DETERMINED SOURCE SEPARATION USING THE SPARSITY OF IMPULSE RESPONSES

Yuki Takahashi *Daichi Kitahara* *Koichiro Matsuura* *Akira Hirabayashi*

Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

ABSTRACT

In this paper, we propose an over-determined sound source separation method considering the sparsity of impulse responses. Conventional methods, including *independent low-rank matrix analysis (ILRMA)*, have mainly focused on design of realistic sound generation models, but the separation performance is sometimes not improved due to the incorrectness of the generation models and convergence to some poor local minimum. In the proposed method, we utilize a prior information on the *mixing process*, i.e., the sparsity of impulse responses, to determine the *demixing matrices*. Numerical experiments using publicly available impulse responses demonstrate that the proposed method based on ILRMA with supervised bases can robustly obtain better results compared to the standard and the supervised ILRMAs.

***Index Terms*—** Over-determined source separation, independent low-rank matrix analysis, impulse response, supervised learning.

1. INTRODUCTION

Source separation is an estimation problem of source signals from observed mixed signals. In particular, if information on each source signal and the mixing process is almost none, this problem is called *blind source separation (BSS)* [1]–[7]. This paper supposes the situation that signals are sounds and the mixing process is the convolution of unknown impulse responses. In this situation, most existing works convert the convolution in the time domain into the multiplication in the frequency domain due to easier handling of the later model. In the *over-determined cases* (i.e., the number of microphones $\geq$ the number of sources), major BSS methods such as *frequency-domain independent component analysis (FDICA)* [3], *independent vector analysis (IVA)* [4], and *independent low-rank matrix analysis (ILRMA)* [5] estimate each source sound by applying the *demixing matrix*, which is the inverse mapping of the mixing process, to the observed sounds. In these methods, the demixing matrix is computed on the basis of a carefully designed sound generation model considering the statistical independence and the super-Gaussianity. For example in ILRMA [5], a *time-varying complex Gaussian distribution* is adopted, where the variance is adaptively updated with *nonnegative matrix factorization (NMF)* [8]–[11], as a sound generation model. ILRMA obtains better separation results compared to FDICA and IVA for audio signals.

As mentioned above, the major BSS methods try to improve the separation performance by changing the sound generation model, but the separation performance is sometimes not improved due to the incorrectness of the model and convergence to a poor local minimum. On the other hand, if a microphone array is used for the observation and the direction of arrival of each source sound is known, then we can estimate the mixing process and obtain good separation results [12]–[14]. However, this technique is not available if the microphone array is not used or the direction of arrival of each sound is unknown.

This work was supported in part by JSPS Grants-in-Aid Grant Number for Early-Career Scientists JP19K20361 (e-mail: d-kita@fc.ritsumei.ac.jp).

In this paper, we propose to use the property that the energy of the impulse response gathers in the early part of the time domain, i.e., each impulse response is *sparse in a wide sense* [15], [16], as a prior information on the mixing process. The proposed method computes the demixing matrices and the impulse responses alternately, and obtains better demixing matrices by considering the consistency for the estimated sparse impulse responses. In addition, by assuming that all sources are known musical instruments, we extend the proposed method to a supervised version based on the supervised NMF [17]–[19]. Numerical experiments show that the proposed method can improve the performance of ILRMA both in blind and supervised situations.

2. OVER-DETERMINED BLIND SOURCE SEPARATION

2.1. Problem Formulation

Let N (≥ 2) and M ($\geq N$) be the numbers of sources and microphones, respectively. $s_n[t]$ and $x_m[t]$ denote sounds generated from the nth source and observed by the mth microphone at time t. By letting $h_{m,n}[\tau]$ be an impulse response of length T from the nth source to the mth microphone, the observed sound $x_m[t]$ is expressed as

$$x_m[t] = \sum_{n=1}^{N} (h_{m,n} * s_n)[t] = \sum_{n=1}^{N} \sum_{\tau=0}^{T-1} h_{m,n}[\tau] s_n[t-\tau], \quad (1)$$

where $*$ denotes the convolution operator. We define $\hat{s}_{i,j,n} \in \mathbb{C}$ by applying a short-time Fourier transform (STFT) to the sound $s_n[t]$ as

$$\hat{s}_{i,j,n} := \sum_{\tau=0}^{L-1} \psi[\tau] s_n[(j-1)\eta + \tau] \exp\left(-\imath \frac{2\pi(i-1)\tau}{L}\right) \quad (2)$$

and similarly define $\hat{x}_{i,j,m} \in \mathbb{C}$, where L is the frame length, η is the frame shift, $\psi[\tau]$ is a window function, $\imath \in \mathbb{C}$ is the imaginary unit, $i = 1, 2, \ldots, L$ is the frequency index, and $j = 1, 2, \ldots, J$ is the frame index. We define $\hat{a}_{i,m,n} \in \mathbb{C}$ as a sample, at $\omega_i := \frac{2\pi(i-1)}{L}$, of the discrete-time Fourier transform of the impulse response $h_{m,n}[\tau]$

$$\hat{a}_{i,m,n} := \sum_{\tau=0}^{T-1} h_{m,n}[\tau] \exp\left(-\imath \frac{2\pi(i-1)\tau}{L}\right). \quad (3)$$

If the frame length L is sufficiently longer than the impulse response length T, the convolution of (1) in the time domain can be approximated, for each frame, by the multiplication in the frequency domain

$$\boldsymbol{x}_{i,j} = A_i \boldsymbol{s}_{i,j} + \boldsymbol{\epsilon}_{i,j} \quad (i = 1, 2, \ldots, I;\, j = 1, 2, \ldots, J), \quad (4)$$

where vectors $\boldsymbol{s}_{i,j}$, $\boldsymbol{x}_{i,j}$, $\boldsymbol{\epsilon}_{i,j}$ and matrices A_i are defined by

$$\begin{cases} \boldsymbol{s}_{i,j} := (\hat{s}_{i,j,1}, \hat{s}_{i,j,2}, \ldots, \hat{s}_{i,j,N})^\top \in \mathbb{C}^N \\ \boldsymbol{x}_{i,j} := (\hat{x}_{i,j,1}, \hat{x}_{i,j,2}, \ldots, \hat{x}_{i,j,M})^\top \in \mathbb{C}^M \\ \boldsymbol{\epsilon}_{i,j} := (\hat{\epsilon}_{i,j,1}, \hat{\epsilon}_{i,j,2}, \ldots, \hat{\epsilon}_{i,j,M})^\top \in \mathbb{C}^M \\ \boldsymbol{a}_{i,n} := (\hat{a}_{i,1,n}, \hat{a}_{i,2,n}, \ldots, \hat{a}_{i,M,n})^\top \in \mathbb{C}^M \\ A_i := (\boldsymbol{a}_{i,1}, \boldsymbol{a}_{i,2}, \ldots, \boldsymbol{a}_{i,N}) \in \mathbb{C}^{M \times N} \end{cases}$$

and $\hat{\epsilon}_{i,j,m} \in \mathbb{C}$ denotes the model error whose value becomes larger as the impulse response length T becomes longer. Note that in (4) we consider the frequency index i only from 1 to $I := \lfloor \frac{L+2}{2} \rfloor$ because $\hat{s}_{L-i+2,j,n} = \hat{s}^*_{i,j,n}$, $\hat{x}_{L-i+2,j,m} = \hat{x}^*_{i,j,m}$, $\hat{a}_{L-i+2,m,n} = \hat{a}^*_{i,m,n}$, and $\hat{\epsilon}_{L-i+2,j,m} = \hat{\epsilon}^*_{i,j,m}$ hold for $i = 2, 3, \ldots, I$, where $\hat{s}^*_{i,j,n}$ denotes the complex conjugate of $\hat{s}_{i,j,n}$. Blind source separation (BSS) is an estimation problem[1] of $\boldsymbol{s}_{i,j}$ from $\boldsymbol{x}_{i,j}$ in the situation where information on the sound $\boldsymbol{s}_{i,j}$ and the *mixing matrix* A_i is almost none.

In the over-determined cases where $M \geq N$ holds,[2] major BSS methods [3]–[5] estimate $\boldsymbol{s}_{i,j}$ by applying the *demixing matrix*

$$W_i := (\boldsymbol{w}_{i,1}, \boldsymbol{w}_{i,2}, \ldots, \boldsymbol{w}_{i,N})^{\mathrm{H}} := (A_i^{\mathrm{H}} A_i)^{-1} A_i^{\mathrm{H}} \in \mathbb{C}^{N \times M}, \quad (5)$$

which is the *pseudo-inverse* $A_i^{\dagger}$ of A_i, to the observed vector $\boldsymbol{x}_{i,j}$ as

$$\begin{aligned} \boldsymbol{s}_{i,j} \approx \boldsymbol{y}_{i,j} &:= (\hat{y}_{i,j,1}, \hat{y}_{i,j,2}, \ldots, \hat{y}_{i,j,N})^{\top} \in \mathbb{C}^N \\ &:= (\boldsymbol{w}_{i,1}^{\mathrm{H}} \boldsymbol{x}_{i,j}, \boldsymbol{w}_{i,2}^{\mathrm{H}} \boldsymbol{x}_{i,j}, \ldots, \boldsymbol{w}_{i,N}^{\mathrm{H}} \boldsymbol{x}_{i,j})^{\top} = W_i \boldsymbol{x}_{i,j}. \end{aligned}$$

Especially in case of $\boldsymbol{\epsilon}_{i,j} = \boldsymbol{0}$ (although this situation never happens), $\boldsymbol{s}_{i,j} = \boldsymbol{y}_{i,j}$ holds if we can apply the true demixing matrix to $\boldsymbol{x}_{i,j}$.

2.2. BSS Using Independent Super-Gaussian Sound Generation

Conventional methods such as frequency-domain independent component analysis (FDICA) [3], independent vector analysis (IVA) [4], and independent low-rank matrix analysis (ILRMA) [5] assume that the complex spectrogram $S_n := (\hat{s}_{i,j,n}) \in \mathbb{C}^{I \times J}$ of each sound $s_n[t]$ is *independently generated from a super-Gaussian*[3] *distribution.* We define $\boldsymbol{\mathcal{S}} := (S_n) \in \mathbb{C}^{I \times J \times N}$ and $\boldsymbol{\mathcal{X}} := (X_m) \in \mathbb{C}^{I \times J \times M}$ and assume $\boldsymbol{\epsilon}_{i,j} = \boldsymbol{0}$ in (4), then the mixing matrices A_i $(i = 1, 2, \ldots, I)$ become a one-to-one complex linear mapping from $\boldsymbol{\mathcal{S}}$ to $\boldsymbol{\mathcal{X}}$. Hence, by letting p_s and p_x be probability densities of $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{X}}$, we have

$$p_x(\boldsymbol{\mathcal{X}}) = \frac{p_s(\boldsymbol{\mathcal{S}})}{\prod_{i=1}^{I} (\det(A_i^{\mathrm{H}} A_i))^J} = \prod_{i=1}^{I} (\det(W_i W_i^{\mathrm{H}}))^J \prod_{n=1}^{N} p_n(S_n)$$

from (4), (5), and the independence of each S_n, where p_n is a certain carefully designed super-Gaussian generation model[4] of S_n. By substituting $\hat{s}_{i,j,n} = \hat{y}_{i,j,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{i,j}$ in the above equation, p_x is seen as the likelihood $p_x(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{W}})$ of parameter $\boldsymbol{\mathcal{W}} := (W_i) \in \mathbb{C}^{N \times M \times I}$.

For example, in ILRMA, the negative log-likelihood is given by

$$\begin{aligned} -\log(p_x(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{W}})) := \sum_{n=1}^{N} \sum_{i=1}^{I} \sum_{j=1}^{J} \Bigg[& \frac{\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^{\mathrm{H}} \boldsymbol{w}_{i,n}}{\sum_{k=1}^{K_n} b_{i,k,n} c_{k,j,n}} \\ & + \log \sum_{k=1}^{K_n} b_{i,k,n} c_{k,j,n} + \log \pi \Bigg] - J \sum_{i=1}^{I} \log \det(W_i W_i^{\mathrm{H}}). \quad (6) \end{aligned}$$

[1] In fact, we estimate $\hat{s}_{i,j,n} \boldsymbol{a}_{i,n} \in \mathbb{C}^M$ $(n = 1, 2, \ldots, N)$ in most cases.

[2] Strictly, $\forall i$ $\mathrm{rank}(A_i) = N$ and each condition number is not so large.

[3] For a zero-mean *complex-valued* random variable Z, the *kurtosis* of Z is defined by $\kappa := E[|Z|^4]/\sigma^4 - |E[Z^2]/\sigma^2|^2 - 2$, where $\sigma^2 := E[|Z|^2]$ [20]. Z is called *super-Gaussian* if $\kappa > 0$, and called *sub-Gaussian* if $\kappa < 0$.

[4] For example, the sound generation model p_n is designed as

$$p_n(S_n) := \begin{cases} \displaystyle\prod_{i=1}^{I} \prod_{j=1}^{J} \frac{1}{2\pi\sigma_n^2} \exp\left(-\frac{|\hat{s}_{i,j,n}|}{\sigma_n} \right) & \text{in FDICA,} \\ \displaystyle\prod_{j=1}^{J} \frac{2^{I-1}(I-1)!}{(2\pi\sigma_n^2)^I (2I-1)!} \exp\left(-\frac{\|\bar{\boldsymbol{s}}_{j,n}\|_2}{\sigma_n} \right) & \text{in IVA,} \\ \displaystyle\prod_{i=1}^{I} \prod_{j=1}^{J} \frac{1}{\pi\sigma_{i,j,n}^2} \exp\left(-\frac{|\hat{s}_{i,j,n}|^2}{\sigma_{i,j,n}^2} \right) & \text{in ILRMA,} \end{cases}$$

where $\sigma_n > 0$, $\sigma_{i,j,n} > 0$ and $\bar{\boldsymbol{s}}_{j,n} := (\hat{s}_{1,j,n}, \hat{s}_{2,j,n}, \ldots, \hat{s}_{I,j,n})^{\top} \in \mathbb{C}^I$.

In (6), p_n is designed as a *time-varying complex Gaussian distribution* of variance $\sigma_{i,j,n}^2 := \sum_{k=1}^{K_n} b_{i,k,n} c_{k,j,n}$ as shown in Footnote 4 from the idea that the power spectrogram $|S_n|^2 := (|\hat{s}_{i,j,n}|^2) \in \mathbb{R}_+^{I \times J}$ can be approximated, with the multiplication of a *nonnegative basis matrix* $B_n := (b_{i,k,n}) \in \mathbb{R}_+^{I \times K_n}$ and a *nonnegative coefficient matrix* $C_n := (c_{k,j,n}) \in \mathbb{R}_+^{K_n \times J}$, as a *low-rank matrix* $B_n C_n$, where K_n $(\ll \min(I, J))$ is the number of basis vectors for the nth source. Based on the maximum likelihood estimation, to determine $\boldsymbol{\mathcal{W}}$, the conventional methods [3]–[5] minimize the negative log-likelihood

$$\underset{\boldsymbol{\mathcal{W}},(B),(C)}{\text{minimize}} \; -\log(p_x(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{W}})) \quad (\text{subject to } \forall n \, \|Y_n\|_{\mathrm{F}}^2 = IJ). \quad (7)$$

In (7), the two matrices $B := (B_1, B_2, \ldots, B_N) \in \mathbb{R}_+^{I \times K}$ and $C := (C_1^{\top}, C_2^{\top}, \ldots, C_N^{\top})^{\top} \in \mathbb{R}_+^{K \times J}$ are needed in ILRMA, and the constraint for $Y_n := (\hat{y}_{i,j,n}) \in \mathbb{C}^{I \times J}$ is added in ILRMA to avoid the numerical instability caused by the *scale ambiguity of* $\boldsymbol{s}_{i,j}$ *and* A_i.

To solve the nonconvex problem of (7), the conventional methods often use a majorization-minimization technique called the *auxiliary function method* [21], and each row vector of W_i is updated[5] as

$$\begin{cases} U_{i,n}^{(l+1)} = \dfrac{1}{J} \displaystyle\sum_{j=1}^{J} \frac{\boldsymbol{x}_{i,j} \boldsymbol{x}_{i,j}^{\mathrm{H}}}{\rho_{i,j,n}^{(l+1)}} \\ \boldsymbol{v}_{i,n}^{(l+1)} = \left(W_i^{(l)} U_{i,n}^{(l+1)} \right)^{\dagger} \boldsymbol{e}_n = \left(U_{i,n}^{(l+1)} \right)^{-1} \boldsymbol{a}_{i,n}^{(l)} \\ \boldsymbol{w}_{i,n}^{(l+1)} = \dfrac{\boldsymbol{v}_{i,n}^{(l+1)}}{\sqrt{\boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}} U_{i,n}^{(l+1)} \boldsymbol{v}_{i,n}^{(l+1)}}} \end{cases} \quad (8)$$

where l is the update index, $\rho_{i,j,n}^{(l+1)}$ is computed for each method by

$$\rho_{i,j,n}^{(l+1)} := \begin{cases} 2\sigma_n |\hat{y}_{i,j,n}^{(l)}| & \text{in FDICA,} \\ 2\sigma_n \|\bar{\boldsymbol{y}}_{j,n}^{(l)}\|_2 & \text{in IVA,} \\ \sum_{k=1}^{K_n} b_{i,k,n}^{(l+1)} c_{k,j,n}^{(l+1)} & \text{in ILRMA,} \end{cases}$$

$\boldsymbol{e}_n \in \mathbb{R}^N$ is the nth standard basis vector, and $\boldsymbol{a}_{i,n}^{(l)}$ is the nth column vector of $A_i^{(l)} = W_i^{(l)\mathrm{H}} (W_i^{(l)} W_i^{(l)\mathrm{H}})^{-1}$. Then, if needed, the scale of each variable is adjusted to satisfy the constraint for Y_n. After repeating the above updates certain times, finally by using the *projection back (PB)* $(\hat{\hat{y}}_{i,j,1,n}, \hat{\hat{y}}_{i,j,2,n}, \ldots, \hat{\hat{y}}_{i,j,M,n})^{\top} := \hat{y}_{i,j,n} W_i^{\dagger} \boldsymbol{e}_n = \hat{y}_{i,j,n} \boldsymbol{a}_{i,n} \in \mathbb{C}^M$ [22] and applying the inverse STFT (ISTFT) [23] to $\hat{Y}_{m,n} := (\hat{\hat{y}}_{i,j,m,n}) \in \mathbb{C}^{I \times J}$, estimates of $(h_{m,n} * s_n)[t]$ are given.

3. BSS USING THE SPARSITY OF IMPULSE RESPONSES

The conventional methods in Section 2 assumed that the number of dominant components of S_n is not so large, and they tried to improve the separation performance by design of some super-Gaussian sound generation model p_n which properly expresses the true sound distribution. Hence, they constructed the optimization problem of (7) by utilizing a prior information on the sound sources. Furthermore, if a microphone array is used for the observation in (1) and the direction

[5] In ILRMA, before updates of W_i in (8), entries of B and C are updated as

$$\begin{cases} b_{i,k,n}^{(l+1)} = b_{i,k,n}^{(l)} \sqrt{\dfrac{\sum_{j=1}^{J} |\hat{y}_{i,j,n}^{(l)}|^2 \, c_{k,j,n}^{(l)} \left(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l)} c_{k',j,n}^{(l)} \right)^{-2}}{\sum_{j=1}^{J} c_{k,j,n}^{(l)} \left(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l)} c_{k',j,n}^{(l)} \right)^{-1}}} \\ c_{k,j,n}^{(l+1)} = c_{k,j,n}^{(l)} \sqrt{\dfrac{\sum_{i=1}^{I} |\hat{y}_{i,j,n}^{(l)}|^2 \, b_{i,k,n}^{(l+1)} \left(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l+1)} c_{k',j,n}^{(l)} \right)^{-2}}{\sum_{i=1}^{I} b_{i,k,n}^{(l+1)} \left(\sum_{k'=1}^{K_n} b_{i,k',n}^{(l+1)} c_{k',j,n}^{(l)} \right)^{-1}}} \end{cases}$$

which are called *multiplicative updates* in nonnegative matrix factorization.

of arrival of each sound $s_n[t]$ is known, then we can estimate the column vectors $\boldsymbol{a}_{i,n}$, called the *steering vectors* in this situation, of the mixing matrices A_i. By using the pseudo-inverse of the estimated A_i to determine the demixing matrices W_i, some methods achieve high separation performance [12]–[14]. However, if the microphone array is not used or the direction of $s_n[t]$ is unknown, we cannot use them.

In this paper, we propose to use the sparsity of impulse responses as a more general prior information on the mixing process in BSS. Figure 1 shows an impulse response in RWCP database [24], and we find that the energy of the impulse response gathers in the early part, i.e., the impulse response is *approximately sparse* [15], [16]. Define

$$\boldsymbol{h}_{m,n} := (h_{m,n}[0], h_{m,n}[1], \dots, h_{m,n}[T-1])^\top \in \mathbb{R}^T,$$

and we estimate not only $\boldsymbol{\mathcal{W}}$ but also $\boldsymbol{\mathcal{H}} := (\boldsymbol{h}_{m,n}) \in \mathbb{R}^{T\times M\times N}$ by newly using the prior information on impulse responses. Let $A_i := W_i^\dagger = W_i^{\mathrm{H}}(W_i W_i^{\mathrm{H}})^{-1}$ be mixing matrices computed from W_i, and define $\hat{a}_{L-i+2,m,n} := \hat{a}_{i,m,n}^*$ for $i = 2, 3, \dots, I$ from entries $\hat{a}_{i,m,n}$ of A_i. By letting $\bar{\boldsymbol{a}}_{m,n} := (\hat{a}_{1,m,n}, \hat{a}_{2,m,n}, \dots, \hat{a}_{L,m,n})^\top \in \mathbb{C}^L$, we propose to solve the following nonconvex optimization problem

$$\begin{aligned}&\underset{\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}},(B),(C)}{\text{minimize}} \; -\log(p_x(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{W}})) + \lambda J[\mathcal{F}(\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}}) + L\mathcal{G}(\boldsymbol{\mathcal{H}})] \\ &\text{subject to } \forall n \sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L \text{ and } \forall n \sum_{m=1}^M \|\boldsymbol{h}_{m,n}\|_2^2 = 1. \quad (9)\end{aligned}$$

In (9), $\lambda > 0$, the function $\mathcal{F}$ evaluates the consistency between $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{H}}$, the function $\mathcal{G}$ evaluates the sparsity of $\boldsymbol{\mathcal{H}}$, and the constraints $\sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L$ and $\sum_{m=1}^M \|\boldsymbol{h}_{m,n}\|_2^2 = 1$ are added to remove the scale ambiguity. B and C are needed if we use $-\log(p_x)$ in (6).

3.1. Updates of the Demixing Matrices

Let $\widetilde{A}_i$ $(i = 1, 2, \dots, L)$ be mixing matrices computed from $\boldsymbol{\mathcal{H}}$ by using (3), and let $\widetilde{W}_i := \widetilde{A}_i^\dagger = (\widetilde{A}_i^{\mathrm{H}}\widetilde{A}_i)^{-1}\widetilde{A}_i^{\mathrm{H}}$ be the corresponding demixing matrices. When we update $\boldsymbol{\mathcal{W}}$, $\boldsymbol{\mathcal{H}}$ is fixed to the current estimate $\boldsymbol{\mathcal{H}}^{(l)}$, i.e., $\widetilde{W}_i$ are fixed, and the function $\mathcal{F}$ is defined as

$$\mathcal{F}(\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}}) := \sum_{i=1}^I \left\|W_i - \widetilde{W}_i\right\|_{\mathrm{F}}^2 = \sum_{i=1}^I \sum_{n=1}^N \|\boldsymbol{w}_{i,n} - \widetilde{\boldsymbol{w}}_{i,n}\|_2^2. \quad (10)$$

By ignoring the constraint $\sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L$ and utilizing the auxiliary function method if needed, the problem of (9) is reduced to

$$\begin{aligned}\underset{W_i}{\text{minimize}} \; &\sum_{n=1}^N \left(\boldsymbol{w}_{i,n}^{\mathrm{H}} U_{i,n}^{(l+1)} \boldsymbol{w}_{i,n} + \lambda \|\boldsymbol{w}_{i,n} - \widetilde{\boldsymbol{w}}_{i,n}^{(l)}\|_2^2\right) \\ &- \log\det(W_i W_i^{\mathrm{H}}), \quad (11)\end{aligned}$$

where $U_{i,n}^{(l+1)}$ is defined as in (8) according to p_n. Define $\widetilde{U}_{i,n}^{(l+1)} := U_{i,n}^{(l+1)} + \lambda E_M$ with the identity matrix $E_M \in \mathbb{R}^{M\times M}$ and let $\delta_{p,q} \in \{0, 1\}$ be the Kronecker delta. Then, the minimizer of (11) satisfies $\boldsymbol{w}_{i,p}^{\mathrm{H}} \widetilde{U}_{i,q}^{(l+1)} \boldsymbol{w}_{i,q} = \delta_{p,q} + \lambda\, \boldsymbol{w}_{i,p}^{\mathrm{H}} \widetilde{\boldsymbol{w}}_{i,q}^{(l)}$ for $p, q = 1, 2, \dots, N$, and it can be approximately obtained, for each row vector, by

$$\begin{cases} \boldsymbol{v}_{i,n}^{(l+1)} = \left(\widetilde{U}_{i,n}^{(l+1)}\right)^{-1} \boldsymbol{a}_{i,n}^{(l)}, & \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} = \lambda \left(\widetilde{U}_{i,n}^{(l+1)}\right)^{-1} \widetilde{\boldsymbol{w}}_{i,n}^{(l)} \\ d_{i,n}^{(l+1)} = \boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}} \widetilde{U}_{i,n}^{(l+1)} \boldsymbol{v}_{i,n}^{(l+1)}, & \widetilde{d}_{i,n}^{(l+1)} = \boldsymbol{v}_{i,n}^{(l+1)\mathrm{H}} \widetilde{U}_{i,n}^{(l+1)} \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} \\ \boldsymbol{w}_{i,n}^{(l+1)} = \begin{cases} \dfrac{\boldsymbol{v}_{i,n}^{(l+1)}}{\sqrt{d_{i,n}^{(l+1)}}} + \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} & \text{if } \widetilde{d}_{i,n}^{(l+1)} = 0, \\ \dfrac{\widetilde{d}_{i,n}^{(l+1)}}{2d_{i,n}^{(l+1)}} \left(\sqrt{1 + \dfrac{4d_{i,n}^{(l+1)}}{|\widetilde{d}_{i,n}^{(l+1)}|^2}} - 1\right) \boldsymbol{v}_{i,n}^{(l+1)} + \widetilde{\boldsymbol{v}}_{i,n}^{(l+1)} & \\ \qquad \text{otherwise.} \end{cases} \end{cases} \quad (12)$$

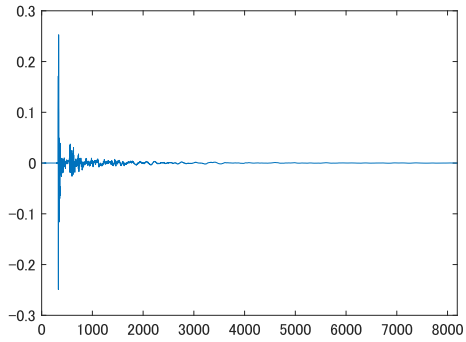

Fig. 1. Example of E2A impulse responses $h[\tau]$ in RWCP database [24], where sampling frequency is 16 [kHz], distance from a sound source to a microphone is 6 [m], and reverberation time is 900 [ms].

The update of (12) was originally derived in [14] as a generalization of (8). Since we ignored the constraint $\sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}\|_2^2 = L$, we have to modify $\boldsymbol{\mathcal{W}}^{(l+1)}$ to satisfy the constraint. First, compute $A_i^{(l+1)} = W_i^{(l+1)\mathrm{H}}(W_i^{(l+1)} W_i^{(l+1)\mathrm{H}})^{-1}$ for $i = 1, 2, \dots, I$ from $W_i^{(l+1)}$. Second, construct $A_{L-i+2}^{(l+1)}$ for $i = 2, 3, \dots, I$ from $\hat{a}_{i,m,n}^*$. Third, define $\gamma_n^{(l+1)} := \sqrt{\sum_{m=1}^M \|\bar{\boldsymbol{a}}_{m,n}^{(l+1)}\|_2^2 / L}$ $(n = 1, 2, \dots, N)$, and divide $\boldsymbol{a}_{i,n}^{(l+1)}$ $(i = 1, 2, \dots, L)$ by $\gamma_n^{(l+1)}$, i.e., multiply $\boldsymbol{w}_{i,n}^{(l+1)}$ by $\gamma_n^{(l+1)}$. Then, we can obtain $\boldsymbol{\mathcal{W}}^{(l+1)}$ satisfying the constraint.

3.2. Updates of the Impulse Responses

When we update $\boldsymbol{\mathcal{H}}$, $\boldsymbol{\mathcal{W}}$ is fixed to $\boldsymbol{\mathcal{W}}^{(l+1)}$ and $\mathcal{F}$ is defined as[6]

$$\mathcal{F}(\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{H}}) := \sum_{i=1}^L \left\|A_i - \widetilde{A}_i\right\|_{\mathrm{F}}^2 = \sum_{m=1}^M \sum_{n=1}^N \|\bar{\boldsymbol{a}}_{m,n} - \Phi \boldsymbol{h}_{m,n}\|_2^2, \quad (13)$$

where the matrix $\Phi := (\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_{T-1}) \in \mathbb{C}^{L\times T}$ denotes the discrete-time Fourier transform in (3), and each column vector $\boldsymbol{\phi}_\tau := (\phi_{1,\tau}, \phi_{2,\tau}, \dots, \phi_{L,\tau})^\top \in \mathbb{C}^L$ is defined from $\phi_{i,\tau} := \exp(\imath\omega_i\tau)$. We define $\mathcal{G}$, to evaluate the sparsity of $\boldsymbol{\mathcal{H}}$, as a *weighted ℓ_0 norm*

$$\mathcal{G}(\boldsymbol{\mathcal{H}}) := \sum_{m=1}^M \sum_{n=1}^N \|\boldsymbol{h}_{m,n}\|_0^\nu := \sum_{m=1}^M \sum_{n=1}^N \sum_{\tau=0}^{T-1} \nu[\tau]\Gamma(h_{m,n}[\tau]), \quad (14)$$

where we design weights $\nu[\tau] > 0$ $(\tau = 0, 1, ..., T-1)$ as a *monotonically increasing sequence* by considering the fact that $|h_{m,n}[\tau]|$ tends to gradually decrease as shown in Fig. 1. The binary function $\Gamma : \mathbb{R} \to \{0, 1\}$ satisfies $\Gamma(h) = 0$ if $h = 0$ and $\Gamma(h) = 1$ if $h \neq 0$.

By ignoring the constraint $\sum_{m=1}^M \|\boldsymbol{h}_{m,n}\|_2^2 = 1$, from $\boldsymbol{\phi}_p^{\mathrm{H}} \boldsymbol{\phi}_q = L\delta_{p,q}$ for $p, q = 0, 1, \dots, T-1$, each $\boldsymbol{h}_{m,n}$ should be updated as

$$\begin{aligned}\boldsymbol{h}_{m,n}^{(l+1)} &= \underset{\boldsymbol{h}_{m,n}}{\operatorname{argmin}} \left\|\frac{\bar{\boldsymbol{a}}_{m,n}^{(l+1)}}{\sqrt{L}} - \frac{\Phi \boldsymbol{h}_{m,n}}{\sqrt{L}}\right\|_2^2 + \|\boldsymbol{h}_{m,n}\|_0^\nu \\ &= \underset{\boldsymbol{h}_{m,n}}{\operatorname{argmin}} \left\|\frac{\Phi^{\mathrm{H}} \bar{\boldsymbol{a}}_{m,n}^{(l+1)}}{L} - \boldsymbol{h}_{m,n}\right\|_2^2 + \|\boldsymbol{h}_{m,n}\|_0^\nu. \quad (15)\end{aligned}$$

By letting $\widetilde{\boldsymbol{h}}_{m,n}^{(l+1)} := \Phi^{\mathrm{H}} \bar{\boldsymbol{a}}_{m,n}^{(l+1)} / L \in \mathbb{R}^T$ be impulse responses computed from $W_i^{(l+1)}$ and $\widetilde{h}_{m,n}^{(l+1)}[\tau]$ be the $(\tau+1)$th entry of $\widetilde{\boldsymbol{h}}_{m,n}^{(l+1)}$, the solution of the problem of (15) can be given by the following hard

[6] For two matrices X and Y of the same size, $\|X - Y\|_{\mathrm{F}}^2 \neq \|X^\dagger - Y^\dagger\|_{\mathrm{F}}^2$ holds in most cases. Therefore, the functions in (10) and (13) are different. Nevertheless, we use (10) and (13) as the functions evaluating the consistency between $\boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{H}}$ since $\|X - Y\|_{\mathrm{F}}^2 = \|X^\dagger - Y^\dagger\|_{\mathrm{F}}^2 = 0$ holds if $X = Y$.

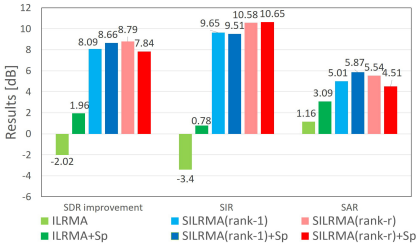

Fig. 2. Results for a bass sound arrived at 50°.

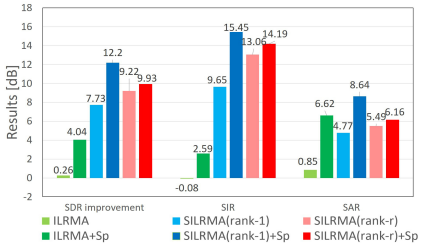

Fig. 3. Results for a piano sound arrived at 90°.

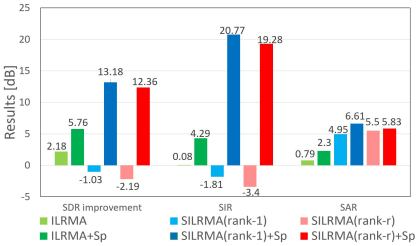

Fig. 4. Results for a drum sound arrived at 130°.

thresholding

$$h_{m,n}^{(l+1)}[\tau] = \begin{cases} \widetilde{h}_{m,n}^{(l+1)}[\tau] & \text{if } \left|\widetilde{h}_{m,n}^{(l+1)}[\tau]\right| \geq \sqrt{\nu[\tau]}, \\ 0 & \text{if } \left|\widetilde{h}_{m,n}^{(l+1)}[\tau]\right| < \sqrt{\nu[\tau]}. \end{cases} \tag{16}$$

Since we ignored the constraint $\sum_{m=1}^{M} \|\boldsymbol{h}_{m,n}\|_2^2 = 1$, we have to modify $\boldsymbol{\mathcal{H}}^{(l+1)}$ to satisfy the constraint in a similar manner to $\boldsymbol{\mathcal{W}}$. Define $\xi_n^{(l+1)} := \sqrt{\sum_{m=1}^{M} \|\boldsymbol{h}_{m,n}^{(l+1)}\|_2^2}$ $(n = 1, 2, \ldots, N)$ and divide $\boldsymbol{h}_{m,n}^{(l+1)}$ $(m = 1, 2, \ldots, M)$ by $\xi_n^{(l+1)}$. Then, we can obtain $\boldsymbol{\mathcal{H}}^{(l+1)}$ satisfying the constraint. After repeating (12) and (16) certain times, by using PB and ISTFT, finally estimates of $(h_{m,n} * s_n)[t]$ are given.

3.3. Supervised Learning of the Nonnegative Basis Matrices

In BSS, since information on each sound and the mixing process is almost unknown, it is very difficult to accurately separate each sound. On the other hand, in supervised source separation, nonnegative matrix factorization (NMF) methods and deep learning methods achieve high separation performance, even in the under-determined cases including $M = 1$, by using training data [17]–[19], [25], [26]. In this paper, we focus on the fact that NMF is used in the sound generation model p_n of ILRMA. Suppose that each sound $s_n[t]$ is an audio signal from a known musical instrument. In such situations, we propose to construct the nonnegative basis matrices B_n in advance from training data. Each B_n is learned, by the method in [27], from *monophonic power spectrograms* of each musical instrument as follows.

Let $|S_{\mathrm{mono}}|^2 \in \mathbb{R}_+^{I \times J}$ be some monophonic power spectrogram of a musical instrument. In the simplest learning method, we create *one basis vector* $\tilde{\boldsymbol{b}}_{\mathrm{mono}}$ from $|S_{\mathrm{mono}}|^2$ by the *rank-1 approximation*

$$(\tilde{\boldsymbol{b}}_{\mathrm{mono}}, \tilde{\boldsymbol{c}}_{\mathrm{mono}}) = \underset{\tilde{\boldsymbol{b}} \in \mathbb{R}_+^I, \tilde{\boldsymbol{c}} \in \mathbb{R}_+^J}{\operatorname{argmin}} \left\| |S_{\mathrm{mono}}|^2 - \tilde{\boldsymbol{b}}\tilde{\boldsymbol{c}}^\top \right\|_{\mathrm{F}}. \tag{17}$$

$\tilde{\boldsymbol{b}}_{\mathrm{mono}}$ is computed as the first left singular vector of $|S_{\mathrm{mono}}|^2$ with the singular value decomposition (SVD). On the other hand, in [27], we create *multiple basis vectors* $(\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \ldots, \tilde{\boldsymbol{b}}_{r_*}) =: \widetilde{B}_{r_*} \in \mathbb{R}_+^{I \times r_*}$ from $|S_{\mathrm{mono}}|^2$ to express sound segments such as *attack* and *sustain* in detail. The number r_* of basis vectors is the smallest r satisfying

$$\exists \widetilde{B}_r \in \mathbb{R}_+^{I \times r} \text{ such that } \min_{\widetilde{C}_r \in \mathbb{R}_+^{r \times J}} \left\| |S_{\mathrm{mono}}|^2 - \widetilde{B}_r \widetilde{C}_r \right\|_{\mathrm{F}} \leq \varepsilon, \tag{18}$$

where $\varepsilon > 0$ is the acceptable error. $\widetilde{B}_{r_*}$ is computed, with SVD and a greedy algorithm, by choosing r_* column vectors of the *rank-r_* approximation* $|S_{\mathrm{mono}}|_{r_*}^2$ for $|S_{\mathrm{mono}}|^2$ (see [27] for more detail).

4. NUMERICAL EXPERIMENTS

We show the effectiveness of the proposed method by numerical experiments where the numbers of sound sources and microphones are $N = M = 3$. As sounds $s_n[t]$ of sampling frequency 16 [kHz], we used three audio signals (bass, piano, and drum sounds) which were created with the musical instrument digital interface by the third author. As real impulse responses $h_{m,n}[\tau]$, we used E2A of reverberation time 900 [ms], as shown in Fig. 1, in RWCP database [24], where we used microphones of numbers 17, 23, and 30. In STFT of (2), we set $L = 8192$, $\eta = 2048$, and $\psi[\tau] = 0.54 - 0.46\cos(2\pi\tau/L)$. We compared ILRMA to the proposed method with (6) (ILRMA + Sp) as blind methods. We compared ILRMA with supervised bases of (17) (SILRMA (rank-1)) & (18) (SILRMA (rank-r)) to the corresponding proposed ones (SILRMA (rank-1) + Sp) & (SILRMA (rank-r) + Sp) as supervised methods, where $\varepsilon = 0.1\||S_{\mathrm{mono}}|^2\|_{\mathrm{F}}$ in (18). We set $K_n = 30$ in ILRMA and ILRMA + Sp. In ILRMA + Sp, we set $T = 4096$, $\lambda = 0.075$ in (9), and $\nu[\tau] = -\log_{10}(1 - \exp(-432/(\tau+1)))$ in (14). In SILRMA + Sp, we only changed the parameter $\lambda = 0.02$.

All the methods updated variables 100 times from the same initial values, where $W_i^{(0)} = E_3$, $\boldsymbol{h}_{m,n}^{(0)} = \boldsymbol{0}$, entries of $B_n^{(0)}$ and $C_n^{(0)}$ were generated from the uniform distribution between $[0, 1]$. We used source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [28] as evaluation indices, and we compared the average of each index in 10 experiments.

Figures 2, 3, and 4 show the separation results of each method for the bass, piano, and drum sounds, respectively. Among the blind methods, ILRMA + Sp improved the standard ILRMA by averagely 3.78 [dB] in SDR that represents the overall separation performance, 3.67 [dB] in SIR that represents the removal performance of the non-target sounds, and 3.07 [dB] in SAR that represents the suppression performance of artifacts. By listening to the sounds, we found that the assignment of each musical instrument was relatively difficult for the results of ILRMA while it was easy for those of ILRMA + Sp though its overall separation performance was also not high yet. Among the supervised methods, both SILRMA and SILRMA + Sp achieved high separation performance for the bass and piano sounds. However, for the drum sound, the results of SILRMA were very bad, i.e., the drum sound almost disappeared, since we might fail to learn the basis matrices due to the *difficulty for percussion instruments*. In spite of the failure of the learning, SILRMA + Sp obtained very good results for the drum sound by newly using the sparsity of the impulse responses. Among the all methods, SILRMA (rank-1) + Sp achieved the highest separation performance, and its separation sounds were very clear.[7]

5. CONCLUSION

In this paper, we proposed an over-determined sound source separation method using the sparsity of impulse responses as a prior information on the mixing process. In the proposed method, we estimate not only the demixing matrices but also the impulse responses while considering the consistency with each other. We extended the proposed method to a supervised one by creating the nonnegative basis matrices in advance. The experiments showed that the proposed method improves the conventional methods and leads to clearer sounds.

[7] See the results (https://mediasensinglab.bitbucket.io/SoundSource.html).

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: John Wiley & Sons, 2001.

[2] S. Makino, *Audio Source Separation*, ser. Signals and Communication Technology. New York, NY, USA: Springer, 2018.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, no. 1–3, vol. 22, pp. 21–34, 1998.

[4] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[5] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[6] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. e12, 14 pages, 2019.

[7] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, Canada, 2018, pp. 776–780.

[8] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)* Denver, CO, USA, 2000, pp. 535–541.

[9] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Int. Conf. Inf. Sci. Signal Process. Appl. (ISSPA)*, Kuala Lumpur, Malaysia, 2010, pp. 1–4.

[10] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, Canada, 2013, pp. 71–75.

[11] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[12] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 666–678, 2006.

[13] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, 2002.

[14] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis" in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, Canada, 2018, pp. 746–750.

[15] Z. Koldovský, F. Nesta, P. Tichavský, and N. Ono, "Frequency-domain blind speech separation using incomplete de-mixing transform," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Budapest, Hungary, 2016, pp. 1663–1667.

[16] J. Janský, Z. Koldovský, and N. Ono, "A computationally cheaper method for blind speech separation based on AuxIVA and incomplete demixing transform," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Xian, China, 2016, 5 pages.

[17] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. Int. Conf. Indep. Comp. Anal. Signal Separat. (ICA)*, London, UK, 2007, pp. 414–421.

[18] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Music signal separation by supervised nonnegative matrix factorization with basis deformation," in *Proc. Int. Conf. Digit. Signal Process. (DSP)*, Fira, Greece, 2013, 6 pages.

[19] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing," in *Proc. IEEE Int Symp. Signal Process. Inf. Technol. (ISSPIT)*, Athens, Greece, 2013, pp. 392–397.

[20] E. Ollila, V. Koivunen, and H. V. Poor, "Complex-valued signal processing—essential models, tools and statistics," in *Proc. Inf. Theory Appl. Workshop (ITA)*, La Jolla, CA, USA, 2011, 10 pages.

[21] N. Ono, and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal Signal Separat. (LVA/ICA)*, St. Malo, France, 2010, pp. 165–172.

[22] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. SICE Annual Conf.*, Osaka, Japan, 2002, pp. 2138–2143.

[23] B. Yang, "A study of inverse short-time Fourier transform," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 3541–3544.

[24] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Int. Conf. Lang. Res. Eval. (LREC)*, Athens, Greece, 2000, 4 pages.

[25] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, 2018, pp. 1557–1561.

[26] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[27] T. Fujiwara, M. Yamagishi, and I. Yamada, "Reduced-rank modeling of time-varying spectral patterns for supervised source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 3307–3311.

[28] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.