Blind extraction of dominant target sources using ICA and time-frequency masking

Hiroshi Sawada*, *Senior Member, IEEE,* Shoko Araki, *Member, IEEE,*
Ryo Mukai, *Senior Member, IEEE,* Shoji Makino, *Fellow, IEEE*

***Abstract*— This paper presents a method for enhancing target sources of interest and suppressing other interference sources. The target sources are assumed to be close to sensors, to have dominant powers at these sensors, and to have non-Gaussianity. The enhancement is performed blindly, i.e. without knowing the position and active time of each source. We consider a general case where the total number of sources is larger than the number of sensors, and neither the number of target sources nor the total number of sources is known. The method is based on a two-stage process where independent component analysis (ICA) is first employed in each frequency bin and then time-frequency masking is used to improve the performance further. We propose a new sophisticated method for deciding the number of target sources and then selecting their frequency components. We also propose a new criterion for specifying time-frequency masks. Experimental results for simulated cocktail party situations in a room, whose reverberation time was 130 ms, are presented to show the effectiveness and characteristics of the proposed method.**

***Index Terms*— Blind source separation, blind source extraction, independent component analysis, convolutive mixture, frequency domain, permutation problem, time-frequency masking**

I. Introduction

The technique for estimating individual source components from their mixtures at sensors is known as blind source separation (BSS) [1]–[4]. With 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, with audio applications such as speech enhancement, the sources do not necessarily have equal significance. We often want to extract only a few sources that are close to sensors, have dominant powers, and/or have interesting features.

This paper presents a method for extracting source signals of interest and suppressing other interference sources blindly. Let us formulate the task. Suppose that a few target sources $\mathrm{s}_1, \ldots, \mathrm{s}_Q$ and other background sources $\mathrm{s}_{Q+1}, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at $M \geq 2$ sensors

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \mathrm{x}_{jk}(t),\ j=1,\ldots,M, \tag{1}$$

where

$$\mathrm{x}_{jk}(t) = \sum_{l} \mathrm{h}_{jk}(l)\,\mathrm{s}_k(t-l) \tag{2}$$

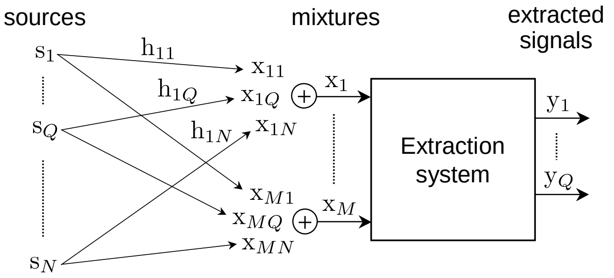

Fig. 1. Signal notations

is the component of s_k measured at sensor j. The impulse response from source k to sensor j is denoted as $\mathrm{h}_{jk}(l)$. The goal is to have output signals $\mathrm{y}_i(t)$, $i = 1, \ldots, Q$ that are close to the components

$$\mathrm{x}_{Ji}(t) = \sum_{l} \mathrm{h}_{Ji}(l)\,\mathrm{s}_i(t-l),\ \ i = 1, \ldots, Q, \tag{3}$$

of target sources measured at a selected sensor J. The task should be performed only with the M mixtures $\mathrm{x}_1, \ldots, \mathrm{x}_M$. The number of target sources Q and the total number of sources N are unknown. The number Q is assumed to be no more than the number of sensors M, and N may be larger than M. Figure 1 shows the signal notations.

The first problem is how to extract target sources $\mathrm{s}_1, \ldots, \mathrm{s}_Q$ blindly. Even if the total number of sources N could be larger than M, independent component analysis (ICA) [1]–[4] with an $N = M$ assumption produces M components that maximize an ICA criterion such as non-Gaussianity. We assume that the target sources are non-Gaussian, close to sensors, and dominant in the mixtures. Therefore, we expect that some of the M components produced by ICA correspond to target sources $\mathrm{s}_1, \ldots, \mathrm{s}_Q$ whose ICA criteria are high.

We employ ICA in the time-frequency domain [5]–[11]. The reason is that it is efficient [11] and also fits time-frequency masking, which is discussed below. An additional operation that should be performed is the selection of each target component in every frequency bin. This is considered to be the permutation problem of frequency-domain BSS [12]. It has been reported that the selection of a component with maximum kurtosis works well when the target is speech and the interferences are babble sources [13]. However, this does not always work well for a case where the interferences are also speech.

In order to solve the permutation problem for mixtures with many speeches, we exploit the information of basis vectors

The authors are with NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan (e-mail: sawada@cslab.kecl.ntt.co.jp; shoko@cslab.kecl.ntt.co.jp; ryo@cslab.kecl.ntt.co.jp; maki@cslab.kecl.ntt.co.jp, phone: +81-774-93-5272, fax: +81-774-93-5158). EDICS: AUD-LMAP

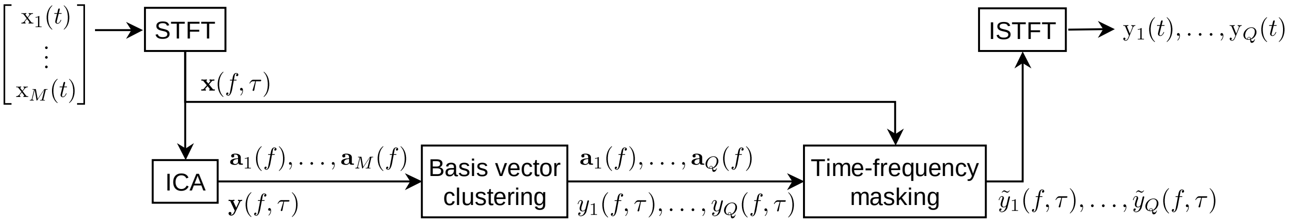
$\begin{bmatrix} \mathrm{x}_1(t) \\ \vdots \\ \mathrm{x}_M(t) \end{bmatrix}$
STFT
$\mathbf{x}(f,\tau)$
ICA
$\mathbf{a}_1(f),\ldots,\mathbf{a}_M(f)$
$\mathbf{y}(f,\tau)$
Basis vector clustering
$\mathbf{a}_1(f),\ldots,\mathbf{a}_Q(f)$
$y_1(f,\tau),\ldots,y_Q(f,\tau)$
Time-frequency masking
$\tilde{y}_1(f,\tau),\ldots,\tilde{y}_Q(f,\tau)$
ISTFT
$\mathrm{y}_1(t),\ldots,\mathrm{y}_Q(t)$

Fig. 2. Flow of proposed method

(10) produced by ICA. Our previously reported methods estimate the directions [12], [14] and/or the distances [14] of the sources from the basis vectors, and then cluster the estimated directions and/or distances to solve the permutation problem. However, the system needs to know the locations of sensors to estimate such geometric information about the sources. In Sec. IV, we propose a new method for solving the permutation problem by clustering the basis vectors themselves after some normalization. With this approach, we do not need to know the sensor locations, simply the maximum distance between a sensor and any other sensor. This relaxation makes it easy to use a non-uniform arrangement of sensors, and also eliminates the need for sensor calibration.

The next issue is that some interference still remains in the extracted frequency components when $N > M$. Post filtering [13], [15] can be used to reduce such residual interference. 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 [16]–[22], which is efficient for sources with sparseness in the time-frequency domain, such as speech. Time-frequency masking has been well studied in the research area of computational auditory scene analysis [22]–[26]. The performance of time-frequency masking depends on how well we can specify the time-frequency slots where the target source is active. A simple way to specify such slots is to calculate the phase and/or amplitude difference between the observations of different sensors [16]–[18]. Another recently proposed approach involves calculating the power ratio between an input and outputs of a spatial filter (beamformer [20], [21] or ICA [21]). However, such a power-based criterion depends on the scaling ambiguity of ICA or beamformer outputs. In Sec. V, we propose a new criterion for specifying masks. It is based on the cosine distance between a sample vector and the basis vector corresponding to a target. The distance is calculated in a spatially whitened space where the target basis vector is expected to be almost orthogonal to those of interferences. Therefore, the new criterion does not suffer from the problem of scaling ambiguity.

This paper is organized as follows. The next section provides an overview of our proposed method. Section III discusses how ICA can be applied to our situation and what should be done for the ICA results. Section IV describes how basis vector clustering works to solve the permutation problem and to decide the number of target sources to be extracted. Section V discusses how to specify time-frequency masks. Section VI presents experimental results, and Sec. VII concludes this paper.

II. Overview of proposed method

Figure 2 shows the flow of the proposed method. First, time-domain observed 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):

$$x_j(f,\tau) \leftarrow \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(\tau+r)\,\mathrm{win}(r)\,e^{-\jmath 2\pi f r}, \qquad (4)$$

where $f \in \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$ is a frequency, $\mathrm{win}(r)$ 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 following operations are performed in the frequency domain. There are two advantages to this. First, the convolutive mixtures (1) can be approximated as instantaneous mixtures at each frequency:

$$x_j(f,\tau) \approx \sum_{k=1}^{N} h_{jk}(f) s_k(f,\tau), \qquad (5)$$

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 $\mathrm{s}_k(t)$ obtained by the same operation as (4). The frequency-domain counterpart of (3) is

$$x_{Ji}(f,\tau) \approx h_{Ji}(f) s_i(f,\tau), \;\; i = 1,\ldots,Q, \qquad (6)$$

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 and non-stationary such as speech. The possibility of $s_k(f,\tau)$ being close to zero is much higher than that of $\mathrm{s}_k(t)$.

Then, we apply ICA (Sec. III) to the STFT results

$$\mathbf{x}(f,\tau) = [x_1(f,\tau),\ldots,x_M(f,\tau)]^T, \qquad (7)$$

which we call a sample vector, and obtain basis vectors $\mathbf{a}_1(f),\ldots,\mathbf{a}_M(f)$ defined by (10) and independent components

$$\mathbf{y}(f,\tau) = [y_1(f,\tau),\ldots,y_M(f,\tau)]^T. \qquad (8)$$

Some of these independent components correspond to the components of dominant sources. However, the correspondence is not clear at this stage because of the permutation ambiguity of ICA. Thus, basis vector clustering (Sec. IV) is performed to decide the number Q of target sources and produce basis vectors $\mathbf{a}_1(f),\ldots,\mathbf{a}_Q(f)$ and independent components $y_1(f,\tau),\ldots,y_Q(f,\tau)$, whose correspondences to the target sources are specified.

If the number of total sources N is larger than the number of sensors M, independent components $y_1(f,\tau),\ldots,y_Q(f,\tau)$ produced by ICA have some residuals caused by the limitation of spatial filtering. Time-frequency masking (Sec. V) is used to reduce such residuals and to obtain outputs $\tilde{y}_1(f,\tau),\ldots,\tilde{y}_Q(f,\tau)$, which should be close to (6) in each frequency bin. At the end of the flow, time-domain output signals $\mathrm{y}_i(t)$ are obtained by an inverse STFT (ISTFT):

$$\mathrm{y}_i(\tau+r) \leftarrow \frac{1}{L\cdot\mathrm{win}(r)} \sum_{f\in\{0,\,\frac{1}{L}f_s,\,\ldots,\,\frac{L-1}{L}f_s\}} \tilde{y}_i(f,\tau)\,e^{\jmath\,2\pi f r}$$

for $i=1,\ldots,Q$.

III. Independent component analysis (ICA)

To extract the components of dominant sources, we apply ICA to sample vectors $\mathbf{x}(f,\tau)$. Even though the total number of sources N may be larger than the number of sensors M, we employ ICA by assuming that the number of independent components is equal to M:

$$\mathbf{y}(f,\tau) = \mathbf{W}(f)\,\mathbf{x}(f,\tau), \tag{9}$$

where $\mathbf{W}=[\mathbf{w}_1,\ldots,\mathbf{w}_M]^H$ is an $M\times M$ separation matrix. In the experiments shown in Sec. VI, we calculated $\mathbf{W}$ by using a complex-valued version of FastICA [3], [27], and improved it further by using InfoMax [28] combined with the natural gradient [29] whose nonlinear function is based on the polar coordinate [30].

Then, we calculate the inverse of $\mathbf{W}$ to obtain basis vectors

$$[\mathbf{a}_1,\cdots,\mathbf{a}_M] = \mathbf{W}^{-1},\ \mathbf{a}_i=[a_{1i},\ldots,a_{Mi}]^T. \tag{10}$$

It is not difficult to make $\mathbf{W}$ invertible by using an appropriate ICA procedure, such as whitening followed by unitary transformation (e.g. FastICA [3]). By multiplying both sides of (9) by $\mathbf{W}^{-1}$, the sample vector $\mathbf{x}(f,\tau)$ is represented by a linear combination of basis vectors $\mathbf{a}_1,\ldots,\mathbf{a}_M$:

$$\mathbf{x}(f,\tau) = \sum_{i=1}^{M} \mathbf{a}_i(f) y_i(f,\tau). \tag{11}$$

By comparing (11) and the vector notation of the mixing model (5):

$$\mathbf{x}(f,\tau) \approx \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,\tau), \tag{12}$$

where $\mathbf{h}_k=[h_{1k},\ldots,h_{Mk}]^T$, we observe the following fact. Since $s_1,\ldots,s_Q$ are assumed to be dominant non-Gaussian sources, it is strongly expected that some of $y_1,\ldots,y_M$ correspond to $s_1,\ldots,s_Q$ and thus some of $\mathbf{a}_1,\ldots,\mathbf{a}_M$ correspond to $\mathbf{h}_1,\ldots,\mathbf{h}_Q$. The correspondence between $\mathbf{a}_i y_i$ and $\mathbf{h}_k s_k$ as well as the number Q of target sources are unknown at this time, because of the permutation ambiguity of ICA. They will be specified by basis vector clustering as described in Sec. IV.

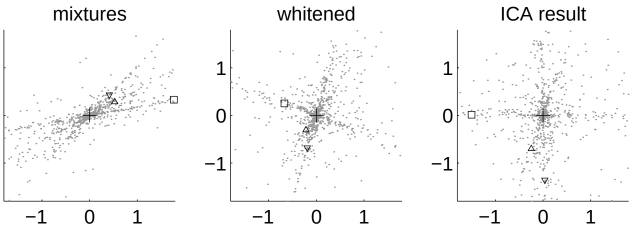

Fig. 3. ICA example for real instantaneous mixtures ($N=3, M=2, Q=1$)

Once the correspondence between basis vectors and target sources are identified, i.e. the permutation problem is solved, we solve the scaling ambiguity in (11):

$$\mathbf{a}_i(f) y_i(f,\tau) = (\alpha_i \mathbf{a}_i(f))\,(y_i(f,\tau)/\alpha_i), \tag{13}$$

for any non-zero complex scalar α_i. This is easily solved by

$$y_i(f,\tau) \leftarrow a_{Ji}(f) y_i(f,\tau), \tag{14}$$

where J is the index of the sensor specified in (6). The reason is as follows. The goal in each frequency bin is to make $y_i(f,\tau)$ as close to $x_{Ji}(f,\tau)$ defined in (6) as possible. And we can derive relations

$$x_{Ji}(f,\tau) \approx h_{Ji}(f)s_i(f,\tau) \approx a_{Ji}(f)y_i(f,\tau) \tag{15}$$

from (6), the $\mathbf{h}_i$ term in (12) and the $\mathbf{a}_i$ term in (11).

Here, we have a simple example to see how well standard ICA works for such a case where the number of total sources N is larger than the number of sensors M but the number of dominant sources Q is no more than M. Figure 3 shows some plots for real instantaneous mixtures. The left hand plot shows the mixtures. The square shows the mixing vector of the dominant target source and the two triangles show those of the other less dominant sources. The center plot shows whitened mixtures, where separation is not achieved. The right hand plot shows the ICA result, where the mixing vector of the dominant target source is identified.

IV. Basis vector clustering

The purpose of basis vector clustering is to solve the permutation problem and also to decide the number Q of dominant target sources. As shown in [12], integrating the basis vector $\mathbf{a}_i(f)$ and signal envelope $|y_i(f,\tau)|$ information solves the permutation problem robustly and precisely, and we also employ this approach in the experiments shown in Sec. VI. In the rest of this section, we discuss a new method for exploiting the basis vector information.

A. Frequency normalization

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

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

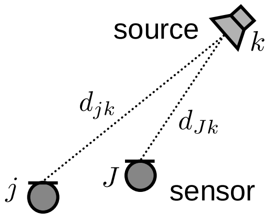

Fig. 4. Direct-path (nearfield) model

where c is the propagation velocity and $d_{\max}$ is the maximum distance between the reference sensor J and any sensor $^{\forall}j \in \{1,\ldots,M\}$. Then, we apply unit-norm normalization

$$\bar{\mathbf{a}}_i(f) \leftarrow \frac{\bar{\mathbf{a}}_i(f)}{||\bar{\mathbf{a}}_i(f)||} \tag{17}$$

for $\bar{\mathbf{a}}_i(f) = [\bar{a}_{1i}(f),\ldots,\bar{a}_{Mi}(f)]^T$.

Here we explain why normalized basis vectors $\bar{\mathbf{a}}_i(f)$ form a cluster for a source. Let us approximate the multi-path mixing model used in (1) and (5) by using a direct-path (nearfield) model (Fig. 4)

$$h_{jk}(f) \approx \frac{q(f)}{d_{jk}} \exp\left[-\jmath\, 2\pi f c^{-1}(d_{jk} - d_{Jk})\right], \tag{18}$$

where $d_{jk} > 0$ is the distance between source k and sensor j. We assume that the phase $-2\pi f c^{-1}(d_{jk} - d_{Jk})$ depends on the distance normalized with the distance to the reference sensor J. This assumption makes the phase zero at the reference sensor J. We also assume that the attenuation $q(f)/d_{jk}$ depends on both the distance and a frequency-dependent constant $q(f) > 0$.

By considering the permutation and scaling ambiguity of ICA, a basis vector and its elements are represented as

$$\mathbf{a}_i \approx \alpha_i \mathbf{h}_k, \quad a_{ji} \approx \alpha_i h_{jk}, \tag{19}$$

where α_i is a non-zero complex scalar representing the scaling ambiguity, and index k, which may be different from index i, represents the permutation ambiguity. Substituting (18) and (19) into (16) and (17) yields

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

which is independent of frequency, and dependent only on the positions of the sources and sensors. Therefore, normalized basis vectors $\bar{\mathbf{a}}_i(f)$ form a cluster for a source that is placed at a specific position.

Let us discuss the intention of using $4fc^{-1}d_{\max}$ for the denominator of (16). From the fact that $\max_{j,k} |d_{jk} - d_{Jk}| \le d_{\max}$, an inequality

$$-\pi/2 \le \arg[\bar{a}_{ji}(f)] \le \pi/2 \tag{20}$$

holds. This property is important for the distance measure (22), which will be used for the clustering algorithm. In the range specified by (20), the distance between two elements $|\bar{a} - \bar{a}'|$ increases monotonically as the difference between two arguments $|\arg(\bar{a}) - \arg(\bar{a}')|$ increases. If argument $\arg(\bar{a})$ falls out of the range (20), the distance $|\bar{a} - \bar{a}'|$ would decrease as the difference $|\arg(\bar{a}) - \arg(\bar{a}')|$ increases, which would adversely affect the distance measure (22).

B. Clustering basis vectors and solving permutation ambiguity

After normalizing all basis vectors $\mathbf{a}_i(f)$, $i = 1,\ldots,M$ and $f = 0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s$, we perform a clustering algorithm to find clusters $C_1,\ldots,C_M$ formed by normalized vectors $\bar{\mathbf{a}}_i(f)$. The centroid $\mathbf{c}_k$ of a cluster C_k is calculated by

$$\mathbf{c}_k \leftarrow \sum_{\bar{\mathbf{a}} \in C_k} \frac{\bar{\mathbf{a}}}{|C_k|},\ \ \mathbf{c}_k \leftarrow \frac{\mathbf{c}_k}{||\mathbf{c}_k||}, \tag{21}$$

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

$$\mathcal{J} = \sum_{k=1}^{M} \mathcal{J}_k,\ \mathcal{J}_k = \sum_{\bar{\mathbf{a}} \in C_k} ||\bar{\mathbf{a}} - \mathbf{c}_k||^2. \tag{22}$$

This minimization can be performed efficiently with the k-means clustering algorithm [31]. Some examples of clustering results are shown in Figs. 10 and 11.

Once we have found M clusters $C_1,\ldots,C_M$, we need to identify clusters that correspond to dominant target sources $s_1,\ldots,s_Q$. We decide that a cluster C_k with a small variance $\mathcal{J}_k/|C_k|$ belongs to the set of target sources. The rationale behind this criterion is that the mixing model (18) is more valid for $s_1,\ldots,s_Q$ than for the other sources. For target sources $k = 1,\ldots,Q$, the direct-path components of impulse responses h_{jk} are distinct since s_k is assumed to be close to the sensors.

To identify target source clusters, we sort the clusters $\mathbf{c}_1,\ldots,\mathbf{c}_M$ so that their variances are sorted in ascending order:

$$\frac{\mathcal{J}_1}{|C_1|} \le \ldots \le \frac{\mathcal{J}_Q}{|C_Q|} \le th_{var} \le \ldots \le \frac{\mathcal{J}_M}{|C_M|}, \tag{23}$$

where th_{var} is a predefined threshold for specifying the set of target sources. Then, to align the permutation ambiguity of ICA, we renumber the indexes of the basis vectors by

$$\mathbf{a}_k(f) \leftarrow \mathbf{a}_{\Pi_f(k)}(f), \tag{24}$$

where $\Pi_f : \{1,\ldots,Q\} \rightarrow \{1,\ldots,M\}$ is a one-to-one mapping decided for each frequency f by

$$\Pi_f = \mathrm{argmin}_{\Pi} \sum_{k=1}^{Q} ||\bar{\mathbf{a}}_{\Pi(k)}(f) - \mathbf{c}_k||^2. \tag{25}$$

We also renumber independent components $y_1(f,\tau),\ldots,$ $y_M(f,\tau)$ accordingly.

V. Time-frequency masking

A. Motivation

Let us discuss the motivation for using time-frequency masking. Suppose that the permutation ambiguities of ICA

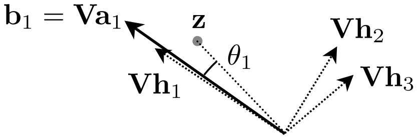

Fig. 5. Angle θ_1 calculated in whitened space

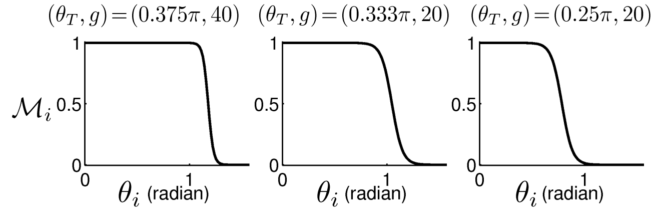

Fig. 6. Masking functions (30) with three sets of parameters (θ_T, g)

solutions are solved at this stage. Then, the extraction of a dominant source s_i by ICA (9) is represented by

$$\begin{aligned} y_i(\tau) &= \mathbf{w}_i^H \mathbf{x}(\tau) && (26)\\ &= \mathbf{w}_i^H \mathbf{h}_i s_i(\tau) + \sum_{k \neq i} \mathbf{w}_i^H \mathbf{h}_k s_k(\tau). && (27) \end{aligned}$$

If $N \leq M$, $\mathbf{w}_i$ satisfies $\mathbf{w}_i^H \mathbf{h}_k = 0, {}^\forall k \neq i$ and makes the second term zero. However, we assume that the total number N of sources is generally larger than M. In this case, there exists a set $\mathcal{K} \subseteq \{1, \ldots, i-1, i+1, \ldots, N\}$ such that $\mathbf{w}_i^H \mathbf{h}_k \neq 0, {}^\forall k \in \mathcal{K}$. Thus, $y_i(\tau)$ contains unwanted residuals $\sum_{k \in \mathcal{K}} \mathbf{w}_i^H \mathbf{h}_k s_k(\tau)$. The purpose of time-frequency masking is to obtain another version of output $\tilde{y}_i(\tau)$ that contains less power of the residuals $\sum_{k \in \mathcal{K}} \mathbf{w}_i^H \mathbf{h}_k s_k(\tau)$ than $y_i(\tau)$.

B. Proposed procedure

Time-frequency masking is performed by

$$\tilde{y}_i(f, \tau) = \mathcal{M}_i(f, \tau)\, y_i(f, \tau), \;\; i = 1, \ldots, Q \qquad (28)$$

where $0 \leq \mathcal{M}_i(f, \tau) \leq 1$ is a mask specified for each time-frequency slot (f, τ). We specify masks based on the angle $\theta_i(f, \tau)$ between $\mathbf{a}_i(f)$ 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 $\mathbf{b}_i(f) = \mathbf{V}(f)\mathbf{a}_i(f)$ be the basis vector in the whitened space. The angle is calculated by

$$\theta_i(f, \tau) = \arccos \frac{|\mathbf{b}_i^H(f)\, \mathbf{z}(f, \tau)|}{||\mathbf{b}_i(f)|| \cdot ||\mathbf{z}(f, \tau)||} \qquad (29)$$

for each time-frequency slot (Fig. 5). Then, we calculate a mask by using a logistic function [26] (Fig. 6)

$$\mathcal{M}_i(\theta_i(f, \tau)) = \frac{1}{1 + e^{g(\theta_i - \theta_T)}}, \qquad (30)$$

where θ_T and g are parameters specifying the transition point and its steepness, respectively. As θ_T becomes smaller, the residual power that appears in $\tilde{y}_i$ decreases but the musical noise in y_i increases.

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, (12) can be approximated as

$$\mathbf{x}(f, \tau) \approx \mathbf{h}_k(f) s_k(f, \tau), \;\; k \in \{1, \ldots, N\}, \qquad (31)$$

where k depends on each time-frequency slot (f, τ). Let us consider the whitened-space counterpart of (31), while distinguishing between cases where s_i is the only active source (32) and other cases (33):

$$\begin{aligned} \mathbf{z}(f, \tau) &\approx \mathbf{V}(f)\mathbf{h}_i(f) s_i(f, \tau) \approx \mathbf{V}(f)\mathbf{a}_i(f) y_i(f, \tau) && (32)\\ \mathbf{z}(f, \tau) &\approx \sum_{k \neq i} \mathbf{V}(f)\mathbf{h}_k(f) s_k(f, \tau). && (33) \end{aligned}$$

If the number of sources N is equal to or less than the number of sensors M, vectors $\mathbf{V}\mathbf{h}_1, \ldots, \mathbf{V}\mathbf{h}_N$ in the whitened space are orthogonal to each other. Even if $N > M$, the vector $\mathbf{b}_i = \mathbf{V}\mathbf{a}_i$ of a dominant source s_i, which points in almost the same direction as $\mathbf{V}\mathbf{h}_i$, tends to have large angles with the other vectors $\mathbf{V}\mathbf{h}_1, \ldots, \mathbf{V}\mathbf{h}_{i-1}, \mathbf{V}\mathbf{h}_{i+1}, \ldots, \mathbf{V}\mathbf{h}_N$. Figure 5 (and also Fig. 3) shows such a case. Therefore, calculating the angle (29) provides information about whether or not s_i is the only active source at a time-frequency slot (f, τ), and specifies the corresponding mask $\mathcal{M}_i(f, \tau)$ accordingly.

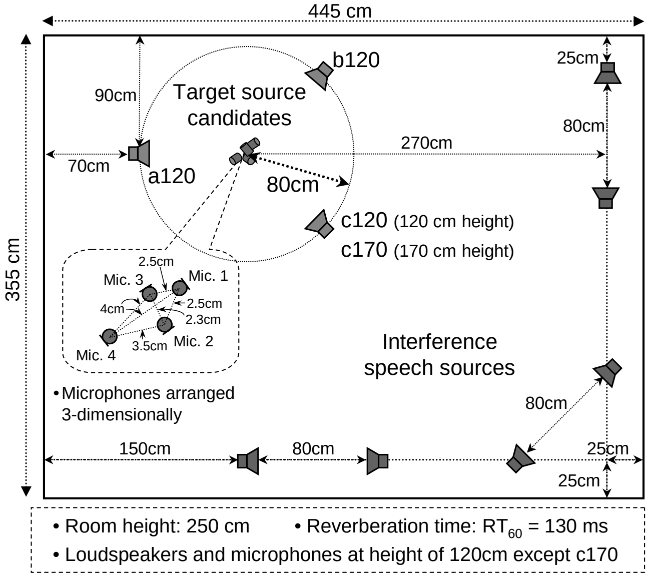

Fig. 7. Experimental conditions

VI. Experiments

A. Experimental conditions and evaluation measures

We performed experiments to enhance dominant speeches that were close to microphones. We measured impulse responses $\mathrm{h}_{jk}(l)$ under the conditions shown in Fig. 7. The speaker positions simulated a cocktail party situation. Mixtures at the microphones were made by convolving the impulse responses and 6-second English and Japanese speeches sampled at 8 kHz. We used four microphones ($M = 4$), whose arrangement was 3-dimensional and non-uniform. The system knew only the maximum distance (4 cm) between the reference microphone (Mic. 1) and the others. For each setup,

TABLE I

AVERAGE SIR IMPROVEMENT FOR SINGLE-TARGET CASES (dB)

Target position	a120	b120	c120	c170
$\mathsf{InputSIR}_i$	1.3	1.5	1.9	-0.0
Only ICA	11.7	11.8	9.0	13.0
ICA and T-F masking $(0.375\pi, 40)$	15.4	14.6	12.5	16.9
ICA and T-F masking $(0.333\pi, 20)$	16.8	15.8	14.1	18.3
ICA and T-F masking $(0.25\pi, 20)$	19.5	18.2	16.9	21.0

TABLE II

AVERAGE SDR FOR SINGLE-TARGET CASES (dB)

Target position	a120	b120	c120	c170
Only ICA	9.3	10.1	10.7	10.8
ICA and T-F masking $(0.375\pi, 40)$	8.1	8.9	9.3	9.4
ICA and T-F masking $(0.333\pi, 20)$	7.4	8.1	8.3	8.5
ICA and T-F masking $(0.25\pi, 20)$	5.4	5.9	5.8	6.2

we selected some of the four speakers (a120, b120, c120, c170) as dominant target sources, and the others were kept silent. The six speakers away from the microphones were used to provide interference sources for every setup. The frame size L of STFT (4) was 1024 (128 ms). We used $th_{var} = 0.015$ to specify the set of target sources by (23).

The performance was evaluated in terms of the signal-to-interference ratio (SIR) improvement for each output i. The improvement was calculated by $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$. These two types of SIRs are defined by

$$\mathsf{InputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{x}_{Ji}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{x}_{Jk}(t)|^2\rangle_t} \quad \text{(dB)},$$

$$\mathsf{OutputSIR}_i = 10\log_{10}\frac{\langle|\mathrm{y}_{ii}(t)|^2\rangle_t}{\langle|\sum_{k\neq i}\mathrm{y}_{ik}(t)|^2\rangle_t} \quad \text{(dB)},$$

where $\mathrm{x}_{Ji}(t)$ is defined in (3) and $\mathrm{y}_{ik}(t)$ is the component of s_k that appears at output $\mathrm{y}_i(t)$: $\mathrm{y}_i(t) = \sum_{k=1}^{N}\mathrm{y}_{ik}(t)$.

The performance was also evaluated by signal-to-distortion ratio (SDR) to allow us to observe the unwanted effects of time-frequency masking, such as musical noises. SDR is defined by

$$\mathsf{SDR}_i = 10\log_{10}\frac{\langle|\alpha_i\,\mathrm{x}_{Ji}(t-D_i)|^2\rangle_t}{\langle|\mathrm{y}_{ii}(t)-\alpha_i\,\mathrm{x}_{Ji}(t-D_i)|^2\rangle_t} \quad \text{(dB)},$$

where D_i and α_i are scalars for aligning delay and amplitude. They are specified so that the power of the denominator $\langle|\mathrm{y}_{ii}(t)-\alpha_i\mathrm{x}_{Ji}(t-D_i)|^2\rangle_t$ is minimized.

B. Single-target cases

First we show the results for single-target cases ($Q = 1$). Experiments were conducted with 16 combinations of 7 speeches (1 target + 6 background interferences) for each target position. The computational time was around 12 seconds for 6-second speech mixtures. The program was coded in Matlab and run on Athlon 64 FX-53 (2.4 GHz CPU clock).

Table I shows the average SIR improvements obtained only with ICA, and by the combination of ICA and time-frequency (T-F) masking. The SIR improvements clearly depend on the position of the target source. Positions a120 and b120 were fairly good for enhancement. This is because the interferences came from different directions. If we consider the speaker arrangement 2-dimensionally, positions c120 and c170 seems to be a hard position as many interferences came from similar directions. However, the result for position c170 was very good. This is because the height of c170 was different from those of interferences, and the 3-dimensionally arranged microphones enable the system to exploit this height difference.

We used three sets of parameters for function (30) specifying a mask for each time-frequency slot. The shapes of these functions are shown in Fig. 6. Table I shows that a smaller θ_T resulted in greater SIR improvements by T-F masking. However, some sounds with a small θ_T were unnatural. Table II shows that a smaller θ_T provided worse SDRs. Therefore, there is clearly a trade-off between SIR improvement and SDR. We observed by informal listening tests that in many cases parameter $(\theta_T, g) = (0.333\pi, 20)$ produced natural sounds with sufficient interference suppression. Some sound examples can be found on our web site [32].

Figure 8 shows example spectrograms. We can see that the target speech was enhanced to a certain degree only with ICA. With the combination of ICA and T-F masking, some residuals were eliminated and the harmonic structure appeared more clearly in the spectrogram (e.g. at time frames from 80 to 90). Moreover the active and inactive time of the target speech became clearer with T-F masking.

Figure 9 shows examples of envelopes and masks at 969 Hz. Although the ICA output was close to the target component at many time frames, some interference components were contained in the ICA output at some time frames (e.g. from 15 to 20). This shows the limitation of ICA as a spatial filter, namely that the number of interferences that can be eliminated is less than the number of sensors. The lower plot shows masks specified for each time frame. These masks show the activity of the target source at this frequency. By using these masks, the system eliminated the interference components at the time frames from 15 to 20, and improved the SIR at the output.

Figure 10 shows an example clustering result for normalized basis vectors. Of the $M = 4$ clusters, there was only one (the leftmost one) that had a smaller variance than $th_{var} = 0.015$ in this case. From this fact, the system decided that there was one dominant target source. This clustering results were also used to solve the permutation ambiguity of ICA based on (25).

C. Multi-target cases

Next we show the results for multi-target cases ($Q > 1$). Although we have performed experiments under various conditions, here we simply show the results for cases where three sources positioned at a120, b120 and c170 were dominant ($Q = 3$). Experiments were conducted with 10 combinations of 9 speeches (3 targets + 6 background interferences). The computational time was around 15 seconds for 6-second speech mixtures.

Table III shows the average SIR improvements obtained solely with ICA, and by a combination of ICA and T-F masking. Even with such hard input SIRs, the system succeeded in enhancing the target sources to a certain degree.

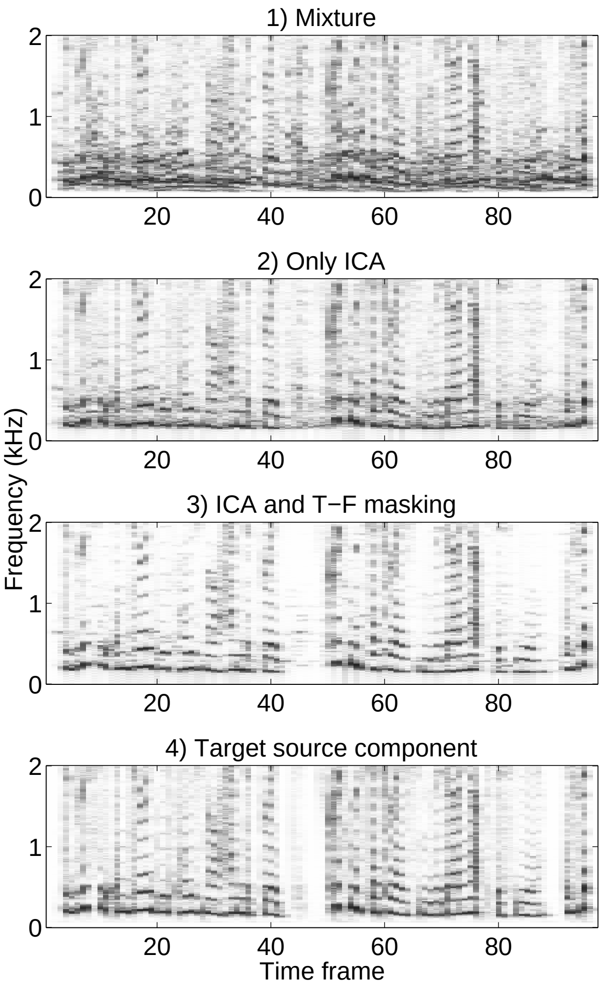

Fig. 8. Spectrograms for 1) a mixture $\mathrm{x}_J(t)$, 2) an output signal $\mathrm{y}_1(t)$ obtained only with ICA, 3) an output signal $\mathrm{y}_1(t)$ obtained by a combination of ICA and T-F masking, and 4) the target source component $x_{J1}(t)$

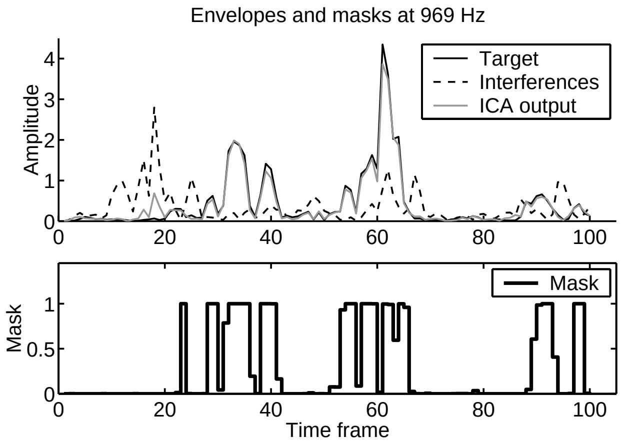

Fig. 9. Envelopes for a target signal, the total sum of all interferences, and the ICA output corresponding to the target signal (above), and masks calculated for each time frame (below).

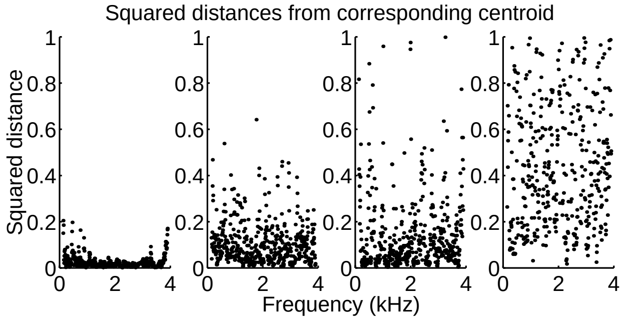

Fig. 10. Single-target clustering result for normalized basis vectors. Each point shows the squared distance $||\bar{\mathbf{a}} - \mathbf{c}_k||^2$ between a normalized basis vector and its corresponding centroid. The four clusters are sorted according to $\mathcal{J}_k/|C_k|$ as shown in (23).

TABLE III

AVERAGE SIR IMPROVEMENT FOR MULTI-TARGET CASES (dB)

Target position	a120	b120	c170
$\mathsf{InputSIR}_i$	-3.9	-3.6	-5.9
Only ICA	12.5	13.6	14.5
ICA and T-F masking $(0.333\pi, 20)$	15.1	16.5	17.6

Figure 11 shows an example clustering result for normalized basis vectors. In this multi-target case, there were three clusters that had smaller variances than $th_{var} = 0.015$, and the system decided that there were three dominant target sources. Again, these clustering results were also used to solve the permutation ambiguity of ICA.

VII. CONCLUSION

We have proposed a new method for extracting dominant target sources and suppressing other interference sources blindly. The method is based on a two-stage process where ICA is first applied to mixtures and then time-frequency masking is used to reduce residuals, which are caused by the limitation of ICA when $N > M$. The main contribution of this paper is to propose the following two new techniques:

1) Basis vector normalization and clustering, which decides the number Q of target sources and aligns the permutation ambiguity of ICA. The new method does not need to know the array geometry and therefore makes it easy to use a 3-dimensional non-uniform arrangement of sensors without exact measurement or calibration.
2) Specifying time-frequency masks from the angle $\theta_i(f,\tau)$ between the basis vector of a target source and a sample vector in a spatially whitened space. The angle indicates whether or not the corresponding target source is active at a time-frequency slot (f,τ).

Both techniques manipulate basis vectors $\mathbf{a}_i(f)$ produced by ICA, which is a statistical tool for blind processing.

We obtained good experimental results for extracting dominant sources out from many interference sources mixed in a real reverberant room. The experiments shown in this paper used impulse responses to evaluate extraction performances in terms of SIR and SDR. We also have tested the system in a live situation where loudspeakers and/or human speakers made

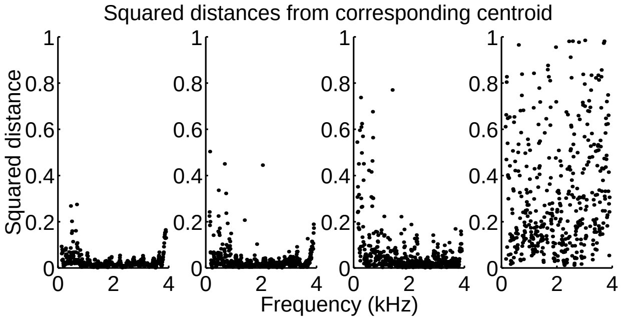

Fig. 11. Multi-target clustering result for normalized basis vectors (three dominant sources). Each point shows the squared distance $||\bar{\mathbf{a}}-\mathbf{c}_k||^2$ between a normalized basis vector and its corresponding centroid. The four clusters are sorted according to $\mathcal{J}_k/|C_k|$ as shown in (23).

speech sounds in a real room. The results were as good as the simulated ones where the impulse responses were used.

References

[1] T. W. Lee, *Independent Component Analysis - Theory and Applications*. Kluwer Academic Publishers, 1998.

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

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

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, 2002.

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

[6] 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.

[7] 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.

[8] L. Schobben and W. Sommen, "A frequency domain blind signal separation method based on decorrelation," *IEEE Trans. Signal Processing*, vol. 50, no. 8, pp. 1855–1865, Aug. 2002.

[9] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 3, pp. 204–215, May 2003.

[10] 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.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, Mar. 2005, pp. 299–327.

[12] ——, "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.

[13] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 539–548, Sept. 2004.

[14] 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*, vol. V, May 2004, pp. 1–4.

[15] 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.

[16] 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.

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

[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[19] 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*, vol. III, May 2004, pp. 881–884.

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

[21] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. ICA 2004 (LNCS 3195)*, Sept. 2004, pp. 832–839.

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

[23] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, Oct. 1994.

[24] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, the Massachusetts Institute of Technology, June 1996.

[25] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.

[26] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP 2000*, vol. 1, Oct. 2000, pp. 373–376.

[27] E. Bingham and A. Hyvärinen, "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.

[28] 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.

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

[30] 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.

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

[32] [Online]. Available: http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/dominant/

PLACE
PHOTO
HERE

Hiroshi Sawada (M'02–SM'04) received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively. In 1993, he joined NTT Communication Science Laboratories. From 1993 to 2000, he was engaged in research on the computer aided design of digital systems, logic synthesis, and computer architecture. Since 2000, he has been engaged in research on signal processing and blind source separation for convolutive mixtures using independent component analysis. He received the 9th TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 1994, and the Best Paper Award of the IEEE Circuit and System Society in 2000. Dr. Sawada is a senior member of the IEEE, a member of the IEICE and the ASJ.

PLACE
PHOTO
HERE

Shoko Araki (M'01) received the B.E. and the M.E. degrees in mathematical engineering and information physics from the University of Tokyo, Japan, in 1998 and 2000, respectively. In 2000, she joined NTT Communication Science Laboratories, Kyoto. Her research interests include array signal processing, blind source separation applied to speech signals, and auditory scene analysis. She received the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Best Paper Award of the IWAENC in 2003 and the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001. She is a member of the IEEE and the ASJ.

PLACE
PHOTO
HERE

Ryo Mukai (A'95–M'01–SM'04) received the B.S. and the M.S. degrees in information science from the University of Tokyo, Japan, in 1990 and 1992, respectively. He joined NTT in 1992. From 1992 to 2000, he was engaged in research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is engaged in research of blind source separation. His current research interests include digital signal processing and its applications. He received the Sato Paper Award of the Acoustical Society of Japan (ASJ) in 2005 and the Paper Award of the IEICE in 2005. He is a senior member of the IEEE, a member of the ACM, ASJ, IEICE, and IPSJ.

PLACE
PHOTO
HERE

Shoji Makino (A'89–M'90–SM'99–F'04) received the B. E., M. E., and Ph. D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively.

He joined NTT in 1981. He is now an Executive Manager at the NTT Communication Science Laboratories. He is also a Guest Professor at the Hokkaido University. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech.

He received the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or co-author of more than 200 articles in journals and conference proceedings and has been responsible for more than 150 patents.

He is a member of the Awards Board and a member of the Conference Board of the IEEE SP Society. He is an Associate Editor of the IEEE Transactions on Speech and Audio Processing and an Associate Editor of the EURASIP Journal on Applied Signal Processing. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society as well as the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is also a member of the International ICA Steering Committee and the Organizing Chair of the ICA2003 in Nara. He is the General Chair of the WASPAA2007 in Mohonk and the General Chair of the IWAENC2003 in Kyoto. He is the Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ.

He is an IEEE Fellow, a council member of the ASJ, and a member of the EURASIP, and a member of the IEICE.