MLSP 2007 DATA ANALYSIS COMPETITION: TWO-STAGE BLIND SOURCE SEPARATION COMBINING SIMO-MODEL-BASED ICA AND BINARY MASKING

Yoshimitsu Mori, Keiichi Osako, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano

Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan
phone: +81-743-72-5287, fax: +81-743-72-5289,
email: {yoshim-m, keiichi-o, shige-m, yuu-t, sawatari, shikano}@is.naist.jp

ABSTRACT

This paper reviews a real-time two-stage blind source separation (BSS) method for convolutive mixtures of speech, in which a single-input multiple-output (SIMO)-model-based independent component analysis (ICA) and a SIMO-model-based binary masking are combined. SIMO-model-based ICA can 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, SIMO-model-based binary masking can be applied to efficiently remove the residual interference components after SIMO-model-based ICA. In addition, the performance deterioration due to the latency problem in ICA can be mitigated by introducing real-time binary masking. We report the parameters used in MLSP 2007 data analisys, and the experimental evaluation of the proposed method's superiority to the conventional BSS methods, regarding static- and moving-sound separation.

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 into *unsupervised* filtering technique, and much attention has been paid to BSS in many fields of signal processing.

In recent researches of BSS based on independent component analysis (ICA), various methods have been presented for acoustic-sound separation [1, 2, 3]. This paper also addresses the BSS problem under highly reverberant conditions (e.g., reverberation time is more than 200 ms) which often arise in many audio applications. The separation performance of the conventional ICA is far from being sufficient in the reverberant case because too long separation filters is required but the unsupervised learning of the filter is not so easy. Therefore, one possible improvement is to partly combine ICA with another signal enhancement technique, but in the conventional ICA, each of the separated outputs is a *single-channel* signal, and this leads to the drawback that many kinds of superior *multichannel* techniques cannot be applied.

In order to attack the tough problem, we have proposed a two-stage BSS algorithm [4]. In this paper, we give a detailed review on the proposed method. This approach resolves the BSS problem into two stages: (a) a Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) proposed by the authors' group [5] and (b) a SIMO-model-based binary masking (SIMO-BM) for the signals obtained from the SIMO-ICA. SIMO-ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of SIMO-ICA can maintain rich spatial qualities of each sound source. After the SIMO-ICA, the residual components of the interference, which are often staying in the output of SIMO-ICA as well as the conventional ICA, can be efficiently removed by the following SIMO-BM.

It should be enhanced that the two-stage method has another important property, i.e., applicability to the real-time processing. In general ICA-based BSS methods require huge calculations, but SIMO-model-based binary masking needs very few computational complexities. Therefore, because of the introduction of binary masking into ICA, the proposed combination can function as the real-time system. In this paper, we introduce the detailed parameters used for "MLSP 2007 Data Analysis Competition" for nonlinear system [6]. Also, we evaluate the "real-time" separation performance for real recording of static and moving sound mixtures under a reverberant condition.

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. In the frequency domain, the observed signals in which multiple sources are mixed are given by

$$\boldsymbol{X}(f) = \boldsymbol{A}(f)\boldsymbol{S}(f), \quad (1)$$

where $\boldsymbol{X}(f) = [X_1(f), \cdots, X_K(f)]^{\mathrm{T}}$ is the observed signal vector, and $\boldsymbol{S}(f) = [S_1(f), \cdots, S_L(f)]^{\mathrm{T}}$ is the source signal vector. Also, $\boldsymbol{A}(f) = [A_{kl}(f)]_{kl}$ is the mixing matrix, where $[X]_{ij}$ denotes the matrix which includes the element X in the i-th row and the j-th column. The mixing matrix $\boldsymbol{A}(f)$ 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 the frequency-domain ICA (FDICA), first, the short-time analysis of observed signals is conducted by frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider them as a time series. Hereafter, we designate the time series as $\boldsymbol{X}(f,t) = [X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}$.

Next, we perform signal separation using the complex-valued unmixing matrix, $\boldsymbol{W}(f) = [W_{lk}(f)]_{lk}$, so that the L time-series

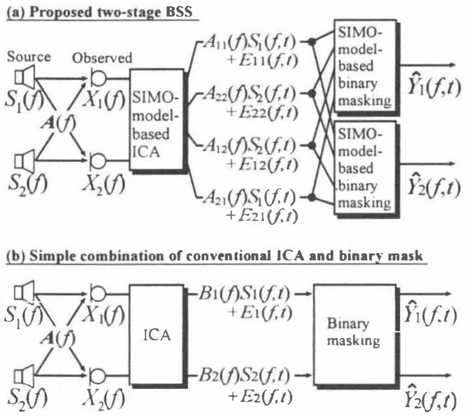

Fig. 1. Input and output relations in (a) proposed two-stage BSS and (b) simple combination of conventional ICA and binary masking. This corresponds to the case of $K = L = 2$.

output $\boldsymbol{Y}(f,t)=[Y_1(f,t),\cdots,Y_L(f,t)]^{\mathrm{T}}$ becomes mutually independent; this procedure can be given as $\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t)$. We perform this procedure with respect to all frequency bins. The optimal $\boldsymbol{W}(f)$ is obtained by, e.g., the following iterative updating equation [1]:

$$\boldsymbol{W}^{[i+1]}(f) = \eta \left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t) \right\rangle_t \right] \boldsymbol{W}^{[i]}(f) + \boldsymbol{W}^{[i]}(f), \quad (2)$$

where $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle_t$ denotes the time-averaging operator, $[i]$ is used to express the value of the i th step in the iterations, η is the step-size parameter, and $\boldsymbol{\Phi}(\cdot)$ is the appropriate nonlinear vector function. After the iterations, the source permutation and the scaling indeterminacy problem can be solved by, e.g., [1, 3].

2.3. Conventional Binary-Mask-Based BSS

Binary mask processing [7] is one of the alternative approach which is aimed to solve 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 speech. This procedure is performed in time-frequency regions, and is to pass the specific regions where target speech is dominant and mask the other regions. Under the assumption that the l-th sound source is close to the l-th microphone, the l-th separated signal is given by

$$\hat{Y}_l(f,t) = m_l(f,t)X_l(f,t), \quad (3)$$

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 few computational complexities, and this property is well applicable to real-time processing. The method, however, needs a sparseness assumption in the sources' spectral components, i.e., there are no overlaps in time-frequency components of the sources. Indeed the assumption does not hold in an usual audio application, e.g., a mixture of speech and a common broadband stationary noise.

3. PROPOSED TWO-STAGE BSS ALGORITHM

3.1. Motivation and Strategy

In the previous research, SIMO-model-based ICA (SIMO-ICA) was proposed by the authors' group [5], who showed that the SIMO-model-based separated signals are still *one set of array signals.*

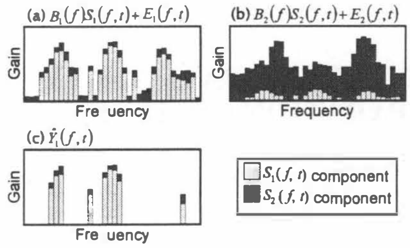

Fig. 2. Examples of spectra in simple combination of ICA and binary masking. (a) ICA's output 1; $B_1(f)S_1(f,t) + E_1(f,t)$, (b) ICA's output 2; $B_2(f)S_2(f,t) + E_2(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.

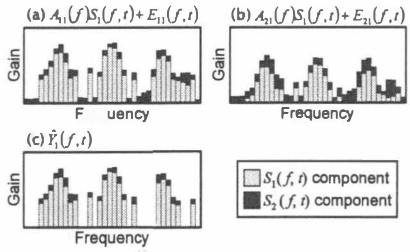

Fig. 3. Examples of spectra in proposed two-stage method. (a) SIMO-ICA's output 1; $A_{11}(f)S_1(f,t) + E_{11}(f,t)$, (b) SIMO-ICA's output 2; $A_{21}(f)S_1(f,t) + E_{21}(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.

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, i.e, the l-th microphone steers to the l-th sound source. Thus the outputs of SIMO-ICA is the estimated (separated) SIMO-model-based signals, and they keep the relation that the l-th source component is the most dominant in the l-th microphone. This finding has motivated us to combine SIMO-ICA and binary masking. Moreover we have proposed a binary masking strategy, so-called *SIMO-model-based binary masking* (SIMO-BM). That is, the 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 Fig. 1(a). SIMO-BM, which subsequently follows SIMO-ICA, can 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 the SIMO-BM. 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 Fig. 1(b)) [8].

In general, conventional ICAs can only supply the source signals $Y_l(f,t) = B_l(f)S_l(f,t) + E_l(f,t)$ $(l = 1,\cdots,L)$, 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

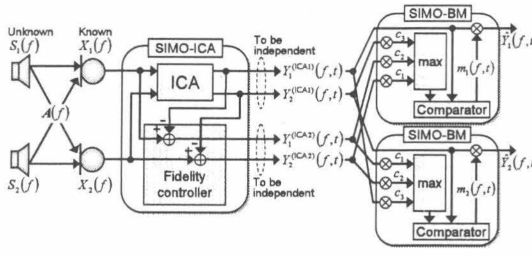

Fig. 4. Input and output relations in proposed two-stage BSS which consists of FD-SIMO-ICA and SIMO-BM, where $K = L = 2$ and exclusively selected permutation matrices are given by $\boldsymbol{P}_1 = \boldsymbol{I}$ and $\boldsymbol{P}_2 = [1]_{ij} - \boldsymbol{I}$. in (8)

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 Fig. 2(a),(b)), i.e.,

$$|B_1(f)S_1(f,t) + E_1(f,t)| \simeq |B_2(f)S_2(f,t) + E_2(f,t)|, \quad (4)$$

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 Fig. 2(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 with respect to each of sources, $A_{kl}(f)S_l(f,t)$, up to the possible residual error $E_{kl}(f,t)$ (see Fig. 3). Needless to say, 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 $A_{11}(f)S_1(f,t)+E_{11}(f,t)$ (Fig. 3(a)) and $A_{21}(f)S_1(f,t)+E_{21}(f,t)$ (Fig. 3(b)) is more acoustically reasonable than the conventional combination because the spatial properties, in which the separated SIMO component at the specific microphone closer to the target sound still maintains a large gain, are kept; i.e.,

$$|A_{11}(f)S_1(f,t) + E_{11}(f,t)| > |A_{21}(f)S_1(f,t) + E_{21}(f,t)|. \quad (5)$$

In this case we can correctly pick up the target signal candidate $A_{11}(f)S_1(f,t) + E_{11}(f,t)$ (see Fig. 3(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 [8], our proposed binary masking is not affected by the amplitude balance among sources. Overall, after obtaining the SIMO components, we can introduce the SIMO-BM for the efficient reduction of the remaining error in ICA, even when the complete sparseness assumption does not hold.

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 a realization of the robust BSS. The detailed algorithm is described in the next subsection.

3.2. Algorithm: SIMO-ICA in 1st Stage

Time-domain SIMO-ICA [5] 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 Fig. 4). The separated signals of the l-th ICA ($l = 1, \cdots L-1$) in FD-SIMO-ICA are defined by

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = [Y_k^{(\mathrm{ICA}l)}(f,t)]_{k1} = \boldsymbol{W}_{(\mathrm{ICA}l)}(f)\boldsymbol{X}(f,t), \quad (6)$$

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

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

$$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t) = \boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t). \quad (7)$$

Hereafter, we regard $\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ as an output of a *virtual* "L-th" ICA. The reason we use the word "*virtual*" here is that the L-th ICA does not have its own separation filters unlike the other ICAs, and $\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)$ is subject to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ ($l = 1, \cdots, L-1$). By transposing the second term ($-\sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)$) on the right-hand side to the left-hand side, we can show that (7) suggests a constraint to force the sum of all ICAs' output vectors $\sum_{l=1}^{L} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)$ to be the sum of all SIMO components $[\sum_{l=1}^{L} A_{kl}(f)\, S_l(f,t)]_{k1}$ $(= \boldsymbol{X}(f,t))$.

If the independent sound sources are separated by (6), and simultaneously the signals obtained by (7) are also mutually independent, then the output signals converge on unique solutions, up to the permutation and the residual error, as

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\left[\boldsymbol{A}(f)\boldsymbol{P}_l^{\mathrm{T}}\right]\boldsymbol{P}_l\boldsymbol{S}(f,t) + \boldsymbol{E}_l(f,t), \quad (8)$$

where diag$[\boldsymbol{X}]$ is the operation for setting every off-diagonal element of the matrix $\boldsymbol{X}$ to zero, $\boldsymbol{E}_l(f,t)$ represents the residual error vector, and $\boldsymbol{P}_l$ ($l = 1, \cdots, L$) are exclusively-selected permutation matrices which satisfy $\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}$. For a proof of this, see [5] with an appropriate modification into the frequency-domain representation. Obviously, the solutions provide necessary and sufficient SIMO components, $A_{kl}(f)S_l(f,t)$, for each l-th 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

$$\left[Y_1^{(\mathrm{ICA1})}(f,t),\ Y_2^{(\mathrm{ICA1})}(f,t)\right]^{\mathrm{T}} = \left[A_{11}(f)S_1(f,t),\ A_{22}(f)S_2(f,t)\right]^{\mathrm{T}}, \quad (9)$$

$$\left[Y_1^{(\mathrm{ICA2})}(f,t),\ Y_2^{(\mathrm{ICA2})}(f,t)\right]^{\mathrm{T}} = \left[A_{12}(f)S_2(f,t),\ A_{21}(f)S_1(f,t)\right]^{\mathrm{T}}, \quad (10)$$

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

In order to obtain (9) and (10), the natural gradient of Kullback-Leibler divergence of (7) with respect to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ should be added to the existing nonholonomic iterative learning rule [1] of the separation filter in the l-th ICA ($l = 1, \cdots, L-1$). The new iterative algorithm of the l-th ICA part ($l = 1, \cdots, L-1$) in FD-SIMO-ICA is given as (see Appendix A.1)

$$\boldsymbol{W}_{(\mathrm{ICA}l)}^{[j+1]}(f) = \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f) - \alpha\left[\left\{ \text{off-diag}\left\langle \boldsymbol{\Phi}(\boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)) \boldsymbol{Y}_{(\mathrm{ICA}l)}^{[j]}(f,t)^{\mathrm{H}}\right\rangle_t \right\} \cdot \boldsymbol{W}_{(\mathrm{ICA}l)}^{[j]}(f)\right.$$

$$- \left\{ \text{off-diag}\left\langle \boldsymbol{\Phi}\Big(\boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}^{[j]}_{(\mathrm{ICA}l)}(f,t)\Big) \cdot \Big(\boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}^{[j]}_{(\mathrm{ICA}l)}(f,t)\Big)^{\mathrm{H}} \right\rangle_t \right\} \cdot \Big(\boldsymbol{I} - \sum_{l=1}^{L-1} \boldsymbol{W}^{[j]}_{(\mathrm{ICA}l)}(f)\Big) \Bigg], \qquad (11)$$

where α is the step-size parameter. Also, the initial values of $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ for all l values should be different.

Note that there exists an alternative method [1] of obtaining the SIMO components in which the separated signals are projected back onto the microphones by using the inverse of $\boldsymbol{W}(f)$ after conventional ICA. The difference and advantage of SIMO-ICA relative to this method are described in Appendix A.2.

3.3. Algorithm: SIMO-BM in 2nd Stage

After FD-SIMO-ICA, SIMO-model-based binary masking is applied (see Fig. 4). Here, we consider the case of (9) and (10). The resultant output signal corresponding to source 1 is determined in the proposed SIMO-BM as follows:

$$\hat{Y}_1(f,t) = m_1(f,t) Y_1^{(\mathrm{ICA1})}(f,t), \qquad (12)$$

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

$$Y_1^{(\mathrm{ICA1})}(f,t) > \max\big[c_1|Y_2^{(\mathrm{ICA2})}(f,t)|,\ c_2|Y_1^{(\mathrm{ICA2})}(f,t)|,\ c_3|Y_2^{(\mathrm{ICA1})}(f,t)|\big]; \qquad (13)$$

otherwise $m_1(f,t) = 0$. Here, $\max[\cdot]$ represents the function of picking up the maximum value among the arguments, and $c_1, \cdots, c_3$ are the weights for enhancing the contribution of each SIMO component to the masking decision process. For example, $[c_1, c_2, c_3] = [0, 0, 1]$ yields the simple combination of conventional ICA and conventional binary mask [8]. Otherwise, if we set $[c_1, c_2, c_3] = [1, 0, 0]$, we can utilize better (acoustically reasonable) SIMO information regarding each source as described in Sect. 3.1. 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

$$\hat{Y}_2(f,t) = m_2(f,t) Y_2^{(\mathrm{ICA1})}(f,t), \qquad (14)$$

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

$$Y_2^{(\mathrm{ICA1})}(f,t) > \max\big[c_1|Y_1^{(\mathrm{ICA2})}(f,t)|,\ c_2|Y_2^{(\mathrm{ICA2})}(f,t)|,\ c_3|Y_1^{(\mathrm{ICA1})}(f,t)|\big]; \qquad (15)$$

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 that the permutation matrices are given as

$$\boldsymbol{P}_l = [\delta_{in(k,l)}]_{ki}, \qquad (16)$$

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

$$n(k,l) = \begin{cases} k+l-1 & (k+l-1 \le L) \\ k+l-1-L & (k+l-1 > L) \end{cases}. \qquad (17)$$

In this case, (8) yields

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \big[A_{kn(k,l)}(f) S_{n(k,l)}(f,t) + E_{kn(k,l)}(f,t)\big]_{k1}. \qquad (18)$$

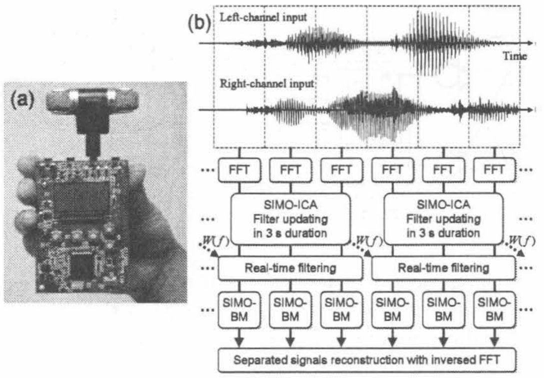

Fig. 5. (a) Overview of pocket-size real-time BSS module, where proposed two-stage BSS algorithm works on TEXAS INSTRUMENTS TMS320C6713 DSP. (b) Signal flow in real-time implementation of proposed method.

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

$$\hat{Y}_1(f,t) = m_1(f,t) Y_1^{(\mathrm{ICA1})}(f,t), \qquad (19)$$

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

$$Y_1^{(\mathrm{ICA1})}(f,t) > \max\big[c_1|Y_2^{(\mathrm{ICA}L)}(f,t)|,\ c_2|Y_3^{(\mathrm{ICA}L-1)}(f,t)|,\ c_3|Y_4^{(\mathrm{ICA}L-2)}(f,t)|, \cdots, c_{L-1}|Y_L^{(\mathrm{ICA2})}(f,t)|, \cdots, c_{LL-1}|Y_L^{(\mathrm{ICA1})}(f,t)|\big]; \qquad (20)$$

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

3.4. Real-Time Implementation

We have already built a pocket-size real-time BSS module in cooperate with Kobe Steel, Ltd., where the proposed two-stage BSS algorithm can work on a general-purpose DSP as shown in Fig. 5(a). Figure 5(b) 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-s-duration data for estimating the separation matrix, that is applied to the next (*not current*) 3 s 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 s 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. Owing to the advantage, the problem of performance decrease is prevented, especially in the case of rapid change of the mixing condition, e.g., the target sources are moving. This fact will appear via experiments in the next section.

4. SOUND SEPARATION EXPERIMENTS

4.1. Real-Time Separation Experiment for Moving Sound Source

In this section, a real-recording-based BSS experiment is 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 Fig. 6, where two loudspeakers and the real-time BSS system (Fig. 5(a)) with a directional microphone (SONY stereo microphone ECM-DS70P) are set. The reverberation time in this room is 200 ms, and the levels of background noise and each of the sound sources measured at the array origin were 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 $\theta_1 = -40°$, and move source 2 as follows:

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

Two kinds of sentences, spoken by two male and two female speakers, are used as the original speech samples. Using these sentences, we obtain 12 combinations with respect to speakers and source directions. The sampling frequency is 8 kHz. The DFT size of $W(f)$ is 1024. We used a null-beamformer-based initial value [3] which is steered to $(-60°, 60°)$. The step-size parameter was optimized for each method to obtain the best separation performance.

We compare four methods as follows: (A) the conventional binary-mask-based BSS, (B) the conventional ICA-based BSS, where the scaling ambiguity can be properly solved by [1]. (C) the simple combination of the conventional ICA and binary masking [8], 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 under this background noise condition.

Figure 7 shows the averaged segmental NRR for 12 speaker combinations, which was calculated along the time axis at every 0.5 s period. The first 3 s duration is spent on the initial filter learning of ICA in the methods (B), (C) and (D), and thus the valid ICA-based separation filter is absent here. Therefore, at 0–3 s, we simply applied binary masking in the methods (C) and (D). The successive duration (at 3–32 s) shows the separation results for *open* data sample, which is to be evaluated in this experiment. From Fig. 7, we can confirm that the proposed two-stage BSS (D) outperforms other methods at almost all the time during 3–32 s. It is worth noting that the conventional ICA shows heavy deteriorations especially in the 2nd source's moving periods, i.e., around 10 s and 21 s, but the proposed method can mitigate the degradations. These results can conclude the proposed method to be beneficial to many real-time BSS applications in the real world.

4.2. For MLSP 2007 Data Analysis Competition

We tried MLSP 2007 data analysis competition to attack the convolutive blind source separation. We entered the competition with non-linear section and achieved victory. In this section, we introduce the parameters employed in the competition as follows: The DFT size of ICA filter $W(f)$ and binary masking is 1024. The initial value of SIMO-ICA is identity matrix. We use 0.05 as the

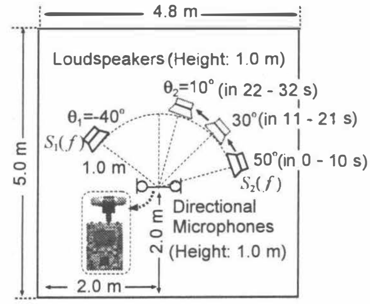

Fig. 6. Layout of reverberant room used in real-recording-based experiment. Reverberation time is 200 ms.

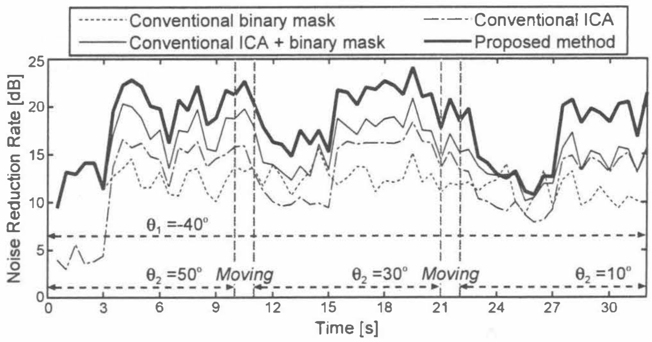

Fig. 7. Results of segmental NRR calculated along time axis at every 0.5 s period, where real recording data and real-time BSS are used. Each line is average for 12 speaker combinations.

step size parameter α, and the number of iterations is 1100. Also, we adopted the SIMO-BM's coefficients $[c_1, c_2, c_3]$ as, $[1, 0, 0.4]$ for speech, and $[1, 0, 0]$ for music.

5. CONCLUSION

We have proposed BSS framework in which SIMO-ICA and SIMO-BM are efficiently combined. The 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 the loss of their spatial qualities. Thus, after the SIMO-ICA, we can introduce the SIMO-model-based binary masking and succeed in removing the residual interference components. In order to evaluate its effectiveness, on-line separation experiment using DSP module was carried out for real recording data under a 200-ms-reverberant condition. The experimental results revealed that the SNR can be considerably improved by the proposed two-stage BSS algorithm compared with the conventional methods. Also, we introduce the detailed parameters used for "MLSP 2007 Data Analysis Competition" for non-linear system.

6. ACKNOWLEDGMENT

This work was partly supported by MEXT e-Society leading project, and NEDO project for strategic development of advanced robotics elemental technologies in Japan. Also, we are grateful that this paper is the special invitation to MLSP 2007 because of the winner of the data analysis competition.

A. APPENDIX

A.1. Derivation of (11)

Here, Kullback-Leibler divergence between the joint probability density function (PDF) of $Y(f, t)$ and the product of marginal

PDFs of $Y_l(f,t)$ is defined by $\mathrm{KLD}(\boldsymbol{Y}(f,t))$. The gradient of $\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))$ with respect to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ should be added to the iterative learning rule of the separation filter in the l-th ICA ($l=1,\cdots,L-1$). We obtain the partial differentiation (standard gradient) of $\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))$ with respect to $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ ($l=1,\ldots,L-1$) as

$$\frac{\partial\,\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial \boldsymbol{W}_{(\mathrm{ICA}l)}(f)} = \left[\frac{\partial\,\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial W_{ij}^{(\mathrm{ICA}L)}(f)}\cdot\frac{\partial W_{ij}^{(\mathrm{ICA}L)}(f)}{\partial W_{ij}^{(\mathrm{ICA}l)}(f)}\right]_{ij} = \left[\frac{\partial\,\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial W_{ij}^{(\mathrm{ICA}L)}(f)}\cdot(-1)\right]_{ij}, \tag{21}$$

where $W_{ij}^{(\mathrm{ICA}L)}(f)$ is the element of $\boldsymbol{W}_{(\mathrm{ICA}L)}(f)$. By replacing $\partial\,\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))\,/\partial\boldsymbol{W}_{(\mathrm{ICA}L)}(f)$ with its natural gradient, we modify (21) as

$$-\frac{\partial\,\mathrm{KLD}(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t))}{\partial \boldsymbol{W}_{(\mathrm{ICA}L)}(f)}\cdot\boldsymbol{W}^{\mathrm{H}}_{(\mathrm{ICA}L)}(f)\boldsymbol{W}_{(\mathrm{ICA}L)}(f) = \left\{\boldsymbol{I}-\left\langle\boldsymbol{\Phi}\left(\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t)\right)\cdot\boldsymbol{Y}^{\mathrm{H}}_{(\mathrm{ICA}L)}(f,t)\right\rangle_t\right\}\cdot\boldsymbol{W}_{(\mathrm{ICA}L)}(f). \tag{22}$$

By inserting (7) and the relation $\boldsymbol{W}_{(\mathrm{ICA}L)}(f)=\boldsymbol{I}-\sum_{l=1}^{L-1}\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ into (22), we obtain

$$\left\{\left(\boldsymbol{I}-\left\langle\boldsymbol{\Phi}\left(\boldsymbol{X}(f,t)-\sum_{l=1}^{L-1}\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)\right)\cdot\left(\boldsymbol{X}(f,t)-\sum_{l=1}^{L-1}\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)\right)^{\mathrm{H}}\right\rangle_t\right\}\cdot\left(\boldsymbol{I}-\sum_{l=1}^{L-1}\boldsymbol{W}_{(\mathrm{ICA}l)}(f)\right). \tag{23}$$

In order to deal with non-i.i.d. signals, we apply the nonholonomic constraint to (23). The natural gradient with the nonholonomic constraint is given as

$$-\left\{\text{off-diag}\left\langle\boldsymbol{\Phi}\left(\boldsymbol{X}(f,t)-\sum_{l=1}^{L-1}\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)\right)\cdot\left(\boldsymbol{X}(f,t)-\sum_{l=1}^{L-1}\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)\right)^{\mathrm{H}}\right\rangle_t\right\}\cdot\left(\boldsymbol{I}-\sum_{l=1}^{L-1}\boldsymbol{W}_{(\mathrm{ICA}l)}(f)\right). \tag{24}$$

Thus, the new iterative algorithm of the l-th ICA part ($l=1,\cdots,L-1$) in SIMO-ICA is given by adding (24) into the existing ICA equation, and we obtain (11).

A.2. Difference between SIMO-ICA and Projection-Back Method

In the projection-back (PB) method, the following operation is performed after (2):

$$Y_l^{(k)}(f,t) = \left\{\boldsymbol{W}(f)^{-1}[\overbrace{0,\cdots,0}^{l-1},Y_l(f,t),\overbrace{0,\cdots,0}^{L-l}]^{\mathrm{T}}\right\}_k = (\det\boldsymbol{W}(f))^{-1}\Delta_{lk}\cdot Y_l(f,t), \tag{25}$$

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

This method is simpler than SIMO-ICA, but its inversion often fails and yields harmful results because the invertibility of every $\boldsymbol{W}(f)$ cannot be guaranteed [9]. 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 $(\det\boldsymbol{W}(f))^{-1}\Delta_{lk}$ is applied to not only the target signal component but also the interference component in $Y_l(f,t)$. For example, similar to Sect. 3.1, (25) leads to

$$Y_l^{(k)}(f,t) = (\det\boldsymbol{W}(f))^{-1}\Delta_{lk}\cdot(B_l(f)S_l(f,t)+E_l(f,t)) = (\det\boldsymbol{W}(f))^{-1}\Delta_{lk}\cdot B_l(f)S_l(f,t) + (\det\boldsymbol{W}(f))^{-1}\Delta_{lk}\cdot E_l(f,t), \tag{26}$$

where we can assume that $|(\det\boldsymbol{W}(f))^{-1}\Delta_{ll}|$ is the largest value among $|(\det\boldsymbol{W}(f))^{-1}\Delta_{lk}|$ ($k=1,\cdots,K$) for the l-th 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 (26) 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 (26) 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, i.e., 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 $Y_l^{(k)}(f,t)$.

B. REFERENCES

[1] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA and BSS (ICA99)*, pp. 365–371, 1999.
[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Processing*, vol. 8, pp. 320–327, 2000.
[3] H. Saruwatari, *et al.*, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1135–1146, 2003.
[4] Y. Mori, *et al.*, "Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 34970, 2006.
[5] T. Takatani, *et al.*, "High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," *Proc. IWAENC2003*, pp. 251–254, 2003.
[6] MLSP 2007 Data Analysis Competition, http://mlsp2007.conwiz.dk/index.php?id=43 .
[7] R. Lyon, "A computational model of binaural localization and separation," *Proc. ICASSP83*, pp. 1148–1151, 1983.
[8] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," *Proc. ICA2004*, pp.832–839, 2004.
[9] T. Nishikawa *et al.*, "Stable learning algorithm for blind separation of temporally correlated acoustic signals combining multistage ICA and linear prediction", *IEICE Trans. Fundamentals*, vol.E86-A, no.8, pp.2028–2036, 2003.