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

Hiroaki Yamajo, Hiroshi Saruwatari, Tomoya Takatani, Tsuyoki Nishikawa, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology 8916-5, Takayama-cho, Ikoma-shi, Nara, Japan
hiroa-ya@is.aist-nara.ac.jp

Abstract
We propose a novel two-stage blind separation and deconvolution (BSD) algorithm for a convolutional mixture of speech, in which a new 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 into monaural source signals but into SIMO-model-based signals from independent sources as they are at the microphones. After SIMO-ICA, a simple blind deconvolution technique for the SIMO model can be applied even when each source signal is temporally correlated. The simulation results reveal that the proposed method can successfully achieve the separation and deconvolution for a convolutional mixture of speech.

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. In the BSD framework, not only the source separation but also the deconvolution of the transmission channel characteristics are considered. For the BSD based on independent component analysis (ICA), various methods have been proposed to deal with the separation and deconvolution for the convolutional mixture of independent, identically distributed (i.i.d.) source signals [1, 2]. 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 in many practical acoustic mixtures of sound signals such as speech. 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.

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: new blind separation technique using a Single-Input Multiple-Output (SIMO)-model-based ICA (SIMO-ICA) 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. 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. 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 simulation results reveal that the proposed method can successfully achieve the separation and deconvolution for a convolutional mixture of speech.

2. Mixing process and conventional BSD
In this study, the number of microphones is $K$ and the number of multiple sound sources is $L$. The observed signals in which multiple source signals are mixed linearly are expressed as

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

where $s(t) = [s_1(t), \ldots, s_L(t)]^T$ is the source signal vector, and $x(t) = [x_1(t), \ldots, x_K(t)]^T$ 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

\[
a(n) = [ak_1(n)]_{K1},
\]

and

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

where $z^{-1}$ is used as the unit-delay operator, i.e., $z^{-n} x(t) = x(t-n)$. $a_{kl}(n)$ is the impulse response between the $k$-th microphone and the $l$-th sound source, and $[X]_j$ 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.

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

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

where $w(n)$ 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 [1] proposed the holonomic TDICA algorithm which optimizes the separation filter by minimizing the Kullback-Leibler divergence between the joint probability density function (PDF) of $y(t)$ and the product of marginal PDFs of $y_i(t)$. The iterative learning rule is given by

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

where $f_\delta(n - d)$ is the $\delta$-function between $n$ and $d$. This equation is used for updating the separation filter $w^{[j]}(n)$ at the $j$-th iteration. The proposed method can successfully achieve the separation and deconvolution for a convolutional mixture of speech.
where $\eta$ is the step-size parameter, the superscript $[j]$ is used to express the value of the $j$-th step in the iterations. $\phi()$ denotes the time-averaging operator, and $I$ is the identity matrix. $\delta(n)$ is a delta function, where $\delta(0) = 1$ and $\delta(n) = 0$ ($n \neq 0$). $\varphi()$ is the nonlinear vector function, e.g., the $l$-th element is set to $\tanh(y(t))$.

Many conventional ICA-based BSD algorithms, e.g., Amari's algorithm, often force the separated signals to be 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. 4.2.

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. In the proposed method, the separation/deconvolution problems can be solved efficiently using the following reasonable assumption and properties.

(A1) The assumption of the mutual independence among the acoustic source signals usually holds, and consequently, this can be utilized in the SIMO-ICA-based separation.

(A2) 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.

The detailed process using the proposed algorithm is as follows.

3.1. First stage: SIMO-ICA for source separation

In this stage, a new blind separation method using SIMO-ICA is conducted. 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

$$y_{ICA}(t) = [s_k(t)]_{k=1} = \sum_{n=0}^{D/2-1} y_{ICA}(n)x(t-n) = W_{ICA}(z)x(t),$$

where $W_{ICA}(n)$ is the separation filter matrix in the $l$-th ICA, and $W_{ICA}(x)$ is the $z$-transform of $W_{ICA}(n)$. Regarding the fidelity controller, we introduce the following cost function to be minimized,

$$\left\langle \left \| \sum_{l=1}^{L} y_{ICA}(t) - \bar{x}(t-D/2) \right \|^2 \right \rangle_t,$$

where $\| \cdot \|$ is the Euclidean norm of vector $x$. Using (6) and (7), we can obtain the appropriate separated signals and maintain their spatial qualities as follows.

**Theorem:** If the independent sound sources are separated by (6), and simultaneously (7) is minimized to be zero, then the output signals converge on unique solutions, up to the permutation, as

$$y_{ICA}(t) = \text{diag}[A(z)P_l^T]P_l \bar{s}(t-D/2),$$

where $P_l$ ($l=1, \ldots, L$) are exclusively-selected permutation matrices which satisfy

$$\sum_{l=1}^{L} P_l = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$ (9)

**Proof of Theorem:** The necessity is obvious. The sufficiency is shown below. Let $D_l(z)$ ($l=1, \ldots, L$) be arbitrary diagonal polynomial matrices and $Q_l$ be arbitrary permutation matrices. The general expression of the $l$-th ICA’s output is given by

$$y_{ICA}(t) = D_l(z)Q_l \bar{s}(t-D/2),$$

If $Q_l$ are not exclusively-selected matrices, i.e., $\sum_{l=1}^{L} Q_l \neq \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then there exists at least one element of $\sum_{l=1}^{L} y_{ICA}(t)$ which does not include all components of $s_l(t-D/2)$ ($l=1, \ldots, L$). This obviously makes the cost function (7) be nonzero because the observed signal vector $x(t-D/2)$ includes all components of $s_l(t-D/2)$ in each element. Accordingly, $Q_l$ should be $P_l$ specified by (9), and we obtain

$$y_{ICA}(t) = D_l(z)P_l \bar{s}(t-D/2),$$

In (11) under (9), the arbitrary diagonal matrices $D_l(z)$ can be substituted by $\text{diag}[B(z)P_l^T]$, where $B(z) = [B_{ij}(z)]_{ij}$ is a single arbitrary matrix, because all diagonal entries of $\text{diag}[B(z)P_l^T]$ for all $l$ are also exclusive. Thus,

$$y_{ICA}(t) = \text{diag}[B(z)P_l^T]P_l \bar{s}(t-D/2),$$

and consequently

$$\sum_{l=1}^{L} y_{ICA}(t) = \sum_{l=1}^{L} B_{kl}(z)s_k(t-D/2).$$

Substitution of (13) in (7) leads to the following equation.

$$\left\langle \left \| \sum_{l=1}^{L} B_{kl}(z)s_k(t-D/2) \right \|^2 \right \rangle_t = 0.$$

where we used the relation, $(s_k(t-D/2)s_{l'}(t-D/2))_t = 0$ ($l \neq l'$). Since $(s_k(t-D/2))_t$ are positive, the cost function given by (14) becomes zero if and only if $B_{kl}(z) = A_{kl}(z)$ for all $k$ and $l$. Thus, (12) results in (8). This completes the proof of Theorem.

Obviously the solutions given by (8) provide necessary and sufficient SIMO components, $A_{kl}(z) \bar{s}_l(t-D/2)$, for each $l$-th
source. There, however, is an arbitrariness in a selection of \( P_1 \). For example, one possible selection is shown in Fig. 2 and this corresponds to

\[
P_1 = \{ \delta_{m(k,l)} \}_{k=1}^{1},
\]

where \( \delta_{ij} \) is Kronecker’s delta function, and

\[
m(k, l) = \begin{cases} 
  k + l - 1 & \text{if } k + l - 1 \leq L \\
  k + l - 1 - L & \text{if } k + l - 1 > L 
\end{cases}
\]

In this case, (8) yields

\[
y_{ICA}(t) = [A_{km}(z)s_m(t - D/2)]_{k=1}^{1}.
\]

In order to obtain (8), the gradient of (7) with respect to \( w_{ICA}(n) \) should be added to the iterative learning rule of the separation filter. The natural gradient [1] of (7) is given as

\[
\frac{\partial}{\partial w_{ICA}(n)} \left\{ \| \sum_{l=1}^{L} y_{ICA}(t) - x(t - D/2) \| \|^2 \right\} = \frac{1}{2} \left( \sum_{l=1}^{L} y_{ICA}(t) - x(t - D/2) \right)^T W_{ICA}(z) W_{ICA}(z)^T \cdot w_{ICA}(d).
\]

By combining (18) with the nonholonomic TDICA [3], we can obtain a new iterative algorithm in the \( l \)-th ICA of SIMO-ICA as

\[
w_{ICA}^{(l), i+1}(n) = w_{ICA}^{(l), i}(n) - \alpha \sum_{d=0}^{D-1} \left\{ \text{off-diag} \left( \varphi(y_{ICA}^{(l), i}(t)) y_{ICA}^{(l), i}(t - n + d) \right)^T \right\} \cdot \left( \sum_{l=1}^{L} y_{ICA}^{(l), i}(t) - x(t - D/2) \right) + \beta \left( \sum_{l=1}^{L} y_{ICA}^{(l), i}(t) - x(t - D/2) \right)^T \right\} \cdot w_{ICA}^{(l), i}(d),
\]

where \( \alpha \) and \( \beta \) are the step-size parameters; \( \alpha \) is for the control of the total update quantity and \( \beta \) is for fidelity control. In (19), updating of \( w_{ICA}(n) \) for all \( l \) should be simultaneously performed in parallel in terms of \( l \) because each iterative equation is associated with the others via \( \sum_{l=1}^{L} y_{ICA}^{(l), i}(t) = \sum_{l=1}^{L} W_{ICA}^{(l), i}(z)x(t) \). Also, the initial values of \( w_{ICA}(n) \) for all \( l \) should be different. If not, each ICA has the same set of inputs and will produce the same outputs. This results in an undesired solution. However, if we use different initial values, then the convergence on the appropriate SIMO solution is guaranteed by the simultaneous minimization of (6) and (7).

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, \( \max_{\delta l} \langle y_{i}^{(l)}(t)y_{k'}^{(l')}(t - \delta l) \rangle \), where \( l \neq l' \) and \( k \neq k' \). The large value of the correlation indicates that \( y_{i}^{(l)}(t) \) 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 \( s_1(t) \), where we deal with the case of \( K = L = 2 \). Note that this can be easily extended to the general case \( (K > 2) \) by picking up the arbitrary two SIMO components from SIMO-ICA’s outputs. In this process, the transfer functions, \( A_{11}(z) \) and \( A_{21}(z) \), can be estimated by the conventional subchannel matching approach [4, 5] in 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. The subchannel matching approach can work even for the temporally correlated signal. Regarding the blind channel identification corresponding to another sound source \( s_2(t) \), we can estimate \( A_{12}(z) \) and \( A_{22}(z) \) using the same approach.

Finally, we determine the multichannel inverse filters, \( G_{11}(z) \) and \( G_{21}(z) \) for the estimated \( A_{11}(z) \) and \( A_{21}(z) \), and \( G_{12}(z) \) and \( G_{22}(z) \) for the estimated \( A_{12}(z) \) and \( A_{22}(z) \), based on the multiple-input/output inverse theorem (MINT) [5]. In the MINT method, the exact inverse of the room acoustics can be uniquely determined, even when \( A_{kl}(z) \) has the nonminimum phase properties, if \( A_{kl}(z) \) does not have any common zeros in the z-plane. For example, the recovered signals \( \hat{s}_1(t) \) under (17) are

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

\[
\hat{s}_2(t) = G_{12}(z) y_1^{(2)}(t) + G_{22}(z) y_2^{(1)}(t),
\]

where \( \sum_{k=1}^{K} G_{kl}(z) A_{kl}(z) = 1 \).
4. Simulations

4.1. Conditions for experiment

The mixing filter matrix \( A(z) \) is taken to be \( A_{11}(z) = 1 - 0.7z^{-1} - 0.3z^{-2} \), \( A_{12}(z) = -0.1 + 0.7z^{-2} + 0.4z^{-3} \), \( A_{21}(z) = z^{-1} + 0.7z^{-2} + 0.4z^{-3} \), and \( A_{22}(z) = 1 - 0.7z^{-1} - 0.3z^{-2} \). 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 length of the separation filter, \( D \), is set to be 64. The number of iterations in ICA is 15000. The length of the deconvolution filter is blindly estimated by using an existing Furuya’s method [5].

In this experiment, two objective evaluation scores are defined as described below. First, noise reduction rate (NRR) [6], 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. 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 \( s_i(t - D/2) \) and that of the separated output. The 40th-order Mel-scaled cepstrum based on the smoothed FFT spectrum is used. The CD will be decreased to zero if the separation-deconvolution processing is performed perfectly.

4.2. Results and discussion

We compare three methods as follows: (a) conventional holonomic ICA (ICA-based BSD) [1] given by (5), (b) conventional nonholonomic ICA [3] given by (19) with setting \( \beta=\alpha \), and (c) proposed two-stage BSD. In SIMO-ICA, the step-size parameter \( \alpha \) is 2 \( \times 10^{-6} \) and \( \beta \) is 6 \( \times 10^{-4} \). Also, \( \eta \) is 1 \( \times 10^{-6} \) in the holonomic ICA, and \( \alpha \) is 1 \( \times 10^{-6} \) in the nonholonomic ICA; these were optima which provide the best performances.

When the channel identification was performed in the second stage, the proposed method could blindly estimate the length of \( a(n) \) at four taps successfully by using an existing Furuya’s method [5] for SIMO model.

Figures 3(a) shows the results of NRR for different methods. From this figure, it is evident that the separation performance of the holonomic ICA is too poor, however those of the proposed method and the nonholonomic ICA are high and comparable as far as the only separation performance is concerned. As for the distortion of the separated speech, which is an important issue from the practical viewpoint, there is a considerable difference between these methods, and this will be discussed in the next.

Figures 3(b) shows the results of CD. First, it is evident that the CD of the holonomic ICA is obviously high, i.e., the resultant speech is whitened by the decorrelation in the conventional method. Next, the result of the nonholonomic ICA shows that there are still some distortions in the separated signals. Finally, regarding the results of the proposed method, there is a considerable reduction of CD. This fact shows that the proposed BSD algorithm can successfully achieve the separation and deconvolution for a convolutive mixture of temporally correlated signals.

5. Conclusion

We proposed a new BSD framework in which SIMO-ICA and blind multichannel inverse filtering are efficiently combined. In order to evaluate its effectiveness, a separation-deconvolution experiment was carried out assuming 2 microphones and 2 speech sources. The simulation results revealed that the conventional ICA-based method includes adverse spectral distortion due to the inherent whitening effect, and the spectral distortion can be considerably reduced by using the proposed two-stage BSD algorithm.

6. Acknowledgement

This work was partly supported by CREST (Core Research for Evolutional Science and Technology) in Japan.

7. References