IEICE TRANS. FUNDAMENTALS, VOL.E88–A, NO.7 JULY 2005

INVITED PAPER *Special Section on Multi-channel Acoustic Signal Processing*

Blind Source Separation of Convolutive Mixtures of Speech in Frequency Domain

Shoji MAKINO[†a)], **Hiroshi SAWADA**[†], **Ryo MUKAI**[†], ***Members***, ***and*** **Shoko ARAKI**[†], ***Nonmember***

SUMMARY This paper overviews a total solution for frequency-domain blind source separation (BSS) of convolutive mixtures of audio signals, especially speech. Frequency-domain BSS performs independent component analysis (ICA) in each frequency bin, and this is more efficient than time-domain BSS. We describe a sophisticated total solution for frequency-domain BSS, including permutation, scaling, circularity, and complex activation function solutions. Experimental results of 2×2, 3×3, 4×4, 6×8, and 2×2 (moving sources), (#sources × #microphones) in a room are promising.

key words: *blind source separation, convolutive mixtures, independent component analysis, frequency-domain BSS, microphone array, adaptive beamformer*

1. Introduction

Blind source separation (BSS) [1]–[3] is an approach for estimating source signals by using only the information of mixed signals observed at each input channel. The estimation is performed blindly, i.e., without possessing information on each source, such as its location and active time. Typical examples of such source signals include mixtures of simultaneous speech signals that have been picked up by several microphones. Its potential audio signal applications include speech enhancement for speech recognition, teleconferences, and hearing aids. In such applications, signals are mixed in a convolutive manner with reverberations. This makes the BSS problem difficult. We need very long finite impulse response (FIR) filters (e.g., around a thousand taps for 8 kHz sampling) to separate the acoustic signals mixed under such conditions.

Independent component analysis (ICA) [4], [5] is a major statistical tool for dealing with the BSS problem. If signals are mixed instantaneously, we can directly employ an instantaneous ICA algorithm to separate them. However, signals are mixed in a convolutive manner in the applications mentioned above. Therefore, we need to extend the ICA/BSS technique so that it is applicable to convolutive mixtures.

The first approach is time-domain BSS, where ICA is directly extended to the convolutive mixture model [6]–[11]. This approach is theoretically sound and achieves good separation once an algorithm converges, since the algorithm correctly evaluates the independence of separated signals. However, an ICA algorithm for convolutive mixtures is not as simple as an ICA algorithm for instantaneous mixtures, and is computationally expensive for long FIR filters because it includes convolution operations.

The second approach is frequency-domain BSS, where complex-valued ICA for instantaneous mixtures is employed in each frequency bin [12]–[29]. The merit of this approach is that the ICA algorithm remains simple and can be performed separately at each frequency. Also, any complex-valued instantaneous ICA algorithm can be employed with this approach. The computational time for BSS can be reduced by employing a fast algorithm such as FastICA [30], [31], and/or by performing parallel computation for multiple frequency bins. However, the permutation ambiguity of the ICA solution becomes a serious problem. We need to align the permutation in each frequency bin so that a separated signal in the time domain contains frequency components from the same source. This problem is well known as the permutation problem of frequency-domain BSS [12]–[21], [25]–[27], which is the main focus of this paper. Another problem relates to the circularity effect of discrete frequency representation. Frequency responses calculated in the frequency domain assume a periodic time-domain filter for their implementation. However, such a periodic filter is unrealistic, and we usually use its one-period realization for the separation filter. Therefore, the frequency responses should be smoothed so that the one-period realization does not rely on the circularity effect [18], [29]. This paper also discusses this problem.

The third approach uses both the time and frequency domains. In some time-domain BSS methods, convolutions in the time domain are speeded up by the overlap-save method in the frequency domain [10], [32]. Furthermore, in some methods [33]–[35], filter coefficients are updated in the frequency domain while nonlinear functions for evaluating independence are applied in the time domain. The permutation problem does not occur in either case since the independence of separated signals is evaluated in the time domain. Nor does the circularity problem occur when there is an appropriate constraint for filter coefficients [36] by such means as rectangular windowing. However, the algorithm moves back and forth between the two domains at every iteration, spending non-negligible time on discrete Fourier transforms (DFTs) and inverse DFTs. Therefore, we consider that the permutation and circularity problems are inevitable if we hope to benefit from the merits of frequency-domain BSS.

Manuscript received February 16, 2005.
Final manuscript received March 9, 2005.

[†]The authors are with NTT Communication Science Laboratories, NTT Corporation, Kyoto-fu, 619-0237 Japan.
a) E-mail: maki@cslab.kecl.ntt.co.jp

DOI: 10.1093/ietfec/e88–a.7.1640

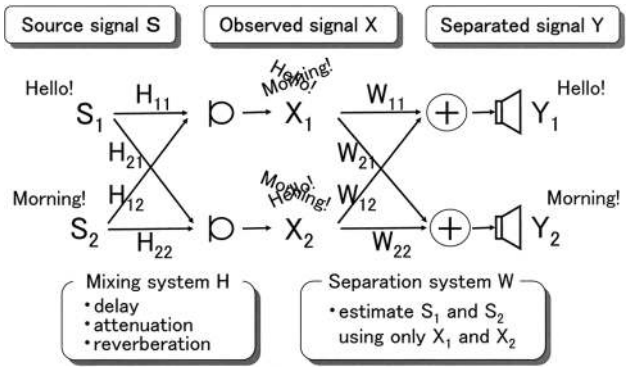

Fig. 1 BSS system configuration.

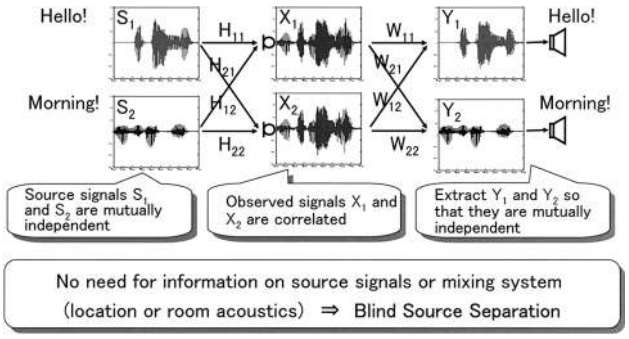

Fig. 2 Task of blind source separation of speech signals.

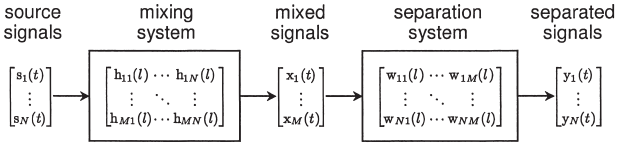

Fig. 3 BSS for convolutive mixtures.

This paper deals with the second approach, i.e., frequency-domain BSS. We begin by formulating the BSS problem for convolutive mixtures in Sect. 2. Section 3 provides an overview of frequency-domain BSS. We then present several important techniques that enable this approach to achieve effective separation of many sources mixed in a reverberant environment. Section 4 discusses complex-valued ICA for instantaneous mixtures. Understanding the separation mechanism of BSS in Sect. 5 greatly helps us to cope with the problem. Section 7 presents a method for solving the permutation problem, which is the most important technique for frequency-domain BSS. To solve the permutation problem, information on source location is very useful. This can be estimated from ICA solutions as shown in Sect. 6. The key point with respect to source localization is that the estimation of the mixing system is easily obtained. This is because the ICA algorithm is just for instantaneous mixtures, and therefore it is straightforward to calculate the (pseudo)-inverse of a separation matrix, which corresponds to the mixing system. This fact also makes it easy to solve the scaling ambiguity as shown in Sect. 8. Section 9 discusses a spectral smoothing technique designed to solve the circularity problem. The experimental results shown in Sect. 10 are very promising. Section 11 concludes this paper.

2. BSS for Convolutive Mixtures

In the case of audio source separation, several sensor microphones are placed in different positions so that each records a mixture of the original source signals at a slightly different time and level. In the real world, where the source signals are speech and the mixing system is a room, the signals that are picked up by the microphones are affected by reverberation. Suppose that N source signals $\mathrm{s}_i(t)$ are mixed and observed at M sensors

$$\mathrm{x}_j(t) = \sum_{i=1}^{N} \sum_{l} \mathrm{h}_{ji}(l)\, \mathrm{s}_i(t-l), \quad j = 1, \ldots, M, \tag{1}$$

where $\mathrm{h}_{ji}(l)$ represents the impulse response from source i to sensor j. We assume that the number of sources N is known or can be estimated in some way (e.g., by [37]), and the number of sensors M is more than or equal to N ($N \leq M$).

The separation system typically consists of a set of FIR filters $\mathrm{w}_{ij}(l)$ of length L to produce N separated signals

$$\mathrm{y}_i(t) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} \mathrm{w}_{ij}(l)\, \mathrm{x}_j(t-l), \quad i = 1, \ldots, N \tag{2}$$

at the outputs. The separation filters are estimated so that the separated signals become mutually independent. The separation filters $\mathrm{w}_{ij}(l)$ should be obtained blindly, i.e., without knowing $\mathrm{s}_i(t)$ or $\mathrm{h}_{ji}(l)$.

A two-input, two-output convolutive BSS problem, i.e., $N = M = 2$, is shown in Figs. 1 and 2. It is assumed that the source signals s_1 and s_2 are mutually independent. This assumption usually holds for sounds in the real world. There are two microphones which pick up the mixed speech. Only the observed signals x_1 and x_2 are available, and they are correlated. The goal is to adapt the separation systems w_{ij} and to extract y_1 and y_2 so that they are mutually independent. With this operation, we can obtain s_1 and s_2 in the output y_1 and y_2. No information is needed on the source positions or period of source existence/absence. Nor is any information on the mixing systems $\mathrm{h}_{ji}(l)$ required. Thus, this task is called *blind* source separation.

Figure 3 shows a block diagram of BSS. The ideal goal of BSS is to separate and deconvolve the mixtures $\mathrm{x}_j(t)$, and to obtain a delayed version of source $\mathrm{s}_i(t)$ at each output i. However, this is very difficult if $\mathrm{s}_i(t)$ is a colored signal, which is the case when separating natural sounds such as speech [9]. A practical alternative goal [8], [11] is to obtain the convolved version of a source $\mathrm{s}_i(t)$ measured at a sensor J_i:

$$\mathrm{y}_i(t) = \sum_{l} \mathrm{h}_{J_i i}(l)\, \mathrm{s}_i\left(t - \frac{L}{2} - l\right), \tag{3}$$

where the sensor index J_i can be selected according to each

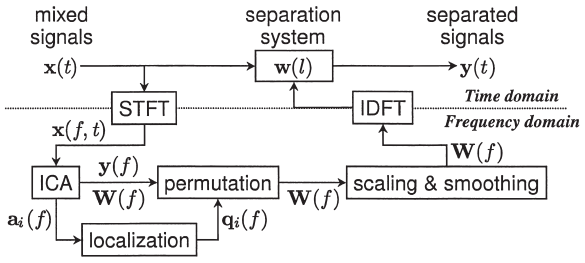

Fig. 4 Flow of frequency-domain BSS.

output i. The way used to attain this goal will be discussed in Sect. 8.

3. Overview of Frequency-Domain Approach

Figure 4 shows the flow of frequency-domain BSS. Time-domain signals $\mathrm{x}_j(t)$ sampled at frequency f_s are converted into frequency-domain time-series signals $x_j(f,\tau)$ with an L-point short-time Fourier transform (STFT):

$$x_j(f,\tau) = \sum_{r=-\frac{L}{2}}^{\frac{L}{2}-1} \mathrm{x}_j(\tau+r)\,\mathrm{win}(r)\,e^{-j2\pi fr}, \tag{4}$$

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

The remaining operations are performed in the frequency domain. The advantage is that the convolutive mixtures in (1) can be approximated as instantaneous mixtures in each frequency bin:

$$x_j(f,\tau) = \sum_{i=1}^{N} h_{ji}(f)s_i(f,\tau), \tag{5}$$

where $h_{ji}(f)$ is the frequency response from source i to sensor j, and $s_i(f,\tau)$ is a frequency-domain time-series signal of $\mathrm{s}_i(t)$ obtained by the same operation as (4). The vector notation of the mixing model (5) is

$$\mathbf{x}(f,\tau) = \sum_{i=1}^{N} \mathbf{h}_i(f)s_i(f,\tau), \tag{6}$$

where $\mathbf{x} = [x_1,\ldots,x_M]^T$ is a sensor sample vector and $\mathbf{h}_i = [h_{1i},\ldots,h_{Mi}]^T$ is the vector of the frequency responses from source s_i to all M sensors.

To obtain the frequency responses $w_{ij}(f)$ of separation filters $\mathrm{w}_{ij}(l)$ in (2), complex-valued ICA

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

is solved, where $\mathbf{y} = [y_1,\ldots,y_N]^T$ is a vector of separated signals, $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_N]^H$ is an $N\times M$ separation matrix, $\mathbf{w}_i = [w_{i1},\ldots,w_{iM}]^H$, and $w_{ij} = [\mathbf{W}]_{ij}$. The details of the ICA algorithm are discussed in Sect. 4.

Calculating the Moore-Penrose pseudoinverse $\mathbf{W}^+$ (reduced to the inverse $\mathbf{W}^{-1}$ if $N=M$) of $\mathbf{W}$ as

$$[\mathbf{a}_1,\cdots,\mathbf{a}_N] = \mathbf{W}^+, \tag{8}$$

$$\mathbf{a}_i = [a_{1i},\ldots,a_{Mi}]^T, \tag{9}$$

is very useful for source localization and scaling alignment, as described in Sects. 6 and 8, respectively. It should be noted that it is not difficult to make $\mathbf{W}$ invertible by using an appropriate ICA procedure (for an example, see Sect. 4). By multiplying both sides of (7) by $\mathbf{W}^+$, the sensor sample vector $\mathbf{x}(\tau)$ is represented by a linear combination of basis vectors $\mathbf{a}_1,\ldots,\mathbf{a}_N$:

$$\mathbf{x}(f,\tau) = \sum_{i=1}^{N} \mathbf{a}_i(f)y_i(f,\tau). \tag{10}$$

It is well-known that an ICA solution (7) has permutation and scaling ambiguities: even if we permute the rows of $\mathbf{W}(f)$ or multiply a row by a constant, it is still an ICA solution. In matrix notation,

$$\mathbf{W}(f) \leftarrow \mathbf{\Lambda}(f)\mathbf{P}(f)\mathbf{W}(f) \tag{11}$$

is also an ICA solution for any permutation $\mathbf{P}(f)$ and diagonal $\mathbf{\Lambda}(f)$ matrix. Permutation alignment is to decide $\mathbf{P}(f)$ so that a time-domain separated signal contains frequency components from the same source. Section 7 presents a method for solving this problem. Scaling alignment is to decide $\mathbf{\Lambda}(f)$ so that a time-domain separated signal satisfies the goal (3), as discussed in Sect. 8.

Then, we perform spectral smoothing so that a time-domain separation filter tapers smoothly to zero at each end. This is typically achieved by multiplying the time-domain filter by a Hanning window, which is equivalent to smoothing the frequency-domain separation matrices as

$$\mathbf{W}(f) \leftarrow \frac{1}{4}\left[\,\mathbf{W}(f-\Delta f) + 2\mathbf{W}(f) + \mathbf{W}(f+\Delta f)\,\right],$$

where $\Delta f = \frac{f_s}{L}$ is the difference from the adjacent frequency. However, this smoothing changes the ICA solution and causes an error. Section 9 discusses the error and how to minimize it.

Finally, separation filters $\mathrm{w}_{ij}(l)$ are obtained by applying inverse DFT to $w_{ij}(f) = [\mathbf{W}(f)]_{ij}$:

$$\mathrm{w}_{ij}(l) = \sum_{f\in\{0,\,\frac{1}{L}f_s,\,\ldots,\,\frac{L-1}{L}f_s\}} w_{ij}(f)e^{j2\pi f(l-\frac{L}{2})},$$

where $l = 0,\ldots,L-1$. The reason for using $e^{j2\pi f(l-\frac{L}{2})}$ instead of $e^{j2\pi fl}$ is to make the separation filter $\mathrm{w}_{ij}(l)$ causal. Then, the separated signals $\mathrm{y}_i(t)$ are produced by (2).

4. Complex-Valued ICA

This section discusses how to solve the ICA Eq. (7). One of the advantages of frequency-domain BSS is that we can employ any ICA algorithm for instantaneous mixtures, such as the information maximization approach (InfoMax) [38] combined with the natural gradient [39], FastICA [30], JADE [40], or an algorithm based on the non-stationarity of

signals [41]. Here, we explain a procedure that was shown to be efficient by the experiments described in Sect. 10. The procedure consists of the following three steps:

1. Dimension reduction and whitening by eigenvalue decomposition
2. ICA by a unitary matrix (FastICA)
3. ICA by InfoMax combined with the natural gradient

The first step performs a linear transformation

$$\mathbf{z}(\tau) = \mathbf{V}\mathbf{x}(\tau)$$

for M-dimensional sensor observations $\mathbf{x}(\tau)$ such that the dimension of $\mathbf{z}(\tau)$ is reduced (if necessary) to the number of sources N and $\mathbf{z}(\tau)$ is spatially whitened (sphered), i.e., $\langle \mathbf{z}(\tau)\mathbf{z}(\tau)^H \rangle_\tau = \mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix. The linear transformation $\mathbf{V}$ is typically obtained by eigenvalue decomposition. Let $\lambda_1 \geq \ldots \geq \lambda_M$ be sorted eigenvalues of the spatial correlation matrix $\mathbf{R} = \langle \mathbf{x}(\tau)\mathbf{x}(\tau)^H \rangle_\tau$ and $\mathbf{e}_1, \ldots, \mathbf{e}_M$ be their corresponding eigenvectors. Then, the linear transformation is

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^H,$$

where $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of the N largest eigenvalues, $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]$ is the matrix of their corresponding eigenvectors, and $\mathbf{e}_i = [e_{1i}, \ldots, e_{Mi}]^T$.

This step has practical importance for the following two reasons. First, the outputs $\mathbf{y}(\tau)$ of ICA (7) adhere to the signal subspace that is identified by the N eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_N$. This means that the following ICA algorithm does not pursue its solution in the noise subspace, which consequently stabilizes the algorithm and also has a noise/reverberation reduction effect [18]. A geometrical interpretation of the dimension reduction is given in [28]. Second, the whitening $\langle \mathbf{z}\mathbf{z}^H \rangle_\tau = \mathbf{I}$ is necessary for FastICA, and also provides an efficient convergence for InfoMax even if the step size is constant over all frequency bins.

The second step performs ICA in a constrained form:

$$\mathbf{y}(\tau) = \mathbf{B}\mathbf{z}(\tau),$$

where $\mathbf{B}$ is an $N \times N$ unitary matrix: $\mathbf{B}\mathbf{B}^H = \mathbf{I}$. This is performed by a complex-valued version of FastICA [30], [31]. It is very efficient because a fairly good solution can be obtained with only several iterations. The efficiency comes from the fact that $\mathbf{z}$ is whitened and $\mathbf{B}$ is unitary. However, there remains room for improving the solution by using another ICA algorithm. One of the reasons is that the output $\mathbf{y}$ of FastICA is whitened $\langle \mathbf{y}(f,\tau)\mathbf{y}(f,\tau)^H \rangle_\tau = \mathbf{I}$ and therefore uncorrelated, whereas original sources $s_1(f,\tau), \ldots, s_N(f,\tau)$ are not always completely uncorrelated with a limited number of samples.

The third step improves the ICA solution obtained so far as an initial value

$$\mathbf{y}(\tau) = \mathbf{W}\mathbf{x}(\tau) = \mathbf{B}\mathbf{V}\mathbf{x}(\tau)$$

by employing another ICA algorithm that does not have the unitary constraint. Based on the use of InfoMax combined with the natural gradient, a separation matrix $\mathbf{W}$ is gradually improved by the learning rule:

$$\mathbf{W} \leftarrow \mathbf{W} + \mu \left[\mathbf{I} - \langle \mathbf{\Phi}(\mathbf{y}(\tau))\mathbf{y}(\tau)^H \rangle_\tau \right] \mathbf{W}, \tag{12}$$

where μ is a step-size parameter. $\mathbf{\Phi}(\mathbf{y}) = [\Phi(y_1), \ldots, \Phi(y_N)]^T$ is an element-wise nonlinear function defined by

$$\Phi(y_i) = -\frac{\partial}{\partial y_i} \log p(y_i), \tag{13}$$

where $p(y_i)$ is the probability density function (pdf) of a complex-valued signal $y_i = |y_i| e^{j \cdot \arg(y_i)}$. Since y_i is a frequency-domain signal whose phase can be shifted arbitrarily by shifting the STFT window position (4), a feasible assumption is that the pdf is independent of the phase $p(y_i) = \alpha \cdot p(|y_i|)$, where α is a constant. This assumption reduces (13) to

$$\Phi(y_i) = \varphi(|y_i|)\, e^{j \cdot \arg(y_i)}, \tag{14}$$

$$\varphi(|y_i|) = -\frac{\partial}{\partial |y_i|} \log p(|y_i|). \tag{15}$$

If we assume the Laplacian distribution $p(|y_i|) = \frac{1}{2} e^{-|y_i|}$, which is typical for speech modeling, we have $\varphi(|y_i|) = 1$ and thus a simple nonlinear function

$$\Phi(y_i) = e^{j \cdot \arg(y_i)}.$$

A nonlinear function of the form (14) has a better convergence property [22] than one where the nonlinearity is applied separately to the real and imaginary parts of a complex-valued signal y_i.

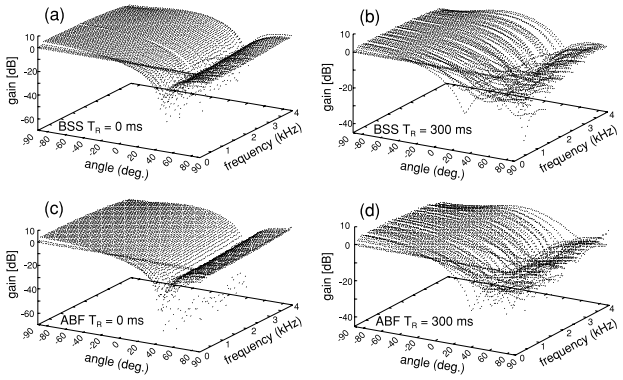

Fig. 5 Directivity patterns (a) obtained by BSS (T_R=0 ms), (b) obtained by BSS (T_R=300 ms), (c) obtained by ABF (T_R=0 ms), and (d) obtained by ABF (T_R=300 ms).

5. Separation Mechanism of BSS

The mechanism of BSS based on ICA has been shown to be equivalent to that of an adaptive microphone array system, i.e., N sets of adaptive beamformers (ABFs) with an adaptive null directivity aimed in the direction of unnecessary sounds [23], [24]. From the equivalence between BSS and ABF, it becomes clear that the physical behavior of BSS reduces the jammer signal by making a spatial null toward

the jammer, and extract the target.

The separation performance of BSS is compared with that of ABF. Figure 5 shows the directivity patterns obtained by BSS and ABF. In Fig. 5, (a) and (b) show directivity patterns by $\mathbf{W}$ obtained by BSS, and (c) and (d) show directivity patterns by $\mathbf{W}$ obtained by ABF. When $T_{\mathrm{R}} = 0$, a sharp spatial null is obtained by both BSS and ABF [see Figs. 5(a) and (c)]. When $T_{\mathrm{R}} = 300$ ms, the directivity pattern becomes duller for both BSS and ABF [see Figs. 5(b) and (d)].

BSS can be regarded as an intelligent version of ABF in the sense that it can adapt without any information on the source positions or period of source existence/absence [1].

6. Source Localization

This section presents a source localization method by analyzing the ICA solution (7) or equivalently (10). The information on source locations can be used to solve the permutation problem, as described in the next section. Many source localization methods have been proposed. A widely used method is MUSIC (MUltiple SIgnal Classification) [42], which employs subspace analysis with second-order statistics. The ICA-based method, on the other hand, employs higher-order statistics (or multiple second-order statistics based on non-stationarity). In this sense, the ICA-based method has certain advantages over the subspace-based method [43].

The source localization technique that employs ICA is a by-product of research on frequency-domain BSS. Direction-of-arrival (DOA) estimation methods [19]–[21] have been proposed that are based on beamforming theory [44]. They calculate directivity patterns as shown in Fig. 5 from the separation matrix $\mathbf{W}$, and then search the null directions, which correspond to the directions of sources [24]. However, it is simpler and more effective to estimate the directions directly from the basis vectors $\mathbf{a}_i$, which are given by the pseudoinverse of $\mathbf{W}$. The source localization method [25]–[27], [43] presented in this section is based on this idea. Such an idea was taken for granted in research on blind identification [45], [46], where the mixing system is estimated directly.

6.1 Basic Theory of Nearfield Model

Let us assume a mixing model that is suitable for source localization. Although the mixing model (1) in the time domain is a multi-path mixing model, we approximate the frequency response $h_{ji}(f)$ in (5) with a nearfield (direct-path) model (Fig. 6):

$$h_{ji}(f) \approx \frac{1}{\|\mathbf{q}_i - \mathbf{p}_j\|} e^{j2\pi f c^{-1}(\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i\|)}, \tag{16}$$

where $\mathbf{p}_j$ and $\mathbf{q}_i$ are 3-dimensional vectors representing the locations of sensor j and source i, respectively, and c is the propagation velocity of the signals. We assume that the amplitude is attenuated based on the distance $\|\mathbf{q}_i - \mathbf{p}_j\|$. We also assume that the phase depends on the difference between the distances $\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i\|$ from the source to the sensor and to the origin $\mathbf{o} = [0, 0, 0]^T$. This makes the phase zero at the origin. If the phase $2\pi f c^{-1}(\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i\|)$ is outside the range $(-\pi, \pi)$, this model suffers from spatial aliasing. Therefore, the model is feasible as long as the condition

$$f < \left| \frac{c}{2 \cdot (\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i\|)} \right|$$

is satisfied.

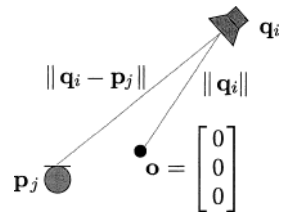

Fig. 6 Nearfield (direct-path) model.

The ICA-based source localization discussed in this section estimates the location $\mathbf{q}_i$ of source i from information on sensor locations $\mathbf{p}_j$ and the separation matrix $\mathbf{W}(f)$ obtained by ICA (7). Let us assume here that the decomposition (10) of observations $\mathbf{x}(f, \tau)$ has been obtained in each frequency bin by the pseudoinverse of $\mathbf{W}(f)$. By comparing (6) and (10), we observe the following fact. If the ICA algorithm works well and the outputs $y_1, \ldots, y_N$ are the estimation of the sources $s_1, \ldots, s_N$, then the basis vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N$ are also estimations of the mixing vectors $\mathbf{h}_1, \ldots, \mathbf{h}_N$ up to the permutation and scaling ambiguity.

Following the model (16), the ratio between two elements $a_{ji}, a_{j'i}$ of the same basis vector $\mathbf{a}_i$ provides the key equation for source localization:

$$\begin{aligned} \frac{a_{ji}}{a_{j'i}} &= \frac{\alpha_i h_{ji}}{\alpha_i h_{j'i}} \\ &= \frac{\|\mathbf{q}_i - \mathbf{p}_{j'}\|}{\|\mathbf{q}_i - \mathbf{p}_j\|} e^{j2\pi f c^{-1}(\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i - \mathbf{p}_{j'}\|)}, \end{aligned} \tag{17}$$

where the scaling ambiguity α_i is canceled out by calculating the ratio. The permutation ambiguity still remains. However, if we estimate the location $\mathbf{q}_i$ for all $i = 1, \ldots, N$, the set of all estimated locations does not depend on the permutation.

With respect to the phase differences, the set of vectors $\mathbf{q}_i$ in the argument of (17),

$$\|\mathbf{q}_i - \mathbf{p}_j\| - \|\mathbf{q}_i - \mathbf{p}_{j'}\| = \frac{\arg(a_{ji}/a_{j'i})}{2\pi f c^{-1}}, \tag{18}$$

defines a surface where the difference between the distances from $\mathbf{p}_j$ and $\mathbf{p}_{j'}$ is constant. The surface is one sheet of a two-sheet hyperboloid.

Alternatively, with respect to the level differences, the set of vectors $\mathbf{q}_i$ in the modulus of (17),

$$\frac{\|\mathbf{q}_i - \mathbf{p}_{j'}\|}{\|\mathbf{q}_i - \mathbf{p}_j\|} = \left| \frac{a_{ji}}{a_{j'i}} \right|, \tag{19}$$

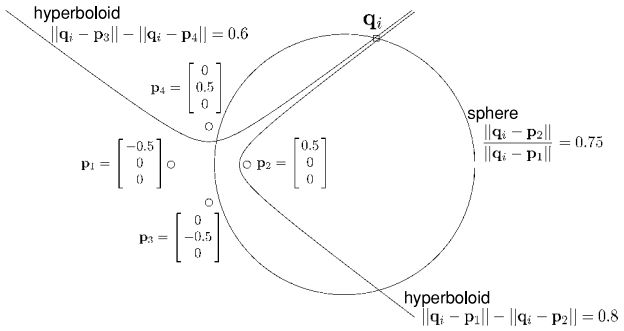

Fig. 7 Source localization by intersection of two hyperboloids and a sphere.

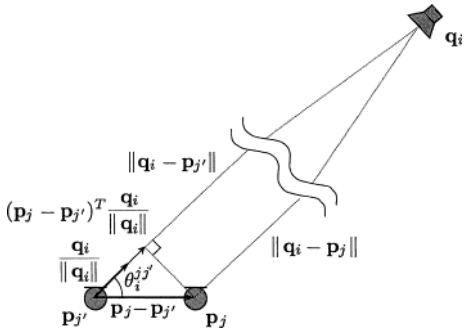

Fig. 8 Farfield model.

defines a sphere where the ratio of the distances from $\mathbf{p}_j$ and $\mathbf{p}_{j'}$ is constant. Therefore, with these two Eqs. (18) and (19), we can estimate the possible location $\mathbf{q}_i$ of source s_i. Such hyperboloid and sphere are defined by a pair of sensors j and j'. If we select another pair of sensors, a different hyperboloid and sphere are obtained. In this way, the location $\mathbf{q}_i$ is estimated as the intersection of several hyperboloids and spheres. An example is shown in Fig. 7.

6.2 DOA Estimation with Farfield Model

Although it is useful to estimate a 3-dimensional location, calculating the intersections of hyperboloids and spheres is computationally demanding. In many cases it is sufficient to estimate just the direction-of-arrival (DOA) of source s_i. If we assume the source location $\mathbf{q}_i$ is far from sensors $\mathbf{p}_j$ and $\mathbf{p}_{j'}$, (18) can be approximated as a farfield model (Fig. 8):

$$(\mathbf{p}_j - \mathbf{p}_{j'})^T \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|} = \frac{\arg(a_{ji}/a_{j'i})}{2\pi f c^{-1}}, \tag{20}$$

and the cosine of angle $\theta_i^{jj'}$ between the two vectors $\mathbf{q}_i$ and $\mathbf{p}_j - \mathbf{p}_{j'}$ can be calculated as

$$\begin{aligned} \cos\theta_i^{jj'} &= \frac{(\mathbf{p}_j - \mathbf{p}_{j'})^T \mathbf{q}_i}{\|\mathbf{p}_j - \mathbf{p}_{j'}\| \cdot \|\mathbf{q}_i\|} \\ &= \frac{\arg(a_{ji}/a_{j'i})}{2\pi f c^{-1} \|\mathbf{p}_j - \mathbf{p}_{j'}\|}. \end{aligned} \tag{21}$$

The set of vectors $\mathbf{q}_i$ that satisfy (20) represents a cone

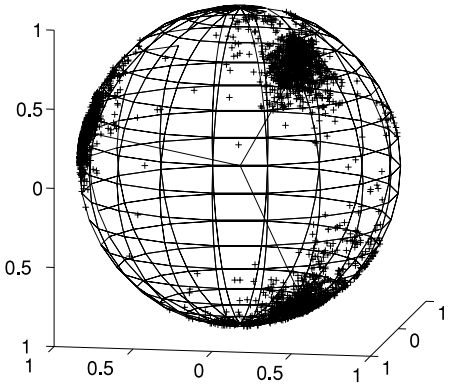

Fig. 9 3-dimensional arrangement of eight microphones and three loudspeakers (upper) and DOA estimation results for this case (lower).

[26], which is the asymptotic surface of the corresponding hyperboloid (18). To estimate the DOA of a source, the intersections of several cones should be obtained. Let us assume that we select u cones whose corresponding sensor pairs are $(j_1, j'_1), \ldots, (j_u, j'_u)$. The set of Eqs. (20) for u sensor pairs is represented as

$$\mathbf{D}\frac{\mathbf{q}_i}{\|\mathbf{q}_i\|} = \frac{\mathbf{r}_i}{2\pi f c^{-1}}, \tag{22}$$

where

$$\begin{aligned} \mathbf{D} &= [\mathbf{p}_{j_1} - \mathbf{p}_{j'_1}, \ldots, \mathbf{p}_{j_u} - \mathbf{p}_{j'_u}]^T, \\ \mathbf{r}_i &= [\arg(a_{j_1 i}/a_{j'_1 i}), \ldots, \arg(a_{j_u i}/a_{j'_u i})]^T. \end{aligned}$$

In practical situations, there is no exact solution for (22) because the u conditions do not coincide exactly. Therefore, we typically solve it in the least-square sense by using the Moore-Penrose pseudoinverse [27]:

$$\frac{\mathbf{q}_i}{\|\mathbf{q}_i\|} = \frac{\mathbf{D}^+ \mathbf{r}_i}{2\pi f c^{-1}}. \tag{23}$$

If rank($\mathbf{D}$) $\geq$ 3, the set of vectors $\mathbf{q}_i$ that satisfy (23) represents a line in 3-dimensional space, which represents the DOA of a source i.

The upper photo in Fig. 9 shows the case where eight microphones and three loudspeakers are arranged 3-dimensionally, and the lower plot shows the DOA estimation results for this case. Each point shows a location vector

$\mathbf{q}_i(f)$ that is normalized to unit norm $\mathbf{q}_i(f) \leftarrow \frac{\mathbf{q}_i(f)}{\|\mathbf{q}_i(f)\|}$. The estimations are obtained for all frequencies f and all output indexes i. As shown in the plot, they form clusters, each of which corresponds to the location of each source.

If the sensor and source locations are limited to a 2-dimensional plane, the dimensionality of location vectors, such as $\mathbf{p}_i$ and $\mathbf{q}_i$, can be reduced to two. In this case, rank($\mathbf{D}$) ≥ 2 is sufficient to reach a solution in (23). Moreover, the DOA of source i can be represented simply by the angle θ_i that satisfies

$$\mathbf{q}_i = [\cos(\theta_i), \sin(\theta_i)]^T, \quad -180^\circ < \theta_i \leq 180^\circ. \tag{24}$$

Figure 15 shows the case where the sensor and source locations are limited to 2-dimensions. The DOA estimations in this case are shown in Figs. 16 and 17.

If the sensors are arranged linearly and the potential source location is in a 2-dimensional half-plane, which is to one side of the sensor arrangement line, the angle $\theta_i^{jj'}$ ($0^\circ \leq \theta_i^{jj'} \leq 180^\circ$) by (21) provides sufficient information on the source location. For example, Fig. 13 shows DOA estimation results for such a case with the conditions shown in Fig. 12.

7. Permutation Alignment

This section discusses how to solve the permutation problem. Various methods have already been proposed. With reference to the ICA Eq. (7) as well as to the decomposition (10) of observations $\mathbf{x}(f, \tau)$, we classify these methods into four categories based on the following strategies:

1. Applying an operation to the separation matrix $\mathbf{W}(f)$,
2. Utilizing the information on the separation matrix $\mathbf{W}(f)$ itself,
3. Utilizing the information on the basis vectors $\mathbf{a}_1(f), \ldots, \mathbf{a}_N(f)$,
4. Utilizing the information on the separated signals $y_1(f, \tau), \ldots, y_N(f, \tau)$.

The operation of the first strategy basically involves smoothing the separation matrices in the frequency domain. This has been realized by reducing the filter length by rectangular windowing in the time domain [10], [13]–[15], or by averaging the separation matrices with adjacent frequencies [13]. However, this operation makes the separation matrix $\mathbf{W}(f)$ different from the ICA solution (7), which may have a detrimental effect on the separation performance. A possible way to solve this problem is to interleave the ICA update, e.g., (12), and this operation until convergence. In this sense, this strategy is related to the third approach to BSS discussed in the Introduction.

The second category includes the beamforming approach [19]–[21], where the directivity patterns formed by the separation matrix are analyzed to identify the DOA of each source. The third category includes an approach that utilizes the results of source localization with the basis vectors [25]–[27], [46]. The theory and operation for source localization were discussed in Sect. 6. These two approaches from the second and the third categories utilize basically the same information because the separation matrix $\mathbf{W}(f)$ and the basis vectors $\mathbf{a}_1(f), \ldots, \mathbf{a}_N(f)$ are directly connected by the pseudoinverse operation (8). However, the information used in the third category is easier to handle since it directly represents the mixing system (6). The last category includes an approach that employs the inter-frequency correlations of output signal envelopes [16], [17]. This is particularly effective for a non-stationary signal such as speech.

In the next two subsections, we explain the approaches of the third and the fourth categories, respectively. Since these two approaches have different but complementary characteristics, integrating them is a good way to pursue a better solution to the permutation problem [25]. Subsection 7.3 presents a method that effectively integrates the two approaches to solve the permutation problem in a better way. In the following subsections, let Π_f be a permutation corresponding to the inverse $\mathbf{P}^{-1}(f)$ of the permutation matrix of (11). The permutation problem can be formulated to obtain Π_f for every frequency f, which is a mapping from source index k to output index i:

$$i = \Pi_f(k).$$

7.1 Localization Approach

The basic idea of this approach is to estimate the locations of sources and then cluster them to decide the permutation. ICA-based source localization (Sect. 6) estimates the location $\mathbf{q}_i(f)$ of a source that corresponds to the i-th basis vector $\mathbf{a}_i(f)$ for each frequency f. Let the following function *localize* estimate the location in this way:

$$\mathbf{q}_i(f) = \mathit{localize}(f, \mathbf{a}_i(f)).$$

If just the DOA estimation is adequate, the location vector $\mathbf{q}_i(f)$ should be normalized to the unit norm $\mathbf{q}_i(f) \leftarrow \frac{\mathbf{q}_i(f)}{\|\mathbf{q}_i(f)\|}$. If the locations of sensors and sources are limited to a 2-dimensional plane, we simply obtain $\theta_i(f)$ that satisfies (24) as a DOA estimation.

Then, we employ a clustering algorithm to find N clusters $C_1, \ldots, C_N$ formed by estimated locations $\mathbf{q}_i(f)$ or $\theta_i(f)$. Each C_k corresponds to the location of source k. Let the following function *clustering* perform clustering for all of the estimated locations $\mathbf{q}_i(f)$ and return the centroid $\mathbf{c}_k$ and the variance σ_k^2 of each cluster C_k:

$$[\mathbf{c}_1, \sigma_1, \ldots, \mathbf{c}_N, \sigma_N] = \mathit{clustering}\,(^\forall f,\ \mathbf{q}_1(f), \ldots, \mathbf{q}_N(f)),$$

$$\mathbf{c}_k = \sum_{\mathbf{q} \in C_k} \frac{\mathbf{q}}{|C_k|},$$

$$\sigma_k^2 = \sum_{\mathbf{q} \in C_k} \frac{\|\mathbf{c}_k - \mathbf{q}\|^2}{|C_k|},$$

where $|C_k|$ is the number of vectors in the cluster. The optimization criterion for clustering is to minimize the total sum $\sum_{k=1}^{N} \sigma_k^2$ of the variances. This optimization is efficiently

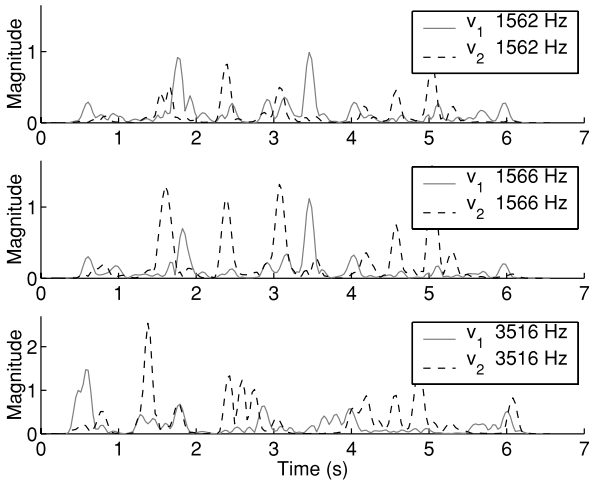

Fig. 10 Envelopes of two output signals at different frequencies.

performed with the k-means clustering algorithm [47]. Once we have N clusters, permutations for all frequencies f can be decided by

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

The advantage of this source localization approach is that it is very simple to decide the permutation Π_f for each frequency once the centroids of N clusters are obtained. However, the downside of this approach is that the estimated locations or DOAs, and thus the permutations Π_f are not accurate for some frequencies. Such situations typically happen at low frequencies, where the phase difference caused by the sensor spacing is very small, as shown in Fig. 13.

7.2 Correlation Approach

This subsection presents an approach to permutation alignment based on the inter-frequency correlation of separated signals. The correlation should be calculated for the amplitude $|y_i(f,\tau)|$ or (log-scaled) power $|y_i(f,\tau)|^2$ of separated signals. The correlation of raw complex-valued signals $y_i(f,\tau)$ would be very low due to the STFT property. Here, we use the amplitude (so-called envelope)

$$v_i^f(\tau) = |y_i(f,\tau)|$$

of a separated signal $y_i(f,\tau)$. The correlation of two sequences $x(\tau)$ and $y(\tau)$ is usually calculated by the correlation coefficient

$$\mathrm{cor}(x,y) = (\mu_{x\cdot y} - \mu_x \cdot \mu_y)/(\sigma_x \cdot \sigma_y),$$

where μ_x is the mean and σ_x is the standard deviation of x. Based on this definition, $\mathrm{cor}(x,x) = 1$, and $\mathrm{cor}(x,y) = 0$ if x and y are uncorrelated.

Envelopes have high correlations at neighboring frequencies if separated signals correspond to the same source signal. Figure 10 shows an example. Two envelopes v_1^{1562} and v_1^{1566}, as well as v_2^{1562} and v_2^{1566}, are highly correlated. Thus, calculating such correlations helps us to align permutations.

A simple criterion for deciding Π_f is to maximize the sum of the correlations between neighboring frequencies within distance δ:

$$\Pi_f = \mathrm{argmax}_\Pi \sum_{|g-f|\le\delta} \sum_{i=1}^{N} \mathrm{cor}(v_{\Pi(i)}^f, v_{\Pi_g(i)}^g), \tag{26}$$

where Π_g is the permutation at frequency g. This criterion is based on local information and has a drawback in that mistakes in a narrow range of frequencies may lead to the complete misalignment of the frequencies beyond the range.

To avoid this problem, the method in [17] does not limit the frequency range in which correlations are calculated. It decides permutations one by one based on the criterion

$$\Pi_f = \mathrm{argmax}_\Pi \sum_{i=1}^{N} \mathrm{cor}\left(v_{\Pi(i)}^f, \sum_{g\in\mathcal{F}} v_{\Pi_g(i)}^g\right),$$

where $\mathcal{F}$ is a set of frequencies in which the permutation is decided. This method assumes high correlations of envelopes even between frequencies that are not close neighbors. This assumption is not satisfied for all pairs of frequencies, e.g., v_i^{1566} and v_i^{3516} in Fig. 10 do not have a high correlation. Therefore, this method still has the drawback of permutations possibly being misaligned at many frequencies.

If a source signal has a harmonic structure, as in the case of speech, there are strong correlations between the envelopes of a fundamental frequency f and its harmonics $2f$, $3f$, ... Therefore, maximizing the correlation among harmonics is another idea for permutation alignment [25]:

$$\Pi_f = \mathrm{argmax}_\Pi \sum_{g\in\mathcal{H}(f)} \sum_{i=1}^{N} \mathrm{cor}(v_{\Pi(i)}^f, v_{\Pi_g(i)}^g), \tag{27}$$

where $\mathcal{H}(f)$ provides a set of harmonic frequencies of f. The permutation accuracy improves if we take the signal's harmonic structure into consideration. However, maximizing (26) and (27) simultaneously is not very straightforward and is computationally expensive.

7.3 Integrated Method

This subsection presents a method that integrates the two approaches discussed in the last two subsections. The intention behind this integration is to solve the permutation problem robustly and precisely. Let us review the characteristics of the above two approaches.

- **robustness**: The localization approach is robust since a misalignment at one frequency does not affect other frequencies. The correlation approach is not robust since a misalignment at one frequency affects the results of other frequencies and may cause consecutive misalignments.
- **preciseness**: The localization approach is not precise

since the evaluation is based on a direct-path approximation (16) of the mixing system. The correlation approach is precise as long as signals are well separated by ICA, since the measurement is based on the separated signals themselves.

To benefit from both advantages, namely the robustness of the localization approach and the preciseness of the correlation approach, the integrated method first decides permutations with the localization approach and then refines the solution with the correlation approach. An implementation of the integrated method consists of the following four steps [25]:

1. Decide the permutations by the localization approach (25) at certain frequencies where the confidence of source localization is sufficiently high,
2. Decide the permutations based on neighboring correlations (26) as long as the criterion gives a clear-cut decision,
3. Decide the permutations at certain frequencies where the correlation among harmonics (27) is sufficiently high,
4. Decide the permutations for the remaining frequencies based on neighboring correlations (26).

The key to the first step is fixing a permutation only if the confidence of source localization is sufficiently high. We assume that the confidence is high if the squared distance between an estimated location and its corresponding centroid is smaller than the variance, i.e., $\|\mathbf{c}_k - \mathbf{q}_{\Pi(k)}(f)\|^2 < \sigma_k^2$. In the second step, permutations are decided one by one for the frequency f where the sum of the correlations with fixed frequencies $g \in \mathcal{F}$ within distance $|g - f| \leq \delta$ is the maximum. This is repeated as long as the maximum correlation sum is larger than a threshold th_{cor}. In the third step, the permutations are decided for frequencies f where the sum of the correlations among harmonics is larger than a threshold th_{ha}. The last step decides the permutations for the remaining frequencies with the same criterion as the second step.

Let us discuss the advantages of the integrated method. The main advantage is that it does not cause a large misalignment as long as the permutations fixed by the localization approach are correct. Moreover, the correlation part compensates for the lack of preciseness of the localization approach. The correlation part consists of three steps (step 2,3,4) for two reasons. First, the harmonics part works well if most of the other permutations are fixed. Second, the method becomes more robust by quitting step 2 if there is no clear-cut decision. With this structure, we can avoid fixing the permutations for consecutive frequencies without high confidence. As shown in the experimental results (Sect. 10), this integrated method is effective in separating many sources.

8. Scaling Alignment

The scaling ambiguity $\mathbf{\Lambda}(f)$ in (11) is easily solved by calculating the (pseudo)-inverse of a separation matrix $\mathbf{W}(f)$ [8],

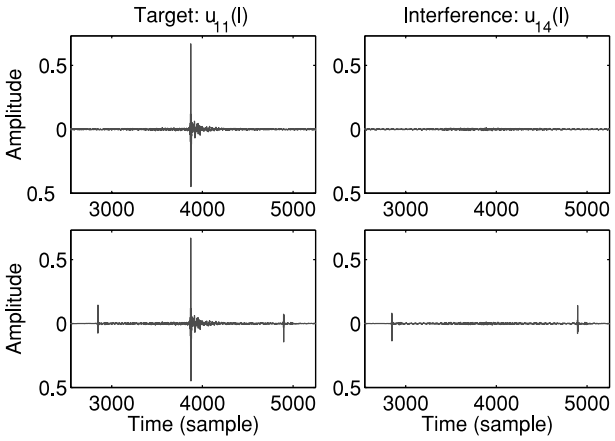

Fig. 11 Impulse responses $u_{ik}(l)$ obtained with periodic filters (above) and with their one-period realization (below).

[17]. The frequency-domain counterpart of the BSS goal (3) is

$$y_i(f, \tau) = h_{J_i i}(f) s_i(f, \tau), \tag{28}$$

where J_i can be selected according to each output i but should be the same for all frequencies f. Let us assume that the ICA and the permutation problem have been solved. Then the $\mathbf{a}_i$ term in (10) is close to the $\mathbf{h}_i$ term in (6):

$$\mathbf{h}_i(f) s_i(f, \tau) \approx \mathbf{a}_i(f) y_i(f, \tau). \tag{29}$$

By substituting (28) into (29), we have the condition for scaling alignment:

$$\mathbf{h}_i(f) \approx \mathbf{a}_i(f) h_{J_i i}(f) \Leftrightarrow a_{J_i i}(f) \approx 1.$$

This condition, i.e., $a_{J_i i}(f) = 1$, is attained by

$$\mathbf{W}(f) \leftarrow \mathbf{\Lambda}(f)\mathbf{W}(f),$$
$$\mathbf{\Lambda}(f) = \mathrm{diag}(a_{J_1 1}(f), \ldots, a_{J_N N}(f)),$$

where $a_{ji}(f) = [\mathbf{W}^+(f)]_{ji}$ is an element of the pseudoinverse of $\mathbf{W}(f)$.

9. Spectral Smoothing

The frequency-domain BSS described in this paper is influenced by the circularity of discrete frequency representation. The circularity refers to the fact that frequency responses sampled at L points with an interval f_s/L (f_s: sampling frequency) represent a periodic time-domain signal whose period is L/f_s. Since this filter is unrealistic, we usually use its one-period realization. However, such one-period filters may cause a problem. Figure 11 shows impulse responses from a source $s_k(t)$ to an output $y_i(t)$ defined by (39). Responses on the left $u_{11}(l)$ correspond to the extraction of a target signal, and those on the right $u_{14}(l)$ correspond to the suppression of an interference signal. The upper responses are obtained with infinite-length filters, and the lower ones with one-period filters. We can see that the one-period filters create spikes, which distort the target signal and degrade

the separation performance. Note that these spikes are inevitable in the frequency-domain BSS since we have an ICA solution in the frequency domain.

9.1 Windowing

To solve this problem, we need to control the frequency responses $w_{ij}(f)$ so that the corresponding time-domain filter $\mathrm{w}_{ij}(l)$ does not rely on the circularity effect whereby adjacent periods work together to perform some filtering. The most widely used approach is spectral smoothing, which is realized by multiplying a window $\mathrm{g}(l)$ that tapers smoothly to zero at each end, such as a Hanning window $\mathrm{g}(l) = \frac{1}{2}(1 + \cos\frac{2\pi l}{L})$. This makes the resulting time-domain filter $\mathrm{w}_{ij}(l) \cdot \mathrm{g}(l)$ fit length L and have small amplitude around the ends [18]. As a result, the frequency responses $w_{ij}(f)$ are smoothed as

$$\tilde{w}_{ij}(f) = \sum_{\phi=0}^{f_s-\Delta f} g(\phi) w_{ij}(f-\phi),$$

where $g(f)$ is the frequency response of $\mathrm{g}(l)$ and $\Delta f = \frac{f_s}{L}$. If a Hanning window is used, the frequency responses are smoothed as

$$\tilde{w}_{ij}(f) = \frac{1}{4}\left[w_{ij}(f-\Delta f) + 2w_{ij}(f) + w_{ij}(f+\Delta f)\right], \quad (30)$$

since the frequency responses $g(f)$ of the Hanning window are $g(0) = \frac{1}{2}$, $g(\Delta f) = g(f_s - \Delta f) = \frac{1}{4}$, and zero for the other frequency bins.

The windowing successfully eliminates the spikes. However, it changes the frequency response from $w_{ij}(f)$ to $\tilde{w}_{ij}(f)$ and causes an error. Let us evaluate the error for each row $\mathbf{w}_i(f) = [w_{i1}(f), \ldots, w_{iM}(f)]^T$ of the ICA solution $\mathbf{W}(f)$. The error is

$$\begin{aligned}\mathbf{e}_i(f) &= \min_{\alpha_i}[\tilde{\mathbf{w}}_i(f) - \alpha_i \mathbf{w}_i(f)] \\ &= \tilde{\mathbf{w}}_i(f) - \frac{\tilde{\mathbf{w}}_i(f)^H \mathbf{w}_i(f)}{\|\mathbf{w}_i(f)\|^2}\mathbf{w}_i(f), \end{aligned} \quad (31)$$

where $\tilde{\mathbf{w}}_i(f) = [\tilde{w}_{i1}(f), \ldots, \tilde{w}_{iM}(f)]^T$ and α_i is a complex-valued scalar representing the scaling ambiguity of the ICA solution. The minimization $\min_{\alpha_i}$ is based on least-squares, and can be represented by the projection of $\tilde{\mathbf{w}}_i$ to $\mathbf{w}_i$. We can evaluate the error for the Hanning window case by substituting (30) for $\tilde{\mathbf{w}}$ of (31):

$$\mathbf{e}_i(f) = \frac{1}{4}\left[\,\mathbf{e}_i^-(f) + \mathbf{e}_i^+(f)\,\right], \quad (32)$$

where

$$\mathbf{e}_i^-(f) = \mathbf{w}_i(f-\Delta f) - \frac{\mathbf{w}_i(f-\Delta f)^H \mathbf{w}_i(f)}{\|\mathbf{w}_i(f)\|^2}\mathbf{w}_i(f), \quad (33)$$

$$\mathbf{e}_i^+(f) = \mathbf{w}_i(f+\Delta f) - \frac{\mathbf{w}_i(f+\Delta f)^H \mathbf{w}_i(f)}{\|\mathbf{w}_i(f)\|^2}\mathbf{w}_i(f). \quad (34)$$

This $\mathbf{e}_i^-$ (or $\mathbf{e}_i^+$) represents the difference between two vectors $\mathbf{w}_i(f)$ and $\mathbf{w}_i(f-\Delta f)$ (or $\mathbf{w}_i(f+\Delta f)$). Since these differences are usually not very large, the error $\mathbf{e}_i$ does not seriously affect the separation if we use a Hanning window for spectral smoothing.

9.2 Minimizing Error by Adjusting Scaling Ambiguity

Even if the error caused by the windowing is not very large, the separation performance is improved by minimizing the error [29]. The minimization is performed by adjusting the scaling ambiguity of the ICA solution before the windowing. Let $d_i(f)$ be a complex-valued scalar for the scaling adjustment:

$$\mathbf{w}_i(f) \leftarrow d_i(f)\mathbf{w}_i(f). \quad (35)$$

We want to find $d_i(f)$ such that the error (31) is minimized. The scalar $d_i(f)$ should be close to 1 to avoid any great change in the predetermined scaling. Thus, an appropriate total cost to be minimized is

$$\mathcal{J} = \sum_f J_i(f), \quad J_i(f) = \frac{\|\mathbf{e}_i(f)\|^2}{\|\mathbf{w}_i(f)\|^2} + \beta|d_i(f) - 1|^2,$$

where β is a parameter indicating the importance of maintaining the predetermined scaling. With the Hanning window, the error after the scaling adjustment is easily calculated by substituting (35) for (32):

$$\mathbf{e}_i(f) = \frac{1}{4}\left[\,d_i(f-\Delta f)\mathbf{e}_i^-(f) + d_i(f+\Delta f)\mathbf{e}_i^+(f)\,\right], \quad (36)$$

where $\mathbf{e}_i^-$ and $\mathbf{e}_i^+$ are defined in (33) and (34), respectively.

The minimization of the total cost can be performed iteratively by

$$d_i(f) = d_i(f) - \mu\frac{\partial \mathcal{J}}{\partial d_i(f)} \quad (37)$$

with a small step-size μ. With the Hanning window, the gradient is

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial d_i(f)} &= \frac{\partial J_i(f-\Delta f)}{\partial d_i(f)} + \frac{\partial J_i(f+\Delta f)}{\partial d_i(f)} + \frac{\partial J_i(f)}{\partial d_i(f)} \\ &= \frac{\mathbf{e}_i(f-\Delta f)^H \mathbf{e}_i^+(f-\Delta f) + \mathbf{e}_i(f+\Delta f)^H \mathbf{e}_i^-(f+\Delta f)}{8 \cdot \|\mathbf{w}_i(f)\|^2} \\ &\quad +2\beta(d_i(f) - 1). \end{aligned} \quad (38)$$

With equations from (36) to (38), we can optimize the scalar $d_i(f)$ for the scaling adjustment, and minimize the error caused by the spectral smoothing (30) with the Hanning window.

10. Experimental Results

The performance of BSS is evaluated by a signal-to-interference ratio (SIR), which is the power ratio between the target component and the interference components. Let $\mathrm{u}_{ik}(l)$ be the impulse responses from source $\mathrm{s}_k(t)$ to separated signal $\mathrm{y}_i(t)$:

Table 1 Separation performance with linear array.

#sources / position	2 / a c		3 / a b d		4 / a b c d	
Spectral smoothing	no	yes	no	yes	no	yes
Average SIR at microphones (dB)	0.1		-2.9		-4.6	
Average SIR of output (dB)	20.1	22.3	14.7	17.0	9.3	11.5
Execution time (s)	5.2	5.2	8.0	8.1	12.3	12.4

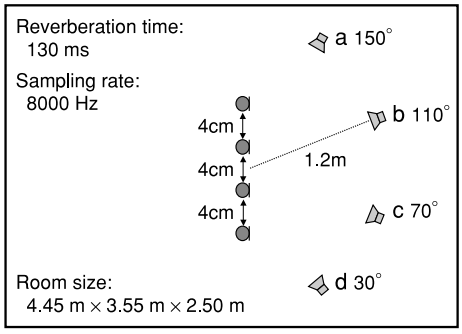

Fig. 12 Experimental conditions with linear array.

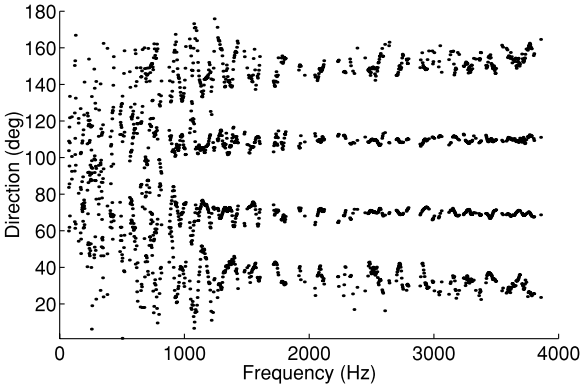

Fig. 13 DOA estimations by (21) with four sources.

$$\mathrm{u}_{ik}(l) = \sum_{j=1}^{M} \sum_{\tau=0}^{L-1} \mathrm{w}_{ij}(\tau)\mathrm{h}_{jk}(l-\tau). \tag{39}$$

Then, the SIR of output i is calculated as

$$\mathrm{SIR}_i = 10\log_{10} \frac{\langle |\sum_l \mathrm{u}_{ii}(l)\mathrm{s}_i(t-l)|^2 \rangle_t}{\langle |\sum_{k\neq i}\sum_l \mathrm{u}_{ik}(l)\mathrm{s}_k(t-l)|^2 \rangle_t}(\mathrm{dB}), \tag{40}$$

where $\langle\cdot\rangle_t$ denotes the averaging operator over time t.

10.1 Linear Array

We performed experiments to separate speech signals in an environment whose conditions are summarized in Fig. 12. Our experiments involved two, three and four sources whose locations are indicated in Table 1. The sensors were arranged linearly, and the number of sensors used was the same as the number of sources. We used filters of length $L = 2048$ because this length provided the best performance under the conditions. The BSS program was coded in Matlab and run on Athlon XP 3200+.

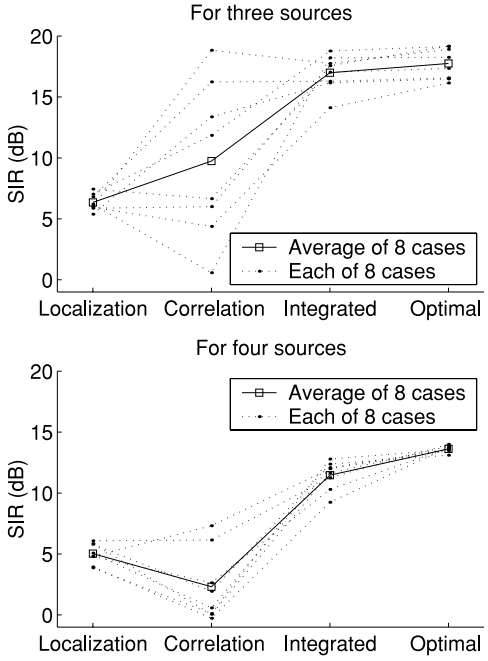

Fig. 14 Comparison of different methods for solving permutation problem.

The results shown in Table 1 are the average SIRs of output for eight combinations of 7-second speeches. We can see that the spectral smoothing discussed in Sect. 9 improves the average SIR for every setup. The short execution time, as shown in Table 1, enables the BSS system to perform in real time if the number of source signals is not very large.

Figure 13 shows DOA estimations for mixtures of four sources obtained with (21). Figure 14 shows SIRs for three and four sources with the different methods for solving the permutation problem discussed in Sect. 7. Here, "Localization" is the localization (DOA) approach (25) alone, "Correlation" is the correlation approach (26) alone, "Integrated" is the integrated method, and "Optimal" is the optimal solution obtained by utilizing the $\mathrm{s}_i(t)$ and $\mathrm{h}_{ji}(l)$ information. The performance of "Localization" was stable but insufficient. The performance of "Correlation" was unstable and very poor in the four-source cases. The "Integrated" method performed very well and was close to "Optimal."

10.2 Planar Array

Next, we carried out experiments on separating six sources with a planar array of eight microphones. The room layout and other experimental conditions are shown in Fig. 15. All six sources were 6-second speech signals, and two came

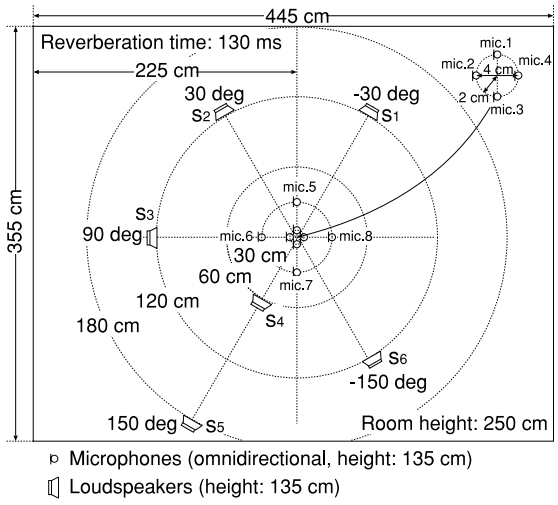

Fig. 15 Experimental conditions for planar array case.

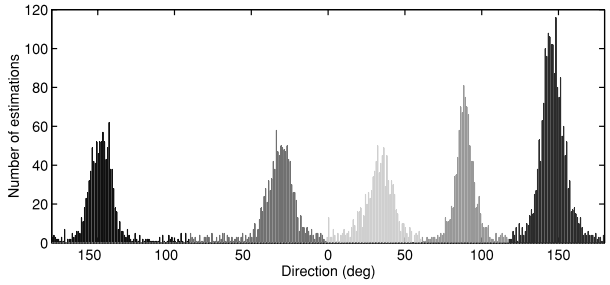

Fig. 16 Histogram of DOAs estimated with small spacing microphone pairs.

from the same direction. The filter length was again $L = 2048$ for an 8-kHz sampling rate.

Let us explain the method for solving the permutation problem in this situation. First, the source directions were estimated with small-spacing microphone pairs (1-3, 2-4, 1-2 and 2-3 shown in the right-top corner of Fig. 15). This was performed based on (20), (22) and (23). Figure 16 shows a histogram of the estimated DOAs. There are five clusters in this histogram, and one cluster is twice the size of the others. This implies that two sources came from the same direction (about 150°). We solved the permutation problem for the other four sources by using this DOA information as shown on the upper plot of Fig. 17.

Then, to distinguish between the two sources that came from the same direction, the spheres of these sources were estimated with large-spacing microphone pairs (7-5, 7-8, 6-5 and 6-8 shown in the center of Fig. 15). This was performed based on (19). The lower plot of Fig. 17 shows the radiuses of the spheres estimated with microphone pair 7-5. Although the radius estimations had large variances, it provided sufficient information to distinguish between the two sources. Consequently, the signal components of all frequencies were classified into six clusters. We decided the permutation only for frequency bins where the classification was reliable, as discussed in Sect. 7.3.

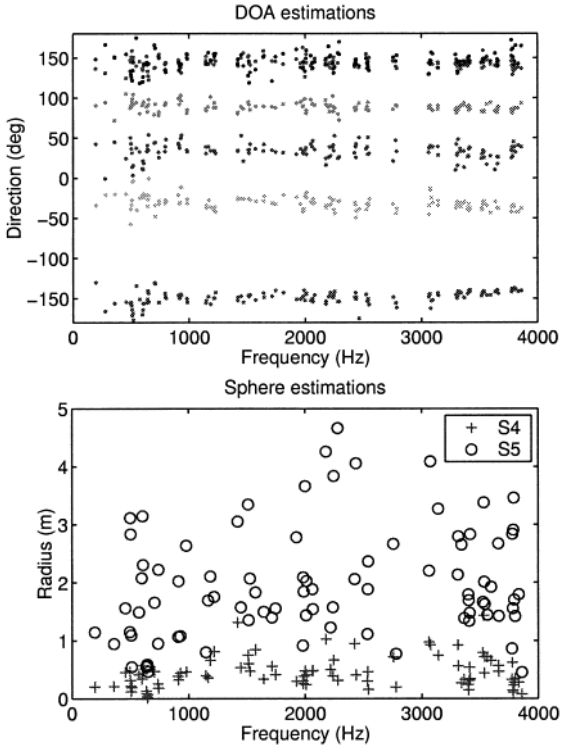

Fig. 17 Permutation solved by using estimated DOAs (upper) and spheres (lower).

To show the effectiveness of this method, we compared SIRs by three different methods for the permutation problem. Table 2 shows the results. The last row, "DOA + Sphere + Correlation," shows the results obtained with the integrated method. The two methods for comparison were "Correlation" where only the correlations (26) were maximized, and "DOA + Correlation" where only the DOA information was used for the source localization step in the integrated method. To see how much the SIRs were improved, we also measured the SIR of the mixture observed at microphone 1 ("SIR at microphone 1"). The effectiveness of the two integrated methods can again be observed. If we compare the results of "DOA + Correlation" with "DOA + Sphere + Correlation," the improvement of the latter over the former is apparent for sources 4 and 5, which came from the same direction. This means that the sphere information was important in terms of distinguishing between sources coming from the same direction. The BSS program was again coded in Matlab and run on Athlon XP 3200+. The computational time for separating six speeches of 6 seconds was around one minute.

10.3 Moving Sources

In most realistic applications, the source location may change. A mixing system is time-varying when source signals move. A naive approach for tracking a time-varying system is an online algorithm that updates the separation system sample by sample [48], [49].

Table 2 Separation performance with planar array measured by SIR (dB).

	SIR_1	SIR_2	SIR_3	SIR_4	SIR_5	SIR_6	average
SIR at microphone 1	-8.3	-6.8	-7.8	-7.7	-6.7	-5.2	-7.1
Correlation	4.4	2.6	4.0	9.2	3.6	-2.0	3.7
DOA + Correlation	9.6	9.3	14.7	2.7	6.5	14.0	9.4
DOA + Sphere + Correlation	10.8	10.4	14.5	7.0	11.0	12.2	11.0

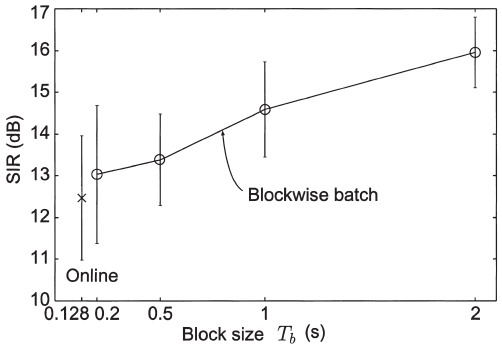

Fig. 18 Average and standard deviation of SIR for fixed sources.

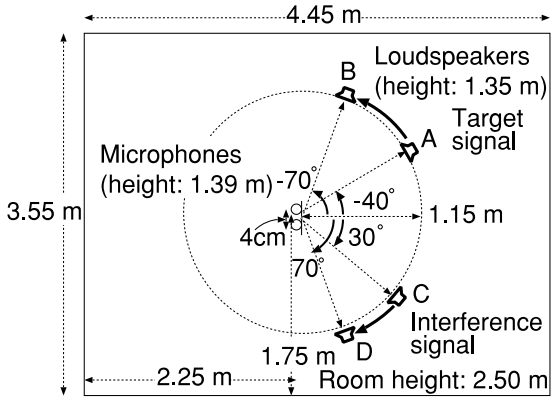

Fig. 19 Layout of room used in experiments. T_R = 130 ms.

Indeed, an online algorithm can track a time-varying system; however, its performance is generally worse than a batch algorithm, which can employ a number of samples, when the system is stationary. Although we are dealing with moving sources, we do not want to degrade the performance for fixed sources.

In this section, we describe a real-time BSS method [50] that employs frequency domain ICA with a blockwise batch algorithm. This algorithm achieves better separation performance than an online algorithm for fixed source signals.

We measured the BSS performance using ICA. Figure 18 shows the average and standard deviation of SIR for fixed sources (the target is at A and the interference at C in Fig. 19). This indicates that the blockwise batch algorithm outperforms the online algorithm (in which μ is tuned to optimize the performance) when we use the update Eq. (12). In addition, the deviation of the batch algorithm is smaller than that of the online algorithm, which is why we adopt the blockwise batch algorithm. We used block size T_b = 1.0 s in the experiments.

Table 3 Experimental conditions.

Common	Sampling rate = 8 kHz Window = hanning Reverberation time T_R=130 ms
ICA part	Frame length T_{ICA} = 1024 points (128 ms) Frame shift = 256 points (32 ms) g = 100.0 μ = optimized for block size T_b Number of iterations N_I = 100

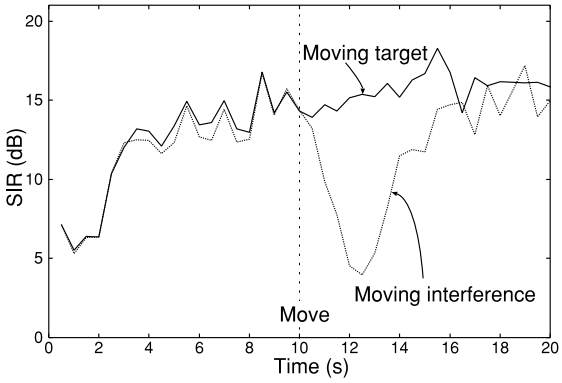

Fig. 20 SIR of blockwise batch algorithm without postprocessing. Target and interference signals moved at 10 s (T_b = 1.0 s).

We carried out experiments using speech signals recorded in a room. The reverberation time of the room was 130 ms. We used two omni-directional microphones with an inter-element spacing of 4 cm. The layout of the room is shown in Fig. 19. The target source signal was first located at A and then moved to B at a speed of 30 deg/s. The interference signal was located at C and moved to D at a speed of 40 deg/s.

The step size parameter μ in (12) affects the separation performance of BSS when the block size changes. We carried out preliminary experiments and chose μ to optimize the performance for each block size. The other conditions are summarized in Table 3. We measured SIRs with 30 combinations of source signals using three male and three female speakers, and averaged them.

We investigate the BSS performance for moving sources using the blockwise batch algorithm. Figure 20 shows the SIR for a moving target (solid line) and that for a moving interference (dotted line). We can see that the SIR is not degraded even when the target moves. By contrast, interference movement causes a decline in the SIR.

This can be explained by the directivity pattern of the separation system obtained by ICA. The solution of frequency-domain BSS works in the same way as an adap-

tive beamformer, which forms a spatial null toward an interference signal (Fig. 5). Because of this characteristic, BSS using ICA is robust with a moving target signal but fragile with a moving interference signal. Taking advantage of this nature, we can estimate residual crosstalk components even when the interference signal moves by employing postprocessing in the second stage [50].

11. Conclusion

This paper presented a comprehensive description of frequency-domain BSS as well as various techniques that enable frequency-domain BSS to be used for separating many speech signals mixed in a real-room environment. The permutation problem has been a major concern with the frequency domain approach. However, with the methods described in Sect. 7, this problem can be solved even in a practical situation. Moreover, the locations of sources can be estimated by the method described in Sect. 6. This ability is unique to the frequency domain approach, and cannot be seen in time-domain BSS. Our experimental results show that the separation performance was fairly good and the computational cost was feasible. These results demonstrate the effectiveness of frequency-domain BSS.

References

[1] S. Makino, "Blind source separation of convolutive mixtures of speech," in Adaptive Signal Processing: Applications to Real-World Problems, eds. J. Benesty and Y. Huang, Springer, 2003.
[2] S. Haykin, ed., Unsupervised Adaptive Filtering (Volume I: Blind Source Separation), John Wiley & Sons, 2000.
[3] A. Cichocki and S. Amari, Adaptive Blind Signal and Image Processing, John Wiley & Sons, 2002.
[4] A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
[5] T.W. Lee, Independent Component Analysis—Theory and Applications, Kluwer Academic Publishers, 1998.
[6] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications, pp.101–104, April 1997.
[7] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved non-stationary signals," Neurocomputing, vol.22, pp.157–171, 1998.
[8] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," Proc. ICA 2001, pp.722–727, Dec. 2001.
[9] S.C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," Speech Commun., vol.39, pp.65–78, 2003.
[10] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in Audio Signal Processing for Next-Generation Multimedia Communication Systems, eds. Y. Huang and J. Benesty, Kluwer Academic Publishers, pp.255–293, 2004,
[11] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis," IEICE Trans. Fundamentals, vol.E87-A, no.8, pp.2063–2072, Aug. 2004.
[12] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in Speech Enhancement, eds. J. Benesty, S. Makino, and J. Chen, Springer, 2005.
[13] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomputing, vol.22, pp.21–34, 1998.
[14] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech Audio Process., vol.8, no.3, pp.320–327, May 2000.
[15] L. Schobben and W. Sommen, "A frequency domain blind signal separation method based on decorrelation," IEEE Trans. Signal Process., vol.50, no.8, pp.1855–1865, Aug. 2002.
[16] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," Proc. ICA 2000, pp.215–220, June 2000.
[17] 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.
[18] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," IEEE Trans. Speech Audio Process., vol.11, no.3, pp.204–215, May 2003.
[19] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," Proc. ICASSP 2000, pp.3140–3143, June 2000.
[20] 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, no.11, pp.1135–1146, 2003.
[21] M.Z. Ikram and D.R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," Proc. ICASSP 2002, pp.881–884, May 2002.
[22] 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, March 2003.
[23] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. Speech Audio Process., vol.11, no.2, pp.109–116, March 2003.
[24] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," EURASIP Journal on Applied Signal Processing, vol.2003, no.11, pp.1157–1166, 2003.
[25] 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 Process., vol.12, pp.530–538, Sept. 2004.
[26] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation using small and large spacing sensor pairs," Proc. ISCAS 2004, vol.V, pp.1–4, May 2004.
[27] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation for many speech signals," Proc. ICA 2004 (LNCS 3195), pp.461–469, Sept. 2004.
[28] S. Winter, H. Sawada, and S. Makino, "Geometrical understanding of the PCA subspace method for overdetermined blind source separation," Proc. ICASSP 2003, pp.769–772, April 2003.
[29] H. Sawada, R. Mukai, S. de la Kethulle, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," Proc. IWAENC2003, pp.311–314, Sept. 2003.
[30] A. Hyvärinen, "Fast and robust fixed-point algorithm for independent component analysis," IEEE Trans. Neural Netw., vol.10, no.3, pp.626–634, 1999.
[31] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," Int. J. Neural Systems, vol.10, no.1, pp.1–8, Feb. 2000.
[32] M. Joho and P. Schniter, "Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient," Proc. ICA2003, pp.543–548, April 2003.

[33] A.D. Back and A.C. Tsoi, "Blind deconvolution of signals using a complex recurrent network," Proc. Neural Networks for Signal Process., pp.565–574, 1994.
[34] R.H. Lambert and A.J. Bell, "Blind separation of multiple speakers in a multipath environment," Proc. ICASSP'97, pp.423–426, April 1997.
[35] T.W. Lee, A.J. Bell, and R. Orglmeister, "Blind source separation of real world signals," Proc. ICNN, pp.2129–2135, June 1997.
[36] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," IEEE Signal Process. Mag., vol.9, no.1, pp.14–37, Jan. 1992.
[37] H. Sawada, S. Winter, R. Mukai, S. Araki, and S. Makino, "Estimating the number of sources for frequency-domain blind source separation," Proc. ICA 2004 (LNCS 3195), pp.610–617, Sept. 2004,
[38] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Comput., vol.7, no.6, pp.1129–1159, 1995.
[39] S. Amari, "Natural gradient works efficiently in learning," Neural Comput., vol.10, no.2, pp.251–276, 1998.
[40] J.F. Cardoso, "Blind beamforming for non-Gaussian signals," IEE Proceedings-F, pp.362–370, Dec. 1993.
[41] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," Neural Netw., vol.8, no.3, pp.411–419, 1995.
[42] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag., vol.34, pp.276–280, March 1986.
[43] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," Proc. International Symposium on Signal Processing and its Applications, pp.411–414, July 2003.
[44] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE ASSP Mag., vol.5, pp.4–24, April 1988.
[45] J.-F. Cardoso, "Source separation using higher order moments," Proc. ICASSP'89, vol.4, pp.2109–2112, May 1989.
[46] V.C. Soon, L. Tong, Y.F. Huang, and R. Liu, "A robust method for wideband signal separation," Proc. ISCAS'93, vol.1, pp.703–706, May 1993.
[47] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed., Wiley Interscience, 2000.
[48] J. Anemüller and T. Gramss, "On-line blind separation of moving sound sources," Proc. ICA'99, pp.331–334, 1999.
[49] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environment," Proc. ICASSP 2000, pp.1133–1136, 2000.
[50] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," IEICE Trans. Fundamentals, vol.E87-A, no.8, pp.1941–1948, Aug. 2004.

Shoji Makino received the B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now an Executive Manager at the NTT Communication Science Laboratories. He is also a Guest Professor at the Hokkaido University. His research interests include adaptive filtering technologies and realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech. He received the TELECOM System Technology Award of the TAF in 2004, the Best Paper Award of the IWAENC in 2003, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002, the Achievement Award of the IEICE in 1997, and the Outstanding Technological Development Award of the ASJ in 1995. He is the author or co-author of more than 200 articles in journals and conference proceedings and has been responsible for more than 150 patents. He is a member of the Conference Board of the IEEE SP Society and an Associate Editor of the IEEE Transactions on Speech and Audio Processing. He is also an Associate Editor of the EURASIP Journal on Applied Signal Processing. He is a member of the Technical Committee on Audio and Electroacoustics of the IEEE SP Society as well as the Technical Committee on Blind Signal Processing of the IEEE Circuits and Systems Society. He is also a member of the International ICA Steering Committee and the Organizing Chair of the ICA2003 in Nara. He is the General Chair of the IWAENC2003 in Kyoto. He was a Vice Chair of the Technical Committee on Engineering Acoustics of the IEICE and the ASJ. He is an IEEE Fellow, a council member of the ASJ, and a member of the EURASIP.

Hiroshi Sawada received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively. In 1993, he joined NTT Communication Science Laboratories. From 1993 to 2000, he was engaged in research on the computer aided design of digital systems, logic synthesis, and computer architecture. Since 2000, he has been engaged in research on array signal processing, blind source separation for convolutive mixtures, and speech enhancement. He received the best paper award of the IEEE Circuit and System Society in 2000. He is a senior member of the IEEE and a member of the ASJ.

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

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