Blind Separation of Convolutive Speech Mixtures Using SIMO-Model-Based ICA and Binary Mask Processing

Yoshimitsu Mori*, Tomoya Takatani*, Satoshi Ukai*, Hiroshi Saruwatari*, Kiyohiro Shikano*, Takashi Hiekkata†, and Takashi Morita†

†Graduate School of Information Science
Nara Institute of Science and Technology, Nara 630-0192 Japan
Tel: +81-743-72-5287, Fax: +81-743-72-5289
E-mail: yoshim-m@is.naist.jp

‡Kobe Steel, Ltd., 5-5 Takatsukadai-1, Nishi-ku, Kobe, 651-2271 Japan
Tel: +81-78-992-5654, Fax: +81-78-992-5650
E-mail: t-hiekkata@kobelco.jp

Abstract — A new two-stage blind source separation (BSS) for convolutive mixtures of speech is proposed, in which a Single-Input Multiple-Output (SIMO)-model-based 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. Owing to the attractive property, binary mask processing 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 using the proposed method in comparison to the conventional source separation methods.

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. In recent works of BSS based on independent component analysis (ICA) [1], various methods have been proposed for acoustic-sound separation [2], [3], [4]. 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 [5], [6]. 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. In this approach, the BSS problem is resolved into two stages: (a) a Single-Input Multiple-Output (SIMO)-model-based ICA [2], [7] and (b) binary mask processing [8], [9], [10] 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-base 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 \( \theta_l \) (\( l = 1, \cdots, L \)) (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

\[
\mathbf{X}(f) = \mathbf{A}(f)\mathbf{S}(f),
\]

where \( \mathbf{X}(f) = [X_1(f), \cdots, X_K(f)]^T \) is the observed signal vector, and \( \mathbf{S}(f) = [S_1(f), \cdots, S_L(f)]^T \) is the source signal.
vector. Also, \( \mathbf{A}(f) = [\mathbf{a}_{kl}(f)]_{kl} \) 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 \( \mathbf{A}(f) \) 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 \( \mathbf{X}(f, t) = [X_1(f, t), \ldots, X_K(f, t)]^T \). 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 \( \mathbf{Y}(f, t) = [Y_1(f, t), \ldots, Y_L(f, t)]^T \) becomes mutually independent; this procedure can be given as

\[
\mathbf{Y}(f, t) = \mathbf{W}(f) \mathbf{X}(f, t).
\]  

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

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

where \( I \) is the identity matrix, \( \langle \cdot \rangle_i \) denotes the time-averaging operator, \( [i] \) is used to express the value of the \( i \)-th step in the iterations, and \( \eta \) is the step-size parameter. In our research, we define the nonlinear vector function \( \Phi(\cdot) \) as [11]:

\[
\Phi(\mathbf{Y}(f, t)) = \left[ e^{j\text{arg}(Y_1(f, t))}, \ldots, e^{j\text{arg}(Y_L(f, t))} \right]^T,
\]  

where \( \text{arg}[\cdot] \) 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 [12].

**C. Conventional Binary-Mask-Based BSS**

Binary mask processing [8], [9], [10] 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

\[
\hat{Y}_l(f, t) = m_l(f, t) X_l(f, t),
\]

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. However, the method assumes the sparseness in the spectral components of the sound sources. 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.
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. [7], [13], and they showed that SIMO-model-based ICA can separate the mixed signals into SIMO-model-based signals at 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-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 with the simple two-stage combination of a conventional monaural-output ICA and binary mask processing (see Fig. 2(b)) [14].

In general, the conventional ICAs can only supply the source signals $Y_l(f, t) = B_l(f)S_l(f, t) + E_l(f, t) \ (l = 1, \ldots, L)$, where $B_l(f)$ is an unknown arbitrary distortion 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 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. Some sources have a 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 are 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. In this paper, we can provide two types of the algorithms in terms of the different SIMO-model-based ICAs. The detailed process of using the proposed algorithm is as follows.

B. Proposed Method 1 Using FDICA with Projection Back

As the SIMO-model-based ICA, FDICA with projection back processing (FDICA-PB) [2] is introduced. In order to obtain the SIMO components, the separated signals $Y_l(f, t)$ in FDICA are projected back onto the microphones by using the inverse of $W(f)$. In this method, the following operation is performed.

$$Y_l^{(k)}(f, t) = \left\{ W(f)^{-1} T \right\}_{\{0, \ldots, 0, Y_l(f, t), 0, \ldots, 0\}^T}^T_{\{0, \ldots, 0, 0, \ldots, 0\}^T}$$

where $Y_l^{(k)}(f, t)$ represents the l-th resultant separated source signal which is projected back onto the k-th microphone, and $\{\}$ denotes the $k$-th element of the argument. The FDICA-PB has the advantages that it is very fast because the calculation of FDICA given by (3) and the projection-back processing given by (6) are simple. There exist, however, the disadvantages that the inversion of $W(f)$ often fails and yields harmful results because the invertibility of every $W(f)$ cannot be guaranteed [15].

After FDICA-PB, binary masking processing is applied, and the resultant output signal is obtained as follows:

$$\hat{Y}_l(f, t) = m_l(f, t)Y_l^{(l)}(f, t),$$

where $m_l(f, t)$ is the binary mask operation which is defined as $m_l(f, t) = 1$ if $Y_l^{(l)}(f, t) = \max_k(Y_l^{(k)}(f, t))$; otherwise $m_l(f, t) = 0$ (note that the algorithm is the extended version of the original binary mask algorithm for binaural signals to deal with the case of $L = K > 2$).

C. Proposed Method 2 Using Frequency-Domain SIMO-ICA

Time-domain SIMO-ICA [7] 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 ($l = 1, \ldots, L - 1$) in FD-SIMO-ICA are defined by

$$Y_{(ICA)}(f, t) = [Y_{(ICA)}^l(f, t)]_{l=1}^{L-1} = W_{(ICA)}(f)X(f, t),$$

(8)

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

Regarding the fidelity controller, we calculate the following signal vector $Y_{(ICAL)}(f, t)$, in which all the elements are to be mutually independent,

$$Y_{(ICAL)}(f, t) = X(f, t) - \sum_{l=1}^{L-1} Y_{(ICA)}(f, t).$$

(9)

Hereafter, we regard $Y_{(ICAL)}(f, t)$ as an output of a virtual “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 $Y_{(ICAL)}(f, t)$ is subject to $W_{(ICAL)}(f)$ ($l = 1, \ldots, L - 1$). By transposing the second term ($-\sum_{l=1}^{L-1} Y_{(ICAL)}(f, t)$) in the right-hand side into the left-hand side, we can show that (9) means a constraint to force the sum of all ICAs’ output vectors $\sum_{l=1}^{L} Y_{(ICAL)}(f, t)$ to be the sum of all SIMO components $\sum_{l=1}^{L} A_{kl}(f)S_{l}(f, t) = X(f, t)$.

If the independent sound sources are separated by (8), and simultaneously the signals obtained by (9) are mutually independent, then the output signals converge on unique solutions, up to the permutation, as

$$Y_{(ICAL)}(f, t) = \text{diag}([A(f)P_{l}^{T}])P_{l}S(f, t),$$

(10)

where $P_{l}$ ($l = 1, \ldots, L$) are exclusively-selected permutation matrices which satisfy $\sum_{l=1}^{L} P_{l} = [1]_{ij}$. Regarding a proof of this, see [7] with an appropriate modification into the frequency-domain representation. Obviously the solutions given by (10) 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

$$\begin{bmatrix} Y_{1}^{(ICA)}(f, t), Y_{2}^{(ICA)}(f, t) \\ Y_{1}^{(ICA)}(f, t), Y_{2}^{(ICA)}(f, t) \end{bmatrix}^{T} = \begin{bmatrix} A_{11}(f)S_{1}(f, t), A_{22}(f)S_{2}(f, t) \\ A_{12}(f)S_{2}(f, t), A_{21}(f)S_{1}(f, t) \end{bmatrix}^{T},$$

(11)

$$\begin{bmatrix} Y_{1}^{(ICA)}(f, t), Y_{2}^{(ICA)}(f, t) \\ Y_{1}^{(ICA)}(f, t), Y_{2}^{(ICA)}(f, t) \end{bmatrix}^{T} = \begin{bmatrix} A_{11}(f)S_{1}(f, t), A_{22}(f)S_{2}(f, t) \\ A_{12}(f)S_{2}(f, t), A_{21}(f)S_{1}(f, t) \end{bmatrix}^{T},$$

(12)

where $P_{1} = I$ and $P_{2} = [1]_{ij} - I$.

In order to obtain (10), the natural gradient of Kullback-Leibler divergence of (9) with respect to $W_{(ICA)}(f)$ should be added to the existing nonholonomic iterative learning rule [2] of the separation filter in the $l$-th ICA ($l = 1, \ldots, L - 1$). The new iterative algorithm of the $l$-th ICA part ($l = 1, \ldots, L - 1$) in FD-SIMO-ICA is given as

$$W_{(ICAL)}^{[j+1]}(f) = W_{(ICAL)}^{[j]}(f) - \alpha \left\{ \text{off-diag} \left( \Phi \left( Y_{(ICAL)}^{[j]}(f, t) \right) \right) \right\} \cdot W_{(ICAL)}^{[j]}(f)$$

$$- \left\{ \text{off-diag} \left( \Phi \left( X(f, t) - \sum_{l=1}^{L-1} Y_{(ICAL)}^{[j]}(f, t) \right) \right) \right\} \cdot \left( X(f, t) - \sum_{l=1}^{L-1} Y_{(ICAL)}^{[j]}(f, t)^{T} \right) \right\}_{t}$$

$$- \left\{ \Phi \left( X(f, t) - \sum_{l=1}^{L-1} Y_{(ICAL)}^{[j]}(f, t) \right) \right\} \cdot \left( I - \sum_{l=1}^{L-1} W_{(ICAL)}^{[j]}(f) \right),$$

(13)

where $\alpha$ is the step-size parameter, and we define the nonlinear vector function $\Phi(\cdot)$ as [11]:

$$\Phi( Y_{(f, t)} ) = \begin{bmatrix} \tanh( \| Y_{1}(f, t) \| e^{j \text{arg}(Y_{1}(f, t))} ), \ldots, \\ \tanh( \| Y_{L}(f, t) \| e^{j \text{arg}(Y_{L}(f, t))} ) \end{bmatrix}^{T}. $$

(14)

Also, the initial values of $W_{(ICAL)}(f)$ for all $l$ should be different.

After FDICA-PB, binary masking processing is applied. For example in the case of (11) and (12), the resultant output signal corresponding to the source 1 is obtained as follows:

$$Y_{1}(f, t) = m_{1}(f, t)Y_{1}^{(IA1)}(f, t),$$

(15)

where $m_{1}(f, t)$ is the binary mask operation which is defined as $m_{1}(f, t) = 1$ if $Y_{1}^{(ICA1)}(f, t)$ is greater than $Y_{2}^{(ICA2)}(f, t)$; otherwise $m_{1}(f, t) = 0$. Also, the resultant output signal corresponding to the source 2 is given by

$$Y_{2}(f, t) = m_{2}(f, t)Y_{2}^{(ICA1)}(f, t),$$

(16)

where $m_{2}(f, t)$ is the binary mask operation which is defined as $m_{2}(f, t) = 1$ if $Y_{2}^{(ICA1)}(f, t)$ is greater than $Y_{1}^{(ICA2)}(f, t)$; otherwise $m_{2}(f, t) = 0$. The extension to the general case of $L = K > 2$ can be easily implemented in the same manner.

IV. EXPERIMENT UNDER REVERBERANT CONDITION

4. Conditions for Experiments

We carried out binaural-sound-separation experiments using source signals which are convolved with impulse responses recorded with a head and torso simulator (HATS) (Brüel & Kjær) in the experimental room illustrated in Fig. 4. The reverberation time in this room is 200 ms. Two acoustic signals are assumed to arrive from different directions, $\theta_{1}$ and $\theta_{2}$, where we prepare three kinds of source direction patterns as follows; $(\theta_{1}, \theta_{2}) = (-60^\circ, 60^\circ)$, $(-60^\circ, 0^\circ)$ or $(0^\circ, 60^\circ)$. We used the speech signals spoken by two male and two female speakers as the source samples. Using these sentences, we obtain 12 combinations. The sampling frequency is 8 kHz and the length of each speech sample is limited to 3 seconds. The
DFT size of $W(f)$ in each method is 1024. We use three types of initial values which are given by the inverse of the HRTF matrices whose directions of sources are $(-15°, 15°)$, $(-30°, 30°)$, and $(-45°, 45°)$.

Noise reduction rate (NRR) [4], 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.

B. Experimental Results

We compare five 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, (D) the proposed method 1 based on FDICA-PB, and (E) the proposed method 2 based on FD-SIMO-ICA.

Figures 5–7 show the results of NRR for different speaker allocations and initial value conditions. These scores are the averages of 12 speaker combinations. From the results, we can confirm that the proposed two-stage BSS (proposed method 1 and/or proposed method 2) can consistently and significantly improve the separation performance regardless the speaker directions and initial value conditions. This fact is a promising evidence on the feasibility of the proposed combination technique of SIMO-ICA and binary mask processing. It should be mentioned that the resultant separation performances in the proposed method change remarkably in accordance with the choice of SIMO-model-based ICA algorithm, e.g., the proposed method 2 (using FD-SIMO-ICA) is preferable to the proposed method 1 (using FDICA-PB) when the sound sources are located at $-60°$ and $60°$, but vice versa in the other sound source allocations. The selection of the ICA algorithm in the proposed two-stage BSS still remains as an open problem for future study.

V. 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 FDICA and binary mask processing.

VI. Acknowledgment

This work was partly supported by CREST “Advanced Media Technology for Everyday Living” of JST in Japan.

REFERENCES

Fig. 5. Results of NRR for different initial values whose directions of sources are (a) $(-45^\circ, 45^\circ)$, (b) $(-30^\circ, 30^\circ)$, and (c) $(-15^\circ, 15^\circ)$. The source signals are two speech, and located at $(-60^\circ, 60^\circ)$.

Fig. 6. Results of NRR for different initial values whose directions of sources are (a) $(-45^\circ, 45^\circ)$, (b) $(-30^\circ, 30^\circ)$, and (c) $(-15^\circ, 15^\circ)$. The source signals are two speech, and located at $(-60^\circ, 0^\circ)$.

Fig. 7. Results of NRR for different initial values whose directions of sources are (a) $(-45^\circ, 45^\circ)$, (b) $(-30^\circ, 30^\circ)$, and (c) $(-15^\circ, 15^\circ)$. The source signals are two speech, and located at $(0^\circ, 60^\circ)$. 

---

- 160 -