Blind extraction of a dominant source from mixtures of many sources using ICA and time-frequency masking

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

***Abstract*— This paper presents a method for enhancing a target source of interest and suppressing other interference sources. The target source is assumed to be close to sensors, to have dominant power at these sensors, and to have non-Gaussianity. The enhancement is performed blindly, i.e. without knowing the total number of 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 process where independent component analysis (ICA) is first employed in each frequency bin and time-frequency masking is then used to improve the performance further. We propose a new sophisticated method for selecting the target source frequency components, and also a new criterion for specifying time-frequency masks. Experimental results for simulated cocktail party situations in a room (reverberation time was 130 ms) are presented to show the effectiveness and characteristics of the proposed method.**

I. Introduction

The technique for estimating individual source components from their mixtures at sensors is known as blind source separation (BSS) [1], [2]. 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 one source that is close to sensors, has dominant power, and/or has interesting features.

This paper presents a method for extracting a source signal of interest and suppressing other interference sources blindly. Let us formulate the task. Suppose that a target source s_1 and other interference sources $\mathrm{s}_2, \ldots, \mathrm{s}_N$ are convolutively mixed and observed at M sensors

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

where $\mathrm{h}_{jk}(l)$ represents the impulse response from source k to sensor j. The goal is to have an output signal $\mathrm{y}_1(t)$ that is close to the component of s_1 measured at a selected sensor J:

$$\mathrm{x}_{J1}(t) = \sum_l \mathrm{h}_{J1}(l)\, \mathrm{s}_1(t-l). \qquad (2)$$

Note that $\mathrm{x}_j(t) = \sum_{k=1}^{N} \mathrm{x}_{jk}(t)$. 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 the 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 that maximize an ICA criterion such as non-Gaussianity. We assume that the target source s_1 is non-Gaussian, close to sensors, and dominant in the mixtures. Therefore, we expect that one of the M components corresponds to s_1 whose ICA criterion is high.

We employ ICA in the time-frequency domain. The reason is that it is efficient and also fits time-frequency masking, which is discussed in the next paragraph. An additional operation that should be performed is the selection of the s_1 component in every frequency bin. This is considered to be the permutation problem of frequency-domain BSS. 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 [3]. However, this does not always work well for a case where the interferences are also speech. Thus, we exploit the information of basis vectors (8) produced by ICA. Our previously reported methods estimate the directions [4], [5] and/or the distances [5] of the sources from the basis vectors, and then cluster the estimations 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. II-C, we propose a new method for solving the permutation problem. 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 [3], [6] 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 [7]–[11], which is efficient for sources with 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 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 [7], [8]. Another recently proposed approach involves calculating the power ratio between an input and outputs of a spatial filter (beamformer [9], [11] or ICA [11]). However, such a power-based criterion depends on the scaling ambiguity of ICA or beamformer outputs. We propose a new criterion for specifying masks in Sec. II-D. It is based on the cosine distance between a sample vector and the basis vector corresponding to the target. The closeness 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.

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

II. The proposed method

A. Frequency domain operations

Figure 1 shows the flow of the method discussed 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):

$$x_j(f,\tau) \leftarrow \sum_{r=-L/2}^{L/2-1} \mathrm{x}_j(\tau+r)\, \mathrm{win}(r)\, e^{-j2\pi f r}, \qquad (3)$$

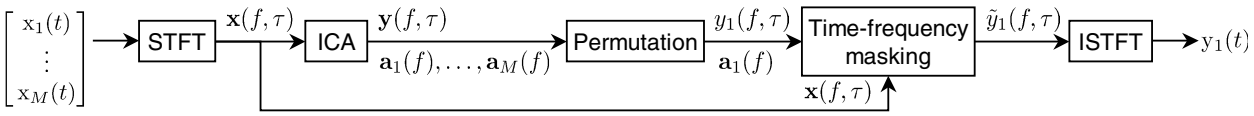

Fig. 1. Flow of proposed method

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 remaining 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:

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

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 $\mathrm{s}_k(t)$ obtained by the same operation as (3). The frequency-domain counterpart of (2) is

$$x_J(f,\tau) \approx h_{J1}(f) s_1(f,\tau), \tag{5}$$

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

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):

$$\mathrm{y}_1(\tau+r) \leftarrow \frac{1}{L\cdot\mathrm{win}(r)} \sum_{f\in\{0,\frac{1}{L}f_s,\ldots,\frac{L-1}{L}f_s\}} \tilde{y}_1(f,\tau)\, e^{\jmath 2\pi f r}.$$

B. Independent component analysis (ICA)

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

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

where $\mathbf{x} = [x_1,\ldots,x_M]^T$ is a sample vector and $\mathbf{h}_k = [h_{1k},\ldots,h_{Mk}]^T$ 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:

$$\mathbf{y}(f,\tau) = \mathbf{W}(f)\,\mathbf{x}(f,\tau), \tag{7}$$

where $\mathbf{y} = [y_1,\ldots,y_M]^T$ is a vector of independent components and $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_M]^H$ is an $M\times M$ separation matrix. In the experiments shown in Sec. III, 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].

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

$$[\mathbf{a}_1,\cdots,\mathbf{a}_M] = \mathbf{W}^{-1},\ \mathbf{a}_i = [a_{1i},\ldots,a_{Mi}]^T. \tag{8}$$

By multiplying both sides of (7) by $\mathbf{W}^{-1}$, the sample vector $\mathbf{x}(f,\tau)$ is represented by a linear combination of basis vectors $\mathbf{a}_1,\ldots,\mathbf{a}_M$:

$$\mathbf{x}(f,\tau) = \textstyle\sum_{i=1}^{M} \mathbf{a}_i(f) y_i(f,\tau). \tag{9}$$

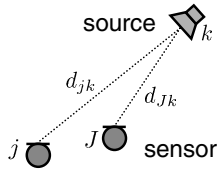

Fig. 2. Direct-path (nearfield) model

Since s_1 is assumed to be a dominant non-Gaussian source, it is strongly expected that one of $y_1,\ldots,y_M$ corresponds to s_1 and thus one of $\mathbf{a}_1,\ldots,\mathbf{a}_M$ corresponds to $\mathbf{h}_1$.

C. Permutation

The next operation is to find i for each frequency f such that $\mathbf{a}_i(f)$ corresponds to $\mathbf{h}_1(f)$. As shown in [4], 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 here. In the rest of this subsection, we discuss a new method for exploiting the basis vector information.

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

$$\bar{a}_{ji}(f) \leftarrow |a_{ji}(f)| \exp\left[\jmath \frac{\arg[a_{ji}(f)/a_{Ji}(f)]}{4fc^{-1}d_{\max}}\right] \tag{10}$$

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

$$\bar{\mathbf{a}}_i(f) \leftarrow \bar{\mathbf{a}}_i(f) \,/\, ||\bar{\mathbf{a}}_i(f)|| \tag{11}$$

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

The next step is 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

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

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

$$\mathcal{J} = \textstyle\sum_{k=1}^{M} \mathcal{J}_k,\ \mathcal{J}_k = \sum_{\bar{\mathbf{a}}\in C_k} ||\bar{\mathbf{a}} - \mathbf{c}_k||^2. \tag{12}$$

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 (nearfield) model (Fig. 2)

$$h_{jk}(f) \approx \frac{q(f)}{d_{jk}} \exp\left[\jmath 2\pi f c^{-1}(d_{jk} - d_{Jk})\right], \tag{13}$$

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

$$\mathbf{a}_i \approx \alpha_i \mathbf{h}_k, \quad a_{ji} \approx \alpha_i h_{jk}, \tag{14}$$

where α_i represents the scaling ambiguity, and index k, which may be different from index i, represents the permutation ambiguity. Substituting (13) and (14) into (10) and (11) yields

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

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

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

holds. This property is important for the distance measure (12), since $|\bar{a} - \bar{a}'|$ increases monotonically as $|\arg(\bar{a}) - \arg(\bar{a}')|$ increases.

Once we have found M clusters $C_1, \ldots, C_M$, we need to identify a cluster that corresponds to the target source s_1. We decide that a cluster C_K with the minimum variance $K = \mathrm{argmin}_k \, \mathcal{J}_k / |C_k|$ corresponds to s_1. The rationale behind this is that the mixing model (13) is more valid for s_1 than for the other sources. The direct-path components of impulse responses h_{j1} are distinct since s_1 is assumed to be close to the sensors. Finally, the output index i for each frequency f is selected by

$$I(f) = \mathrm{argmin}_i \, ||\bar{\mathbf{a}}_i(f) - \mathbf{c}_K||^2.$$

This means that basis vector $\mathbf{a}_{I(f)}(f)$ corresponds to $\mathbf{h}_1(f)$.

After we align the index as $\mathbf{a}_1(f) \leftarrow \mathbf{a}_{I(f)}(f)$ and $y_1(f,\tau) \leftarrow y_{I(f)}(f,\tau)$, we solve the scaling ambiguity in (9):

$$\mathbf{a}_1(f) y_1(f,\tau) = (\alpha_1 \mathbf{a}_1(f)) \, (y_1(f,\tau)/\alpha_1),$$

for any non-zero complex scalar α_1. This is easily solved by

$$y_1(f,\tau) \leftarrow a_{J1}(f) y_1(f,\tau),$$

where J is the index of the sensor specified in (5). The reason is as follows. The goal in each frequency bin is to make $y_1(f,\tau)$ as close to $x_{J1}(f,\tau)$ defined in (5) as possible. And we can derive relations

$$x_{J1}(f,\tau) \approx h_{J1}(f) s_1(f,\tau) \approx a_{J1}(f) y_1(f,\tau).$$

from (5), the $\mathbf{h}_1$ term in (6) and the $\mathbf{a}_1$ term in (9).

D. Time-frequency masking

Suppose that the permutation ambiguity of ICA is solved at this stage. The extraction of s_1 by ICA (7) is represented by

$$\begin{aligned} y_1(\tau) &= \mathbf{w}_1^H \mathbf{x}(\tau) \\ &= \mathbf{w}_1^H \mathbf{h}_1 s_1(\tau) + \textstyle\sum_{k=2}^{N} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau). \end{aligned}$$

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 $\mathcal{K} \subseteq \{2, \ldots, N\}$ such that $\mathbf{w}_1^H \mathbf{h}_k \neq 0, {}^\forall k \in \mathcal{K}$. Thus, $y_1(\tau)$ contains an unwanted residual $\sum_{k \in \mathcal{K}} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau)$. The purpose of time-frequency masking is to obtain another output $\tilde{y}_1(\tau)$ that contains less power of the residual $\sum_{k \in \mathcal{K}} \mathbf{w}_1^H \mathbf{h}_k s_k(\tau)$ than $y_1(\tau)$.

Time-frequency masking is performed by

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

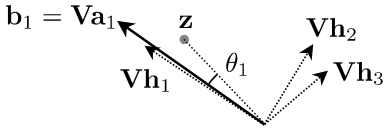

Fig. 3. Angle θ_1 calculated in whitened space

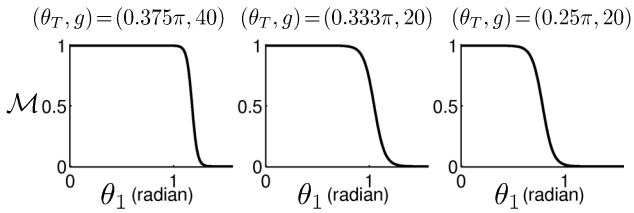

Fig. 4. Masking functions with three sets of parameters (θ_T, g)

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 in the whitened space. The angle is calculated by

$$\theta_1(f,\tau) = \arccos \frac{|\mathbf{b}_1^H(f) \cdot \mathbf{z}(f,\tau)|}{||\mathbf{b}_1(f)|| \cdot ||\mathbf{z}(f,\tau)||} \tag{15}$$

for each time-frequency slot (Fig. 3). Then, we calculate a mask by using a logistic function (Fig. 4)

$$\mathcal{M}(\theta_1(f,\tau)) = \frac{1}{1 + e^{g(\theta_1 - \theta_T)}}, \tag{16}$$

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 $\tilde{y}_1$ decreases but the musical noise in y_1 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, (6) can be approximated as

$$\mathbf{x}(f,\tau) \approx \mathbf{h}_k(f) s_k(f,\tau), \; k \in \{1, \ldots, N\}, \tag{17}$$

where k depends on each time-frequency slot (f,τ). Let us consider the whitened-space counterpart of (17), while distinguishing between cases where s_1 is the only active source (18) and other cases (19):

$$\mathbf{z}(f,\tau) \approx \mathbf{V}(f)\mathbf{h}_1(f)s_1(f,\tau) \approx \mathbf{V}(f)\mathbf{a}_1(f)y_1(f,\tau) \tag{18}$$

$$\mathbf{z}(f,\tau) \approx \textstyle\sum_{k=2}^{N} \mathbf{V}(f)\mathbf{h}_k(f)s_k(f,\tau). \tag{19}$$

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 $\mathbf{b}_1 = \mathbf{V}\mathbf{a}_1$ of a dominant source s_1, which points in almost the same direction as $\mathbf{V}\mathbf{h}_1$, tends to have large angles with the other vectors $\mathbf{V}\mathbf{h}_2, \ldots, \mathbf{V}\mathbf{h}_N$. Figure 3 shows such a case. Therefore, calculating the angle (15) provides information about whether or not s_1 is the only active source at a time-frequency slot (f,τ).

III. Experiments

We performed experiments to enhance a dominant speech that was close to microphones. We measured impulse responses $\mathrm{h}_{jk}(l)$ under the conditions shown in Fig. 5. The speaker positions simulated a cocktail party situation. Mixtures at the microphones were made by convolving the impulse responses and 6-second English speeches sampled at 8 kHz. Microphone arrangement was 3-dimensional

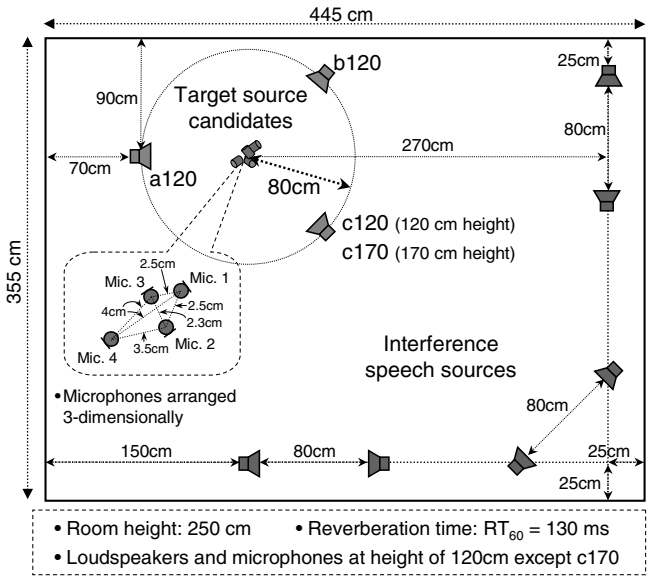

Fig. 5. Experimental conditions

and non-uniform. The system knew only the maximum distance (4 cm) between the reference microphone (Mic. 1) and others. For each setup, we selected one of the four speakers (a120, b120, c120, c170) as a dominant target source, and the others were kept silent. The six speakers away from the microphones were used as interferences for every setup. The frame size L of STFT (3) was 1024 (128 ms). 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. The performance was evaluated in terms of the signal-to-interference ratio (SIR) improvement, which is OutputSIR − InputSIR. These two types of SIRs are defined by

$$\mathsf{InputSIR} = 10\log_{10}\frac{\langle|\mathrm{x}_{J1}(t)|^2\rangle_t}{\langle|\sum_{k\neq 1}\mathrm{x}_{Jk}(t)|^2\rangle_t} \quad \text{(dB)},$$

$$\mathsf{OutputSIR} = 10\log_{10}\frac{\langle|\mathrm{y}_{11}(t)|^2\rangle_t}{\langle|\sum_{k\neq 1}\mathrm{y}_{1k}(t)|^2\rangle_t} \quad \text{(dB)},$$

where $\mathrm{x}_{Jk}(t)$ is defined in (2), and $\mathrm{y}_{1k}(t)$ is the component of s_k that appears at output $\mathrm{y}_1(t)$, i.e. $\mathrm{y}_1(t) = \sum_{k=1}^{N}\mathrm{y}_{1k}(t)$.

Experiments were conducted with 16 combinations of 7 speeches for each target position. 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 depend on the target position. 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 the height difference.

We used three sets of parameters for function (16) specifying a mask for each time-frequency slot. The shapes of these functions are shown in Fig. 4. 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. We observed 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 [16].

TABLE I

AVERAGE SIR IMPROVEMENT FOR EACH POSITION (dB)

Target position	a120	b120	c120	c170
InputSIR	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

IV. CONCLUSION

We have presented a method for extracting a dominant target source and suppressing other interferences. The process of ICA and following permutation alignment extracts the target source, and estimates the corresponding basis vector. The new method for permutation alignment makes it easy to use a 3-dimensional non-uniform arrangement of sensors without exact measurement or calibration. Time-frequency masking in the second stage reduces the power of the residuals caused by ICA when $N > M$. It exploits the sparseness of sources. We have proposed a new criterion for specifying masks. It is based on the angle between the target basis vector and a sample vector, and gives information about whether or not the target source is active. The experiments showed good results for extracting a dominant source out from six interferences mixed in a real room.

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*, vol. I, May 2004, 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, no. 5, 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*, vol. V, May 2004, 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] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. ICASSP 2004*, vol. II, May 2004, pp. 373–376.

[10] 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.

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

[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*, 2nd ed. Wiley Interscience, 2000.

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