EVALUATION OF BLIND SEPARATION AND DECONVOLUTION FOR BINAURAL-SOUND MIXTURES USING SIMO-MODEL-BASED ICA

Hiroaki Yamajo, Hiroshi Saruwatari, Tomoya Takatani, Tsuyoki Nishikawa and Kiyohiro Shikano

Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, Japan (Asia)
phone: +81 743 72 5287, fax: +81 743 72 5289, email: sawatari@is.aist-nara.ac.jp
web: isw3.aist-nara.ac.jp/IS/Shikano-lab/e-home.html

ABSTRACT

In this paper, blind separation and deconvolution (BSD) problem with binaural-sound mixtures is addressed. We have proposed two-stage blind separation and deconvolution algorithm, which consists of Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) and blind multichannel inverse filtering. In the previous report, we carried out simulations in the artificial mixing system and only showed that the proposed BSD can work theoretically. In order to evaluate the proposed method in more actual situations, we carried out BSD experiments assuming that speech sources are convolved with head related transfer functions (HRTFs). The simulation results reveal that the proposed BSD method can be effective in the separation and deconvolution even with binaural-sound mixtures.

1. INTRODUCTION

Blind separation and deconvolution (BSD) of sources is an approach taken to estimate original source signals using only the information of mixed signals observed in each input channel. For the BSD based on independent component analysis (ICA), various methods have been proposed to deal with the separation and deconvolution for the convolutive mixture of independently, identically distributed (i.i.d.) source signals [2, 3]. These ICA-based BSD methods often whiten the separated signals, because they use the assumption of temporally independency of the signals. Therefore they cannot be applied to acoustic signals which is colored generally. We have proposed a novel BSD approach [1] that combines information-geometry theory and multichannel signal processing. In this approach, the BSD problem is resolved into two stages: new blind separation technique using a Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) and the deconvolution in the SIMO-model framework.

BSD for colored sources is a very hard problem. Although common room reverberation is generally regarded as an FIR filter with thousands of taps, existing BSD methods can deal with only few-tap transfer channels. Consequently, conventional BSD techniques are demonstrated only in the case of artificial transfer functions. In the previous report, we also dealt with artificial transfer functions, and only showed that the proposed BSD method can work theoretically.

In this paper, we mainly address the BSD problem with *head related transfer function* (HRTF) [4] as a more actual transfer channel. HRTF is relatively shorter among real acoustical channels (see Fig. 1). Also the left ear channel and the right ear channel are very distinct from each other. These properties are very favorable to BSD. The sound convolved with HRTF is generally called *binaural sound*, which plays the main role in human hearing, and we attempt separation and deconvolution of binaural-sound mixtures. We carry out the simulation using the HRTFs obtained from the CIPIC database [5]. Simulation results show the effectiveness of the proposed BSD method for binaural-sound mixtures.

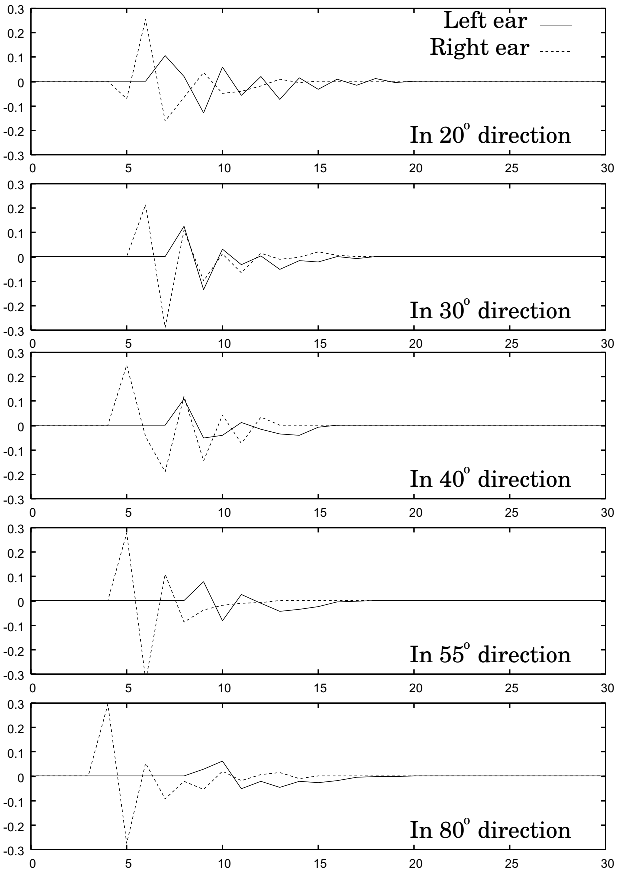

Figure 1: Examples of HRTFs used in this paper. These are down-sampled HRTFs from 44100 Hz to 8000 Hz.

2. MIXING PROCESS IN BINAURAL-SOUND MIXTURES AND CONVENTIONAL BSD

2.1 Mixing process

In this study, mixing process is assumed as the binaural-sound mixtures which are described as Fig. 2. These correspond to special cases of convolutive mixtures with two microphones (left ear and right ear) and two sound sources.

This research is (partly) supported by Core Research for Evolution Science and Technology (CREST) Program "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency (JST).

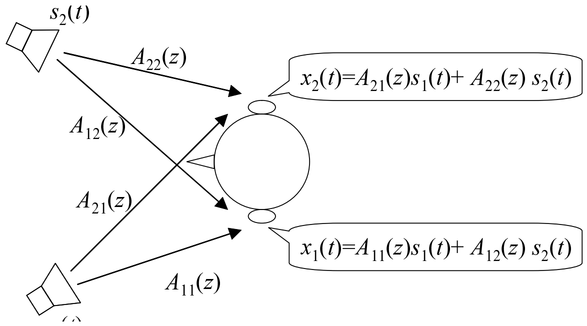
$s_2(t)$
$A_{22}(z)$
$A_{12}(z)$
$A_{21}(z)$
$A_{11}(z)$
$x_2(t)=A_{21}(z)s_1(t)+ A_{22}(z)\, s_2(t)$
$x_1(t)=A_{11}(z)s_1(t)+ A_{12}(z)\, s_2(t)$

Figure 2: Illustration of binaural-sound mixtures.

The observed signals are expressed as

$$\boldsymbol{x}(t) = \sum_{n=0}^{N-1} \boldsymbol{a}(n)\boldsymbol{s}(t-n) = \boldsymbol{A}(z)\boldsymbol{s}(t), \qquad (1)$$

where $\boldsymbol{s}(t) = [s_1(t), s_2(t)]^{\mathrm{T}}$ is the source signal vector, and $\boldsymbol{x}(t) = [x_1(t), x_2(t)]^{\mathrm{T}}$ is the observed signal vector. Also, $\boldsymbol{a}(n)$ is the mixing filter matrix with the length of N, and $\boldsymbol{A}(z)$ is the z-transform of $\boldsymbol{a}(n)$; these are given as

$$\begin{aligned} \boldsymbol{a}(n) &= [a_{kl}(n)]_{kl}, & (2)\\ \boldsymbol{A}(z) &= [A_{kl}(z)]_{kl} = \left[\sum_{n=0}^{N-1} a_{kl}(n) z^{-n}\right]_{kl}, & (3)\end{aligned}$$

where z^{-1} is used as the unit-delay operator, i.e., $z^{-n} \cdot x(t) = x(t-n)$, $a_{kl}(n)$ is the HRTF in the direction of the l-th sound source with the k-th ear (1: left ear, 2: right ear). $[X]_{ij}$ denotes the matrix which includes the element X in the i-th row and the j-th column. Since we make a free-field assumption, $a_{kl}(n)$ represents only diffraction on the head, reflection on the torso and effects of the earlobe. In general, the binaural system has the following notable features.

(a) These channels are very distinct from each other.

(b) The length of the $\boldsymbol{a}(n)$, N, is relatively short rather than that of a common room impulse response. Typical length of $\boldsymbol{a}(n)$ in 8 kHz sampling is less than 15 taps as can be seen in Fig. 1.

Owing to the attractive characteristics, we can speculate that BSD can be applied to the binaural-mixtures problem.

2.2 Conventional BSD

In the time-domain ICA (TDICA), the separated signal $\boldsymbol{y}(t) = [y_1(t), y_2(t)]^{\mathrm{T}}$ is expressed as

$$\boldsymbol{y}(t) = \sum_{n=0}^{D-1} \boldsymbol{w}(n)\boldsymbol{x}(t-n), \qquad (4)$$

where $\boldsymbol{w}(n)$ is the separation filter matrix, and D is the filter length of $\boldsymbol{w}(n)$. In the ICA-based BSD, Amari [2] proposed the holonomic TDICA algorithm which optimizes the separation filter by minimizing the Kullback-Leibler divergence between the joint probability density function (PDF) of $\boldsymbol{y}(t)$ and the product of marginal PDFs of $y_l(t)$. The iterative learning rule is given by

$$\begin{aligned} &\boldsymbol{w}^{[j+1]}(n)\\ &= \boldsymbol{w}^{[j]}(n) + \eta \sum_{d=0}^{D-1} \Big\{ \boldsymbol{I}\delta(n-d)\\ &\quad -\Big\langle \boldsymbol{\varphi}(\boldsymbol{y}^{[j]}(t))\boldsymbol{y}^{[j]}(t-n+d)^{\mathrm{T}} \Big\rangle_t \Big\} \cdot \boldsymbol{w}^{[j]}(d), \qquad (5)\end{aligned}$$

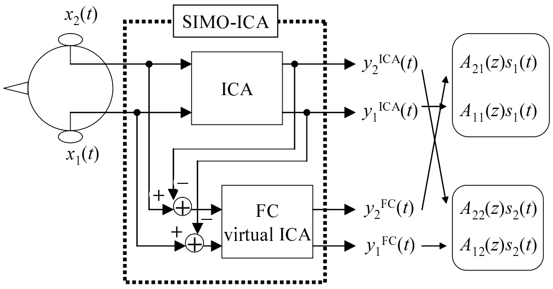

Figure 3: Example of input and output relations in SIMO-ICA used in binaural separation.

where η is the step-size parameter, the superscript $[j]$ is used to express the value of the j-th step in the iterations, $\langle\cdot\rangle_t$ denotes the time-averaging operator, and $\boldsymbol{I}$ is the identity matrix. $\delta(n)$ is a delta function, where $\delta(0) = 1$ and $\delta(n) = 0$ $(n \neq 0)$. $\boldsymbol{\varphi}(\cdot)$ is the nonlinear vector function. This BSD based on ICA, however, might whiten the separate signals. Therefore it cannot be applied to colored sources although most of audio signals are colored.

3. TWO-STAGE BSD FOR BINAURAL-SOUND MIXTURES

In this section, we explain our proposed two-stage BSD algorithm [1] combining SIMO-ICA and blind multichannel inverse filtering, which is specified for binaural-sound BSD. In the proposed method, the separation and deconvolution procedures are performed under the following assumptions.

(A1) The source signals, $s_1(t)$ and $s_2(t)$, are mutually independent, and unknown.

(A2) Each source signal is temporally correlated (colored), i.e.,

$$\frac{\big\langle s_l(t)s_l(t-n)\big\rangle_t}{\big\langle s_l(t)^2\big\rangle_t} \neq \delta(n) \quad (l = 1, 2), \qquad (6)$$

but its coloration characteristics are unknown.

(A3) The mixing system $\boldsymbol{A}(z)$ is unknown, and probably has the nonminimum phase property. However, every column of $\boldsymbol{A}(z)$ is guaranteed not to have any common zeros in the z-plane.

(A4) The order of the mixing system, N, is unknown.

These are reasonable assumptions in binaural-sound mixtures driven by audio signals. Details of the process using the proposed algorithm are as follows.

3.1 First stage: SIMO-ICA for source separation

In this stage, a blind separation method using SIMO-ICA in binaural system is conducted. In the binaural system (see Fig. 3) SIMO-ICA consists of an ICA part and a *fidelity controller*, and the ICA runs under fidelity control of the entire system. The separated signals of the ICA in SIMO-ICA are defined by

$$\boldsymbol{y}_{\mathrm{ICA}}(t) = \begin{bmatrix} y_1^{\mathrm{ICA}}(t) \\ y_2^{\mathrm{ICA}}(t) \end{bmatrix} = \sum_{n=0}^{D-1} \boldsymbol{w}_{\mathrm{ICA}}(n)\boldsymbol{x}(t-n), \qquad (7)$$

where $\boldsymbol{w}_{\mathrm{ICA}}(n)$ is the separation filter matrix in the ICA. Regarding the fidelity controller, we calculate the following signal vector, in which the all elements are to be mutually independent,

$$\boldsymbol{y}_{\mathrm{FC}}(t) = \boldsymbol{x}(t - D/2) - \boldsymbol{y}_{\mathrm{ICA}}(t). \qquad (8)$$

Hereafter, we regard $\boldsymbol{y}_{\mathrm{FC}}(t)$ as an output of a *virtual* ICA, and define its virtual separation filter matrix as

$$\boldsymbol{w}_{\mathrm{FC}}(n) = \boldsymbol{I}\delta(n - \frac{D}{2}) - \boldsymbol{w}_{\mathrm{ICA}}(n). \qquad (9)$$

From (9) we can rewrite (8) as

$$\boldsymbol{y}_{\mathrm{FC}}(t) \quad = \quad \sum_{n=0}^{D-1} \boldsymbol{w}_{\mathrm{FC}}(n) \cdot \boldsymbol{x}(t-n). \qquad (10)$$

The reason why we use the word "virtual" here is that fidelity controller does not have own separation filters unlike the ICA.

In order to make $\boldsymbol{y}_{\mathrm{ICA}}(t)$ independent and simultaneously $\boldsymbol{y}_{\mathrm{FC}}(t)$ independent, the natural gradient [2] of KLD of (10) with respect to $\boldsymbol{w}_{\mathrm{ICA}}(n)$ should be added to the iterative learning rule of the separation filter in ICA. The new iterative learning rule of the ICA in SIMO-ICA is given as

$$\begin{aligned}
&\boldsymbol{w}_{\mathrm{ICA}}^{[j+1]}(n) \\
&= \boldsymbol{w}_{\mathrm{ICA}}^{[j]}(n) - \alpha \sum_{d=0}^{D-1} \Bigg[\Bigg\{ \text{off-diag} \Big\langle \boldsymbol{\varphi}(\boldsymbol{y}_{\mathrm{ICA}}^{[j]}(t)) \\
&\quad \boldsymbol{y}_{\mathrm{ICA}}^{[j]}(t-n+d)^{\mathrm{T}} \Big\rangle_t \Bigg\} \cdot \boldsymbol{w}_{\mathrm{ICA}}^{[j]}(d) \\
&\quad - \Bigg\{ \text{off-diag} \Big\langle \boldsymbol{\varphi}(\boldsymbol{x}(t - \frac{D}{2}) - \boldsymbol{y}_{\mathrm{ICA}}^{[j]}(t)) \\
&\quad \cdot (\boldsymbol{x}(t-n+d-\frac{D}{2}) - \boldsymbol{y}_{\mathrm{ICA}}^{[j]}(t-n+d)^{\mathrm{T}}) \Big\rangle_t \Bigg\} \\
&\quad \cdot (\boldsymbol{I}\delta(d - \frac{D}{2}) - \boldsymbol{w}_{\mathrm{ICA}}^{[j]}(d)) \Bigg], \qquad (11)
\end{aligned}$$

where α is a step-size parameter. Under (11) the separated signals converge on the following solutions;

$$\begin{bmatrix} y_1^{\mathrm{ICA}}(t) \\ y_2^{\mathrm{ICA}}(t) \end{bmatrix} = \begin{bmatrix} A_{11}(z)s_1(t-D/2) \\ A_{22}(z)s_2(t-D/2) \end{bmatrix}, \qquad (12)$$

$$\begin{bmatrix} y_1^{\mathrm{FC}}(t) \\ y_2^{\mathrm{FC}}(t) \end{bmatrix} = \begin{bmatrix} A_{12}(z)s_2(t-D/2) \\ A_{21}(z)s_1(t-D/2) \end{bmatrix}, \qquad (13)$$

or

$$\begin{bmatrix} y_1^{\mathrm{ICA}}(t) \\ y_2^{\mathrm{ICA}}(t) \end{bmatrix} = \begin{bmatrix} A_{12}(z)s_2(t-D/2) \\ A_{21}(z)s_1(t-D/2) \end{bmatrix}, \qquad (14)$$

$$\begin{bmatrix} y_1^{\mathrm{FC}})(t) \\ y_2^{\mathrm{FC}}(t) \end{bmatrix} = \begin{bmatrix} A_{11}(z)s_1(t-D/2) \\ A_{22}(z)s_2(t-D/2) \end{bmatrix}. \qquad (15)$$

The proof of theorem and more details are given in [6].

3.2 Second stage: blind multichannel inverse filtering for deconvolution

In this stage consider the blind channel identification corresponding to the first sound source $s_1(t)$. In this process, the HRTFs, $A_{11}(z)$ and $A_{21}(z)$, can be estimated by a subchannel matching approach [7, 8, 9] in an SIMO framework because we have already resolved the mixing process of the sources into a simple SIMO model through SIMO-ICA in the previous stage. The subchannel matching approach can work even for the temporally correlated signal. Regarding the blind channel identification corresponding to another sound source $s_2(t)$, we can estimate $A_{12}(z)$ and $A_{22}(z)$ using the same approach.

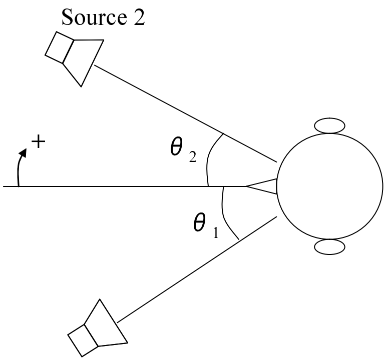

Figure 4: Locations of sources and dummy head in the simulation. These systems are called Lθ_1Rθ_2.

We can estimate the multichannel inverse filters, $G_{11}(z)$ and $G_{21}(z)$ for $\hat{A}_{11}(z)$ and $\hat{A}_{21}(z)$, and $G_{12}(z)$ and $G_{22}(z)$ for $\hat{A}_{12}(z)$ and $\hat{A}_{22}(z)$, based on the multiple-input/output inverse theorem (MINT) [10]. In the MINT method, the exact inverse of the transfer functions can be uniquely determined, even when $\hat{A}_{kl}(z)$ has the nonminimum phase properties, if $\hat{A}_{kl}(z)$ does not have any common zeros in the z-plane. For example, the recovered signals $\hat{s}_l(t)$ under Fig. 3 are given as

$$\begin{aligned}
\hat{s}_1(t) &= G_{11}(z)y_1^{(1)}(t) + G_{21}(z)y_2^{(2)}(t), \qquad (16) \\
\hat{s}_2(t) &= G_{12}(z)y_1^{(2)}(t) + G_{22}(z)y_2^{(1)}(t). \qquad (17)
\end{aligned}$$

The accurate estimation of the filter length N of the impulse responses is indispensable for improving the system identification performance. There are various methods for filter-length estimation and we use the Furuya's method [9] in this work.

4. SIMULATIONS

4.1 Conditions for experiment

The mixing filter matrix $\boldsymbol{A}(x)$ is taken to be the HRTFs which were measured by CIPIC. The CIPIC HRTF database [5] was measured with KEMAR dummy head in an anechoic room. It is a public domain HRTF database with the high spatial resolution. We chose six azimuths from the database and down-sampled them from 44100 Hz to 8000 Hz (shown in Fig. 1). The locations of sound sources and the dummy head are set as shown in Fig. 4. Elevation is set to 0 degree. θ_1 is fixed to -30 degrees, and θ_2 is varied from 20 to 80 (20, 30, 40, 55, and 80) degrees. These systems are called Lθ_1Rθ_2, e.g. "L$-$30R20". Two sentences spoken by two male speakers are used as the original speech samples $\boldsymbol{s}(t)$. The sampling frequency is 8 kHz and the length of speech is limited to 30 seconds.

We compare two methods as follows: conventional **holonomic ICA** (ICA-based BSD) [2] given by (5), and **proposed two-stage BSD**. The step-size parameter η is 1×10^{-6} in the holonomic ICA and α is 1×10^{-6} in SIMO-ICA; these are optima which provide the best performance. The length of the separation filter is set to be 512 taps.

In the experiment, two objective evaluation scores are defined as follows. First, *noise reduction rate* (NRR) [11], defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB, is used as the objective indication of separation performance, where we do not take into account the distortion of the separated signal. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise. Secondly, *mel cepstral distortion* (melCD) is used as the indication of deconvolution performance. In this study, we defined the melCD

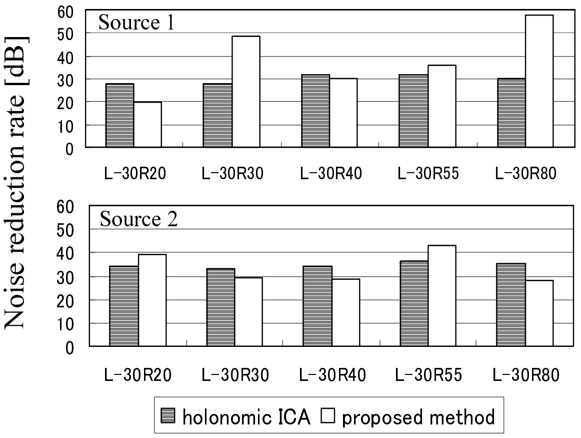

Figure 5: Simulation results of noise reduction rate.

as the distance between the spectral envelope of the original source signal $s_l(t - D/2)$ and that of the separated output. The 16th-order mel-scaled cepstrum based on the smoothed FFT spectrum is used. The melCD will be decreased to zero if the separation-deconvolution processing is performed perfectly.

4.2 Results and discussion

Figures 5 and 6 show the results of NRR and melCD for different methods. From the results of NRR, the separation performance of the holonomic ICA is comparable to those of the proposed method and almost all of them are over 20 dB. Accordingly the holonomic ICA and the proposed method are both effective as far as the only separation performance is concerned. As for the distortion of the separated speech, which is an important issue from the practical viewpoint, there is a considerable difference between two methods. As can be seen in Fig. 6, it is evident that the melCD of the holonomic ICA is obviously high, i.e., the resultant speech is whitened by the decorrelation in the conventional method. On the other hand, the melCD of the proposed method are around 3 dB except for L−30R20 and L−30R40. Since the melCD of the each binaural sound was around 4 dB, it can be asserted that SIMO deconvolution part in the second stage of the proposed BSD works effectively. These results indicate that the proposed BSD has the possibility to achieve the separation and deconvolution for binaural-sound mixtures.

5. CONCLUSION

In order to evaluate our proposed method in more actual situations, we carried out BSD experiments assuming that speech sources are convolved with HRTFs. The simulation results reveal that the proposed two stage BSD method [1] can achieve the sufficient separation performance as much as holonomic ICA and can recover the source signals from binaural-sound mixtures.

6. ACKNOWLEDGEMENT

All of authors thank to Center for Image Processing and Integrated Computing Interface lab. in University of California Davis. Their HRTF database is very useful for our research.

REFERENCES

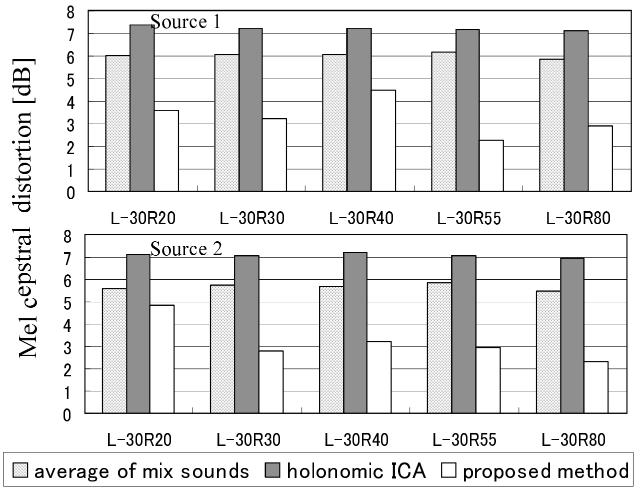

Figure 6: Simulation results of mel cepstral distortion.

[1] H. Saruwatari, H. Yamajo, T. Takatani, T. Nishikawa, K. Shikano, "Blind separation and deconvolution of MIMO-FIR system with colored sound inputs using SIMO-model-based ICA", *Proc. IEEE Workshop on Statistical Signal Processing,* pp.421–424, Sept. 2003.

[2] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient", *Proc. IEEE Int. Workshop on Wireless Communication,* pp.101–104, April 1997.

[3] S. Haykin (ed.), *Unsupervised Adaptive Filtering*, John Wiley & Sons, Ltd., New York, 2000.

[4] J. Blauert, *Spatial Hearing (revised ed.)*, Cambridge, MA: The MIT Press, 1997.

[5] CIPIC HRTF Database Files, Release 1.1, August,22,2001, available at `http://interface.cipic.ucdavis.edu/CIL_html/`

[6] T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning", *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, pp.251–254, Sept. 2003.

[7] H. Xu and L. Tong, "A deterministic approach to blind identification of multi-channel FIR systems," *Proc. ICASSP94*, pp.581–584, 1994.

[8] Z. Ding and Y. Li, *Blind Equalization and Identification*, Marcel Dekker, Inc., New York, 2001.

[9] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution of nonminimum phase FIR system," *IEICE Trans. Fundamentals*, vol. E80-A, no. 5, pp.804–808, 1997.

[10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp.145–152, Feb. 1988.

[11] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp.1135–1146.