BLIND SIGNAL SEPARATION OF AUDIO SIGNALS

Hiroshi Saruwatari

Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, JAPAN

ABSTRACT

This paper reviews a real-time blind source separation (BSS) method for convolutive mixtures of audio signals, 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 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, novel 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 develop a pocket-size real-time DSP module implementing the new BSS method, and report the experimental evaluation of the proposed method's superiority to the conventional BSS methods, regarding 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 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 BSS in many fields of signal processing such as speech enhancement.

In recent researches of BSS based on independent component analysis (ICA), various methods have been presented for acoustic-sound separation [1, 2, 3, 4]. This paper also addresses 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 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 *monaural* 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 novel two-stage BSS algorithm [5] which is applicable to an array of directional microphones. The main aim of this paper is to introduce and review our BSS method. This approach resolves the BSS problem into two stages: (a) a Single-Input Multiple-Output (SIMO)-model-based ICA proposed by the authors [6] and (b) SIMO-model-based binary masking for the SIMO signals obtained from the preceding SIMO-model-based ICA. 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 rich 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 SIMO-model-based binary masking.

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 mainly address the real-time implementation issue on the proposed BSS with our developed pocket-size DSP module, and evaluate the "real-time" separation performance for real recording of 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. 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:

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

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. Here, $\boldsymbol{N}(f)$ is the additive noise term which generally represents, for example, a background noise and/or a sensor noise. 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 $\boldsymbol{X}(f,t)$ $=[X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}$.

We thank KOBE STEEL LTD. for funding and collaboration.

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 $\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t)$. 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]:

$$\begin{aligned} \boldsymbol{W}^{[i+1]}(f) &= \eta\Big[\boldsymbol{I} - \big\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\big\rangle_t\Big]\boldsymbol{W}^{[i]}(f) \\ &\quad + \boldsymbol{W}^{[i]}(f), \end{aligned} \tag{2}$$

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, 8] 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 and $L = 2$, the l-th separated signal is given by

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

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

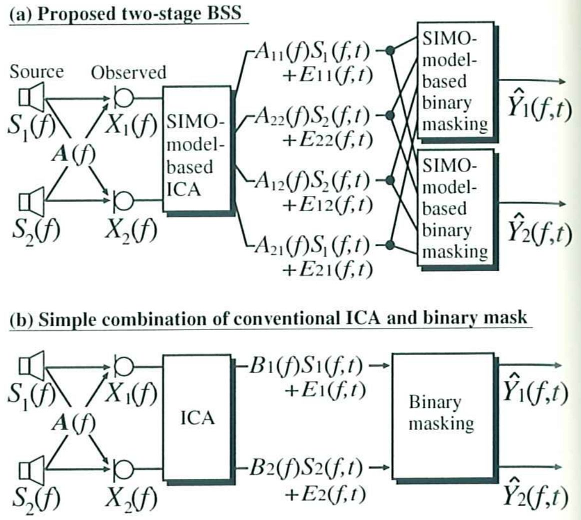
3.1. What is SIMO-model-based ICA?

In a previous study, SIMO-model-based ICA (SIMO-ICA) was proposed by some of the authors [6], 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:

$$\begin{aligned} \boldsymbol{X}(f) = {}& [A_{11}(f)S_1(f), \cdots, A_{K1}(f)S_1(f)]^{\mathrm{T}} \\ & \vdots \\ & + [A_{1L}(f)S_L(f), \cdots, A_{KL}(f)S_L(f)]^{\mathrm{T}}, \end{aligned} \tag{4}$$

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



(a) Proposed two-stage BSS
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
$\hat{Y}_1(f,t)$
$\hat{Y}_2(f,t)$
(b) Simple combination of conventional ICA and binary mask
$S_1(f)$
$X_1(f)$
$A(f)$
$S_2(f)$
$X_2(f)$
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
$\hat{Y}_1(f,t)$
$\hat{Y}_2(f,t)$

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 $K = L = 2$.

The aim of SIMO-ICA is to decompose the mixed observations $\boldsymbol{X}(f)$ into the SIMO components of each independent sound source; i.e., 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.

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, 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 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 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-model-based binary mask. To illustrate the novelty of

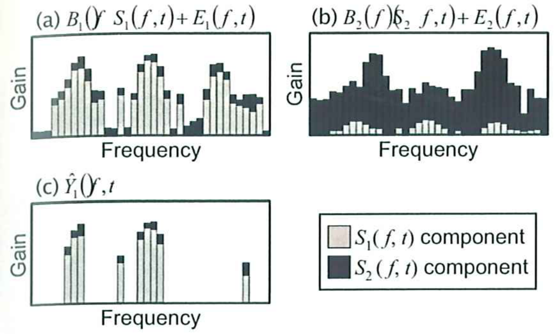


Fig. 2. Examples of spectra in simple combination of ICA and binary masking. (a) ICA's output 1; $B_1(f)S_1(f,t)+E_1(f,t)$, (b) ICA's output 2; $B_2(f)S_2(f,t)+E_2(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.

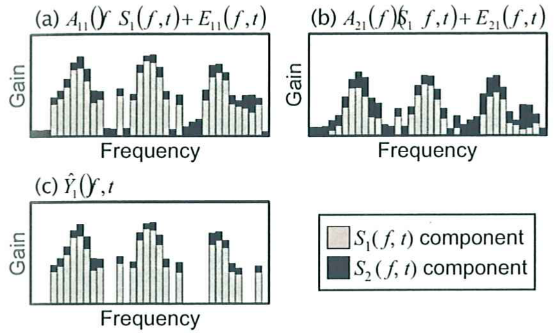


Fig. 3. Examples of spectra in proposed two-stage method. (a) SIMO-ICA's output 1; $A_{11}(f)S_1(f,t)+E_{11}(f,t)$, (b) SIMO-ICA's output 2; $A_{21}(f)S_1(f,t)+E_{21}(f,t)$, and (c) result of binary masking between (a) and (b); $\hat{Y}_1(f,t)$.

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)) [9].

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 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.,

$$|B_1(f)S_1(f,t)+E_1(f,t)| \simeq |B_2(f)S_2(f,t)+E_2(f,t)|, \quad (5)$$

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.,

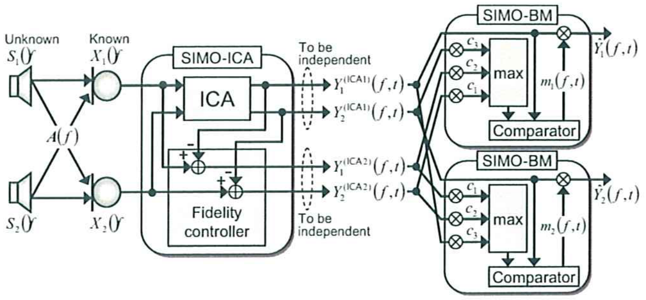


Fig. 4. Input and output relations in proposed two-stage BSS which consists of FD-SIMO-ICA and SIMO-BM, where $K=L=2$ and exclusively selected permutation matrices are given by $\boldsymbol{P}_1=\boldsymbol{I}$ and $\boldsymbol{P}_2=[1]_{ij}-\boldsymbol{I}$. in (9)

$$|A_{11}(f)S_1(f,t)+E_{11}(f,t)| > |A_{21}(f)S_1(f,t)+E_{21}(f,t)|. \quad (6)$$

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 [9], 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.

3.3. Algorithm: SIMO-ICA in 1st stage

Time-domain SIMO-ICA [6] 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

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t)=[Y_k^{(\mathrm{ICA}l)}(f,t)]_{k1}=\boldsymbol{W}_{(\mathrm{ICA}l)}(f)\boldsymbol{X}(f,t), \quad (7)$$

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,

$$\boldsymbol{Y}_{(\mathrm{ICA}L)}(f,t) \quad = \quad \boldsymbol{X}(f,t) - \sum_{l=1}^{L-1} \boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t). \tag{8}$$

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 (8) 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 (7), and simultaneously the signals obtained by (8) are also mutually independent, then the output signals converge on unique solutions, up to the permutation and the residual error, as

$$\boldsymbol{Y}_{(\mathrm{ICA}l)}(f,t) = \mathrm{diag}\ \boldsymbol{A}(f)\boldsymbol{P}_l^{\mathrm{T}}\ \boldsymbol{P}_l\boldsymbol{S}(f,t) + \boldsymbol{E}_l(f,t), \tag{9}$$

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 [6] 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

$$\begin{aligned} & Y_1^{(\mathrm{ICA1})}(f,t),\ Y_2^{(\mathrm{ICA1})}(f,t)\ ^{\mathrm{T}} \\ & = A_{11}(f)S_1(f,t){+}E_{11}(f,t),\ A_{22}(f)S_2(f,t){+}E_{22}(f,t)\ ^{\mathrm{T}}, \end{aligned} \tag{10}$$

$$\begin{aligned} & Y_1^{(\mathrm{ICA2})}(f,t),\ Y_2^{(\mathrm{ICA2})}(f,t)\ ^{\mathrm{T}} \\ & = A_{12}(f)S_2(f,t){+}E_{12}(f,t),\ A_{21}(f)S_1(f,t){+}E_{21}(f,t)\ ^{\mathrm{T}}, \end{aligned} \tag{11}$$

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

In order to obtain (10) and (11), the natural gradient of Kullback-Leibler divergence of (8) 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

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

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

3.4. 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 (10) and (11). The resultant output signal corresponding to source 1 is determined in the proposed SIMO-BM as follows:

$$\hat{Y}_1(f,t) \quad = \quad m_1(f,t)Y_1^{(\mathrm{ICA1})}(f,t), \tag{13}$$

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

$$\begin{aligned} & Y_1^{(\mathrm{ICA1})}(f,t) \\ & > \max\ c_1|Y_2^{(\mathrm{ICA2})}(f,t)|,\ c_2|Y_1^{(\mathrm{ICA2})}(f,t)|,\ c_3|Y_2^{(\mathrm{ICA1})}(f,t)| \end{aligned} \tag{14}$$

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 [9]. 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. 3.2. 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

$$\hat{Y}_2(f,t) \quad = \quad m_2(f,t)Y_2^{(\mathrm{ICA1})}(f,t), \tag{15}$$

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

$$\begin{aligned} & Y_2^{(\mathrm{ICA1})}(f,t) \\ & > \max\ c_1|Y_1^{(\mathrm{ICA2})}(f,t)|,\ c_2|Y_2^{(\mathrm{ICA2})}(f,t)|,\ c_3|Y_1^{(\mathrm{ICA1})}(f,t)| \end{aligned} \tag{16}$$

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

$$\boldsymbol{P}_l \quad = \quad [\delta_{in(k,l)}]_{ki}, \tag{17}$$

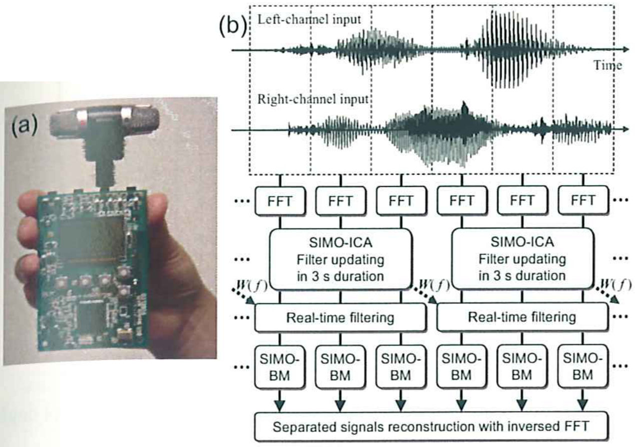


Fig. 5. (a) Overview of pocket-size real-time BSS module, where proposed two-stage BSS algorithm works on TEXAS INSTRUMENTS TMS320C6713 DSP. (b) Signal flow in real-time implementation of proposed method.

Table 1. Specifications of pocket-size real-time BSS module

Processor	TI TMS320VC6713 (clock frequency: 200 MHz)
Input/output interfaces	2 ch mic. in (expandable to 4 ch)
	2 ch speacker/line out
Sampling frequency	8 kHz (expandable to 16 / 32 kHz)
Power supply	AA cell battery × 2
Amount of memory	Flash ROM: 100 KByte used
	SDRAM: 1 MByte used
Weight	150 g (including buttery)

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

$$n(k,l) \;=\; \begin{matrix} k+l-1 & (k+l-1 \le L) \\ k+l-1-L & (k+l-1 > L) \end{matrix} \;. \qquad (18)$$

In this case, (9) yields

$$Y_{(\mathrm{ICA}l)}(f,t) = \; A_{kn(k,l)}(f) S_{n(k,l)}(f,t) + E_{kn(k,l)}(f,t) \;\; _{k1} . \quad (19)$$

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

$$\hat{Y}_1(f,t) \;=\; m_1(f,t) Y_1^{(\mathrm{ICA1})}(f,t), \qquad (20)$$

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

$$\begin{aligned} Y_1^{(\mathrm{ICA1})}(f,t) > \max\; & c_1|Y_2^{(\mathrm{ICA}L)}(f,t)|,\; c_2|Y_3^{(\mathrm{ICA}L-1)}(f,t)|, \\ & c_3|Y_4^{(\mathrm{ICA}L-2)}(f,t)|, \cdots, c_{L-1}|Y_L^{(\mathrm{ICA2})}(f,t)|, \\ & \cdots, c_{LL-1}|Y_L^{(\mathrm{ICA1})}(f,t)| \;; \end{aligned} \qquad (21)$$

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

3.5. Real-time implementation

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 as shown in Fig. 5(a) and Table 1. 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.

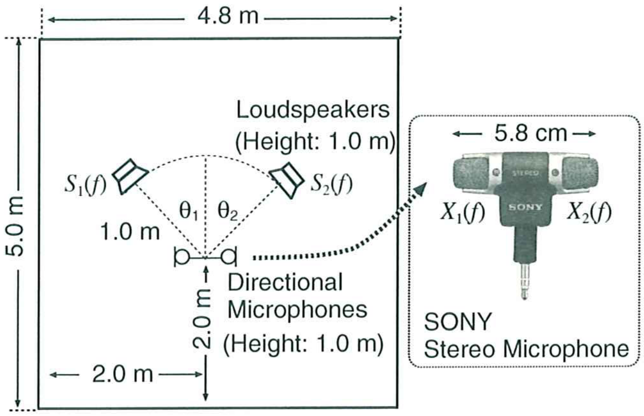


Fig. 6. Layout of reverberant room used in computer-simulation-based BSS experiment, where room impulse responses are recorded for generation of convolutive mixtures. The reverberation time is 200 ms.

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 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 (Fig. 6) for the generation of convolutive mixtures. The reverberation time in this room is 200 ms. We neglect the additive noise term $N(f)$ in (1).

First, to evaluate the feasibility for general hands-free applications, we carried out sound-separation experiments with two sources

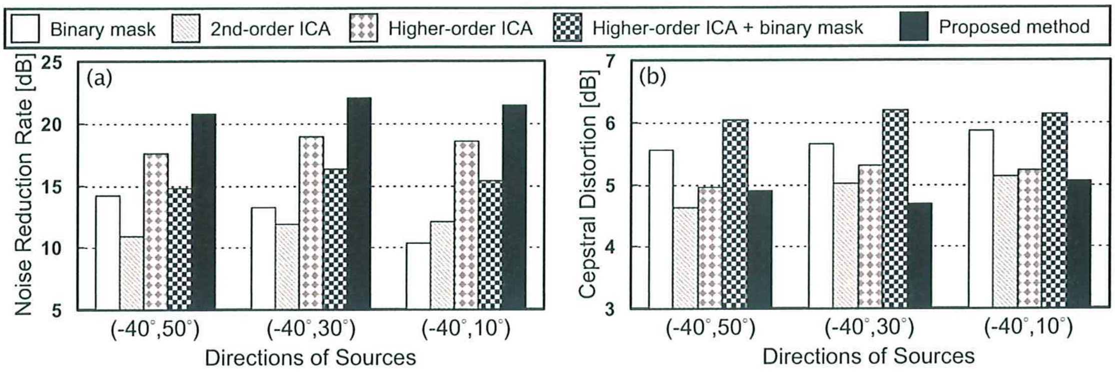


Fig. 7. (a) Results of NRR and (b) results of CD under different speaker configurations and methods, where background noise is neglected. Each score is an average for 12 speaker combinations.

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; $(\theta_1, \theta_2) = (-40°, 50°)$, $(-40°, 30°)$, or $(-40°, 10°)$. Two kinds of sentences, spoken by two male and two female speakers selected from the ASJ continuous speech corpus for research [10], 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 $\boldsymbol{W}(f)$ is 1024. We used a null-beamformer-based initial value [3] which is steered to $(-60°, 60°)$. This experiment corresponds to the *off-line* 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.

4.2. Experimental evaluation of separation performance

We compare the following methods.

(A) Conventional binary-mask-based BSS given in Sect. 2.3.

(B) Conventional second-order-ICA-based BSS proposed by Parra [2], where scaling ambiguity can be properly solved by method used in [1]. Also, permutation is solved by [3].

(C) Conventional higher-order-ICA-based BSS given in Sect. 2.2 with scaling ambiguity solver [1]. Also, permutation is solved by [3].

(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) [3], 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

$$\text{ISNR [dB]} = \frac{1}{L}\sum_{l=1}^{L} 10\log_{10}\frac{\langle|A_{ll}(f)S_l(f,t)|^2\rangle_t}{\langle|X_l(f,t) - A_{ll}(f)S_l(f,t)|^2\rangle_t}, \quad (22)$$

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 $[A_{1l}(f)S_l(f,t), \cdots, A_{Kl}(f)S_l(f,t)]$ for each source to the separation system, where the separation filter matrices and binary-mask patterns estimated in the preceding blind process with $\boldsymbol{X}(f,t)$ are used.

Figure 7(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 of 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, i.e., the degree of sound distortion, in the previous paragraph. To assess the distortion of the separated signals, we measure *cepstral distortion* (CD) [11], 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

$$\text{CD [dB]} \equiv \frac{1}{J}\sum_{j=1}^{J} D_b\sqrt{\sum_{i=1}^{p} 2\left(C_{\text{out}}(i,j) - C_{\text{ref}}(i,j)\right)^2}, \quad (23)$$

where J denotes the number of speech frames, $C_{\text{out}}(i,j)$ is the i-th FFT-based cepstrum of the target component in the separated output at the j-th 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 7(b) shows the results of CD (average of 12 speaker combinations) for all speaker directions. As can be confirmed, the CDs

Table 2. Parameters of Speech Recognition Experiment

Database	JNAS [12], 306 speakers (150 sentences / speaker)
Task	20-k newspaper dictation
Acoustic model	phonetic tied mixture [13] (clean model)
Feature vectors	12-order MFCCs, 12-order ΔMFCCs, 1-order Δ energy
Training data	260 speakers' utterances (150 sentences / speaker)
Testing data	46 speakers' utterances (200 sentences)
Decoder	Julius [13] ver.3.4.2
Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms

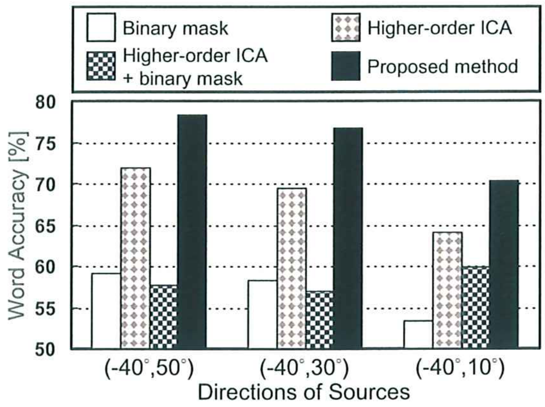


Fig. 8. Result of word accuracy for different speaker allocations and methods. The recognition task is 20k-word newspaper dictation. Julius decoder [13] is used, where a phonetic tied mixture model was trained via 260 speakers selected from JNAS database [12]. Test sets include 46 speakers' utterances (200 sentences).

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 Sect. 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, e.g., hands-free telecommunication via mobile phones.

4.3. Speech recognition experiment

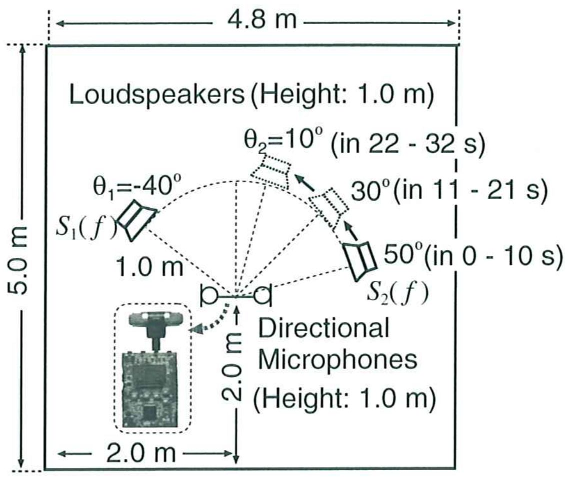


Fig. 9. Layout of reverberant room used in real-recording-based experiment. the reverberation time is 200 ms.

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 2 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 8 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.

5. REAL-TIME SEPARATION EXPERIMENT FOR MOVING SOUND SOURCE

In this section, we discus 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 Fig. 9, where two loudspeakers and the real-time BSS system (Fig. 5) 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 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 s duration, source 2 is set to $\theta_2 = 50^\circ$,
2. in the 10–11 s duration, source 2 moves from $\theta_2 = 50^\circ$ to 30°,
3. in the 11–21 s duration, source 2 is settled in $\theta_2 = 30^\circ$,
4. in the 21–22 s duration, source 2 moves from $\theta_2 = 30^\circ$ to 10°,
5. in 22–32 s 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.

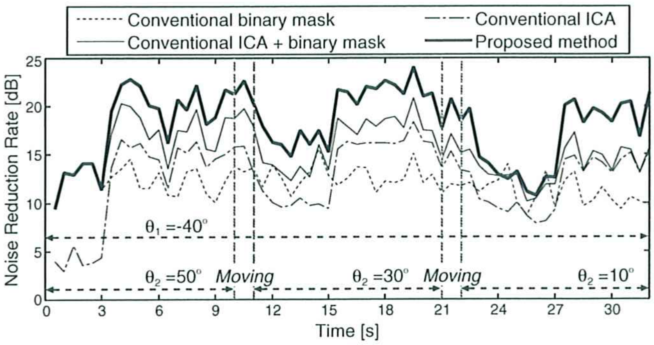


Fig. 10. Results of segmental NRR calculated along time axis at 0.5 s intervals, where real recording data and real-time BSS are used. Each line is an average for 12 speaker combinations. The levels of background noise and sound source are 39 dB(A) and 65 dB(A), respectively.

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 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 10 shows the averaged segmental NRR for 12 speaker combinations, which was calculated along the time axis at 0.5 s intervals. The first 3 s 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 s, we simply applied binary masking in methods (C) and (D). The successive duration (in the period of 3–32 s) shows the separation results for the *open* data sample, which is to be evaluated in this experiment. From Fig. 10, we can confirm that the proposed two-stage BSS (D) outperforms other methods throughout almost the entire duration of 3–32 s. It is worth noting that conventional ICA shows appreciable 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. On the basis of these results, we can assess the proposed method to be beneficial to many practical real-time BSS applications.

6. 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-ms-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, i.e., an off-line test and an on-line test using a DSP module applied for real recording data.

7. REFERENCES

[1] 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 (ICA99)*, 1999, pp. 365–371.

[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] H. Saruwatari, et al., "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[5] 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, 17 pages, 2006.

[6] 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

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

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

[9] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," *Proc. ICA2004*, pp.832–839, 2004.

[10] 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, (in Japanese).

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

[12] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS : Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of The Acoustic Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[13] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, pp.1691–1694, 2001.