PAPER *Special Section on Adaptive Signal Processing and Its Applications*

Multistage SIMO-Model-Based Blind Source Separation Combining Frequency-Domain ICA and Time-Domain ICA

Satoshi UKAI[†a)], ***Nonmember***, **Tomoya TAKATANI**[†], ***Student Member***, **Hiroshi SARUWATARI**[†], **Kiyohiro SHIKANO**[†], **Ryo MUKAI**[††], ***and* Hiroshi SAWADA**[††], ***Members***

SUMMARY In this paper, single-input multiple-output (SIMO)-model-based blind source separation (BSS) is addressed, where unknown mixed source signals are detected at microphones, and can be separated, not into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. This technique is highly applicable to high-fidelity signal processing such as binaural signal processing. First, we provide an experimental comparison between two kinds of SIMO-model-based BSS methods, namely, conventional frequency-domain ICA with projection-back processing (FDICA-PB), and SIMO-ICA which was recently proposed by the authors. Secondly, we propose a new combination technique of the FDICA-PB and SIMO-ICA, which can achieve a higher separation performance than the two methods. The experimental results reveal that the accuracy of the separated SIMO signals in the simple SIMO-ICA is inferior to that of the signals obtained by FDICA-PB under low-quality initial value conditions, but the proposed combination technique can outperform both simple FDICA-PB and SIMO-ICA.

key words: *blind source separation, microphone array, independent component analysis, SIMO model*

1. Introduction

Blind source separation (BSS) is the approach taken to estimate original source signals using only the data of the mixed signals observed in each input channel. This technique is based on *unsupervised adaptive filtering* [1], 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 various fields of signal processing such as digital communications systems and acoustic signal processing systems. In this paper, we mainly address the BSS problem encountered in acoustic signal processing.

In recent studies based on independent component analysis (ICA) [2], various methods have been proposed for dealing with the BSS for acoustical sounds [3]–[8]. However, the existing ICA-based BSS approaches are basically means of extracting each of the independent sound sources as a *monaural* signal. Accordingly, they have a serious drawback in that the separated sounds cannot maintain information about the directivity, localization, or spatial qualities of each sound source. This prevents any BSS methods from being applied to binaural signal processing [9], or any high-fidelity acoustic signal processing.

In order to solve this problem, we must adopt a new blind separation framework in which Single-Input Multiple-Output (SIMO)-model-based BSS is considered. 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. In the SIMO-model-based separation 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, SIMO-model-based separated signals can maintain the spatial qualities of each sound source. Clearly, this attractive feature makes SIMO-model-based BSS highly applicable to high-fidelity acoustic signal processing, e.g., binaural sound separation [10]. In addition, owing to the fact that SIMO-model-based separated signals are still one set of array signals, there exist alternative applications in which SIMO-model-based separation is combined with other types of multichannel signal processing; the Multiple-Input Multiple-Output (MIMO) system deconvolution [11], and the combination of SIMO-model-based BSS with adaptive beamforming [12].

The first objective of this paper is to provide an experimental comparison between two kinds of SIMO-model-based BSS methods, as follows; (a) conventional frequency-domain ICA (FDICA) with projection-back processing (hereafter we call this *FDICA-PB*), proposed by Murata and Ikeda [13], and (b) *SIMO-ICA* which consists of multiple time-domain ICAs (TDICAs), recently proposed by the authors [14], [15]. The second objective of this paper is to propose a new combination technique of the FDICA-PB and SIMO-ICA, which can achieve a higher separation performance with low computational complexity in comparison to each of the two separate methods. It is worth mentioning that this study is well inspired by Nishikawa's multistage ICA approach [16] although Nishikawa's ICA was still monaural-output ICA; indeed, the research presented in this paper is an extension of the multistage ICA to SIMO-model-based BSS framework. The experiments are carried out under a reverberant condition, and the results explicitly reveal

Manuscript received June 24, 2004.
Manuscript revised October 2, 2004.
Final manuscript received November 15, 2004.

[†] The authors are with Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan.

[††] The authors are with NTT Communication Science Laboratories, Kyoto-fu, 619-0237 Japan.

a) E-mail: sato-uk@is.naist.jp

DOI: 10.1093/ietfec/e88–a.3.642

the advantages and disadvantages of each method, and the superiority of the proposed combination technique over the FDICA-PB and SIMO-ICA techniques.

The rest of this paper is organized as follows. In Sect. 2, the sound mixing model is described. In Sect. 3, the conventional SIMO-model-based BSS methods are explained in detail. In Sect. 4, the complementarity among the conventional methods is pointed out, and the combination method is newly proposed. In Sect. 5, the signal-separation experiments are described and the results are compared with those for the conventional methods. Following a discussion on the results of the experiments, we give conclusions in Sect. 6.

2. Sound Mixing Process

In this study, the number of microphones is K and the number of sound sources is L. 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 $\boldsymbol{s}(t) = [s_1(t), \cdots, s_L(t)]^{\mathrm{T}}$ is the source signal vector, and $\boldsymbol{x}(t) = [x_1(t), \cdots, x_K(t)]^{\mathrm{T}}$ is the observed signal vector. Also, $\boldsymbol{a}(n) = [a_{kl}(n)]_{kl}$ is the mixing filter matrix with the length of N, and $\boldsymbol{A}(z) = [A_{kl}(z)]_{kl} = [\sum_{n=0}^{N-1} a_{kl}(n)z^{-n}]_{kl}$ is the z-transform of $\boldsymbol{a}(n)$, where z^{-1} is used as the unit-delay operator, i.e., $z^{-n} \cdot x(t) = x(t-n)$, $a_{kl}(n)$ 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.

3. Conventional SIMO-Model-Based BSS

3.1 What is SIMO-Model-Based BSS?

In general, the observed signal can be represented as a superposition of the SIMO-model-based signals as follows:

$$\begin{aligned}\boldsymbol{x}(t) = {} & [A_{11}(z)s_1(t), \cdots, A_{K1}(z)s_1(t)]^{\mathrm{T}} \\ & + [A_{12}(z)s_2(t), \cdots, A_{K2}(z)s_2(t)]^{\mathrm{T}} \\ & \vdots \\ & + [A_{1L}(z)s_L(t), \cdots, A_{KL}(z)s_L(t)]^{\mathrm{T}},\end{aligned} \tag{2}$$

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

The aim of SIMO-model-based BSS is to decompose the mixed observations $\boldsymbol{x}(t)$ into the SIMO components of each independent sound source, i.e., we estimate $A_{kl}(z)s_l(t)$ for all k and l (up to the permissible time delay in the separation filtering). The SIMO-model-based BSS has the advantages that the separated signals is less distorted and maintain the spatial qualities of each sound source in comparison to the conventional ICA-based BSS. Note that Matsuoka et al. have proposed a modified ICA based on the Minimal Distortion Principle (MDP) [17] to reduce the distortion in the separated signals. However, Matsuoka's method only estimates the limited part of the SIMO components, $A_{ll}(z)s_l(t)$ for all l. Consequently, this method is valid only for monaural outputs, and the spatial qualities of the output signals cannot be obtained.

In the following section, two kinds of existing SIMO-model-based BSS methods are described in detail, and their advantages and disadvantages are pointed out.

3.2 Conventional FDICA-PB [13]

In the conventional FDICA-PB (see Fig. 1), first, the short-time analysis of observed signals is conducted by frame-by-frame discrete Fourier transform (DFT). By plotting the spectral values in a frequency bin for each microphone input frame by frame, we consider them as a time series. Hereafter, we designate the time series as $\boldsymbol{X}(f,t) = [X_1(f,t), \cdots, X_K(f,t)]^{\mathrm{T}}$.

Next, we perform signal separation using the complex-valued unmixing matrix, $\boldsymbol{W}(f) = [W_{lk}(f)]_{lk}$, so that the L time-series output $\boldsymbol{Y}(f,t) = [Y_1(f,t), \cdots, Y_L(f,t)]^{\mathrm{T}}$ becomes mutually independent; this procedure can be given as

$$\boldsymbol{Y}(f,t) = \boldsymbol{W}(f)\boldsymbol{X}(f,t). \tag{3}$$

We perform this procedure with respect to all frequency

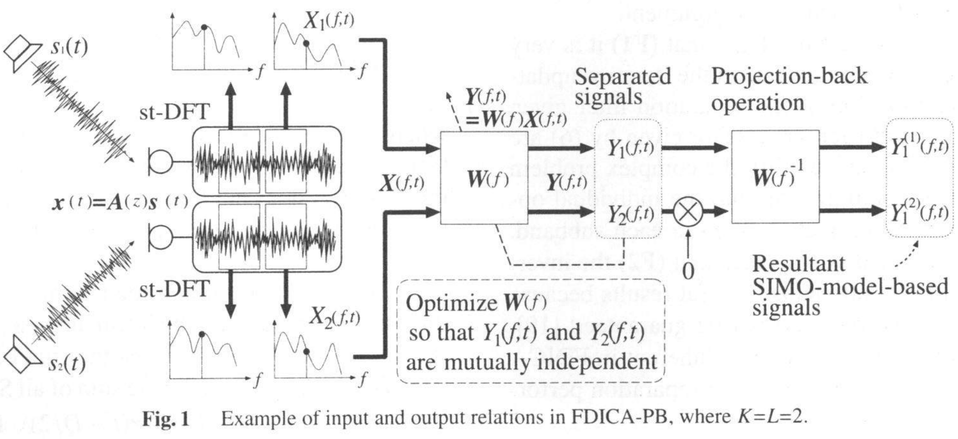


Fig. 1 Example of input and output relations in FDICA-PB, where K=L=2.

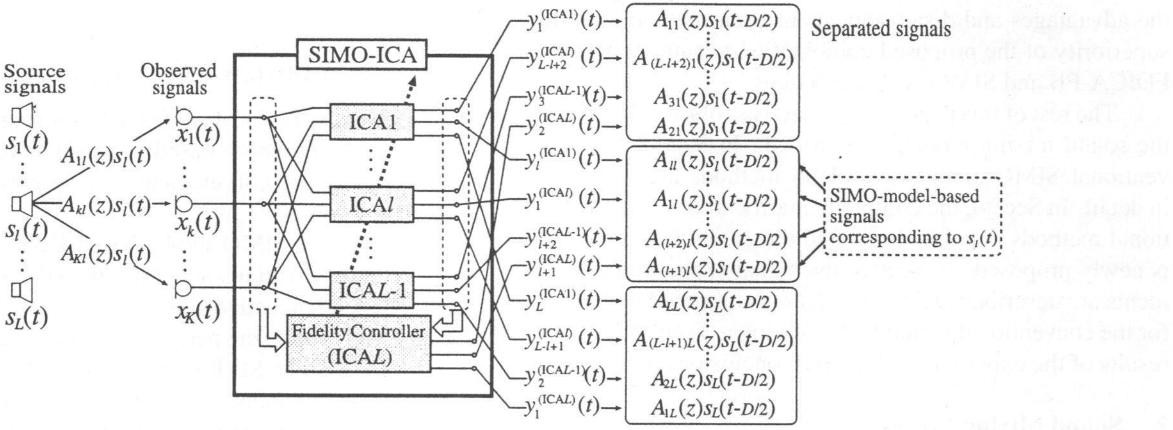

Source signals
$s_1(t)$
$s_l(t)$
$s_L(t)$
$A_{1l}(z)s_l(t)$
$A_{kl}(z)s_l(t)$
$A_{Kl}(z)s_l(t)$
Observed signals
$x_1(t)$
$x_k(t)$
$x_K(t)$
SIMO-ICA
ICA1
ICAl
ICAL-1
Fidelity Controller (ICAL)
$y_1^{(ICA1)}(t)$ → $A_{11}(z)s_1(t\text{-}D/2)$
$y_{L-l+2}^{(ICAl)}(t)$ → $A_{(L-l+2)1}(z)s_1(t\text{-}D/2)$
$y_3^{(ICAL-1)}(t)$ → $A_{31}(z)s_1(t\text{-}D/2)$
$y_2^{(ICAL)}(t)$ → $A_{21}(z)s_1(t\text{-}D/2)$
$y_l^{(ICA1)}(t)$ → $A_{ll}(z)s_l(t\text{-}D/2)$
$y_1^{(ICAl)}(t)$ → $A_{1l}(z)s_l(t\text{-}D/2)$
$y_{l+2}^{(ICAL-1)}(t)$ → $A_{(l+2)l}(z)s_l(t\text{-}D/2)$
$y_{l+1}^{(ICAL)}(t)$ → $A_{(l+1)l}(z)s_l(t\text{-}D/2)$
$y_L^{(ICA1)}(t)$ → $A_{LL}(z)s_L(t\text{-}D/2)$
$y_{L-l+1}^{(ICAl)}(t)$ → $A_{(L-l+1)L}(z)s_L(t\text{-}D/2)$
$y_2^{(ICAL-1)}(t)$ → $A_{2L}(z)s_L(t\text{-}D/2)$
$y_1^{(ICAL)}(t)$ → $A_{1L}(z)s_L(t\text{-}D/2)$
Separated signals
SIMO-model-based signals corresponding to $s_l(t)$

Fig. 2 Example of input and output relations in SIMO-ICA, where permutation $\boldsymbol{P}_l$ is given by (12).

bins. The optimal $\boldsymbol{W}(f)$ is obtained by, for example, the following iterative updating equation:

$$\boldsymbol{W}^{[i+1]}(f) = \eta\left[\boldsymbol{I} - \left\langle \boldsymbol{\Phi}(\boldsymbol{Y}(f,t))\boldsymbol{Y}^{\mathrm{H}}(f,t)\right\rangle_t\right]\boldsymbol{W}^{[i]}(f) + \boldsymbol{W}^{[i]}(f), \tag{4}$$

where $\langle\cdot\rangle_t$ denotes the time-averaging operator, $[i]$ is used to express the value of the i th step in the iterations, and η is the step-size parameter. In our research, we define the nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ as [7]:

$$\boldsymbol{\Phi}(\boldsymbol{Y}(f,t)) \equiv \left[e^{j\cdot\arg(Y_1(f,t))}, \cdots, e^{j\cdot\arg(Y_L(f,t))}\right]^{\mathrm{T}}, \tag{5}$$

where $\arg[\cdot]$ represents an operation to take the argument of the complex value. After the iterations, the permutation problem, i.e., indeterminacy in ordering sources, can be solved by [18].

Finally, in order to obtain the SIMO components, the separated signals are projected back onto the microphones by using the inverse of $\boldsymbol{W}(f)$ [13]. In this method, the following operation is performed.

$$Y_l^{(k)}(f,t) = \left\{\boldsymbol{W}(f)^{-1}[\overbrace{0,\cdots,0}^{l-1}, Y_l(f,t), \overbrace{0,\cdots,0}^{L-l}]^{\mathrm{T}}\right\}_k, \tag{6}$$

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

The FDICA-PB has the advantages that **(F1)** it is very fast and insensitive to the initial value in the iterative updating because the optimization of the separation filter given by (4) and the projection-back processing given by (6) are simple. That is, FDICA can simplify the complex problem of the separation filter optimization into the individual optimization of the separation matrix $\boldsymbol{W}(f)$ in each subband. There exist, however, the disadvantages that **(F2)** the inversion of $\boldsymbol{W}(f)$ often fails and yields harmful results because the invertibility of every $\boldsymbol{W}(f)$ cannot be guaranteed [19], and **(F3)** the circular convolution effect inherent in FDICA is likely to cause the deterioration of the separation performance.

3.3 SIMO-ICA [14], [15]

SIMO-ICA has recently been proposed by one of the authors as a means of obtaining SIMO-model-based signals directly in the ICA updating. The SIMO-ICA consists of $(L-1)$ TDICA parts and a *fidelity controller*, and each ICA runs in parallel under the fidelity control of the entire separation system (see Fig. 2). The separated signals of the l-th ICA $(l = 1, \cdots L-1)$ in SIMO-ICA are defined by

$$\boldsymbol{y}_{(\mathrm{ICA}l)}(t) = [y_k^{(\mathrm{ICA}l)}(t)]_{k1} = \sum_{n=0}^{D-1} \boldsymbol{w}_{(\mathrm{ICA}l)}(n)\boldsymbol{x}(t-n), \tag{7}$$

where $\boldsymbol{w}_{(\mathrm{ICA}l)}(n) = [w_{ij}^{(\mathrm{ICA}l)}(n)]_{ij}$ is the separation filter matrix in the l-th ICA, and D is the filter length.

Regarding the fidelity controller, we calculate the following signal vector, in which all elements are to be mutually independent:

$$\boldsymbol{y}_{(\mathrm{ICA}L)}(t) = \boldsymbol{x}(t-D/2) - \sum_{l=1}^{L-1} \boldsymbol{y}_{(\mathrm{ICA}l)}(t). \tag{8}$$

Hereafter, we regard $\boldsymbol{y}_{(\mathrm{ICA}L)}(t)$ as an output of a *virtual* "L-th" ICA, and define its *virtual* separation filter matrix as

$$\boldsymbol{w}_{(\mathrm{ICA}L)}(n) = \boldsymbol{I}\delta\left(n - \frac{D}{2}\right) - \sum_{l=1}^{L-1} \boldsymbol{w}_{(\mathrm{ICA}l)}(n), \tag{9}$$

where $\delta(n)$ is a delta function, i.e., $\delta(0) = 1$ and $\delta(n) = 0$ $(n \neq 0)$. The reason we use the term *virtual* here is that the L-th ICA does not have its own separation filters, unlike the other ICAs, and $\boldsymbol{w}_{(\mathrm{ICA}L)}(n)$ is subject to $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ $(l = 1, \cdots, L-1)$.

By transposing the second term $(-\sum_{l=1}^{L-1}\boldsymbol{y}_{(\mathrm{ICA}l)}(t))$ in the right-hand side into the left-hand side, we can show that (8) means a constraint to force the sum of all ICAs' output vectors $\sum_{l=1}^{L}\boldsymbol{y}_{(\mathrm{ICA}l)}(t)$ to be the sum of all SIMO components $[\sum_{l=1}^{L} A_{kl}(z)s_l(t-D/2)]_{k1}(= \boldsymbol{x}(t-D/2))$. Here, the delay of

$D/2$ is used to deal with nonminimum phase systems. Using (7) and (8), we can obtain the appropriate separated signals and maintain their spatial qualities as follows.

Theorem: 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, as

$$\boldsymbol{y}_{(\mathrm{ICA}l)}(t) = \mathrm{diag}\left[\boldsymbol{A}(z)\boldsymbol{P}_l^{\mathrm{T}}\right]\boldsymbol{P}_l\boldsymbol{s}(t-D/2), \tag{10}$$

where diag$[\boldsymbol{X}]$ is the operation for setting every off-diagonal element of the matrix $\boldsymbol{X}$ to zero, and $\boldsymbol{P}_l$ $(l = 1, \cdots, L)$ are exclusively-selected permutation matrices which satisfy

$$\sum_{l=1}^{L} \boldsymbol{P}_l = [1]_{ij}. \tag{11}$$

Regarding proof of the theorem, see [14].

Clearly, the solutions given by (10) provide necessary and sufficient SIMO components, $A_{kl}(z)s_l(t-D/2)$, for each l-th source. For example, one possibility is shown in Fig. 2 and this corresponds to

$$\boldsymbol{P}_l = [\delta_{im(k,l)}]_{ki}, \tag{12}$$

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

$$m(k,l) = \begin{cases} k+l-1 & (k+l-1 \le L) \\ k+l-1-L & (k+l-1 > L) \end{cases} \tag{13}$$

In this case, (10) yields

$$\begin{aligned} &\boldsymbol{y}_{(\mathrm{ICA}l)}(t) \\ &\quad = [A_{km(k,l)}s_{m(k,l)}(t-D/2)]_{k1} \quad (l=1,\cdots,L). \end{aligned} \tag{14}$$

In order to obtain (10), the natural gradient of the Kullback-Leibler divergence of (8) with respect to $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ should be added to the existing TDICA-based iterative learning rule [4] 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 SIMO-ICA is given as

$$\begin{aligned} &\boldsymbol{w}^{[i+1]}_{(\mathrm{ICA}l)}(n) \\ &= \boldsymbol{w}^{[i]}_{(\mathrm{ICA}l)}(n) - \alpha \sum_{d=0}^{D-1}\Bigg[\Bigg\{\text{off-diag}\left\langle \boldsymbol{\varphi}\left(\boldsymbol{y}^{[i]}_{(\mathrm{ICA}l)}(t)\right)\right. \\ &\quad \left.\boldsymbol{y}^{[i]}_{(\mathrm{ICA}l)}(t-n+d)^{\mathrm{T}}\right\rangle_t\Bigg\} \cdot \boldsymbol{w}^{[i]}_{(\mathrm{ICA}l)}(d) \\ &\quad - \Bigg\{\text{off-diag}\left\langle \boldsymbol{\varphi}\left(\boldsymbol{x}(t-\frac{D}{2}) - \sum_{l=1}^{L-1}\boldsymbol{y}^{[i]}_{(\mathrm{ICA}l)}(t)\right)\right. \\ &\quad \left. \cdot \left(\boldsymbol{x}(t-n+d-\frac{D}{2}) - \sum_{l=1}^{L-1}\boldsymbol{y}^{[i]}_{(\mathrm{ICA}l)}(t-n+d)\right)^{\mathrm{T}}\right\rangle_t\Bigg\} \\ &\quad \cdot \left(\boldsymbol{I}\delta(d-\frac{D}{2}) - \sum_{l=1}^{L-1}\boldsymbol{w}^{[i]}_{(\mathrm{ICA}l)}(d)\right)\Bigg], \end{aligned} \tag{15}$$

where off-diag$[\boldsymbol{X}]$ is the operation for setting every diagonal element of the matrix $\boldsymbol{X}$ to zero, α is the step-size parameter, and $\boldsymbol{\varphi}(\cdot)$ is the nonlinear vector function where the l-th element is set to be $\tanh(y_l(t))$. The initial values of $\boldsymbol{w}_{(\mathrm{ICA}l)}(n)$ for all l should be different.

The SIMO-ICA has the following advantage and disadvantage. **(T1)** This method is free from both the circular convolution effect and the invertibility of the separation filter matrix. **(T2)** Since the SIMO-ICA is based on TDICA which involves more complex calculations than FDICA, the convergence of the SIMO-ICA is very slow, and its sensitivity to the initial settings of separation filter matrices is very high.

4. Proposed Method

4.1 Motivation: Complementarity between FDICA-PB and SIMO-ICA

As described in the previous section, the two SIMO-model-based BSS methods have some disadvantages. However, we note that the advantages and disadvantages of FDICA-PB and SIMO-ICA are mutually complementary, i.e., (F2) and (F3) can be resolved by (T1), and (T2) can be resolved by (F1). In order to explicitly illustrate this complementarity, we carried out a preliminary experiment on SIMO-ICA's sensitivity to the initial value of the separation filter matrices. We artificially generated multiple distinct initial values of the separation filter matrices with various qualities by using the following equation:

$$\begin{aligned} \boldsymbol{W}^{[0]}_{(\mathrm{ICA}l)}(z) = {} & \beta\mathrm{diag}\left[\boldsymbol{A}(z)\boldsymbol{P}_l^{\mathrm{T}}\right]\boldsymbol{P}_l\boldsymbol{A}(z)^{-1}z^{-D/2} \\ & + (1-\beta)\boldsymbol{P}_l z^{-D/2}, \end{aligned} \tag{16}$$

where $\boldsymbol{W}^{[0]}_{(\mathrm{ICA}l)}(z)$ is the z-transform of $\boldsymbol{w}^{[0]}_{(\mathrm{ICA}l)}(t)$, and β (=0–1) is a parameter for controling the quality of the initial values; e.g., the ideal separation is achieved if $\beta = 1$, but the separation performance decreases as β decreases.

The experimental conditions are summarized in Table 1. To simulate the convolutive mixtures, clean speech samples are convolved with impulse responses recorded in an experimental room (see Fig. 3) in which the reverberation time (RT) is set at 150 ms. Two kinds of sentences spoken by two male and two female speakers are used as the source speech samples. Using these sentences, we obtain 12 combinations.

As an objective evaluation score, *SIMO-model accuracy* (SA) is used to indicate the degree of similarity (mean-squared-error) between the SIMO-model-based BSSs' out-

Table 1 Experimental conditions for signal separation (see also Fig. 3).

Number of Microphones	2
Interelement Spacing	4 cm
Number of Sound Sources	2
Sound Source Directions	−30° and 40°
Sampling Frequency	8 kHz
Speech Data Length	7.5 s
FFT Length in FDICA-PB	2048 samples
Filter Length in SIMO-ICA	2048 taps

puts and the original SIMO-model-based signals ($A_{kl}(z)s_l(t-D/2)$). The detailed calculation of SA is described in Appendix.

Figure 4 shows the results of SAs for FDICA-PB and SIMO-ICA with different initial values. We got 612 results with 12 combinations × 51 types of β, and plot the average of the results whose initial SAs are within the same range. From this figure, it is evident that the performances of SIMO-ICA are inferior to those of FDICA-PB under low-quality initial value conditions (0–14 dB), but SIMO-ICA can outperform FDICA-PB in particular when the initial value is improved over 14 dB. This result is highly consistent with (F1) and (T2). Also, this motivates us to propose a promising combination technique of FDICA-PB and SIMO-ICA, i.e., we can obtain accurate SIMO signals by using SIMO-ICA which follows a saturated FDICA-PB with sufficient iterative updating.

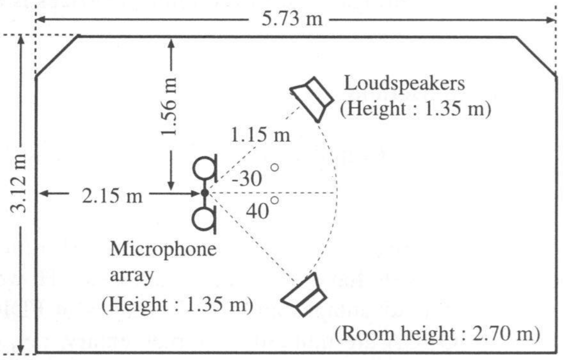


Fig. 3 Layout of reverberant room used in experiments.

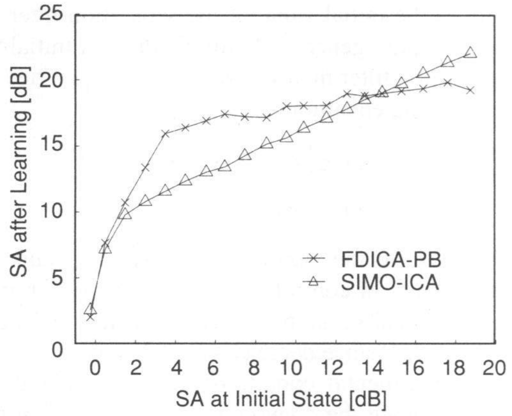


Fig. 4 SIMO-model accuracies of FDICA-PB and SIMO-ICA under different initial value conditions.

4.2 Combination Algorithm of FDICA-PB and SIMO-ICA

We propose a new multistage technique combining FDICA-PB and SIMO-ICA (see Fig. 5). The proposed multistage technique is conducted with the following steps.

Step 0: Set an arbitrary initial value of the separation matrix $\boldsymbol{W}(f)$ in FDICA. For example, an appropriate null-beamformer [8] can be used.

Step 1: Perform FDICA (see (4)) to separate the source signals to some extent with the fast- and robust-convergence advantage (F1).

Step 2: After the FDICA, we generate a specific initial value $\boldsymbol{w}^{[0]}_{(\mathrm{ICA}l)}(n)$ for SIMO-ICA to be performed in the next step, by using $\boldsymbol{W}(f)$ obtained from FDICA. This procedure is given by

$$\boldsymbol{w}^{[0]}_{(\mathrm{ICA}l)}(n)$$
$$= \mathrm{IFFT}\Big[\mathrm{diag}\Big[\boldsymbol{W}(f)^{-1}\boldsymbol{P}_l^{\mathrm{T}}\Big]\boldsymbol{P}_l\boldsymbol{W}(f)\Big], \qquad (17)$$

where $\boldsymbol{P}_l$ is set to be, e.g., (12), and IFFT[·] represents an inverse DFT with the time shift of $D/2$ samples.

Step 3: Perform SIMO-ICA (see (15)) to obtain resultant SIMO components with the advantage (T1).

Compared with the simple SIMO-ICA, this combination algorithm is not as sensitive to the initial value of the separation filter because FDICA is used for the estimation of a good initial value. Actually, the performance of the separation filter optimized by FDICA is high enough for SIMO-ICA to start leaning (see the result of FDICA-PB in

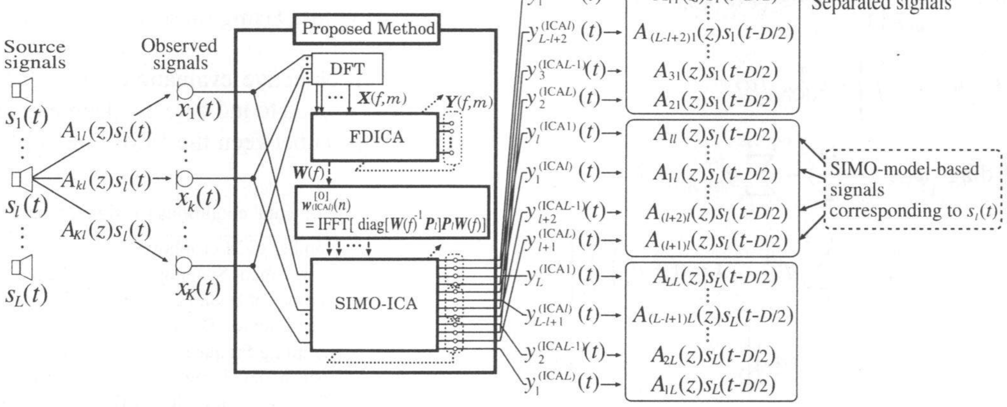

Source
signals
Observed
signals
$s_1(t)$
$s_l(t)$
$s_L(t)$
$A_{1l}(z)s_l(t)$
$A_{kl}(z)s_l(t)$
$A_{Kl}(z)s_l(t)$
$x_1(t)$
$x_k(t)$
$x_K(t)$
Proposed Method
DFT
$\boldsymbol{X}(f,m)$
$\boldsymbol{Y}(f,m)$
FDICA
$\boldsymbol{W}(f)$
$w^{[0]}_{(\mathrm{ICA}l)}(n)$
$= \mathrm{IFFT}[\ \mathrm{diag}[\boldsymbol{W}(f)^{-1}\boldsymbol{P}_l]\boldsymbol{P}_l\boldsymbol{W}(f)]$
SIMO-ICA
$y_1^{(\mathrm{ICA}1)}(t)$
$y_{L-l+2}^{(\mathrm{ICA}l)}(t)$
$y_3^{(\mathrm{ICA}L-1)}(t)$
$y_2^{(\mathrm{ICA}L)}(t)$
$y_l^{(\mathrm{ICA}1)}(t)$
$y_1^{(\mathrm{ICA}l)}(t)$
$y_{l+2}^{(\mathrm{ICA}L-1)}(t)$
$y_{l+1}^{(\mathrm{ICA}L)}(t)$
$y_L^{(\mathrm{ICA}1)}(t)$
$y_{L-l+1}^{(\mathrm{ICA}l)}(t)$
$y_2^{(\mathrm{ICA}L-1)}(t)$
$y_1^{(\mathrm{ICA}L)}(t)$
$A_{11}(z)s_1(t-D/2)$
$A_{(L-l+2)1}(z)s_1(t-D/2)$
$A_{31}(z)s_1(t-D/2)$
$A_{21}(z)s_1(t-D/2)$
$A_{ll}(z)s_l(t-D/2)$
$A_{1l}(z)s_l(t-D/2)$
$A_{(l+2)l}(z)s_l(t-D/2)$
$A_{(l+1)l}(z)s_l(t-D/2)$
$A_{LL}(z)s_L(t-D/2)$
$A_{(L-l+1)L}(z)s_L(t-D/2)$
$A_{2L}(z)s_L(t-D/2)$
$A_{1L}(z)s_L(t-D/2)$
Separated signals
SIMO-model-based
signals
corresponding to $s_l(t)$

Fig. 5 Input and output relations in proposed multistage method.

Fig. 4). Also, this technique offers the possibility of providing a more accurate separation result than the simple FDICA because the resultant quality of the output signal is determined by the separation ability of the SIMO-ICA starting from a good initial state.

5. Experiments and Results

5.1 Conditions for Experiment

The experimental conditions are the same as those provided in Table 1 and Fig. 3. The RT is set to 150 ms and 300 ms. Two kinds of sentences spoken by two male and two female speakers are used as the source speech samples. Using these sentences, we obtain 12 combinations. The initial value in all methods is fixed to null-beamformer whose directional null is steered to ±45° [8]. Note that the null-beamformer is commonly used as an ICA's initial value in several recent works on ICA-based BSS [16], [20], [21] because of the superiority to the traditional setting of the initial value such as random values. This may be due to the close relationship between ICA and the null-beamformer [22], which has reported that ICA with the small number of sensors often provides directional nulls against the undesired source signals.

5.2 Results

Figures 6 and 7 show the results of SAs for FDICA-PB, SIMO-ICA, and the proposed combination technique in all speaker combinations, for each of the reverberation conditions. In the results of the proposed combination technique, there exists a consistent improvement of SA compared with the results of FDICA-PB as well as those of the simple SIMO-ICA. At RT = 150 ms, the average score of the improvement is 8.3 dB over SIMO-ICA, and 2.9 dB over FDICA-PB. Also, at RT = 300 ms, the average score of the improvement is 5.1 dB over SIMO-ICA, and 2.9 dB over FDICA-PB. In these figures, we can also confirm that SIMO-ICA step in the proposed method can start leaning with the good initial separation filter whose performance is high enough for SIMO-ICA step to outperform simple FDICA-PB.

Figure 8 shows the sensitivity to the initial state of the proposed method. The experimental conditions are the same as those in Sect. 4.1. From this figure, we can confirm that the proposed method outperforms both of the conventional methods with any initial states and the proposed method is not sensitive to the initial state. On the basis of these results, we can conclude that the proposed combination technique can assist the SIMO-ICA in improving the separation performance, and successfully achieve the SIMO-model-based BSS under reverberant conditions.

5.3 Discussion on Combination Order

The previous section shows that the cascade connection of FDICA-PB and SIMO-ICA improves the separation performance. This is mainly due to the fact that the FDICA-PB in the first stage can provide a better separation filter matrix for SIMO-ICA in the second stage, and subsequently SIMO-ICA can improve the quality of the resultant separated signals. In this section, to discuss the validity of the proposed combination order, we compare the proposed combination with another combination in which SIMO-ICA is used in the first stage and FDICA-PB is used in the second stage (hereafter we designate this combination as "swapped combination").

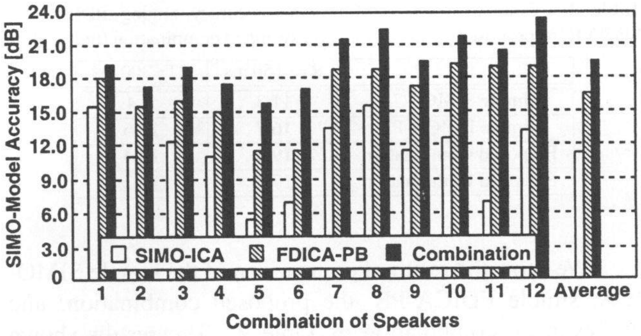


Fig. 6 Comparison of SIMO-model accuracy among conventional FDICA-PB, SIMO-ICA, and the proposed combination technique (RT is 150 ms).

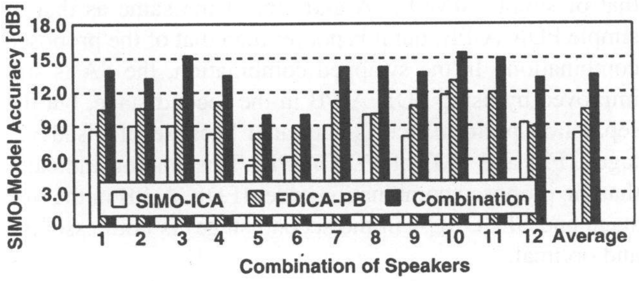


Fig. 7 Comparison of SIMO-model accuracy among conventional FDICA-PB, SIMO-ICA, and the proposed combination technique (RT is 300 ms).

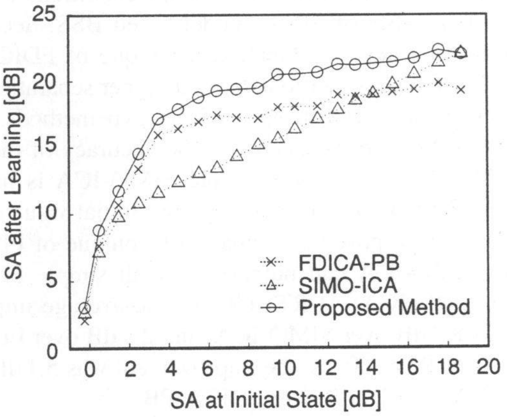


Fig. 8 SIMO-model accuracies of FDICA-PB and SIMO-ICA and proposed method under different initial value conditions.

Table 2 Comparison of SIMO-model accuracy among FDICA-PB, SIMO-ICA, proposed combination, and swapped combination (unit is dB).

	RT=150 ms	RT=300 ms
Simple SIMO-ICA	11.3	8.4
Simple FDICA-PB	16.7	10.6
Proposed Combination	**19.6**	**13.4**
Swapped Combination	17.1	11.2

We show the result of comparison of the simple SIMO-ICA, simple FDICA-PB, the proposed combination, and the swapped combination in Table 2. The results shown in Table 2 are the averages of 12 experiments with different combinations of speakers. The average SA of 17.1 dB (RT=150 ms) or 11.2 dB (RT=300 ms) is obtained in the swapped combination. This performance is still better than that of simple SIMO-ICA and almost the same as that of simple FDICA-PB, but it is poorer than that of the proposed combination. In the swapped combination, the SA is still improved by using FDICA-PB in the second stage, but the separation performance is saturated because of disadvantages (F2) and (F3) of FDICA-PB. This finding indicates that the proposed combination order (FDICA-PB in the first stage and SIMO-ICA in the second stage) is both essential and optimal.

6. Conclusion

In this paper, first, the conventional FDICA-PB and SIMO-ICA were compared under reverberant conditions to evaluate the feasibility of SIMO-model-based BSS. Secondly, we proposed a new combination technique of FDICA-PB and SIMO-ICA, in order to achieve a higher separation performance compared with each of the two methods. The experimental results revealed that the accuracy of the separated SIMO signals in the simple SIMO-ICA is inferior to that of FDICA-PB under low-quality initial value conditions, but the proposed combination technique of FDICA-PB and SIMO-ICA can outperform both simple FDICA-PB and SIMO-ICA. At RT=150 ms, the average improvement was 8.3 dB over SIMO-ICA, and 2.9 dB over FDICA-PB. Also, at RT=300 ms, the improvement was 5.1 dB over SIMO-ICA, and 2.9 dB over FDICA-PB.

Needless to say, the computional complexity of the proposed method is larger than those of simple FDICA-PB and simple SIMO-ICA because the proposed method includes both of them. The reduction of the computional cost still remains as an open problem for future study. Fortunately several kinds of real-time (low-computational-cost) ICA algorithms have been proposed (see, e.g., [23]) in recent works. In future, we should utilize such a fast algorithm in our multistage method to realize the speed-up.

Acknowledgement

The authors are grateful to Dr. Shoji Makino, Dr. Masato Miyoshi and Ms. Shoko Araki of NTT Corp., and Mr. Tsuyoki Nishikawa of NAIST for their fruitful discussions on this work. This work was partly supported by Core Research for Evolutional Science and Technology Program, "Advanced Media Technology for Everyday Living" of Japan Science and Technology Agency.

References

[1] S. Haykin, Unsupervised Adaptive Filtering, John Wiley & Sons, NY, 2000.

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

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

[4] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," Proc. Int. Workshop on ICA and BSS (ICA'99), pp.371–376, 1999.

[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech Audio Process., vol.8, no.3, pp.320–327, 2000.

[6] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," Proc. ICASSP2002, pp.1789–1792, 2002.

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

[8] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," EURASIP J. Applied Signal Processing, vol.2003, pp.1135–1146, 2003.

[9] J. Blauert, Spatial Hearing (revised edition), The MIT Press, Cambridge, MA, 1997.

[10] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind separation of binaural sound mixtures using SIMO-model-based independent component analysis," Proc. ICASSP2004, vol.IV, pp.113–116, 2004.

[11] H. Saruwatari, H. Yamajo, T. Takatani, T. Nishikawa, and K. Shikano, "Blind separation and deconvolution of MIMO system driven by colored inputs using SIMO-model-based ICA with information-geometric learning," Proc. IEEE Neural Network for Signal Processing Workshop 2003 (NNSP2003), pp.379–388, 2003.

[12] S. Ukai, H. Saruwatari, T. Takatani, and K. Shikano, "Blind source separation using SIMO-model-signal extraction and adaptive beamforming," Technical Report of IEICE, vol.EA2004-24, 2004.

[13] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," Proc. Int. Sympo. on Nonlinear Theory and its Application (NOLTA'98), vol.3, pp.923–926, 1998.

[14] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind separation of acoustic signals using SIMO-model-based ICA with information-geometric learning," Proc. Int. Workshop on Acoustic Echo and Noise Control, pp.251–254, 2003.

[15] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "High-fidelity blind source separation using SIMO-model-based ICA with information-geometric learning algorithm," IEEE Trans. Speech Audio Process. (in submitting).

[16] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," IEICE Trans. Fundamentals, vol.E86-A, no.4, pp.846–858, April 2003.

[17] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," Proc. International Conference on Independent Component Analysis and Blind Signal Separation, pp.722–727, 2001.

[18] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and

precise method for solving the permutation problem of frequency-domain blind source separation," Proc. Int. Sympo. on ICA and BSS, pp.505–510, 2003.

[19] T. Nishikawa, H. Saruwatari, and K. Shikano, "Stable learning algorithm for blind separation of temporally correlated acoustic signals combining multistage ICA and linear prediction," IEICE Trans. Fundamentals, vol.E86-A, no.8, pp.2028–2036, Aug. 2003.

[20] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband based blind source separation with appropriate processing for each frequency band," Proc. Int. Sympo. on ICA and BSS, pp.137–142, 2003.

[21] M. Furukawa, Y. Hioka, T. Ema, and N. Hamada, "Introducing new mechanism in the learning process of FDICA-based speech separation," Proc. Int. Workshop on Acoustic Echo and Noise Control, pp.291–294, 2003.

[22] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," IEEE Trans. Speech Audio Process., vol.11, no.2, pp.109–116, 2003.

[23] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," IEICE Trans. Fundamentals, vol.E87-A, no.8, pp.1941–1948, Aug. 2004.

Appendix: Calculation of SA

This section describes a calculation of SA under the specific assumption that the permutation matrices $\boldsymbol{P}_l$ $(l = 1, 2)$ are given by (12). If another permutation condition arises, the sound source number should be swapped. Note that the unit of all scores is the *decibel* (dB), but hereafter we omit the unit in equations.

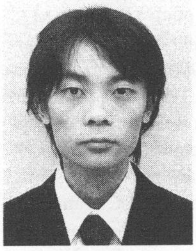
The SA for sound source 1 is defined as

$$\begin{aligned}\mathrm{SA}_1 &\\ &= 10\log_{10}\left(\frac{\left\langle|A_{11}(z)s_1(t-D/2)|^2\right\rangle_t}{\left\langle|y_1^{(\mathrm{ICA1})}(t)-A_{11}(z)s_1(t-D/2)|^2\right\rangle_t}\right)\\ &\quad+10\log_{10}\left(\frac{\left\langle|A_{21}(z)s_1(t-D/2)|^2\right\rangle_t}{\left\langle|y_2^{(\mathrm{ICA2})}(t)-A_{21}(z)s_1(t-D/2)|^2\right\rangle_t}\right).\end{aligned} \tag{A$\cdot$1}$$

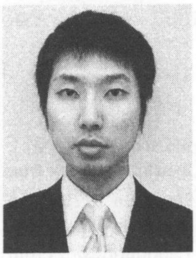
The SA for sound source 2 is defined as

$$\begin{aligned}\mathrm{SA}_2 &\\ &= 10\log_{10}\left(\frac{\left\langle|A_{22}(z)s_2(t-D/2)|^2\right\rangle_t}{\left\langle|y_2^{(\mathrm{ICA1})}(t)-A_{22}(z)s_2(t-D/2)|^2\right\rangle_t}\right)\\ &\quad+10\log_{10}\left(\frac{\left\langle|A_{12}(z)s_2(t-D/2)|^2\right\rangle_t}{\left\langle|y_1^{(\mathrm{ICA2})}(t)-A_{12}(z)s_2(t-D/2)|^2\right\rangle_t}\right).\end{aligned} \tag{A$\cdot$2}$$

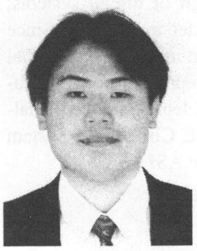
The resultant SA is an average of SA_1 and SA_2.



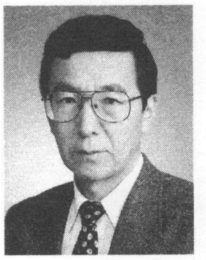
Satoshi Ukai was born in Shiga, Japan on 1980. He received the B.E. degree in electronic engineering from Kobe University in 2003. He is now a master-course student at Graduate School of Information Science, NAIST. His research interests include array signal processing and blind source separation. He is a member of the the Acoustical Society of Japan.



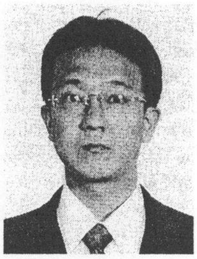
Tomoya Takatani was born in Hyogo, Japan on 1977. He received the B.E. degree in electronics from Doshisha University in 2001. He received the M.E. degree in information science from Nara Institute of Science and Technology in 2003. He is currently a Ph.D. candidate of Nara Institute of Science and Technology. His research interests include array signal processing and blind source separation. He is a member of the Acoustical Society of Japan.



Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO.,LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on the ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2000, and from TAF in 2004. He is a member of the IEEE, the VR Society of Japan, and the Acoustical Society of Japan.



Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialog system, speech enhancement, adaptive microphone array, and acoustic field reproduction. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990–1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984–1986, he was a visiting scientist in Carnegie Mellon University, where he was working on distance measures, speaker adaptation, and statistical language modeling. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001. He is a member of Information Processing Society of Japan, the Acoustical Society of Japan (ASJ), Japan VR Society, the Institute of Electrical and Electronics, Engineers (IEEE), and International Speech Communication Society.



Ryo Mukai received the B.S. and the M.S. degrees in information science from the University of Tokyo, Japan, in 1990 and 1992, respectively. He joined NTT in 1992. From 1992 to 2000, he was engaged in research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is engaged in research of blind source separation. His current research interests include digital signal processing and its applications. He is a member of the IEEE, ACM, the ASJ, and the IPSJ.

Hiroshi Sawada received the B.E., M.E. and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 1991, 1993 and 2001, respectively. In 1993, he joined NTT Communication Science Laboratories. From 1993 to 2000, he was engaged in research on the computer aided design of digital systems, logic synthesis and computer architecture. Since 2000, he has been engaged in research on signal processing, blind source separation for convolutive mixtures and independent component analysis. He received the best paper award of the IEEE Circuit and System Society in 2000. He is a member of the IEEE and the ASJ.