Virtual Listening Point Audio Synthesis using Anechoic Chamber Recording

Shih-Jie Chien (簡士傑)
Department of Electronics Engineering
National Chiao Tung University
Hsinchu, Taiwan, R.O.C
shih.jie30@gmail.com

Hsueh-Ming Hang (杭學鳴)
Department of Electronics Engineering
National Chiao Tung University
Hsinchu, Taiwan, R.O.C
hmhang@mail.nctu.edu.tw

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

Keywords Blind Source Separation; Direction of Arrival; SLAB;

I. INTRODUCTION

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

For the first step, we use the blind source separation (BSS) technique to separate individual sound source from the mixed signals. We model and use the known mathematical tools [1] to solve the separation problem. The subspace of interest is extracted by the principal component analysis (PCA) method [2]. For solving the permutation problem, we adopt [3]. The scaling problem is solved by the minimum distortion principle (MDP) [4]. There are many well-known BSS methods and one of the most popular methods is the so-called independent vector analysis (IVA) [3]. The IVA method has different learning rules [5] and different properties from the conventional ICA methods.

For the second step, we use the direction of arrival (DOA) technique to locate individual sound source from the mixed signals. The time difference of arrival (TDOA) is a basic concept to explain the technique. It also has to satisfy some conditions in order to avoid the spatial aliasing [6]. The DOA technique can be solved under the invariant property assumption [7]. We adopt [8] to estimate DOA estimation for 3-D sources.

For the third step, we separate sources and identify their locations using the methods described in the first step and the second step. We adopt the software developed by the NASA Ames Research Center [9] to synthesize the audio at a virtual listening point.

Because recording the audio signal in a real world environment, we also need to consider the noise effect. We adopt [10] to reduce the noise in our system.

II. MIXED SIGNAL MODEL

In this paper, we use the multiple-input multiple-output system to model the sound signals with microphone array, and we assume that there is no room reflections and ambient noises in an anechoic chamber. Considering the mixture model, we convert the time-domain signals into frequency-domain by Short-Time Fourier Transform (STFT). Assume the system model involves *K* input signals and *N* output signals, which can be modeled as:

$$\mathbf{x}(f,t)=\mathbf{A}(f)\mathbf{s}(f,t) \tag{1}$$

$$\mathbf{y}(f,t)=\mathbf{W}(f)\mathbf{x}(f,t) \tag{2}$$

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

where F denotes the number of frequency bin; $\mathbf{x}(f,t)$ and $\mathbf{y}(f,t)$ denotes the source signals and the separated signals at frequency *f*; $\mathbf{A}(f)$ is the $N\times K$ mixed matrix or also called the steering vector matrix; $\mathbf{W}(f)$ is the $K\times N$ demixing matrix.

III. ACOUSTIC SIGNAL PROCESSING AND SYNTHESIS

In this section, we review the BSS technique and DOA estimation procedures using the IVA algorithm [3] and ICA-based algorithm [8]. We also describe the adopted denosing method [10] for improving audio quality.

This work was supported in part by the NSC, Taiwan under Grants 98-2221-E-009-087 -MY3 and by the Aim for the Top University Project of National Chiao Tung University, Taiwan.

A. *Blind Source Separation (BSS)*

We adopt [3] as the ICA method. The Independent Vector Analysis (IVA) method uses a different approach to solve the BSS problem by assuming that the source signals have certain dependency in the frequency domain. Under this hypothesis, the original sources are dependent together as a group by using the multidimensional prior. The model is a maximum likelihood approach to the multidimensional ICA (MICA), which is called independent vector analysis. For BSS problem, the main target is to find the demixing matrix $\mathbf{W}(f)$.

The ICA algorithm based on IVA consists of two steps. The first step is to find contrast function as the input learning function. The second step is to choose optimization method. Once the contrast function is selected, we can derive the separating matrix by selecting the optimization method. [6] uses the Newton's method, which is called FastICA algorithm. Here, we assume $g(\cdot)$ is the input learning function, which can be expressed as:

$$g(\mathbf{W}_i(f)) = \hat{E}\left[G\left(\sum_f |\mathbf{W}_i(f)^H \mathbf{W}_i(f)|^2\right) - \sum_f \lambda_i(f)\left(\mathbf{W}_i(f)^H \mathbf{W}_i(f) - 1\right)\right] \tag{3}$$

where the Lagrange multiplier $\lambda_i(f)$ can be expressed by

$$\lambda_i(f) = \hat{E}\left[|y_{i,o}(f)|^2\, G'\left(\sum_f |y_{i,o}(f)|^2\right)\right] \tag{4}$$

The function can be approximated by the quadratic Taylor polynomial. The optimization of $g(\cdot)$ will set the gradient $\partial g(\cdot)/\partial$ to zero. The iterative algorithm becomes as the following equation:

$$\begin{aligned}\mathbf{W}_i(f) = \hat{E}\left[G'(\sum_f |y_{i,o}(f)|^2) + |y_{i,o}(f)|^2\, G''(\sum_f |y_{i,o}(f)|^2)\right]\mathbf{W}_{i,o}(f) \\ -\hat{E}\left[(y_{i,o}(f)^*)G'(\sum_f |y_{i,o}(f)|^2)\mathbf{x}(f)\right]\end{aligned} \tag{5}$$

where $G(z) = \sqrt{\frac{2z}{f}} + (F - \frac{1}{2})\log z$ with the constraint that $\mathbf{W}_i(f)$ are normalized; G' and G'' denotes the first order and second order differentials. In addition to normalization, the rows of the demixing matrix $\mathbf{W}$ have to be decorrelated. The learning rules of $\mathbf{W}$ can be expressed as:

$$\mathbf{W}(f) \leftarrow \left(\mathbf{W}(f)\mathbf{W}(f)^H\right)^{-\frac{1}{2}} \mathbf{W}(f) \tag{6}$$

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

B. *Direction of Arrival (DOA)*

ICA-based algorithm is used to estimate the demixing matrix $\mathbf{W}$ in solving the BSS problem. We assemble three microphones as a microphone array for estimating the azimuth and elevation of the source signal as shown in Fig. 1.

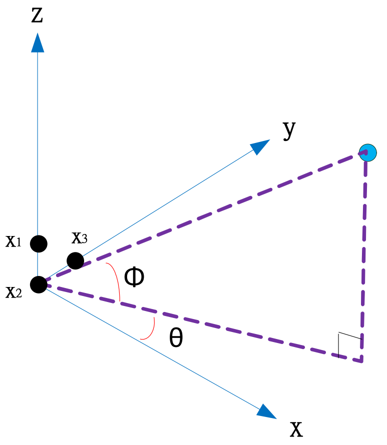


Figure 1. Spatial Relationship of a Microphone Array and a Source Signal

Considering the mixture model, we convert the time-domain signals into frequency-domain, and the mixed matrix can be expressed as:

$$\mathbf{A}(f) = \left[\mathbf{a}_1(f,\theta_1,\phi_1) \quad \cdots \quad \mathbf{a}_K(f,\theta_K,\phi_K)\right] \tag{7}$$

and

$$\mathbf{a}_k(f,\theta_k,\phi_k) = \left[a_{1k}(f,\theta_k,\phi_k) \quad \cdots \quad a_{Nk}(f,\theta_k,\phi_k)\right]^T \tag{8}$$

$$a_{nk}(f,\theta_k,\phi_k) = g_{nk} \exp\left\{ j\frac{2\pi f}{c} \vec{r}^{\,T} \cdot \vec{v}(\theta_k,\phi_k)\right\} \tag{9}$$

where $\mathbf{A}(f)$ is the mixing matrix, whose k-th column vector represents the transfer function of the k-th source signal, which is the so-called steering vector matrix. The g_{nk} denotes the gain of a_{nk}, $\vec{r} = (x_n, y_n, z_n)^T$ denotes the coordinate vector of the n-th microphone, and $\vec{v}(\theta_k,\phi_k) = \left(\cos\theta_k \cos\phi_k, \sin\theta_k \cos\phi_k, \sin\phi_k\right)$ represents the look direction vector of the k-th signal as shown in Fig. 1. Then, we can obtain the equation by dividing two elements [8] as shown in following equation:

$$\frac{a_{1k}}{a_{2k}} = \left|\frac{a_{1k}}{a_{2k}}\right| \exp\left\{ j\frac{2\pi f}{c}\left[(y_1 - y_2)\sin\theta_k \cos\phi_k + (z_1 - z_2)\sin\phi_k\right]\right\} \tag{10}$$

$$\frac{a_{1k}}{a_{3k}} = \left|\frac{a_{1k}}{a_{3k}}\right| \exp\left\{ j\frac{2\pi f}{c}\left[(y_1 - y3)\sin\theta_k \cos\phi_k + (z_1 - z_3)\sin\phi_k\right]\right\} \tag{11}$$

where

$$A = \frac{angle\left(a_{1k} / a_{2k}\right)}{2\pi f c^{-1}} \tag{12}$$

$$B = \frac{angle\left(a_{1k} / a_{3k}\right)}{2\pi fc^{-1}} \tag{13}$$

Then, we extract the angles ϕ and θ:

$$\theta_k = \cos^{-1}\left\{\frac{(y_1-y_3)A-(y_1-y_2)B}{\left[(x_1-x_2)(y_1-y_3)-(x_1-x_3)(y_1-y_2)\right]\cos\phi_k}\right\} \tag{14}$$

$$\phi_k = \sin^{-1}\left\{\frac{(x_1-x_2)B-(x_1-x_3)A}{(x_1-x_2)(y_1-y_3)-(x_1-x_3)(y_1-y_2)}\right\} \tag{15}$$

C. Audio Denoising

In a real acoustic environment, the environment parameters including the air absorption, the surface reflection and microphone intrinsic distortion, and others, all generate audio noises. However, in many cases, it is assumed that there is no reverberation effect, which is called the single-path assumption. We adopt [10] to solve the denoising problem in our system.

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

$$z_i = x_i + n_i \tag{16}$$

where n_i denotes a zero-mean white Gaussian random sequence and x_i denotes the audio signals.

There are two adaptive filters to be combined into a contextual adaptive Wiener filter [10], which is represented by

$$\hat{x}_i = \bar{x}_i + \left(\frac{\bar{\sigma}_x^2}{\bar{\sigma}_x^2+\sigma_n^2}\right)\left[\alpha(z_i-\bar{x}_i)+(1-\alpha)\sum_{z_j\in\eta_i}(z_j-\bar{x}_i)\right] \tag{17}$$

where $\hat{x}_i$ denotes an estimate (filter output) of the i-th sample point; $\bar{x}_i$, $\bar{\sigma}_x^2$ and σ_n^2 denote the sample mean, the sample variance and the noise variance in a W-size window; η_i denotes the neighboring element of the i-th sample. And $\alpha \in [0,1]$. According to the Peak Signal-To-Noise Ratio (PSNR) measured in [10], $\alpha = 0.79$ is the optimum tradeoff value between two adaptive filters.

D. Virtual Listening Point Audio Synthesis

Fig. 2 shows the acoustic signal synthesis flowchart. We are able to construct the acoustic signal at the desired virtual listening position, which is the so-called *Virtual Listening Point Audio Synthesis*. We assume that there are two source signals and one microphone array in our experiment. The task includes three major steps. First, we adopt [3] to separate the mixture signals recorded by the microphone array. Second, by employing the IVA method, we can obtain the demixing matrix $\mathbf{W}(f)$. Thus, we derive the steering matrix $\mathbf{A}(f)=\mathbf{W}(f)^{+}$, where $\mathbf{W}(f)^{+}$ denotes the pseudo-inverse of $\mathbf{W}(f)$. Then, we use the steering matrix $\mathbf{A}(f)$ and [8] to estimate the DOA of two source signals. Third, we select an arbitrary point to synthesis the virtual audio in the space.

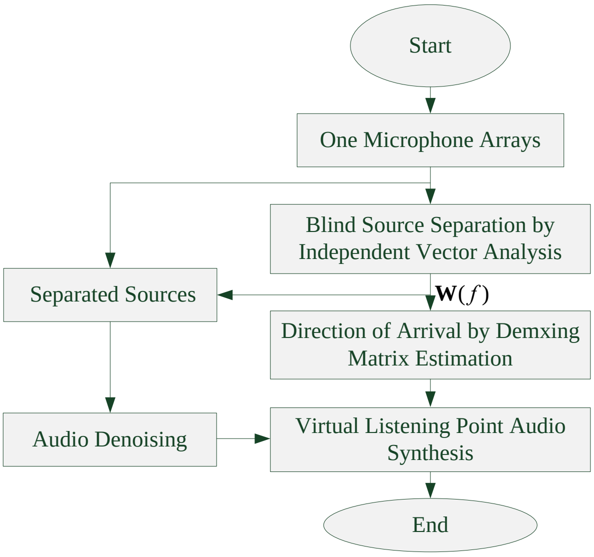

Start
One Microphone Arrays
Blind Source Separation by Independent Vector Analysis
Separated Sources
$\mathbf{W}(f)$
Direction of Arrival by Demxing Matrix Estimation
Audio Denoising
Virtual Listening Point Audio Synthesis
End

Figure 2. Flowchart of 3D Acoustic Signal Synthesis

We adopt the software developed by [9] to synthesize the virtual listening point. The software implements the spatial 3D-sound processing procedure. We perform BSS to separate signals from the recorded mixture signals. Then, we take separated signals as inputs. Fig. 3 shows the arrangement of separated signals and the microphone array on the X-Y plane. S_1, S_2 and P_o respectively represent the first source, the second source and the position of the original microphone array. θ_1, θ_2 respectively represent the azimuth angles of the first source and the second source. d_1, d_2 respectively represent the distances of the first source and the second source from the microphone. Here, the distances are true outcomes. We then synthesize the audio signals at P_1, P_2, P_3 and etc. Thus, we obtain the virtual listening point audio signals.

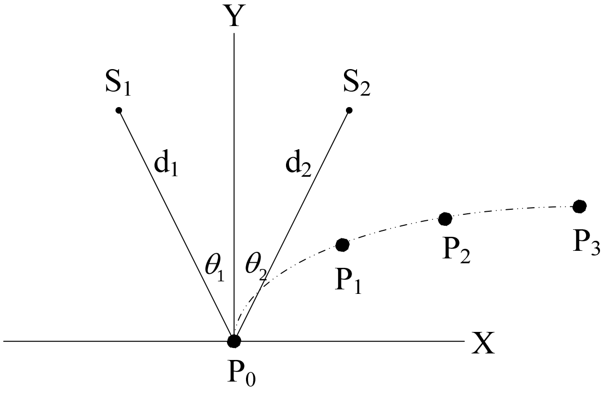


Figure 3. Schematic Diagram of Audio Synthesis

IV. Experimental Results

An anechoic chamber is a room with special walls designed to prevent the sound reflection. It can also insulate the outside interference or noise. An anechoic chamber is commonly used to conduct experiments for simulating “free field” conditions or noise reduction.

For convenience, we move the speech source instead of moving microphone array in recording. In addition, we set up a source and a sensor at the same horizontal plane in our experiments, which means that the elevation angle ϕ is 0^o.The azimuth angle θ of speech source varies from -30 ^{o} to 30 ^{o} with a 15 ^{o} step. It represents that the angles include $\pm30^o$, $\pm15^o$, 0^o with different directions. The set-up of the microphone array and the sources is shown in Fig. 4. In principle, the estimations would be more accurate when there are more sensors. Here, we show test sequences with seven sensors.

TABLE I. Experimental Setting

Sampling rate	8 kHz
Number of source	4 (two female speech and two male speech)
Frame length	512
Frame shift	128
Window function	Hamming
Array spacing	0.03 m
Room size	4m × 4m × 4m

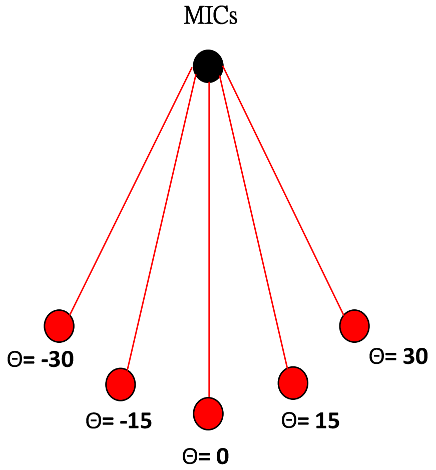


Figure 4. The Locations of the Sources and the Microphone Array

A. *Blind Source Separation*

In this section, we focus on the effect of input data size, that is, we choose different data length of mixture signals to test the BSS algorithm. Starting from a small size inputs, increasing data size can significantly improve the performance. When the input data reach a certain amount, we get less improvement. Therefore, we ought to limit data to a proper size to reduce delay and processing cost. In our experiments, we have one hundred and twenty test sequences. The sequences contain ten combinations of $\theta = 0^o, \pm15^o, \pm30^o$. Each combination has twelve groups (four sources). There are many popular metrics of evaluating the BBS quality, and one way is to measure the Signal to Interference Ratio (SIR).

The data points in Fig. 5 are collected from 120 test sequences. The x-axis represents the size of data length (sampling rate: 8KHz). According to Fig. 5, we observe that the performance of one-second data length, the shortest length, is the worst. When the data length increases to two seconds, the performance gets better obviously. However, we notice that the performance saturates at about four-second data length. In other words, when the data length gets beyond four seconds, the SIR does not gain much.

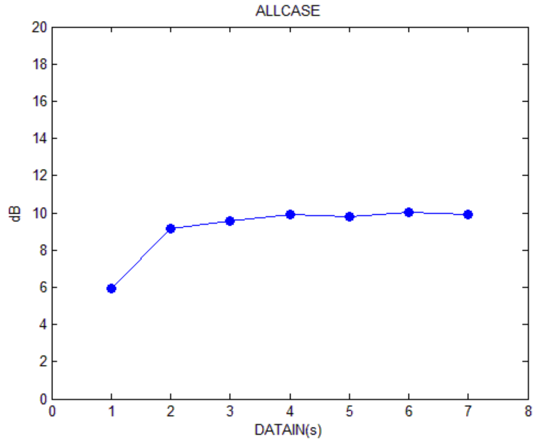
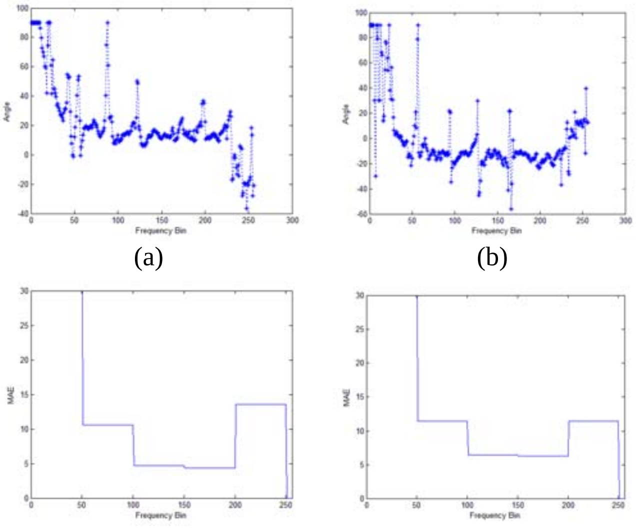


Figure 5. Data Length Test with Seven MICs

B. *Direction of Arrival*

In this section, we focus on the effect of frequency bins in the DOA estimation algorithm. In our experiments, we show a case to derive our selection. We measure the estimation accuracy by using the mean absolute error (MAE).



(a) (b)

(c) (d)

Figure 6. DOA estimates in various bins (Two Males)

Fig. 6(a)~(b) show the angle estimated at each frequency bin for two males. The x-axis represents 256 bins from low frequency to high frequency. The bin size is 15.625 Hz (sampling rate: 8KHz). The source speeches come from 15 ^{o} and -15 ^{o}. We divide the 256 frequency bins into five intervals. Each interval contains 50 frequency bins, and we discard the final 16 bins since a typical speech signal contains less high frequency components. Fig. 6(c)~(d) show the Mean Absolute Error (MAE) corresponding to Fig. 6(a)~(b). In Fig. 6(a)~(b), we observe that the median frequency bins have better estimations. In fact, the situation is reasonable. The low frequency has large wavelength. Theoretically, the wavelength should be smaller than the distance between source and sensor; otherwise, the angle (phase shift) cannot be accurately estimated. Furthermore, there are also high estimation errors in high frequency bins. This is particularly true for two males test sequence. In general, the male voice seldom includes high frequency components. According to the above discussions, we should avoid using low frequencies and high frequencies in DOA estimation. Because of the high estimation errors on the elevation angles, we do not consider the estimation of the elevation angles.

C. Audio Synthesis

In our proposed audio synthesis system, we first perform BSS to separate signals from the recorded mixture signals. Then, we take separated signals as inputs to SLAB. Fig. 7 shows the arrangement of separated signals and the microphone array on the X-Y plane. *Female*_1, *Male*_1 and P_o represent the original recording layout. They are respectively the first separated source, the second separated source and the position of original microphone array. The azimuth angles of the first source and the second source, estimated by the DOA algorithm, are 14^o and 14^o respectively. Here, we do not estimate the distance of the two sources. The distance of these sources from the microphone are the true values, 1.5M and 1.5M respectively. With all the above set-up, we can synthesize the virtual listening point audio using SLAB. Fig. 8 (a)~(d) show the audio signals we synthesize at P_1, P_2, P_3, P_4. The X-Y coordinates in Table II sent the exact positions in Fig. 7.

TABLE II. Spatial Locations

	Female_1	Male_1	P_1	P_2	P_3	P_4
Coordinate	(-0.36, 1.46)	(0.36, 1.46)	(-1, 1.3)	(-0.4, 1.1)	(0.4, 1.1)	(1, 1.3)

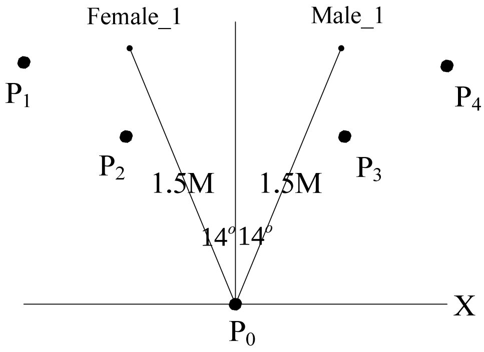


Figure 7. Locations of Synthesized Audio

I. Conclusions

The main propose of this paper is to synthesis virtual listening point audio from the recorded mixture signals in an anechoic chamber. For the BSS technique, the BBS quality provides the better results when the input data are sufficiently abundant. We obtain rather good performance with four-second data length. For the DOA technique, low frequency components are unreliable in DOA estimate because of their long period in time. However, the high frequency components often have high noise. Therefore, from our statistics, the median frequency bins offer more reliable estimates. For the denoising technique, this technique can improve the subjective hearing quality in the BSS method but does not help in the source direction estimations in the DOA method.

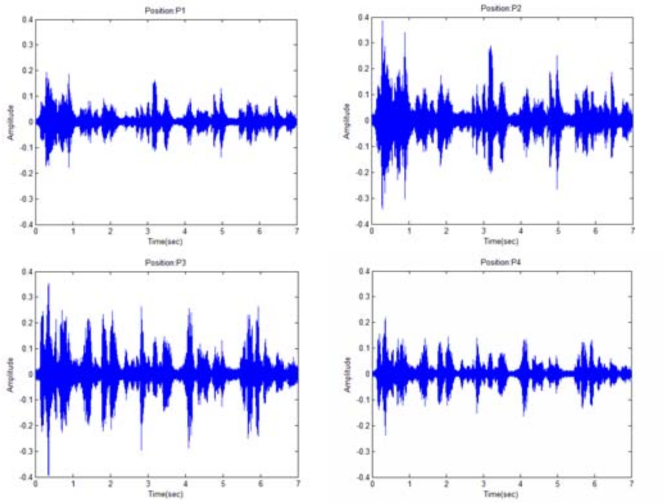


Figure 8. Virtual Listening Point Audio

References

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

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

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

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

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

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

[7] H. Sawada, et al., “Direction of arrival estimation for multiple source signals using independent component analysis,” in *Proc. International Symposium on Signal Processing and its Applications*, pp. 411–414, July 2003.

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

[9] 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.

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