BLIND EXTRACTION OF A DOMINANT SOURCE SIGNAL FROM MIXTURES OF MANY SOURCES

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
{sawada,shoko,ryo,maki}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper presents a method for enhancing a dominant target source that is close to sensors, and suppressing other interferences. The enhancement is performed blindly, i.e. without knowing the number of total sources or information about each source, such as position and active time. We consider a general case where the number of sources is larger than the number of sensors. We employ a two-stage processing technique where a spatial filter is first employed in each frequency bin and time-frequency masking is then used to improve the performance further. To obtain the spatial filter we employ independent component analysis and then select the component of the target source. Time-frequency masks in the second stage are obtained by calculating the angle between the basis vector corresponding to the target source and a sample vector. The experimental results for a simulated cocktail party situation were very encouraging.

1. INTRODUCTION

The technique for estimating individual source components from their mixtures at sensors is known as blind source separation (BSS) [1]. In some applications such as brain imaging or wireless communications, it makes sense to extract as many source components as possible, because many sources are equally important. However, in audio applications such as speech enhancement, the significance of each source is not necessarily equal. We often want to extract only one source that is close to sensors, has a dominant power, and/or has interesting features.

This paper presents a method for extracting a source signal of interest and suppressing other interferences blindly. Let us formulate the task. Suppose that a target source s_1 and other interference sources $\mathrm{s}_k,\ k=2,\dots,N$ are convolutively mixed and observed at M sensors

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

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. The goal is to have an output signal

$$\mathrm{y}_1(t) = \sum_l \mathrm{h}_{J1}(l)\, \mathrm{s}_1(t-l) \quad (2)$$

which is the component of s_1 measured at sensor J. The task should be performed only with the M observed signals. The number of sources N is unknown and may be larger than M.

The first problem is how to extract a target source s_1 blindly. Even if N could be larger than M, independent component analysis (ICA) [2] with an $N=M$ assumption produces M components which maximize an ICA criterion such as non-Gaussianity. We assume that a target source s_1 is non-Gaussian, close to sensors, and dominant in the mixtures. Therefore, it is expected that one of the M components corresponds to s_1 whose ICA criterion is high.

We apply ICA in the time-frequency domain. The reason is that it is efficient and also it fits time-frequency masking, which will be discussed in the next paragraph. An additional operation is the selection of the s_1 component in every frequency bin. It has been reported that the selection of a component with maximum kurtosis works well when the target is speech and interferences are babble source [3]. However, we consider a case where interferences are also speech. Thus, we exploit the information of basis vectors produced by ICA. In Sec. 2.3, we propose a new idea that improves our previously reported methods [4, 5].

The next issue is the fact that some interference still remains in the target component in a case where $N > M$. Post filtering [3, 6] can be used to reduce such residuals. However, it needs additional adaptation where the step size should be controlled based on the short-term power of the target. Another approach is time-frequency masking [7–11], which is effective for sources having sparseness in the time-frequency domain, such as speech. The performance of time-frequency masking depends on how well we can specify the time-frequency slots where a target source is active. A simple way to specify such slots is to calculate the phase and/or amplitude difference of sensor observations [7, 8]. In [10], the output of a beamformer is also used to improve mask accuracy. We propose a new criterion for specifying masks in Sec. 2.5. It is based on how close a sample vector is to the basis vector corresponding to a target. The closeness is calculated in a whitened space where the basis vectors corresponding to interferences are expected to be far from that of the target.

The next section describes our proposed method. Section 3 shows experimental results, and Section 4 concludes this paper.

2. THE PROPOSED METHOD

2.1. Frequency domain operations

Figure 1 shows the flow of the method proposed here. First, time-domain signals $\mathrm{x}_j(t)$ sampled at frequency f_s are converted into frequency-domain time-series signals $x_j(f,\tau)$ with an L-point short-time Fourier transform (STFT):

$$x_j(f,\tau) = \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(\tau+r)\, \mathit{win}(r)\, e^{-\jmath 2\pi f r}, \quad (3)$$

where $f = 0, \frac{1}{L}f_s, \dots, \frac{L-1}{L}f_s$ is a frequency, $\mathit{win}(r)$ is a window that tapers smoothly to zero at each end, such as a Hanning window $\frac{1}{2}(1+\cos\frac{2\pi r}{L})$, and τ is a new index representing time.

The remaining operations are performed in the frequency domain. There are two advantages to this. First, the convolutive

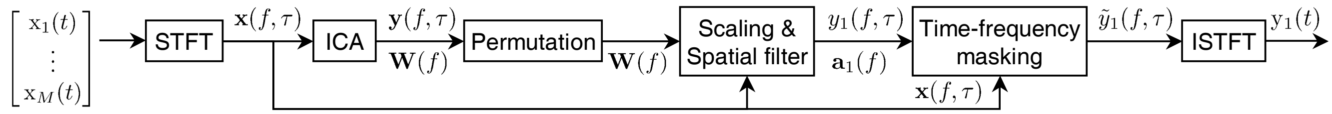

Fig. 1. Flow of the proposed method

mixtures in (1) can be approximated as instantaneous mixtures in each frequency bin:

$$x_j(f,\tau) = \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau), \tag{4}$$

where $h_{jk}(f)$ is the frequency response from source k to sensor j, and $s_k(f,\tau)$ is a frequency-domain time-series signal of $s_k(t)$ obtained by the same operation as (3). The frequency-domain counterpart of (2) is

$$y_1(f,\tau) = h_{J1}(f) s_1(f,\tau), \tag{5}$$

where J should be the same for all frequency bins f. The second advantage is that the sparseness of a source signal becomes prominent in the time-frequency domain if the source is colored such as speech. The possibility of $s_k(f,\tau)$ being close to zero is much higher than that of $\mathrm{s}_k(t)$.

Through several operations, which will be discussed in the following subsections, we have an output $\tilde{y}_1(f,\tau)$, which should be close to (5) in each frequency bin. At the end of the flow, we have an output $\mathrm{y}_1(t)$ by an inverse STFT (ISTFT) :

$$\mathrm{y}_1(\tau+r) = \frac{1}{L \cdot win(r)} \sum_{l=0}^{L-1} \tilde{y}_1(\frac{l}{L} f_s, \tau) \exp(\jmath 2\pi \frac{l}{L} f_s r).$$

2.2. Independent component analysis (ICA)

Let us have a vector notation of the mixing model (4) :

$$\mathbf{x}(f,\tau) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,\tau), \tag{6}$$

where $\mathbf{x} = [x_1, \dots, x_M]^T$ is a sample vector and $\mathbf{h}_k = [h_{1k}, \dots, h_{Mk}]^T$ is the vector of frequency responses from source s_k to all sensors. Independent component analysis (ICA) is used as a first step to identify the vector $\mathbf{h}_1$ of a dominant source s_1.

Even though the number of independent components N may be larger than the number of sensors M, we employ ICA by assuming that N is equal to M:

$$\mathbf{y}(f,\tau) = \mathbf{W}(f)\mathbf{x}(f,\tau),$$

where $\mathbf{y} = [y_1, \dots, y_M]^T$ is a vector of separated signals and $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_M]^H$ is an $M \times M$ separation matrix. In the experiments shown in Sec. 3, we calculated $\mathbf{W}$ by using a complex-valued version of FastICA [2], and improved it further by using InfoMax [12] combined with the natural gradient [13] whose nonlinear function is based on the polar coordinate [14].

By calculating the inverse of $\mathbf{W}$

$$[\mathbf{a}_1, \cdots, \mathbf{a}_M] = \mathbf{W}^{-1}, \ \mathbf{a}_i = [a_{1i}, \dots, a_{Mi}]^T, \tag{7}$$

we have the decomposition of $\mathbf{x}$

$$\mathbf{x}(f,\tau) = \sum_{i=1}^{M} \mathbf{a}_i(f) y_i(f,\tau), \tag{8}$$

where $\mathbf{a}_i$ is a basis vector. Since s_1 is assumed to be a dominant non-Gaussian source, it is expected that one of $y_1, \dots, y_M$ corresponds to s_1 and therefore one of $\mathbf{a}_1, \dots, \mathbf{a}_M$ corresponds to $\mathbf{h}_1$.

2.3. Permutation

The next operation is to find i for each frequency f such that $\mathbf{a}_i(f)$ corresponds to $\mathbf{h}_1(f)$. This is considered to be the permutation problem of frequency-domain BSS. Integrating the basis vector $\mathbf{a}_i(f)$ and signal envelope $|y_i(f,\tau)|$ information solves the problem effectively [4]. We also employ this approach here.

In the rest of this subsection, we discuss how to exploit the basis vector $\mathbf{a}_i(f)$. The former methods estimate geometric information (direction [4,5] and distance [5]) about the sources by selecting two elements from $\mathbf{a}_i(f)$, and then cluster the directions or the distances to solve the permutation problem. However, the directions and distances are clustered separately, and such estimations depend on the selection of two elements. Consequently, different kinds of estimations by different pairs of elements need to be reconciled. To solve this issue, we here propose a new method that considers all the elements of $\mathbf{a}_i(f)$ simultaneously.

The new idea is to normalize all basis vectors $\mathbf{a}_i(f)$, $i = 1, \dots, M$, for all frequency bins $f = 0, \frac{1}{L} f_s, \dots, \frac{L-1}{L} f_s$ such that they form clusters, each of which corresponds to each source. The normalization is performed by selecting a reference sensor J and calculating

$$\bar{a}_{ji}(f) \leftarrow |a_{ji}(f)| \exp\left[\jmath \frac{\arg[a_{ji}(f)/a_{Ji}(f)]}{4fc^{-1}d}\right] \tag{9}$$

where c is the propagation velocity and d is the maximum distance between the reference sensor J and a sensor $^\forall j \in \{1, \dots, M\}$. The rationale of this operation will be explained afterwards. Then, we apply unit-norm normalization

$$\bar{\mathbf{a}}_i(f) \leftarrow \bar{\mathbf{a}}_i(f) \, / \, ||\bar{\mathbf{a}}_i(f)|| \tag{10}$$

for $\bar{\mathbf{a}}_i(f) = [\bar{a}_{1i}(f), \dots, \bar{a}_{Mi}(f)]^T$. The next step is to find M clusters $C_1, \dots, C_M$ formed by normalized vectors $\bar{\mathbf{a}}_i(f)$. The centroid $\mathbf{c}_k$ of a cluster C_k is calculated by

$$\mathbf{c}_k \leftarrow \sum_{\bar{\mathbf{a}} \in C_k} \bar{\mathbf{a}} / |C_k|, \ \ \mathbf{c}_k \leftarrow \mathbf{c}_k / ||\mathbf{c}_k||,$$

where $|C_k|$ is the number of vectors in C_k. The criterion of clustering is to minimize the total sum $\mathcal{J}$ of the squared distances between cluster members and their centroid

$$\mathcal{J} = \sum_{k=1}^{M} \mathcal{J}_k, \ \mathcal{J}_k = \sum_{\bar{\mathbf{a}} \in C_k} ||\bar{\mathbf{a}} - \mathbf{c}_k||^2. \tag{11}$$

This minimization can be performed efficiently with the k-means clustering algorithm [15].

This paragraph explains the reason why normalized basis vectors $\bar{\mathbf{a}}_i(f)$ form a cluster for a source. Let us approximate the multi-path mixing model (1) by using a direct-path mixing model

$$h_{jk}(f) = \frac{q(f)}{d_{jk}} \exp\left[\jmath 2\pi f c^{-1}(d_{jk} - d_{Jk})\right], \tag{12}$$

where $d_{jk} > 0$ is the distance between source k and sensor j. We assume that the attenuation $q(f)/d_{jk}$ depends on both the distance and a frequency-dependent constant $q(f) > 0$, and that the delay $(d_{jk} - d_{Jk})/c$ depends on the distance normalized with the reference sensor J. By considering the permutation and scaling ambiguity of ICA, a basis vector and its elements are represented as

$$\mathbf{a}_i = \alpha_i \mathbf{h}_k, \quad a_{ji} = \alpha_i h_{jk}, \tag{13}$$

where α_i is a scalar representing the scaling ambiguity, and index k that may be different from index i represents the permutation ambiguity. Substituting (12) and (13) into (9) and (10) yields

$$\bar{a}_{ji}(f) = \frac{1}{d_{jk}D} \exp\left[\jmath \frac{\pi}{2} \frac{(d_{jk} - d_{Jk})}{d}\right], \; D = \sqrt{\sum_{i=1}^{M} \frac{1}{{d_{ik}}^2}}$$

which is independent of frequency, and dependent only on the positions of the sources and sensors. From the fact that $\max_{j,k} |d_{jk} - d_{Jk}| \leq d$, an inequality holds:

$$-\pi/2 \leq \arg[\bar{a}_{ji}(f)] \leq \pi/2.$$

This property is important for the distance measure (11), since $|\bar{a} - \bar{a}'|$ increases monotonically as $|\arg(\bar{a}) - \arg(\bar{a}')|$ increases.

After we have found M clusters $C_1, \ldots, C_M$, we need to identify a cluster that corresponds to a dominant source s_1. Since we assume that s_1 is close to the sensors, the mixing model (12) is more valid for s_1 than for the other sources. Therefore, we decide that a cluster that has the minimum variance $\sigma_k^2 = \mathcal{J}_k / |C_k|$ corresponds to s_1.

2.4. Scaling and Spatial filter

Here the columns of (7) have been permuted such that $\mathbf{a}_1(f)$ corresponds to $\mathbf{h}_1(f)$. Then, we solve the scaling ambiguity in (8):

$$\mathbf{a}_i y_i(\tau) = (\alpha_i \mathbf{a}_i)(y_i(\tau)/\alpha_i), \text{ for any complex scalar } \alpha_i.$$

This is easily solved by

$$\mathbf{a}_i \leftarrow \mathbf{a}_i / a_{Ji} \text{ or equivalently } \mathbf{w}_i \leftarrow [\mathbf{W}^{-1}]_{Ji} \mathbf{w}_i$$

to make $a_{Ji} = 1$ for a selected sensor J. The reason can be seen by comparing the $\mathbf{h}_1$ term in (6) and the $\mathbf{a}_1$ term in (8), and taking (5) into consideration

$$\mathbf{h}_1 s_1(\tau) \approx \mathbf{a}_1 y_1(\tau) = \mathbf{a}_1 h_{J1} s_1(\tau) \quad \Leftrightarrow \quad \mathbf{h}_1 \approx \mathbf{a}_1 h_{J1}.$$

Now the permutation and the scaling ambiguities are solved. The extraction of s_1 by a spatial filter is performed by

$$y_1(\tau) \; = \; \mathbf{w}_1^H \mathbf{x}(\tau) \tag{14}$$

$$\; = \; \mathbf{w}_1^H \mathbf{h}_1 s_1(\tau) + \textstyle\sum_{k=2}^{N} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau). \tag{15}$$

If $N \leq M$, $\mathbf{w}_1$ satisfies $\mathbf{w}_1^H \mathbf{h}_k = 0, {}^\forall k \in \{2, \ldots, N\}$ and makes the second term zero. However, we assume that the number of sources N is generally larger than M. In this case, there exists a set $K \subseteq \{2, \ldots, N\}$ such that $\mathbf{w}_1^H \mathbf{h}_k \neq 0, {}^\forall k \in K$. Thus, $y_1(\tau)$ contains an unwanted residual $\sum_{k \in K} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau)$.

2.5. Time-frequency masking

The purpose here is to have another output $\tilde{y}_1(\tau)$ that contains less power of the residual $\sum_{k \in K} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau)$ than $y_1(\tau)$. This is performed by time-frequency masking

$$\tilde{y}_1(f, \tau) = \mathcal{M}(f, \tau) \cdot y_1(f, \tau), \tag{16}$$

where $0 \leq \mathcal{M}(f, \tau) \leq 1$ is a mask specified for each time-frequency slot (f, τ). We specify masks based on the angle $\theta_1(f, \tau)$ between $\mathbf{a}_1(f)$ and $\mathbf{x}(f, \tau)$ calculated in the space transformed by a whitening matrix $\mathbf{V}(f) = \mathbf{R}^{-1/2}$, $\mathbf{R} = \langle \mathbf{x}(\tau)\, \mathbf{x}(\tau)^H \rangle_\tau$. Let $\mathbf{z}(f, \tau) = \mathbf{V}(f)\mathbf{x}(f, \tau)$ be whitened samples and $\mathbf{b}_1(f) = \mathbf{V}(f)\mathbf{a}_1(f)$ be the basis vector of $y_1(\tau)$ in the whitened space. The angle is calculated by

$$\theta_1(f, \tau) = \cos^{-1} \frac{|\mathbf{b}_1^H(f) \cdot \mathbf{z}(f, \tau)|}{||\mathbf{b}_1(f)|| \cdot ||\mathbf{z}(f, \tau)||} \tag{17}$$

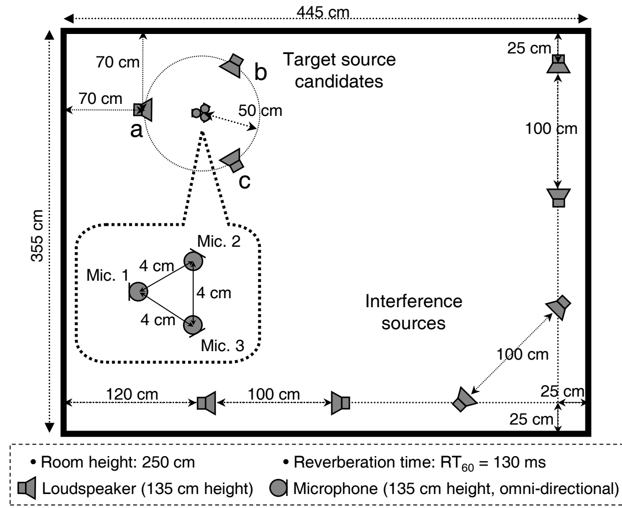

Fig. 2. Experimental conditions

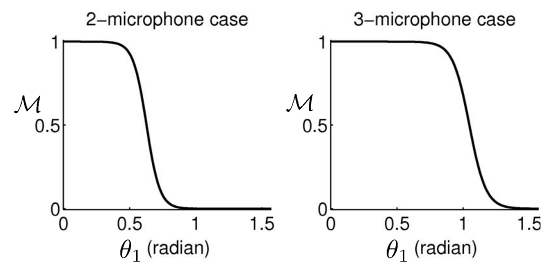

Fig. 3. Masking functions: parameters $(\theta_T, g) = (\pi/5, 20)$ and $(\pi/3, 16)$ were used for 2- and 3-microphone cases, respectively

for each time-frequency slot. Then, we calculate a mask by using a logistic function

$$\mathcal{M}(\theta_1(f, \tau)) = 1/(1 + e^{g(\theta_1 - \theta_T)}), \tag{18}$$

where θ_T and g are parameters specifying the transition point and its steepness, respectively. The smaller θ_T, the less power of the residual appears in $\tilde{y}_1$ but the more musical noises sound in y_1.

The effectiveness of the above operation depends on the sparseness of sources. If we assume that the possibility of $s_k(f, \tau)$ being close to zero is very high, (6) can be approximated as

$$\mathbf{x}(f, \tau) \approx \mathbf{h}_k(f) s_k(f, \tau), \; k \in \{1, \ldots, N\}, \tag{19}$$

where k depends on each time-frequency slot (f, τ). Let us consider the whitened-space counterpart of (19), with distinguishing the cases where s_1 is the only active source and other cases:

$$\mathbf{z}(\tau) \; \approx \; \mathbf{V}\mathbf{h}_1 s_1(\tau) \approx \mathbf{V}\mathbf{a}_1 y_1(\tau) \tag{20}$$

$$\mathbf{z}(\tau) \; \approx \; \textstyle\sum_{k=2}^{N} \mathbf{V}\mathbf{h}_k s_k(\tau). \tag{21}$$

If the number of sources N is equal to or less than the number of sensors M, vectors $\mathbf{V}\mathbf{h}_k$ in whitened space are orthogonal to each other. Even if $N > M$, the vector $\mathbf{b}_1 = \mathbf{V}\mathbf{a}_1 \approx \mathbf{V}\mathbf{h}_1$ of a dominant source s_1 tends to have large angles with the other vectors $\mathbf{V}\mathbf{h}_k, k = 2, \ldots, N$. Therefore, calculating the angle (17) provides information about whether or not s_1 is the only active source at a time-frequency slot (f, τ).

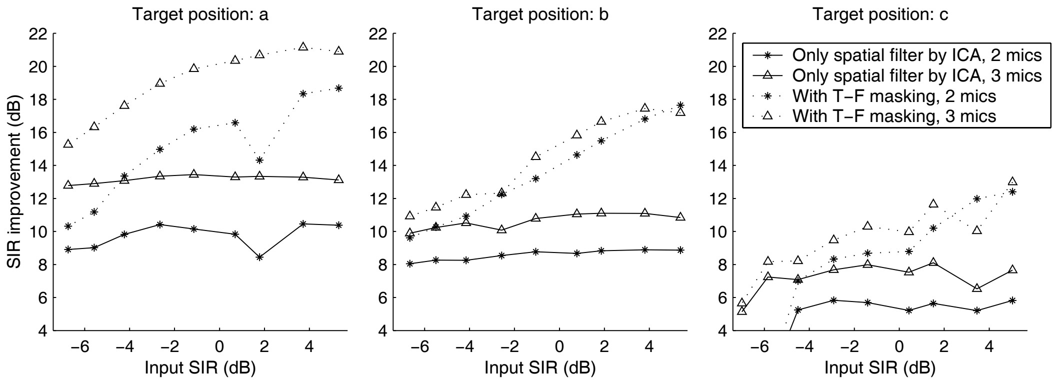

Fig. 4. SIR improvements for various input SIRs

3. EXPERIMENTS

We performed experiments to examine the effectiveness of the proposed method. We measured impulse responses under the conditions shown in Fig. 2. The speaker positions simulated a cocktail party situation. Mixtures at microphones were made by convolving the impulse responses and 6-second speeches sampled at 8 kHz. For each setup, we selected one of the three speakers (a, b, c) as a dominant target source, and the remaining two speakers were kept silent. The six speakers away from the microphones were used as interferences for every setup. We set parameters of (18) as shown in Fig. 3, which were empirically proved to be good.

Figure 4 shows experimental results for various input SIRs, which were controlled by multiplying a positive constant for all interferences equally. The vertical axis shows the SIR improvements obtained solely by using a spatial filter made by ICA, and by using the combination of the spatial filter and time-frequency (T-F) masking. The frame size L of STFT (3) was 1024 (128 ms). SIR improvements depend heavily on the position of the target source. Position a was good for enhancement, whereas c was a hard position as a lot of interferences came from the similar directions. The improvement realized by using a spatial filter was almost the same independent of input SIRs. This means that ICA and the following permutation alignment worked similarly well for various input SIRs. In contrast, the further improvement achieved by time-frequency masking depends on the input SIRs. This is why the vectors $\mathbf{V}\mathbf{h}_k$ of interferences in the whitened space approaches that $\mathbf{V}\mathbf{h}_1$ of a target source, as the interference power increases. With the masking functions shown in Fig. 3, musical noises hardly sound in the output signals, especially 3-microphone cases.

4. CONCLUSION

We have presented a method for extracting a dominant target source and suppressing interferences. The process of ICA and following permutation alignment extracts the target source by a spatial filter, and estimates the basis vector corresponding to the target source. The performance depends on the accuracy of the permutation alignment, which has been improved by the new idea shown in subsection 2.3. Time-frequency masking in the second stage reduces the power of the residual caused by the limitations of a spatial filter. It exploits the sparseness of sources. The angle between the basis vector corresponding to the target source and a sample vector gives information about whether or not the target is active. The experiments showed good results for extracting a dominant source out from six interferences only with 2 or 3 microphones, in a room whose reverberation time was 130 ms.

5. REFERENCES

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

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

[3] S. Y. Low, R. Togneri, and S. Nordholm, "Spatio-temporal processing for distant speech recognition," in *Proc. ICASSP 2004*, May 2004, vol. I, pp. 1001–1004.

[4] 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, pp. 530–538, Sept. 2004.

[5] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation using small and large spacing sensor pairs," in *Proc. ISCAS 2004*, May 2004, vol. V, pp. 1–4.

[6] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual crosstalk components in blind source separation using LMS filters," in *Proc. NNSP 2002*, Sept. 2002, pp. 435–444.

[7] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[8] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. ICA2001*, Dec. 2001, pp. 651–656.

[9] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. ICASSP 2004*, May 2004, vol. III, pp. 881–884.

[10] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. ICASSP 2004*, May 2004, vol. II, pp. 373–376.

[11] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic Publishers, 2004.

[12] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[13] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[14] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, Mar. 2003.

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