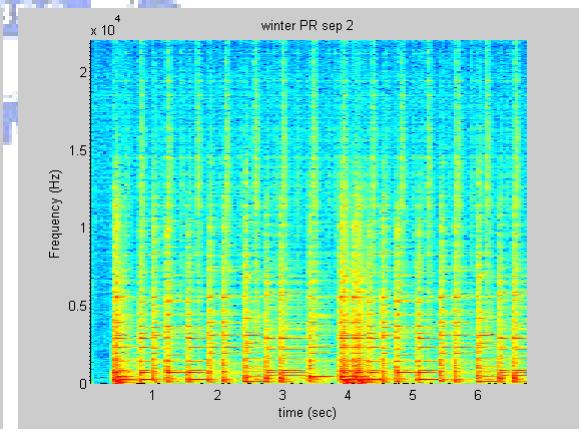
(c)



(d)



(e)



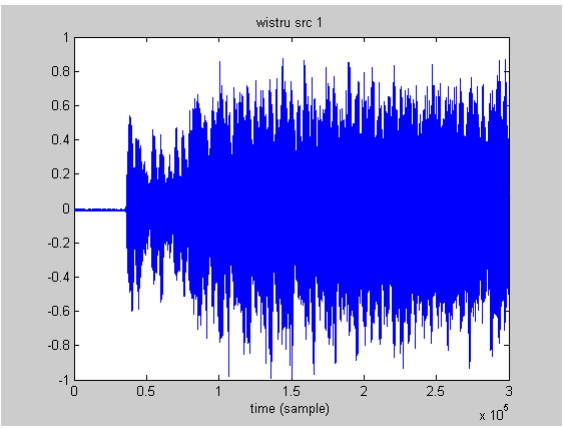
(f)

Fig. 4.21 Sequence “winter” Spectrograms in Time-Frequency Domain

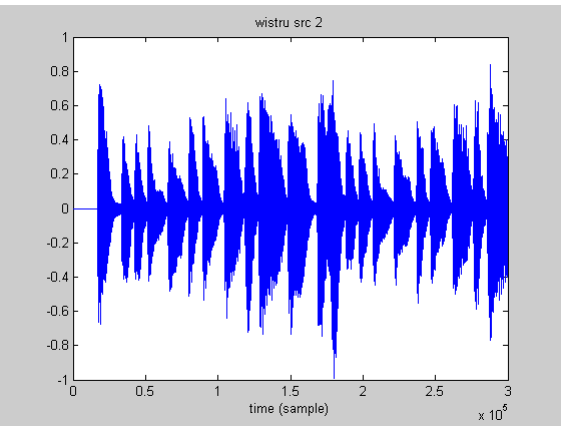
(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

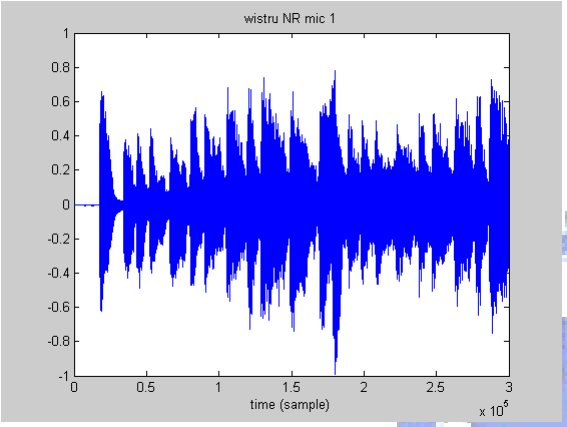
(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR



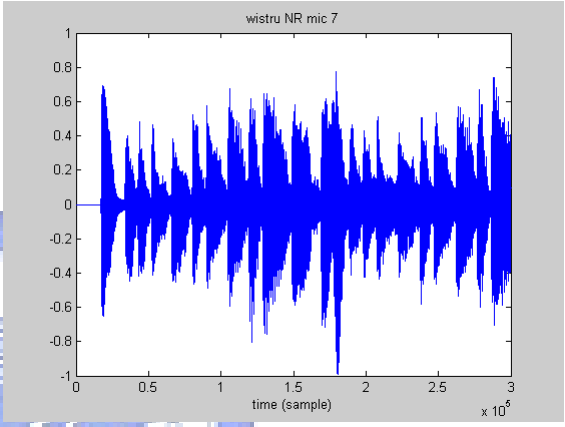
(a)



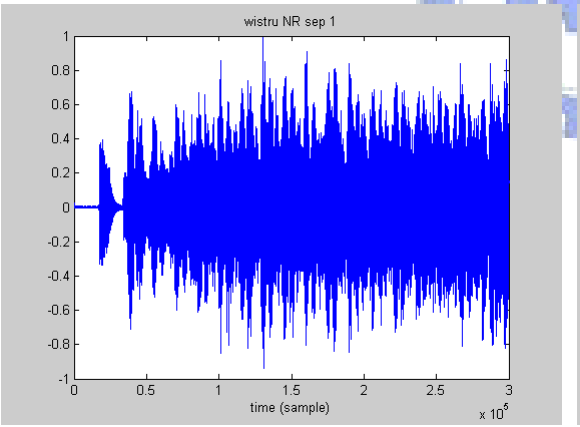
(b)



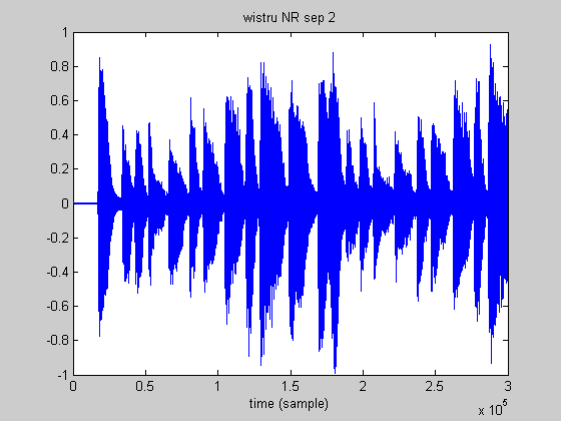
(c)



(d)



(e)



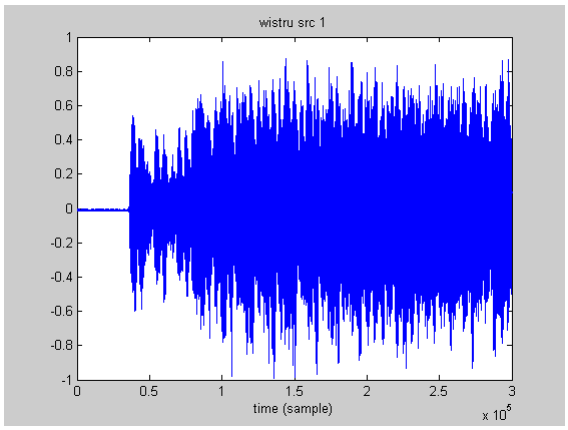
(f)

Fig. 4.22 Sequence “wistru” Waveforms in Time Domain

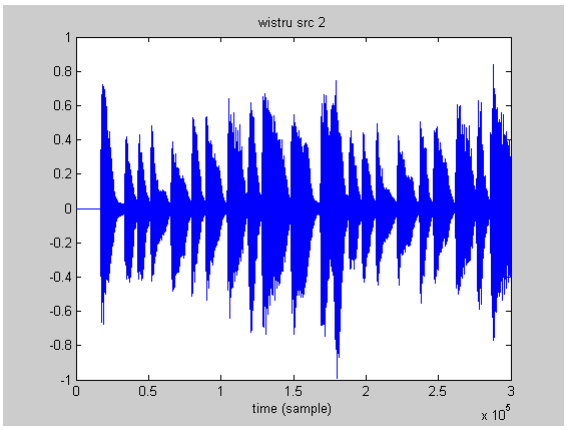
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

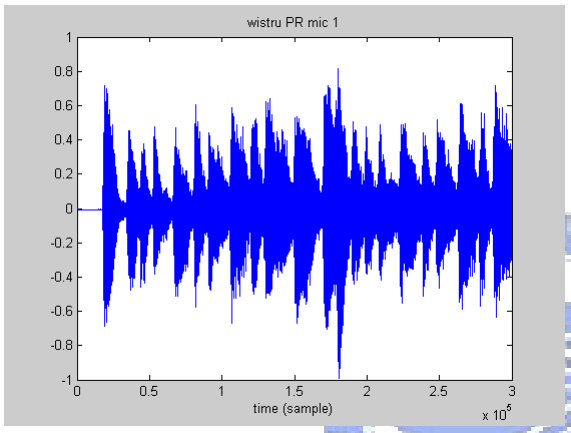
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



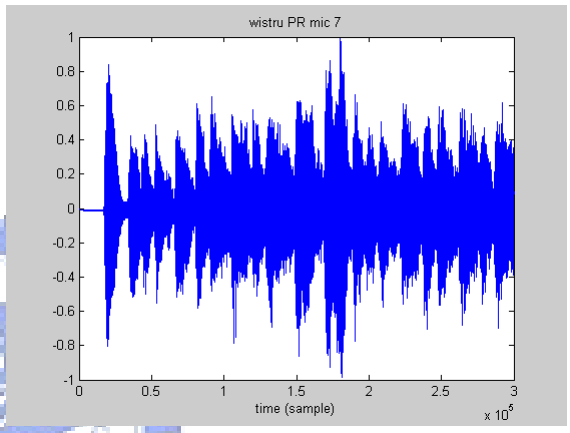
(a)



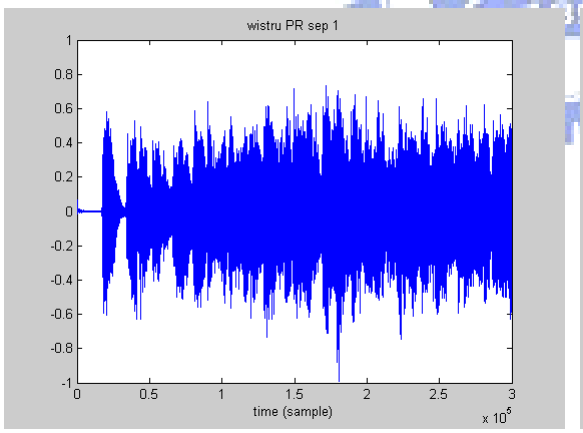
(b)



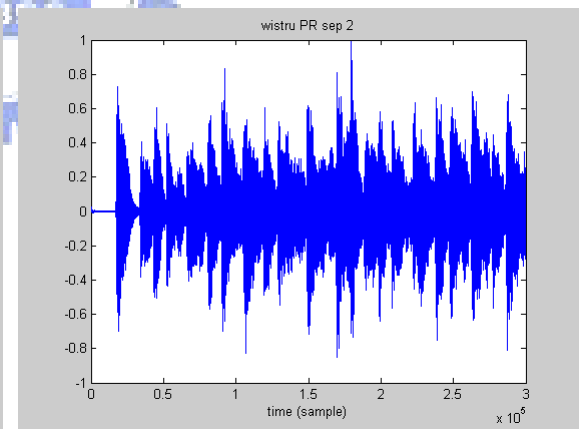
(c)



(d)



(e)



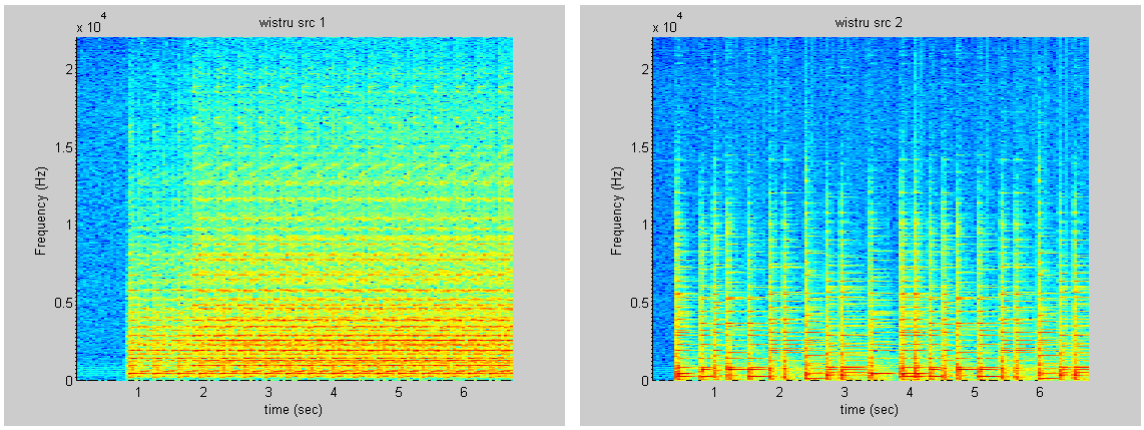
(f)

Fig. 4.23 Sequence “wistru” Waveforms in Time Domain

(a) Source 1 (b) Source 2

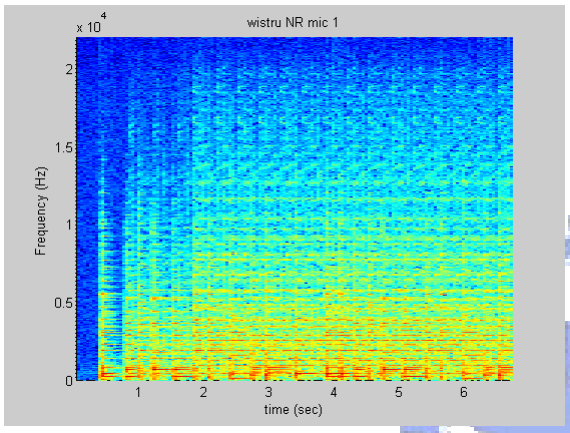
(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR

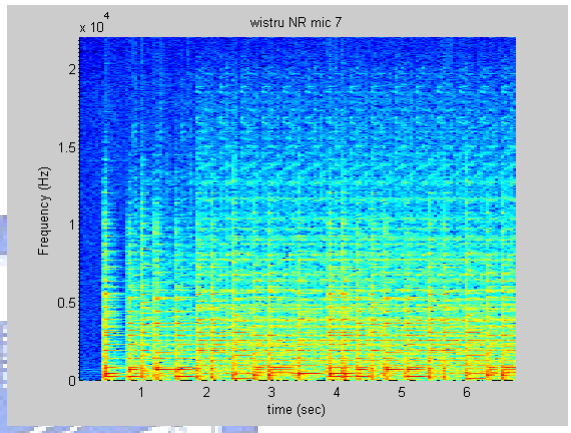


(a)

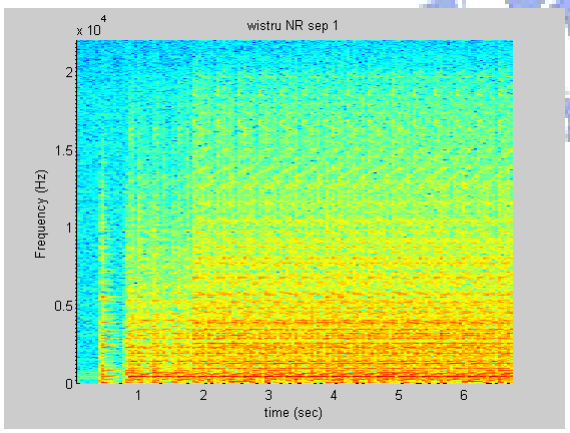
(b)



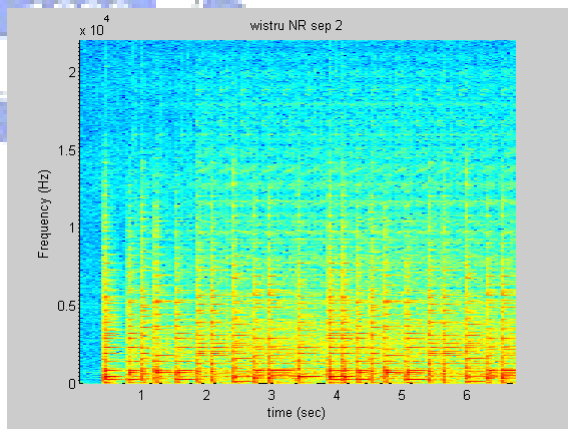
(c)



(d)



(e)



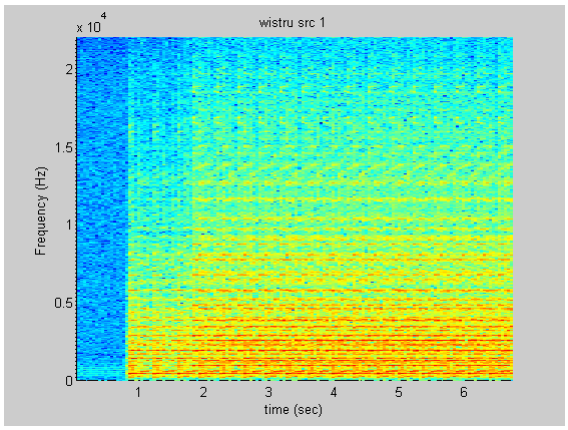
(f)

Fig. 4.24 Sequence “wistru” Spectrograms in Time-Frequency Domain

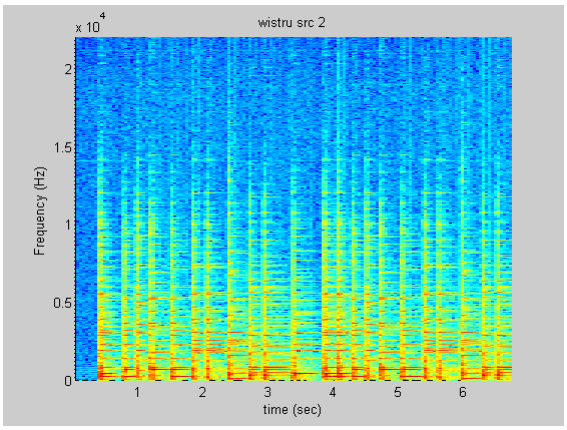
(a) Source 1 (b) Source 2

(c) Microphone 1 with NR (d) Microphone 7 with NR

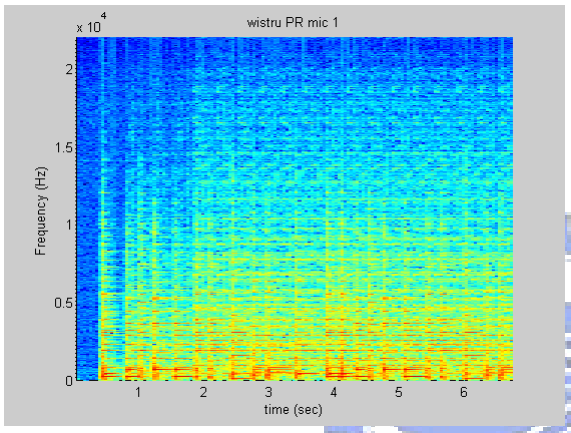
(e) Separated Signal 1 with NR (f) Separated Signal 2 with NR



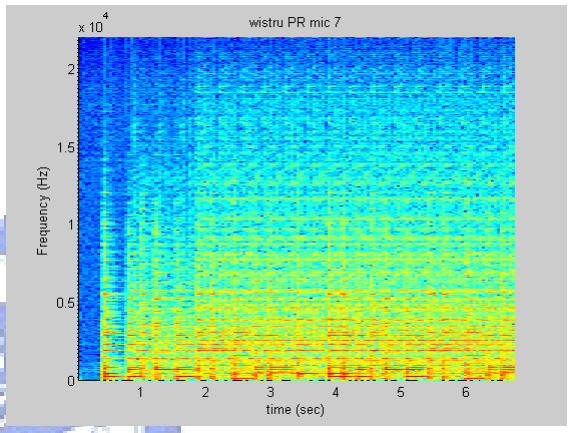
(a)



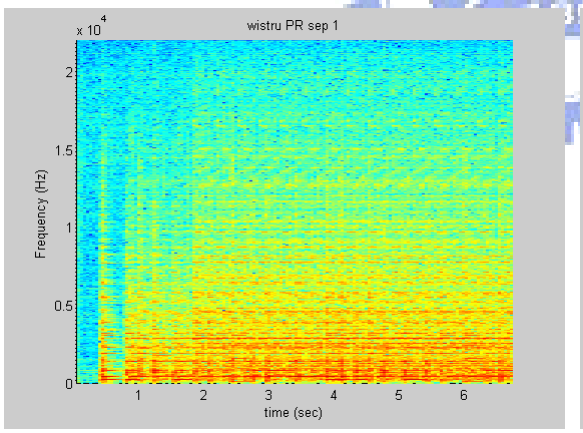
(b)



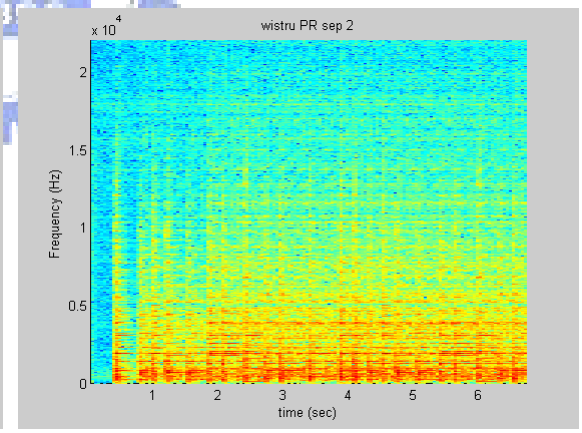
(c)



(d)



(e)



(f)

Fig. 4.25 Sequence “wistru” Spectrograms in Time-Frequency Domain

(a) Source 1 (b) Source 2

(c) Microphone 1 with PR (d) Microphone 7 with PR

(e) Separated Signal 1 with PR (f) Separated Signal 2 with PR

4.2 Virtual Acoustic Environment

4.2.1 Introduction to NASA Sound Lab (SLAB) Software

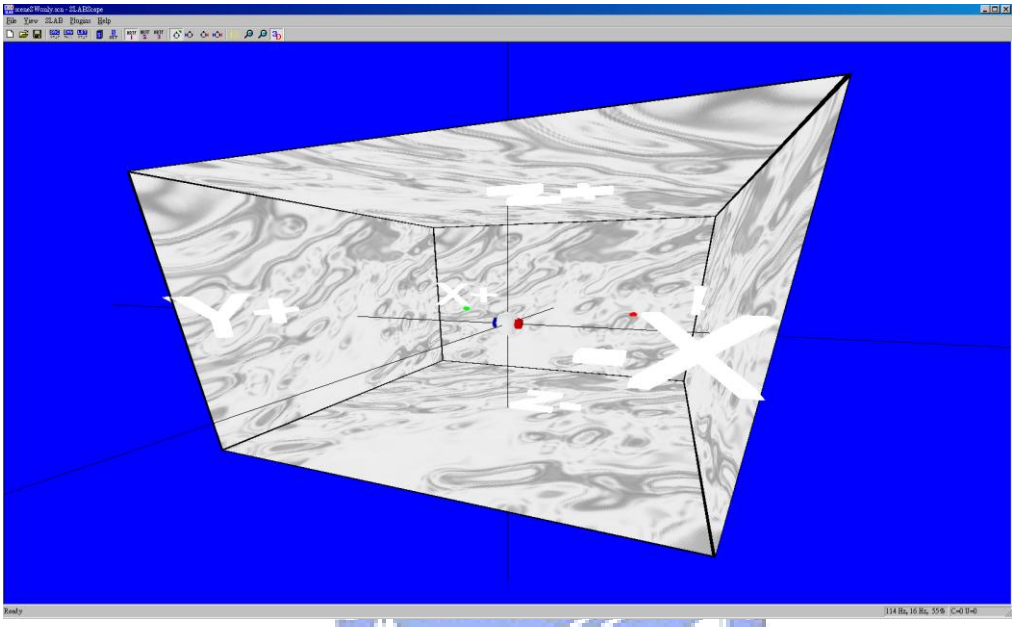


Fig. 4.26 Snapshot of the 3D Virtual Acoustic Room in SLAB

SLAB is a software-based real-time virtual acoustic environment rendering system developed by the NASA Ames Research Center. This software provides an offline acoustic environment for spatial hearing and psychoacoustic studies. The acoustic scenario parameters considered in the SLAB include three main categories: the source, the environment, and the listener. The source parameters include the source locations, the source waveforms, the radiation pattern and radius of each source, etc. The environment parameters include the sound speed, the air absorption, the surface locations, the room dimension and the surface reflections, etc. The listener parameters include the listener location, the HRTF model and the interaural time difference (ITD), etc. There are some other specifications about the SLAB software which are presented in the following section.

4.2.2 SLAB Acoustic Scenario

SLAB Specifications [25]:

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 4.3 Scenario Specifications [25]

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)**

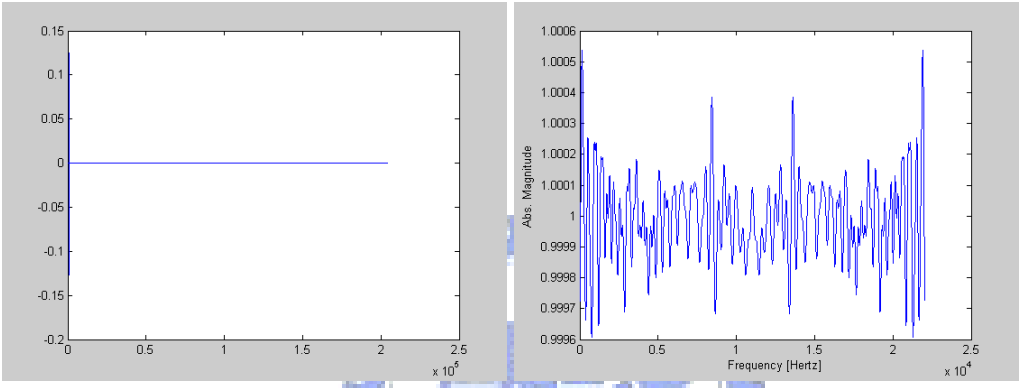
Table 4.4 System Dynamics Specifications [25]

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

Table 4.5 Numerical Precision Specifications [25]

4.3 Wall Material ATF Characteristics

There are seven kinds of wall materials provided by the SLAB software. The ATF spectrum is estimated by the TSP signal changes along with different wall materials. The tail of the time domain TSP signal with N = 2048 and M = 64 appends some padding zeros in order to observe the effect of reflections from the six-sided wall materials. As in Fig. 4.27(b) shown, the padding zeros introduce some tolerable amplitude distortions.

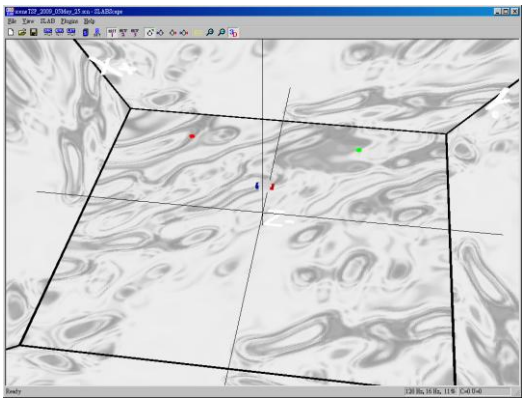


(a) (b)

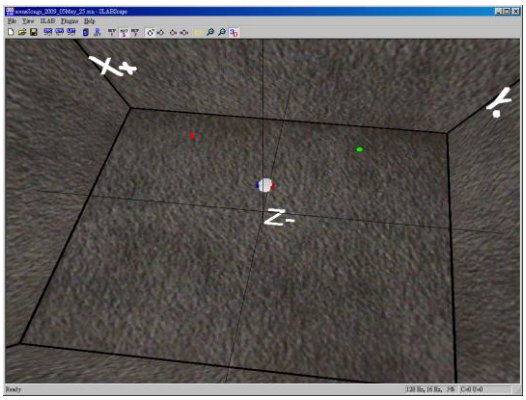
Fig. 4.27 TSP Signal with Padding Zeros

(a) Time Domain (b) Frequency Domain Amplitude

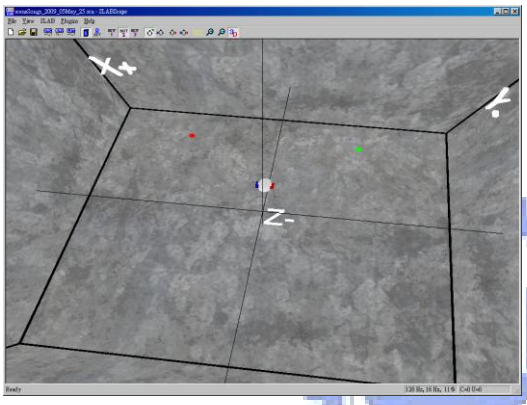
The frequency spectrum characteristics for the seven materials and the no reflection scene are shown as Fig. 4.29 from (a) through (h). All the data of Fig. 4.29 are the ATFs measured from the source 1 (red point) to the virtual listening point at (1.25, 0, 1.5) in the median room of the dimension 10 x 10 x 10 in meters. The left column of Fig. 4.29 shows the frequency domain log10 amplitudes and the right column shows the frequency domain unwrapped phase. The name list of the eight wall properties are no reflection (NR), perfect reflector (PR), heavy carpet (HC), concrete (Co), heavy glass (HG), gypsum board (GB), wood with airspace (WA) and plaster on metal (PM), which are shown in Fig. 4.28.



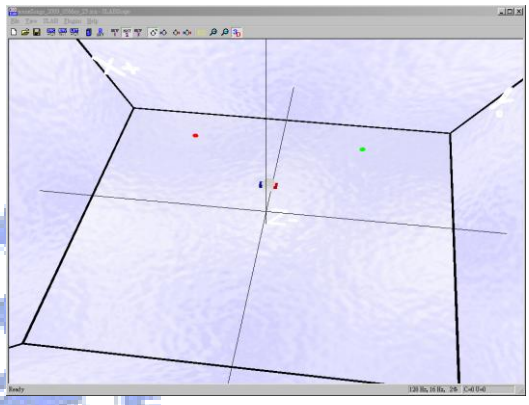
(a)



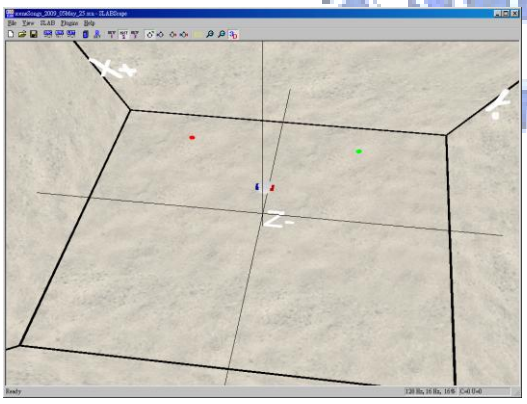
(b)



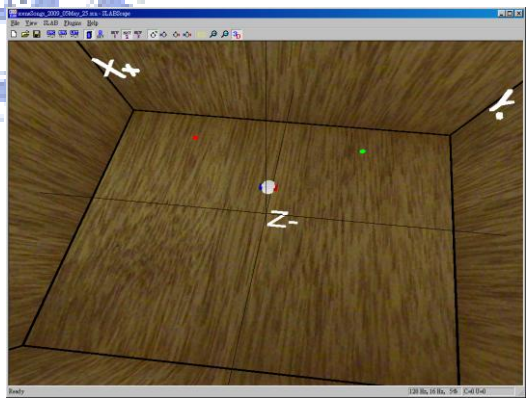
(c)



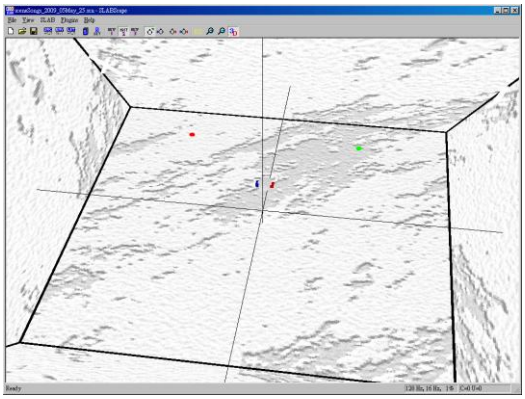
(d)



(e)



(f)

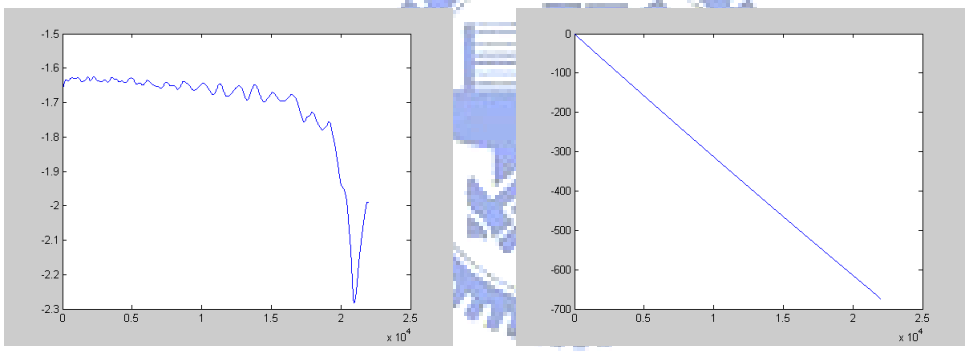


(g)

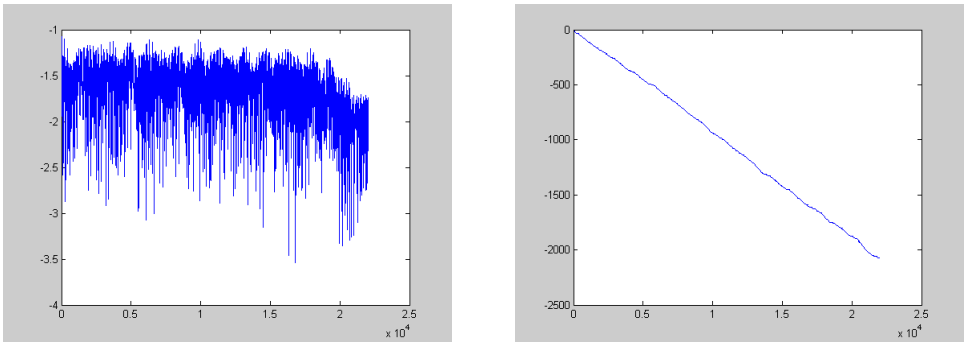
Fig. 4.28 Wall Materials

(a) Perfect Reflector (b) Heavy Carpet (c) Concrete (d) Heavy Glass

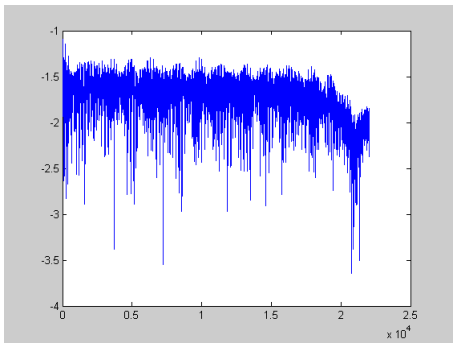
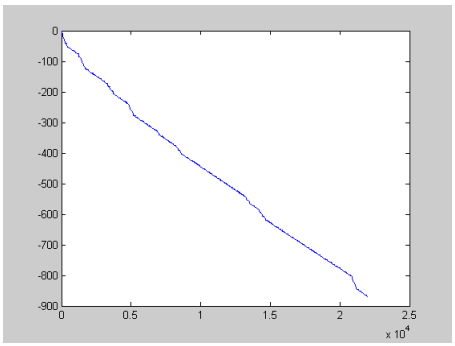
(e) Gypsum Board (f) Wood with Airspace (g) Plaster on Metal



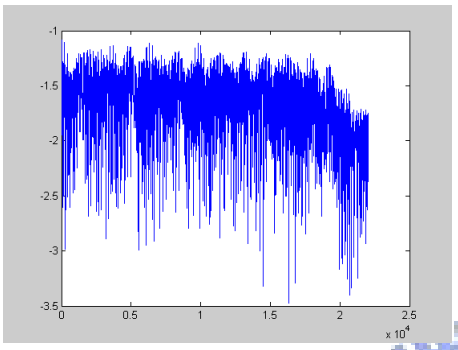
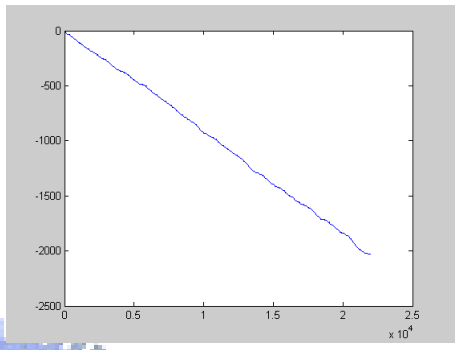
(a)



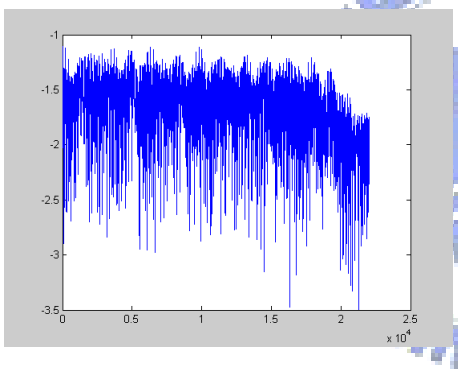
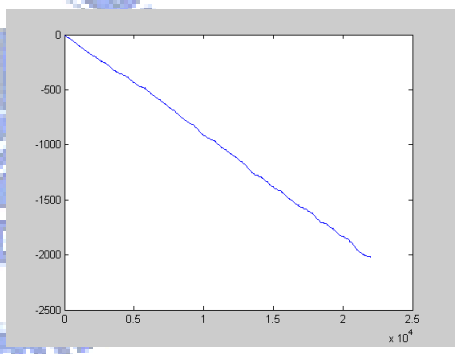
(b)

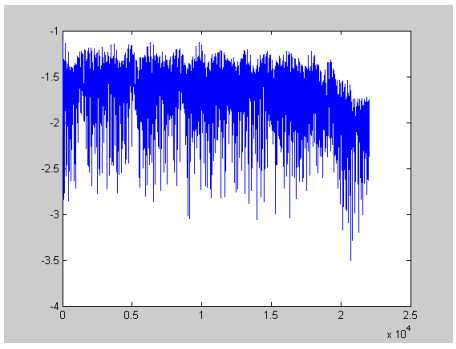
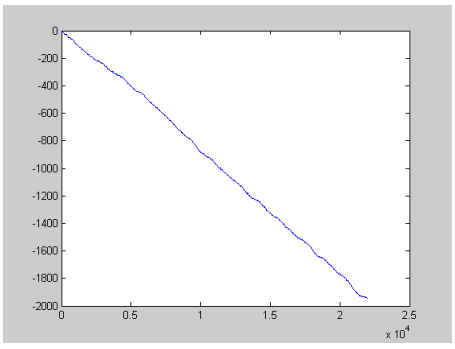
(c)

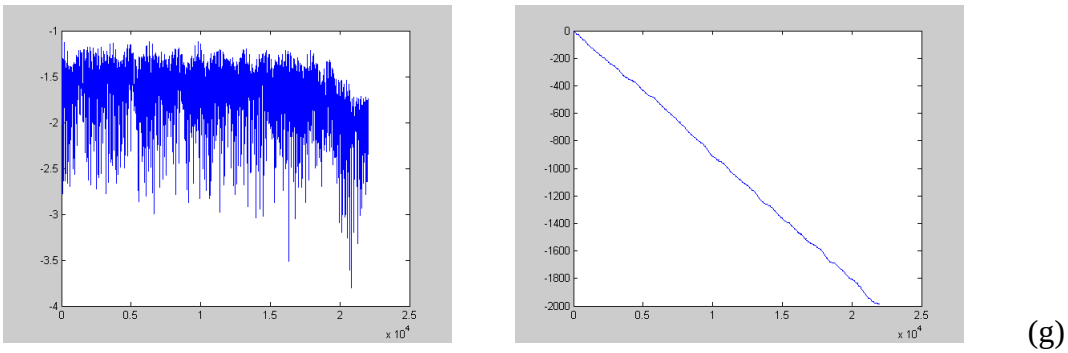
(d)

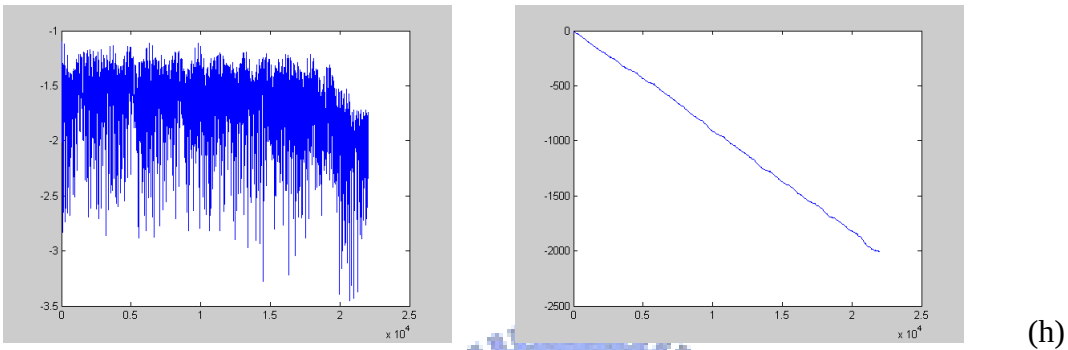
(e)

(f)



(g)



(h)

Fig. 4.29 ATF Characteristic with Different Wall Materials,

Left: Freq. log10 Magnitude, Right: Unwrapped Phase

(a) No Reflection (b) Perfect Reflector (c) Heavy Carpet (d) Concrete

(e) Heavy Glass (f) Gypsum Board (g) Wood with Airspace (h) Plaster on Metal

4.4 Demonstrations of 3D Acoustic Signal Synthesis Results

In Fig. 4.30, we show the 3D acoustic signal synthesis flow. By dividing the separated signals into parts, we are able to build the 3D acoustic signal as the designed HRTF scenario. It can be done by filtering each divided parts with its corresponding ATF and HRTF. The order of ATF filtering and HRTF filtering does not affect the output signal but the computational complexity since the HRTF filtering produce a two channel signal for each input signal.

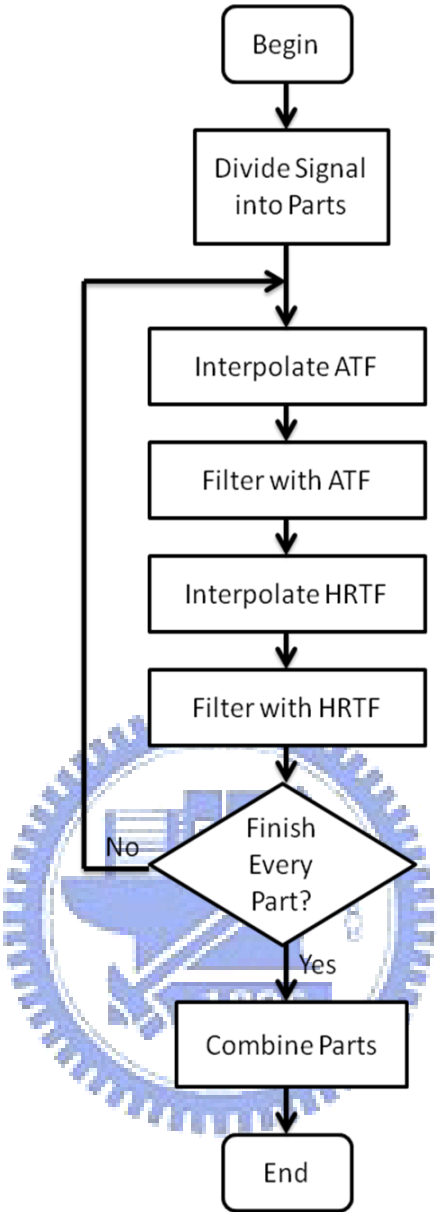
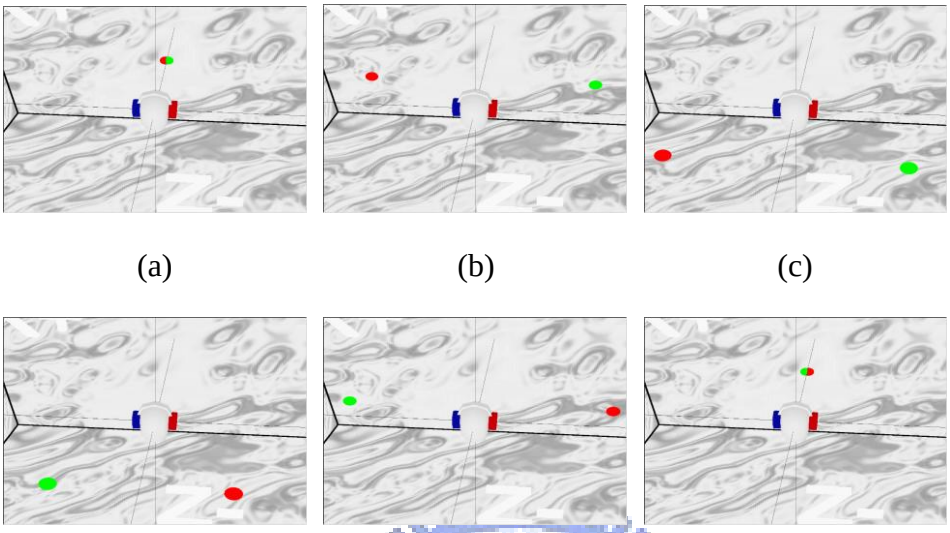


Fig. 4.30 Flow Diagram of 3D Acoustic Signal Synthesis

For each sequence data, we provide three kinds of waveforms: the SLAB synthesis waveform, the HRTF+ATF waveform from the original source signals and the HRTF+ATF waveform from the separated signals.

The demonstrations show two kinds of HRTF scenarios. The first scenario which is shown as Fig. 4.31 has 25 frames and the frame interval is about 0.5 second. The second scenario which is shown as Fig. 4.32 has 27 frames and the frame interval is also about 0.5

second. The red point represents the source 1, the green point represents the source 2 and the blue and red parts of the headphone represent the left and right ear of HRTF respectively.



(a) (b) (c)

(d) (e) (f)

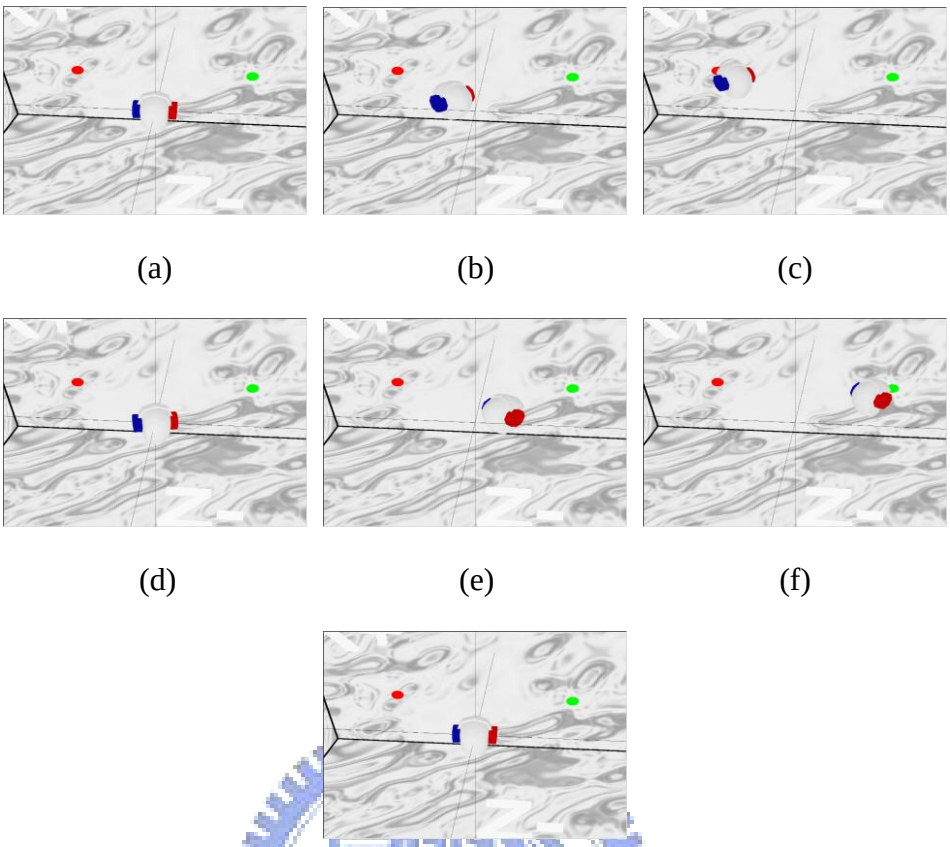
Fig. 4.31 HRTF Scenario 1,

25 Frames, Frame Interval ≈ 0.5 sec,

Red: Source 1, Green: Source 2

(a) Frame 1 (b) Frame 5 (c) Frame 10

(d) Frame 15 (e) Frame 20 (f) Frame 25



(a) (b) (c)

(d) (e) (f)

(g)

Fig. 4.32 HRTF Scenario 2,

27 Frames, Frame Interval ≈ 0.5 sec,

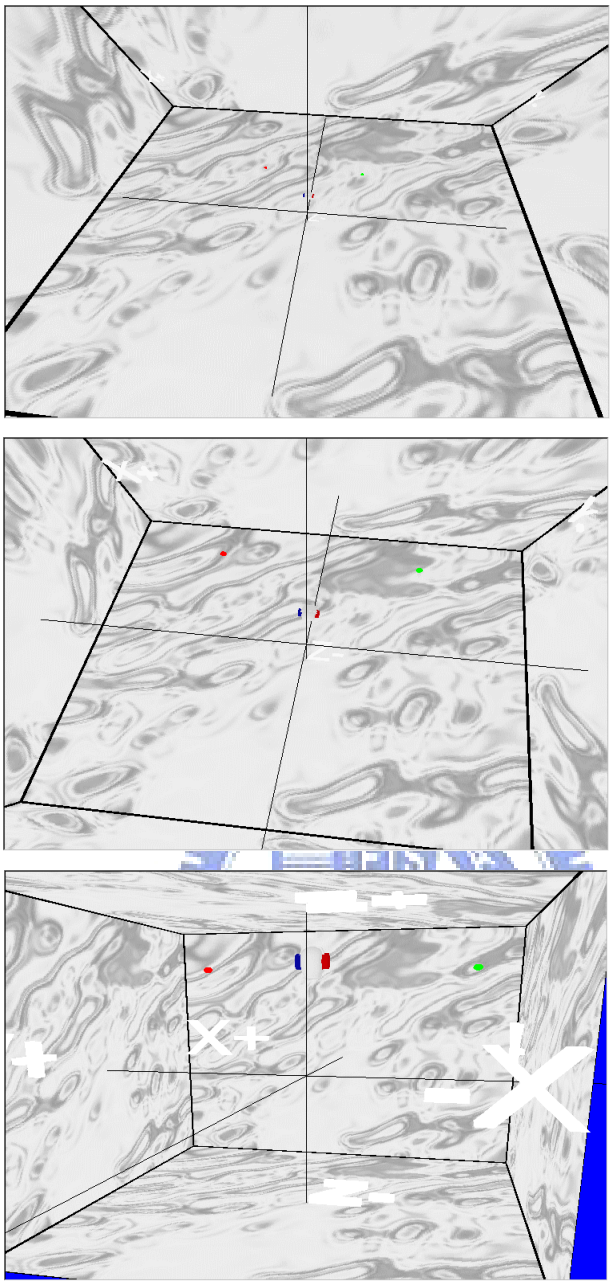
Red: Source 1, Green: Source 2

(a) Frame 1 (b) Frame 5 (c) Frame 8

(d) Frame 13 (e) Frame 18 (f) Frame 21

(g) Frame 27

In order to amplify the noticeable effect of the ATF, we demonstrate the 3D acoustic signals for three different room sizes: large room with 20 x 20 x 20 (m), median room with 10 x 10 x 10 (m) and small room with 4 x 4 x 4 (m), which are shown in Fig. 4.33.



(a)

(b)

(c)

Fig. 4.33 Different Room Sizes

(a) Large Room (b) Medium Room (c) Small Room

From Fig. 4.34 to Fig. 4.41, we can observe the effects of ATF to the waveforms and the spectrograms. By the comparisons of the figures in (a) and the ones in (c), it can be identified that the ATFs change the waveforms of the separated signals; the difference is implicit without reflection (NR), but it is visible for perfect reflectors (PR) as the wall material in the three different room sizes (Small, Medium, Large). The effect of room sizes

to ATFs can be observed in (f). The longer the reverberation time is, the faster the changes in the adjacent frequencies are. The explanation comes from the sum of different time domain shifting of signals cause the frequency domain magnitude variation:

$$\left| DFT\left\{ \sum_{k=0}^{n} a_k s(t-t_k) \right\} \right| = \left| \sum_{k=0}^{n} a_k e^{-j2\pi f \cdot t_k} S(f) \right|$$

$$= \sqrt{\sum_{k=0}^{n} a_k^2 + \sum_{k=0}^{n}\sum_{m=0}^{n} a_k a_m \cos\left(2\pi f\left(t_k - t_m\right)\right)} \cdot |S(f)| .$$

Therefore, for a larger room, there exists some larger value of $t_k - t_m$ which cause a faster oscillation of the spectrum. By comparing the spectrograms in (e) with those in (b), we are able to see some blue slices at the frequencies with lower spectrum magnitudes in (f). After the HRTF filtering, the interchannel level difference (ILD) is noticeable in (d), which is related to the HRTF azimuth angle. For the signals at 45°, the left channel amplitude is much larger than the right one; in the other hand, for those at -45°, the right channel amplitude is larger than the left one.

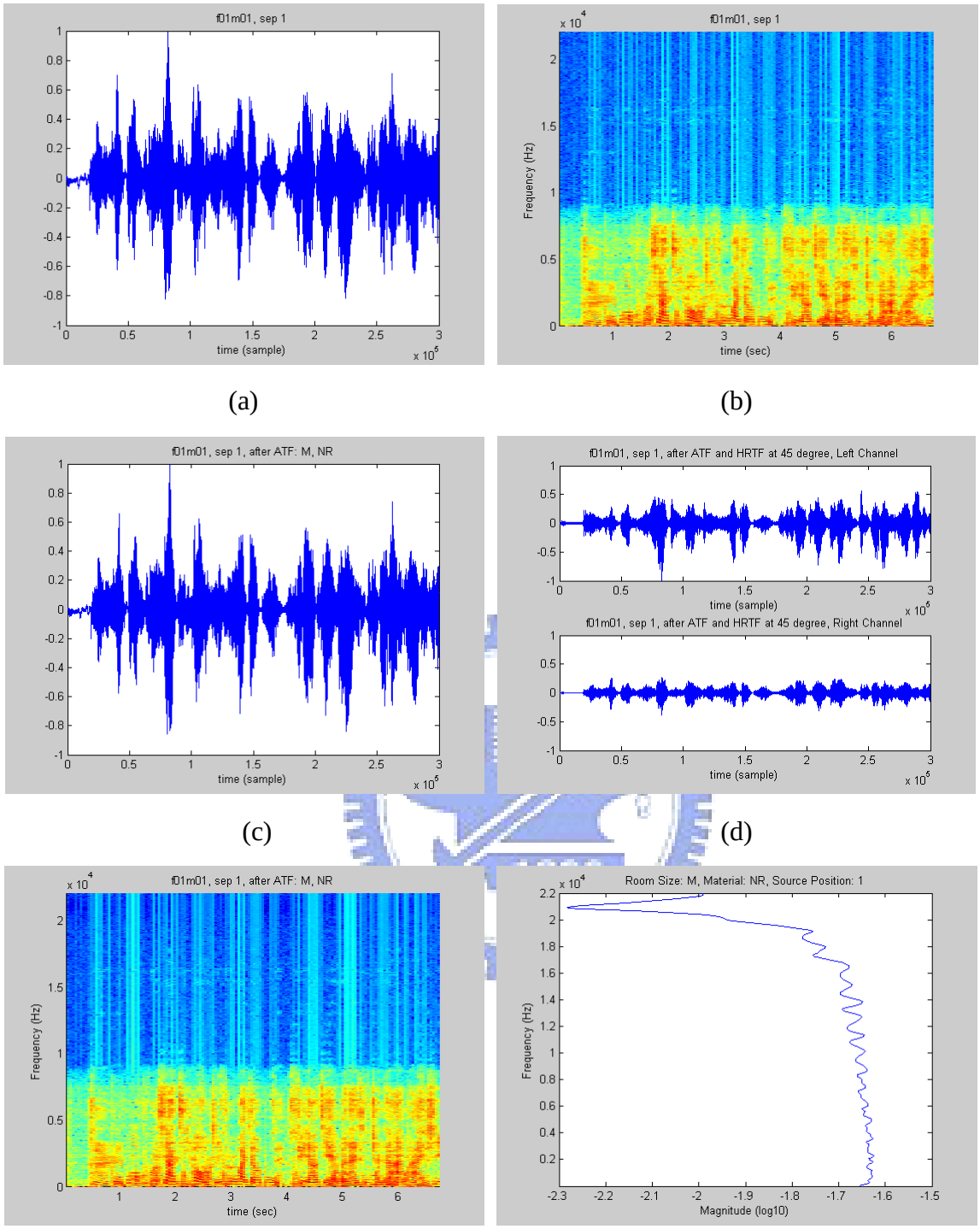


Fig. 4.34 “f01m01”, Separated Signal 1, NR, HRTF at 45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF

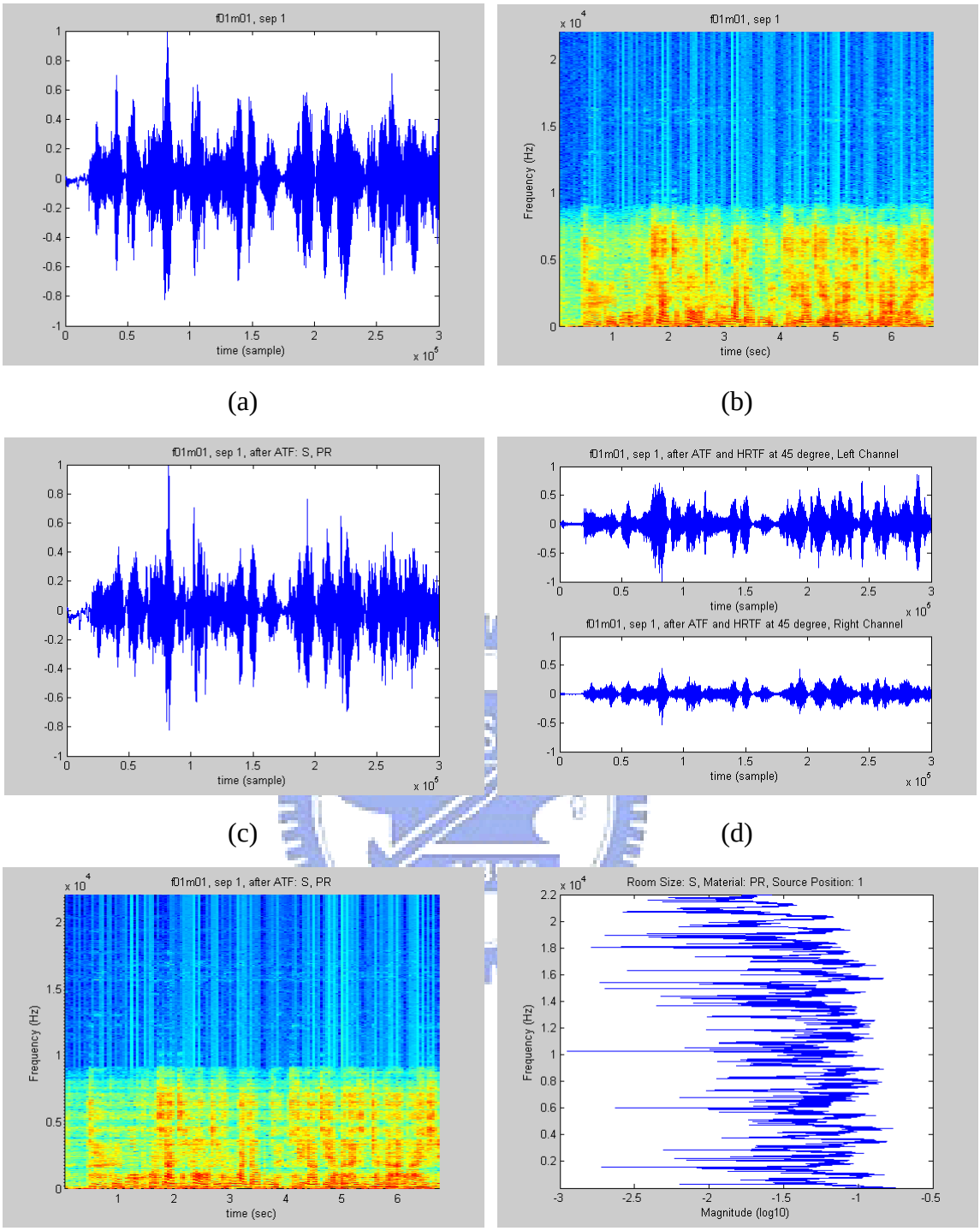


Fig. 4.35 "f01m01", Separated Signal 1, Small Room, PR, HRTF at 45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF

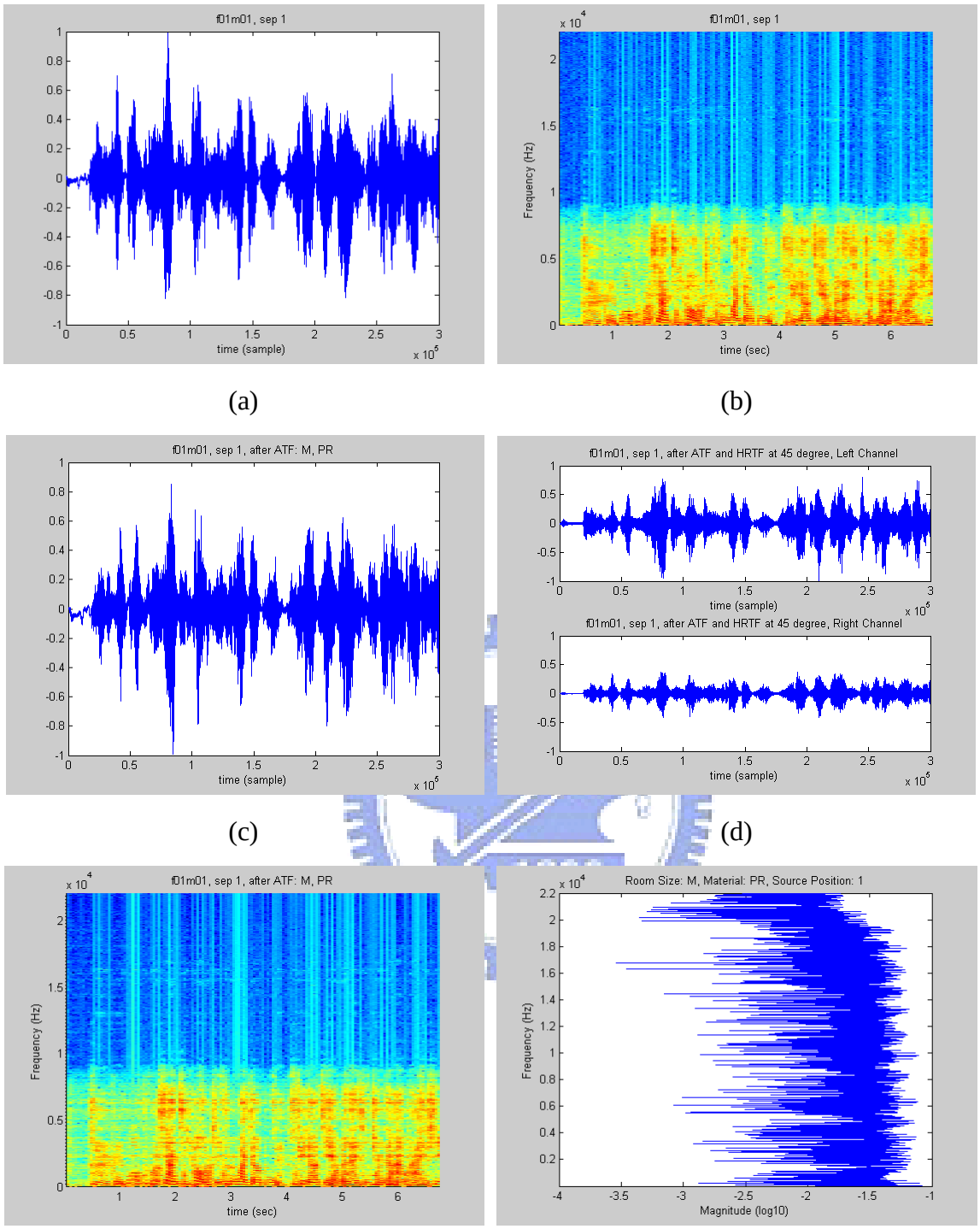


Fig. 4.36 "f01m01", Separated Signal 1, Medium Room, PR, HRTF at 45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF

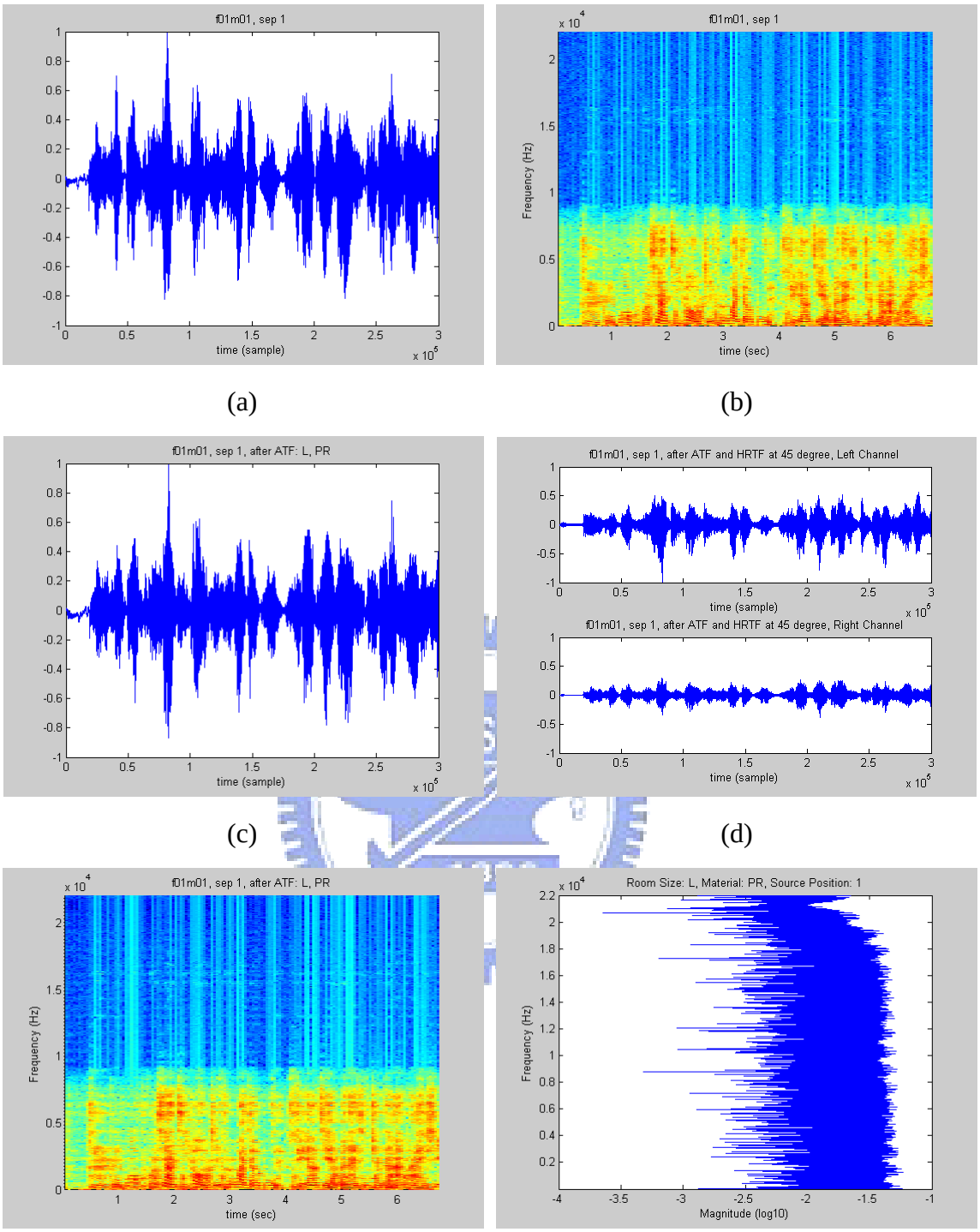


Fig. 4.37 “f01m01”, Separated Signal 1, Large Room, PR, HRTF at 45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF

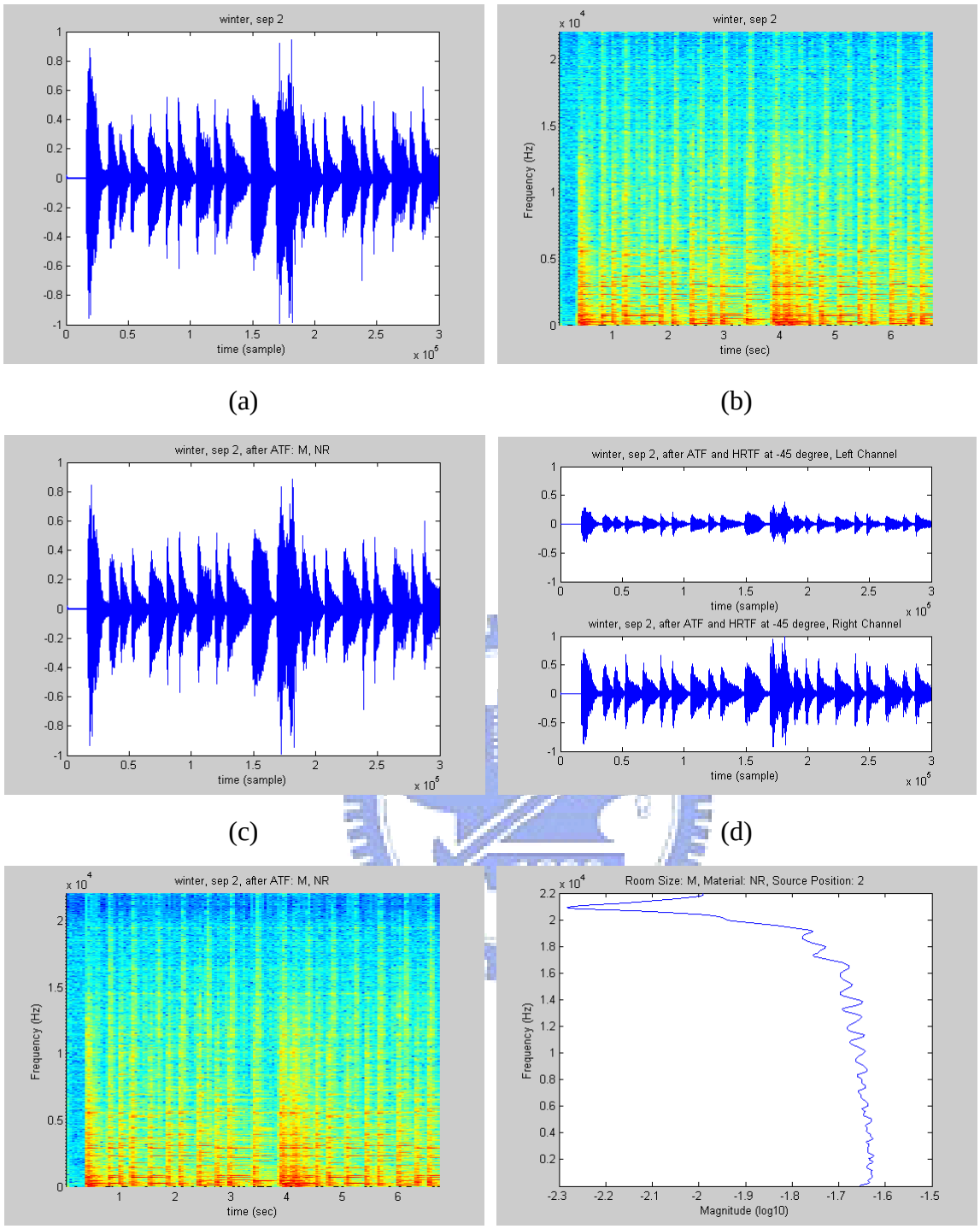
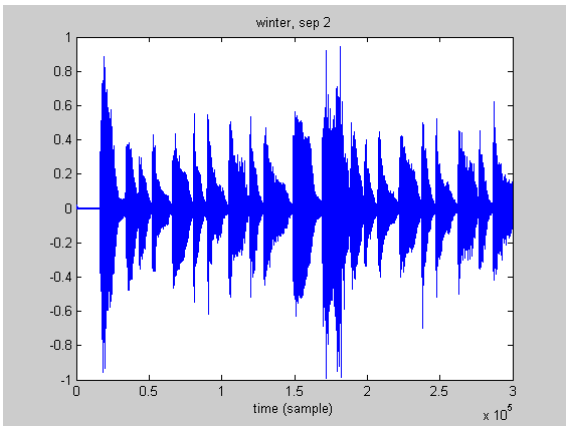


Fig. 4.38 "winter", Separated Signal 2, NR, HRTF at –45°

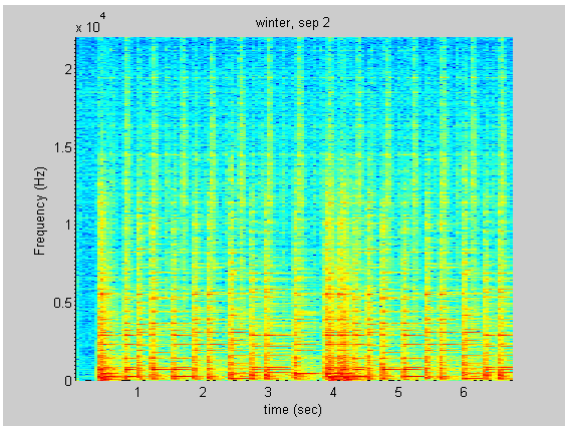
(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

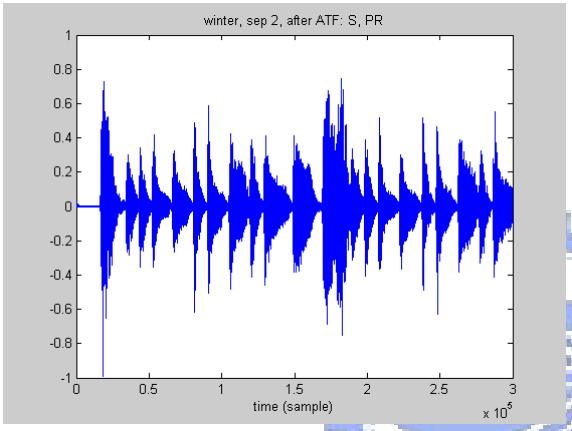
(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF



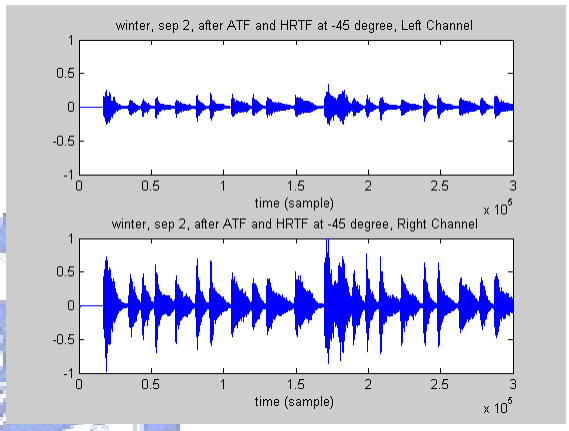
(a)



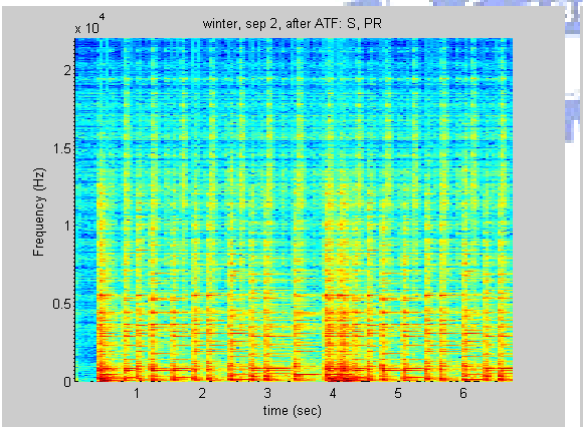
(b)



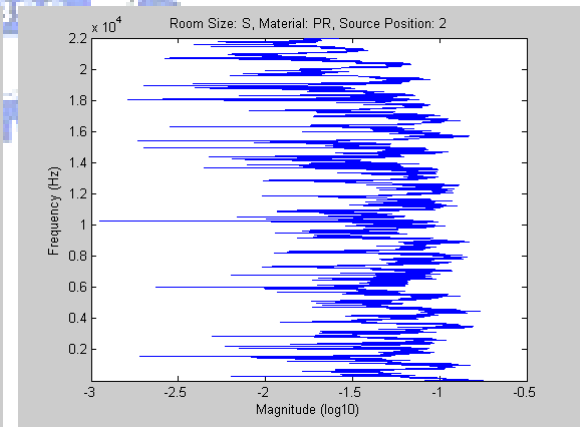
(c)



(d)



(e)



(f)

Fig. 4.39 "winter", Separated Signal 2, Small Room, PR, HRTF at –45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF

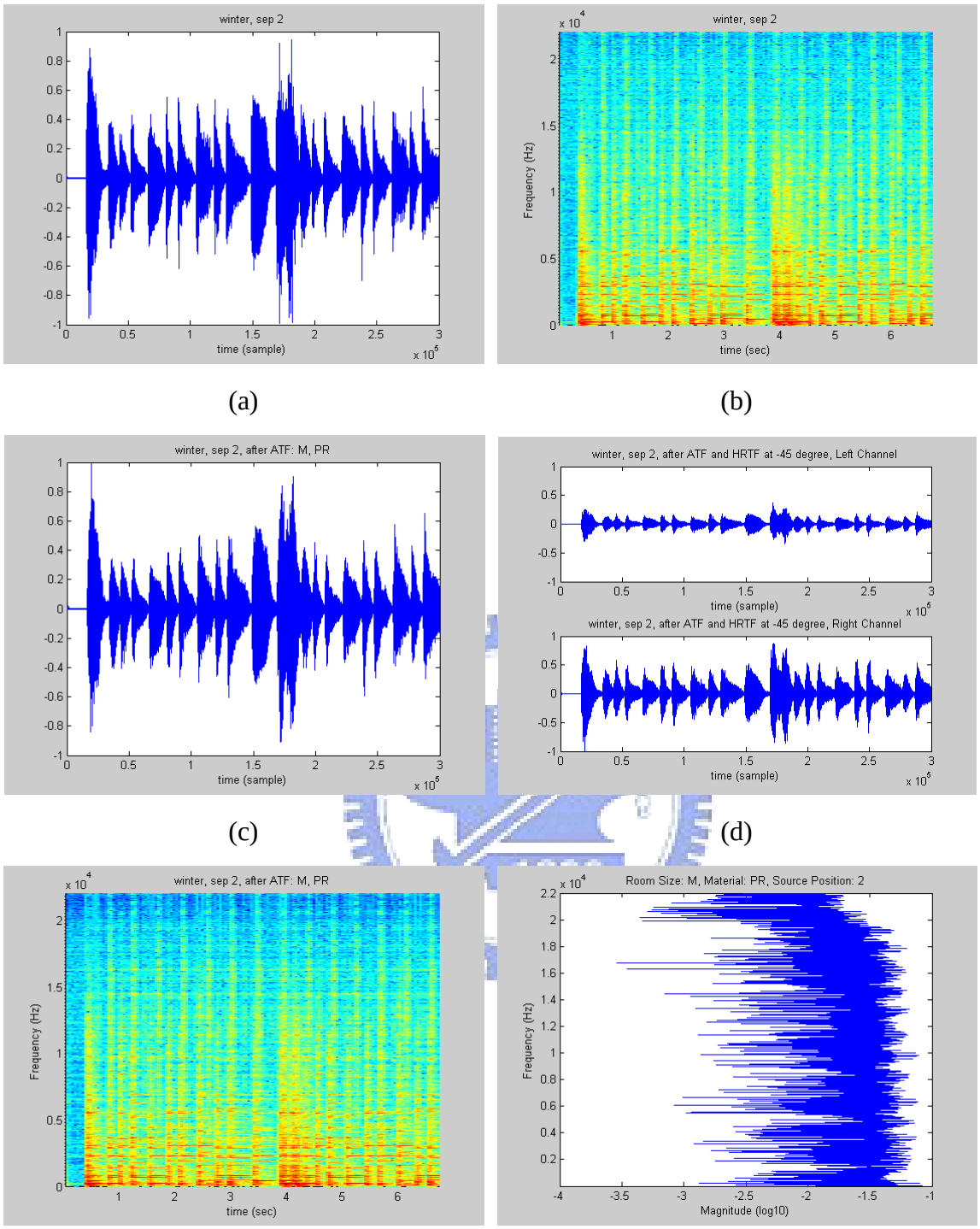
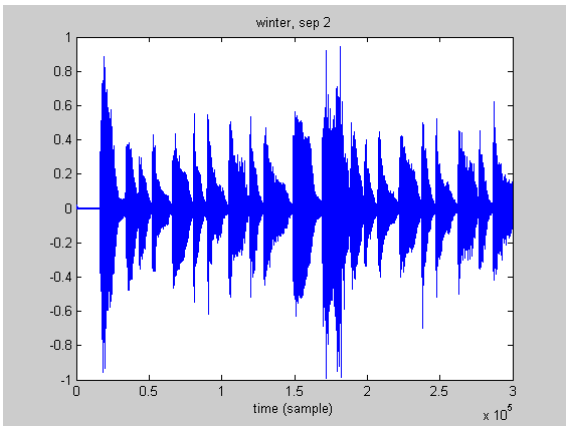


Fig. 4.40 "winter", Separated Signal 2, Medium Room, PR, HRTF at –45°

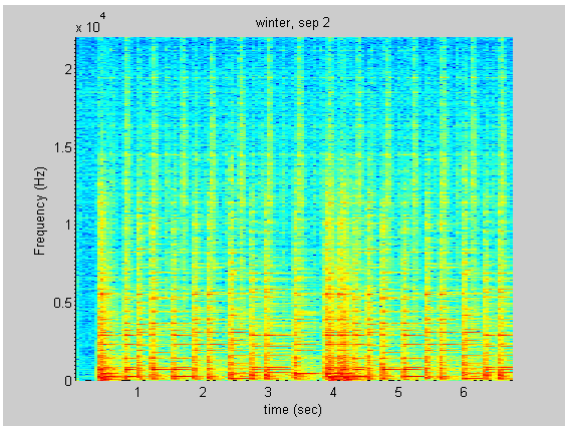
(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

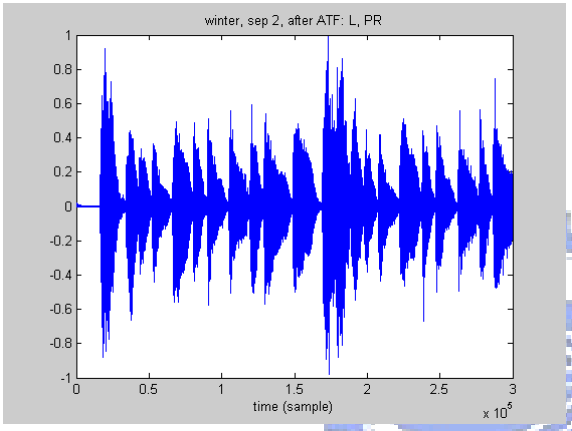
(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF



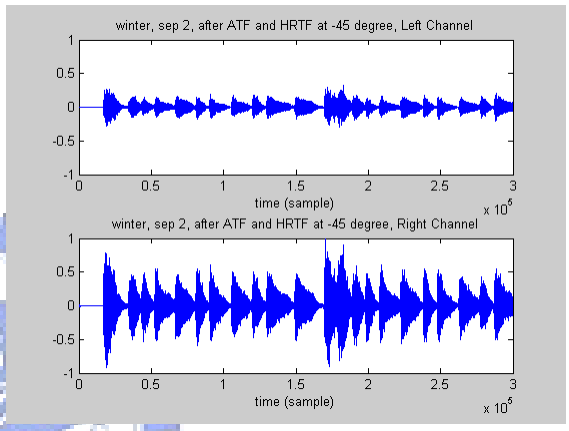
(a)



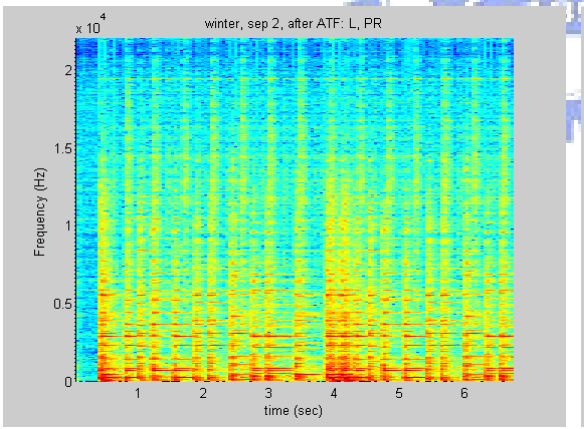
(b)



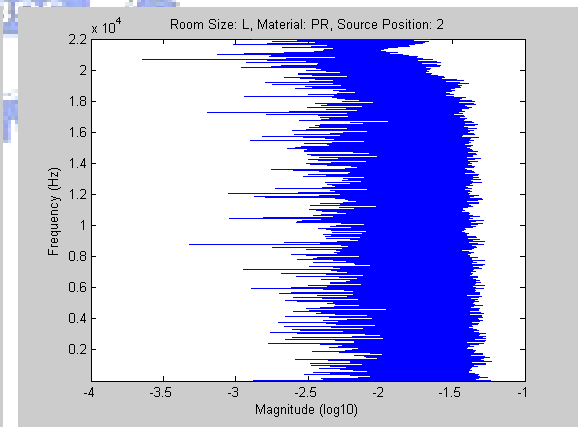
(c)



(d)



(e)



(f)

Fig. 4.41 "winter", Separated Signal 2, Large Room, PR, HRTF at –45°

(a) Separated Signal in Time Domain (b) Separated Signal in Time-Frequency Domain

(c) After ATF in Time Domain (d) After HRTF in Time Domain

(e) After ATF in Time-Frequency Domain (f) Log 10 Magnitude of ATF


E
S
A
1896

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main propose of this thesis is to synthesize the 3D acoustic signal at a virtual listening point from the captured microphone array signals. We adopt the known BSS method to separate the sound source signals from the received microphone array signals. The PCA method is used to extract on the direct components of the source signals and discard the reverberant components and the noise energy. The permutation and scaling problems of FD-ICA are solved by the hybrid DOA and correlation method and the MDP, respectively. A least squares optimization technique based on the cross-power-spectrum approach with the gradient descent algorithm is used for the blind separation of the convolutive mixture signals. The separated signal quality is evaluated by SIR. The simulation and discussion on the SIR values, waveforms, and spectrograms of each input sequence are presented in section 4.1.

To construct a 3D audio on the headphone, the separated signals are filtered by the HRTF and the ATF at the virtual listening point. The interpolation methods of the HRTF and the ATF at the virtual listening point are derived in chapter 3. Chapter 4 discusses the ATFs of different room sizes and different wall materials. The spatial impression, which is given by the combination of the HRTF and the ATF, is demonstrated with the resulting 3D acoustic signals.

The SLAB software is used to generate the audio signals in a room, to capture the microphone array signals, and to measure the ATFs in different room sizes and wall materials. The afterward signal processing implementation is done in MATLAB. The

spectrograms of signals in each stage are shown to visualize the signal envelope transition process. Different HRTF scenarios are employed to demonstrate the 3D acoustic feeling of the synthesized signals.

5.2 Future Work

This thesis concentrates on the overall combination of BSS, HRTF and ATF to produce the 3D acoustic signal at a virtual listening point. Yet there are many extensions can be made to improve the quality of the 3D acoustic signal. For example, the source signal location detection can complete the sound field reconstruction and it is also helpful to obtain the corresponding ATF. Another possible subsequent work is the synthesis of moving 3D acoustic signals considering the Doppler effect of the frequency variation along with the relative velocity of each source signal to the virtual listening point. It is also expected to reduce the computational complexity of the overall process, which aims at the real time synthesis of the 3D acoustic signals. Additionally, the graphic user interface (GUI) can improve the interaction of selecting the virtual listening point in the specific acoustic room.

References

[1] S. Choi, et al., "Blind Source Separation and Independent Component Analysis: A Review," *Neural Information Processing - Letters and Reviews,* vol. 6, no. 1, Jan. 2005.

[2] 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.

[3] E. Binghan and A. Hyvarinen, "A Fast Fixed-point Algorithm for Independent Component Analysis of Complex Valued Signals," *International Journal of Neural Systems,* vol. 10, no. 1, pp. 1-8, Feb. 2000.

[4] A. Bell and T. Sejnowski, "An Information-maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation,* vol. 7, pp. 1129-1159, 1995.

[5] S. Haykin, Ed., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation),* John Wiley & Sons, 2000.

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

[7] A. Hyvarinen and E. Oja, "A Fast Fixed-point Algorithm for Independent Component Analysis," *Neural Computation 9,* 1483-1492, 1997.

[8] S. Ikeda and N. Murata, "A Method of ICA in Time-Frequency Domain," in *Proc. ICA'99,* pp. 365-371, Jan. 1999.

[9] S. Amari, et al., "Stability Analysis of Learning Algorithms for Blind Source Separation," *Neural Networks,* vol. 10, no. 8, pp. 1345-1351, 1997.

[10] P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain," in *Proc. Int. Workshop on Independence and Artificial Neural Networks,* 1998.

[11] S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computing*, vol. 10, no. 2, pp. 251-276, 1998.

[12] S. Winter, et al., "Geometrical Understanding of the PCA Subspace Method for Overdetermined Blind Source Separation," in *Proc. ICASSP*, pp. 769-772, Apr. 2003.

[13] K. Niwa, et al., “Encoding Large Array Signals into a 3D Sound Field Representation for Selective Listening Point Audio based on Blind Source Separation,” *ICASSP2008*(AE-P2.E10), pp. 181-184, 2008.

[14] K.Niwa, et al., "Selective Listening Point Audio Based on Blind Source Separation and Stereophonic Technology", *IEICE Trans. of Information and System*, vol.E92-D, no.3, Mar. 2009.

[15] 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.

[16] K. Matsuoka and S. Nakashima, “Minimal Distortion Principle for Blind Source Separation,” in *Proc. ICA*, pp. 722-727, Dec. 2001.

[17] 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.

[18] K. Niwa, et al., "Development of Selectable Viewpoint and Listening Point System for Musical Performance," *ICA2007*, PPA-06-011, 2007.

[19] M. P. Tehrani, et al., "3DAV Integrated System Featuring Arbitrary Listening-point and Viewpoint Generation," in *Proc. of IEEE Multimedia Signal Processing*, MMSP 2008, PID-213, pp. 855-860, Oct. 2008.

[20] Wikipedia of HRTF: http://en.wikipedia.org/wiki/HRTF.

[21] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, pp. 2-24, Apr. 1988.

[22] Y. Suzuki, et al, "An Optimum Computer-Generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses", *J. Acoust. Soc. Am.*, vol.97(2), pp. 1119-1123, 1995.

[23] TSP design:

http://tosa.mri.co.jp/sounddb/tsp/tsp_design_e.htm.

[24] An online HRTF database:

http://recherche.ircam.fr/equipes/salles/listen/index.html.

[25] 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.


E
S
A
1896

自傳

張欽淵，1985 年 7 月 21 日出生於新竹市。2007 年畢業於國立交通大學電機資訊學院學士班，之後進入國立交通大學電子研究所攻讀碩士學位，研究方向為多媒體訊號處理，論文題目為「由麥克風陣列訊號合成出虛擬聆聽點的３Ｄ音訊」。