Grouping Separated Frequency Components by Estimating Propagation Model Parameters in Frequency-Domain Blind Source Separation

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

***Abstract*— This paper proposes a new formulation and optimization procedure for grouping frequency components in frequency-domain blind source separation (BSS). We adopt two separation techniques, independent component analysis (ICA) and time-frequency (T-F) masking, for the frequency-domain BSS. With ICA, grouping the frequency components corresponds to aligning the permutation ambiguity of the ICA solution in each frequency bin. With T-F masking, grouping the frequency components corresponds to classifying sensor observations in the time-frequency domain for individual sources. The grouping procedure is based on estimating anechoic propagation model parameters by analyzing ICA results or sensor observations. More specifically, the time delays of arrival and attenuations from a source to all sensors are estimated for each source. The focus of this paper includes the applicability of the proposed procedure for a situation with wide sensor spacing where spatial aliasing may occur. Experimental results show that the proposed procedure effectively separates two or three sources with several sensor configurations in a real room, as long as the room reverberation is moderately low.**

***Index Terms*— Blind source separation, convolutive mixture, frequency domain, independent component analysis, permutation problem, sparseness, time-frequency masking, time delay estimation, generalized cross correlation**

I. Introduction

The technique for estimating individual source components from their mixtures at multiple sensors is known as blind source separation (BSS) [3]–[6]. With acoustic applications of BSS, such as solving a cocktail party problem, signals are generally mixed in a convolutive manner with reverberations. Let $\mathrm{s}_1, \ldots, \mathrm{s}_N$ be source signals and $\mathrm{x}_1, \ldots, \mathrm{x}_M$ be sensor observations. The convolutive mixture model is formulated as

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

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

An efficient and practical approach for such convolutive mixtures is frequency-domain BSS [7]–[25], where we apply a short-time Fourier transform (STFT) to the sensor observations $\mathrm{x}_j(t)$. In the frequency domain, the convolutive mixture (1) can be approximated as an instantaneous mixture at each frequency:

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

where f represents frequency, $h_{jk}(f)$ is the frequency response from source k to sensor j, and $s_k(f,t)$ is the time-frequency representation of a source signal s_k.

Independent component analysis (ICA) [3]–[6] is a major statistical tool for BSS. With the frequency-domain approach, ICA is employed in each frequency bin with the instantaneous mixture model (2). This makes the convergence of ICA stable and fast. However, the permutation ambiguity of the ICA solution in each frequency bin should be aligned so that the frequency components of the same source are grouped together. This is known as the permutation problem of frequency-domain BSS. Various methods have been proposed to solve this problem. Early work [7], [8] considered the smoothness of the frequency response of separation filters. For non-stationary sources such as speech, it is effective to exploit the mutual dependence of separated signals across frequencies either with simple second order correlation [9]–[12] or with higher order statistics [17], [18].

Spatial information of sources is also useful for the permutation problem, such as the direction-of-arrival of a source [12]–[14] or the ratio of the distances from a source to two sensors [15]. Our recent work [16] generalizes these methods so that the two types of geometrical information (direction and distance) are treated in a single scheme and also the BSS system does not need to know the sensor array geometry. When we are concerned with the directions of sources, we generally prefer the sensor spacing to be no larger than half the minimum wavelength of interest to avoid the effect of spatial aliasing [26]. We typically use 4 cm sensor spacing for an 8 kHz sampling rate. However, there are cases where widely spaced sensors are used to achieve better separation for low frequencies. Or, if we increase the sampling rate, for example up to 16 kHz, to obtain better speech recognition accuracy for separated signals, spatial aliasing occurs even with 4 cm spacing. If spatial aliasing occurs at high frequencies, the ICA

Earlier versions of this work were presented in [1] and [2] as conference papers. 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-SSEN, AUD-LMAP solutions in these frequencies imply multiple possibilities for a source direction. Such a problem is troublesome for frequency-domain BSS as previously pointed out [14], [27].

There is another method for frequency-domain BSS, which is based on time-frequency (T-F) masking [19]–[23]. It does not employ ICA to separate mixtures, but relies on the sparseness of source signals exhibited in time-frequency representations. The method groups sensor observations together for each source based on spatial information extracted from them. In [22], we applied a technique similar to that used with ICA [16] to classify sensor observations for T-F masking separation. From this experience, we consider the two methods, ICA-based separation and T-F masking separation, to be very similar in terms of exploiting the spatial information of sources.

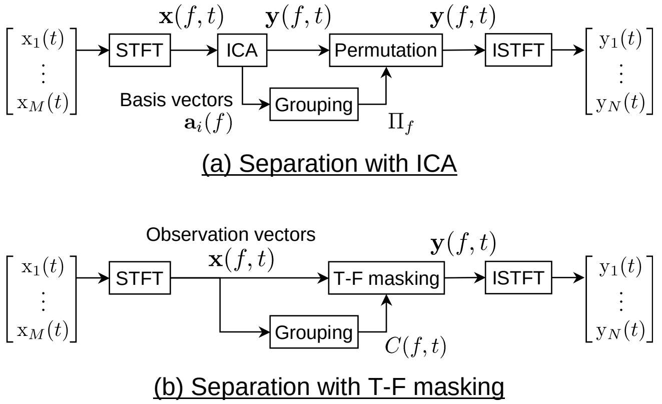
Based upon the above review of previous work and related methods, this paper proposes a new formulation and optimization procedure for grouping frequency components in the context of frequency-domain BSS. Grouping frequency components corresponds to solving the permutation problem in ICA-based separation, and to classifying sensor observations in T-F masking separation. In the formulation, we use relative time delays and attenuations from sources to sensors as parameters to be estimated. The idea of parameterizing time delays and attenuations has already been proposed in previous studies [20], [21], [24], where only simple two-sensor cases were considered without the possibility of spatial aliasing. The novelty of this paper compared with these previous studies and our recent work [16], [22] can be summarized as follows:

1) Two methods of ICA-based separation and T-F masking separation are considered uniformly in terms of grouping frequency components.
2) The problem of spatial aliasing is solved by the proposed procedure, not only for ICA-based separation but also for T-F masking separation, thanks to 1).
3) It is shown that the time delay parameters in the formulation are estimated with a function similar to the Generalized Cross Correlation PHAse Transform (GCC-PHAT) function [23], [28]–[30].

And the proposed procedure inherits the attractive properties of our recently proposed approaches [16], [22]:

4) The procedure can be applied to any number of sensors, and is not limited to two sensors.
5) The complete sensor array geometry does not have to be known, only the information about the maximum distance between sensors. If the complete geometry were known, the location (direction and/or distance from the sensors) of each source could be estimated [31], [32].

This paper is organized as follows. The next section provides an overview of frequency-domain BSS. It includes both the ICA-based method and the T-F masking method. Section III presents an anechoic propagation model with the time delays and attenuations from a source to sensors, and also cost functions for grouping frequency components. Section IV proposes a procedure for optimizing the cost function for permutation alignment in ICA-based separation. Section V shows a similar optimization procedure for classifying sensor observations in T-F masking separation, together with the relationship with the GCC-PHAT function. Experimental results for various setups are summarized in Sec. VI. Section VII concludes this paper.



(a) Separation with ICA

(b) Separation with T-F masking

Fig. 1. System structure of frequency-domain BSS. We consider two methods for separating the mixtures, (a) ICA and (b) T-F masking. For both methods, grouping frequency components, basis vectors or observation vectors, is the key technique discussed in this paper.

II. Frequency-domain BSS

This section presents an overview of frequency-domain BSS. Figure 1 shows the system structure. First, the sensor observations (1) sampled at frequency f_s are converted into frequency-domain time-series signals (2) by a short-time Fourier transform (STFT) of frame size L:

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

for all discrete frequencies $f \in \{0, \frac{1}{L}f_s, \ldots, \frac{L-1}{L}f_s\}$, and for time t, which is now down-sampled with the distance of the frame shift. We denote the imaginary unit as $\imath = \sqrt{-1}$ in this paper. We typically use a window $\mathrm{win}(q)$ that tapers smoothly to zero at each end, such as a Hanning window $\mathrm{win}(q) = \frac{1}{2}(1 + \cos\frac{2\pi q}{L})$.

Let us rewrite (2) in a vector notation:

$$\mathbf{x}(f,t) = \sum_{k=1}^{N} \mathbf{h}_k(f) s_k(f,t), \qquad (4)$$

where $\mathbf{h}_k = [h_{1k}, \ldots, h_{Mk}]^T$ is the vector of frequency responses from source s_k to all sensors, and $\mathbf{x} = [x_1, \ldots, x_M]^T$ is called an observation vector in this paper. We consider two methods for separating the mixtures as shown in Fig. 1. They are described in the following two subsections. In either case, we can limit the set of frequencies $\mathcal{F}$ where the operation is performed by

$$\mathcal{F} = \{0, \frac{1}{L}f_s, \ldots, \frac{1}{2}f_s\} \qquad (5)$$

due to the relationship of the complex conjugate:

$$x_j(\tfrac{n}{L}f_s, t) = x_j^*(\tfrac{L-n}{L}f_s, t)\,, \quad n = 1, \ldots, \tfrac{L}{2}-1\,. \qquad (6)$$

A. Independent Component Analysis (ICA)

The first method employs complex-valued instantaneous ICA in each frequency bin $f \in \mathcal{F}$:

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

where $\mathbf{y} = [y_1, \ldots, y_N]^T$ is the vector of separated frequency components and $\mathbf{W}$ is an $N \times M$ separation matrix. There are many ICA algorithms known in the literature [3]–[6]. We do not describe these ICA algorithms in detail. More importantly, here let us explain how to estimate the mixing situation, such as (4), from the ICA solution. We calculate a matrix $\mathbf{A}$ whose columns are basis vectors $\mathbf{a}_i$,

$$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N],\ \mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T, \tag{8}$$

in order to represent the vector $\mathbf{x}$ by a linear combination of the basis vectors:

$$\mathbf{x}(f,t) = \mathbf{A}(f)\,\mathbf{y}(f,t) = \sum_{i=1}^{N} \mathbf{a}_i(f) y_i(f,t)\,. \tag{9}$$

If $\mathbf{W}$ has an inverse, the matrix $\mathbf{A}$ is given simply by the inverse $\mathbf{A} = \mathbf{W}^{-1}$. Otherwise it is calculated as a least-mean-square estimator [33]

$$\mathbf{A} = \mathrm{E}\{\mathbf{x}\mathbf{y}^H\}(\mathrm{E}\{\mathbf{y}\mathbf{y}^H\})^{-1}\,,$$

which minimizes $\mathrm{E}\{||\mathbf{x} - \mathbf{A}\mathbf{y}||^2\}$. The above procedure is effective only when there are enough sensors ($N \leq M$). Under-determined ICA ($N > M$) is still difficult to solve, and we do not usually follow the above procedure, but directly estimate basis vectors $\mathbf{a}_i(f)$, as shown in e.g. [25].

In any case, if ICA works well, we expect the separated components $y_1(f,t), \ldots, y_N(f,t)$ to be close to the original source components $s_1(f,t), \ldots, s_N(f,t)$ up to permutation and scaling ambiguity. Based on this, we see that a basis vector $\mathbf{a}_i(f)$ in (9) is close to $\mathbf{h}_k(f)$ in (4) again up to permutation and scaling ambiguity. The use of different subscripts, i and k, indicates the permutation ambiguity. They should be related by a permutation $\Pi_f : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\}$ for each frequency bin f as

$$i = \Pi_f(k) \tag{10}$$

so that the separated components y_i originating from the same source s_k are grouped together. Section IV presents a procedure for deciding a permutation Π_f for each frequency. After permutations have been calculated, separated frequency components and basis vectors are updated by

$$y_k(f,t) \leftarrow y_{\Pi_f(k)}(f,t),\ \ \mathbf{a}_k(f) \leftarrow \mathbf{a}_{\Pi_f(k)}(f),\ \ \forall\, k, f, t. \tag{11}$$

Next, the scaling ambiguity of ICA solution is aligned. The exact recovery of the scaling corresponds to blind dereverberation [34], [35], which is a challenging task especially for colored sources such as speech. A much easier way has been proposed in [10], [11], [36], which involves adjusting to the observation $x_J(f,t)$ of a selected reference sensor $J \in \{1, \ldots, M\}$:

$$y_k(f,t) \leftarrow a_{Jk}(f) y_k(f,t),\ \ \forall\, k, f, t. \tag{12}$$

We see in (9) that $a_{Jk}(f) y_k(f,t)$ is a part of $x_J(f,t)$ that originates from source s_k.

Finally, time-domain output signals $\mathrm{y}_k(t)$ are calculated with an inverse STFT (ISTFT) to the separated frequency components $y_k(f,t)$.

B. Time-Frequency (T-F) Masking

The second method considered in this paper is based on T-F masking, in which we assume the sparseness of source signals, i.e., at most only one source makes a large contribution to each time-frequency observation $\mathbf{x}(f,t)$. Based on this assumption, the mixture model (4) can simply be approximated as

$$\mathbf{x}(f,t) = \mathbf{h}_k(f) s_k(f,t),\ \ k \in \{1, \ldots, N\} \tag{13}$$

where the index k of the dominant source depends on each time-frequency slot (f,t).

The method classifies observation vectors $\mathbf{x}(f,t)$ of all time-frequency slots (f,t) into N classes so that the k-th class consists of mixtures where the k-th source is the dominant source. The notation

$$C(f,t) = k \tag{14}$$

is used to represent a situation that an observation vector $\mathbf{x}(f,t)$ belongs to the k-th class. Section V provides a procedure for classifying observation vectors $\mathbf{x}$. Once the classification is completed, time domain separated signals $\mathrm{y}_k(t)$ are calculated with an inverse STFT (ISTFT) to the following classified frequency components

$$y_k(f,t) = \begin{cases} x_J(f,t) & \text{if } C(f,t) = k, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

C. Relationship between ICA based and T-F Masking Methods

As mentioned in the Introduction, this paper handles the cases of ICA and T-F masking uniformly in terms of grouping frequency components. Let us discuss the relationship between the two [1]. If the approximation (13) in T-F masking is satisfied, the linear combination form (9) obtained by ICA is reduced to

$$\mathbf{x}(f,t) = \mathbf{a}_i(f) y_i(f,t),\ \ i \in \{1, \ldots, N\} \tag{16}$$

where i depends on each time-frequency slot (f,t). Thus, the spatial information expressed in an observation vector $\mathbf{x}(f,t)$ with the approximation (13) is the same as that of the basis vector $\mathbf{a}_i(f)$ up to scaling ambiguity, with $y_i(f,t)$ being dominant in the time-frequency slot. Therefore, we can use similar techniques for extracting spatial information from observation vectors $\mathbf{x}$ and basis vectors $\mathbf{a}_i$.

III. Propagation Model and Cost Functions

A. Problem Statement

The problem of grouping frequency components considered in this paper is stated as follows:

> Classify all basis vectors $\mathbf{a}_i(f)$, $\forall\, i, f$ or all observation vectors $\mathbf{x}(f,t)$, $\forall\, f, t$ into N groups so that each

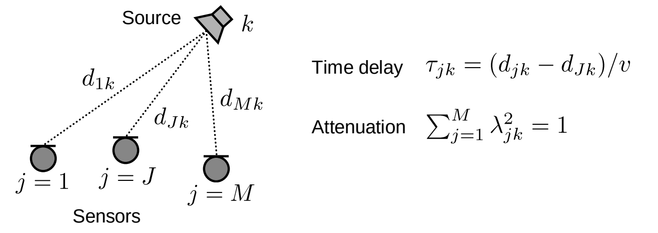

Source
k
d_{1k}
d_{Jk}
d_{Mk}
Time delay $\tau_{jk} = (d_{jk} - d_{Jk})/v$
Attenuation $\sum_{j=1}^{M} \lambda_{jk}^2 = 1$
$j = 1$
$j = J$
$j = M$
Sensors

Fig. 2. Anechoic propagation model with the time delay τ_{jk} and the attenuation λ_{jk} from source k to sensor j. The time delay τ_{jk} depends on the distance d_{jk} from source k to sensor j, and is normalized with the distance d_{Jk} of a selected reference sensor $J \in \{1, \ldots, M\}$. The attenuation λ_{jk} has no explicit dependence on the distance, and is normalized so that the squared sum over all the sensors is 1.

group consists of frequency components originating from the same source.

Solving this problem corresponds to deciding permutations Π_f in ICA-based separation, and to obtaining classification information $C(f,t)$ in T-F masking separation, respectively.

As discussed in the previous section, from (4) and (9), basis vectors $\mathbf{a}_1(f), \ldots, \mathbf{a}_N(f)$ obtained by ICA are close to $\mathbf{h}_1(f), \ldots, \mathbf{h}_N(f)$ up to permutation and scaling ambiguity. Also from (13), an observation vector $\mathbf{x}(f,t)$ is a scaled version of $\mathbf{h}_k(f)$ with k being specific to the time-frequency slot (f,t). Therefore, we see that modeling the vector $\mathbf{h}_k(f)$ of frequency responses is an important issue as regards solving the grouping problem.

B. Propagation Model with Time Delays and Attenuations

We model the propagation from a source to a sensor with the time delay and attenuation (Fig. 2), i.e., with an anechoic model. This model considers only direct paths from sources to sensors, even though in reality signals are mixed in a multi-path manner (1) with reverberations. Such an anechoic assumption has been used in many previous studies exploiting spatial information of sources, some of which are enumerated in the Introduction. As shown by the experimental results in Sec. VI, modeling only direct paths is still effective for a real room situation as long as the room reverberation is moderately low. With this model, we approximate the frequency response $h_{jk}(f)$ in (2) with

$$c_{jk}(f) = \lambda_{jk} \cdot \exp(-\imath\, 2\pi f \tau_{jk}), \tag{17}$$

where τ_{jk} and $\lambda_{jk} > 0$ are the time delay and attenuation from source k to sensor j, respectively. In the vector form, $\mathbf{h}_k(f)$ in (4) is approximated with

$$\mathbf{c}_k(f) = \begin{bmatrix} \lambda_{1k} \cdot \exp(-\imath\, 2\pi f \tau_{1k}) \\ \vdots \\ \lambda_{Mk} \cdot \exp(-\imath\, 2\pi f \tau_{Mk}) \end{bmatrix}. \tag{18}$$

Since we cannot distinguish the phase (or amplitude) of $s_k(f,t)$ and $h_{jk}(f)$ of the mixture (2) in a blind scenario, the two types of parameters τ_{jk} and λ_{jk} can be considered to be relative. Thus, without loss of generality, we normalize them by

$$\tau_{jk} = (d_{jk} - d_{Jk})/v, \tag{19}$$

$$\textstyle\sum_{j=1}^{M} \lambda_{jk}^2 = 1, \tag{20}$$

where d_{jk} is the distance from source k to sensor j (Fig. 2), and v is the propagation velocity of the signal. Normalization (19) makes $\tau_{Jk} = 0$ and $\arg(c_{Jk}) = 0$, i.e., the relative time delay is zero at a selected reference sensor $J \in \{1, \ldots, M\}$. Normalization (20) makes the model vector $\mathbf{c}_k$ have unit-norm $||\mathbf{c}_k|| = 1$.

If we do not want to treat reference sensor J as a special case, we normalize the time delay in a more general way:

$$\tau_{jk} = (d_{jk} - d_{\mathrm{pair}(j)k})/v, \tag{21}$$

where $\mathrm{pair}(j) \neq j$ is the sensor that is pairing with sensor j. We can arbitrarily specify the $\mathrm{pair}(\cdot)$ function . An example is a simple pairing with the next sensor:

$$\mathrm{pair}(j) = \begin{cases} 1 & \text{if } j = M, \\ j+1 & \text{otherwise.} \end{cases} \tag{22}$$

In either case, the normalized time delay τ_{jk} can now be considered as the time difference of arrival (TDOA) [30], [31] of source s_k between sensor j and sensor J or $\mathrm{pair}(j)$.

C. Phase & Amplitude Normalization

As mentioned in Sec. III-A, basis vectors $\mathbf{a}_i$ and observation vectors $\mathbf{x}$ have scaling (phase and amplitude) ambiguity. To align the ambiguity, we apply the same kind of normalization as discussed in the previous subsection, and then obtain phase/amplitude normalized vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{x}}$.

As regards phase ambiguity, if we follow (19), we apply

$$\tilde{\mathbf{a}}_i \leftarrow \mathbf{a}_i \cdot \exp[-\imath \arg(a_{Ji})]\,, \text{ or} \tag{23}$$

$$\tilde{\mathbf{x}} \leftarrow \mathbf{x} \cdot \exp[-\imath \arg(x_J)] \tag{24}$$

leading to $\arg(\tilde{a}_{Ji}) = 0$ or $\arg(\tilde{x}_J) = 0$. If we prefer (21), we apply

$$\tilde{a}_{ji} \leftarrow a_{ji} \cdot \exp[-\imath \arg(a_{\mathrm{pair}(j)i})]\,, \text{ or} \tag{25}$$

$$\tilde{x}_j \leftarrow x_j \cdot \exp[-\imath \arg(x_{\mathrm{pair}(j)})], \tag{26}$$

for $j = 1, \ldots, M$ to construct $\tilde{\mathbf{a}}_i = [\tilde{a}_{1i}, \ldots, \tilde{a}_{Mi}]^T$ or $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_M]^T$. Next, the amplitude ambiguity is aligned based on (20) by

$$\tilde{\mathbf{a}}_i \leftarrow \tilde{\mathbf{a}}_i \,/\, ||\tilde{\mathbf{a}}_i||\,, \text{ or} \tag{27}$$

$$\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} \,/\, ||\tilde{\mathbf{x}}|| \tag{28}$$

leading to $||\tilde{\mathbf{a}}_i|| = 1$ or $||\tilde{\mathbf{x}}|| = 1$.

D. Cost Functions

Given that the phase and amplitude are normalized according to the above procedures, the task for grouping frequency components can be formulated as minimizing a cost function.

With ICA-based separation, the task is to determine a permutation Π_f for each frequency $f \in \mathcal{F}$ that relates the subscripts i and k with (10), and to estimate parameters τ_{jk}, λ_{jk} in the model (18) so that the cost function is minimized:

$$\mathcal{D}_{\mathbf{a}}(\{\tau_{jk}\}, \{\lambda_{jk}\}, \{\Pi_f\}) = \sum_{k=1}^{N} \sum_{f \in \mathcal{F}} ||\tilde{\mathbf{a}}_i(f) - \mathbf{c}_k(f)||^2 \,\big|_{i=\Pi_f(k)} \tag{29}$$

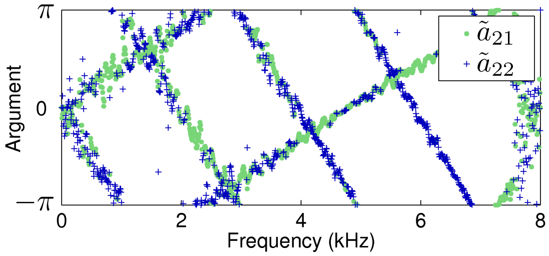


Fig. 3. Arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ before permutation alignment.

where $\{\tau_{jk}\}$ denotes the set $\{\tau_{11},\ldots,\tau_{MN}\}$ of time delay parameters, and similarly for $\{\lambda_{jk}\}$ and $\{\Pi_f\}$.

With T-F masking separation, the task is to determine classification $C(f,t)$ defined in (14) for each time-frequency slot, and to estimate parameters τ_{jk}, λ_{jk} in the model (18) so that the cost function is minimized:

$$\mathcal{D}_{\mathbf{x}}(\{\tau_{jk}\},\{\lambda_{jk}\},C) = \sum_{k=1}^{N} \sum_{C(t,f)=k} ||\tilde{\mathbf{x}}(f,t) - \mathbf{c}_k(f)||^2, \tag{30}$$

where the right-hand summation is across all the time-frequency slots (f,t) that belong to the k-th class.

The cost function $\mathcal{D}_{\mathbf{a}}$ or $\mathcal{D}_{\mathbf{x}}$ can become zero if 1) the real mixing situation follows the assumed anechoic model (17) perfectly and 2) the ICA is perfectly solved or the sparseness assumption (13) is satisfied in a T-F masking case. However, in real applications, none of these conditions is perfectly satisfied. Thus, these cost functions end up with a positive value, which corresponds to the variance in the mixing situation modeling. Yet minimizing them provides a solution to the grouping problem stated in Sec. III-A.

E. Simple Example

To make the discussion here intuitively understandable, let us show a simple example performed with setup A. We have three setups (A, B and C) shown in Fig. 9, and their common experimental configurations are summarized in Table I. Setup A was a simple $M = N = 2$ case, but the sensor spacing was 20 cm, which induced spatial aliasing for a 16 kHz sampling rate. The example here is with ICA-based separation, and Fig. 3 shows the arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ after the normalization (23) where we set $J = 1$ as a reference sensor. The arguments of $\tilde{a}_{1i}$ are not shown because they are all zero. The time delays τ_{21} and τ_{22} can be estimated from these data, as we see the two lines with different slopes corresponding to τ_{21} and τ_{22}. However, the following two factors complicate the time delay estimation. The first is that different symbols ('•' and '+') constitute each of the two lines, because of the permutation ambiguity of the ICA solutions. The second is the circular jumps of the lines at high frequencies, which are due to phase wrapping caused by spatial aliasing. We will explain how to group such frequency components in the next section.

IV. Permutation Alignment for ICA Results

This section presents a procedure for minimizing the cost function $\mathcal{D}_{\mathbf{a}}$ in (29), and for obtaining a permutation Π_f for each frequency. Figure 4 shows the flow of the procedure. We adopt an approach that first considers only the frequency range where spatial aliasing does not occur, and then considers the whole range $\mathcal{F}$.

A. For Frequencies without Spatial Aliasing

Let us first consider the lower frequency range

$$\mathcal{F}_L = \{f : -\pi < 2\pi f \tau_{jk} < \pi, \forall j,k\} \cap \mathcal{F} \tag{31}$$

where we can guarantee that spatial aliasing does not occur. Let $d_{\max}$ be the maximum distance between the reference sensor J and any other sensor if we take (19), or between sensor pairs of j and $\mathrm{pair}(j)$ if we take (21). Then the relative time delay is bounded by

$$\max_{j,k} |\tau_{jk}| \le d_{\max}/v \tag{32}$$

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

$$\mathcal{F}_L = \{f : 0 < f < \frac{v}{2\,d_{\max}}\} \cap \mathcal{F}\,. \tag{33}$$

For the frequency range $\mathcal{F}_L$, appropriate permutations Π_f can be obtained by minimizing another cost function

$$\bar{\mathcal{D}}_{\mathbf{a}}(\{\tau_{jk}\},\{\lambda_{jk}\},\{\Pi_f\}) = \sum_{k=1}^{N} \sum_{f\in\mathcal{F}_L} ||\bar{\mathbf{a}}_i(f) - \bar{\mathbf{c}}_k||^2 \,\Big|_{i=\Pi_f(k)} \tag{34}$$

as proposed in our previous work [16]. The cost function $\bar{\mathcal{D}}_{\mathbf{a}}$ is different from (29) in that $\bar{\mathbf{a}}_i(f)$ and $\bar{\mathbf{c}}_k$ are frequency normalized versions of basis vectors and the model vector. They are obtained by a procedure that divides their elements' argument by a scalar proportional to the frequency:

$$\bar{\mathbf{a}}_i(f) = [\bar{a}_{1i}(f),\ldots,\bar{a}_{Mi}(f)]^T\,,$$
$$\bar{a}_{ji}(f) \leftarrow |\tilde{a}_{ji}(f)| \exp\left(\imath\,\frac{\beta \arg[\tilde{a}_{ji}(f)]}{f}\right) \tag{35}$$

and

$$\bar{\mathbf{c}}_k = \begin{bmatrix} \bar{c}_{1k} \\ \vdots \\ \bar{c}_{Mk} \end{bmatrix} = \begin{bmatrix} \lambda_{1k}\cdot\exp(-\imath 2\pi\beta\tau_{1k}) \\ \vdots \\ \lambda_{Mk}\cdot\exp(-\imath 2\pi\beta\tau_{Mk}) \end{bmatrix}. \tag{36}$$

where β is a constant scalar (its role will be discussed afterwards). Since the original model (17) has a linear phase, the above procedure removes the frequency dependency so that the resultant model vector $\bar{\mathbf{c}}_k$ does not depend on frequency.

The advantage of introducing the frequency-normalized cost function $\bar{\mathcal{D}}_{\mathbf{a}}$ is that it can be minimized efficiently by the following clustering algorithm similar to the k-means algorithm [37]. The algorithm iterates the following two updates until convergence:

$$\Pi_f \leftarrow \mathrm{argmin}_{\Pi} \sum_{k=1}^{N} ||\bar{\mathbf{a}}_{\Pi(k)}(f) - \bar{\mathbf{c}}_k||^2, \ \ \forall f \in \mathcal{F}_L, \tag{37}$$

$$\bar{\mathbf{c}}_k \leftarrow \frac{1}{|\mathcal{F}_L|} \sum_{f\in\mathcal{F}_L} \bar{\mathbf{a}}_i(f)\,\Big|_{i=\Pi_f(k)},\ \ \bar{\mathbf{c}}_k \leftarrow \bar{\mathbf{c}}_k/||\bar{\mathbf{c}}_k||, \ \ \forall k \tag{38}$$

where $|\mathcal{F}_L|$ is the number of elements (cardinality) of the set. The first update (37) optimizes the permutation Π_f for

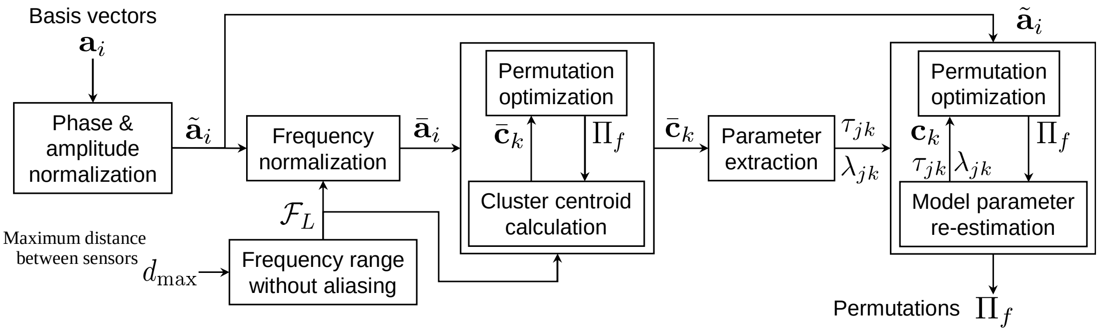


Fig. 4. Flow of the permutation alignment procedure presented in Sec. IV, which corresponds to the grouping part of (a) separation with ICA in Fig. 1.

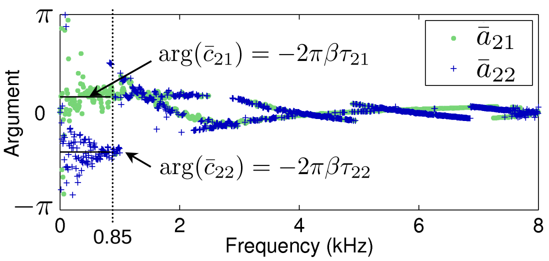


Fig. 5. Arguments of $\bar{a}_{21}$ and $\bar{a}_{22}$ after permutations are aligned only for frequency range $\mathcal{F}_L = \{f : 0 < f < 850 \text{ Hz}\} \cap \mathcal{F}$.

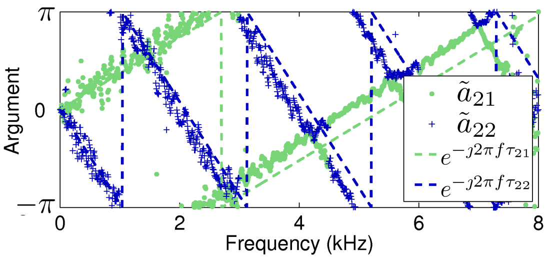


Fig. 6. Arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ after permutation alignment using model parameters estimated with low frequency range $\mathcal{F}_L$ data. Because τ_{21} and τ_{22} are not precisely estimated, there are some permutation errors at high frequencies.

each frequency with the current model $\bar{\mathbf{c}}_k$. The second update (38) calculates the most probable model $\bar{\mathbf{c}}_k$ with the current permutations.

The constant scalar β in (35) and (36) affects how much the phase part is emphasized compared to the amplitude part in frequency-normalized vectors $\bar{\mathbf{a}}_i(f)$ and $\bar{\mathbf{c}}_k$. In general microphone setups, time delays provide more reliable information than attenuations for distinguishing frequency components that originate from different source signals. Thus, it is advantageous to emphasize the phase part by using as large a β value as possible. However, too large a β value may cause phase wrapping. We use $\beta = v/(4\,d_{\max})$ as an appropriate value. The reason for using this value is discussed in [16].

Figure 5 shows the arguments of $\bar{a}_{21}$ and $\bar{a}_{22}$ calculated by operation (35) in the setup A experiment. For frequency range $\mathcal{F}_L$, the clustering algorithm of iterating (37) and (38) was performed to decide the permutations Π_f and the subscripts were updated by (11). We see two clusters whose centroids are the two lines represented by $\arg(\bar{c}_{21})$ and $\arg(\bar{c}_{22})$. For frequencies higher than 850 Hz, we see that operation (35) did not work effectively because of the effect of spatial aliasing. We need another algorithm to minimize the cost function (29) for such higher frequencies.

B. For Frequencies where Spatial Aliasing may Occur

This subsection presents a procedure for deciding permutations Π_f for frequencies where spatial aliasing may occur. Thus far, the frequency-normalized model $\bar{\mathbf{c}}_k$ has been calculated by (38), and it contains model parameters τ_{jk}, λ_{jk} as shown in (36). They can be extracted from the elements of $\bar{\mathbf{c}}_k$ as

$$\tau_{jk} = -\frac{\arg(\bar{c}_{jk})}{2\pi\beta}, \quad \lambda_{jk} = |\bar{c}_{jk}|, \quad \forall\, j, k. \qquad (39)$$

A simple way of deciding permutations for higher frequencies is to use these extracted parameters for the vector form $\mathbf{c}_k(f)$ in (18) and calculate a permutation Π_f based on the original cost function (29) with

$$\Pi_f \leftarrow \text{argmin}_\Pi \sum_{k=1}^{N} ||\tilde{\mathbf{a}}_{\Pi(k)}(f) - \mathbf{c}_k(f)||^2, \quad \forall\, f \in \mathcal{F}. \qquad (40)$$

However, τ_{jk} and λ_{jk} estimated only with frequencies in $\mathcal{F}_L$ may not be very accurate. Figure 6 shows $\arg(\tilde{a}_{21})$ and $\arg(\tilde{a}_{22})$ after the permutations had been calculated by (40) using the model parameters extracted by (39). We see some estimation error for τ_{21} and τ_{22}, as the data (shown in marks '•' and '+') are not lined up along the model line (shown as dashed lines) at high frequencies.

A better way is to re-estimate parameters τ_{jk} and λ_{jk} by minimizing the original cost function $\mathcal{D}_{\mathbf{a}}$ in (29), where the frequency range is not limited to $\mathcal{F}_L$. In our earlier work [2], we used a gradient descent approach to refine these parameters, where we needed to carefully select a step size parameter that guaranteed a stable convergence. In this paper, we adopt the following direct approach instead. With a simple mathematical manipulation (see Appendix VIII-A), the cost function $\mathcal{D}_{\mathbf{a}}$ becomes

$$\sum_{k=1}^{N} \sum_{f\in\mathcal{F}} \sum_{j=1}^{M} \left\{ \frac{1}{M} + \lambda_{jk}^2 - 2\lambda_{jk}\text{Re}[\tilde{a}_{ji}(f)\, e^{\imath 2\pi f \tau_{jk}}]\,\Big|_{i=\Pi_f(k)} \right\} \qquad (41)$$

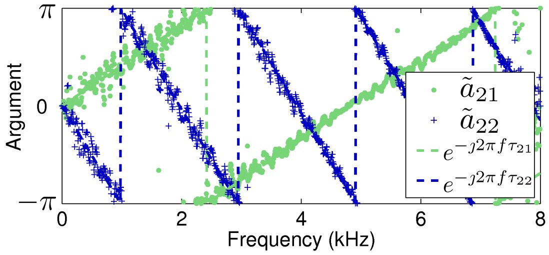


Fig. 7. Arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ after permutation alignment using model parameters re-estimated with data from the whole frequency range $\mathcal{F}$. Now τ_{21} and τ_{22} are precisely estimated, and permutations are aligned correctly.

where $\mathrm{Re}[\cdot]$ takes only the real parts of a complex number. Thus, the optimum time delay τ_{jk} for minimizing the cost function with the current permutations Π_f is given by

$$\tau_{jk} \leftarrow \mathrm{argmax}_{\tau} \sum_{f\in\mathcal{F}} \mathrm{Re}[\tilde{a}_{ji}(f)\, e^{\imath 2\pi f\tau}] \,|_{i=\Pi_f(k)}\,, \quad \forall\, j,k. \tag{42}$$

And, the optimum attenuation λ_{jk} with the current permutations Π_f and the delay parameter τ_{jk} is given by

$$\lambda_{jk} \leftarrow \frac{1}{|\mathcal{F}|} \sum_{f\in\mathcal{F}} \mathrm{Re}[\tilde{a}_{ji}(f)\, e^{\imath 2\pi f\tau_{jk}}] \,|_{i=\Pi_f(k)}\,, \quad \forall\, j,k. \tag{43}$$

This is because the gradient of (41) with respect to λ_{jk} is

$$\frac{\partial \mathcal{D}_{\mathbf{a}}}{\partial \lambda_{jk}} = 2\sum_{f\in\mathcal{F}} \left\{ \lambda_{jk} - \mathrm{Re}[\tilde{a}_{ji}(f)\, e^{\imath 2\pi f\tau_{jk}}] \,|_{i=\Pi_f(k)} \right\}$$

and setting the gradient zero gives the equation (43).

We can iteratively update Π_f by (40) and τ_{jk}, λ_{jk} by (42)-(43) to obtain better estimations of the model parameters and consequently better permutations. Note that the structure that iterates (40) and (42)-(43) has the same structure as (37) and (38). Figure 7 shows $\arg(\tilde{a}_{21})$ and $\arg(\tilde{a}_{22})$ after Π_f and τ_{jk}, λ_{jk} were refined by (40) and (42)-(43). We see that τ_{21} and τ_{22} were precisely estimated and the permutations were aligned correctly even for high frequencies.

V. Classification of Observations for T-F Masking

This section presents a procedure for minimizing the cost function $\mathcal{D}_{\mathbf{x}}$ in (30), and for obtaining a classification $C(f,t)$ of observation vectors $\mathbf{x}(f,t)$ for the T-F masking separation described in Sec. II-B.

A. Procedure

The structure of the procedure is shown in Fig. 8. It is almost the same as that of the permutation alignment (Fig. 4) presented in the last section. The modification made for T-F masking separation involves replacing $\mathbf{a}_i$, $\tilde{\mathbf{a}}_i$, $\bar{\mathbf{a}}_i$, Π_f and "Permutation optimization" with $\mathbf{x}$, $\tilde{\mathbf{x}}$, $\bar{\mathbf{x}}$, C and "Classification optimization," respectively.

Let us assume here that observation vectors $\mathbf{x}$ have been converted into $\tilde{\mathbf{x}}$ by the phase and amplitude normalization presented in Sec. III-C. For frequency range $\mathcal{F}_L$ where spatial aliasing does not occur, frequency normalization [22] is applied to the elements of $\tilde{\mathbf{x}}(f,t)$:

$$\bar{x}_j(f,t) \leftarrow |\tilde{x}_j(f,t)| \exp\left(\imath\, \frac{\beta \arg[\tilde{x}_j(f,t)]}{f}\right)\,, \quad \forall\, j,f,t. \tag{44}$$

With the frequency normalization, the cost function (30) is converted into

$$\bar{\mathcal{D}}_{\mathbf{x}}(\{\tau_{jk}\},\{\lambda_{jk}\},C) = \sum_{k=1}^{N} \sum_{C(f,t)=k} ||\bar{\mathbf{x}}(f,t) - \bar{\mathbf{c}}_k||^2, \tag{45}$$

where $\bar{\mathbf{x}} = [\bar{x}_1,\ldots,\bar{x}_M]^T$, and the right-hand summation with $C(f,t)=k$ is limited to the frequency range $\mathcal{F}_L$ given by (33). The cost function $\bar{\mathcal{D}}_{\mathbf{x}}$ can be minimized efficiently by iterating the following two updates until convergence:

$$C(f,t) \leftarrow \mathrm{argmin}_k ||\bar{\mathbf{x}}(f,t) - \bar{\mathbf{c}}_k||^2\,, \quad \forall\, f,t, \tag{46}$$

$$\bar{\mathbf{c}}_k \leftarrow \frac{1}{N_k} \sum_{C(f,t)=k} \bar{\mathbf{x}}(f,t), \quad \bar{\mathbf{c}}_k \leftarrow \bar{\mathbf{c}}_k / ||\bar{\mathbf{c}}_k||\,, \quad \forall\, k, \tag{47}$$

where N_k is the number of time-frequency slots (f,t) that satisfy $C(f,t)=k$.

For higher frequencies where spatial aliasing may occur, model parameters τ_{jk} and λ_{jk} are first extracted from $\bar{\mathbf{c}}_k$ as shown in (39), and then substituted into the vector form $\mathbf{c}_k(f)$ in (18). Then, the classification of the observation vectors can be decided by

$$C(f,t) \leftarrow \mathrm{argmin}_k ||\tilde{\mathbf{x}}(f,t) - \mathbf{c}_k(f)||^2\,, \quad \forall\, f,t. \tag{48}$$

As with (42)-(43) for permutation alignment in the previous section, the parameters are better estimated according to the original cost function $\mathcal{D}_{\mathbf{x}}$ in (30) by

$$\tau_{jk} \leftarrow \mathrm{argmax}_{\tau} \sum_{C(f,t)=k} \mathrm{Re}[\tilde{x}_j(f,t)\, e^{\imath 2\pi f\tau}]\,, \quad \forall\, j,k, \tag{49}$$

$$\lambda_{jk} \leftarrow \frac{1}{N_k} \sum_{C(f,t)=k} \mathrm{Re}[\tilde{x}_j(f,t)\, e^{\imath 2\pi f\tau_{jk}}]\,, \quad \forall\, j,k, \tag{50}$$

where the summation with $C(f,t)=k$ is not limited to $\mathcal{F}_L$ but covers the whole range $\mathcal{F}$. We can iteratively update $C(f,t)$ by (48) and τ_{jk}, λ_{jk} by (49)-(50) to obtain better estimations of the model parameters and consequently better classification.

B. Relationship to GCC-PHAT

This subsection discusses the relationship between (49) and the GCC-PHAT function [23], [28], [29]. Let us assume that only the first source s_1 is active in an STFT frame centered at time t. The TDOA $\tau_{[j,J]}(t)$ of the source between sensor j and J can be estimated with the GCC-PHAT function as

$$\tau_{[j,J]}(t) = \mathrm{argmax}_{\tau} \sum_{f} \frac{x_j(f,t) x_J^*(f,t)}{|x_j(f,t) x_J^*(f,t)|} e^{\imath 2\pi f\tau} \tag{51}$$

where the summation is over all discrete frequencies.

If the same assumption holds for T-F masking separation, all the observation vectors at time frame t are classified into

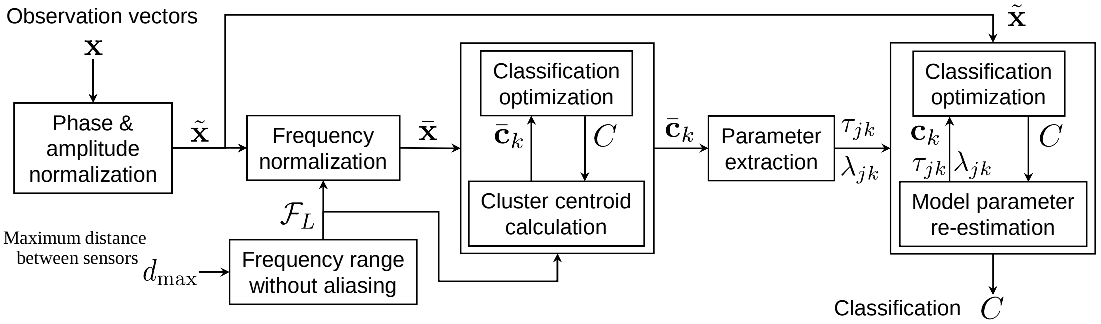


Fig. 8. Flow of the classification procedure presented in Sec. V, which corresponds to the grouping part of (b) separation with T-F masking in Fig. 1.

the first one, i.e., $C(f,t) = 1, \forall\, f$. Then, the delay parameter estimation by (49) using only the time frame is reduced to

$$\tau_{j1} \leftarrow \text{argmax}_\tau \sum_{f\in\mathcal{F}} \text{Re}[\tilde{x}_j(f,t)\, e^{\imath 2\pi f\tau}]\,, \quad \forall\, j, \qquad (52)$$

where $\tilde{x}_j(f,t)$ can be expressed in

$$\tilde{x}_j(f,t) = \frac{x_j(f,t)x_J^*(f,t)}{||\mathbf{x}(f,t)|| \cdot |x_J^*(f,t)|}$$

if we follow the phase and amplitude normalization (24) and (28). Time delay τ_{j1} can be considered as the TDOA of source s_1 between sensors j and J.

We see that (51) and (52) are very similar. The summation in (51) and (52) has the same effect because of the conjugate relationship (6). Thus, the only difference is in the denominator part, $||\mathbf{x}(f,t)||$ or $|x_j(f,t)|$, but this difference has very little effect in the argmax operation if we can approximate $||\mathbf{x}(f,t)|| \approx \alpha \cdot |x_j(f,t)|$ with the same constant α for all frequencies. In [23], T-F masking separation and time delay estimation with GCC-PHAT were discussed, but there was no mathematical statement relating these two.

Based on this observation, we recognize that iterative updates with (48) and (49) perform time delay estimation with the GCC-PHAT function by selecting frequency components of the source. The estimations τ_{jk} are improved by a better classification $C(f,t)$ of the frequency components, and conversely the classification $C(f,t)$ is also improved by better time delay estimations τ_{jk}.

VI. Experiments

A. Experimental setups and evaluation measure

To verify the effectiveness of the proposed formulation and procedure, we conducted experiments with the three setups A, B and C shown in Fig. 9. They differs as regards number of sources and sensors, and sensor spacing. The configurations common to all setups are summarized in Table I. We tested the BSS system mainly with a low reverberation time (130 ms) so that the system can exploit spatial information of the sources accurately when grouping frequency components, but we also tested the system in more reverberant conditions to observe how the separation performance degrades as the reverberation time increases (reported in Sec. VI-E).

TABLE I
Common experimental configurations

Room size	$4.45 \times 3.55 \times 2.5$ m
Reverberation time	RT$_{60}$ = 130 ms 130 $\sim$ 450 ms for setup A
Sampling rate	16 kHz
STFT frame size	2048 points (128 ms)
STFT frame shift	512 points (32 ms)
Source signals	Speeches of 3 s
Propagation velocity	$v = 340$ m/s

The separation performance was evaluated in terms of signal-to-interference ratio (SIR) improvement. The improvement was calculated by $\mathsf{OutputSIR}_i - \mathsf{InputSIR}_i$ for each output i, and we took the average over all output $i = 1, \ldots, N$. These two types of SIRs are defined by

$$\mathsf{InputSIR}_i = 10\log_{10}\frac{\sum_t |\sum_l \mathrm{h}_{Ji}(l)\mathrm{s}_i(t-l)|^2}{\sum_t |\sum_{k\neq i}\sum_l \mathrm{h}_{Jk}(l)\mathrm{s}_k(t-l)|^2} \quad \text{(dB)},$$

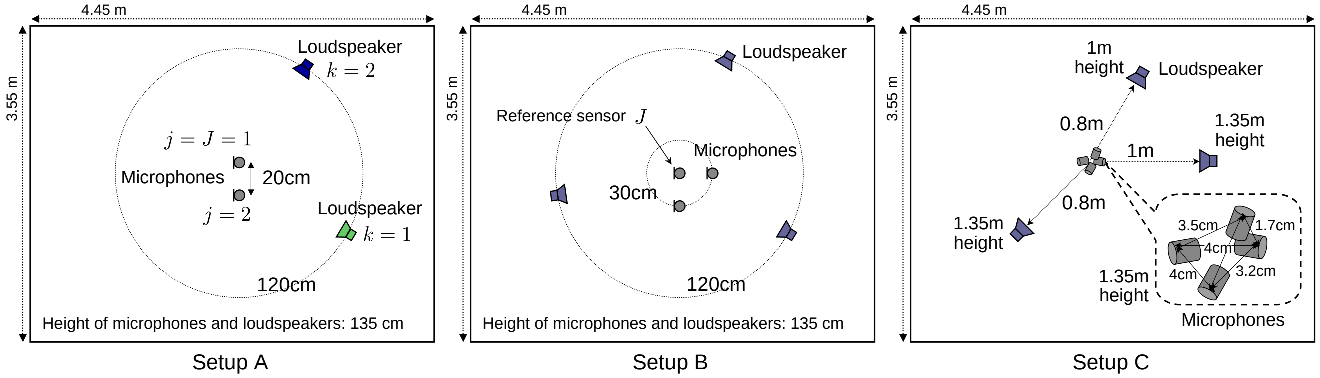
$$\mathsf{OutputSIR}_i = 10\log_{10}\frac{\sum_t |\mathrm{y}_{ii}(t)|^2}{\sum_t |\sum_{k\neq i}\mathrm{y}_{ik}(t)|^2} \quad \text{(dB)},$$

where $J \in \{1,\ldots,M\}$ is the index of a selected reference sensor, and $\mathrm{y}_{ik}(t)$ is the component of s_k that appears at output $\mathrm{y}_i(t)$, i.e., $\mathrm{y}_i(t) = \sum_{k=1}^N \mathrm{y}_{ik}(t)$.

B. Main experiments

Figure 10 summarizes the experimental results with a reverberation time of 130 ms. We performed experiments with eight combinations of 3-second speeches, for pairs consisting of each method (ICA or T-F masking) and setup (A, B or C). As regards phase normalization, a reference sensor was selected (19) for setups A and B, and pairing with the next sensor (21) was employed in setup C. To observe the effect of the multi-stage procedures presented in Secs. IV and V, we measured the SIR improvements at three different stages and for two special options:

Stage I Grouping frequency components only at low frequency range $\mathcal{F}_L$ where spatial aliasing does not occur, by (37) and (38) for permutations Π_f, or by (46) and (47) for classification $C(f,t)$. At the remaining frequencies, the permutations or classification were random.



Setup A

Setup B

Setup C

Fig. 9. Three experimental setups. Setup A: two sources and two sensors with large spacing. Setup B: three sources and three sensors with large spacing. Setup C: three sources and four sensors with small spacing. All the microphones were omni-directional.

Stage II After Stage I, grouping frequency components at the remaining high frequencies by (40) or (48) with the model parameters τ_{jk}, λ_{jk} extracted by (39), which were not so accurate because they were estimated only with the data from the low frequency range $\mathcal{F}_L$.

Stage III After Stage II, re-estimating model parameters τ_{jk}, λ_{jk} by (42)-(43) with $\mathbf{a}_i$, or by (49)-(50) with $\mathbf{x}$. This re-estimation was interleaved with grouping frequency components at the high frequencies by (40) or (48).

Only III Only the core part of stage III was applied. Grouping frequency components by interleaving (40) and (42)-(43) for permutations Π_f, or (48) and (49)-(50) for classification $C(f,t)$, starting from random initial permutations or classification.

Optimal Optimal permutations Π_f or classification $C(f,t)$ was calculated using the information on source signals. This is not a practical solution, but is to enable us to see the upper limit of the separation performance.

SIR improvements became better as the stage proceeded from I to III. This is noticeable in setups A and B where the sensor spacing was large and the frequency range $\mathcal{F}_L$ without spatial aliasing was very small. On the other hand, in setup C, the difference was not so large because the sensor spacing was small and the range $\mathcal{F}_L$ occupied more than half the whole range $\mathcal{F}$.

Even if only stage III was employed with random initial permutations or classification, the results were sometimes good. In some cases, however, especially for setup B with T-F masking, the results were not good. These results show that the classification problem for T-F masking has a much larger possible solution space than the permutation problem for ICA, and it is easy to get stuck in a local minimum of the cost function $\mathcal{D}_{\mathbf{x}}$. Therefore, the multi-stage procedure has an advantage in that it is not likely to become stuck in local minima.

Table II shows the total computational time for the BSS procedure, and also those of the ICA and Grouping subcomponents depicted in Fig. 1. They are for 3-second source signals, and are averaged over the eight different source combinations. The BSS program was coded in Matlab and run on an AMD 2.4 GHz Athlon 64 processor. The computational time of the Grouping procedure was not very large and was smaller than that of ICA. Table II also shows the average number of iterations to converge for the Grouping procedure, (40) and (42)-(43) with ICA, or (48) and (49)-(50) with T-F masking. The T-F masking grouping procedure requires more iterations than that of ICA because of the larger solution space, but it converges within a reasonable number of iterations.

TABLE II
COMPUTATIONAL TIME

	Total	ICA	Grouping	(#iterations)
Setup A, ICA	4.87 s	4.07 s	0.48 s	(4.9)
Setup B, ICA	8.05 s	6.85 s	0.80 s	(6.4)
Setup C, ICA	7.71 s	6.81 s	0.42 s	(4.2)
Setup A, T-F masking	1.64 s	-	1.44 s	(9.4)
Setup B, T-F masking	2.68 s	-	2.37 s	(11.5)
Setup C, T-F masking	4.18 s	-	3.83 s	(8.1)

C. Comparison with null beamforming

Let us compare the separation capability of the proposed methods (ICA and T-F masking) with that of null beamforming, which is a conventional source separation method that similarly exploits the spatial information of sources. In null beamforming, filter coefficients are designed by assuming the anechoic propagation model (17). In this sense, all these three methods rely on delay τ_{jk} and attenuation λ_{jk} parameters.

We designed the null beamformer in the frequency domain. The separation matrix $\mathbf{W}(f)$ in each frequency bin was given by the inverse (or Moore-Penrose pseudo inverse if $N < M$) of the assumed mixing matrix

$$\begin{bmatrix} c_{11}(f) & \dots & c_{1N}(f) \\ \vdots & \ddots & \vdots \\ c_{M1}(f) & \dots & c_{MN}(f) \end{bmatrix},$$

where $c_{jk}(f)$ is the propagation model defined in (17). The delay τ_{jk} and attenuation λ_{jk} parameters were accurately estimated in the experiment, from the individual source contributions on the microphones for each source.

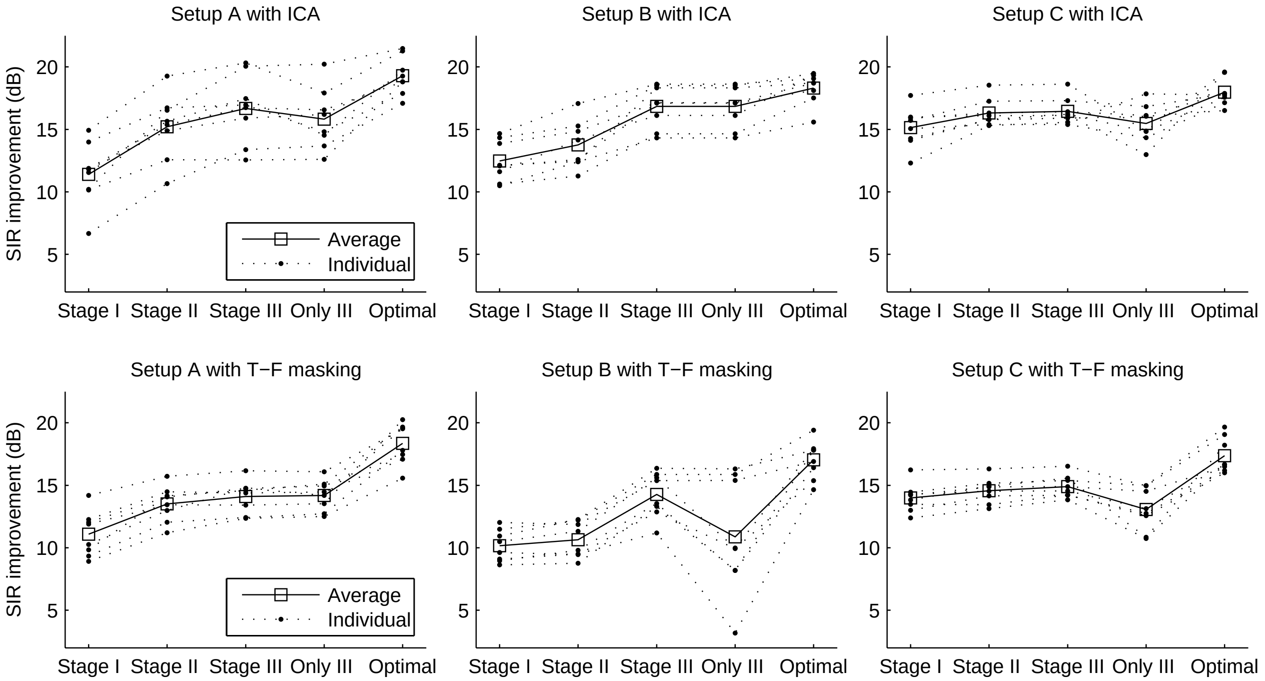


Fig. 10. SIR improvements at different stages. The first and second rows correspond to ICA-based separation and T-F masking separation, respectively. The first, second, and third columns correspond to setups A, B, and C, respectively. Each dotted line shows an individual case, and a solid line with squares shows the average of the eight individual cases.

TABLE III
SIR IMPROVEMENTS (dB) WITH DIFFERENT SEPARATION METHODS

	Anechoic	Setup A	Setup B	Setup C
Null beamforming	37.29	8.14	7.93	6.94
ICA	27.53	16.67	16.85	16.44
T-F masking	17.92	14.10	14.27	14.90

Table III reports SIR improvements with these methods for four different setups. An anechoic setup was added to the existing three setups (A, B, and C) to contrast the characteristics of these three methods. In the anechoic setup, the positions of loudspeakers and microphones were the same as those of setup A.

We observe the following from the table. Null beamforming performs the best in the anechoic setup, but worse than the other two methods in the three real-room setups. With null beamforming, propagation model parameters are used for designing the filter coefficients in the separation system. Thus, even a small discrepancy between the propagation model and a real room situation directly affects the separation. With ICA or T-F masking, on the other hand, the propagation model is used only for grouping separated frequency components. The discrepancy between the propagation model and a real room situation is reflected in the cost function $\mathcal{D}_{\mathbf{a}}$ or $\mathcal{D}_{\mathbf{x}}$ as discussed in Sec. III-D. Therefore, these methods are robust to such a discrepancy if it is not very severe.

D. Comparison of ICA and T-F masking

In terms of grouping frequency components, the ICA-based and T-F masking methods have a lot in common as discussed above. However, they are of course different in terms of the whole BSS procedure. Here we compare these two methods.

With ICA, separated frequency components are generated by the ICA formula (7). The separation matrix $\mathbf{W}(f)$ is designed for each frequency so that it adapts to a mixing situation (anechoic or real reverberant). This is why ICA performs well in all the setups in Table III and also in Fig. 10.

In contrast, with T-F masking, separated frequency components are simply frequency-domain sensor observations calculated by an STFT (3). How well these components are separated depends on how well the sparseness assumption (13) holds for the original source signals. In general, a speech signal follows the sparseness assumption to a certain degree, but it does less accurately than the anechoic situation follows the propagation model (17). This is why the SIR improvement of T-F masking for the anechoic setup saturated compared with the other two in Table III. It should also be noted that violation of the sparseness assumption leads to an undesirable musical noise effect.

In summary, if the number of sensors is sufficient for the number of sources as shown in Table III, the ICA based method performs better than the T-F masking method. However, a T-F masking approach has a separation capability for an under-determined case where the number of sensors is insufficient.

E. Experiments in more reverberant conditions

We also performed experiments in more reverberant conditions. The reverberation time was controlled by changing the area of cushioned wall in the room. We considered five

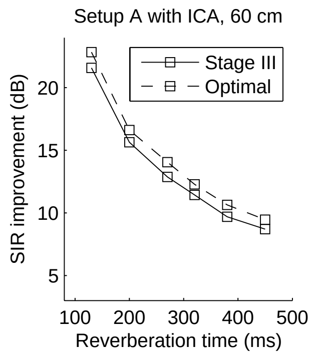


Fig. 11. SIR improvements with ICA-based BSS for setup A for various reverberation times (RT_{60} = 130, 200, 270, 320, 380, and 450 ms) and two different distances (60 and 120 cm) from the sources to the microphones. Each square shows the average SIR improvement of the eight different combinations of speech sources.

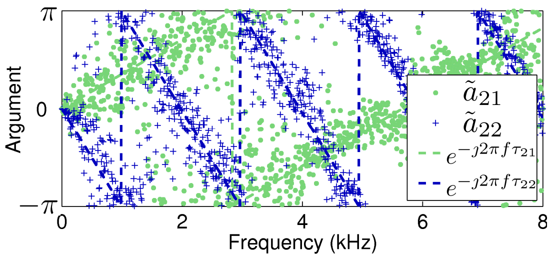


Fig. 12. Arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ after permutations were aligned at stage III. The room reverberation time was 380 ms and the distance from the sources to the microphones was 120 cm, which made the situation very different from the assumed anechoic model. Consequently, the samples of the arguments were widely scattered around the estimated model parameters. However, the model parameters were reasonably estimated so the source directions can be approximately estimated together with the information about the microphone array geometry.

additional different reverberation times for setup A, namely 200, 270, 320, 380, and 450 ms. We also considered another distance of 60 cm from the sources to the microphones. As regards the experiments reported here, let us focus on ICA-based separation for simplicity.

Figure 11 shows SIR improvements at stage III and also with optimal permutations. Reverberation affects the ICA solutions as well as the permutation alignment. Even with optimal permutations, the ICA separation performance degrades as the reverberation time increases. The difference between "Optimal" and "Stage III" SIR improvements indicates the performance degradation caused by permutation misalignment. In the shorter distance case (60 cm), the degree of degradation was uniformly small for various reverberation times. This is because the contribution of the direct path from a source to a microphone is dominant compared with those of the reverberations, and thus the situation is well approximated with the anechoic propagation model. However, with the original distance (120 cm), the degradation became large as the reverberation time became long. These results show the applicability/limitation of the proposed method for permutation alignment in more reverberant conditions as a case study.

Figure 12 shows the arguments of $\tilde{a}_{21}$ and $\tilde{a}_{22}$ after the permutations were aligned at stage III, in an experiment with a reverberation time of 380 ms and a distance of 120 cm. Compared with Fig. 7 (where the reverberation time was 130 ms), we see that the basis vector elements were widely scattered around the estimated anechoic model due to the long reverberation time, and thus permutation misalignments occurred more frequently. However, the model parameters were reasonably estimated, capturing the center of the scattered samples to minimize the cost function (29).

VII. Conclusion

We proposed a procedure for grouping frequency components, which are basis vectors $\mathbf{a}_i(f)$ in ICA-based separation, or observation vectors $\mathbf{x}(f,t)$ in T-F masking separation. The grouping result is expressed in permutations Π_f for ICA-based separation, or in classification information $C(f,t)$ for T-F masking separation. The grouping is decided based on the estimated parameters of time delays τ_{jk} and attenuations λ_{jk} from source to sensors. The proposed procedure interleaves the grouping of frequency components and the estimation of the parameters, with the aim of achieving better results for both. We adopt a multi-stage approach to attain a fast and robust convergence to a good solution. Experimental results show the validity of the procedure, especially when spatial aliasing occurs due to wide sensor spacing or a high sampling rate. The applicability/limitation of the proposed method under reverberant conditions is also demonstrated experimentally.

The primary objective of this work was blind source separation of acoustic sources. However, with the proposed scheme, the time delays and attenuations from sources to sensors are also estimated with a function similar to that of GCC-PHAT. If we have information on the sensor array geometry, we can also estimate the locations of multiple sources. This point should be interesting also to researchers working in the field of source localization.

VIII. Appendix

A. Calculating and simplifying the cost functions

The squared distance $||\tilde{\mathbf{a}}_i - \mathbf{c}_k||^2$ that appears in (29) can be transformed into

$$(\tilde{\mathbf{a}}_i - \mathbf{c}_k)^H(\tilde{\mathbf{a}}_i - \mathbf{c}_k) = \tilde{\mathbf{a}}_i^H\tilde{\mathbf{a}}_i + \mathbf{c}_k^H\mathbf{c}_k - \tilde{\mathbf{a}}_i^H\mathbf{c}_k - \mathbf{c}_k^H\tilde{\mathbf{a}}_i$$

where

$$\tilde{\mathbf{a}}_i^H\tilde{\mathbf{a}}_i = ||\tilde{\mathbf{a}}_i||^2 = 1\,,$$

$$\mathbf{c}_k^H\mathbf{c}_k = \sum_{j=1}^{M}\lambda_{jk}^2 = 1$$

from the assumptions, and

$$-\tilde{\mathbf{a}}_i^H\mathbf{c}_k - \mathbf{c}_k^H\tilde{\mathbf{a}}_i = -2\mathrm{Re}(\mathbf{c}_k^H\tilde{\mathbf{a}}_i)\,.$$

Thus, the minimization of the squared distance $||\tilde{\mathbf{a}}_i - \mathbf{c}_k||^2$ is equivalent to the maximization of the real part of the inner product $\mathbf{c}_k^H\tilde{\mathbf{a}}_i$, whose calculation is less demanding in terms of computational complexity. We follow this idea in calculating the argmin operators in (37), (40), (46) and (48).

The mathematical manipulations conducted for obtaining (41) were the above equations and

$$\mathrm{Re}[\mathbf{c}_k^H(f)\tilde{\mathbf{a}}_i(f)] = \sum_{j=1}^{M} \lambda_{jk} \mathrm{Re}[\tilde{a}_{ji}(f)\, e^{\imath 2\pi f \tau_{jk}}]\,.$$

References

[1] H. Sawada, S. Araki, R. Mukai, and S. Makino, "On calculating the inverse of separation matrix in frequency-domain blind source separation," in *Independent Component Analysis and Blind Signal Separation*, ser. LNCS, vol. 3889. Springer, 2006, pp. 691–699.

[2] ——, "Solving the permutation problem of frequency-domain BSS when spatial aliasing occurs with wide sensor spacing," in *Proc. ICASSP 2006*, vol. V, May 2006, pp. 77–80.

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

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

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

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

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

[8] 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.

[9] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. ICA 2000*, June 2000, pp. 215–220.

[10] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Jan. 1999, pp. 365–371.

[11] 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.

[12] 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.

[13] 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.

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

[15] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Near-field frequency domain blind source separation for convolutive mixtures," in *Proc. ICASSP 2004*, vol. IV, 2004, pp. 49–52.

[16] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. Audio, Speech and Language Processing*, pp. 2165–2173, Nov. 2006.

[17] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," in *Proc. ICA 2006 (LNCS 3889)*. Springer, Mar. 2006, pp. 601–608.

[18] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech and Language Processing*, pp. 70–79, Jan. 2007.

[19] 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.

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

[21] 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.

[22] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC 2005)*, Sept. 2005, pp. 117–120.

[23] M. Swartling, N. Grbić, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *Proc. ICASSP 2006*, vol. IV, May 2006, pp. 833–836.

[24] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, pp. 627–641, 2003.

[25] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–12, Article ID 24 717, 2007.

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

[27] W. Kellermann, H. Buchner, and R. Aichner, "Separating convolutive mixtures with TRINICON," in *Proc. ICASSP 2006*, vol. V, May 2006, pp. 961–964.

[28] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[29] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 288–292, May 1997.

[30] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation," in *Audio Signal Processing*, Y. Huang and J. Benesty, Eds. Kluwer Academic Publishers, 2004, pp. 197–227.

[31] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[32] Y. Huang, J. Benesty, and G. Elko, "Source localization," in *Audio Signal Processing*, Y. Huang and J. Benesty, Eds. Kluwer Academic Publishers, 2004, pp. 229–253.

[33] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[34] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[35] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multi-channel linear prediction," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[36] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. ICA 2001*, Dec. 2001, pp. 722–727.

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

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.

He joined NTT in 1993. He is now a senior research scientist at the 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. In 2000, he stayed at the Computation Structures Group of MIT for six months. From 2002 to 2005, he taught a class on computer architecture at Doshisha University, Kyoto.

Since 2000, he has been engaged in research on signal processing, microphone array, and blind source separation (BSS). More specifically, he is working on the frequency-domain BSS for acoustic convolutive mixtures using independent component analysis (ICA). He is an associate editor of the IEEE Transactions on Audio, Speech & Language Processing, and a member of the Audio and Electroacoustics Technical Committee of the IEEE SP Society. He was a tutorial speaker at ICASSP 2007. He serves as the publications chairs of the WASPAA 2007 in Mohonk, and served as an organizing committee member for ICA 2003 in Nara and the communications chair for IWAENC 2003 in Kyoto.

He is the author or co-author of three book chapters, more than 20 journal articles, and more than 80 conference papers. 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 from the University of Tokyo, Japan, in 1998 and 2000, respectively, and the Ph. D degree from Hokkaido University, Japan in 2007. 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 is a member of the Organizing Committee of the ICA 2003, the Finance Chair of IWAENC 2003, the Registration Chair of WASPAA 2007. She received the 19th Awaya Prize from Acoustical Society of Japan (ASJ) in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, and the Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2005. She is a member of the IEEE, IEICE, 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 is a member of the ACM, the Acoustical Society of Japan (ASJ), Institute of Electronics, Information and Communication Engineers (IEICE), and Information Processing Society of Japan (IPSJ). He is also a member of the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society, the Organizing Committee of the ICA 2003 in NARA, and the Publications Chair of the IWAENC 2003 in Kyoto. He received the Sato Paper Award of the ASJ in 2005 and the Paper Award of the IEICE in 2005.

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 ICA Unsupervised Learning Pioneer Award in 2006, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002, the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, 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 is responsible for more than 150 patents. He is a Tutorial speaker at ICASSP2007, and a Panelist at HSCMA2005.

He is a member of both the Awards Board and 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 Guest Editor of the Special Issue of the IEEE Transactions on Audio, Speech and Language Processing and a Guest Editor of the Special Issue of the IEEE Transactions on Computers. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society and the Chair-Elect of the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is the Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. He is a member of the International IWAENC Standing committee and a member of the International ICA Steering Committee.

He is the General Chair of the WASPAA2007 in Mohonk, the General Chair of the IWAENC2003 in Kyoto, the Organizing Chair of the ICA2003 in Nara.

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