SOLVING THE PERMUTATION PROBLEM OF FREQUENCY-DOMAIN BSS WHEN SPATIAL ALIASING OCCURS WITH WIDE SENSOR SPACING

Hiroshi Sawada *Shoko Araki* *Ryo Mukai* *Shoji Makino*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper describes a method for solving the permutation problem of frequency-domain blind source separation (BSS). The method analyzes the mixing system information estimated with independent component analysis (ICA). When we use widely spaced sensors or increase the sampling rate, spatial aliasing may occur for high frequencies due to the possibility of multiple cycles in the sensor spacing. In such cases, the estimated information would imply multiple possibilities for a source location. This causes some difficulty when analyzing the information. We propose a new method designed to overcome this difficulty. This method first estimates the model parameters for the mixing system at low frequencies where spatial aliasing does not occur, and then refines the estimations by using data at all frequencies. This refinement leads to precise parameter estimation and therefore precise permutation alignment. Experimental results show the effectiveness of the new method.

1. INTRODUCTION

The technique for estimating individual source components from their mixtures at multiple sensors is known as blind source separation (BSS) [1]. With acoustical applications of BSS, such as solving a cocktail party problem, signals are generally mixed in a convolutive manner with reverberations. Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be sensor observations. The convolutive mixture model is formulated as

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_{l} \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l), \quad j=1, \ldots, M, \qquad (1)$$

where t represents time and $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. In a practical room situation, impulse responses $\mathrm{h}_{jk}(l)$ can have thousands of taps even with an 8 kHz sampling rate. This makes the convolutive BSS problem very difficult compared with the BSS of simple instantaneous mixtures.

An efficient and practical approach for such convolutive mixtures is frequency-domain BSS [2–9], where we apply a short-time Fourier transform (STFT) to the sensor observations $\mathrm{x}_j(t)$. In the frequency domain, the convolutive model (1) can be approximated as an instantaneous mixture model at each frequency:

$$x_j(f,t) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,t), \quad j=1, \ldots, M, \qquad (2)$$

where f represents frequency, t is now down-sampled with the distance of the frame shift, $h_{jk}(f)$ is the frequency response from source k to sensor j, and $s_k(f,t)$ is a frequency-domain representation of a source signal $\mathrm{s}_k(t)$.

Independent component analysis (ICA) [10] is a major statistical tool for BSS. In the frequency-domain approach, ICA is employed in each frequency bin with the instantaneous mixture model (2). This makes the convergence of ICA relatively fast compared with convolutive ICA where the mixture model (1) is explicitly assumed. However, the ICA solution has permutation ambiguity. Even if we change the order of the separated signals at the ICA output, it is still an ICA solution. This causes the permutation problem of frequency-domain BSS. We need to align the permutation of each frequency bin so that the frequency components of the same source are grouped together.

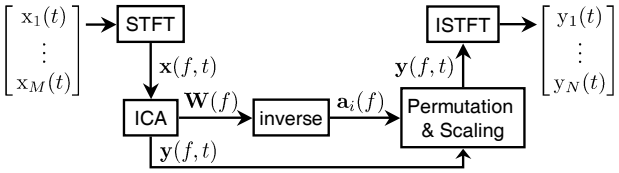

Fig. 1. Flow of frequency-domain BSS

Various methods have been proposed for solving the permutation problem: 1) making the separation matrices smooth [2, 3], 2) maximizing the correlation of separated signal envelopes [4], 3) analyzing the directivity patterns calculated from the separation matrices [5, 6] and 4) analyzing the mixing system information estimated with ICA [7, 8]. The third and fourth methods utilize similar kinds of information, such as estimated directions of sources. However, the fourth method is more general than the third one, as it can easily be applied to a situation where there are more than two sources [7, 8]. We have experimentally shown that the fourth (or the third) method provides a robust solution for the permutation problem [9]. When we employ such a method, we prefer the sensor spacings to be no larger than half the minimum wavelength of interest. We typically use a 4 cm sensor spacing for an 8 kHz sampling rate to satisfy this condition. If sensor spacing is wider, spatial aliasing might occur at high frequencies [11], and the ICA solution in such a frequency bin implies multiple possibilities for a source location.

In this paper, we propose a new method for dealing with a situation where the distance between sensors is larger than half the wavelength. Although a wider sensor spacing is discussed in [6], it is based on the third method, which is hard to generalize for more than two sources. With the new proposed method working effectively, we can use widely spaced sensors to achieve better separation for low frequencies or we can increase a sampling rate, for example up to 16 kHz, to obtain better speech recognition accuracy for separated signals.

2. FREQUENCY-DOMAIN BSS

This section presents an overview of frequency-domain BSS. Figure 1 shows the flow that we consider in this paper. First, the sen-

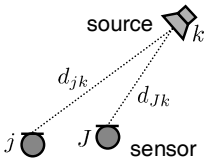

Fig. 2. Direct-path model

sor observations (1) are converted into frequency-domain time-series signals (2) by an STFT. Let us rewrite (2) in a vector notation:

$$\mathbf{x}(f,t) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t), \tag{3}$$

where $\mathbf{x} = [x_1, \ldots, x_M]^T$ is the vector of observed signals and $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ is the vector of frequency responses from source s_k to all sensors.

Then, complex-valued instantaneous ICA [10] is applied to the mixtures $\mathbf{x}(f,t)$ to obtain separated frequency components:

$$\mathbf{y}(f,t) = \mathbf{W}(f)\,\mathbf{x}(f,t), \tag{4}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ is the vector of separated frequency components and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]^H$ is an $N \times M$ separation matrix.

With the method considered here, the inverse of separation matrix $\mathbf{W}$ (or the Moore-Penrose pseudoinverse if $\mathbf{W}$ is not square) is calculated in each frequency bin. The inverse is represented as

$$[\mathbf{a}_1, \cdots, \mathbf{a}_N] = \mathbf{W}^{-1},\ \mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T, \tag{5}$$

which we call basis vectors obtained by ICA, because the mixture $\mathbf{x}(f,t)$ is represented with a linear combination of basis vectors by multiplying $\mathbf{W}^{-1}$ and (4):

$$\mathbf{x}(f,t) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(f,t). \tag{6}$$

By comparing (6) and (3), we see that a basis vector $\mathbf{a}_i(f)$ represents the same information as $\mathbf{h}_k(f)$ up to permutation and scaling ambiguity if ICA works well. The use of different subscripts, i and k, indicates the permutation ambiguity.

Next, the permutation ambiguity is aligned. This paper focuses on the fourth method discussed in the introduction, which analyzes the mixing system information represented by the basis vectors $\mathbf{a}_i(f)$. Section 3 will discuss the way to analyze the information and then align the permutation.

Then, the scaling ambiguity of ICA is aligned by

$$y_i(f,t) \leftarrow a_{Ji}(f) y_i(f,t), \tag{7}$$

where J is the index of a reference sensor (see [8] for the rationale of this operation). Finally, time-domain output signals $\mathrm{y}_i(t)$ are obtained from separated frequency components $y_i(f,t)$ by an inverse STFT (ISTFT).

3. PERMUTATION ALIGNMENT

In this section, we propose a new method for solving the permutation problem. It is based on an analysis of the mixing system information estimated by ICA and represented in basis vectors (5). The method can handle a situation where spatial aliasing occurs because of a wide sensor spacing or a high sampling rate.

Table 1. Experimental setup A

Source directions	70° and 150° (2 sources)
Sensor spacing	$d_{\max} = 20$ cm (2 sensors)
Source distance from sensors	120 cm
Sampling rate	16 kHz
Reverberation time	RT_{60} = 130 ms
Frame size of STFT	128 ms
Source signal	speech of 3 s
Propagation velocity	$c = 340$ m/s

3.1. Assumption and basic idea

We assume a simple direct-path model (Fig. 2) for the mixing system, even though in reality signals are mixed in a multi-path model (1). This simplified model is expressed in the frequency domain:

$$h_{jk}(f) = \lambda_{jk} \cdot e^{-\jmath\, 2\pi f \tau_{jk}}, \tag{8}$$

where τ_{jk} and $\lambda_{jk} \geq 0$ are the time delay and attenuation from source k to sensor j, respectively. Since we cannot distinguish the phase (or amplitude) of $s_k(f,t)$ and $h_{jk}(f)$, these two parameters can be considered to be relative. Thus, without loss of generality, we normalize them by

$$\tau_{jk} = (d_{jk} - d_{Jk})/c, \tag{9}$$

$$\textstyle\sum_{j=1}^{M} \lambda_{jk}^2 = 1, \tag{10}$$

where d_{jk} is the distance from source k to sensor j (Fig. 2), and c is the propagation velocity of the signal. Normalization (9) makes $\tau_{Jk} = 0$, i.e. the relative time delay is zero at the reference sensor J.

As explained in Sec. 2, a basis vector $\mathbf{a}_i(f)$ represents the same information as $\mathbf{h}_k(f)$ up to permutation and scaling ambiguity. Following (9) and (10), the scaling ambiguity is aligned by the operation

$$\mathbf{a}_i \leftarrow \frac{\mathbf{a}_i}{||\mathbf{a}_i||} e^{-\jmath \arg(a_{Ji})} \tag{11}$$

which makes $\arg(a_{Ji}) = 0$ and $||\mathbf{a}_i|| = 1$. Now, the task for the permutation problem is to determine a permutation Π_f that relates the subscript i and k with $i = \Pi_f(k)$, and to estimate parameters τ_{jk}, λ_{jk} that make the model (8) match the $a_{ji}(f)$ element of the basis vector. This can be formulated so as to find Π_f, τ_{jk} and λ_{jk} that minimize the cost function:

$$\mathcal{J} = \sum_{f \in \mathcal{F}} \sum_{k=1}^{N} \sum_{j=1}^{M} |a_{ji}(f) - \lambda_{jk} \cdot e^{-\jmath\, 2\pi f \tau_{jk}}|^2,\ \ i = \Pi_f(k), \tag{12}$$

where $\mathcal{F}$ is the set of frequencies that we have to consider.

To make the discussion here intuitively understandable, we performed an experiment with setup A shown in Table 1. This was a simple $M = N = 2$ case, but the sensor spacing was 20 cm, which induced spatial aliasing for a 16 kHz sampling rate. Figure 3 shows the argument of a_{21} and a_{22} after the normalization (11) where we set $J = 1$ as a reference sensor. The arguments of $a_{1i}(f)$ are not shown because they are all zero. The relative time delay τ_{21} and τ_{22} can be estimated from these data. However, we see some circular jumps at high frequencies, which are caused by spatial aliasing. They complicate the estimation of τ_{21} and τ_{22}.

3.2. For frequencies without spatial aliasing

Let us first consider the lower frequency range

$$\mathcal{F}_L = \{f :\ -\pi < 2\pi f \tau_{jk} < \pi, {}^{\forall} j, k\} \tag{13}$$

where we can guarantee that spatial aliasing does not occur. Let $d_{\max}$ be the maximum distance between the reference sensor J and

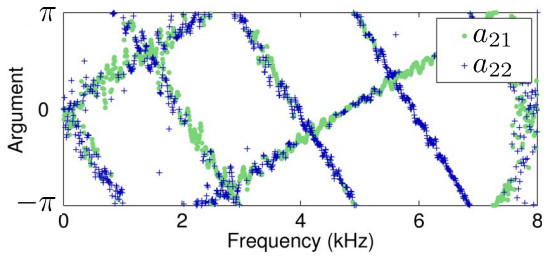

Fig. 3. Arguments of a_{21} and a_{22} before permutation alignment

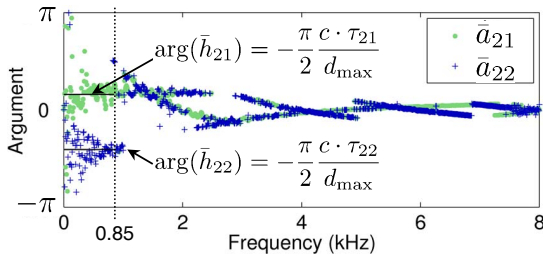

Fig. 4. Arguments of $\bar{a}_{21}$ and $\bar{a}_{22}$ after permutations are aligned only for frequency range $\mathcal{F}_L = \{f : 0 < f < 850 \text{ Hz}\}$

any other sensor. Then the relative time delay is bounded by

$$\max_{jk} |\tau_{jk}| \leq d_{\max}/c \tag{14}$$

and therefore $\mathcal{F}_L$ can be defined as

$$\mathcal{F}_L = \{f : 0 < f < \frac{c}{2d_{\max}}\}. \tag{15}$$

For frequency range $\mathcal{F}_L$, appropriate Π_f for (12) can be obtained by the method shown in our previous work [8], which further normalizes the basis vectors (11) to remove frequency dependence:

$$\bar{a}_{ji}(f) \leftarrow |a_{ji}(f)| \exp\left[\jmath \frac{\arg[a_{ji}(f)]}{4fc^{-1}d_{\max}}\right]. \tag{16}$$

The rationale of dividing the argument by $4fc^{-1}d_{\max}$ is discussed in [8]. With this operation, the cost function (12) is converted into

$$\bar{\mathcal{J}} = \sum_{f\in\mathcal{F}_L}\sum_{k=1}^{N}\sum_{j=1}^{M} |\bar{a}_{ji}(f) - \bar{h}_{jk}|^2, \quad i = \Pi_f(k) \tag{17}$$

where

$$\bar{h}_{jk} = \lambda_{jk} \cdot \exp[-\jmath \frac{\pi}{2}\frac{c\cdot\tau_{jk}}{d_{\max}}] \tag{18}$$

is a frequency-normalized model. In a vector notation,

$$\bar{\mathcal{J}} = \sum_{f\in\mathcal{F}_L}\sum_{k=1}^{N} ||\bar{\mathbf{a}}_i(f) - \bar{\mathbf{h}}_k||^2, \quad i = \Pi_f(k), \tag{19}$$

where $\bar{\mathbf{a}}_i = [\bar{a}_{1i}, \ldots, \bar{a}_{Mi}]^T$ and $\bar{\mathbf{h}}_k = [\bar{h}_{1k}, \ldots, \bar{h}_{Mk}]^T$. Here we see that frequency-normalized basis vectors $\bar{\mathbf{a}}_i(f)$ and the frequency-normalized model $\bar{\mathbf{h}}_k$ do not depend on frequency. Therefore, $\bar{\mathcal{J}}$ can be minimized efficiently by a clustering algorithm that iterates the following two updates until convergence:

$$\Pi_f \leftarrow \text{argmin}_\Pi \sum_{k=1}^{N} ||\bar{\mathbf{a}}_{\Pi(k)}(f) - \bar{\mathbf{h}}_k||^2, \tag{20}$$

$$\bar{\mathbf{h}}_k \leftarrow \sum_{f\in\mathcal{F}_L} \bar{\mathbf{a}}_{\Pi_f(k)}(f), \quad \bar{\mathbf{h}}_k \leftarrow \bar{\mathbf{h}}_k/||\bar{\mathbf{h}}_k||. \tag{21}$$

The first update (20) optimizes the permutation Π_f for each frequency with the current model $\bar{\mathbf{h}}_k$. The second update (21) calculates the most probable model $\bar{\mathbf{h}}_k$ with the current permutations. This set of updates is very similar to that of the k-means algorithm [12]. After the algorithm has converged, we update the subscript of the basis vectors by

$$\mathbf{a}_k(f) \leftarrow \mathbf{a}_{\Pi_f(k)}(f), \quad k = 1, \ldots, N. \tag{22}$$

Figure 4 shows the arguments of $\bar{a}_{21}$ and $\bar{a}_{22}$ calculated by operation (16) in the setup A experiment. For frequency range $\mathcal{F}_L$, the permutations Π_f were aligned and updated by (22). We see two clusters whose centroids are the two lines represented by $\arg(\bar{h}_{21})$ and $\arg(\bar{h}_{22})$. For frequencies higher than 850 Hz, we see the effect

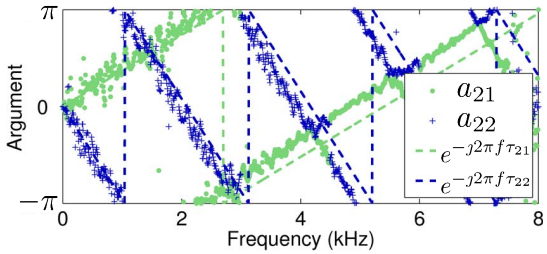

Fig. 5. Arguments of a_{21} and a_{22} after permutation alignment using model parameters estimated with data in the low frequency range $\mathcal{F}_L$. Because τ_{21} and τ_{22} are not precisely estimated, there are some permutation errors at high frequencies.

of spatial aliasing. This means that the frequency normalization (16) does not work for these higher frequencies, and therefore we need a new method to minimize the cost function (12).

3.3. For frequencies where spatial aliasing may occur

This subsection proposes a new method for deciding the permutation Π_f for frequencies where spatial aliasing may occur. The model parameters τ_{jk}, λ_{jk} can be extracted from the frequency-normalized model $\bar{h}_{jk}$ (18) calculated by (21):

$$\tau_{jk} = -\frac{2}{\pi}\frac{d_{\max}}{c}\arg(\bar{h}_{jk}), \quad \lambda_{jk} = |\bar{h}_{jk}|. \tag{23}$$

Thus, a simple way is to use these extracted parameters for the cost function (12), and decide the permutations Π_f by

$$\Pi_f = \text{argmin}_\Pi \sum_{k=1}^{N}\sum_{j=1}^{M} |a_{j\Pi(k)}(f) - \lambda_{jk}\cdot e^{-\jmath 2\pi f\tau_{jk}}|^2. \tag{24}$$

However, τ_{jk} and λ_{jk} estimated only with frequencies in $\mathcal{F}_L$ may not be very accurate. Figure 5 shows $\arg(a_{21})$ and $\arg(a_{22})$ after permutations were calculated by (24) using model parameters extracted by (23). We see some estimation error for τ_{21} and τ_{22}.

A better way is to refine the model parameters τ_{jk} and λ_{jk} by the gradient of the cost function $\mathcal{J}$ (12):

$$\tau_{jk} \leftarrow \tau_{jk} - \mu\frac{\partial\mathcal{J}}{\partial\tau_{jk}}, \quad \lambda_{jk} \leftarrow \lambda_{jk} - \mu\frac{\partial\mathcal{J}}{\partial\lambda_{jk}}, \tag{25}$$

where μ is a step size parameter, and

$$\frac{\partial\mathcal{J}}{\partial\tau_{jk}} \propto \lambda_{jk}\sum_{f\in\mathcal{F}} f\cdot\text{imag}[a_{j\Pi(k)}(f)\cdot e^{\jmath 2\pi f\tau_{jk}}], \tag{26}$$

$$\frac{\partial\mathcal{J}}{\partial\lambda_{jk}} \propto \sum_{f\in\mathcal{F}}\left\{\lambda_{jk} - \text{real}[a_{j\Pi(k)}(f)\cdot e^{\jmath 2\pi f\tau_{jk}}]\right\}, \tag{27}$$

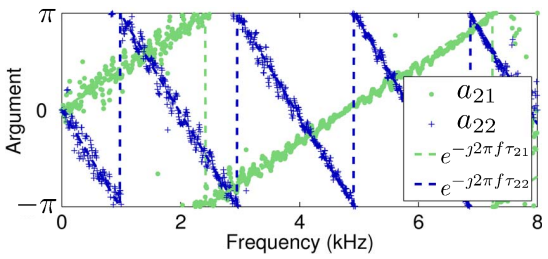

Fig. 6. Argument of a_{21} and a_{22} after permutation alignment using model parameters refined with data at all frequencies. Now τ_{21} and τ_{22} are precisely estimated, and permutations are aligned correctly.

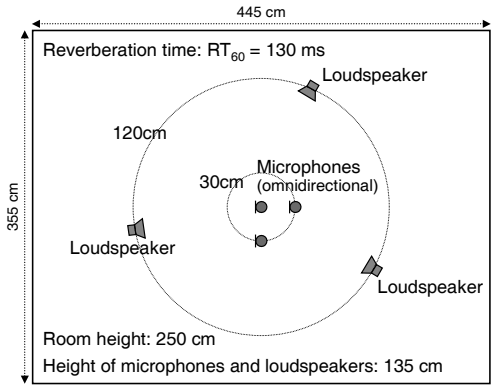

Fig. 7. Experimental setup B

are the partial derivatives of $\mathcal{J}$ with respect to τ_{jk} and λ_{jk}. The operations imag[·] and real[·] extract the imaginary and real parts of a complex number, respectively. We can iteratively update Π_f by (24) and $(\tau_{jk}, \lambda_{jk})$ by (25) to obtain better estimations of the model and consequently better permutations. Note that the structure that iterates (24) and (25) has the same structure as (20) and (21). Figure 6 shows $\arg(a_{21})$ and $\arg(a_{22})$ after τ_{jk} and Π_f were refined by (24) and (25). We see that τ_{21} and τ_{22} were precisely estimated and the permutations were aligned correctly even for high frequencies.

4. EXPERIMENTS

To see the effectiveness of the new method, we conducted experiments to separate 3-second English speeches blindly. We used two setups. Setup A (Table 1) was used to illustrate the process of the method with Figs. 3-6. Setup B (Fig. 7) was used to examine the validity of the method for a more complicated case. One attractive feature of the method shown in Sec. 3 is that the system does not have to know the sensor geometry, but simply the maximum distance $d_{\max}$ from a reference sensor J to the other sensors. In setup B, we assigned the center microphone as the reference sensor, and provided the system with the information $d_{\max} = 30$ cm.

Table 2 shows the BSS results measured with the average signal-to-interference ratio (SIR) for 16 combinations of speeches. The SIR was calculated as the ratio of the power of a target component and interference components [9]. In method I, permutations were aligned only for a low frequency range $\mathcal{F}_L$ (corresponding to Fig. 4). In method II, permutations were aligned for all frequency bins, but with model parameters τ_{jk} and λ_{jk} estimated with data in the low frequency range $\mathcal{F}_L$ (corresponding to Fig. 5). In method III, permutations were aligned for all frequency bins with τ_{jk} and λ_{jk} refined with data at all frequencies (corresponding to Fig. 6). We see that method III provides superior results to the other methods. Since separated frequency components generated by ICA were the same for all methods, the differences among these average SIRs were due solely to the preciseness of the permutation alignment.

Table 2. Comparison of average signal-to-interference ratio (SIR) obtained with three different methods for permutation alignment.

	Input SIR	Method I	Method II	Method III
Setup A	0.0 dB	11.5 dB	15.4 dB	17.2 dB
Setup B	−3.1 dB	9.0 dB	11.2 dB	13.8 dB

5. CONCLUSION

We have proposed a new method for the permutation problem of frequency-domain BSS, which analyzes mixing system information estimated with ICA. The method can handle a situation where the sensor spacing is larger than half the minimum wavelength and thus spatial aliasing occurs. Experimental results clearly show the effectiveness of the proposed method.

6. REFERENCES

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

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

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

[4] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct. 2001.

[5] 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, Nov. 2003.

[6] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 1–13, Jan. 2005.

[7] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation for many speech signals," in *Proc. ICA 2004 (LNCS 3195)*, pp. 461–469. Springer-Verlag, Sept. 2004.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," in *Proc. ICASSP 2005*, Mar. 2005, vol. III, pp. 61–64.

[9] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[11] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice-Hall, 1993.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.