2005 IEEE/RSJ International Conference on Intelligent Robots and Systems

Two-Stage Blind Source Separation Based on ICA and Binary Masking for Real-Time Robot Audition System*

Hiroshi Saruwatari, Yoshimitsu Mori,
Tomoya Takatani, Satoshi Ukai, Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
sawatari@is.naist.jp

Takashi Hiekata, Takashi Morita
Kobe Steel,Ltd.
Kobe, 651-2271, Japan
t-hiekata@kobelco.jp

***Abstract*— We newly propose a real-time two-stage blind source separation (BSS) for binaural mixed signals observed at the ears of humanoid robot, in which a Single-Input Multiple-Output (SIMO)-model-based independent component analysis (ICA) and binary mask processing 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 as they are at the microphones. Thus, the separated signals of SIMO-model-based ICA can maintain the spatial qualities of each sound source, and this yields that binary mask processing can be applied to efficiently remove the residual interference components after SIMO-model-based ICA. The experimental results obtained with a human-like head reveal that the separation performance can be considerably improved by using the proposed method in comparison to the conventional ICA-based and binary-mask-based BSS methods.**

***Index Terms*— Robot audition, blind source separation, ICA, binary masking.**

I. 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. This technique is based on *unsupervised* filtering 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 paid to the BSS technique in many fields of signal processing. One promising example in acoustic signal processing is a humanoid robot auditory system [1], i.e., separation of binaural mixed signals observed at the ears of the robot, which constructs an indispensable basis for intelligent robot technology [2], [3].

In recent works of BSS based on independent component analysis (ICA) [4], various methods have been proposed for acoustic-sound separation [5], [6], [7], [8]. In this paper, we mainly address the BSS problem under highly reverberant conditions which often arise in many practical audio applications. The separation performance of the conventional ICA is far from being sufficient in such a 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 supervised signal enhancement technique, e.g., spectral subtraction [9]. However, in the conventional ICA framework, each of the separated outputs is a *monaural* signal, and this leads to the drawback that many kinds of superior *multichannel* techniques cannot be applied.

To solve the problem, we propose a novel two-stage BSS algorithm which is applicable to an augmentation of the humanoid robot audition. In this approach, the BSS problem is resolved into two stages: (a) a Single-Input Multiple-Output (SIMO)-model-based ICA [10], [11] and (b) binary mask processing [12], [13], [14] 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 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-model-based ICA can maintain the spatial qualities of each sound source. After the SIMO-model-based ICA, the residual components of the interference, which are often staying in the output of SIMO-model-based ICA as well as the conventional ICA, can be efficiently removed by the following binary mask processing. The experimental results reveal that the proposed method can successfully achieve the BSS for speech mixtures even under a realistic reverberant condition.

II. Mixing Process and Conventional BSS

A. Mixing Process

In this study, the number of microphones is K and the number of multiple sound sources is L. The directions of arrival of multiple L sound sources are designated as θ_l ($l = 1, \cdots, L$) (see Fig. 1), where we deal with the case of $K = L$.

In the frequency domain, the observed signals in which multiple source signals are mixed linearly are given by

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

*This work is partially supported by CREST "Advanced Media Technology for Everyday Living" of JST in Japan.

0-7803-8912-3/05/$20.00 ©2005 IEEE

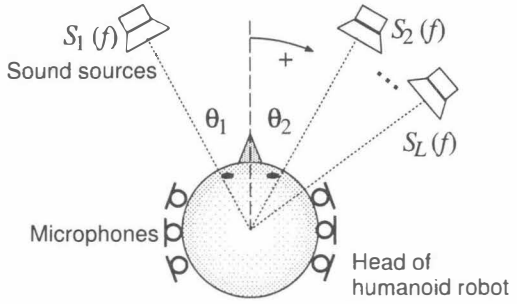

Fig. 1. Configuration of a multi-microphone system in robot head and source signals.

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 assumed to be complex-valued because we introduce a model to deal with the arrival lags among each of the elements of the microphone array and room reverberations.

B. 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 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

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

We perform this procedure with respect to all frequency bins. The optimal $\boldsymbol{W}(f)$ is obtained by, for example, the following iterative updating equation:

$$\boldsymbol{W}^{[i+1]}(f) = \eta\left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\right\rangle_t\right]\boldsymbol{W}^{[i]}(f) + \boldsymbol{W}^{[i]}(f), \tag{3}$$

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, and η is the step-size parameter. In our research, we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as [15]:

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv \left[e^{j\cdot\arg(Y_1(f,t))}, \cdots, e^{j\cdot\arg(Y_L(f,t))}\right]^{\mathrm{T}}, \tag{4}$$

where $\arg[\cdot]$ represents an operation to take the argument of the complex value. After the iterations, the permutation problem, i.e., indeterminacy in ordering sources, can be solved by, e.g., [8], [16].

C. Conventional Binary-Mask-Based BSS

Binary mask processing [12], [13], [14] is one of the alternative approach which is aimed to solve the BSS problem, but is not based on ICA. This method is basically introducing the auditory masking effect which tells that the stronger signal masks the weaker one. We estimate a binary mask by comparing the amplitudes of the observed (binaural) signals, and pick up the target sound component which arrives at the *better ear* (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 and $L = 2$, the l-th 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 few computational complexities, and this property is well applicable to real-time processing. The method, however, assumes the sparseness in the spectral components of the sound sources, which is often introduced in Computational Auditory Scene Analysis (CASA)-based BSS. That is, in binary mask processing, it should be assumed that there are no overlaps in time-frequency components of the sources, but the assumption does not hold in an usual application to the acoustic sound separation (indeed, e.g., a mixture of speech and common broadband stationary noise has many overlaps).

III. Proposed Two-Stage BSS Algorithm

A. Motivation and Strategy

In the previous research, SIMO-model-based ICA was proposed by, e.g., Takatani et al. [10], [11], and they showed that SIMO-model-based ICA can separate the mixed signals into SIMO-model-based signals at the microphone points. This finding has motivated us to combine the SIMO-model-based ICA and binary mask processing. That is, the binary mask technique can be applied to the SIMO components of each source obtained from SIMO-model-based ICA. The configuration of the proposed method is depicted in Fig. 2(a). Binary mask processing which follows SIMO-model-based ICA can remove the residual component of the interference effectively without adding huge computational complexities.

It is worth mentioning that the novelty of this strategy mainly lies in the two-stage idea of the unique combination of SIMO-mode-based ICA and the SIMO-model-based binary mask. To illustrate the novelty of the proposed method, we hereinafter compare the proposed combination with a simple two-stage combination of a conventional monaural-output ICA and binary mask processing (see Fig. 2(b)) [17].

In general, the 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 distortion filter and $E_l(f,t)$ is a residual separation error which is mainly

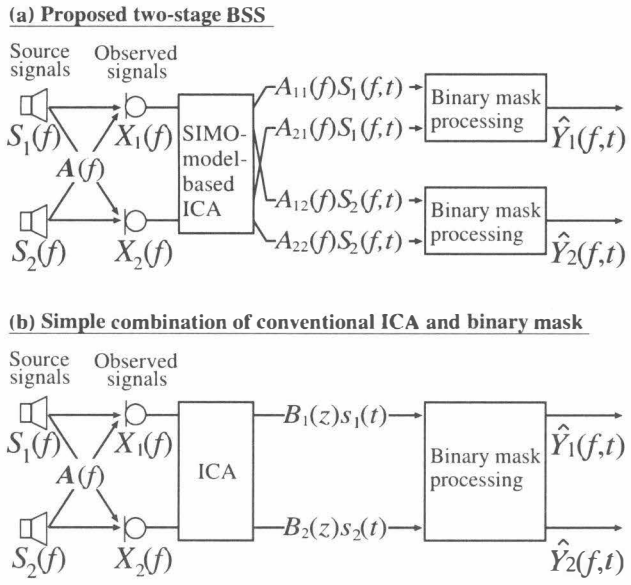

Fig. 2. Input and output relations in (a) proposed two-stage BSS and (b) simple combination of conventional ICA and binary mask processing. This corresponds to the case of $K = L = 2$.

caused by an insufficient convergence in ICA. The residual error $E_l(f,t)$ should be removed by binary mask processing in the next post-processing stage. However, the combination is very problematic and cannot function well because of the existence of the spectral overlaps in the time-frequency domain. For instance, if all sources have nonzero spectral components (i.e., sparseness assumption does not hold) in the specific frequency subband and these are comparable, the decision in binary mask processing for $Y_1(f,t)$ and $Y_2(f,t)$ is vague and the output results in a ravaged signal. Thus the simple combination of the conventional ICA and binary mask processing is not valid for solving the BSS problem.

On the other hand, our proposed combination contains the special SIMO-model-based ICA in the first stage. The aim of the SIMO-model-based ICA is to supply the specific SIMO signals with respect to each of sources, $A_{kl}(f)S_l(f,t)$, up to the possible delay of the filters and the residual error. Needless to say, the obtained SIMO components is well applicable to binary mask processing because of the spatial properties that the separated SIMO component at the specific microphone closer to the target sound still maintains the large gain. Thus, after having the SIMO components, we can introduce the binary mask for the efficient reduction of the remaining error in ICA, even when the sparseness assumption does not hold.

In summary, the novelty of the proposed two-stage idea is due to the introduction of SIMO-model-based framework into both separation and post processes, and this offers a realization of the robust BSS. The detailed process of using the proposed algorithm is as follows.

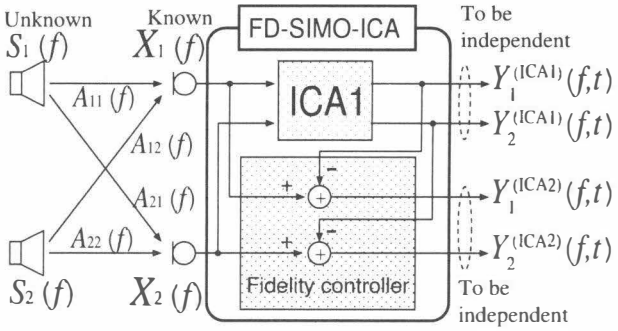

Fig. 3. Input and output relations in the proposed FD-SIMO-ICA, where $K = L = 2$.

B. Algorithm

Time-domain *SIMO-ICA* [10] has recently been proposed by one of the authors as a means of obtaining SIMO-model-based signals directly in the ICA updating. In this paper, we extend the 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 sources. The 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. 3). The separated signals of the l-th ICA $(l = 1, \cdots L-1)$ in FD-SIMO-ICA are defined by

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

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,

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

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 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 $(-\sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t))$ in the right-hand side into the left-hand side, we can show that (7) means 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, as

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

where $\boldsymbol{P}_l$ $(l = 1, \cdots, L)$ are exclusively-selected permutation matrices which satisfy $\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}$. Regarding a proof of this, see [10] with an appropriate modification into the frequency-domain representation. Obviously the solutions given by (8) 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

$$\begin{aligned} &\left[Y_1^{(\mathrm{ICA}1)}(f,t),\ Y_2^{(\mathrm{ICA}1)}(f,t)\right]^{\mathrm{T}} \\ &\quad = \left[A_{11}(f)S_1(f,t),\ A_{22}(f)S_2(f,t)\right]^{\mathrm{T}}, \qquad (9) \\ &\left[Y_1^{(\mathrm{ICA}2)}(f,t),\ Y_2^{(\mathrm{ICA}2)}(f,t)\right]^{\mathrm{T}} \\ &\quad = \left[A_{12}(f)S_2(f,t),\ A_{21}(f)S_1(f,t)\right]^{\mathrm{T}}, \qquad (10) \end{aligned}$$

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

In order to obtain (8), 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 [5] 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

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

where α is the step-size parameter, and we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as [15]:

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \ \equiv\ &\left[\tanh(|Y_1(f,t)|)e^{j\cdot\arg(Y_1(f,t))}, \cdots,\right. \\ &\left.\tanh(|Y_L(f,t)|)e^{j\cdot\arg(Y_L(f,t))}\right]^{\mathrm{T}}. \qquad (12) \end{aligned}$$

Also, the initial values of $\boldsymbol{W}_{(\mathrm{ICA}l)}(f)$ for all l should be different.

After FD-SIMO-ICA, binary masking processing is applied. For example in the case of (9) and (10), the resultant output signal corresponding to the source 1 is obtained as follows:

$$\hat{Y}_1(f,t) \ = \ m_1(f,t)Y_1^{(\mathrm{ICA}1)}(f,t), \qquad (13)$$

where $m_1(f,t)$ is the binary mask operation which is defined as $m_1(f,t) = 1$ if $Y_1^{(\mathrm{ICA}1)}(f,t)$ is greater than $Y_2^{(\mathrm{ICA}2)}(f,t)$; otherwise $m_1(f,t) = 0$. Also, the resultant output signal corresponding to the source 2 is given by

$$\hat{Y}_2(f,t) \ = \ m_2(f,t)Y_2^{(\mathrm{ICA}1)}(f,t), \qquad (14)$$

where $m_2(f,t)$ is the binary mask operation which is defined as $m_2(f,t) = 1$ if $Y_2^{(\mathrm{ICA}1)}(f,t)$ is greater than $Y_1^{(\mathrm{ICA}2)}(f,t)$; otherwise $m_2(f,t) = 0$. The extension to the general case of $L = K > 2$ can be easily implemented in the same manner.

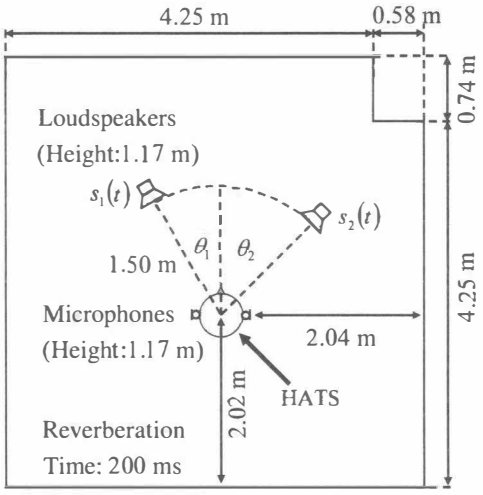

Fig. 4. Layout of reverberant room used in experiments.

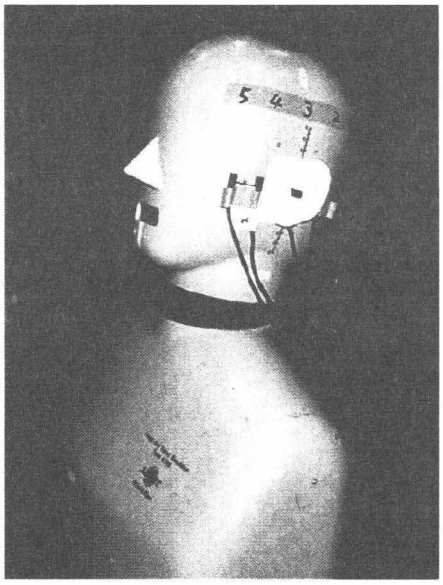

Fig. 5. Head and torso simulator used in experiment.

IV. Experiment under Real Acoustic Environment

A. Conditions for Experiments

We carried out binaural-sound-separation experiments using acoustical source signals recorded in the experimental

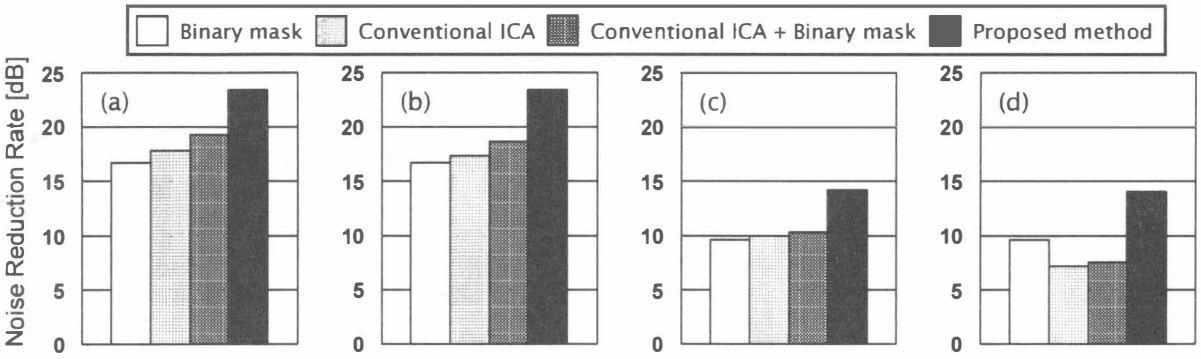

Fig. 6. Results of NRR in speech-speech mixing: (a) source DOAs are $(-60°, 60°)$ and initial value DOAs are $(-30°, 30°)$, (b) source DOAs are $(-60°, 60°)$ and initial value DOAs are $(-15°, 15°)$, (c) source DOAs are $(-60°, 0°)$ and initial value DOAs are $(-30°, 30°)$, and (d) source DOAs are $(-60°, 0°)$ and initial value DOAs are $(-15°, 15°)$.

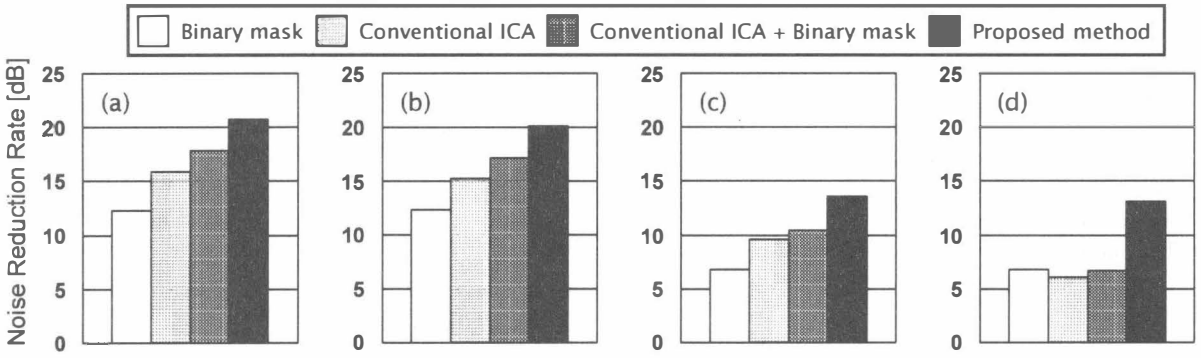

Fig. 7. Results of NRR in speech-noise mixing: (a) source DOAs are $(-60°, 60°)$ and initial value DOAs are $(-30°, 30°)$, (b) source DOAs are $(-60°, 60°)$ and initial value DOAs are $(-15°, 15°)$, (c) source DOAs are $(-60°, 0°)$ and initial value DOAs are $(-30°, 30°)$, and (d) source DOAs are $(-60°, 0°)$ and initial value DOAs are $(-15°, 15°)$.

room illustrated in Fig. 4, where two sources and two microphones are set. The reverberation time in this room is 200 ms. A head and torso simulator (HATS; see Fig. 5) by Brüel & Kjær is used as the recording apparatus, which simulates a robot auditory system. Two acoustic signals are assumed to arrive from different directions, θ_1 and θ_2, where we prepare two kinds of source direction patterns as follows; $(\theta_1, \theta_2) = (-60°, 60°)$, or $(-60°, 0°)$. We used the speech signals spoken by two male and two female speakers, and human-speech-colored stationary noise as the source samples. The sampling frequency is 8 kHz and the length of each speech sample is limited to 3 seconds. The DFT size of $\boldsymbol{W}(f)$ in each method is 1024. We use two types of initial values which are given by the HRTF-based null beamformers [8] whose directions of sources are $(-15°, 15°)$, or $(-30°, 30°)$.

B. Experimental Results

We compare four methods as follows: (A) the conventional binary-mask-based BSS given by (5), (B) the conventional ICA-based BSS given by (2), (C) simple combination of the conventional ICA and binary mask processing, and (D) the proposed two-stage BSS method. Here we did not use any a priori information on the true DOA of sources, room transfer functions, positions of microphones, and acoustic characteristics of HATS (the robot head) in the separation procedure of each method. These information can not be used, especially for the robot which moves around the user.

Noise reduction rate (NRR) [8], defined as the output signal-to-noise ratio (SNR) in dB minus the input SNR in dB, is used as the objective indication of separation performance. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise.

Figure 6 show the results of NRR for speech-speech mixing under different speaker allocations and initial value conditions. These scores are the averages of 12 speaker combinations. Also, Fig. 7 shows the results of NRR for the mixing of speech and stationary noise; this corresponds to the case in that the spectral sparseness assumption does not hold. From the results, we can confirm that the proposed two-stage BSS can consistently and significantly improve the separation performance regardless the speaker directions, noise and initial value conditions. It is also worth mentioning that the proposed method can provide the improvements even under unsparse-source mixing conditions, unlike the conventional binary mask processing. This fact is a promising evidence

on the feasibility of the proposed combination technique of SIMO-model-based ICA and binary mask processing.

V. Real-Time Implementation

We have already built a real-time two-stage BSS demo system running on a very light palmtop PC (SONY VAIO type-U with Pentium-M 1.1 GHz processor, 550 g weight). Figure 8 shows a configuration of a real-time implementation for the proposed two-stage BSS. Signal processing in this implementation is performed as the following instructions.

1) Inputted binaural signals are converted to time-frequency series by using frame-by-frame fast Fourier transform (FFT).
2) SIMO-ICA is conducted using a current 3 s-duration data for estimating the separation matrix which 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 huge computational complexities and cannot provide the optimal separation filter for the current 3 s data.
3) Binary mask processing 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 binary mask processing are converted to the resultant time-domain waveforms by using an inverse FFT.

Although the separation filter update in SIMO-ICA part is not real-time processing but includes a 3 s latency, the whole two-stage system still seems real-time because the binary masking can work in the current segment with no delay. Generally the latency in the conventional ICAs is problematic and reduces the applicability of the 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 mask processing.

VI. Conclusion

We proposed a new BSS framework in which the SIMO-model-based ICA and binary mask processing are efficiently combined. In order to evaluate its effectiveness, a separation experiment was carried out under a reverberant condition. The experimental results revealed that the NRR can be considerably improved by using the proposed two-stage BSS algorithm. In addition, we could find the fact that the proposed method outperforms the combination of the conventional ICA and binary mask processing as well as the simple ICA and binary mask processing.

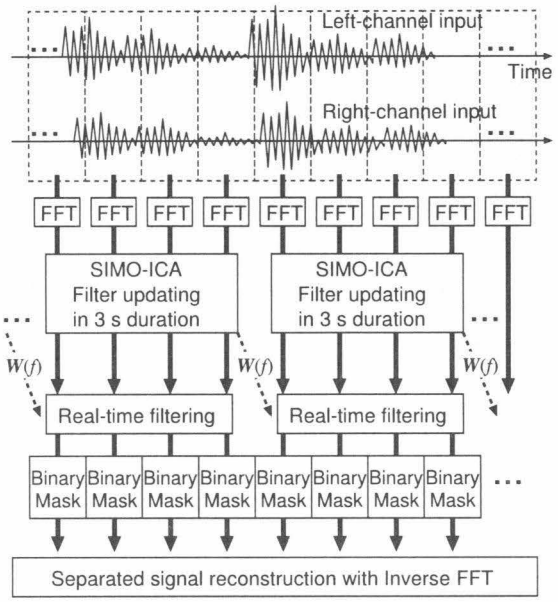

Fig. 8. Signal flow in real-time implementation of proposed method.

References

[1] K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: robust sound source localization and extraction," *Proc. IROS-2003*, pp.1147–1152, 2003.

[2] R. Nishimura, T. Uchida, A. Lee, H. Saruwatari, K. Shikano, and Y. Matsumoto, "ASKA: Receptionist robot with speech dialogue system," *Proc. IROS-2002*, pp.1314–1317, 2002.

[3] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol.18, pp.533–564, 2004.

[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol.36, pp.287–314, 1994.

[5] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Proc. NOLTA98*, vol.3, pp.923–926, 1998.

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

[7] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech & Audio Processing*, vol.8, pp.320–327, 2000.

[8] 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, pp.1135–1146, 2003.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech & Signal Process.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[10] 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," *Proc. IWAENC2003*, pp.251–254, 2003.

[11] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis," *IEICE Trans. Fundamentals*, vol.E87-A, no.8, pp.2063–2072, 2004.

[12] R. Lyon, "A computational model of binaural localization and separation," *Proc. ICASSP83*, pp.1148–1151, 1983.

[13] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Proc. IJCNN01*, pp.2861–2866, 2001.

[14] 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.

[15] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency domain blind source separation," *IEICE Trans. Fundamentals*, vol.E86-A, no.3, pp.590–596, 2003.

[16] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Proc. Int. Sympo. on ICA and BSS*, pp.505–510, 2003.

[17] M. Aoki and K. Furuya, "Using spatial information for speech enhancement," *Technical Report of IEICE*, vol.EA2002-11, pp.23–30, 2002 (in Japanese).