BLIND SEPARATION AND DECONVOLUTION FOR REAL CONVOLUTIVE MIXTURE OF TEMPORALLY CORRELATED ACOUSTIC SIGNALS USING SIMO-MODEL-BASED ICA

Hiroshi SARUWATARI, Tomoya TAKATANI, Hiroaki YAMAJO, Tsuyoki NISHIKAWA, and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN (E-mail: sawatari@is.aist-nara.ac.jp)

ABSTRACT

We propose a new novel two-stage blind separation and deconvolution (BSD) algorithm for a real convolutive mixture of temporally correlated signals, in which a new Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) and blind multichannel inverse filtering are combined. SIMO-ICA consists of multiple ICAs and a fidelity controller, and each ICA runs in parallel under fidelity control of the entire separation system. SIMO-ICA can separate the mixed signals, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of SIMO-ICA can maintain the spatial qualities of each sound source. After the separation by SIMO-ICA, a simple blind deconvolution technique based on multichannel inverse filtering for the SIMO model can be applied even when the mixing system is the nonminimum phase system and each source signal is temporally correlated. The experimental results obtained under the reverberant condition reveal that the sound quality of the separated signals in the proposed method is superior to that in the conventional ICA-based BSD.

1. INTRODUCTION

Blind separation and deconvolution (BSD) of sources is an approach taken to estimate original source signals using only the information of mixed signals observed in each input channel. The difference between the BSD and blind source separation (BSS) of the convolutive mixture [1, 2] is that not only the source separation but also the deconvolution of the transmission channel characteristics are considered in the BSD framework. Therefore, the BSD technique is highly applicable to robust hands-free speech recognition systems, where the distortion due to the room transfer function should be reduced. For the BSD based on independent component analysis (ICA), various methods have been proposed to deal with the separation and deconvolution for the convolutive mixture of independently, identically distributed (i.i.d.) source signals [3, 4]. These BSD methods require the specific assumptions that the source signals are mutually independent and each source signal is also temporally independent. However, the latter assumption does not hold, particularly in many practical acoustic mixtures of sound signals which often correspond to the temporally correlated signals. The application of the conventional ICA-based BSD to speech often yields the negative results, e.g., the separated speech is adversely decorrelated and whitened [5].

In order to solve the problem, we propose a novel BSD approach that combines information-geometry theory and multichannel signal processing. In this approach, the separation-deconvolution problem is resolved into two stages: the Single-Input Multiple-Output (SIMO)-model-based separation and the deconvolution in the SIMO-model framework. 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 sensors. First, we propose a new blind separation framework using a SIMO-model-based ICA algorithm, SIMO-ICA. In the SIMO-ICA scenario, unknown multiple source signals which are mixed through unknown acoustical transmission channels are detected at the microphones, and these signals can be separated, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. Thus, the separated signals of SIMO-ICA can maintain the spatial qualities of each sound source. After the separation by the SIMO-ICA, a simple blind deconvolution technique based on the multichannel inverse filtering for the SIMO model can be applied. In the proposed method, the separation/deconvolution problems can be solved efficiently using the following reasonable assumption and properties. (1) The assumption of the mutual independence among the acoustic sound sources usually holds, and consequently, this can be used in the SIMO-ICA-based separation. (2) The temporal-correlation property of the source signals and the nonminimum phase property of the mixing system can be taken into account in the blind multichannel inverse filtering. Thus, the proposed algorithm can provide a more feasible performance for the separation and deconvolution of real acoustic signals, compared with the conventional ICA-based BSD. This can be confirmed from the experimental results obtained under the real reverberant condition.

2. MIXING PROCESS AND CONVENTIONAL BSD

2.1. Mixing process

In this study, the number of array elements (microphones) is K and the number of multiple sound sources is L. In general, the observed signals in which multiple source signals are mixed linearly are expressed as

$$\boldsymbol{x}(t) = \sum_{n=0}^{N-1} \boldsymbol{a}(n)\boldsymbol{s}(t-n) = \boldsymbol{A}(z)\boldsymbol{s}(t), \tag{1}$$

where $\boldsymbol{s}(t)$ is the source signal vector, $\boldsymbol{x}(t)$ is the observed signal vector, $\boldsymbol{a}(n)$ is the mixing filter matrix with the length of N, and $\boldsymbol{A}(z)$ is the z-transform of $\boldsymbol{a}(n)$; these are given as

$$\boldsymbol{s}(t) = [s_1(t), \cdots, s_L(t)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{x}(t) = [x_1(t), \cdots, x_K(t)]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{a}(n) = [a_{kl}(n)]_{kl}, \tag{4}$$

$$\boldsymbol{A}(z) = [A_{kl}(z)]_{kl} = \left[\sum_{n=0}^{N-1} a_{kl}(n) z^{-n}\right]_{kl}, \tag{5}$$

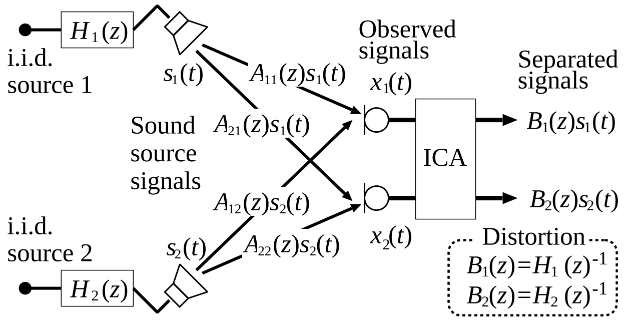

Fig. 1. Input and output relations in conventional ICA-based BSD.

where z^{-1} is used as the unit-delay operator, i.e., $z^{-n} \cdot x(t) = x(t-n)$, a_{kl} is the impulse response between the k-th microphone and the l-th sound source, and $[X]_{ij}$ denotes the matrix which includes the element X in the i-th row and the j-th column. Hereafter, we only deal with the case of $K = L$ in this paper.

2.2. Conventional ICA-based BSD and its problem

In the BSS method, we consider the time-domain ICA (TDICA), in which each element of the separation matrix is represented as a FIR filter. In the TDICA, we optimize the separation matrix using only the full-band observed signals without subband processing. The separated signal $\boldsymbol{y}(t) = [y_1(t), \cdots, y_L(t)]^{\mathrm{T}}$ is expressed as

$$\begin{aligned} \boldsymbol{y}(t) &= \sum_{n=0}^{D-1} \boldsymbol{w}(n)\boldsymbol{x}(t-n) = \boldsymbol{W}(z)\boldsymbol{x}(t) \\ &= \boldsymbol{W}(z)\boldsymbol{A}(z)\boldsymbol{s}(t), \end{aligned} \tag{6}$$

where $\boldsymbol{w}(n)$ is the separation filter matrix, $\boldsymbol{W}(z)$ is the z-transform of $\boldsymbol{w}(n)$, and D is the filter length of $\boldsymbol{w}(n)$. In the ICA-based BSD framework assuming i.i.d. sources, Amari [3] proposed the TDICA algorithm which optimizes the separation filter by minimizing the Kullbuck-Leibler divergence between the joint probability density function (PDF) of $\boldsymbol{y}(t)$ and the product of marginal PDFs of $y_l(t)$. The iterative learning rule is given by

$$\begin{aligned} \boldsymbol{w}^{[j+1]}(n) &= \boldsymbol{w}^{[j]}(n) + \eta \sum_{d=0}^{D-1} \Big\{ \boldsymbol{I}\delta(n-d) \\ &\quad - \left\langle \boldsymbol{\varphi}(\boldsymbol{y}^{[j]}(t))\boldsymbol{y}^{[j]}(t-n+d)^{\mathrm{T}} \right\rangle_t \Big\} \\ &\quad \cdot \boldsymbol{w}^{[j]}(d), \end{aligned} \tag{7}$$

where η is the step-size parameter, the superscript $[j]$ is used to express the value of the j-th step in the iterations, $\langle \cdot \rangle_t$ denotes the time-averaging operator, and $\boldsymbol{I}$ is the identity matrix. $\delta(n)$ is Dirac's delta function, where $\delta(0) = 1$ and $\delta(n) = 0$ $(n \neq 0)$. Also, we define the nonlinear vector function $\boldsymbol{\varphi}(\cdot)$ as

$$\boldsymbol{\varphi}(\boldsymbol{y}(t)) = [\tanh(y_1(t)), \cdots, \tanh(y_L(t))]^{\mathrm{T}}. \tag{8}$$

The conventional ICA-based BSD algorithm forces the separated signals to have the characteristic that their higher-order autocorrelation is $\delta(t)$, i.e., the signals are temporally decorrelated (see Fig. 1). This might have a negative influence on the quality of the separated signals, particularly when confronted with temporally correlated signals such as speech. For example, separated speech is adversely distorted by an excessive whitening effect due to the temporal decorrelation, as described in Sect. 4.3.

3. PROPOSED TWO-STAGE BSD ALGORITHM

In this section, we propose a new two-stage BSD algorithm combining SIMO-ICA and blind multichannel inverse filtering. The detailed process using the proposed algorithm is as follows.

3.1. First stage: SIMO-ICA for source separation

In this stage, we propose a new blind separation method for SIMO-model-based acoustic signals using SIMO-ICA. SIMO-ICA consists of multiple ICA parts and a *fidelity controller*, and each ICA runs in parallel under fidelity control of the entire separation system (see Fig. 2). The separated signals of the l-th ICA in SIMO-ICA are defined by

$$\begin{aligned} \boldsymbol{y}_{\mathrm{ICA}l}(t) &= [y_k^{(l)}(t)]_{k1} = \sum_{n=0}^{D-1} \boldsymbol{w}_{\mathrm{ICA}l}(n)\boldsymbol{x}(t-n) \\ &= \boldsymbol{W}_{\mathrm{ICA}l}(z)\boldsymbol{x}(t), \end{aligned} \tag{9}$$

where $\boldsymbol{w}_{\mathrm{ICA}l}(n)$ is the separation filter matrix in the l-th ICA, and $\boldsymbol{W}_{\mathrm{ICA}l}(z)$ is the z-transform of $\boldsymbol{w}_{\mathrm{ICA}l}(n)$. Regarding the fidelity controller, we introduce the following new cost function to be minimized,

$$\left\langle \| \sum_{l=1}^{L} \boldsymbol{y}_{\mathrm{ICA}l}(t) - \boldsymbol{x}(t-D/2) \|^2 \right\rangle_t, \tag{10}$$

where $\| \boldsymbol{x} \|$ is the Euclidean norm of vector $\boldsymbol{x}$. Note that Eq. (10) is the extension of the cost function proposed by Matsuoka [6] into SIMO model. Unlike [6], however, our cost function Eq. (10) has a more attractive property that it is sure to converge on zero in the optimal point, and we can *blindly* know the convergence point in the iterative learning. On the contrary, the Matsuoka's cost function converges on nonzero and blindly unknown value.

Using Eq. (9) and Eq. (10), we can obtain the appropriate separated signals and maintain their spatial qualities as follows.

Theorem: If the independent sound sources are separated by (9), and simultaneously Eq. (10) is minimized to be zero, then the output signals converge on unique solutions, up to the permutation, as

$$\boldsymbol{y}_{\mathrm{ICA}l}(t) = \mathrm{diag}\left[\boldsymbol{A}(z)\boldsymbol{P}_l^{\mathrm{T}}\right] \boldsymbol{P}_l \boldsymbol{s}(t-D/2), \tag{11}$$

where $\boldsymbol{P}_l$ $(l = 1, \ldots, L)$ are exclusively-selected permutation matrices which satisfy

$$\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}. \tag{12}$$

Proof of Theorem: The necessity is obvious because the following equation holds with Eq. (11).

$$\begin{aligned} \sum_{l=1}^{L} \boldsymbol{y}_{\mathrm{ICA}l}(t) &= \sum_{l=1}^{L} \mathrm{diag}\left[\boldsymbol{A}(z)\boldsymbol{P}_l^{\mathrm{T}}\right] \boldsymbol{P}_l \boldsymbol{s}(t-D/2) \\ &= \left[\sum_{l=1}^{L} A_{kl}(z) s_l(t-D/2) \right]_{k1} \\ &= \boldsymbol{A}(z)\boldsymbol{s}(t-D/2). \end{aligned} \tag{13}$$

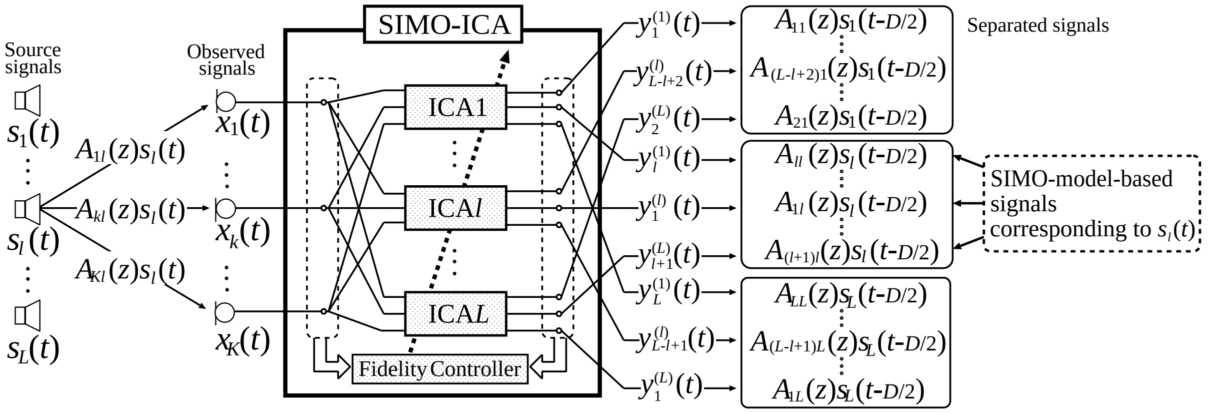

Fig. 2. Example of input and output relations in proposed SIMO-ICA performed in first stage, where permutation $\boldsymbol{P}_l$ is given by Eq. (14).

This results in $\boldsymbol{x}(t-D/2)$, and makes the cost function Eq. (10) be zero. As for the sufficiency, see [7].

Obviously the solutions given by Eq. (11) provide necessary and sufficient SIMO components, $A_{kl}(z)s_l(t-D/2)$, for each l-th source. There, however, is an arbitrariness in a selection of $\boldsymbol{P}_l$. For example, one possible selection is shown in Fig. 2 and this corresponds to

$$\boldsymbol{P}_l \quad = \quad \left[\delta_{im(k,l)}\right]_{ki}, \tag{14}$$

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

$$m(k,l) = \begin{cases} k+l-1 & (k+l-1 \le L) \\ k+l-1-L & (k+l-1 > L) \end{cases}. \tag{15}$$

In this case, Eq. (11) yields

$$\boldsymbol{y}_{\mathrm{ICA}l}(t) \quad = \quad \left[A_{km}(z)s_m(t-D/2)\right]_{k1}. \tag{16}$$

In order to obtain Eq. (11), the gradient of Eq. (10) with respect to $\boldsymbol{w}_{\mathrm{ICA}l}(n)$ should be added to the iterative learning rule of the separation filter. The natural gradient [3] of Eq. (10) is given as

$$\begin{aligned} &\left\{\frac{\partial}{\partial \boldsymbol{w}_{\mathrm{ICA}l}(n)}\left\langle \parallel \sum_{l=1}^{L}\boldsymbol{y}_{\mathrm{ICA}l}(t)-\boldsymbol{x}(t-D/2) \parallel^2 \right\rangle_t\right\} \\ &\qquad \cdot \boldsymbol{W}_{\mathrm{ICA}l}(z^{-1})^{\mathrm{T}}\boldsymbol{W}_{\mathrm{ICA}l}(z) \\ &= \quad 2\sum_{d=0}^{D-1}\left\langle\Big(\sum_{l=1}^{L}\boldsymbol{y}_{\mathrm{ICA}l}(t)-\boldsymbol{x}(t-D/2)\Big)\right. \\ &\qquad \left.\cdot \boldsymbol{y}_{\mathrm{ICA}l}(t-n+d)^{\mathrm{T}}\right\rangle_t \cdot \boldsymbol{w}_{\mathrm{ICA}l}(d). \end{aligned} \tag{17}$$

By combining Eq. (17) with the nonholonomic TDICA [8], we can obtain a new iterative algorithm in the l-th ICA of SIMO-ICA as

$$\begin{aligned} &\boldsymbol{w}_{\mathrm{ICA}l}^{[j+1]}(n) \\ &= \quad \boldsymbol{w}_{\mathrm{ICA}l}^{[j]}(n) \\ &\quad -\alpha\sum_{d=0}^{D-1}\left\{\text{off-diag}\left\langle\boldsymbol{\varphi}\big(\boldsymbol{y}_{\mathrm{ICA}l}^{[j]}(t)\big)\boldsymbol{y}_{\mathrm{ICA}l}^{[j]}(t-n+d)^{\mathrm{T}}\right\rangle_t\right. \\ &\quad +\beta\left\langle\Big(\sum_{l=1}^{L}\boldsymbol{y}_{\mathrm{ICA}l}^{[j]}(t)-\boldsymbol{x}(t-D/2)\Big)\right. \\ &\qquad \left.\left.\cdot \boldsymbol{y}_{\mathrm{ICA}l}^{[j]}(t-n+d)^{\mathrm{T}}\right\rangle_t\right\}\cdot \boldsymbol{w}_{\mathrm{ICA}l}^{[j]}(d), \end{aligned} \tag{18}$$

where α and β are the step-size parameters; α is for the control of the total update quantity and β is for fidelity control. In Eq. (18), updating of $\boldsymbol{w}_{\mathrm{ICA}l}(n)$ for all l should be simultaneously performed in parallel because each iterative equation is associated with the others via $\sum_{l=1}^{L}\boldsymbol{y}_{\mathrm{ICA}l}^{[j]} = \sum_{l=1}^{L}\boldsymbol{W}_{\mathrm{ICA}l}^{[j]}(z)\boldsymbol{x}(t)$.

After the iterations, the separated signals should be classified into SIMO components of each source because the permutation arises. This can be easily achieved by using a cross correlation between time-shifted separated signals, $\max_n \left\langle y_k^{(l)}(t)y_{k'}^{(l')}(t-n)\right\rangle_t$, where $l \ne l'$ and $k \ne k'$. The large value of the correlation indicates that $y_k^{(l)}(t)$ and $y_{k'}^{(l')}(t)$ are SIMO components of the same source.

3.2. Second stage: blind multichannel inverse filtering for deconvolution

In this stage, first, consider the blind channel identification corresponding to the first sound source $s_1(t)$, as shown in Fig. 3(a), where we deal with the case of $K = L = 2$. In this process, the room transfer functions, $A_{11}(z)$ and $A_{21}(z)$, can be estimated by a subchannel matching approach [9, 10, 11] in an SIMO framework because we have already resolved the mixing process of the sources into a simple SIMO model through SIMO-ICA in the previous stage. To take an example under Eq. (14), the output signal of this subchannel matching system is given by

$$\begin{aligned} e_1(t) \quad &= \quad \hat{A}_{21}(z)y_1^{(1)}(t)-\hat{A}_{11}(z)y_2^{(2)}(t) \\ &= \quad \big(\hat{A}_{21}(z)A_{11}(z)-\hat{A}_{11}(z)A_{21}(z)\big)s_1(t-D/2), \end{aligned} \tag{19}$$

where $\hat{A}_{kl}(z)$ is the estimated transfer function of the mixing process $A_{kl}(z)$, and is defined using the estimated impulse responses $\hat{a}_{kl}(n)$ as

$$\hat{A}_{kl}(z) \quad = \quad \sum_{n=0}^{N-1}\hat{a}_{kl}(n)z^{-n}. \tag{20}$$

For a common input $s_1(t)$ even with the temporal correlation, the output signal $e_1(t)$ of the entire system will be zero if and only if the following holds (γ is an arbitrary constant):

$$\hat{A}_{11}(z) \quad = \quad \gamma A_{11}(z), \tag{21}$$

$$\hat{A}_{21}(z) \quad = \quad \gamma A_{21}(z). \tag{22}$$

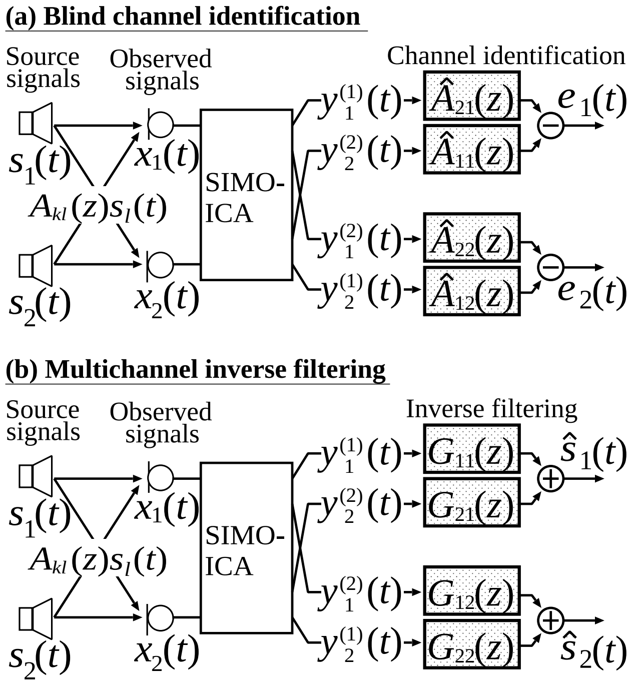

Fig. 3. Input and output relations in (a) blind channel identification and (b) multichannel inverse filtering performed in second stage.

Thus, the minimization of $\langle e_1(t)^2 \rangle_t$ in terms of $\hat{A}_{11}(z)$ and $\hat{A}_{21}(z)$ yields the identification of $A_{11}(z)$ and $A_{21}(z)$. The output signal power $\langle e_1(t)^2 \rangle_t$ is given as

$$\begin{aligned} \langle e_1(t)^2 \rangle_t &= \langle \boldsymbol{a}_1^{\mathrm{T}} \boldsymbol{y}_1(t) \boldsymbol{y}_1^{\mathrm{T}}(t) \boldsymbol{a}_1 \rangle_t \\ &= \boldsymbol{a}_1^{\mathrm{T}} \langle \boldsymbol{y}_1(t) \boldsymbol{y}_1^{\mathrm{T}}(t) \rangle_t \boldsymbol{a}_1, \end{aligned} \tag{23}$$

where

$$\begin{aligned} \boldsymbol{a}_1 &= [\hat{a}_{21}(0),\ \hat{a}_{21}(1),\ \cdots,\ \hat{a}_{21}(N-1), \\ &\quad -\hat{a}_{11}(0),\ -\hat{a}_{11}(1),\ \cdots,\ -\hat{a}_{11}(N-1)]^{\mathrm{T}}, \end{aligned} \tag{24}$$

$$\begin{aligned} \boldsymbol{y}_1(t) &= [y_1^{(1)}(t),\ y_1^{(1)}(t-1),\ \cdots,\ y_1^{(1)}(t-N+1), \\ &\quad y_2^{(2)}(t),\ y_2^{(2)}(t-1),\ \cdots,\ y_2^{(2)}(t-N+1)]^{\mathrm{T}}. \end{aligned} \tag{25}$$

From this, we solve the following minimization problem:

$$\min_{\boldsymbol{a}_1} \boldsymbol{a}_1^{\mathrm{T}} \langle \boldsymbol{y}_1(t) \boldsymbol{y}_1^{\mathrm{T}}(t) \rangle_t \boldsymbol{a}_1, \quad \text{subject to } \|\boldsymbol{a}_1\| = \text{const.} \tag{26}$$

The optimal vector $\boldsymbol{a}_1$ in Eq. (26) can be derived as the specific eigenvector which corresponds to the minimum eigenvalue of the matrix $\langle \boldsymbol{y}_1(t) \boldsymbol{y}_1^{\mathrm{T}}(t) \rangle_t$.

Next, regarding the blind channel identification corresponding to another sound source $s_2(t)$, the optimal vector $\boldsymbol{a}_2$ can be derived as the eigenvector which corresponds to the minimum eigenvalue of the matrix $\langle \boldsymbol{y}_2(t) \boldsymbol{y}_2^{\mathrm{T}}(t) \rangle_t$, where $\boldsymbol{a}_2$ and $\boldsymbol{y}_2(t)$ are defined as

$$\begin{aligned} \boldsymbol{a}_2 &= [\hat{a}_{22}(0),\ \hat{a}_{22}(1),\ \cdots,\ \hat{a}_{22}(N-1), \\ &\quad -\hat{a}_{12}(0),\ -\hat{a}_{12}(1),\ \cdots,\ -\hat{a}_{12}(N-1)]^{\mathrm{T}}, \end{aligned} \tag{27}$$

$$\begin{aligned} \boldsymbol{y}_2(t) &= [y_1^{(2)}(t),\ y_1^{(2)}(t-1),\ \cdots,\ y_1^{(2)}(t-N+1), \\ &\quad y_2^{(1)}(t),\ y_2^{(1)}(t-1),\ \cdots,\ y_2^{(1)}(t-N+1)]^{\mathrm{T}}. \end{aligned} \tag{28}$$

Finally, we can estimate the multichannel inverse filters, $G_{11}(z)$ and $G_{21}(z)$ for $\hat{A}_{11}(z)$ and $\hat{A}_{21}(z)$, and $G_{12}(z)$ and $G_{22}(z)$ for $\hat{A}_{12}(z)$ and $\hat{A}_{22}(z)$, based on the multiple-input/output inverse theorem (MINT) [10]. In the MINT method, the exact inverse of the room acoustics can be uniquely determined, even when $\hat{A}_{kl}(z)$ has the nonminimum phase properties, if $\hat{A}_{kl}(z)$ does not have any common zeros in the z-plane. The optimal multichannel inverse filters $G_{kl}(z)$ are derived by solving the following diophantine equations:

$$G_{11}(z)\hat{A}_{11}(z) + G_{21}(z)\hat{A}_{21}(z) = 1, \tag{29}$$

$$G_{12}(z)\hat{A}_{12}(z) + G_{22}(z)\hat{A}_{22}(z) = 1. \tag{30}$$

The recovered source signals $\hat{s}_l(t)$ can be given by (see Fig. 3(b))

$$\hat{s}_1(t) = G_{11}(z)y_1^{(1)}(t) + G_{21}(z)y_2^{(2)}(t), \tag{31}$$

$$\hat{s}_2(t) = G_{12}(z)y_1^{(2)}(t) + G_{22}(z)y_2^{(1)}(t). \tag{32}$$

3.3. Advantages and disadvantages of proposed algorithm

In the proposed method, the separation-deconvolution problem is resolved into two stages, i.e., SIMO-model-based separation and deconvolution in the SIMO-model framework. Individual problems in each stage can be solved efficiently using the following reasonable assumption and properties of the source signals and the mixing system.

- The assumption of the mutual independence among the acoustic sound sources, such as speech, usually holds in many practical situations, and consequently, this can be used for the SIMO-ICA-based separation.
- The temporal-correlation property of the source signals can be taken into account and retained in the blind multichannel inverse filtering.
- The nonminimum phase property of the mixing system does not prevent the multichannel inverse filtering from achieving the deconvolution.

Thus, the proposed algorithm can provide a feasible performance for the separation and deconvolution of the acoustic signals.

Regarding blind multichannel inverse filtering, the accurate estimation of the filter length N of the room impulse responses is indispensable for improving the system identification performance. In many practical cases, however, it is difficult to estimate the filter length blindly, particularly when room reverberation is long. Various methods for filter-length estimation based on the MDL, AIC, [11] and utilization of the modeling error from MINT filtering [10] have been presented, however they cannot be guaranteed to be applicable in a real acoustical environment. Moreover, the estimation accuracy often degrades because of the existence

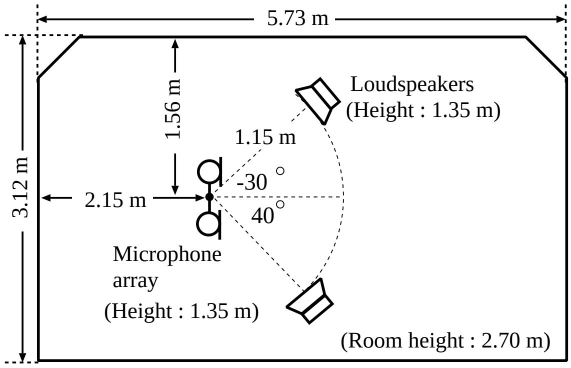

Fig. 4. Layout of reverberant room used in experiments.

of residual crosstalk which arose in the previous source-separation stage. Therefore, we should choose an alternative and easy way of underestimating the filter length rather than the actual length of the impulse responses. This will produce insufficient deconvolution results, however, the excessive decorrelation effect for the source signals can be avoided.

4. EXPERIMENT AND RESULTS

4.1. Conditions for experiment

A two-element array with an interelement spacing of 4 cm is assumed. The speech signals are assumed to arrive from two directions, $-30°$ and $40°$. The distance between the microphone array and the loudspeakers is 1.15 m. Two sentences, spoken by two male speakers selected from the ASJ continuous speech corpus for research, are used as the original speech samples. The sampling frequency is 8 kHz and the length of speech is limited to 3 seconds. The source signals are the original speech convolved with the impulse response specified by the reverberation time (RT) of 150 ms. The impulse response is recorded in the experimental room (see Fig. 4). The length of $\boldsymbol{w}(n)$ is set to be 512, and the initial value is the Null-Beamformer [2] whose directional null is steered to $\pm 60°$. The number of iterations in ICA is 5000.

4.2. Objective evaluation score

In this experiment, two objective evaluation scores are defined as described below. First, *noise reduction rate* (NRR), 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, where we do not take into account the distortion of the separated signal. The SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise; these are defined as

$$\mathrm{OSNR}_l^{(\mathrm{ICA1})} = 10\log_{10}\frac{\sum_t \mid H_{ll}^{\mathrm{ICA1}}(z)s_l(t)\mid^2}{\sum_t \mid H_{ln}^{\mathrm{ICA1}}(z)s_n(t)\mid^2}, \quad (33)$$

$$\mathrm{ISNR}_l^{(\mathrm{ICA1})} = 10\log_{10}\frac{\sum_t \mid A_{ll}(z)s_l(t)\mid^2}{\sum_t \mid A_{ln}(z)s_n(t)\mid^2}, \quad (34)$$

$$\mathrm{OSNR}_l^{(\mathrm{ICA2})} = 10\log_{10}\frac{\sum_t \mid H_{ln}^{\mathrm{ICA2}}(z)s_n(t)\mid^2}{\sum_t \mid H_{ll}^{\mathrm{ICA2}}(z)s_l(t)\mid^2}, \quad (35)$$

$$\mathrm{ISNR}_l^{(\mathrm{ICA2})} = 10\log_{10}\frac{\sum_t \mid A_{ln}(z)s_n(t)\mid^2}{\sum_t \mid A_{ll}(z)s_l(t)\mid^2}, \quad (36)$$

Table 1. NRR results for each sound source [dB]

	Conventional BSD	Proposed SIMO-ICA
Source 1	13.0	10.5
Source 2	12.8	10.1

where $\mathrm{OSNR}_l^{(\mathrm{ICA}k)}$ and $\mathrm{ISNR}_l^{(\mathrm{ICA}k)}$ are the output SNR and the input SNR for ICAk, respectively, and $l \neq n$. Also, $H_{ij}^{\mathrm{ICA}k}(z)$ is the element in the i-th row and the j-th column of the matrix $\boldsymbol{H}^{\mathrm{ICA}k}(z) = \boldsymbol{W}_{\mathrm{ICA}k}(z)\boldsymbol{A}(z)$.

Secondly, *cepstral distortion* (CD) is used as the indication of deconvolution performance. In this study, we defined the CD as the distance between the spectral envelope of the original source signal $s_l(t)$ and that of the separated output. The CD will be decreased to zero if the separation-deconvolution processing is performed perfectly, and the CD reduction of 1 dB roughly corresponds to an SNR improvement of more than 5 dB in the case that the additional noise is white Gaussian. It is known that the CD has a good correspondence to the subjective evaluation score of the speech quality or the recognition score of the speech recognizer. The 40th-order Mel-scaled cepstrum based on the smoothed FFT spectrum is used.

4.3. Results and discussion

The step-size parameter α is changed from 5.0×10^{-8} to 1.0×10^{-6} and β is changed from 5.0×10^{-3} to 2.0×10^{-2} in order to find the optima which minimize Eq. (10). Table 1 shows the results of NRR for each speaker. In these scores, the proposed SIMO-ICA can achieve NRRs of more than 10 dB, however, the deterioration of NRR in SIMO-ICA is more than 2.5 dB compared with that in the conventional ICA-based BSD. We speculate that the *specious* performance of the conventional ICA-based BSD is due to the excessive emphasis on high-frequency components by the effect of temporal decorrelation; the evidence for this will be shown in the next discussion on CD. In general, separation in the high-frequency region is easier than that in the low-frequency region [12] because the reverberation shortens as the frequency increases. Thus, the conventional ICA-based BSD gains the improvement of the NRR only in the high-frequency region, however the accompanying whitening effect leads to an adverse result for separating the speech signals from the practical viewpoint.

Figures 5 (a) and (b) show the results of CD for different lengths of the deconvolution filter. The results concerning the conventional ICA-based BSD, the proposed SIMO-ICA itself (without deconvolution), and the proposed two-stage BSD algorithm are shown in this figure. First, it is evident that the CD of the conventional ICA-based BSD is obviously high, and that of the SIMO-ICA is rather low, i.e., the sound quality in the separation/deconvolution results is degraded in the conventional method. Next, regarding the results of the proposed BSD, there is a considerable reduction of CD with the deconvolution filter length of less than 200 taps, compared with reduction in the conventional method or simple SIMO-ICA. This indicates that the proposed BSD algorithm can achieve reasonable separation and deconvolution for a real convolutive mixture of speech, even when the mixing system is the nonminimum phase system and each source signal is temporally correlated. However, regarding the results of the proposed BSD with a deconvolution filter of more than 200 taps, a significant increase of CD is shown. The main reason for this is

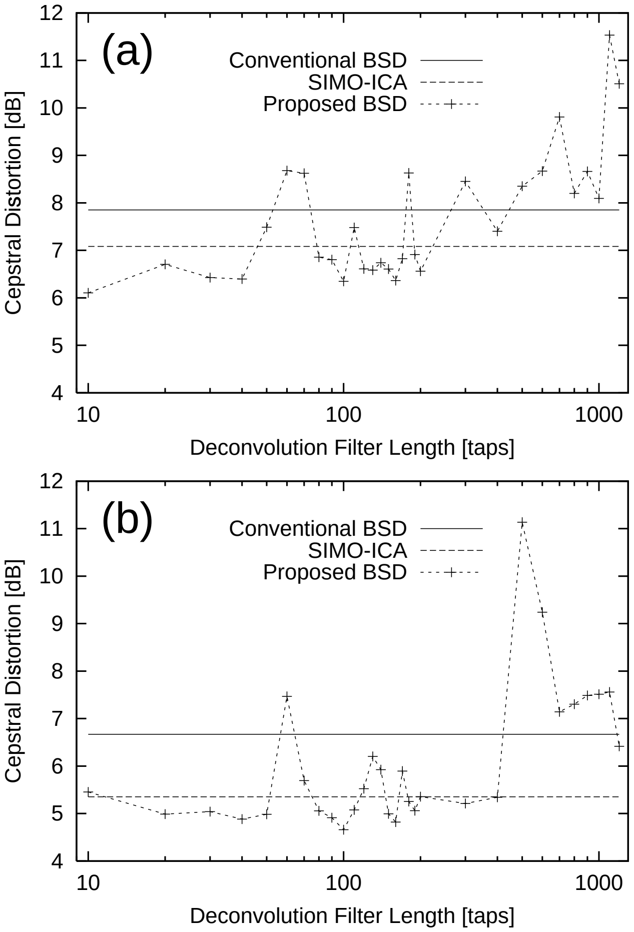

Fig. 5. Results of CD for different length of the deconvolution filter with regard to (a) sound source 1 and (b) sound source 2.

the instability of the multichannel inverse filtering in the deconvolution stage.

On the basis of these overall results, we can conclude that the use of the proposed two-stage BSD algorithm is an effective approach for recovering the original sound source signals in a real acoustical environment; however, the length of the deconvolution filter should be carefully determined without loss of quality of the separated sound. The blind estimation of the appropriate filter length remains an open problem for future study.

5. CONCLUSION

We proposed a new novel BSD framework in which SIMO-ICA and blind multichannel inverse filtering 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 the loss of their spatial qualities. Thus, after SIMO-ICA, we can easily use the simple blind channel identification and multichannel inverse filtering technique. In order to evaluate its effectiveness, a separation-deconvolution experiment was carried out using 2 microphones and 2 speech sources under the condition that the RT is set to 150 ms. The experimental results revealed that (1) the conventional ICA-based BSD includes adverse spectral distortion due to the inherent whitening effect, and (2) the spectral distortion can be considerably reduced by using the proposed two-stage BSD algorithm. Therefore, we can conclude that the proposed BSD algorithm is applicable to high-fidelity sound recovery processing.

6. ACKNOWLEDGEMENT

The authors are grateful to Dr. Kiyotoshi Matsuoka of Kyusyu Institute of Technology, and Dr. Shoji Makino and Ms. Shoko Araki of NTT Co., Ltd. for their fruitful discussions. This work was partly supported by CREST (Core Research for Evolutional Science and Technology) in Japan.

7. REFERENCES

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

[2] H. Saruwatari, T. Kawamura, Tsuyoki Nishikawa, and K. Shikano, “Fast-convergence algorithm for blind source separation based on array signal processing,” *IEICE Trans. Fundamentals*, vol.E86-A, no.3, pp.286–291, 2003.

[3] S. Amari, S. Douglas, A. Cichocki, and H. H. Yang, “Multichannel blind deconvolution and equalization using the natural gradient,” *Proc. of IEEE international Workshop on Wireless Communication*, pp.101–104, April 1997.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, Ltd., West Sussex, 2002.

[5] T. Nishikawa, H. Saruwatari, and K. Shikano, “Stable learning algorithm for blind separation of temporally correlated signals combining multistage ICA and linear prediction,” *Proc. International Symposium on ICA and BSS (ICA2003)*, April 2003.

[6] K. Matsuoka and S. Nakashima, “Minimal distortion principle for blind source separation,” *Proc. International Conference on ICA and BSS (ICA2001)*, pp.722–727, Dec. 2001.

[7] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, “SIMO-model-based independent component analysis for high-fidelity blind separation of acoustic signals,” *Proc. International Symposium on ICA and BSS (ICA2003)*, April 2003.

[8] S. Choi, S. Amari, A. Cichocki, and R. Liu, “Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels,” *Proc. International Workshop on ICA and BSS (ICA’99)*, pp.371–376, 1999.

[9] H. Xu and L. Tong, “A deterministic approach to blind identification of multi-channel FIR systems,” *Proc. ICASSP94*, vol.IV, pp.581–584, 1994.

[10] K. Furuya and Y. Kaneda, “Two-channel blind deconvolution of nonminimum phase FIR system,” *IEICE Trans. Fundamentals*, vol.E80-A, no.5, pp.804–808, 1997.

[11] Z. Ding and Y. Li, *Blind Equalization and Identification*, Marcel Dekker, Inc., New York, 2001.

[12] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, “Time domain ICA blind source separation of non-stationary convolved signals by utilizing geometric beamforming,” *Proc. IEEE International Workshop on Neural Networks for Signal Processing*, pp.445–454, 2002.