Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 34970, Pages 1–17
DOI 10.1155/ASP/2006/34970

Blind Separation of Acoustic Signals Combining SIMO-Model-Based Independent Component Analysis and Binary Masking

Yoshimitsu Mori,[1] Hiroshi Saruwatari,[1] Tomoya Takatani,[1] Satoshi Ukai,[1] Kiyohiro Shikano,[1] Takashi Hiekata,[2] Youhei Ikeda,[2] Hiroshi Hashimoto,[2] and Takashi Morita[2]

[1] *Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma 630-0192, Japan*
[2] *Kobe Steel, Ltd., Kobe 651-2271, Japan*

Received 1 January 2006; Revised 22 June 2006; Accepted 22 June 2006

A new two-stage blind source separation (BSS) method for convolutive mixtures of speech is proposed, in which a single-input multiple-output (SIMO)-model-based independent component analysis (ICA) and a new SIMO-model-based binary masking are combined. SIMO-model-based ICA enables us to separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources in their original form at the microphones. Thus, the separated signals of SIMO-model-based ICA can maintain the spatial qualities of each sound source. Owing to this attractive property, our novel SIMO-model-based binary masking can be applied to efficiently remove the residual interference components after SIMO-model-based ICA. The experimental results reveal that the separation performance can be considerably improved by the proposed method compared with that achieved by conventional BSS methods. In addition, the real-time implementation of the proposed BSS is illustrated.

1. INTRODUCTION

Blind source separation (BSS) is the approach taken to estimate original source signals using only the information of the mixed signals observed in each input channel. Basically, BSS is classified as an *unsupervised* filtering technique [1] in that the source separation procedure requires no training sequences and no a priori information on the directions-of-arrival (DOAs) of the sound sources. Owing to the attractive features of BSS, much attention has been given to BSS in many fields of signal processing such as speech enhancement. This technique will provide an indispensable basis of realizing noise-robust speech recognition and high-quality hands-free telecommunication systems.

The early contributory studies of BSS are mainly based on the utilization of high-order statistics [2, 3] or independent component analysis (ICA) [4–6], where the independence among source signals is used for separation. In recent years, various methods have been presented for acoustic-sound separation [7–11] in which the sound mixing model is referred to as *convolutive mixtures*. In this paper, we also address the BSS problem under highly reverberant conditions, which often arise in many practical audio applications. The separation performance of conventional ICA is far from being sufficient in the reverberant case because excessively long separation filters are required but the unsupervised learning of the filters is difficult. Therefore, the development of high-accuracy BSS in a real-world application is a problem demanding prompt attention. One possible improvement is to partly combine ICA with another signal enhancement technique; however, in conventional ICA, each of the separated outputs is a *monaural* signal, which leads to the drawback that many types of superior *multichannel* techniques cannot be applied.

In order to attack this difficult problem, we propose a novel two-stage BSS algorithm that is applicable to an array of directional microphones. This approach resolves the BSS problem into two stages: (a) a single-input multiple-output (SIMO)-model-based ICA proposed by some of the authors [12] and (b) a new SIMO-model-based binary masking in the time-frequency domain for the SIMO signals obtained from the preceding SIMO-model-based ICA. Here, the term "SIMO" represents the specific transmission system in which the input is a single source signal and the outputs are its transmitted signals observed at multiple microphones. SIMO-model-based ICA enables us to separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as if these sources were at the microphones. Thus, the separated signals of SIMO-model-based ICA can maintain the rich spatial qualities of each sound source. After SIMO-model-based ICA, the residual components of interference, which often appear at the output of SIMO-model-based ICA as well as of the conventional ICA, can be efficiently removed by the following binary masking. The experimental results show the proposed method's efficacy under realistic reverberant conditions. The proposed method can achieve enhanced interference reduction while keeping the distortion low for the target signals, compared with many existing BSS methods.

In the similar context of a technique that combines ICA and binary masking, Kolossa and Orglmeister have proposed the method [13] in which conventional binary masking [14–16] is cascaded after conventional monaural-output ICA as a postprocessing for residual interference reduction. Indeed the method is slightly more effective in obtaining further separation performances than ICA, especially when the ICA part has an insufficient performance. However, unlike our proposed method, it will be revealed that the existing combination method produces very large sound distortions in the resultant signals, and thus yields a deterioration. This drawback is not acceptable in several acoustical sound applications, for example, speech recognition, because the recognition rate is affected by the separated sounds' distortions.

It should be emphasized that the proposed two-stage method has another important property, that is, applicability to *real-time* processing. In general, ICA-based BSS methods require enormous calculations, but binary masking needs very low computational complexities. Therefore, because of the introduction of binary masking into ICA, the proposed combination can function as a real-time system. In this paper, we also discuss the real-time implementation issue on the proposed BSS, and evaluate the "real-time" separation performance for speech mixtures under real reverberant conditions.

The rest of this paper is organized as follows. In Sections 2 and 3, the formulation for the general BSS problems and the principle of the proposed method are explained. In Sections 4-5, various signal separation experiments are described to assess the proposed method's superiority to conventional BSS methods. Following the discussion on the results of the experiments, we present our conclusions in Section 7.

2. MIXING PROCESS AND CONVENTIONAL BSS

2.1. Mixing process

In this study, the number of microphones is K and the number of multiple sound sources is L, where we deal with the case of $K = L$.

Multiple mixed signals are observed at the microphone array, and these signals are converted into discrete-time series via an A/D converter. By applying the discrete-time Fourier transform, we can express the observed signals, in which multiple source signals are linearly mixed with additive noise, as follows in the frequency domain:

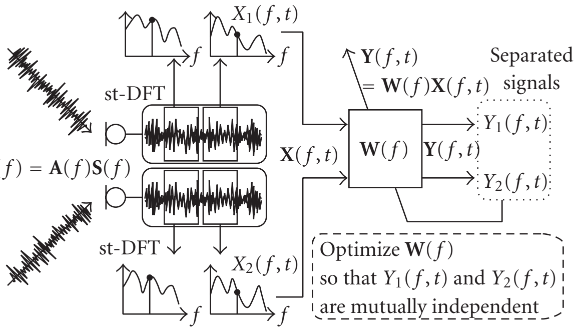


Figure 1: Blind source separation procedure performed in frequency-domain ICA.

$$\mathbf{X}(f) = \mathbf{A}(f)\mathbf{S}(f) + \mathbf{N}(f), \tag{1}$$

where $\mathbf{X}(f) = [X_1(f), \ldots, X_K(f)]^{\mathrm{T}}$ is the observed signal vector, and $\mathbf{S}(f) = [S_1(f), \ldots, S_L(f)]^{\mathrm{T}}$ is the source signal vector. Also, $\mathbf{A}(f) = [A_{kl}(f)]_{kl}$ is the mixing matrix, where $[X]_{ij}$ denotes the matrix which includes the element X in the ith row and the jth column. Here, $\mathbf{N}(f)$ is the additive noise term which generally represents, for example, a background noise and/or a sensor noise. The mixing matrix $\mathbf{A}(f)$ is complex-valued because we introduce a model to deal with the relative time delays among the microphones and room reverberations.

2.2. Conventional ICA-based BSS

In frequency-domain ICA (FDICA) [7–10], first, the short-time analysis of observed signals is conducted by a frame-by-frame discrete Fourier transform (DFT) (see Figure 1). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider these values as a time series. Hereafter, we designate the time series as $\mathbf{X}(f,t) = [X_1(f,t), \ldots, X_K(f,t)]^{\mathrm{T}}$.

Next, we perform signal separation using the complex-valued unmixing matrix $\mathbf{W}(f) = [W_{lk}(f)]_{lk}$, so that the L time-series output $\mathbf{Y}(f,t) = [Y_1(f,t), \ldots, Y_L(f,t)]^{\mathrm{T}}$ becomes mutually independent; this procedure can be given as

$$\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t). \tag{2}$$

We perform this procedure with respect to all frequency bins.

The optimal $\mathbf{W}(f)$ is obtained by many types of ICA. For example, second-order ICA has the following iterative updating equation [9]:

$$\begin{aligned}\mathbf{W}^{[i+1]}(f) = -\eta \sum_{\tau} \alpha(f)\, \text{off-diag}\left(\mathbf{R}_{yy}(f,\tau)\right) \\ \cdots \mathbf{W}^{[i]}(f)\mathbf{R}_{xx}(f,\tau) + \mathbf{W}^{[i]}(f),\end{aligned} \tag{3}$$

where η is the step-size parameter, off-diag$[\mathbf{X}]$ is the operation for setting every diagonal element of the matrix $\mathbf{X}$ to zero, $[i]$ is used to express the value of the ith step in the iterations, and $\alpha(f) = (\sum_\tau \|\mathbf{R}_{xx}(f,\tau)\|^2)^{-1}$ is a normalization factor ($\|\cdot\|$ represents the Frobenius norm). $\mathbf{R}_{xx}(f,\tau)$ and $\mathbf{R}_{yy}(f,\tau)$ are the cross-power spectra of the input $\mathbf{x}(f,t)$ and the output $\mathbf{y}(f,t)$, respectively, which are calculated around the multiple time indices τ.

On the other hand, higher-order ICA typically involves the following updating [7]:

$$\mathbf{W}^{[i+1]}(f) = \eta\left[\mathbf{I} - \left\langle \mathbf{\Phi}(\mathbf{Y}(f,t))\mathbf{Y}^{\mathrm{H}}(f,t)\right\rangle_t\right]\mathbf{W}^{[i]}(f) + \mathbf{W}^{[i]}(f), \tag{4}$$

where $\mathbf{I}$ is the identity matrix, $\langle\cdot\rangle_t$ denotes the time-averaging operator, and $\mathbf{\Phi}(\cdot)$ is the appropriate nonlinear vector function [17]. After the iterations, the source permutation and the scaling indeterminacy problem can be solved, for example, by the methods outlined in [8, 10].

The ICA-based BSS approach seems to be a very flexible and effective technique for the source separation because it does not need a priori information except for the assumption of sources' independence. However, it has an inherent disadvantage in that there is difficulty with the poor and slow convergence of nonlinear optimization [18, 19], particularly when we are confronted with very complex convolutive mixtures as in the case of reverberant acoustic conditions. Furthermore, ordinary ICA-based BSS algorithms require huge computational complexities. The disadvantages reduce the applicability of the approach to the general audio applications which often need real-time processing.

2.3. Conventional binary-mask-based BSS

Binary masking [14–16] is one of the alternative approaches aimed at solving the BSS problem, but is not based on ICA. We estimate a binary mask by comparing the amplitudes of the observed signals, and pick up the target sound component which arrives at the *better microphone* closer to the target sound (this is easier even for the far-field sources when we use directional microphones whose directivities are steered distinctly from each other). This procedure is performed in time-frequency regions; it allows the specific regions where the target sound is dominant to pass and mask the other regions. Under the assumption that the lth sound source is close to the lth microphone and $K = L = 2$, the lth separated signal is given by

$$\hat{Y}_l(f,t) = m_l(f,t)X_l(f,t), \tag{5}$$

where $m_l(f,t)$ is the binary mask operation which is defined as $m_l(f,t) = 1$ if $|X_l(f,t)| > |X_k(f,t)|$ $(k \neq l)$; otherwise $m_l(f,t) = 0$.

This method requires very low computational complexities, thereby making it well applicable to real-time processing. The method, however, needs an assumption of sparseness in the sources' spectral components; that is, there should be no overlaps in the time-frequency components of the sources. However, strictly speaking, the assumption does not hold in a usual audio application, and in that case the method often produces very harmful noise, so-called *musical noise*. In particular, for the speech-speech mixing, the breach of the sparseness assumption can be partly mitigated [20], but it still retains the overlapped spectral components greater than several dozens of percent. This yields a considerable signal distortion, which will be experimentally shown in Section 4.

3. PROPOSED TWO-STAGE BSS ALGORITHM

3.1. What is SIMO-model-based ICA?

In a previous study, SIMO-model-based ICA (SIMO-ICA) was proposed by some of the authors [12], who showed that SIMO-ICA enables the separation of mixed signals into SIMO-model-based signals at microphone points.

In general, the observed signals at the multiple microphones can be represented as a superposition of the SIMO-model-based signals as follows:

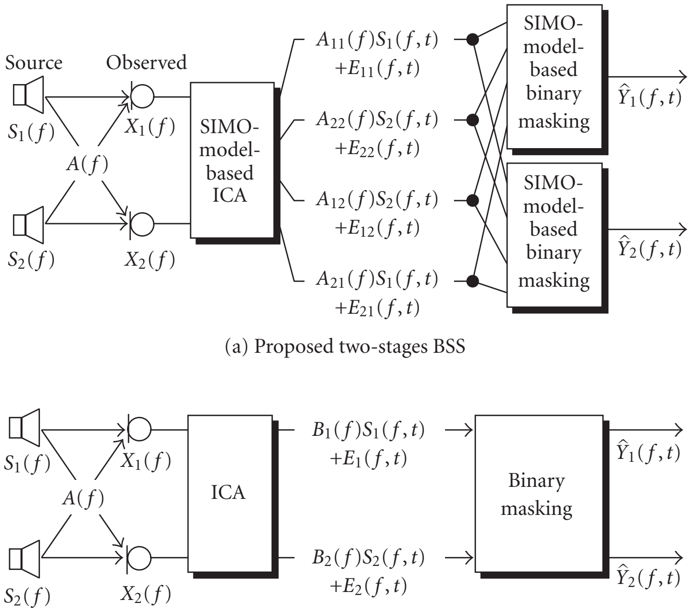
$$\begin{aligned}\mathbf{X}(f) = {} & \left[A_{11}(f)S_1(f), \dots, A_{K1}(f)S_1(f)\right]^{\mathrm{T}} \\ & + \left[A_{12}(f)S_2(f), \dots, A_{K2}(f)S_2(f)\right]^{\mathrm{T}} \\ & \vdots \\ & + \left[A_{1L}(f)S_L(f), \dots, A_{KL}(f)S_L(f)\right]^{\mathrm{T}},\end{aligned} \tag{6}$$

where $[A_{1l}(f)S_l(f), \dots, A_{Kl}(f)S_l(f)]^{\mathrm{T}}$ is a vector which corresponds to the SIMO-model-based signals with respect to the lth sound source; the kth element corresponds to the kth microphone's signal.

The aim of SIMO-ICA is to decompose the mixed observations $\mathbf{X}(f)$ into the SIMO components of each independent sound source; that is, we estimate $A_{kl}(f)S_l(f)$ for all k and l values (up to the permissible time delay in separation filtering). SIMO-ICA has the advantage that the separated signals still maintain the spatial qualities of each sound source, in comparison with conventional ICA-based BSS methods. Clearly, this attractive feature makes SIMO-ICA highly applicable to high-fidelity acoustic signal processing, for example, binaural sound separation [21].

3.2. Motivation and strategy

Owing to the fact that SIMO-model-based separated signals are still *one set of array signals*, there exist new applications in which SIMO-model-based separation is combined with other types of multichannel signal processing. In this paper, hereinafter we address a specific BSS consisting of directional microphones in which each microphone's directivity is steered to a distinct sound source, that is, the lth microphone steers to the lth sound source. Thus, the outputs of SIMO-ICA are the estimated (separated) SIMO-model-based signals, and they keep the relation that the lth source component is the most dominant in the lth microphone. This finding has motivated us to combine SIMO-ICA and binary masking. Moreover, we propose to extend the simple binary masking to a new binary masking strategy, so-called *SIMO-model-based binary masking* (SIMO-BM). That is, the


Source
Observed
$S_1(f)$
$X_1(f)$
$A(f)$
$S_2(f)$
$X_2(f)$
SIMO-model-based ICA
$A_{11}(f)S_1(f,t)$ $+E_{11}(f,t)$
$A_{22}(f)S_2(f,t)$ $+E_{22}(f,t)$
$A_{12}(f)S_2(f,t)$ $+E_{12}(f,t)$
$A_{21}(f)S_1(f,t)$ $+E_{21}(f,t)$
SIMO-model-based binary masking
SIMO-model-based binary masking
$\hat{Y}_1(f,t)$
$\hat{Y}_2(f,t)$
ICA
$B_1(f)S_1(f,t)$ $+E_1(f,t)$
$B_2(f)S_2(f,t)$ $+E_2(f,t)$
Binary masking
$\hat{Y}_1(f,t)$
$\hat{Y}_2(f,t)$

(a) Proposed two-stages BSS

(b) Simple combination of conventional ICA and binary mask

Figure 2: Input and output relations in (a) proposed two-stage BSS and (b) simple combination of conventional ICA and binary masking. This corresponds to the case of $K = L = 2$.

masking function is determined by all the information regarding the SIMO components of all sources obtained from SIMO-ICA. The configuration of the proposed method is shown in Figure 2(a). SIMO-BM, which subsequently follows SIMO-ICA, enables us to remove the residual component of the interference effectively without adding enormous computational complexities. This combination idea is also applicable to the realization of the proposed method's real-time implementation.

It is worth mentioning that the novelty of this strategy mainly lies in the two-stage idea of the unique combination of SIMO-ICA and SIMO-model-based binary masking. To illustrate the novelty of the proposed method, we hereinafter compare the proposed combination with a simple two-stage combination of conventional monaural-output ICA and conventional binary masking (see Figure 2(b)) [13].

In general, conventional ICAs can only supply the source signals $Y_l(f,t) = B_l(f)S_l(f,t) + E_l(f,t)$ $(l = 1,\ldots,L)$, where $B_l(f)$ is an unknown arbitrary filter and $E_l(f,t)$ is a residual separation error which is mainly caused by an insufficient convergence in ICA. The residual error $E_l(f,t)$ should be removed by binary masking in the subsequent postprocessing stage. However, the combination is very problematic and cannot function well because of the existence of spectral overlaps in the time-frequency domain. For instance, if all sources have nonzero spectral components (i.e., when the sparseness assumption does not hold) in the specific frequency subband and are comparable (see Figures 3(a) and 3(b)), that is,

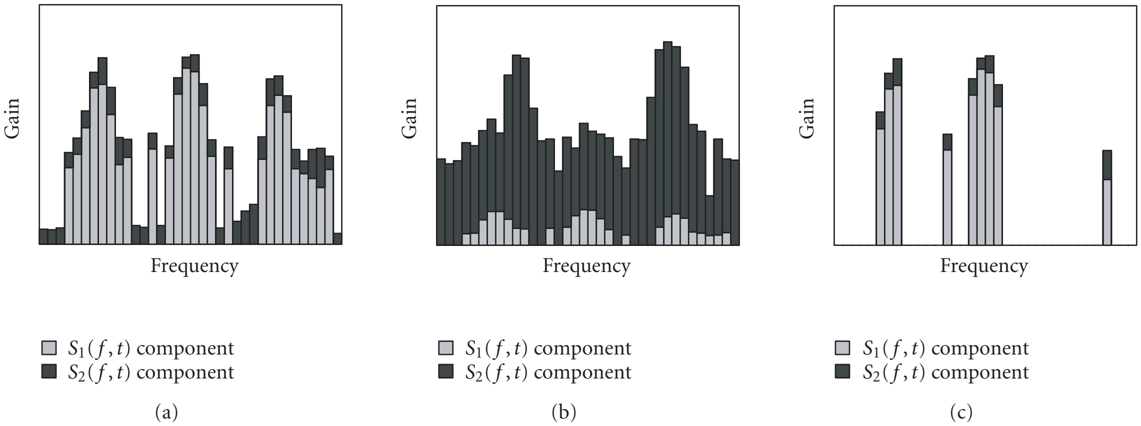
$$\left|B_1(f)S_1(f,t) + E_1(f,t)\right| \simeq \left|B_2(f)S_2(f,t) + E_2(f,t)\right|, \tag{7}$$

the decision in binary masking for $Y_1(f,t)$ and $Y_2(f,t)$ is vague and the output results in a ravaged (highly distorted) signal (see Figure 3(c)). Thus, the simple combination of conventional ICA and binary masking is not suited for achieving BSS with high accuracy.

On the other hand, our proposed combination contains the special SIMO-ICA in the first stage, where the SIMO-ICA can supply the specific SIMO signals corresponding to each of the sources, $A_{kl}(f)S_l(f,t)$, up to the possible residual error $E_{kl}(f,t)$ (see Figure 4). Needless to say that the obtained SIMO components are very beneficial to the decision-making process of the masking function. For example, if the residual error $E_{kl}(f,t)$ is smaller than the main SIMO component $A_{kl}(f)S_l(f,t)$, the binary masking between $A_{11}(f)S_1(f,t)+E_{11}(f,t)$ (Figure 4(a)) and $A_{21}(f)S_1(f,t) + E_{21}(f,t)$ (Figure 4(b)) is more acoustically reasonable than the conventional combination because the spatial properties, in which the separated SIMO component at the specific microphone close to the target sound still maintains a large gain, are kept; that is,

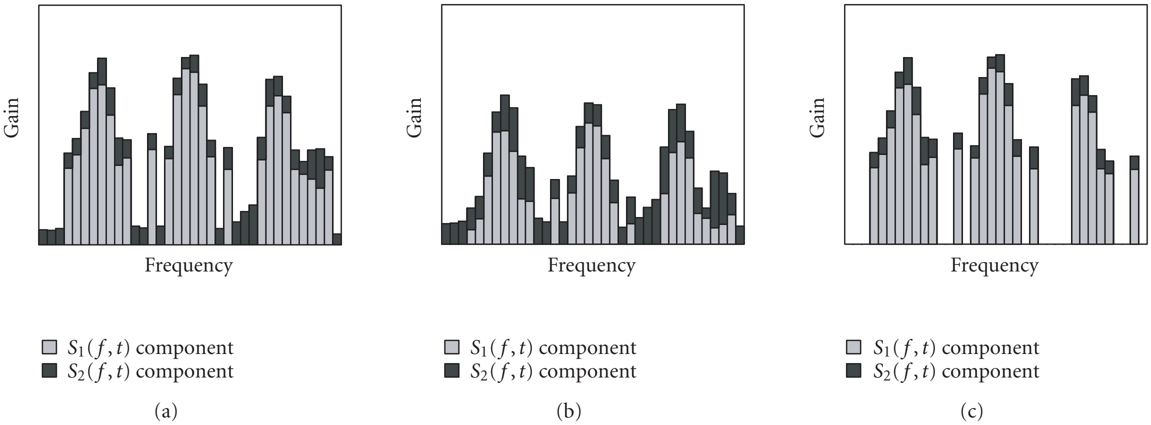
$$\left|A_{11}(f)S_1(f,t) + E_{11}(f,t)\right| > \left|A_{21}(f)S_1(f,t) + E_{21}(f,t)\right|. \tag{8}$$

In this case, we can correctly pick up the target signal candidate $A_{11}(f)S_1(f,t) + E_{11}(f,t)$ (see Figure 4(c)). When the target components $A_{k1}(f)S_1(f,t)$ are absent in the target-speech silent duration, if the errors have a possible amplitude relation of $|E_{11}(f,t)| < |E_{21}(f,t)|$, then our binary masking forces the period to be zero and can remove the residual errors. Note that unlike the simple combination method [13] our proposed binary masking is not affected by the



(a) (b) (c)

FIGURE 3: Examples of spectra in simple combination of ICA and binary masking. (a) ICA's output 1; $B_1(f)S_1(f,t) + E_1(f,t)$, (b) ICA's output 2; $B_2(f)S_2(f,t) + E_2(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.



(a) (b) (c)

FIGURE 4: Examples of spectra in proposed two-stage method. (a) SIMO-ICA's output 1; $A_{11}(f)S_1(f,t) + E_{11}(f,t)$, (b) SIMO-ICA's output 2; $A_{21}(f)S_1(f,t) + E_{21}(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.

amplitude balance among sources. Overall, after obtaining the SIMO components, we can introduce SIMO-BM for the efficient reduction of the remaining error in ICA, even when the complete sparseness assumption does not hold.

3.3. Illustrative example

To illustrate the proposed theory with examples, we performed a preliminary experiment in which the binary mask is applied to the ideal solutions of the two types of ICAs (SIMO-ICA and the simple conventional ICA) under a real acoustic condition which will be described in Section 4. First we consider the case in which binary masking is directly applied to straight-pass components of each source ($A_{11}(f)S_1(f,t)$ and $A_{22}(f)S_2(f,t)$). The following resultant outputs are calculated:

$$\hat{Y}_1(f,t) = m_1(f,t)A_{11}(f)S_1(f,t), \tag{9}$$

where $m_1(f,t) = 1$ if $|A_{11}(f)S_1(f,t)| > |A_{22}(f)S_2(f,t)|$; otherwise $m_1(f,t) = 0$, and

$$\hat{Y}_2(f,t) = m_2(f,t)A_{22}(f)S_2(f,t), \tag{10}$$

where $m_2(f,t) = 1$ if

$$|A_{22}(f)S_2(f,t)| > |A_{11}(f)S_1(f,t)|; \tag{11}$$

otherwise $m_2(f,t) = 0$. As a result, a large distortion of about 5 dB was observed, which means that the simple combination of ICA and binary masking is likely to involve sound distortion. On the other hand, when binary masking is applied to the SIMO components of $S_1(f,t)(A_{11}(f)S_1(f,t)$ and $A_{21}(f)S_1(f,t))$ for picking up source 1, we obtain

$$\hat{Y}_1(f,t) = m_1(f,t)A_{11}(f)S_1(f,t), \tag{12}$$

where $m_1(f,t) = 1$ if $|A_{11}(f)S_1(f,t)| > |A_{21}(f)S_1(f,t)|$; otherwise $m_1(f,t) = 0$. Also, for picking up source 2, we obtain

$$\hat{Y}_2(f,t) = m_2(f,t)A_{22}(f)S_2(f,t), \tag{13}$$

where $m_2(f,t) = 1$ if $|A_{22}(f)S_2(f,t)| > |A_{12}(f)S_2(f,t)|$; otherwise $m_2(f,t) = 0$. This processing yields a small distortion of less than 1 dB. Thus, the proposed idea, the use of binary masking after obtaining SIMO components of each source, is well suited to the realization of low-distortion BSS.

In summary, the novelty of the proposed two-stage idea is attributed to the introduction of the SIMO-model-based framework into both separation and postprocessing, and this offers the realization of a robust BSS. The detailed algorithm is described in the next subsection.

3.4. *Algorithm: SIMO-ICA in 1st stage*

Time-domain SIMO-ICA [12] has recently been proposed by some of the authors as a means of obtaining SIMO-model-based signals directly in ICA updating. In this study, we extend time-domain SIMO-ICA to frequency-domain SIMO-ICA (FD-SIMO-ICA). FD-SIMO-ICA is conducted for extracting the SIMO-model-based signals corresponding to each of the sources. FD-SIMO-ICA consists of $(L-1)$ FDICA parts and a *fidelity controller*, and each ICA runs in parallel under the fidelity control of the entire separation system (see Figure 5). The separated signals of the lth ICA $(l = 1,\ldots,L-1)$ in FD-SIMO-ICA are defined by

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \left[Y_k^{(\mathrm{ICA}l)}(f,t)\right]_{k1} = \mathbf{W}_{(\mathrm{ICA}l)}(f)\mathbf{X}(f,t), \tag{14}$$

where $\mathbf{W}_{(\mathrm{ICA}l)}(f) = [W_{ij}^{(\mathrm{ICA}l)}(f)]_{ij}$ is the separation filter matrix in the lth ICA.

Regarding the fidelity controller, we calculate the following signal vector $\mathbf{Y}_{(\mathrm{ICA}L)}(f,t)$, in which all the elements are to be mutually independent:

$$\begin{aligned}\mathbf{Y}_{(\mathrm{ICA}L)}(f,t) &= \left(\mathbf{I} - \sum_{l=1}^{L-1}\mathbf{W}_{(\mathrm{ICA}l)}(f)\right)\mathbf{X}(f,t)\\ &= \mathbf{X}(f,t) - \sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t).\end{aligned} \tag{15}$$

Hereafter, we regard $\mathbf{Y}_{(\mathrm{ICA}L)}(f,t)$ as an output of a *virtual* "Lth" ICA. The word " *virtual*" is used here because the Lth ICA does not have its own separation filters unlike the other ICAs, and $\mathbf{Y}_{(\mathrm{ICA}L)}(f,t)$ is subject to $\mathbf{W}_{(\mathrm{ICA}l)}(f)$ $(l=1,\ldots,L-1)$. By transposing the second term $(-\sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t))$ on the right-hand side to the left-hand side, we can show that (15) suggests a constraint that forces the sum of all ICAs' output vectors $\sum_{l=1}^{L}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t)$ to be the sum of all SIMO components $\left[\sum_{l=1}^{L} A_{kl}(f)\ S_l(f,t)\right]_{k1}(=\mathbf{X}(f,t))$.

If the independent sound sources are separated by (14), and simultaneously the signals obtained by (15) are also mutually independent, then the output signals converge towards unique solutions, up to the permutation and the residual error, as

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\left[\mathbf{A}(f)\ \ \mathbf{P}_l^{\mathrm{T}}\right]\mathbf{P}_l\mathbf{S}(f,t) + \mathbf{E}_l(f,t), \tag{16}$$

where diag[$\mathbf{X}$] is the operation for setting every off-diagonal element of the matrix $\mathbf{X}$ to zero, $\mathbf{E}_l(f,t)$ represents the residual error vector, and $\mathbf{P}_l$ $(l = 1,\ldots,L)$ are exclusively-selected permutation matrices [22] which satisfy

$$\sum_{l=1}^{L}\mathbf{P}_l = [1]_{ij}. \tag{17}$$

For a proof of this, see Appendix A. Obviously, the solutions provide necessary and sufficient SIMO components, $A_{kl}(f)S_l(f,t)$, for each lth source. Thus, the separated signals of SIMO-ICA can maintain the spatial qualities of each sound source. For example, in the case of $L = K = 2$, one possibility is given by

$$\begin{aligned}&\left[Y_1^{(\mathrm{ICA1})}(f,t), Y_2^{(\mathrm{ICA1})}(f,t)\right]^{\mathrm{T}}\\ &\quad= \left[A_{11}(f)S_1(f,t) + E_{11}(f,t), A_{22}(f)S_2(f,t)\right.\\ &\qquad\left. + E_{22}(f,t)\right]^{\mathrm{T}},\end{aligned} \tag{18}$$

$$\begin{aligned}&\left[Y_1^{(\mathrm{ICA2})}(f,t), Y_2^{(\mathrm{ICA2})}(f,t)\right]^{\mathrm{T}}\\ &\quad= \left[A_{12}(f)S_2(f,t) + E_{12}(f,t), A_{21}(f)S_1(f,t)\right.\\ &\qquad\left. + E_{21}(f,t)\right]^{\mathrm{T}},\end{aligned} \tag{19}$$

where $\mathbf{P}_1 = \mathbf{I}$ and $\mathbf{P}_2 = [1]_{ij} - \mathbf{I}$.

In order to obtain (18), the natural gradient of Kullback-Leibler divergence on probability density functions of (15) with respect to $\mathbf{W}_{(\mathrm{ICA}l)}(f)$ should be added to the existing nonholonomic iterative learning rule [8] of the separation filter in the lth ICA$(l = 1,\ldots,L-1)$. The new iterative algorithm of the lth ICA part $(l = 1,\ldots,L-1)$ in FD-SIMO-ICA is given as (see Appendix B)

$$\begin{aligned}&\mathbf{W}_{(\mathrm{ICA}l)}^{[j+1]}(f)\\ &\quad= \mathbf{W}_{(\mathrm{ICA}l)}^{[j]}(f) - \alpha\\ &\qquad\times\Bigg[\left\{\text{off-diag}\left\langle\mathbf{\Phi}\left(\mathbf{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)\right)\mathbf{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)^{\mathrm{H}}\right\rangle_t\right\}\\ &\qquad\quad\cdot\mathbf{W}_{(\mathrm{ICA}l)}^{[j]}(f)\\ &\qquad\quad-\left\{\text{off-diag}\left\langle\mathbf{\Phi}\left(\mathbf{X}(f,t) - \sum_{l'=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l')}^{[j]}(f,t)\right)\right.\right.\\ &\qquad\qquad\left.\left.\cdot\left(\mathbf{X}(f,t) - \sum_{l'=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l')}^{[j]}(f,t)\right)^{\mathrm{H}}\right\rangle_t\right\}\\ &\qquad\quad\cdot\left(\mathbf{I} - \sum_{l'=1}^{L-1}\mathbf{W}_{(\mathrm{ICA}l')}^{[j]}(f)\right)\Bigg],\end{aligned} \tag{20}$$

where α is the step-size parameter, and we define the nonlinear vector function $\mathbf{\Phi}(\cdot)$ as $[\tanh(|Y_l(f,t)|)e^{J\cdot\arg(Y_l(f,t))}]_{l1}$ [17]. Also, the initial values of $\mathbf{W}_{(\mathrm{ICA}l)}(f)$ for all l values should be different.

After the iterations, we should solve two types of permutation problems, namely, (1) frequency-inside permutation specific to SIMO-ICA, and (2) inter-frequency permutation which commonly arises in FDICA. As for the frequency-inside permutation, the separated signals should be classified into the SIMO components of each source because the permutation corresponding to $\mathbf{P}_l$ possibly arises, even within

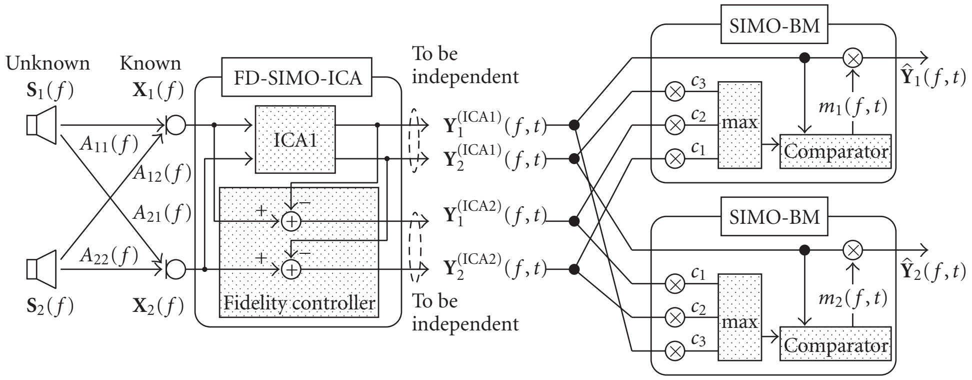


Figure 5: Input and output relations in proposed two-stage BSS which consists of FD-SIMO-ICA and SIMO-BM, where $K = L = 2$ and exclusively selected permutation matrices are given by $\mathbf{P}_1 = \mathbf{I}$ and $\mathbf{P}_2 = [1]_{ij} - \mathbf{I}$ in (16).

each frequency bin f. This can be easily achieved using a cross-correlation between time-shifted separated signals,

$$C(l,l',k,k') = \max_n \left\langle Y_k^{(\mathrm{ICA}l)}(f,t) Y_{k'}^{(\mathrm{ICA}l')}(f,t-n) \right\rangle_t, \tag{21}$$

where $l \neq l'$ and $k \neq k'$. The large value of $C(l,l',k,k')$ indicates that $Y_k^{(\mathrm{ICA}l)}(f,t)$ and $Y_{k'}^{(\mathrm{ICA}l')}(f,t)$ are SIMO components from the same source. As for the inter-frequency permutation, we can solve this problem between different f's by comparing the amplitude differences of the SIMO components in our scenario with directional microphones.

Note that there exists an alternative method [8] of obtaining the SIMO components in which the separated signals are projected back onto the microphones by using the inverse of $\mathbf{W}(f)$ after conventional ICA. The difference and advantage of SIMO-ICA relative to the projection-back method are described in Appendix C.

3.5. Algorithm: SIMO-BM in 2nd stage

After FD-SIMO-ICA, SIMO-model-based binary masking is applied (see Figure 5). Here, we consider the case of (18). The resultant output signal corresponding to source 1 is determined in the proposed SIMO-BM as follows:

$$\hat{Y}_1(f,t) = m_1(f,t) Y_1^{(\mathrm{ICA1})}(f,t), \tag{22}$$

where $m_1(f,t)$ is the *SIMO-model-based* binary mask operation which is defined as $m_1(f,t) = 1$ if

$$\begin{aligned} &\left| Y_1^{(\mathrm{ICA1})}(f,t) \right| \\ &\quad > \max\Big[c_1 \left| Y_2^{(\mathrm{ICA2})}(f,t) \right|, c_2 \left| Y_1^{(\mathrm{ICA2})}(f,t) \right|, \\ &\qquad c_3 \left| Y_2^{(\mathrm{ICA1})}(f,t) \right| \Big]; \end{aligned} \tag{23}$$

otherwise $m_1(f,t) = 0$. Here, $\max[\cdot]$ represents the function of picking up the maximum value among the arguments, and $c_1, \ldots, c_3$ are the weights for enhancing the contribution of each SIMO component to the masking decision process. For example, in the case of $[c_1, c_2, c_3] = [0,0,1]$, (23) becomes $|Y_1^{(\mathrm{ICA1})}(f,t)| > |Y_2^{(\mathrm{ICA1})}(f,t)|$, that is,

$$\left| A_{11}(f) S_1(f,t) + E_{11}(f,t) \right| > \left| A_{22}(f) S_2(f,t) + E_{22}(f,t) \right|. \tag{24}$$

This yields the simple combination of conventional ICA and conventional binary masking as described in Section 3.2. Otherwise, if we set $[c_1, c_2, c_3] = [1,0,0]$, (23) is turned to $|Y_1^{(\mathrm{ICA1})}(f,t)| > |Y_2^{(\mathrm{ICA2})}(f,t)|$, that is,

$$\left| A_{11}(f) S_1(f,t) + E_{11}(f,t) \right| > \left| A_{21}(f) S_1(f,t) + E_{21}(f,t) \right|. \tag{25}$$

This equation is identical to (8), where we can utilize better (acoustically reasonable) SIMO information regarding each source as described in Sections 3.2 and 3.3. If we change another pattern of c_i, we can generate various SIMO-model-based maskings with different separation and distortion properties.

The resultant output corresponding to source 2 is given by

$$\hat{Y}_2(f,t) = m_2(f,t) Y_2^{(\mathrm{ICA1})}(f,t), \tag{26}$$

where $m_2(f,t)$ is defined as $m_2(f,t) = 1$ if

$$\begin{aligned} &\left| Y_2^{(\mathrm{ICA1})}(f,t) \right| \\ &\quad > \max\Big[c_1 \left| Y_1^{(\mathrm{ICA2})}(f,t) \right|, c_2 \left| Y_2^{(\mathrm{ICA2})}(f,t) \right|, \\ &\qquad c_3 \left| Y_1^{(\mathrm{ICA1})}(f,t) \right| \Big]; \end{aligned} \tag{27}$$

otherwise $m_2(f,t) = 0$.

The extension to the general case of $L = K > 2$ can be easily implemented. Hereafter we consider one example in which the permutation matrices are given as

$$\mathbf{P}_l = \left[\delta_{in(k,l)} \right]_{ki}, \tag{28}$$

where δ_{ij} is the Kronecker's delta function, and

$$n(k,l) = \begin{cases} k + l - 1 & (k + l - 1 \leq L), \\ k + l - 1 - L & (k + l - 1 > L). \end{cases} \tag{29}$$

In this case, (16) yields

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \left[A_{kn(k,l)}(f)S_{n(k,l)}(f,t) + E_{kn(k,l)}(f,t)\right]_{k1}. \tag{30}$$

Thus, the resultant output for source 1 in SIMO-BM is given by

$$\hat{Y}_1(f,t) = m_1(f,t)Y_1^{(\mathrm{ICA1})}(f,t), \tag{31}$$

where $m_1(f,t)$ is defined as $m_1(f,t) = 1$ if

$$\begin{aligned} &\left|Y_1^{(\mathrm{ICA1})}(f,t)\right| \\ &\quad > \max\Big[c_1\left|Y_2^{(\mathrm{ICA}L)}(f,t)\right|, c_2\left|Y_3^{(\mathrm{ICA}L-1)}(f,t)\right|, \\ &\qquad c_3\left|Y_4^{(\mathrm{ICA}L-2)}(f,t)\right|, \dots, c_{L-1}\left|Y_L^{(\mathrm{ICA2})}(f,t)\right|, \\ &\qquad \dots, c_{LL-1}\left|Y_L^{(\mathrm{ICA1})}(f,t)\right|\Big]; \end{aligned} \tag{32}$$

otherwise $m_1(f,t) = 0$. The other sources can be obtained in the same manner.

3.6. Real-time implementation

Several recent research studies [23, 24] have dwelt on the issue of real-time implementation of ICA. The methods used, however, require high-speed personal computers, and a BSS implementation on a small-size LSI still receives much attention in industrial applications.

We have already built a pocket-size real-time BSS module, where the proposed two-stage BSS algorithm can work on a general-purpose DSP (TEXAS INSTRUMENTS TMS320C6713; 200 MHz clock, 100 kB program size, 1 MB working memory) as shown in Figure 6. Figure 7 shows a configuration of a real-time implementation for the proposed two-stage BSS. Signal processing in this implementation is performed in the following manner.

(1) Inputted signals are converted to time-frequency series by using a frame-by-frame fast Fourier transform (FFT).

(2) SIMO-ICA is conducted using current 3-seconds-duration data for estimating the separation matrix, which is applied to the next (*not current*) 3-seconds-samples. This staggered relation is due to the fact that the filter update in SIMO-ICA requires substantial computational complexities (the DSP performs at most 100 iterations) and cannot provide the optimal separation filter for the current 3-seconds-data.

(3) SIMO-BM is applied to the separated signals obtained by the previous SIMO-ICA. Unlike SIMO-ICA, binary masking can be conducted just in the current segment.

(4) The output signals from SIMO-BM are converted to the resultant time-domain waveforms by using an inverse FFT.

Although the separation filter update in the SIMO-ICA part is not real-time processing but includes a latency of 3 seconds, the entire two-stage system still seems to run in real-time because SIMO-BM can work in the current segment with no delay. Generally, the latency in conventional ICAs is problematic and reduces the applicability of such methods to real-time systems. In the proposed method, however, the performance deterioration due to the latency problem in SIMO-ICA can be mitigated by introducing real-time binary masking.

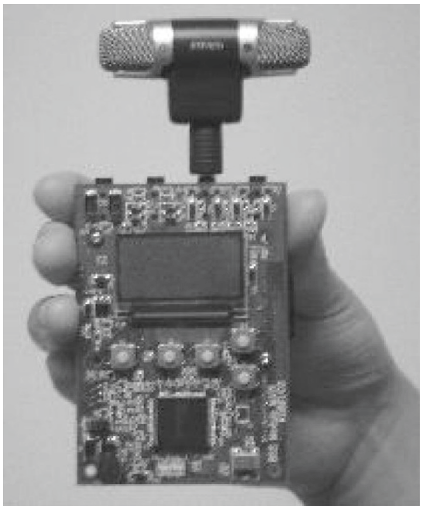


Figure 6: Overview of pocket-size real-time BSS module, where proposed two-stage BSS algorithm works on TEXAS INSTRUMENTS TMS320C6713 DSP.

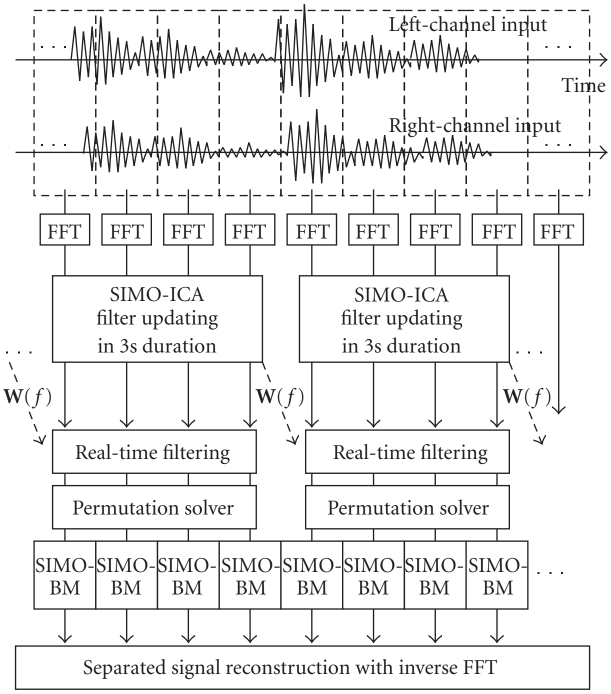


Figure 7: Signal flow in real-time implementation of proposed method.

4. SOUND SEPARATION EXPERIMENT

4.1. Experimental conditions

In this section, computer-simulation-based BSS experiments are discussed to investigate the basic properties of the proposed method. We use realistic (measured) room impulse responses recorded in a reverberant room (Figure 8) for the generation of convolutive mixtures. The reverberation time in this room is 200 milliseconds. We neglect the additive noise term $\mathbf{N}(f)$ in (1).

First, to evaluate the feasibility for general hands-free applications, we carried out sound-separation experiments with two sources and two directional microphones (Sony stereo microphone ECM-DS70P). Two speech signals are assumed to arrive from different directions, θ_1 and θ_2, where we prepare three kinds of source direction patterns as follows: $(\theta_1, \theta_2) = (-40^\circ, 50^\circ)$, $(-40^\circ, 30^\circ)$, or $(-40^\circ, 10^\circ)$. Two kinds of sentences, spoken by two male and two female speakers selected from the ASJ continuous speech corpus for research [25], are used as the original speech samples. Using these sentences, we obtain 12 combinations with respect to speakers and source directions, where the power ratio between every pair of the sound sources is set to 0 dB. The sampling frequency is 8 kHz and the length of each sound sample is limited to 3 seconds. The DFT size of $\mathbf{W}(f)$ is 1024. We used a null-beamformer-based initial value [10] which is steered to $(-60^\circ, 60^\circ)$. This experiment corresponds to the *offline* test, and the number of iterations in the ICA part is 500. The step-size parameter was optimized for each method to obtain the best separation performance.

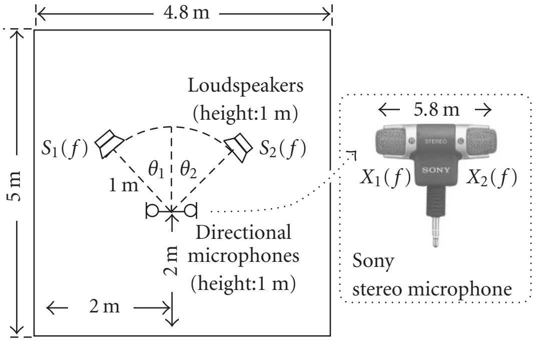


FIGURE 8: Layout of reverberant room used in computer-simulation-based BSS experiment, where room impulse responses are recorded for generation of convolutive mixtures. The reverberation time is 200 milliseconds.

4.2. Experimental evaluation of separation performance

We compare the following methods.

(A) Conventional binary-mask-based BSS that is given in Section 2.3.

(B) Conventional second-order-ICA-based BSS given in Section 2.2, where scaling ambiguity can be properly solved by method used in [8]. Also, permutation is solved by [10]. In this study, we estimate $\mathbf{R}_{xx}(f, \tau)$ and $\mathbf{R}_{yy}(f, \tau)$ at three time instances with each 1 second data,

(C) Conventional higher-order-ICA-based BSS given in Section 2.2 with scaling ambiguity solver [8]. Also, permutation is solved by [9].

(D) Simple combination of conventional higher-order ICA and binary masking.

(E) Proposed two-stage BSS method with $[c_1, c_2, c_3] = [1, 0, 0.1]$; this parameter was determined in the preliminary experiment (performed via various c_i's with 0.1 step) and gave the best performance (high separation but low distortion).

Noise reduction rate (NRR) [10], defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB, is used as the objective measure of separation performance. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise. The input SNR is defined as

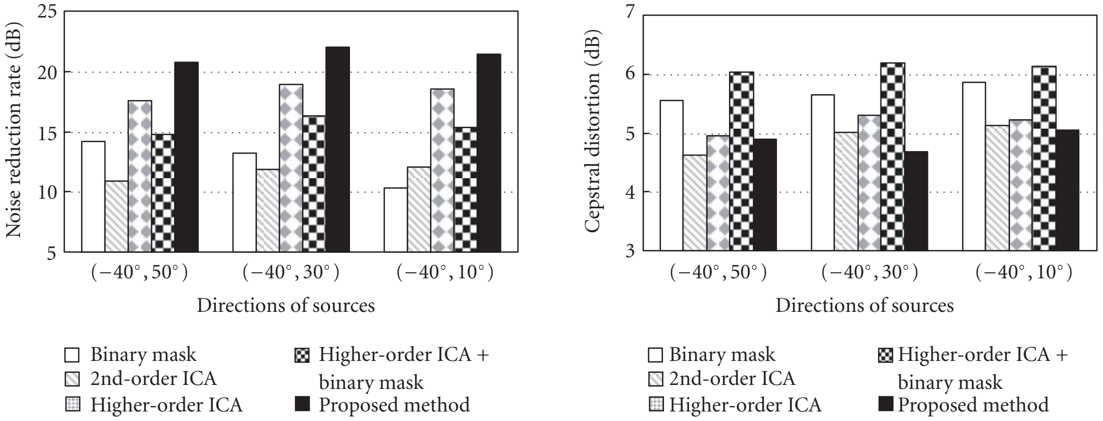
$$\mathrm{ISNR[dB]} = \frac{1}{L}\sum_{l=1}^{L} 10\log_{10}\frac{\langle |A_{ll}(f)S_l(f,t)|^2\rangle_t}{\langle |X_l(f,t) - A_{ll}(f)S_l(f,t)|^2\rangle_t}, \tag{33}$$

and the output SNR is calculated as a ratio between the target component power in the output signal and the interference component power. We obtain these components by inputting SIMO-model-based signals $[A_{1l}(f)S_l(f,t), \ldots, A_{Kl}(f)S_l(f,t)]$ for each source to the separation system, where the separation filter matrices and binary-mask patterns estimated in the preceding blind process with $\mathbf{X}(f,t)$ are used.

Figure 9(a) shows the results of NRR under different speaker configurations. These scores are the averages of 12 speaker combinations. From the results, we can confirm that employing the proposed two-stage BSS can improve the separation performance regardless of the speaker directions, and the proposed BSS outperforms all of the conventional methods. Since the NRR of the SIMO-ICA part in the proposed method was almost the same as that of conventional higher-order ICA, we conclude that the NRR improvements greater than 3 dB can be gained by introducing SIMO-BM.

Since the NRR score indicates only the degree of interference reduction, we could not evaluate the sound quality, that is, the degree of sound distortion, in the previous paragraph. To assess the distortion of the separated signals, we measure *cepstral distortion* (CD) [26], which indicates the distance between the spectral envelopes of the original source signal and the target component in the separated output. CD does not take into account the degree of interference reduction, unlike NRR; thus, CD and NRR are complementary scores. CD is given by

$$\mathrm{CD[dB]} \equiv \frac{1}{J}\sum_{j=1}^{J} D_b \sqrt{\sum_{i=1}^{p} 2\left(C_{\mathrm{out}}(i,j) - C_{\mathrm{ref}}(i,j)\right)^2}, \tag{34}$$



(a)

(b)

Figure 9: (a) Results of NRR and (b) results of CD under different speaker configurations and methods, where background noise is neglected. Each score is an average for 12 speaker combinations.

where J denotes the number of speech frames, $C_{\text{out}}(i, j)$ is the ith FFT-based cepstrum of the target component in the separated output at the jth frame, $C_{\text{ref}}(i, j)$ is the cepstrum of an original source signal, $D_b = 20/\log 10$ indicates the constant value for converting the distance scale to the decibel scale, and the number of liftering points p is 10. CD decreases as the distortion is reduced.

Figure 9(b) shows the results of CD (average of 12 speaker combinations) for all speaker directions. As can be confirmed, the CDs of both conventional ICA and the proposed method are smaller than those of binary masking and its simple combination with ICA. This means that (a) the conventional binary-mask-based methods (A) and (D) involve significant distortion due to the inappropriate time-variant masking arising in the nonsparse frequency subband, (b) but the proposed method cannot be affected by such inappropriateness. It should be mentioned that the simple combination of conventional ICA and binary masking still shows deterioration, and this result is well consistent with the discussion provided in Section 3.2.

These results provide promising evidence that the proposed combination of SIMO-ICA and SIMO-BM is well applicable to low-distortion sound segregation, for example, hands-free telecommunication via mobile phones.

4.3. *Speech recognition experiment*

Next, to evaluate the applicability to speech enhancement, we performed large-vocabulary speech recognition experiments utilizing the proposed BSS as a preprocessing for noise reduction. Table 1 shows the parameter settings in the speech recognition. Sound source 1 ($S_1(f)$) produces 200 sentences of the test sets, and source 2 ($S_2(f)$) produces a different sentence as the interference with a 0 dB mixing condition. Thus, the separation task is to segregate source 1 from the mixtures and recognize it.

Figure 10 shows the results of word recognition performance (word accuracy) for each method, where we can see the proposed method's superiority. The score of the proposed method is obviously better than the scores of binary masking and its simple combination with ICA, and significantly outperforms conventional ICA. Thus, the proposed method is potentially beneficial to noise-robust speech recognition as well as hands-free telephony.

This experiment addressed adverse-condition speech recognition, where the target speech was distorted by improper spectral masking (i.e., artificial spectral hole) as well as contaminated by additive noise. In such a condition, our proposed method is preferable because of the low-distortion property. As an altenative solution, it is reported that missing feature theory can be applicable to the distorted speech [31, 32]. By introducing missing feature theory, we may gain more on the speech recognition accuracy; it still remains as a future work.

Table 1: Parameters of speech recognition experiment.

Database	JNAS [27], 306 speakers (150 sentences/speaker)
Task	20 k newspaper dictation
Acoustic model	Phonetic tied mixture [28] (clean model)
Feature vectors	12-order MFCCs [29], 12-order ΔMFCCs, 1-order Δ energy
Training data	260 speakers' utterances (150 sentences/speaker)
Testing data	46 speakers' utterances (200 sentences)
Decoder	Julius [30] ver.3.4.2
Sampling frequency	16 kHz
Frame length	25 milliseconds
Frame shift	10 milliseconds

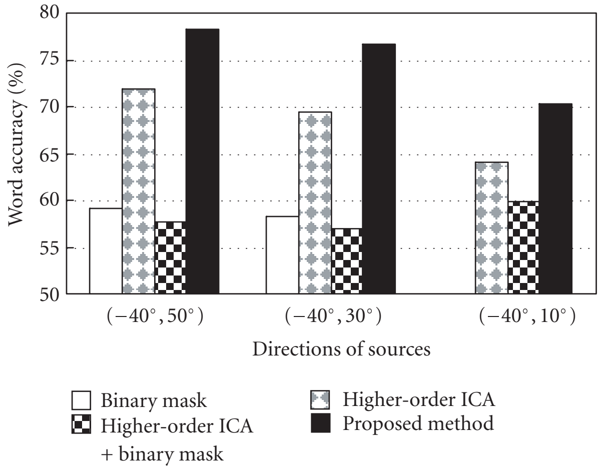


FIGURE 10: Result of word accuracy for different speaker allocations and methods. The recognition task is 20k-word newspaper dictation. Julius decoder [30] is used, where a phonetic tied mixture model was trained via 260 speakers selected from JNAS database [27]. Test sets include 46 speakers' utterances (200 sentences).

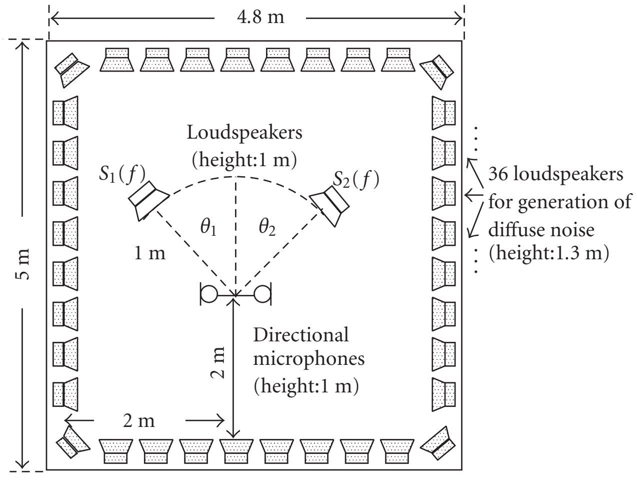


FIGURE 11: Layout of reverberant room used in computer-simulation-based BSS experiment, where 36 loudspeakers simulate heavy background noise. The reverberation time is 200 milliseconds.

5. SPEECH SEPARATION EXPERIMENT UNDER NOISY CONDITIONS

In this section, we consider a specific BSS problem under heavily noisy conditions to assess the proposed method's efficacy in a more challenging situation. As for the additive noise term $\mathbf{N}(f)$ in (1), we create and record a diffuse noise consisting of 36 independent speech signals emitted by surrounding loudspeakers as shown in Figure 11. We add the noise to the two-source two-microphone simulation described in the previous section, where the ratio of mixed two source signals and the noise is set to 20 dB. The other conditions are the same as those of Section 4.1.

We compare the following methods: (A) the conventional binary-mask-based BSS given in Section 2.3, (B) the conventional higher-order-ICA-based BSS given in Section 2.2, (C) the simple combination of conventional ICA and binary masking, and (D) the proposed two-stage BSS method with various c_i parameters.

The results of NRR and CD are shown in Figure 12, where each score is averaged among 12 speaker combinations. We can confirm the following findings. For $(\theta_1, \theta_2) = (-40^\circ, 50^\circ)$, conventional binary masking outperforms the other methods. This is because all the ICA-based methods are harmfully influenced by the separation error due to the background noise, but binary masking is robust against the noise, particularly when the sources are widely apart. For $(\theta_1, \theta_2) = (-40^\circ, 30^\circ)$ or $(-40^\circ, 10^\circ)$, however, the proposed method is superior to the other methods. In comparison with the conventional methods under the same CD level, the proposed method can obtain further NRR improvements with the appropriate c_i parameter settings, for example, $c_3 = 0.5$ for $(-40^\circ, 30^\circ)$ and $c_3 = 0.2$ for $(-40^\circ, 10^\circ)$. Thus, the slight addition of c_3 is preferable in the heavily noisy environment, and can provide higher-quality output signals.

6. REAL-TIME SEPARATION EXPERIMENT FOR MOVING SOUND SOURCE

In this section, we discuss a real-recording-based BSS experiment performed using actual devices in a real acoustic environment. We carried out real-time sound separation using source signals recorded in the real room illustrated in Figure 13, where two loudspeakers and the real-time BSS system (Figure 6) are set. The reverberation time in this room is 200 milliseconds, and the levels of background noise and each of the sound sources measured at the array origin are 39 dB(A) and 65 dB(A), respectively. Two speech signals, whose length is limited to 32 seconds, are assumed to arrive from different directions, θ_1 and θ_2, where we fix source 1 in $\theta_1 = -40^\circ$, and move source 2 as follows:

(1) in the 0–10 seconds duration, source 2 is set to $\theta_2 = 50^\circ$,
(2) in the 10–11 seconds duration, source 2 moves from $\theta_2 = 50^\circ$ to 30°,
(3) in the 11–21 seconds duration, source 2 is settled in $\theta_2 = 30^\circ$,
(4) in the 21–22 seconds duration, source 2 moves from $\theta_2 = 30^\circ$ to 10°,
(5) in 22–32 seconds duration, source 2 is fixed in $\theta_2 = 10^\circ$.

The rest of the experimental conditions are the same as those of the previous experiment described in Section 4.1.

It was difficult to evaluate an accurate NRR in this real environment because we never know the target and interference components separately. In order to calculate NRRs approximately, first, we recorded each sound source individually for making the reference in the SNR calculations, and then we immediately recorded the mixed sounds which are to be processed in the BSS system. We can estimate SNRs by

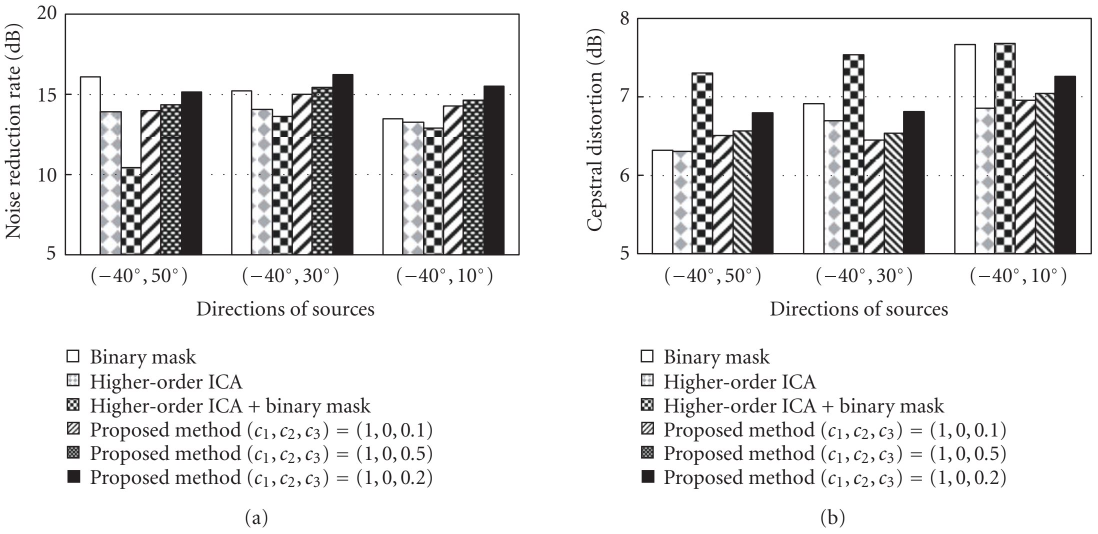


FIGURE 12: (a) Results of NRR and (b) results of CD under different speaker allocations and methods, where background noise (36 independent speech signals) is added with 20 dB SNR. Each score is an average for 12 speaker combinations.

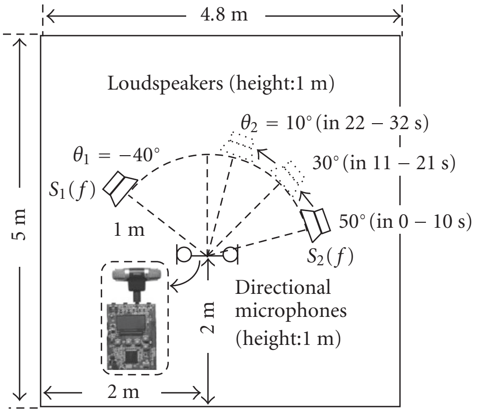

4.8 m
Loudspeakers (height:1 m)
$\theta_2 = 10°$ (in 22 − 32 s)
$\theta_1 = -40°$
30° (in 11 − 21 s)
$S_1(f)$
1 m
50° (in 0 − 10 s)
$S_2(f)$
5 m
Directional microphones (height:1 m)
2 m
2 m

FIGURE 13: Layout of reverberant room used in real-recording-based experiment the reverberation time is 200 milliseconds.

memorizing the separation filter matrices and binary mask patterns along the time axis, and combining them with the individual sound sources.

We compare four methods as follows: (A) the conventional binary-mask-based BSS, (B) the conventional higher-order-ICA-based BSS, (C) the simple combination of conventional ICA and binary masking, and (D) the proposed two-stage BSS method. In the proposed method, we set $[c_1, c_2, c_3] = [1, 0, 0.4]$, which gives the best performance (high NRR but low CD) under this background noise condition.

Figure 14 shows the averaged segmental NRR for 12 speaker combinations, which was calculated along the time axis at 0.5 seconds intervals. The first 3 seconds duration is spent on the initial filter learning of ICA in methods (B), (C), and (D), and thus the valid ICA-based separation filter is absent here. Therefore, in the period of 0–3 seconds, we simply applied binary masking in methods (C) and (D). The successive duration (in the period of 3–32 seconds) shows the separation results for the *open* data sample, which is to be evaluated in this experiment. From Figure 14, we can confirm that the proposed two-stage BSS (D) outperforms other methods throughout almost the entire duration of 3–32 seconds. It is worth noting that conventional ICA shows appreciable deteriorations especially in the 2nd source's moving periods, that is, around 10 seconds and 21 seconds, but the proposed method can mitigate the degradations. On the basis of these results, we can assess the proposed method to be beneficial to many practical real-time BSS applications.

7. CONCLUSION

We proposed a new BSS framework in which SIMO-ICA and a new SIMO-BM are efficiently combined. SIMO-ICA is an algorithm for separating the mixed signals, not into monaural source signals but into SIMO-model-based signals of independent sources without losing their spatial qualities. Thus, after SIMO-ICA, we can introduce the novel SIMO-BM and succeed in removing the residual interference components.

In order to evaluate its effectiveness, many separation experiments were carried out under a 200-milliseconds-reverberation-time condition. The experimental results revealed that the SNR can be considerably improved by the proposed two-stage BSS algorithm with no increase in signal distortion. In addition, we found that the proposed method outperforms the combination of conventional ICA and binary masking as well as of a simple ICA and binary masking. The efficacy of the proposed method was confirmed in various separation tasks, that is, an offline test, a noisy-environment test, and an online test using a DSP module applied for real recording data.

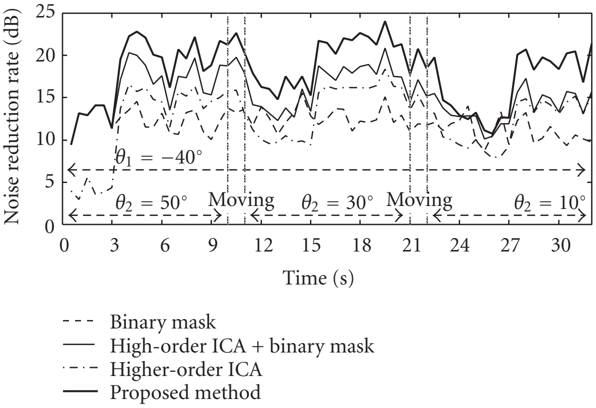


Figure 14: Results of segmental NRR calculated along time axis at 0.5 seconds intervals, where real-recording data and real-time BSS are used. Each line is an average for 12 speaker combinations. The levels of background noise and sound source are 39 dB(A) and 65 dB(A), respectively.

As described in Section 3, there is a possibility, in theory, that the proposed method can deal with the case of $K = L > 2$. However, only the results for $K = L = 2$ were shown in this paper. Therefore, further study for $K = L > 2$ and a method of estimating the number of sources remain as open problems for the future.

Although the proposed method does not require the accurate DOAs of sources in advance, it is still needed to set the array in the proper direction towards the sources. The proposed method has an inherent limitation that the separation performance degrades if the sources are located at the same side of the array, as in the case of conventional binary masking. This is due to the fact that the binary masking in the second stage cannot utilize the original assumption on the source amplitude difference between directional microphones, that is, the left/right-hand-side microphone has a large gain corresponding to the left/right source. For example, in our experiment with $(\theta_1, \theta_2) = (-80°, -10°)$, the NRR scores of binary masking, conventional ICA, and the proposed method are 8.9, 14.6, and 11.1 dB, respectively. Further improvement is requisite in a future work.

APPENDICES

A. UNIQUE SOLUTION IN FD-SIMO-ICA

In this section, we will prove (16) under the condition that the residual error $\mathbf{E}_l(f,t) = 0$. Please note that the original version of this proof has been presented in our previous work [12] with a time-domain representation, but we hereafter show the modified version with a frequency-domain representation for the readers' convenience.

Theorem A.1. *The output signals converge towards unique SIMO solutions* (16) *up to the permutation* $\mathbf{P}_l (l = 1, \ldots, L)$, *given by* (17), *if and only if the independent sound sources are separated as defined by* (14) *and simultaneously the signals obtained using* (15) *are mutually independent.*

Proof. The necessity is obvious. The sufficiency is shown below. Let $\mathbf{D}_l$ be arbitrary diagonal polynomial matrices and $\mathbf{Q}_l$ be arbitrary permutation matrices. The general expression of the lth ICA's output is given by

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \mathbf{D}_l \mathbf{Q}_l \mathbf{S}(f,t). \tag{A.1}$$

If $\mathbf{Q}_l$ are not exclusively selected matrices, that is,

$$\sum_{l=1}^{L} \mathbf{Q}_l \neq [1]_{ij}, \tag{A.2}$$

then there exists at least one element of $\sum_{l=1}^{L} \mathbf{Y}_{(\mathrm{ICA}l)}(f,t)$ which does not include all of the components of $S_l(f,t)$ $(l = 1, \ldots, L)$. This obviously makes the left-hand side of the next equation, which consists of (14) and (15):

$$\mathbf{X}(f,t) - \sum_{l=1}^{L} \mathbf{Y}_{(\mathrm{ICA}l)}(f,t) \equiv [0]_{m1}, \tag{A.3}$$

nonzero because the observed signal vector $\mathbf{X}(f,t)$ includes all of the components of $S_l(f,t)$ in each element. Accordingly, $\mathbf{Q}_l$ should be the $\mathbf{P}_l$ specified by (17), and we obtain

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \mathbf{D}_l \mathbf{P}_l \mathbf{S}(f,t). \tag{A.4}$$

In (A.4) under (17), the arbitrary diagonal matrices $\mathbf{D}_l$ can be substituted with $\mathrm{diag}[\mathbf{B}\mathbf{P}_l^{\mathrm{T}}]$, where $\mathbf{B} = [B_{ij}]_{ij}$ is a single arbitrary matrix, because all diagonal entries of $\mathrm{diag}[\mathbf{B}\mathbf{P}_l^{\mathrm{T}}]$ for all l's are also exclusive. Thus,

$$\mathbf{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\left[\mathbf{B}\mathbf{P}_l^{\mathrm{T}}\right]\mathbf{P}_l \mathbf{S}(f,t). \tag{A.5}$$

Substituting (A.5) into (15) leads to the following equation:

$$\mathrm{diag}\left[\mathbf{B}\mathbf{P}_L^{\mathrm{T}}\right]\mathbf{P}_L \mathbf{S}(f,t) = \mathbf{X}(f,t) - \sum_{l=1}^{L-1} \mathrm{diag}\left[\mathbf{B}\mathbf{P}_l^{\mathrm{T}}\right]\mathbf{P}_l \mathbf{S}(f,t), \tag{A.6}$$

and consequently

$$\begin{aligned}
&\sum_{l=1}^{L} \mathrm{diag}\left[\mathbf{B}\mathbf{P}_l^{\mathrm{T}}\right]\mathbf{P}_l \mathbf{S}(f,t) - \mathbf{X}(f,t) \\
&\quad = \left[\sum_{l=1}^{L} B_{kl} S_l(f,t)\right]_{k1} - \left[\sum_{l=1}^{L} A_{kl}(f) S_l(f,t)\right]_{k1} \\
&\quad = \left[\sum_{l=1}^{L} \{B_{kl} - A_{kl}(f)\} S_l(f,t)\right]_{k1} = [0]_{k1}.
\end{aligned} \tag{A.7}$$

Equation (A.7) is satisfied if and only if $B_{kl} = A_{kl}(f)$ for all values of k and l. Thus, (A.5) results in (16). This completes the proof of the theorem. □

B. DERIVATION OF (20)

Here, Kullback-Leibler divergence between the joint probability density function (PDF) of $\mathbf{Y}(f,t)$ and the product of marginal PDFs of $Y_l(f,t)$ is defined by $\mathrm{KLD}(\mathbf{Y}(f,t))$. The gradient of $\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))$ with respect to $\mathbf{W}_{(\mathrm{ICA}l)}(f)$ should be added to the iterative learning rule of the separation filter in the lth ICA ($l = 1,\ldots,L-1$). We obtain the partial differentiation (standard gradient) of $\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))$ with respect to $\mathbf{W}_{(\mathrm{ICA}l)}(f)$ ($l = 1,\ldots,L-1$) as

$$\begin{aligned}
&\frac{\partial\,\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial \mathbf{W}_{(\mathrm{ICA}l)}(f)}\\
&\quad= \left[\frac{\partial\,\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial W_{ij}^{(\mathrm{ICA}L)}(f)} \cdot \frac{\partial W_{ij}^{(\mathrm{ICA}L)}(f)}{\partial W_{ij}^{(\mathrm{ICA}l)}(f)}\right]_{ij}\\
&\quad= \left[\frac{\partial\,\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial W_{ij}^{(\mathrm{ICA}L)}(f)} \cdot (-1)\right]_{ij},
\end{aligned}\tag{B.1}$$

where $W_{ij}^{(\mathrm{ICA}L)}(f)$ is the element of $\mathbf{W}_{(\mathrm{ICA}L)}(f)$. By replacing $\partial\,\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))/\partial \mathbf{W}_{(\mathrm{ICA}L)}(f)$ with its natural gradient [33], we modify (B.1) as

$$\begin{aligned}
&-\frac{\partial\,\mathrm{KLD}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial \mathbf{W}_{(\mathrm{ICA}L)}(f)} \cdot \mathbf{W}_{(\mathrm{ICA}L)}^{\mathrm{H}}(f)\mathbf{W}_{(\mathrm{ICA}L)}(f)\\
&\quad= \left\{\mathbf{I} - \left\langle \mathbf{\Phi}(\mathbf{Y}_{(\mathrm{ICA}L)}(f,t)) \cdot \mathbf{Y}_{(\mathrm{ICA}L)}^{\mathrm{H}}(f,t)\right\rangle_t\right\}\\
&\qquad \cdot \mathbf{W}_{(\mathrm{ICA}L)}(f).
\end{aligned}\tag{B.2}$$

By inserting (15) and the relation of $\mathbf{W}_{(\mathrm{ICA}L)}(f) = \mathbf{I} - \sum_{l=1}^{L-1}\mathbf{W}_{(\mathrm{ICA}l)}(f)$ into (B.2), we obtain

$$\begin{aligned}
&\left\{\mathbf{I} - \left\langle \mathbf{\Phi}\left(\mathbf{X}(f,t) - \sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t)\right)\right.\right.\\
&\qquad \left.\left. \cdot \left(\mathbf{X}(f,t) - \sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t)\right)^{\mathrm{H}}\right\rangle_t\right\}\\
&\qquad \cdot \left(\mathbf{I} - \sum_{l=1}^{L-1}\mathbf{W}_{(\mathrm{ICA}l)}(f)\right).
\end{aligned}\tag{B.3}$$

In order to deal with non-i.i.d. signals, we apply the nonholonomic constraint [34] to (B.3). The natural gradient with the nonholonomic constraint is given as

$$\begin{aligned}
&-\left\{\text{off-diag}\left\langle \mathbf{\Phi}\left(\mathbf{X}(f,t) - \sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t)\right)\right.\right.\\
&\qquad \left.\left. \cdot \left(\mathbf{X}(f,t) - \sum_{l=1}^{L-1}\mathbf{Y}_{(\mathrm{ICA}l)}(f,t)\right)^{\mathrm{H}}\right\rangle_t\right\}\\
&\quad \cdot \left(\mathbf{I} - \sum_{l=1}^{L-1}\mathbf{W}_{(\mathrm{ICA}l)}(f)\right).
\end{aligned}\tag{B.4}$$

Thus, the new iterative algorithm of the lth ICA part ($l = 1,\ldots,L-1$) in SIMO-ICA is given by adding (B.4) into the existing ICA equation, and we obtain (20).

C. DIFFERENCE BETWEEN SIMO-ICA AND PROJECTION-BACK METHOD

In the projection-back (PB) method, the following operation is performed after (4):

$$\begin{aligned}
Y_l^{(k)}(f,t) &= \left\{\mathbf{W}(f)^{-1}\left[\overbrace{0,\ldots,0}^{l-1}, Y_l(f,t), \overbrace{0,\ldots,0}^{L-l}\right]^{\mathrm{T}}\right\}_k\\
&= \left(\det\mathbf{W}(f)\right)^{-1}\Delta_{lk} \cdot Y_l(f,t),
\end{aligned}\tag{C.1}$$

where $Y_l^{(k)}(f,t)$ represents the lth resultant separated source signal which is projected back onto the kth microphone, $\{\cdot\}_k$ denotes the kth element of the argument, and Δ_{kl} is a cofactor of the matrix $\mathbf{W}(f)$.

This method is simpler than SIMO-ICA, but its inversion often fails and yields harmful results because the invertibility of every $\mathbf{W}(f)$ cannot be guaranteed [35]. Also, there exists another improper issue for the combination of ICA and binary masking as shown below. In PB, spatial information (amplitude difference between directional microphones) in the target signal is just similar to that in the interference because the projection operator $(\det\mathbf{W}(f))^{-1}\Delta_{lk}$ is applied to not only the target signal component but also the interference component in $Y_l(f,t)$. For example, similar to Section 3.2, (C.1) leads to

$$\begin{aligned}
Y_l^{(k)}(f,t) &= \left(\det\mathbf{W}(f)\right)^{-1}\Delta_{lk} \cdot \left(B_l(f)S_l(f,t) + E_l(f,t)\right)\\
&= \left(\det\mathbf{W}(f)\right)^{-1}\Delta_{lk} \cdot B_l(f)S_l(f,t)\\
&\quad + \left(\det\mathbf{W}(f)\right)^{-1}\Delta_{lk} \cdot E_l(f,t),
\end{aligned}\tag{C.2}$$

where we can assume that $|(\det\mathbf{W}(f))^{-1}\Delta_{ll}|$ is the largest value among $|(\det\mathbf{W}(f))^{-1}\Delta_{lk}|$ ($k = 1,\ldots,K$) for the lth source in our directional-microphone-use scenario. Thus, when the target signal component $S_l(f,t)$ is not silent, binary masking can approximately extract $S_l(f,t)$ component because the first term in the right-hand side in (C.2) becomes the most dominant just in $k = l$ among $Y_l^{(k)}(f,t)$ for all k. However, the problem is that, when $S_l(f,t)$ is almost silent, binary masking has to pick up (i.e., cannot mask) the *undesired* $E_l(f,t)$ component because the second term in the right-hand side in (C.2) also becomes the most dominant in $k = l$. This fact yields the negative result that the PB method is *not* available to a residual-noise reduction purpose via the combination of SIMO-model-based signals and binary masking. In contrast to the PB method, SIMO-ICA holds the applicability to the combination with binary masking because the separation filter of SIMO-ICA cannot always be represented in the PB form, that is, we are often confronted with the case that the residual-noise component in the $k(\neq l)$th microphone has the largest amplitude even among $Y_l^{(k)}(f,t)$.

ACKNOWLEDGMENTS

The authors thank Dr. Hiroshi Sawada, Mr. Ryo Mukai, Mrs. Shoko Araki, and Dr. Shoji Makino of NTT CS-Lab. for fruitful discussions on this work. This work was partially supported by CREST "Advanced Media Technology for Everyday Living" of JST, and by MEXT e-Society leading project in Japan.

REFERENCES

[1] S. Haykin, Ed., *Unsupervised Adaptive Filtering*, John Wiley & Sons, New York, NY, USA, 2000.

[2] J. F. Cardoso, "Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '89)*, pp. 2109–2112, Glasgow, UK, May 1989.

[3] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[4] P. Comon, "Independent component analysis. A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

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

[6] T.-W. Lee, *Independent Component Analysis*, Kluwer Academic, Norwell, Mass, USA, 1998.

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

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

[9] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[10] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[11] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 4, pp. 846–858, 2003.

[12] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 251–254, Kyoto, Japan, September 2003, (also submitted to IEEE Transactions on Speech and Audio Processing).

[13] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proceedings of 5th International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 832–839, Granada, Spain, September 2004.

[14] R. Lyon, "A computational model of binaural localization and separation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '83)*, pp. 1148–1151, Boston, Mass, USA, April 1983.

[15] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '01)*, vol. 4, pp. 2861–2866, Washington, DC, USA, July 2001.

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

[17] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 3, pp. 590–596, 2003.

[18] H. Saruwatari, T. Kawamura, T. Nishikawa, and K. Shikano, "Fast-convergence algorithm for blind source separation based on array signal processing," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 4, pp. 286–291, 2003.

[19] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[20] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 529–532, Orlando, Fla, USA, May 2002.

[21] T. Takatani, S. Ukai, T. Nishikawa, H. Saruwatari, and K. Shikano, "A self-generator method for initial filters of SIMO-ICA applied to blind separation of binaural sound mixtures," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 7, pp. 1673–1682, 2005.

[22] A. Poularikas, *The Handbook of Formulas and Tables for Signal Processing*, CRC Press, Boca Raton, Fla, USA, 1999.

[23] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E87-A, no. 8, pp. 1941–1948, 2004.

[24] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[25] T. Kobayashi, S. Itabashi, S. Hayashi, and T. Takezawa, "ASJ continuous speech corpus for research," *The Journal of The Acoustic Society of Japan*, vol. 48, no. 12, pp. 888–893, 1992 (Japanese).

[26] J. J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE Press, New York, NY, USA, 2000.

[27] K. Itou, M. Yamamoto, K. Takeda, et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of The Acoustic Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.

[28] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, pp. 1269–1272, Istanbul, Turkey, June 2000.

[29] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[30] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 1691–1694, Aalborg, Danemark, September 2001.

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

[32] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '05)*, pp. 82–85, New Paltz, NY, USA, October 2005.

[33] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley & Sons, West Sussex, UK, 2002.

[34] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Proceedings of 1st International Workshop on Independent Component Analysis and Blind Source Separation (ICA '99)*, pp. 371–376, Aussois, France, January 1999.

[35] T. Nishikawa, H. Saruwatari, and K. Shikano, "Stable learning algorithm for blind separation of temporally correlated acoustic signals combining multistage ICA and linear prediction," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 8, pp. 2028–2036, 2003.

Yoshimitsu Mori was born in Gifu, Japan, in 1981. He received the B.E. degree in electronic engineering from Nagoya Institute of Technology in 2004 and received the M.E. degree in electronic engineering form Nara Institute of Science and Technology (NAIST) in 2006. He is now a Ph.D. student at Graduate School of Information Science, NAIST. His research interests include array signal processing and blind source separation. He is a Member of the IEICE and the Acoustical Society of Japan.

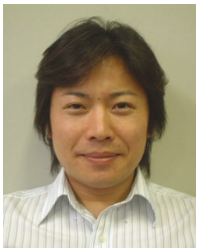
Hiroshi Saruwatari was born in Nagoya, Japan, on 27 July, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he was engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2000 and 2006. He is a Member of the IEEE, the VR Society of Japan, the IEICE, and the Acoustical Society of Japan.

Tomoya Takatani was born in Hyogo, Japan, in 1977. He received the B.E. degree in electronics from Doshisha University in 2001 and received the M.E. and Ph.D. degrees in electronic engineering form NAIST in 2003 and 2006. His research interests include array signal processing and blind source separation. He is a Member of the IEICE and the Acoustical Society of Japan.

Satoshi Ukai was born in Shiga, Japan, in 1980. He received the B.E. degree in electronic engineering from Kobe University in 2003 and received the M.E. degree in electronic engineering form NAIST in 2005. His research interests include array signal processing and blind source separation. He is a Member of the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a Professor of Nara Institute of Science and Technology (NAIST), where he is directing Speech and Acoustics Laboratory. From 1972, he worked at NTT Laboratories, where he was engaged in Speech Recognition Research. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories. During 1984–1986, he was a Visiting Scientist in Carnegie Mellon University. He received the IEICE (Institute of Electronics, Information and Communication Engineers of Japan) Yonezawa Prize in 1975, IEEE Signal Processing Society 1990 Senior Award in 1991, the Technical Development Award from ASJ (Acoustical Society of Japan) in 1994, IPSJ (Information Processing Society of Japan) Yamashita SIG Research Award in 2000, Paper Award from the Virtual Reality Society of Japan in 2001, IEICE Paper Award in 2005 and 2006, and IEICE Inose Best Paper Award in 2005. He is a Fellow Member of IEICE and IPSJ. He is a Member of ASJ, Japan VR Society, IEEE, and International Speech Communication Association.

Takashi Hiekata was born in Kobe, Japan, in 1969. He received the B.E., M.E. degrees in Computer and Systems engineering from Kobe University in 1992 and 1994, respectively. He joined Production Systems Research Laboratory, KOBE STEEL, LTD., Kobe, Japan, where he was engaged in the research and development on the digital signal processing. He is a Member of the IEICE, and the Acoustical Society of Japan (ASJ).



Youhei Ikeda was born in Osaka, Japan, in 1975. He received the B.E. degree in industrial engineering from Osaka Prefecture University in 1999 and the M.E. degree in information and science from Nara Institute of Science and Technology (NAIST) in 2001. He joined Production Systems Research Laboratory, KOBE STEEL, LTD., Kobe, Japan, where he was engaged in the research and development on the digital signal processing. He is a Member of the IEICE.

Hiroshi Hashimoto was born in Hyogo, Japan, in 1966. He received the B.E., M.E. degrees in electrical engineering from Kobe University in 1989 and 1991, respectively. He joined Production Systems Research Laboratory, KOBE STEEL, LTD., Kobe, Japan, where he was engaged in the research and development on the digital signal processing.

Takashi Morita was born in Tottori, Japan, in 1962. He received the B.E. and M.E. degrees in control engineering from Osaka University in 1984 and 1986, respectively. He joined Production Systems Research Laboratory, KOBE STEEL, LTD., Kobe, Japan, where he was engaged in the research and development on the digital signal processing.