Blind Separation and Deconvolution for Convolutive Mixture of Speech Combining SIMO-Model-Based ICA and Multichannel Inverse Filtering

Hiroshi SARUWATARII, Member, Hiroaki YAMAJI', Nonmember, Tomoya TAKATANI', Tsuyoki NISHIKAWA1, Student Members, and Kiyohiro SHIKANO1, Member

SUMMARY We propose a new two-stage blind separation and deconvolution strategy for multiple-input multiple-output (MIMO)-FIR systems driven by colored sound sources, in which single-input multiple-output (SIMO)-model-based ICA (SIMO-ICA) and blind multichannel inverse filtering are combined. SIMO-ICA can separate the mixed signals, not intermonaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. After the separation by the SIMO-ICA, a blind deconvolution technique for the SIMO model can be applied even when each source signal is temporally correlated and the mixing system has a nonminimum phase property. The simulation results reveal that the proposed algorithm can successfully achieve separation and deconvolution of a convolutive mixture of speech, and outperforms a number of conventional ICA-based BSD methods.

key words: blind source separation, blind deconvolution, independent component analysis, SIMO model

1. Introduction

Blind separation and deconvolution (BSD) of sources is an unsupervised adaptive filtering 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 of the convolutive mixture [1]–[6] is that not only the source separation but also the deconvolution of the transmission channel characteristics are considered in the BSD framework. Owing to the attractive features of BSD, much attention has been paid to the BSD technique in various fields of signal processing such as digital communications systems, radar antenna systems, and image and acoustic signal processing systems. One promising example in acoustic signal processing is a hands-free speech recognition system involving a microphone array [7]–[9], where not only the interfering noise but also the distortion due to the acoustical transfer function must be reduced to maintain the high recognition accuracy. The main objective of this paper is to provide a feasible BSD method for wide-band acoustic signals, e.g., speech. Accordingly, we consider a multiple-input multiple-output (MIMO) system with finite-impulse response (FIR) filters as the transmission channels, and temporally corre-
SIMO-model-based blind deconvolution technique based on multichannel inverse filtering can be applied even when the mixing system is the nonminimum phase system and each source signal is temporally correlated. The BSD method based on the proposed two-stage idea can provide a more feasible performance for the separation and deconvolution of acoustic signals than the conventional BSD method. This can be confirmed from the simulation results, which are the world’s first successful examples of the BSD for two colored sources and two sensors without special constraints.

The rest of this paper is organized as follows. In Sect. 2, the sound mixing model and the conventional ICA-based BSD is explained. Also, the problem arising in the conventional method is mentioned. In Sect. 3, the proposed two-stage BSD algorithm is described in detail. First, SIMO-ICA is described, and secondly the SIMO-model-based deconvolution is explained. In Sects. 4 and 5, the BSD experiments based on the computer simulation are described, and the results are compared with those of the conventional method. Following a discussion on the results of the experiments, we give conclusions in Sect. 6.

2. Mixing Process and Conventional BSD

2.1 Mixing Process

In this study, the number of microphones is K and the number of multiple sound sources is L. Hereafter, we deal with discrete time series, and symbols t, n and d are used as the discrete time indexes. Disregarding an additive background noise, we can express the observed signals in which multiple source signals are mixed linearly as

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

where \( s(t) = [s_1(t), \cdots, s_L(t)]^T \) is the source signal vector, and \( x(t) = [x_1(t), \cdots, x_K(t)]^T \) is the observed signal vector. Also, \( a(n) \) is the mixing filter matrix with the length of \( N \), and \( A(z) \) is the z-transform of \( a(n) \); these are given as

\[ a(n) = [a_{ij}(n)]_{KL}, \]

\[ A(z) = [A_{ij}(z)]_{KL} = \left[ \sum_{n=0}^{N-1} a_{ij}(n) z^{-n} \right]_{KL}, \]

where \( z^{-1} \) is used as the unit-delay operator, i.e., \( z^{-n} \cdot x(t) = x(t-n), a_{ij}(n) \) 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.

Regarding the source signals and mixing systems, the following conditions and assumptions are considered.

(C1) The source signals \( s_l(t) \) (\( l = 1, \cdots, L \)) are mutually independent, and unknown.

(C2) Each source signal is temporally correlated (colored), but its coloration characteristics are unknown.

(C3) The mixing system \( A(z) \) is unknown, and probably has the nonminimum phase property. However, every column of \( A(z) \) is guaranteed not to have any common zeros in the z-plane. That is, the transfer function polynomials \( A_{11}(z), \cdots, A_{KL}(z) \) corresponding to the l-th sound source \( s_l(t) \) have no common zeros, and this holds in every l.

(C4) The order of the mixing system, \( N \), is unknown.

2.2 Conventional ICA-Based BSD and Its Problem

In the time-domain ICA (TDICA), the separated signal \( y(t) = [y_1(t), \cdots, y_L(t)]^T \) is expressed as

\[ y(t) = \sum_{n=0}^{D-1} w(n)x(t-n) = W(z)x(t), \]

where \( w(n) = [w_{ij}(n)]_{L \times L} \) is the separation filter matrix, \( W(z) \) is the z-transform of \( w(n) \), and \( D \) is the filter length of \( w(n) \).

In the ICA-based BSD framework assuming i.i.d. sources, Amari [15] proposed the holonomic TDICA algorithm which optimizes the separation filter by minimizing the Kullback-Leibler divergence (KLD) between the joint probability density function (PDF) of \( y(t) \) and the product of marginal PDFs of \( y_l(t) \). The iterative learning rule is given by

\[ w^{[j+1]}(n) = w^{[j]}(n) + \eta \sum_{d=0}^{D-1} \left[ I \delta(n-d) - \left( \varphi(y_l^{[j]}(t)) y_l^{[j]}(t-n+d) \right)^T \right] w^{[j]}(d), \]

where \( \eta \) is the step-size parameter, and the superscript \( [j] \) is used to express the value of the j-th step in the iterations, \( \delta(.) \) denotes the time-averaging operator and \( I \) is the identity matrix. Also, \( \delta(n) \) is a delta function, where \( \delta(0) = 1 \) and \( \delta(n) = 0 \) (\( n \neq 0 \)), and \( \varphi(.) \) is an appropriate nonlinear vector function, e.g., given as

\[ \varphi(y_l(t)) = [\tanh(y_l(t)), \cdots, \tanh(y_L(t))]^T. \]

Note that Amari’s original method [15] was an online learning algorithm, but (5) represents an off-line (batch) learning algorithm, where the update of \( w^{[j]}(n) \) in (5) does not depend on the time index \( t \). Throughout the paper, we only deal with the off-line algorithm because it is easy to treat and is the basis of the on-line algorithm.

Many conventional ICA-based BSD algorithms, e.g., Amari’s algorithm, often force 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). In this figure, e.g., the colored sound sources are represented by \( s_l(t) = H_{l}(z)r_l(t) \), where \( r_l(t) \) is the i.i.d. source and \( H_{l}(z) \) is the coloration filter. The outputs of the conventional ICA-based BSD result in \( r_l(t) = B_{l}(z)s_l(t) \), where \( B_{l}(z) = 1/H_{l}(z) \) is the distortion. Here note that we cannot identify and recover the target sound source \( s_l(t) \) blindly because the distortion \( B_{l}(z) \) is unknown in advance (see also Sect. 3.1).
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. 4.4.

3. Proposed Two-Stage BSD Algorithm

3.1 Motivation and Strategy of Two-Stage BSD

In this paper, we address the BSD problem for colored signals which are very common in many practical situations. However, this was considered as a tough problem because of the difficulty that how we can apply the deconvolution only to the mixing system without destroying the coloration characteristics of the sources.

Meanwhile, in our previous research, SIMO-ICA was proposed by one of the authors [24], and we showed that SIMO-ICA can separate the mixed colored signals into SIMO-model-based signals at the microphone points. This finding has motivated us to combine our SIMO-ICA and an existing blind deconvolution technique for the SIMO model [32]–[34]. The configuration of the proposed method is depicted in Fig. 2(a). In the proposed method, the separation-deconvolution problem is resolved into two stages, an SIMO-model-based separation stage and an SIMO-model-based deconvolution stage. Each of the separation/deconvolution problems can be solved efficiently using the following reasonable properties. (a) 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. (b) 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 for the SIMO model.

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 deconvolution. To illustrate the novelty of the proposed method, we hereinafter compare the proposed combination with a simple two-stage combination of a conventional monaural-output ICA [1]–[6] and a conventional single-channel deconvolution (see Fig. 2(b)).

The conventional ICAs generally supply the distorted version of source signals $B_i(z)s_i(t)$ ($i = 1, \cdots, L$), where $B_i(z)$ is an unknown arbitrary distortion filter. From the point of view in BSD problem, the distortion $B_i(z)$ should be removed in the next deconvolution stage. This corresponds to a single-channel blind deconvolution problem, but the problem can not be solved at all because of the following reasons.

- It is impossible to find the deconvolution filter which cancels $B_i(z)$ preserving the coloration of the source because we have no ways to blindly discriminate $B_i(z)$ and the coloration characteristics included in $s_i(t)$ via only single channel information $B_i(z)s_i(t)$.
- $B_i(z)$ is likely to become a nonminimum-phase system. Since the nonminimum-phase system is not invertible in theory, there are no valid filters to remove the distortion in the single-channel deconvolution framework [25], [26].

Thus the simple combination of the conventional ICA and the conventional single-channel deconvolution is not valid for solving the problem of BSD with colored inputs.

On the other hand, our proposed combination contains the special SIMO-model-based ICA in the first stage. The aim of the SIMO-ICA is to supply the specific SIMO signals with respect to each of sources, $\mathcal{A}_1(z)s_1(t)$, up to the possible delay of the filters. After having the SIMO components, we can identify each source signal by utilizing the blind SIMO-model-based deconvolution as follows.

- Multiple signals $A_1(z)s_1(t), \cdots, A_L(z)s_L(t)$ regarding the $l$-th source $s_l(t)$ have been obtained in SIMO-ICA. In addition, from the assumption (C3), $A_1(z), \cdots, A_L(z)$ have no common zeros, i.e., these transfer functions (polynomials) are distinct each other. From these facts, we can blindly identify $A_1(z), \cdots, A_L(z)$ by carefully finding the difference among $A_1(z)s_1(t), \cdots, A_L(z)s_L(t)$ because the difference is due to the distinct property of $A_i(z)$ and the common part is due to the source signal itself.
- After identification of $A_i(z)$, the source signal $s_i(t)$ can be recovered from the corresponding SIMO components $A_1(z)s_1(t), \cdots, A_L(z)s_L(t)$ even if each $A_i(z)$ is a nonminimum phase filter. Here note that although each $A_i(z)$ cannot be invertible, the whole SIMO system is guaranteed to be invertible in the multiple inverse filter theory under (C3) (i.e., $A_1(z), \cdots, A_L(z)$ have no common zeros).

In summary, the novelty of the proposed two-stage idea is due to the introduction of SIMO-model-based framework into both separation and deconvolution processes, and this offers a realization of the BSD. The detailed process of using the proposed algorithm is as follows.
3.2 First Stage: SIMO-ICA for Source Separation

In this stage, SIMO-ICA [24] is conducted for extracting the SIMO-model-based signals corresponding to each of sources. A brief explanation of the SIMO-ICA is given in the following. 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. 3). The separated signals of the \(l\)-th ICA \((l = 1, \cdots, L - 1)\) in SIMO-ICA are defined by

\[
y_{[\text{ICA}]}^{[l]}(t) = [y_{[\text{ICA}]}^{[l]}(t)]_{k1} = \sum_{n=0}^{D-1} w_{[\text{ICA}]}^{[l]}(n)x(t-n),
\]

where \(w_{[\text{ICA}]}^{[l]}(n) = [w_{[\text{ICA}]}^{[l]}(n)]_{ij}\) is the separation filter matrix in the \(l\)-th ICA.

Regarding the fidelity controller, we calculate the following signal vector \(y_{[\text{ICAL}]}(t)\), in which all elements are to be mutually independent,

\[
y_{[\text{ICAL}]}(t) = [y_{[\text{ICAL}]}^{[l]}(t)]_{k1} = x(t-D/2) - \sum_{l=1}^{L-1} y_{[\text{ICA}]}^{[l]}(t).
\]

Hereafter, we regard \(y_{[\text{ICAL}]}(t)\) 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 \(y_{[\text{ICAL}]}(t)\) is subject to \(w_{[\text{ICA}]}^{[l]}(n) (l = 1, \cdots, L - 1)\). By transposing the second term \((-\sum_{l=1}^{L-1} y_{[\text{ICA}]}^{[l]}(t))\) 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 \(\sum_{l=1}^{L-1} y_{[\text{ICA}]}^{[l]}(t)\) to be the sum of all SIMO components \(\sum_{l=1}^{L-1} A_{[\text{ICA}]}^{[l]}(z) s(t-D/2)\) \(= x(t-D/2)\). Here the delay of \(D/2\) is used to deal with non-minimum phase systems.

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

\[
y_{[\text{ICA}]}^{[l]}(t) = \text{diag} \left[ A_{[\text{ICA}]}^{[l]} \right] P_{[l]} s(t-D/2),
\]

where \(\text{diag}[X]\) is the operation for setting every off-diagonal element of the matrix \(X\) to zero, and \(P_{[l]}\) \((l = 1, \cdots, L)\) are exclusively-selected permutation matrices [27] which satisfy \(\sum_{l=1}^{L} P_{[l]} = [1]_{L}\). Regarding a proof for this, see [24]. Obviously the solutions given by (9) provide necessary and sufficient SIMO components, \(A_{[\text{ICA}]}^{[l]}(z)s(t-D/2)\), for each \(l\)-th source. For example, in the case of \(L = K = 2\), one possibility is given by
where

\[
\begin{align*}
P_1 &= I, \\
P_2 &= [1, 1] - I.
\end{align*}
\]

In order to obtain (9), the natural gradient [28] of Kullback-Leibler divergence of (8) with respect to \( w_{\text{ICAO}(l)}(n) \) should be added to the existing TDICA-based iterative learning rule [29] of the separation filter in the \( i \)-th ICA (\( i = 1, \ldots, L - 1 \)). The new iterative algorithm of the \( i \)-th ICA part (\( i = 1, \ldots, L - 1 \)) in SIMO-ICA is given as

\[
\begin{align*}
\mathbf{w}^{(i+1)}_{\text{ICAO}(l)}(n) &= \mathbf{w}^{(i)}_{\text{ICAO}(l)}(n) - \alpha \sum_{d=0}^{D-1} \left\{ \text{off-diag} \left( \varphi \left( y^{(i)}_{\text{ICAO}(l)}(t) \right) \right) \\
&\quad \cdot \mathbf{w}^{(i)}_{\text{ICAO}(l)}(d) \\
&\quad - \left\{ \text{off-diag} \left( \varphi \left( x(t) - \frac{D}{2} - \sum_{l=1}^{L-1} y^{(i)}_{\text{ICAO}(l)}(t) \right) \right) \\
&\quad \cdot \left( x(t - n + d - \frac{D}{2}) - \sum_{l=1}^{L-1} y^{(i)}_{\text{ICAO}(l)}(t - n + d) \right) \right\} \right\}
\end{align*}
\]

\[
\cdot \left( I \delta(t - D/2) - \sum_{l=1}^{L-1} \mathbf{w}^{(i)}_{\text{ICAO}(l)}(d) \right),
\]

where off-diag[X] is the operation for setting every diagonal element of the matrix X to zero, and \( \alpha \) is the step-size parameter. The initial values of \( \mathbf{w}_{\text{ICAO}(l)}(n) \) for all \( l \) should be different.

Note that there exists an alternative method [30] of obtaining the SIMO components in which the separated signals are projected back onto the microphones by using the inverse of \( \mathbf{W}(z) \). In this method, the following operation is performed after (4):

\[
y^{(i)}_k(t) = \left( \mathbf{W}(z)^{-1} [0, \ldots, 0, y_i(t), 0, \ldots, 0] \right)_{k},
\]

where \( y^{(i)}_k(t) \) represents the \( i \)-th resultant separated source signal which is projected back onto the \( k \)-th microphone, and \( [1]_{k} \) denotes the \( k \)-th element of the argument. This method is simpler than SIMO-ICA, but this inversion often fails and yields harmful results because the invertibility of every \( \mathbf{W}(z) \) cannot be guaranteed [31].

### 3.3 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. 4(a), where we deal with the case of \( K = L = 2 \). Note that this can be easily extended to the general case \( K > 2 \) by selecting the arbitrary two SIMO components from the SIMO-ICA’s outputs. In this process, the transfer functions, \( A_{11}(z) \) and \( A_{12}(z) \), can be estimated by a subchannel matching approach [32]-[34] 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 consider an example under (12), the output signal of this subchannel matching system is given by

\[
\begin{align*}
e_1(t) &= \hat{A}_{11}(z)y^{(1\text{ICA})}_1(t) - \hat{A}_{11}(z)y^{(1\text{ICA})}_2(t) \\
&= \left( \hat{A}_{11}(z)A_{11}(z) - \hat{A}_{11}(z)A_{12}(z) \right)s_1(t - D/2),
\end{align*}
\]

where \( \hat{A}_{i1}(z) \) is the estimated transfer function of the mixing process \( A_{ij}(z) \), and is defined using the estimated impulse responses \( \hat{a}_{i1}(n) \) as

\[
\hat{A}_{i1}(z) = \sum_{n=0}^{N-1} \hat{a}_{i1}(n)z^{-n}.
\]

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 (\( \gamma \) is an arbitrary constant):

\[
\hat{A}_{11}(z) = \gamma A_{11}(z),
\]

\[
\hat{A}_{12}(z) = \gamma A_{12}(z).
\]
Thus, the minimization of $\left\langle e_1(t)^2 \right\rangle$ 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 $\left\langle e_1(t)^2 \right\rangle$ is given as
\[
\left\langle e_1(t)^2 \right\rangle = \left\langle a_1^T y_1(t) y_1^T(t) a_1 \right\rangle = a_1^T \left[ y_1(t) y_1^T(t) \right] a_1 ,
\]
where
\[
a_1 = \begin{bmatrix} \hat{a}_{21}(0), \hat{a}_{21}(1), \cdots, \hat{a}_{21}(N-1), \\
-\hat{a}_{11}(0), -\hat{a}_{11}(1), \cdots, -\hat{a}_{11}(N-1) \end{bmatrix}^T ,
\]
\[
y_1(t) = \begin{bmatrix} y_1^{(ICA1)}(t), y_1^{(ICA1)}(t-1), \cdots, y_1^{(ICA1)}(t-N+1), \\
y_2^{(ICA2)}(t), y_2^{(ICA2)}(t-1), \cdots, y_2^{(ICA2)}(t-N+1) \end{bmatrix}^T .
\]
(21)

On the basis of this, we solve the following minimization problem:
\[
\min_{a_1} a_1^T \left[ y_1(t) y_1^T(t) \right] a_1 , \quad \text{subject to} \quad \|a_1\| = \text{const}. 
\]
(22)
The optimal vector $a_1$ in (22) can be derived as the specific eigenvector which corresponds to the minimum eigenvalue of the matrix $\left[ y_1(t) y_1^T(t) \right]$.

Next, regarding the blind channel identification corresponding to another sound source $s_2(t)$, the optimal vector $a_2$ can be derived as the eigenvector which corresponds to the minimum eigenvalue of the matrix $\left[ y_2(t) y_2^T(t) \right]$, where $a_2$ and $y_2(t)$ are defined as
\[
a_2 = \begin{bmatrix} \hat{a}_{22}(0), \hat{a}_{22}(1), \cdots, \hat{a}_{22}(N-1), \\
-\hat{a}_{12}(0), -\hat{a}_{12}(1), \cdots, -\hat{a}_{12}(N-1) \end{bmatrix}^T ,
\]
\[
y_2(t) = \begin{bmatrix} y_2^{(ICA2)}(t), y_2^{(ICA2)}(t-1), \cdots, y_2^{(ICA2)}(t-N+1), \\
y_1^{(ICA1)}(t), y_1^{(ICA1)}(t-1), \cdots, y_1^{(ICA1)}(t-N+1) \end{bmatrix}^T .
\]
(24)

Finally, we can estimate the multichannel inverse filters, $G_{i1}(z)$ and $G_{i2}(z)$ for $\hat{A}_{11}(z)$ and $\hat{A}_{21}(z)$, and $G_{i2}(z)$ and $G_{i2}(z)$ for $\hat{A}_{12}(z)$ and $\hat{A}_{22}(z)$, based on the multiple-input/output inverse theorem (MINT) [26]. The reason we use the MINT is that the single-channel inverse filter can not be applied in this case because each $\hat{A}_{i}(z)$ is a nonminimum phase system and is not invertible. In the MINT method, the exact inverse of the room acoustics can be uniquely determined, even when $\hat{A}_{i}(z)$ has the nonminimum phase properties, if $\hat{A}_{i}(z)$ does not have any common zeros in the $z$-plane. The optimal multichannel inverse filters $G_{i}(z)$ are derived by solving the following diophantine equations:
\[
G_{i1}(z)\hat{A}_{i1}(z) + G_{i2}(z)\hat{A}_{i2}(z) = 1 ,
\]
(25)
\[
G_{i2}(z)\hat{A}_{i2}(z) + G_{i2}(z)\hat{A}_{i2}(z) = 1 ,
\]
(26)
where $G_{i}(z)$ are FIR filters of less than the $N$-th order. The recovered source signals $\hat{s}_1(t-D/2)$ can be given by (see Fig. 4(b))
\[
\hat{s}_1(t-D/2) = G_{i1}(z)y_1^{(ICA1)}(t) + G_{i2}(z)y_2^{(ICA2)}(t) ,
\]
(27)
\[
\hat{s}_2(t-D/2) = G_{i2}(z)y_1^{(ICA1)}(t) + G_{i2}(z)y_2^{(ICA2)}(t) .
\]
(28)

The accurate estimation of the filter length $N$ of the room impulse responses is indispensable for improving the system identification performance. Various methods for filter-length estimation based on the MDL, AIC, [34] and utilization of the modeling error from MINT filtering [33] have been presented, and we use Furuya’s method in this work.

3.4 Discussion on Filter Length and Identifiability

In this section, we first describe the filter length used in the proposed method. Following the presentation of results, we discuss the identifiability of the proposed BSD.

Using (7) and (8), we can express the recovered source signals (27) and (28) as
\[
\hat{s}_1(t-D/2) = [G_{i1}(z), G_{i2}(z)] \begin{bmatrix} W_{11}^{(ICA1)}(z) \\
W_{21}^{(ICA1)}(z) \\
W_{12}^{(ICA2)}(z) \\
W_{22}^{(ICA2)}(z) \end{bmatrix} \cdot x(t) ,
\]
(29)
\[
\hat{s}_2(t-D/2) = [G_{i2}(z), G_{i2}(z)] \begin{bmatrix} W_{11}^{(ICA1)}(z) \\
W_{21}^{(ICA1)}(z) \\
W_{12}^{(ICA2)}(z) \\
W_{22}^{(ICA2)}(z) \end{bmatrix} \cdot x(t) ,
\]
(30)
where $W_i(z)$ is the element of the $z$-transform of $w^{(ICA1)}(n)$, and $W_i(z)$ is the element of the $z$-transform of $b(n-D/2-w^{(ICA1)}(n))$. Thus, we obtain the entire input-output relation,
\[
[\hat{s}_1(t-D/2), \hat{s}_2(t-D/2)]^T = \hat{W}(z)x(t) ,
\]
(31)
where
\[
\hat{W}(z) = \begin{bmatrix} G_{i1}(z)W_{11}^{(ICA1)}(z) + G_{i2}(z)W_{21}^{(ICA2)}(z) ,
G_{i2}(z)W_{12}^{(ICA1)}(z) + G_{i2}(z)W_{22}^{(ICA2)}(z) ,
G_{i1}(z)W_{11}^{(ICA1)}(z) + G_{i2}(z)W_{21}^{(ICA2)}(z) ,
G_{i2}(z)W_{12}^{(ICA1)}(z) + G_{i2}(z)W_{22}^{(ICA2)}(z) \end{bmatrix} .
\]
(32)

Obviously, $\hat{W}(z)$ is the resultant separation filter matrix, and is represented as a square (2 x 2) polynomial matrix with a finite order of less than $D + N - 1$. Here, $N$ corresponds to the length of the multichannel inverse filter $G_{i}(z)$, and is automatically determined in accordance with the length.
of $A(z)$. On the other hand, $D$, the length of the separation filter $W_{ij}^{(ICA)}(z)$ in SIMO-ICA, can be arbitrarily set by the user.

Many previous studies [22],[23],[35] have indicated that the source identifiability in the case of $L = K$ cannot hold using only an FIR-type separation filter matrix if we do not assume special constraints. That is, at least $K = L + 1$ sensors should be required for identifying the $L$ sources. Therefore, in theory, the proposed BSD cannot identify the sources perfectly because the entire BSD procedure is regarded as the square FIR-type separation filter matrix. Since the deconvolution for the SIMO model in the second stage can be performed without errors, it is concluded that the SIMO-ICA only provides the SIMO components allowing a few residuals.

In practice, however, we can reduce the residuals by setting the separation filter length $D$ to be sufficiently long in the SIMO-ICA; this can be shown in the next simulation. Thus, the SIMO-model-based signals are approximately reproduced in this case. Overall, the identifiability almost holds under the assumption that we are allowed to use the long FIR filters in SIMO-ICA as well as (C1)–(C4). This finding is well supported by the simulation results described in Sect. 4.4. It should be enhanced that the simulation results are the world’s first evidence of success in the colored-source BSD under $L = K$ condition without additional assumptions, as far as we know.

4. Simulation in Artificial Mixing Condition

4.1 Conditions for Experiment

The mixing filter matrix $A(z)$ is taken to be

$$A_{11}(z) = 1 - 0.7z^{-1} - 0.3z^{-2}, \tag{33}$$
$$A_{12}(z) = z^{-1} + 0.7z^{-2} + 0.4z^{-3}, \tag{34}$$
$$A_{12}(z) = z^{-1} + 0.7z^{-2} + 0.4z^{-3}, \tag{35}$$
$$A_{12}(z) = 1 + 0.7z^{-1} - 0.3z^{-2}, \tag{36}$$

The impulse responses $a_{ij}(n)$ corresponding to $A_{ij}(z)$ are shown in Fig. 5. Two sentences spoken by two male speakers are used as the original speech samples $s(t)$. The sampling frequency is 8 kHz and the length of speech is limited to 7 seconds. The number of iterations in ICA is 15000.

We conduct two kinds of experiments as follows.

**Experiment 1:** We evaluate the accuracy of the reproduced SIMO-model-based signals in SIMO-ICA by changing the length of the separation filter, $D$, from 4 to 64 taps. The step-size parameter $\alpha$ was set in the range of $1 \times 10^{-6}$–$2 \times 10^{-6}$, and we select the optimum step size which gives the best performance.

**Experiment 2:** We compare five methods as follows.

(a) The conventional ICA-based BSD [15] (Amari’s BSD) given by (5),
(b) Amari’s BSD with the projection procedure (14) onto the microphones (Amari’s BSD+P),
(c) Amari’s BSD+P with deconvolution [33] for the SIMO components (Amari’s BSD+PD),
(d) SIMO-ICA only, and
(e) the proposed BSD.

The step-size parameter $\alpha$ is $2 \times 10^{-6}$ in the SIMO-ICA, and $\eta$ is $1 \times 10^{-6}$ in Amari’s BSD; these were the optimas which provide the best performances. The length of the separation filter, $D$, is set to 64 taps.

4.2 Objective Evaluation Scores

Three objective evaluation scores are defined as follows.

First, in Experiment 1, **SIMO-model accuracy** (SA) is used to indicate a degree of similarity between the SIMO-ICA’s outputs and the original SIMO-model-based signals. The detailed calculation of SA is described in Appendix A.

Secondly, in Experiment 2, **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. Thus, NRR is a part BSD evaluation score only in terms of the separation performance, and is not a final BSD score. The

![Fig. 5](image-url) Elements of mixing matrix $a(n)$ used in simulation, which are given by inverse z-transform of (33)-(36).
SNRs are calculated under the assumption that the speech signal of the undesired speaker is regarded as noise. The detailed calculation of NRR for each of the five methods (a)–(e) is summarized in Appendix B.

Thirdly, in Experiment 2, cepstral distortion (CD) [36] is used as the indication of both separation and deconvolution performances. In this study, we defined the CD as the distance between the spectral envelope of the original source signal $s(t-D/2)$ and that of the separated output. The 16th-order Mel-scaled cepstrum [37] based on the smoothed FFT spectrum is used. The CD gives the final BSD score, and will be decreased to zero if the separation-deconvolution processing is performed perfectly.

4.3 Results and Discussion in Experiment 1

Figures 6 and 7 show the elements of a composite filter matrix of $w_{jICA}(n)$ and $a(n)$. The composite filter matrix represents the whole system characteristics of the mixing and separation processes in SIMO-ICA; this is defined as

$$ [\hat{h}_{ij}^{(ICA)}(n)]_{ij} = e^{D/2} \sum_{d=0}^{D-1} w_{jICA}(d)a(n-d), \quad (37) $$

![Fig. 8 SIMO-model accuracy (SA) of the SIMO-ICA with different filter lengths D. The definition of SA is given by (A-1) and (A-2). The SA is used as the indication of a degree of similarity between the SIMO-ICA's outputs and the original SIMO-model-based signals.](image)

**Fig. 6** Elements of composite filter matrix $[\hat{h}_{ij}^{(ICA)}(n)]_{ij} = e^{D/2} \sum_{d=0}^{D-1} w_{jICA}(d)a(n-d)$ for different filter lengths $D$. This represents the whole system characteristics of the mixing and separation processes in the first ICA part of the SIMO-ICA.

**Fig. 7** Elements of composite filter matrix $[\hat{h}_{ij}^{(ICA)}(n)]_{ij} = e^{D/2} \sum_{d=0}^{D-1} w_{jICA}(d)a(n-d)$ for different filter lengths $D$. This represents the whole system characteristics of the mixing and separation processes in the second ICA part of the SIMO-ICA.
Table 1 Simulation results for different methods under artificial mixing condition.

<table>
<thead>
<tr>
<th>Method</th>
<th>NRR [dB]</th>
<th>CD [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_1(t)$</td>
<td>$s_2(t)$</td>
</tr>
<tr>
<td>(a) Amari's BSD</td>
<td>27.1</td>
<td>20.2</td>
</tr>
<tr>
<td>(b) Amari's BSD+P</td>
<td>13.2</td>
<td>12.7</td>
</tr>
<tr>
<td>(c) Amari's BSD+P+D</td>
<td>10.1</td>
<td>7.7</td>
</tr>
<tr>
<td>(d) SIMO-ICA</td>
<td>49.7</td>
<td>38.8</td>
</tr>
<tr>
<td>(e) Proposed BSD</td>
<td>42.5</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Fig. 9 Long-term averaged power spectra of (a) sound source 1 and (b) sound source 2 in original sources and Amari's BSD. The short-term analysis with the 256-th order FFT is applied to each speech segment, and all of the short-term spectra are averaged along with the whole data.

$$ \left[ \tilde{R}^{(ICA2)}_{ij}(n) \right]_{jj} = z^{D/2} \sum_{d=0}^{D-1} w_{(ICA2)}(d) a(n-d), $$

where $z^{D/2}$ is used to cancel the time delay of $D/2$ in $w_{(ICA2)}(n)$ for centering the impulse responses in each of the figures. From these figures, it is confirmed that the composite filter matrix for the first ICA, $\tilde{R}^{(ICA1)}_{ij}(n)_{jj}$, becomes $\text{diag}(a(n))$ as the length of the separation filter, $D$, is increased to more than the length of the mixing system. Also, the composite filter matrix for the second ICA, $\tilde{R}^{(ICA2)}_{ij}(n)_{jj}$, becomes off-$\text{diag}(a(n))$ as the length of the separation filter is increased. This obviously indicates that each ICA part in the SIMO-ICA can work so as to separately extract each SIMO component.

Figure 8 shows the results of SA, where the SA increases monotonically as the length of the separation filter, $D$, becomes large. In particular, the SA of more than 35 dB, which is sufficiently accurate for the following deconvolution process, is achieved when the filter length is set to 64 taps. Thus, the SIMO-ICA can reproduce the SIMO-model-based signals using the sufficiently long filter. This result supports the discussion on the identifiability of the proposed method as described in Sect. 3.4.

4.4 Results and Discussion in Experiment 2

Table 1 shows the results of NRR and CD for different methods. Regarding the NRR, it is evident that the conventional methods can separate the source signals to some extent, but their separation performances are inferior to those of SIMO-ICA and two-stage BSD.

Regarding the CD, first, it is evident that the CD of Amari's BSD is very high, i.e., the resultant speech is whitened by the decorrelation in the conventional method. To visualize this finding, we show the long-term averaged power spectrum of the resultant speech obtained by Amari's BSD in Fig. 9. The short-term analysis with the 256-th order FFT is applied to each speech segment, and all of the short-term spectra are averaged along with the whole data. This
figure indicates that the power spectrum in Amari’s BSD is flatter than that of the original source in terms of the spectral envelope.

Next, the result of Amari’s BSD+P in Table 1 shows a certain improvement in CD with the projection onto the microphones. However, Amari’s BSD+PD fails to achieve BSD. This implies that Amari’s BSD+P could not produce accurate SIMO components.

Finally, regarding the results of the proposed BSD, there is a considerable reduction in CD as shown in Table 1. This indicates that the proposed BSD algorithm can successfully achieve separation and deconvolution of a convolutive mixture of temporally correlated signals. This can also be confirmed from the results of the power spectra in Fig. 10; the power spectra of the original source and the proposed BSD are almost the same, up to the gain indeterminacy due to $\gamma$ in (17) and (18).

5. Simulation Using Measured Impulse Responses

5.1 Conditions for Experiment

The mixing filter matrix $A(z)$ is taken to be real head related transfer functions (HRTFs) which were measured by the Center for Image Processing and Integrated Computing (CIPIc), University of California, Davis. The public-domain CIPIc HRTF database [38] was measured with a KEMAR dummy head in an anechoic room. We chose the impulse responses of the HRTFs in four azimuths (see Fig. 11(a)), and down-sampled them from 44100 to 8000 Hz as shown in Fig. 11(b). The elevation is set to 0 degrees. In this experiment, $\theta_1$ (direction of source 1) is fixed to $-30$ degrees, and $\theta_2$ (direction of source 2) is set to 30, 55, or 80 degrees. Hereafter, these systems are designated as $L_{\theta_1}R_{\theta_2}$, e.g., “L$-30$R55.” Two sentences spoken by two male speakers are used as the original speech samples $s(t)$. The sampling frequency is 8 kHz and the length of speech is limited to 30 seconds. The length of the separation filter is set to be 512 taps.

5.2 Results and Discussion

Tables 2–4 show the results of NRR and CD for different methods and different mixing conditions (L$-30$R30, L$-30$R55, and L$-30$R80). From the results of NRR, the separation performance of the conventional BSD (Amari’s BSD) is comparable to that of the proposed method, and almost all of them are over 30 dB. Accordingly the conventional BSD and the proposed method are both effective as

| Table 2 | Simulation results for different methods with measured impulse responses (L$-30$R30). |
|-----------------|-----------------|-----------------|
|                | NRR [dB]        | CD [dB]         |
|                | $s_1(t)$        | $s_2(t)$        |
| (a) Amari’s BSD | 27.1            | 33.6            |
| (b) Amari’s BSD+P | 10.7            | 15.2            |
| (c) Amari’s BSD+PD | 16.4            | 28.1            |
| (d) SIMO-ICA    | 51.0            | 51.0            |
| (e) Proposed BSD | 39.9            | 24.7            |

| Table 3 | Simulation results for different methods with measured impulse responses (L$-30$R55). |
|-----------------|-----------------|-----------------|
|                | NRR [dB]        | CD [dB]         |
|                | $s_1(t)$        | $s_2(t)$        |
| (a) Amari’s BSD | 31.3            | 33.6            |
| (b) Amari’s BSD+P | 13.2            | 20.1            |
| (c) Amari’s BSD+PD | 22.4            | 30.4            |
| (d) SIMO-ICA    | 37.1            | 37.2            |
| (e) Proposed BSD | 38.6            | 43.3            |

| Table 4 | Simulation results for different methods with measured impulse responses (L$-30$R80). |
|-----------------|-----------------|-----------------|
|                | NRR [dB]        | CD [dB]         |
|                | $s_1(t)$        | $s_2(t)$        |
| (a) Amari’s BSD | 30.4            | 34.9            |
| (b) Amari’s BSD+P | 11.9            | 28.4            |
| (c) Amari’s BSD+PD | 0.5             | 25.9            |
| (d) SIMO-ICA    | 50.5            | 52.1            |
| (e) Proposed BSD | 32.9            | 33.0            |

![Fig. 11](image) (a) Locations of sources and dummy head in the simulation (top view). These systems are called $L_{\theta_1}R_{\theta_2}$, e.g., L$-30$R55. (b) examples of measured impulse responses of HRTFs used in this paper.
far as only the separation performance is concerned. As for the distortion of the separated speech, which is an important issue from the practical point of view, there is a considerable difference between the two methods. It is evident that the CDs of Amari’s BSD as well as Amari’s BSD+P and Amari’s BSD+PD are obviously high under every mixing condition. On the other hand, the CDs of the proposed method are consistently less than 3 dB regardless of the mixing condition. It can be asserted that the proposed BSD works effectively and achieves the BSD to some extent.

6. Conclusion

We proposed a new BSD framework in which the SIMO-ICA and blind multichannel inverse filtering are efficiently combined. The 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 the SIMO-ICA, we can introduce the SIMO-model-based blind channel identification and multichannel inverse filtering technique. In order to evaluate its effectiveness, a separation-deconvolution experiments were carried out using two microphones and two speech sources under both artificial and measured mixing condition. The experimental results revealed that the conventional ICA-based BSD includes adverse spectral distortion due to an inherent whitening effect, and the spectral distortion can be considerably reduced using the proposed two-stage BSD algorithm. Therefore, we can conclude that the proposed BSD algorithm is applicable to high-fidelity sound recovery processing.

As described in Sect. 3, there is a possibility that the proposed method can deal with the case of $K = L > 2$ in theory. However, only the results for $K = L = 2$ were shown in this paper. Therefore, we should perform further study for $K = L > 2$, and the method for estimating the number of sources should also be developed. In addition, the applicability to more longer reverberation cases and robustness against the background noise remain as open problems for future study.

Acknowledgement

All of the authors thank CIPIC in University of California, Davis, for their HRTF database which was very useful in our research. This work was partly supported by CREST (Core Research for Evolutionary Science and Technology) in Japan.

References


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


### Appendix A: Calculation of SA

This section describes a calculation of SA under the specific assumption that the permutation matrices $P_l$ ($l = 1, 2$) 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

$$SA_1 = 10 \log_{10} \left( \frac{\langle A_{11}(z)s_1(t-D/2)^2 \rangle}{\langle |y_1^{ICA}(t)| - A_{11}(z)s_1(t-D/2)^2 \rangle} \right) + 10 \log_{10} \left( \frac{\langle A_{21}(z)s_1(t-D/2)^2 \rangle}{\langle |y_2^{ICA}(t)| - A_{21}(z)s_1(t-D/2)^2 \rangle} \right).$$  \hspace{1cm} (A-1)

Also, the SA for sound source 2 is defined as

$$SA_2 \begin{align*}
&= 10 \log_{10} \left( \frac{\langle A_{22}(z)s_2(t-D/2)^2 \rangle}{\langle |y_2^{ICA}(t)| - A_{22}(z)s_2(t-D/2)^2 \rangle} \right) \\
&+ 10 \log_{10} \left( \frac{\langle A_{12}(z)s_2(t-D/2)^2 \rangle}{\langle |y_1^{ICA}(t)| - A_{12}(z)s_2(t-D/2)^2 \rangle} \right). \end{align*} \hspace{1cm} (A-2)

### Appendix B: Calculation of NRR

This section describes the calculation of NRR under the specific assumption that the permutation matrices $P_l$ ($l = 1, 2$) 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.

Since the separation procedures of the five methods (a)–(e) in Experiment 2 are different to each other, the calculation of NRR in each method is also different. The NRR for each of the five methods is described in detail.

(a) Amari's BSD:

The NRR for sound source $l$ ($l = 1, 2$) is defined as

$$\text{NRR}_l = \text{OSNR}_l - \text{ISNR}_l,$$  \hspace{1cm} (A-3)

where OSNR and ISNR are the output SNR and the input SNR, respectively; these are given by

$$\text{OSNR}_l = 10 \log_{10} \left( \frac{\langle H_l(z)s_l(t)^2 \rangle}{\langle |H_l(z)s_l(t)|^2 \rangle} \right),$$  \hspace{1cm} (A-4)

$$\text{ISNR}_l = 10 \log_{10} \left( \frac{\langle |A_{ll}(z)s_l(t)|^2 \rangle}{\langle |A_{ll}(z)s_l(t)|^2 \rangle} \right),$$  \hspace{1cm} (A-5)

where $l \neq l'$. Also, $H_{ll}(z)$ is the element in the $i$-th row and the $j$-th column of the matrix $H(z) = W(z)A(z)$ where $W(z)$ is the z-transform of $w(n)$ given by (5).

(b) Amari's BSD+P:

The NRR for sound source $l$ ($l = 1, 2$) is defined as

$$\text{NRR}_l = \frac{1}{2} \sum_{k=1}^{2} \text{OSNR}^{(k)}_l - \text{ISNR}^{(k)}_l,$$  \hspace{1cm} (A-6)

where OSNR and ISNR are the output SNR and the input SNR for the $k$-th microphone, respectively; these are given by

$$\text{OSNR}^{(k)}_l = 10 \log_{10} \left( \frac{\langle H_l^{(k)}(z)s_l(t)^2 \rangle}{\langle |H_l^{(k)}(z)s_l(t)|^2 \rangle} \right),$$  \hspace{1cm} (A-7)

$$\text{ISNR}^{(k)}_l = 10 \log_{10} \left( \frac{\langle |A_{ll}(z)s_l(t)|^2 \rangle}{\langle |A_{ll}(z)s_l(t)|^2 \rangle} \right),$$  \hspace{1cm} (A-8)
where $l \neq l'$. Also, $H_{i,j}(z)$ is the element in the $i$-th row and the $j$-th column of the matrix $H(z)$ where

$$H_{i,j}(z) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} W(z)A(z).$$

(Eq. 9)

$$H_{i,j}(z) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} W(z)A(z).$$

(Eq. 10)

(c) Amari's BSD+PD:

We define the NRR as being the same as (A.3)–(A.5) by substituting $W(z)$ with $\tilde{W}(z)$ with

$$\tilde{W}(z) = \begin{bmatrix} G_{11}(z), G_{21}(z) \\ G_{12}(z), G_{22}(z) \end{bmatrix} H_{i,j}(z).$$

(Eq. 11)

(d) SIMO-ICA:

The NRR for sound source $l$ ($l = 1, 2$) is defined as

$$\text{NRR}_l = \frac{1}{2} \sum_{k=1}^{2} \left( \text{OSNR}_{l}^{\text{ICA}k} - \text{ISNR}_{l}^{\text{ICA}k} \right),$$

(Eq. 12)

where $\text{OSNR}_{l}^{\text{ICA}k}$ and $\text{ISNR}_{l}^{\text{ICA}k}$ are the output SNR and the input SNR for ICA, respectively; these are given by

$$\text{OSNR}_{l}^{\text{ICA}1} = 10 \log_{10} \left( \frac{\left| H_{l}^{\text{ICA}1}(z) s_{l}(t) \right|^2}{\left| H_{l}^{\text{ICA}2}(z) s_{l}(t) \right|^2} \right),$$

(Eq. 13)

$$\text{ISNR}_{l}^{\text{ICA}1} = 10 \log_{10} \left( \frac{\left| A_{l}^{\text{ICA}1}(z) s_{l}(t) \right|^2}{\left| A_{l}^{\text{ICA}2}(z) s_{l}(t) \right|^2} \right),$$

(Eq. 14)

$$\text{OSNR}_{l}^{\text{ICA}2} = 10 \log_{10} \left( \frac{\left| H_{l}^{\text{ICA}2}(z) s_{l}(t) \right|^2}{\left| H_{l}^{\text{ICA}1}(z) s_{l}(t) \right|^2} \right),$$

(Eq. 15)

$$\text{ISNR}_{l}^{\text{ICA}2} = 10 \log_{10} \left( \frac{\left| A_{l}^{\text{ICA}2}(z) s_{l}(t) \right|^2}{\left| A_{l}^{\text{ICA}1}(z) s_{l}(t) \right|^2} \right).$$

(Eq. 16)

where $l \neq l'$. Also, $H_{i}^{\text{ICA}l}(z)$ is the element in the $i$-th row and the $j$-th column of the matrix $H^{\text{ICA}l}(z) = W^{\text{ICA}l}(z)A(z)$ where $W^{\text{ICA}l}(z)$ is the $k$-th separation filter matrix obtained in the SIMO-ICA.

(e) Proposed BSD:

We define the NRR as being the same as (A.3)–(A.5) by substituting $W(z)$ with $\tilde{W}(z)$ given by (32).
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 the 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.