






Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation

Naoki Makishima , Shinichi Mogami , *Student Member, IEEE*, Norihiro Takamune, Daichi Kitamura , *Member, IEEE*, Hayato Sumino, Shinnosuke Takamichi , *Member, IEEE*, Hiroshi Saruwatari , *Member, IEEE*, and Nobutaka Ono, *Senior Member, IEEE*

Abstract—In this paper, we propose a new framework called independent deeply learned matrix analysis (IDLMA), which unifies a deep neural network (DNN) and independence-based multichannel audio source separation. IDLMA utilizes both pretrained DNN source models and statistical independence between sources for the separation, where the time-frequency structures of each source are iteratively optimized by a DNN while enhancing the estimation accuracy of the spatial demixing filters. As the source generative model, we introduce a complex heavy-tailed distribution to improve the separation performance. In addition, we address a semi-supervised situation; namely, a solo-recorded audio dataset can be prepared for only one source in the mixture signal. To solve the limited-data problem, we propose an appropriate data augmentation method to adapt the DNN source models to the observed signal, which enables IDLMA to work even in the semi-supervised situation. Experiments are conducted using music signals with a training dataset in both supervised and semi-supervised situations. The results show the validity of the proposed method in terms of the separation accuracy.

Index Terms—Audio source separation, independent component analysis, deep neural networks, semi-supervised learning.

I. INTRODUCTION

BLIND source separation (BSS) aims at extracting specific sources from an observed multichannel mixture signal without knowing a priori information about the mixing system. The most commonly used algorithms for BSS in the (over)determined case (the number of microphones is more than that of sources) are independent component analysis (ICA) [1] and its extended algorithms such as independent

vector analysis (IVA) [2], which assume statistical independence between the sources and estimate the demixing system. Recently, auxiliary-function-based algorithms (equivalent to majorization-minimization (MM) algorithms [3]) for ICA (AuxICA) [4] and IVA (AuxIVA) [5] have been derived, in which convergence-guaranteed fast optimization was realized by alternative updates of auxiliary parameters and demixing matrices, where the latter was called iterative projection (IP). On the basis of IP, independent low-rank matrix analysis (ILRMA) [6], [7], which is a unification of AuxIVA and nonnegative matrix factorization (NMF) [8], has been proposed as a state-of-the-art BSS method. ILRMA assumes both statistical independence between sources and a low-rank time-frequency structure for each source, and the frequency-wise demixing matrices are estimated without encountering the so-called permutation problem [9], [10]. ILRMA assumes the local Gaussian model (LGM) [11], [12], which was originally proposed as a probabilistic interpretation of NMF based on Itakura–Saito divergence (ISNMF) [11] and its multichannel extensions [13]–[15]. The LGM consists of a zero-mean and isotropic complex Gaussian distribution independently defined at each time-frequency slot and has been used in many techniques such as the full-rank spatial covariance model (FSCM) [13] and multichannel NMF (MNMF) [14], [15]. In recent studies, the LGM in ILRMA has been generalized to the complex Student’s t distribution (t -ILRMA) [16], [17] for more high-performance BSS. As a more general framework of LGM-based BSS, in [18], demixing matrix optimization based on a given power spectrogram estimate (time-frequency-wise variance) was proposed, showing that the precise source spectrogram model enables accurate spatial model estimation.

In the underdetermined case (the number of microphones is less than that of sources), the FSCM [13] is a commonly used framework. In this model, frequency-wise spatial covariances, which encode source locations and their spatial spreads, are estimated by the expectation-maximization (EM) algorithm, where the permutation problem must be solved after the optimization. MNMF [15] is a technique combining the FSCM and an NMF-based permutation solver, which was extended to ILRMA in the determined case. In [14], [15], NMF was used in a blind setting and its role was to ensure that the separated signals have low-rank time-frequency structures. On the other hand, in [14], [19], NMF was used in a supervised (informed) setting, e.g., the basis matrix was trained in advance. Note that the

Manuscript received March 26, 2019; revised June 14, 2019; accepted June 18, 2019. Date of publication June 27, 2019; date of current version July 15, 2019. This work was supported by the SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers JP17H06101, JP19H01116, and JP19K20306. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sven Erik Nordholm. (*Corresponding author: Naoki Makishima.*)

N. Makishima, S. Mogami, N. Takamune, H. Sumino, S. Takamichi, and H. Saruwatari are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: naoki_makishima@ipc.i.u-tokyo.ac.jp; shinichi_mogami@ipc.i.u-tokyo.ac.jp; norihiro_takamune@ipc.i.u-tokyo.ac.jp; hayato-sumino787@g.ecc.u-tokyo.ac.jp; shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp; hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp).

D. Kitamura is with the National Institute of Technology, Kagawa College, Kagawa 761-8058 Japan (e-mail: kitamura-d@t.kagawa-nct.ac.jp).

N. Ono is with the Graduate School of System Design, Tokyo Metropolitan University, Tokyo 191-0065, Japan (e-mail: onono@tmu.ac.jp).

Digital Object Identifier 10.1109/TASLP.2019.2925450

FSCM and MNMF formulate a *mixing* model, whereas ICA-based methods including ILRMA estimate a *demixing* model for the separation by focusing only on the determined case. It has been experimentally confirmed that the optimization of a demixing model is more efficient and numerically stable than that of a mixing model [6].

In supervised source separation, deep neural networks (DNNs) have shown promising performance in both single-channel [20]–[23] and multichannel source separation [24]–[27]. In fact, when a sufficient number of solo-recorded audio signals (signals recorded when isolated instruments are played) are available as a dataset, a DNN can effectively model their time-frequency structures. However, it is almost impossible to compose an appropriate and generalized spatial model using a DNN from training data observed in a multichannel format. This is because the spatial model depends on many factors, including the source and microphone locations, the recording environment, and reverberation. Therefore, it is reasonable to combine a pretrained DNN source model and a blind estimation of the spatial model. In [25], [26], a DNN-based beamformer was proposed and applied to a speech enhancement task. Although this framework iteratively updates spatial filters and DNN-based source models, the algorithm is not based on a consistent deterministic or probabilistic model; namely, the spatial beamformer and DNN are heuristically combined without any theoretical validity or justification. In [27], Nugraha et al. proposed a DNN-based multichannel source separation framework using the FSCM (hereafter referred to as *FSCM+DNN*). Although this is a convincing approach, a large computational cost is required to estimate the spatial covariance (the EM algorithm in the FSCM) and the performance is not satisfactory owing to the difficulty of spatial parameter optimization, as discussed in [6].

In this paper, we unify the ICA-based blind estimation of the demixing matrix and the DNN-based supervised update of the source spectrogram model. In the proposed method, we introduce a complex Student's t distribution as a generalized source generative model including the LGM, and the demixing matrix (spatial model) is efficiently optimized using an MM algorithm. Since the proposed method utilizes a time-frequency spectrogram matrix estimated by a DNN to optimize the spatial model, we call this method *independent deeply learned matrix analysis (IDLMA)*. In addition, we address a semi-supervised situation where a solo-recorded dataset can be prepared for only one source in the mixture signal and there is no solo-recorded dataset for the other sources. In this situation, since a DNN source model for the other sources cannot be prepared in advance, we propose a new data augmentation scheme, where the augmented data are used to iteratively retrain the DNN source model for all the sources while optimizing the spatial model.

Figure 1 shows the relationship between the existing and proposed multichannel source separation methods. The source spectrogram model is estimated by supervised NMF in [14], [19] and is estimated by a DNN in FSCM+DNN and the proposed IDLMA.

The rest of this paper is organized as follows. In Section II, we formulate the problem and introduce the source generative

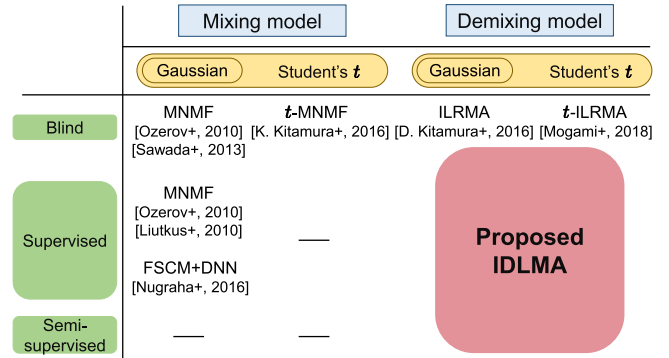


Fig. 1. Relationship between existing and proposed multichannel source separation methods.

model using the existing conventional methods. In Section III, the detailed framework of IDLMA in the supervised situation is described. In Section IV, the application of IDLMA to the semi-supervised situation is described. An experimental evaluation using music signals is given in Section V. The conclusions of this paper are presented in Section VI. Note that this paper is partially based on an international conference paper [28] written by the authors. The contribution of this paper is that we provide a new extended scheme of the proposed method for the semi-supervised situation and report expanded experiments carried out under various conditions.

II. CONVENTIONAL METHOD

A. Formulation

Let N and M be the numbers of sources and channels, respectively. The short-time Fourier transforms (STFTs) of the multichannel source, observed, and estimated signals are defined as

$$\mathbf{s}_{ij} = (s_{ij1}, \dots, s_{ijN})^T, \quad (1)$$

$$\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijM})^T, \quad (2)$$

$$\mathbf{y}_{ij} = (y_{ij1}, \dots, y_{ijN})^T, \quad (3)$$

where $i = 1, \dots, I; j = 1, \dots, J; n = 1, \dots, N; m = 1, \dots, M$ are the indexes of the frequency bins, time frames, sources, and channels, respectively, and T denotes the transpose. We also denote these spectrograms as $\mathbf{S}_n \in \mathbb{C}^{I \times J}$, $\mathbf{X}_m \in \mathbb{C}^{I \times J}$, and $\mathbf{Y}_n \in \mathbb{C}^{I \times J}$, whose elements are s_{ijn}, x_{ijn} , and y_{ijn} , respectively. In ILRMA, the following mixing system is assumed:

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij}, \quad (4)$$

where $\mathbf{A}_i = (\mathbf{a}_{i1}, \dots, \mathbf{a}_{iN}) \in \mathbb{C}^{M \times N}$ is a frequency-wise mixing matrix and \mathbf{a}_{in} is the steering vector for the n th source. The assumption of the mixing system (4) corresponds to restricting the spatial covariance in the FSCM to a rank-1 matrix [13]. When $M = N$ and \mathbf{A}_i is not a singular matrix, the estimated signal \mathbf{y}_{ij} can be represented as

$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij} \approx \mathbf{s}_{ij}, \quad (5)$$

where $\mathbf{W}_i = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{iN})^H$ is the demixing matrix, \mathbf{w}_{in} is the demixing filter for the n th source, and H denotes the Hermitian transpose. ILRMA estimates both \mathbf{W}_i and \mathbf{y}_{ij} from only the observation \mathbf{x}_{ij} assuming statistical independence between s_{ijn} and $s_{ijn'}$, where $n \neq n'$.

B. FSCM and MNMF

In the FSCM [13], a generative model of a multichannel observed signal \mathbf{x}_{ij} is defined by a zero-mean multivariate complex Gaussian distribution as

$$\prod_{i,j} p(\mathbf{x}_{ij}) = \prod_{i,j} \frac{1}{\det(\pi \mathbf{R}_{ij}^{(x)})} \exp\left(-\mathbf{x}_{ij} \mathbf{R}_{ij}^{(x)-1} \mathbf{x}_{ij}^H\right), \quad (6)$$

where $\mathbf{R}_{ij}^{(x)} \in \mathbb{C}^{M \times M}$ is the spatial covariance of the observed signal. This model is often called the LGM [11], [12]. Under the assumption that the sources are mutually uncorrelated, $\mathbf{R}_{ij}^{(x)}$ can be decomposed as

$$\mathbf{R}_{ij}^{(x)} = \sum_n r_{ijn} \mathbf{R}_{in}^{(s)}, \quad (7)$$

where $r_{ijn} > 0$ is the time-frequency-varying variance (power spectrogram) of the n th source and $\mathbf{R}_{in}^{(s)} \in \mathbb{C}^{M \times M}$ is the time-invariant spatial covariance, which encodes the acoustic path from the n th source to the microphones and its spatial spread. The parameters r_{ijn} and $\mathbf{R}_{in}^{(s)}$ can be optimized by the EM algorithm, and the estimated multichannel source signals (source images) can be recovered by multichannel Wiener filtering.

The FSCM incurs the permutation problem if no additional model is assumed for r_{ijn} . MNMF [14], [15] introduces an NMF-based low-rank assumption into r_{ijn} , and permutation-free BSS is achieved. However, its performance is not stable against, e.g., parameter initialization because a huge number of parameters must be optimized. Instead of employing NMF, FSCM+DNN [27] utilizes a DNN-based source model to estimate r_{ijn} ; namely, a DNN that enhances each source is trained using sourcewise (solo-recorded) datasets in advance, and the variance r_{ijn} is updated by the DNN while optimizing the spatial covariance $\mathbf{R}_{in}^{(s)}$. However, the training cost of the DNN used in [27] is large because multiple DNNs should be prepared for each spatial update with the EM algorithm. Also, the performance is still limited because of the difficulty of optimizing the spatial covariance $\mathbf{R}_{in}^{(s)}$.

C. ILRMA and Its Generalization

ILRMA [6], [7] is a fast and stable BSS algorithm that estimates the demixing matrix \mathbf{W}_i instead of the mixing system (spatial covariance $\mathbf{R}_{in}^{(s)}$). The difference between MNMF and ILRMA is the rank of the spatial covariance; namely, ILRMA restricts $\mathbf{R}_{in}^{(s)}$ to be a rank-1 matrix. This rank-1 spatial model is equivalent to assuming the mixing system (4), and the parameter $\mathbf{R}_{in}^{(s)}$ can be converted to the demixing matrix \mathbf{W}_i , resulting in a substantial reduction of the number of spatial parameters from INM^2 to INM .

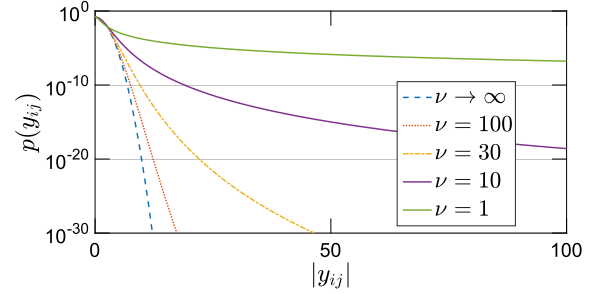


Fig. 2. Tails of Student's t distribution at various ν values.

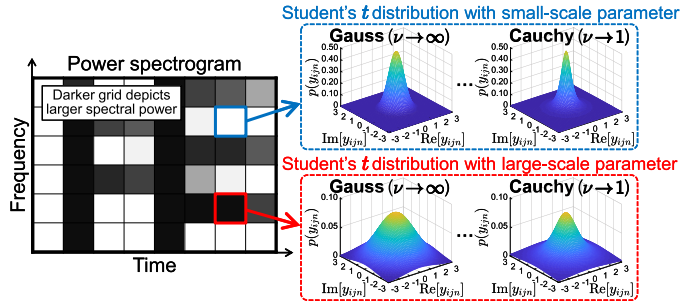


Fig. 3. Generative model based on complex Student's t distribution.

In ILRMA, the following univariate complex Gaussian distribution is assumed as a source generative model (hereafter referred to as Gauss-ILRMA):

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{1}{\pi \sigma_{ijn}^2} \exp\left(-\frac{|y_{ijn}|^2}{\sigma_{ijn}^2}\right), \quad (8)$$

$$\sigma_{ijn}^2 = \sum_k t_{ikn} v_{kjn}, \quad (9)$$

where σ_{ijn} is the scale parameter, $k = 1, \dots, K$ is the index of the NMF bases, and $t_{ikn} > 0$ and $v_{kjn} > 0$ are the elements of the basis matrix $\mathbf{T}_n \in \mathbb{R}_{\geq 0}^{I \times K}$ and the activation matrix $\mathbf{V}_n \in \mathbb{R}_{\geq 0}^{K \times J}$, respectively. We also denote the scale parameter matrix as $\mathbf{\Sigma}_n \in \mathbb{R}_{\geq 0}^{I \times J}$, whose elements are σ_{ijn} . Thus, (9) can also be represented as $|\mathbf{\Sigma}_n|^2 = \mathbf{T}_n \mathbf{V}_n$. The marginal distribution of the time-varying complex Gaussian model w.r.t. the time frame is super-Gaussian when the scale parameter fluctuates and is not constant w.r.t. the time frame. In t -ILRMA [16], the generative model (8) is extended to the complex Student's t distribution as follows:

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{1}{\pi \sigma_{ijn}^2} \left(1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{\sigma_{ijn}^2}\right)^{-\frac{2+\nu}{2}}, \quad (10)$$

$$\sigma_{ijn}^p = \sum_k t_{ikn} v_{kjn}, \quad (11)$$

where ν is the degree-of-freedom parameter defined in the Student's t distribution and p is a domain parameter used in the NMF decomposition. When $\nu \rightarrow \infty$ and $p = 2$, (10) and (11) become identical to (8) and (9), respectively. Also, (10) with $\nu = 1$ represents the Cauchy distribution likelihood. Fig. 2 shows the tails of the complex Student's t distribution at various ν values, and

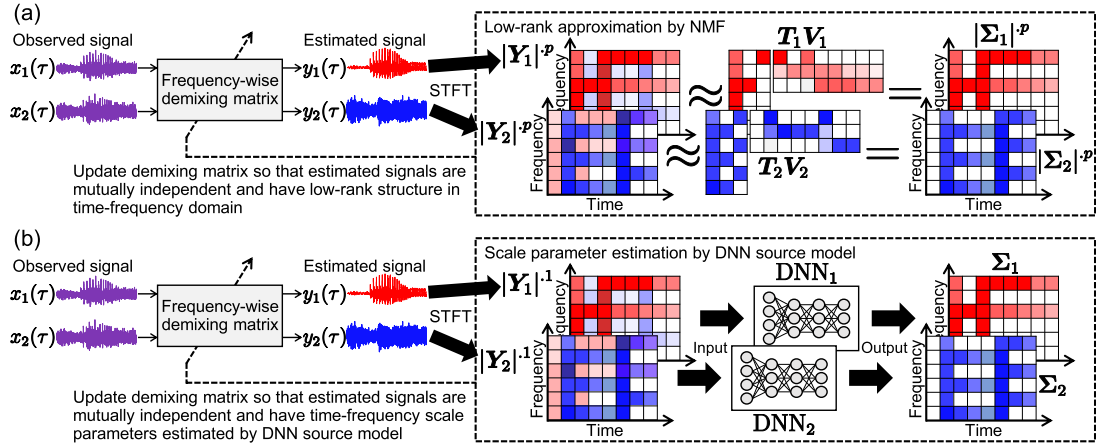


Fig. 4. Separation principle of (a) ILRMA and (b) IDLMA, where $x_n(\tau)$ and $y_n(\tau)$ are observed and estimated time-domain signals, respectively, and τ denotes index of time samples.

Fig. 3 shows the generative model based on (10). A small ν leads to a heavy-tailed distribution, which provides outlier-robust parameter estimation. Also, the distribution is independently defined in each time-frequency slot, and the spectral power corresponds to the variance $\sigma_{ij n}^2$.

Figure 4(a) shows the principle of source separation based on ILRMA. The demixing matrix \mathbf{W}_i and the NMF source model $\mathbf{T}_n \mathbf{V}_n$ can be optimized in the maximum-likelihood (ML) sense based on (8) or (10). Since the low-rank structure of the power spectrogram $|\mathbf{Y}|^2$ is ensured by the NMF decomposition, the permutation problem can be avoided, where $|\cdot|^p$ for matrices denotes the element-wise absolute and p th-power operations.

III. PROPOSED METHOD

A. Motivation

The NMF source model in ILRMA is effective for some sources that have a low-rank time-frequency structure. However, this source model is not always valid. For example, speech signals have continuously varying spectra, which cannot be efficiently modeled by NMF, and the separation performance of ILRMA is degraded for such sources.

If sufficient training data for each source can be prepared in advance, it is possible to construct a suitable source model by employing a DNN [20]. On the other hand, since the spatial parameters depend on many factors, it is impractical to train a general spatial model with a DNN even if huge amounts of multichannel observation data are available; therefore, the spatial parameters should be estimated *blindly*.

In this paper, we propose a new framework, IDLMA, which combines the ICA-based blind estimation of the demixing matrix \mathbf{W}_i and the supervised learning of the scale parameter matrix Σ_n based on a DNN. The separation principle of IDLMA is shown in Fig. 4(b). In IDLMA, the NMF source model is replaced by DNN-based scale parameter estimation for each source, where DNN_n is the DNN source model for the n th source trained in advance. The loss function in DNN_n is designed to

maximize the likelihood of the source generative model and to output the scale parameter matrix Σ_n .

In addition, similarly to t -ILRMA, we use a generalized model based on the zero-mean and isotropic complex Student's t distribution including both the Gaussian and Cauchy distributions. FSCM+DNN also employs a DNN that maximizes the likelihood of the Gaussian or Cauchy distribution. However, since the mixing model $\mathbf{R}_{in}^{(s)}$ in FSCM+DNN is defined by only the Gaussian model, the estimations of the spectral and spatial parameters are inconsistent. In the proposed method, this statistical conflict is resolved by modeling both the spatial and source parameters with the consistent generative model based on the Student's t distribution, and their optimization algorithms are derived.

B. Cost Function in IDLMA

In IDLMA, we assume the same source generative model as in ILRMA, modifying the low-rank modeling of the scale parameter to DNN modeling. On the basis of (8), the cost function (negative log-likelihood of the observed signal \mathbf{x}_{ij}) in IDLMA with the LGM (Gauss-IDLMA) is obtained as

$$\begin{aligned} \mathcal{L}_{\text{Gauss}}(\mathbf{W}) &= -\log p(\mathbf{X}) \\ &= -\log p(\mathbf{Y}) - 2J \sum_i \log |\det \mathbf{W}_i| \\ &= \sum_{i,j,n} \left[\frac{|\mathbf{w}_{in}^H \mathbf{x}_{ij}|^2}{\sigma_{ijn}^2} + 2 \log \sigma_{ijn} \right] \\ &\quad - 2J \sum_i \log |\det \mathbf{W}_i| + IJN \log \pi, \end{aligned} \quad (12)$$

$$\text{s.t. } \sigma_{ijn} = \max([\text{DNN}_n(|\mathbf{Y}_n|^{\cdot 1})]_{i,j}, \varepsilon), \quad (13)$$

where $\mathbf{W} = \{\mathbf{W}_1, \dots, \mathbf{W}_I\}$ is the set of demixing matrices, $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_m\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ are the sets of the observed and estimated signals, respectively, $\text{DNN}_n(\cdot)$

is the output of DNN_n trained to estimate the scale parameter of the n th source (the detail of DNN training is described in Section III-C), $[\cdot]_{i,j}$ is the (i, j) th value of the matrix, and ε is a small value to increase the numerical stability. In the derivation of (12), the independence between sources, $p(\mathbf{Y}) = \prod_n p(\mathbf{Y}_n)$, is assumed, and we used the transformation of random variables from \mathbf{x}_{ij} to \mathbf{y}_{ij} as denoted in (5). Since the joint optimization of (12) and (13) is difficult, we iteratively update \mathbf{W}_i and Σ_n . In the iterations, \mathbf{W}_i is estimated in the ML sense based on the currently estimated scale parameter Σ_n . Although Σ_n given by (13) is not the ML estimate in the strict sense, it is expected to increase the likelihood when DNN_n is trained to output the scale parameter of \mathbf{S}_n from the noisy (interference-remaining) spectrogram, as described in Section III-C. Gauss-IDLMA can be generalized using (10) as IDLMA based on the Student's t distribution (t -IDLMA). The cost function in t -IDLMA is defined as

$$\begin{aligned} \mathcal{L}_t(\mathbf{W}) = & \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \log \left(1 + \frac{2}{\nu} \frac{|\mathbf{w}_{in}^H \mathbf{x}_{ij}|^2}{\sigma_{ijn}^2}\right) + 2 \log \sigma_{ijn} \right] \\ & - 2J \sum_i \log |\det \mathbf{W}_i| + IJN \log \pi, \end{aligned} \quad (14)$$

$$\text{s.t. } \sigma_{ijn} = \max([\text{DNN}_n(|\mathbf{Y}_n|^1)]_{i,j}, \varepsilon). \quad (15)$$

Note that \mathcal{L}_t converges to $\mathcal{L}_{\text{Gauss}}$ when $\nu \rightarrow \infty$.

C. Architecture and Training of DNN Source Model

DNN_n is trained so that the scale parameter of source $\tilde{\mathbf{S}}_n$, Σ_n , based on (8) or (10) is predicted from an input mixture spectrogram \mathbf{B} , where $\tilde{\mathbf{S}}_n \in \mathbb{C}^{I \times J}$ and $\mathbf{B} \in \mathbb{C}^{I \times J}$ are the source and mixture spectrograms in the training data, respectively. When we define the output scale parameter matrix as $\mathbf{D}_n = \text{DNN}_n(|\mathbf{B}|^1) \approx \Sigma_n$, the loss function of DNN_n in (13) can be defined as

$$\mathcal{L}_{\text{Gauss}}(\mathbf{D}_n) = \sum_{i,j} \left(\frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1} - \log \frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1} - 1 \right), \quad (16)$$

where \tilde{s}_{ijn} and d_{ijn} are the elements of $\tilde{\mathbf{S}}_n$ and \mathbf{D}_n , respectively, and δ_1 is a small value to avoid division by zero [27]. Also, the loss function of DNN_n in (15) can be defined as

$$\begin{aligned} \mathcal{L}_t(\mathbf{D}_n) = & \sum_{i,j} \left[\left(1 + \frac{\nu}{2}\right) \log \left(1 + \frac{2}{\nu} \frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1}\right) \right. \\ & \left. + \log(d_{ijn}^2 + \delta_1) \right]. \end{aligned} \quad (17)$$

Since minimizing (16) or (17) corresponds to a simulation for the ML estimation of σ_{ijn} in (12) or (14) (only limited to the training data), DNN_n can be approximately interpreted as an appropriate source model based on (8) or (10), respectively. Similarly to (14), $\mathcal{L}_t(\mathbf{D}_n)$ converges to $\mathcal{L}_{\text{Gauss}}(\mathbf{D}_n)$ up to a constant when $\nu \rightarrow \infty$.

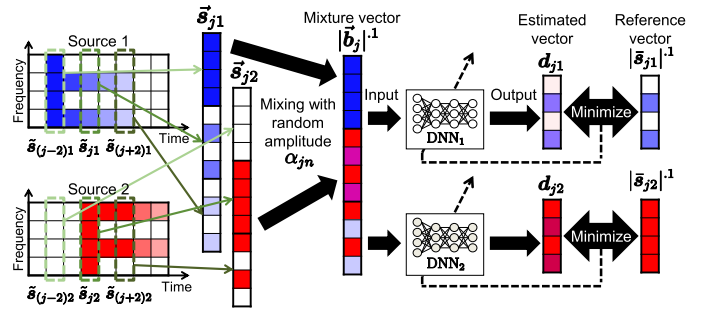


Fig. 5. Outline of DNN training when $I = 4$, $J = 8$, $N = 2$, and $c = 1$.

Although many sophisticated DNN-related methods have been proposed such as convolutional neural networks, generative adversarial networks [29], variational autoencoders [30], U-Nets [22], and deep clustering [23], in this paper, we only focus on the simplest networks, i.e., fully connected DNNs. This is because our aim in this paper is to build a framework of DNN-based BSS under a consistent ML criterion and appropriate utilization of DNNs in terms of the parameter optimization. All types of state-of-the-art DNN architectures and methods are available to further enhance the separation performance, which is a future work and beyond the scope of this paper.

The outline of DNN training is depicted in Fig. 5. To prepare the training data of mixed signals, we define the following vectors:

$$\vec{\mathbf{s}}_{jn} = (\tilde{\mathbf{s}}_{(j-2c)n}^T, \tilde{\mathbf{s}}_{(j-2c+2)n}^T, \dots, \tilde{\mathbf{s}}_{(j+2c)n}^T)^T \quad (18)$$

$$\vec{\mathbf{b}}_j = \frac{\sum_n \alpha_{jn} \vec{\mathbf{s}}_{jn}}{\|\sum_n \alpha_{jn} \vec{\mathbf{s}}_{jn}\|_2 + \delta_2}, \quad (19)$$

$$\bar{\mathbf{s}}_{jn} = \frac{\alpha_{jn} \vec{\mathbf{s}}_{jn}}{\|\sum_n \alpha_{jn} \vec{\mathbf{s}}_{jn}\|_2 + \delta_2}, \quad (20)$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\tilde{\mathbf{s}}_{jn} \in \mathbb{C}^I$ is the STFT vector of the n th source at j (the column vector of $\tilde{\mathbf{S}}_n$), $\vec{\mathbf{s}}_{jn} \in \mathbb{C}^{I(2C+1)}$ is a vector that vertically concatenates $\tilde{\mathbf{s}}_{jn}$ for $2c$ frames around j as shown in Fig. 5, $\vec{\mathbf{b}}_j \in \mathbb{C}^{I(2C+1)}$ is the normalized mixture vector whose amplitude $|\vec{\mathbf{b}}_j|^1$ is an input vector for all DNN_n , $\bar{\mathbf{s}}_{jn} \in \mathbb{C}^I$ is the reference vector for each source, α_{jn} is a random variable in the range $[0.05, 1]$, which controls the signal-to-noise ratio (SNR) in $\vec{\mathbf{b}}_j$, and δ_2 is a small value to avoid division by zero. DNN_n is optimized so that the loss function (16) or (17) between the output (estimated) vector $\mathbf{d}_{jn} \in \mathbb{R}_{\geq 0}^I$ and the reference vector $|\bar{\mathbf{s}}_{jn}|^1$ is minimized, where \mathbf{d}_{jn} is the column vector of \mathbf{D}_n . The input and output vectors of DNN_n are $|\vec{\mathbf{b}}_j|^1$ and \mathbf{d}_{jn} , respectively.

D. Update Rule of Demixing Matrix

The cost function (12) consists of a negative log-determinant term of \mathbf{W}_i and a quadratic form of \mathbf{w}_{in} , and the minimization of (12) w.r.t. \mathbf{W}_i leads to the solution that maximizes the independence between sources, given the scale parameter σ_{ijn} estimated by the DNN as a fixed value. In AuxICA [4] and AuxIVA [5], an efficient and convergence-guaranteed optimization

algorithm called IP was proposed, which can be applied to the sum of a negative log-determinant and a quadratic form. Since IP can find the stationary point of $\mathcal{L}_{\text{Gauss}}$, we can update the demixing matrix \mathbf{W}_i in a vectorwise iterative calculation as follows:

$$\mathbf{U}_{in} = \frac{1}{J} \sum_j \frac{1}{\sigma_{ijn}^2} \mathbf{x}_{ij} \mathbf{x}_{ij}^H. \quad (21)$$

$$\mathbf{w}_{in} \leftarrow (\mathbf{W}_i \mathbf{U}_{in})^{-1} \mathbf{e}_n, \quad (22)$$

$$\mathbf{w}_{in} \leftarrow \frac{\mathbf{w}_{in}}{\sqrt{\mathbf{w}_{in}^H \mathbf{U}_{in} \mathbf{w}_{in}}}, \quad (23)$$

where \mathbf{e}_n denotes the unit vector with the n th element equal to unity. After calculating \mathbf{w}_{in} for all n , the estimated signal y_{ijn} is updated by (5).

In t -IDLMA, IP cannot be applied to (14) because $|\mathbf{w}_{in}^H \mathbf{x}_{ij}|^2$ is intrinsic in the logarithm function. Therefore, we use an MM algorithm to transform the cost function (14) into an IP-applicable form, namely, the sum of the negative log-determinant and the quadratic form. The following derivation is based on [16], [17], but in this paper, we newly show a different form of the update rules to interpret the relation between the DNN output σ_{ijn} and its smoothing effect. To design a majorization function for (14), we apply the tangent line inequality

$$\log \xi \leq \frac{1}{\gamma} (\xi - \gamma) + \log \gamma \quad (24)$$

to the logarithm term in (14), where $\xi > 0$ is the original variable and $\gamma > 0$ is an auxiliary variable. The majorization function can be designed as

$$\begin{aligned} \mathcal{L}_t &\leq \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right) \frac{1}{\gamma_{ijn}} \left(1 + \frac{2}{\nu} \frac{|\mathbf{w}_{in}^H \mathbf{x}_{ij}|^2}{\sigma_{ijn}^2} - \gamma_{ijn}\right) \right. \\ &\quad \left. + \left(1 + \frac{\nu}{2}\right) \log \gamma_{ijn} + 2 \log \sigma_{ijn} \right] \\ &\quad - 2J \sum_i \log |\det \mathbf{W}_i| + IJN \log \pi \\ &= -2J \sum_i \log |\det \mathbf{W}_i| + J \sum_{i,n} \mathbf{w}_{in}^H \mathbf{U}_{in} \mathbf{w}_{in} + \mathcal{C} \\ &=: \mathcal{L}_t^+, \end{aligned} \quad (25)$$

$$\mathbf{U}_{in} = \frac{1}{J} \left(\frac{2}{\nu} + 1 \right) \sum_j \frac{1}{\gamma_{ijn} \sigma_{ijn}^2} \mathbf{x}_{ij} \mathbf{x}_{ij}^H, \quad (26)$$

where \mathcal{C} includes the constant terms that do not depend on \mathbf{w}_{in} , γ_{ijn} is the auxiliary variable, and \mathcal{L}_t and \mathcal{L}_t^+ become equal if and only if

$$\gamma_{ijn} = 1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{\sigma_{ijn}^2}. \quad (27)$$

By substituting (27) into (26), we obtain

$$\mathbf{U}_{in} = \frac{1}{J} \sum_j \frac{1}{\zeta_{ijn}} \mathbf{x}_{ij} \mathbf{x}_{ij}^H, \quad (28)$$

Algorithm 1: Iterative Algorithm of IDLMA.

```

1: function IDLMA ( $\mathbf{X}_1, \dots, \mathbf{X}_M, \text{DNN}_0, \dots, \text{DNN}_N$ )
2:   for  $l$  of number of iterations  $L$  do
3:     for  $l'$  of number of spatial updates  $L'$  do
4:       for all frequency bin  $i$  and source index  $n$  do
5:         Update  $\mathbf{w}_{in}$  by (22) and (23) with (21)
           (Gauss-IDLMA) or (28) ( $t$ -IDLMA)
6:       end for
7:     end for
8:     for all source index  $n$  do
9:       Update  $\mathbf{Y}_n$  by (5)
10:    end for
11:    for all source index  $n$  do
12:      Apply back-projection to obtain  $\hat{\mathbf{Y}}_n$  by (30)
13:    end for
14:    for all source index  $n$  do
15:      Update source model  $\Sigma_n$  by (13)
16:    end for
17:  end for
18:  return separated signals  $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N$ 
19: end function

```

where

$$\zeta_{ijn} = \frac{\nu}{\nu + 2} \sigma_{ijn}^2 + \frac{2}{\nu + 2} |y_{ijn}|^2. \quad (29)$$

Since (25) is the IP-applicable form, the update rule for \mathbf{w}_{in} can be obtained as (22) and (23) with (28). Note that y_{ijn} in ζ_{ijn} comes from the equality condition (27), which means that it is determined by \mathbf{W}_i one step before in the MM algorithm and is a constant when \mathbf{W}_i is updated. In particular, when $\nu \rightarrow \infty$, the majorization function (25) converges to the cost function in Gauss-IDLMA, (12), and (28) also converges to (21). Interestingly, (28) and (29) works as the spectral smoothing of the DNN output, which will be explained in Section III-E in detail.

To fix the scales of y_{ijn} among the frequency bins, the following back-projection technique [31] is applied before updating Σ_n using (13):

$$\hat{y}_{ijn} \leftarrow [\mathbf{W}_i^{-1} (\mathbf{e}_n \circ \mathbf{y}_{ij})]_{m_{\text{ref}}}, \quad (30)$$

where \hat{y}_{ijn} is the scale-fitted estimated signal whose spectrogram form is denoted as $\hat{\mathbf{Y}}_n \in \mathbb{C}^{I \times J}$, \circ is the Hadamard product (element-wise product), $[\cdot]_n$ is the n th value of the vector, and m_{ref} is the index of the reference channel. Note that back-projection technique is essential because DNN training data are scale-fitted w.r.t. the frequency bin. The algorithm of IDLMA is summarized in Algorithm 1.

E. Relation Between Parameter ν and Numerical Stability

In Gauss-IDLMA, \mathbf{U}_{in} defined by (21) can be interpreted as the instantaneous covariance matrix $\mathbf{x}_{ij} \mathbf{x}_{ij}^H$ weighted by σ_{ijn}^{-2} . In general, σ_{ijn} is estimated by DNN_n , whose output likely fluctuates, resulting in many spectral chasms in the time-frequency

plane. Therefore, the weight coefficient σ_{ijn}^{-2} may be excessively large, reducing the numerical stability of IP in Gauss-IDLMA. In t -IDLMA, on the other hand, ζ_{ijn} in (28) is the point internally dividing σ_{ijn}^2 and $|y_{ijn}|^2$ with a ratio of $\nu : 2$. Since y_{ijn} is the output of a spatial linear filter w_{in} , $|y_{ijn}|^2$ contains fewer chasms than σ_{ijn}^2 ; this yields a beneficial spectral smoothing and numerical stability in optimization.

A prospective drawback of t -IDLMA is slower convergence, especially in the case of small ν close to unity, because the inference of the DNN is discounted. Thus, there is a tradeoff when setting ν . The appropriate selection of ν will be discussed in Section V.

IV. SOURCE MODEL ADAPTATION BASED ON DATA AUGMENTATION IN SEMI-SUPERVISED CASE

A. Strategy

In this section, we focus on the semi-supervised case, where two training datasets are available: (a) a dataset of solo-recorded source signals (hereafter referred to as a supervised source) and (b) a dataset of mixed signals that include diverse sources except for the supervised source. To apply the IDLMA framework even in such a semi-supervised case, we introduce a new data augmentation scheme to the framework and propose a semi-supervised version of IDLMA, which is a more practical application with limited data resources. To distinguish the proposed methods for full- and semi-supervised cases, hereafter, Algorithm 1 is referred to as *full-supervised IDLMA (full-IDLMA)*, and the method we discuss in this section is referred to as *semi-supervised IDLMA (semi-IDLMA)*.

For simplicity, let \mathcal{S}_1 ($n = 1$) be the supervised source, and there are no solo-recorded datasets for the other sources ($n \neq 1$). In such a semi-supervised case, only two DNNs (source models) can be trained from the datasets of the supervised source and the mixture of diverse sources; namely, only DNN_1 and $\overline{\text{DNN}}_1$ are obtained in the training stage, where DNN_1 is a source model that estimates the scale parameter of only the supervised source $\tilde{\mathcal{S}}_1$ from the diverse mixture, and $\overline{\text{DNN}}_1$ is a source model that estimates the scale parameter of the other residual sources in the mixture. That is, DNN_1 enhances only $\tilde{\mathcal{S}}_1$ and $\overline{\text{DNN}}_1$ suppresses only $\tilde{\mathcal{S}}_1$. Even if we apply full-IDLMA in the semi-supervised situation, the separation accuracy is highly limited. This is because the parameter estimation of $\overline{\text{DNN}}_1$ does not work well compared with DNN_n ($n \neq 1$) in the full-supervised case.

To cope with this problem, we propose to adapt the pretrained DNN source models DNN_1 and $\overline{\text{DNN}}_1$ to the supervised source \mathcal{S}_1 and the other sources $\mathcal{S}_2, \dots, \mathcal{S}_N$ in the observed mixture \mathbf{X}_m , respectively. Fig. 6 shows the process flow in semi-IDLMA. This method consists of two processes: a source separation process and a data adaptation process, where the source separation process is the same as full-IDLMA. The adaptation process is carried out by the following steps: (a) the DNN source model for the other sources, $\overline{\text{DNN}}_1$, is copied and set to the source models of each unsupervised source ($n \neq 1$) in the observed mixture as $\text{DNN}_2, \dots, \text{DNN}_N$, (b) BSS based on IL-RMA is performed on the observed mixture to initialize the estimated signals $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N$, (c) the new datasets for each source

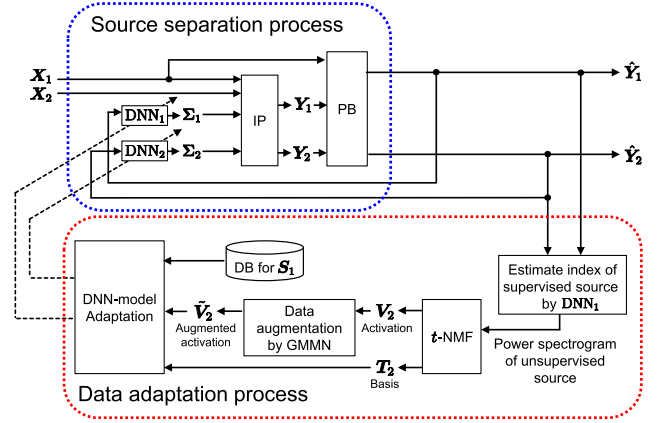


Fig. 6. Process flow of semi-IDLMA when $N = M = 2$, supervised source is $\hat{\mathbf{Y}}_1$, and unsupervised source is $\hat{\mathbf{Y}}_2$.

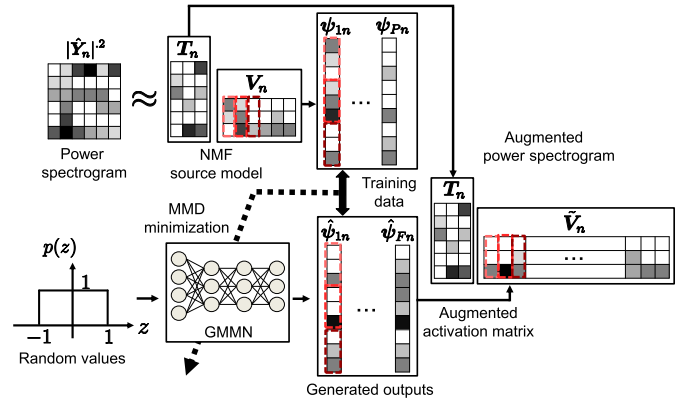


Fig. 7. Data augmentation based on NMF and GMMN when $I = 6$, $J = 6$, $K = 3$, and $\varphi = 3$.

$\mathcal{S}_1, \dots, \mathcal{S}_N$ are produced via a data augmentation technique using the current estimated signals $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N$, (d) all the DNN source models $\text{DNN}_1, \dots, \text{DNN}_N$ are retrained (updated) using the newly produced datasets for $\mathcal{S}_1, \dots, \mathcal{S}_N$, and (e) IDLMA is performed using the up-to-date $\text{DNN}_1, \dots, \text{DNN}_N$. Intriguingly, steps (c)–(e) (data augmentation, DNN retraining, and separation) can be iterated until a satisfactory performance is obtained by IDLMA, which can be interpreted as an adaptation to the test data.

B. Data Augmentation Utilizing Moment Matching Network

In this subsection, we explain the proposed data augmentation technique, which is summarized in Fig. 7. To improve the DNN performance with the augmented data, it is important to preserve both the statistical and structural properties in the data augmentation. The proposed method combines a generative moment matching network (GMMN) [32] and NMF to represent the spectral and time-series structures in $\hat{\mathbf{Y}}_n$, respectively.

First, the low-rank source model $\mathbf{T}_n \mathbf{V}_n$ is obtained by applying NMF to the current estimated signal $\hat{\mathbf{Y}}_n$, where the basis and activation matrices \mathbf{T}_n and \mathbf{V}_n represent the spectral and time-series structures in $\hat{\mathbf{Y}}_n$, respectively. Second, a new activation

matrix $\tilde{\mathbf{V}}_n$ is generated from random values while minimizing the discrepancy between the statistical properties (moments) of \mathbf{V}_n and $\tilde{\mathbf{V}}_n$ using a GMMN. Third, the augmented power spectrogram for the n th source is produced by multiplying \mathbf{T}_n and $\tilde{\mathbf{V}}_n$, resulting in moment- and structure-preserved appropriate data augmentation for audio signals. Note that the GMMN often fails to train networks for large-dimensional input data; however, in the proposed method, the input data of the GMMN are several time frames of K coefficients for the bases in \mathbf{T}_n , as shown in Fig. 7. Thus, the dimension of the input vector is greatly reduced owing to the NMF decomposition. The drawback of a small K is that NMF cannot represent the signal minutely, which degrades the quality of the augmented data. An appropriate K is determined experimentally.

In the data adaptation process shown in Fig. 6, the index of the supervised source n_s is estimated using DNN_1 as follows:

$$n_s = \arg \max_n \text{DNN}_1(|\hat{\mathbf{Y}}_n|^{-1}). \quad (31)$$

This is necessary because the order of the estimated signals in ILRMA applied in advance depends on the initial values of \mathbf{W}_i , \mathbf{T}_n , and \mathbf{V}_n . Then, a low-rank source model $\mathbf{T}_n \mathbf{V}_n$ of the current estimated signal $\hat{\mathbf{Y}}_n$ for the unsupervised sources ($n \neq 1$) is obtained by applying NMF based on the Student's t distribution (t -NMF) [33] as

$$|\hat{\mathbf{Y}}_n|^2 \approx \mathbf{T}_n \mathbf{V}_n. \quad (32)$$

The update rules of \mathbf{T}_n and \mathbf{V}_n are given by the following equations as [33]

$$t_{ikn} = t_{ikn} \left(\frac{\sum_j \frac{|\hat{y}_{ijn}|^2}{\eta_{ijn} \sum_{k'} t_{ik'n} v_{k'jn}} v_{kjn}}{\sum_j \frac{1}{\sum_{k'} t_{ik'n} v_{k'jn}} v_{kjn}} \right)^{\frac{1}{2}}, \quad (33)$$

$$v_{kjn} = v_{kjn} \left(\frac{\sum_i \frac{|\hat{y}_{ijn}|^2}{\eta_{ijn} \sum_{k'} t_{ik'n} v_{k'jn}} t_{ikn}}{\sum_i \frac{1}{\sum_{k'} t_{ik'n} v_{k'jn}} t_{ikn}} \right)^{\frac{1}{2}}, \quad (34)$$

where

$$\eta_{ijn} = \frac{\nu}{\nu + 2} \sum_{k'} t_{ik'n} v_{k'jn} + \frac{2}{\nu + 2} |\hat{y}_{ijn}|^2. \quad (35)$$

Note that the degree-of-freedom parameter ν should be set to be the same as that used in (17) for t -IDLMA or set to ∞ for Gauss-IDLMA to maintain the consistency with the source generative model $p(y_{ijn})$, where (33) and (34) reduce to the update rules of ISNMF [11] when $\nu \rightarrow \infty$. The estimated source models \mathbf{T}_n and \mathbf{V}_n of the unsupervised sources are used for the data augmentation process explained below.

As shown in Fig. 7, a supervector $\boldsymbol{\psi}_{jn} \in \mathbb{R}_{\geq 0}^{K\varphi}$ is composed by concatenating multiple column vectors of the activation matrix \mathbf{V}_n as

$$\boldsymbol{\psi}_{jn} = [\mathbf{v}_{(j\varphi)n}^T, \mathbf{v}_{(j\varphi+1)n}^T \cdots, \mathbf{v}_{(j\varphi+\varphi-1)n}^T]^T, \quad (36)$$

where φ is the length of the concatenation and $\mathbf{v}_{jn} \in \mathbb{R}_{\geq 0}^K$ is the j th column vector of \mathbf{V}_n . The concatenation allows the GMMN to capture the locally temporal dependences. The GMMN generates another sample of $\boldsymbol{\psi}_{jn}$, which is denoted as $\hat{\boldsymbol{\psi}}_{fn}$, from

uniformly distributed random values z as

$$\hat{\boldsymbol{\psi}}_{fn} = \text{GMMN}(z), \quad (37)$$

where $f = 1, \dots, F$ and F is the arbitrary data length of generated outputs. Since the statistical discrepancy between $\boldsymbol{\psi}_{jn}$ and $\hat{\boldsymbol{\psi}}_{fn}$ is minimized by the GMMN, a new activation matrix $\tilde{\mathbf{V}}_n$ constructed from $\hat{\boldsymbol{\psi}}_{fn}$ maintains the statistical properties in \mathbf{V}_n . Thus, we can reproduce the augmented spectrogram of the unsupervised sources as $\mathbf{T}_n \tilde{\mathbf{V}}_n$. This can be interpreted as an appropriate data augmentation for acoustic signals because both the time-frequency structure and the statistical properties are maintained by NMF and the GMMN, respectively. The training criterion for $\text{GMMN}(\cdot)$ is the maximum mean discrepancy (MMD) [32], which is defined as

$$\begin{aligned} L_{\text{MMD}} = & \frac{1}{P^2} \sum_{j=1}^P \sum_{j'=1}^P \kappa(\boldsymbol{\psi}_j, \boldsymbol{\psi}_{j'}) \\ & + \frac{1}{F^2} \sum_{f=1}^F \sum_{f'=1}^F \kappa(\hat{\boldsymbol{\psi}}_f, \hat{\boldsymbol{\psi}}_{f'}) \\ & - \frac{2}{PF} \sum_{j=1}^P \sum_{f=1}^F \kappa(\boldsymbol{\psi}_j, \hat{\boldsymbol{\psi}}_f), \end{aligned} \quad (38)$$

where $P = \lfloor J/\varphi \rfloor$ is the length of the training data, $\lfloor \cdot \rfloor$ is the floor function, $\kappa(\mathbf{b}_1, \mathbf{b}_2) = \exp(-\|\mathbf{b}_1 - \mathbf{b}_2\|_2^2 / (2\rho^2))$ is a Gaussian kernel, and ρ^2 is the variance. The criterion (38) is based on a kernel trick and becomes zero when all the moments of $\hat{\boldsymbol{\psi}}_{fn}$ are identical to those of $\boldsymbol{\psi}_{jn}$.

The augmented activation matrix $\tilde{\mathbf{V}}_n$ can be obtained from $\hat{\boldsymbol{\psi}}_{fn}$ as

$$\begin{aligned} \tilde{\mathbf{V}}_n = & [[\hat{\boldsymbol{\psi}}_{1n}]_{1:K}^T, [\hat{\boldsymbol{\psi}}_{1n}]_{K+1:2K}^T, \\ & [\hat{\boldsymbol{\psi}}_{1n}]_{2K+1:3K}^T, \dots, [\hat{\boldsymbol{\psi}}_{Fn}]_{2K+1:3K}^T]^T, \end{aligned} \quad (39)$$

where $[\cdot]_{q_1:q_2}$ returns a vector whose elements are the q_1 th to q_2 th elements in the input vector.

C. DNN Model Adaptation

After the data augmentation for the unsupervised sources, we prepare a new mixture dataset using the database of the supervised source and the augmented data. Then, the source-wise DNNs are updated (retrained) by a transfer learning technique [34], where the parameters in only a few layers are updated. Detailed information is described in Section V-B. The input and output vectors used for retraining DNNs are the same as those in the full-supervised case described in Section III-C. The algorithm of semi-IDLMA is summarized in Algorithm 2.

V. EXPERIMENTAL EVALUATION

A. Full-Supervised Case

1) *Task, Dataset, and Conditions:* We confirmed the validity of the proposed method by conducting a music source separation task. We compared eight methods: MNMF (blind, $K = 20$), ILRMA (blind, $K = 20$), basis-supervised NMF+WF ($K = 50$),

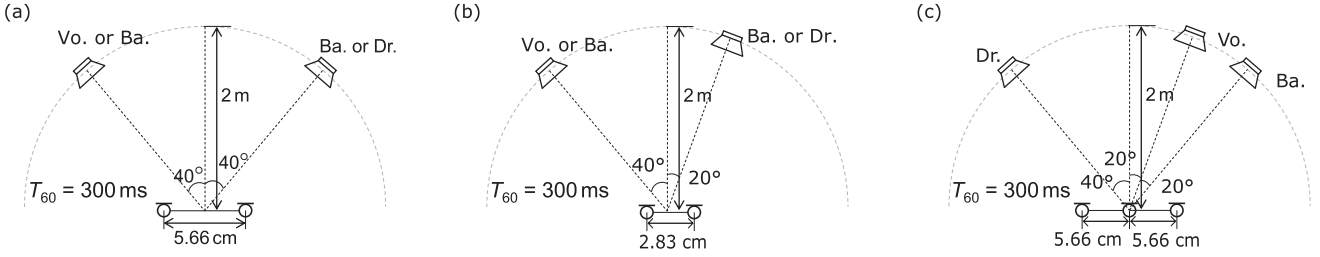


Fig. 8. Recording conditions of impulse responses obtained from RWCP database.

Algorithm 2: Iterative Algorithm of Semi-IDLMA.

```

1: function SEMI-IDLMA( $\mathbf{X}_1, \dots, \mathbf{X}_M, \text{DNN}_1, \text{DNN}_1$ )
2:   for  $n = 2, \dots, N$  do
3:      $\text{DNN}_n \leftarrow \text{DNN}_1$ 
4:   end for
5:   Initialize  $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N$  by ILRMA
6:   for  $l$  of number of iterations  $L$  do
7:      $\text{DNN}_1, \dots, \text{DNN}_N \leftarrow \text{ADAPTATION}$ 
       ( $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N,$ 
         $\text{DNN}_1, \dots, \text{DNN}_N$ )
8:     for  $l'$  of number of iterations  $L'$  do
9:        $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N \leftarrow \text{IDLMA}(\mathbf{X}_1, \dots, \mathbf{X}_M,$ 
         $\text{DNN}_1, \dots, \text{DNN}_N)$ 
10:    end for
11:  end for
12:  return  $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N$ 
13: end function

15: function ADAPTATION ( $\hat{\mathbf{Y}}_1, \dots, \hat{\mathbf{Y}}_N, \text{DNN}_1,$ 
     $\dots, \text{DNN}_N$ )
16:  Estimate supervised source index  $n_s$ :
17:   $n_s = \arg \max_n \text{DNN}_1(|\hat{\mathbf{Y}}_n|^{-1})$ 
18:  for all  $n$  except  $n_s$  do
19:    Decompose  $\hat{\mathbf{Y}}_n$  into  $\mathbf{T}_n$  and  $\mathbf{V}_n$  by (33) and (34)
20:    Train GMMN with  $\mathbf{V}_n$ :
21:    Compose  $\psi_{j_n}$  by (36)
22:    Train GMMN with loss function (38)
23:    Generate augmented activation  $\tilde{\mathbf{V}}_n$  by (39)
24:    Add  $\mathbf{T}_n \tilde{\mathbf{V}}_n$  to training dataset for  $\text{DNN}_n$ 
25:  end for
26:  Adapt  $\text{DNN}_n$  for all  $n$  with new training dataset
27:  return  $\text{DNN}_1, \dots, \text{DNN}_N$ 
28: end function

```

DNN+WF, basis-supervised MNMF ($K = 50$ for each source), basis-supervised ILRMA ($K = 50$), FSCM+DNN, and proposed IDLMA. In basis-supervised NMF+WF, basis-supervised MNMF, and basis-supervised ILRMA, the basis matrices \mathbf{T}_n of NMF for all n were trained using solo-recorded datasets before source separation. Basis-supervised NMF+WF and DNN+WF apply a Wiener filter constructed using all the outputs of the NMF source models and the DNN source models, respectively,

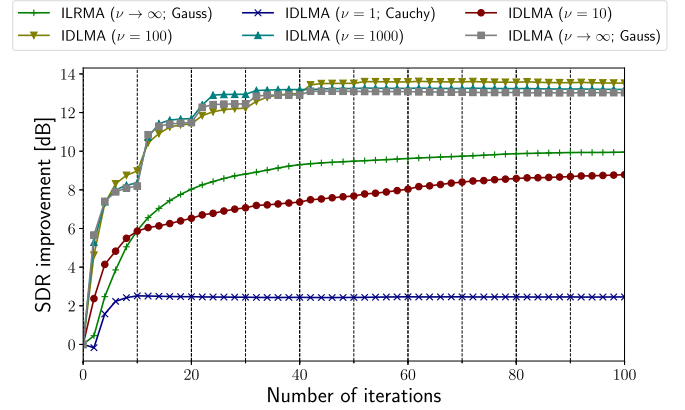


Fig. 9. Example of SDR improvements for each method for Ba./Vo.

to the reference channel signal [21]. Note that MNMF and ILRMA are “blind” (unsupervised) techniques, but we show their performances just for reference to understand to what extent the supervised methods (basis-supervised NMF+WF, DNN+WF, basis-supervised MNMF, basis-supervised ILRMA, FSCM+DNN, and IDLMA) can improve the performance by using pretraining data. For FSCM+DNN and IDLMA, the scale parameter matrix Σ_n was updated by DNN_n after every 10 iterations of the spatial parameter optimization. Note that the DNNs employed in this paper were different from those of original FSCM+DNN [27] as follows: (a) each DNN was prepared for each single source, and (b) each DNN was trained under multiple-SNR conditions. We used the same DNNs for all the methods, where the selection of the best DNN architecture is beyond the scope of this paper. In the experiments described in Section V-A, all the methods except MNMF and ILRMA were conducted in a full-supervised manner, i.e., we trained all basis matrices and DNNs for each of the sources in advance using datasets of a sufficient size.

We used the DSD100 dataset of SiSEC2016 [35] as the dry sources and the training datasets of each basis matrix and DNN. The 50 songs in the dev data were used to train \mathbf{T}_n and DNN_n , and the top 25 songs in alphabetical order in the test data were used for performance evaluation. For the pretraining of NMF, we excerpted a 10-s-long part from each solo-recorded song and concatenated them to prepare the input data of NMF for a specific instrument [36]. The basis matrices \mathbf{T}_n were trained with 200 iterations in t -NMF applied to each isolated spectrogram. The test songs were trimmed only in the interval

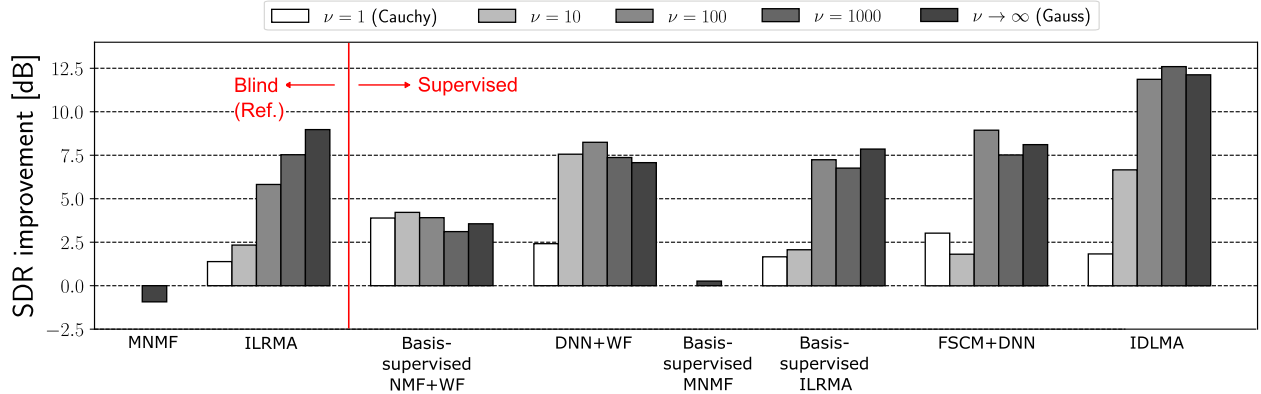


Fig. 10. Average SDR improvements of 25 Ba./Vo. songs.

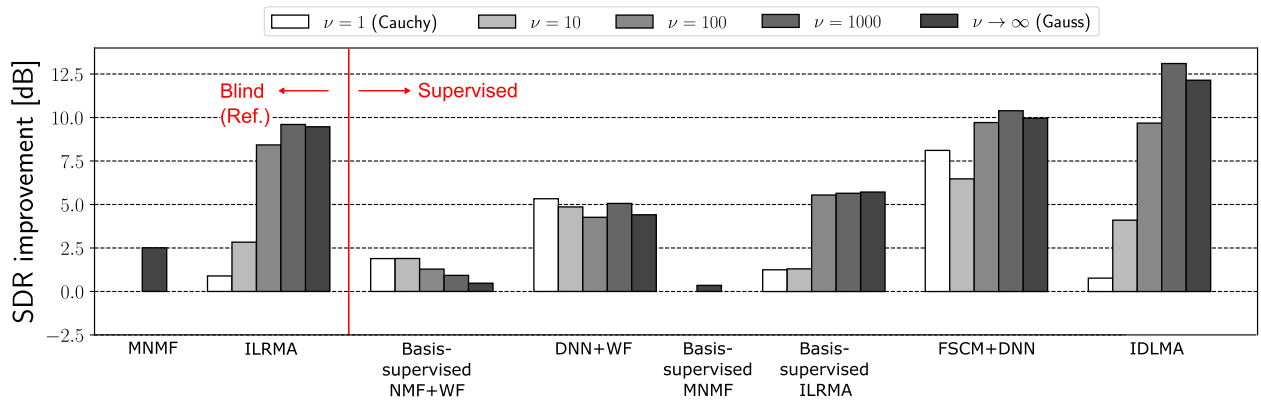


Fig. 11. Average SDR improvements of 25 Dr./Vo. songs.

of 30 to 60 s. To simulate reverberant mixtures, we produced two-channel and three-channel observed signals by convoluting the impulse response E2A ($T_{60} = 300$ ms) obtained from the RWCP database [37] with each source, and mixtures of bass (Ba.) and vocals (Vo.) (Ba./Vo.); drums (Dr.) and Vo. (Dr./Vo.); Ba. and Dr. (Ba./Dr.); Vo. and other (Vo./other); and Ba., Vo., and Dr. (Ba./Vo./Dr.) were created. We compared MNMF and basis-supervised MNMF only for two-source separation because it has been revealed that they always achieve lower separation performance than ILRMA [6], [7]. The recording conditions of E2A are shown in Fig. 8. All the signals were downsampled to 8 kHz. An STFT was performed using a 512-ms-long Hamming window with a 256-ms-long shift in Ba./Vo. separation and a 256-ms-long Hamming window with a 128-ms-long shift in the other cases. We used the signal-to-distortion ratio (SDR) as the total separation performance [38]. Note that, in this paper, we omitted the signal-to-interference ratio (SIR) and the signal-to-artifact ratio (SAR) because their tendencies are the same as that of SDR.

In this paper, the number of hidden layers in the constructed fully connected DNN was set to four. Each layer had 1024 units, and a rectified linear unit was used for the output of each layer. To optimize the DNN, we added the term $(\lambda/2) \sum_q g_q^2$ to (16) or (17) for regularization, where g_q is the weight coefficient in DNN, and ADADELTA [39] with a 128-size minibatch was performed for 2000 epochs. The parameter ε was experimentally

optimized and set to $0.1 \times (IJ)^{-1} \sum_{i,j} \sigma_{ijn}$. The other parameters were set to $\delta_1 = \delta_2 = 10^{-5}$, $c = 3$, and $\lambda = 10^{-5}$.

2) *Comparison of Separation Performance*: Fig. 9 depicts an example of the convergence behaviors of ILRMA and IDLMA. These results show that (a) the DNN source model leads the demixing matrix to more accurate estimation, resulting in a significant leap of SDR improvement, and (b) a larger ν provides a faster spatial model update, but t -IDLMA with an appropriate ν ($=100$) converges to a higher SDR than Gauss-IDLMA ($\nu = \infty$), as mentioned in Section III-E.

Figures 10 to 13 show the average SDR improvements of the 25 test songs for Ba./Vo., Dr./Vo., Ba./Dr., and Vo./other (two-source mixing cases). These results are the average of the cases for the recording conditions (a) and (b) in Fig. 8. We can confirm that the proposed IDLMA outperforms the other methods for Ba./Vo., Dr./Vo., and Vo./other. In particular, t -IDLMA with a weakly super-Gaussian distribution, i.e., $\nu = 100$ or 1000, achieves better separation performance than Gauss-ILDMA ($\nu = \infty$), showing the efficacy of introducing the source generative model using the Student's t distribution.

In Ba./Dr. separation, ILRMA achieves the best separation performance among the methods. This is because both Ba. and Dr. have a very simple time-frequency structure, which fits the low-rank model of ILRMA very well. However, when we separate a source whose time-frequency structure is complex, such as vocals, the low-rank model of ILRMA does not fit well and

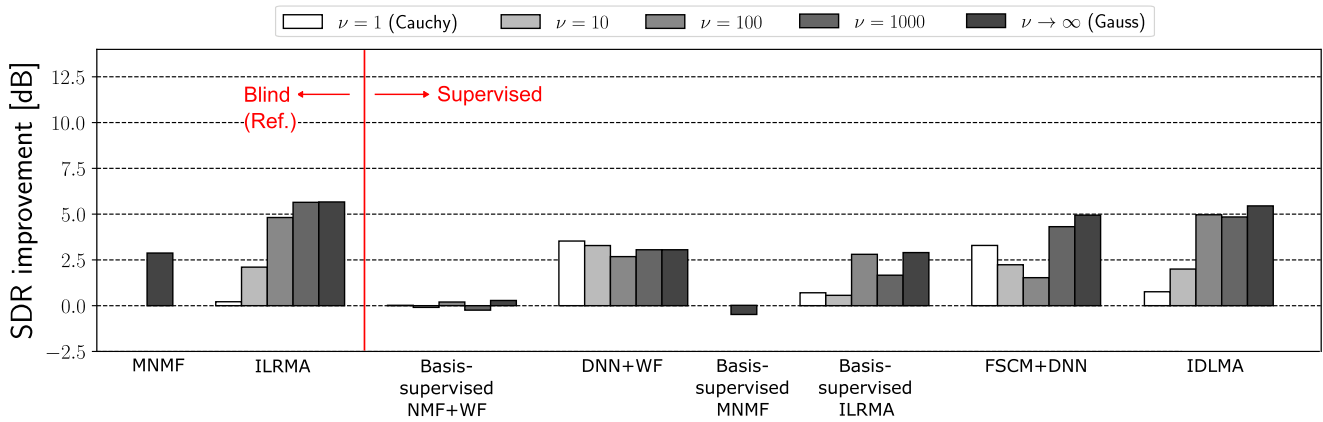


Fig. 12. Average SDR improvements of 25 Ba./Dr. songs.

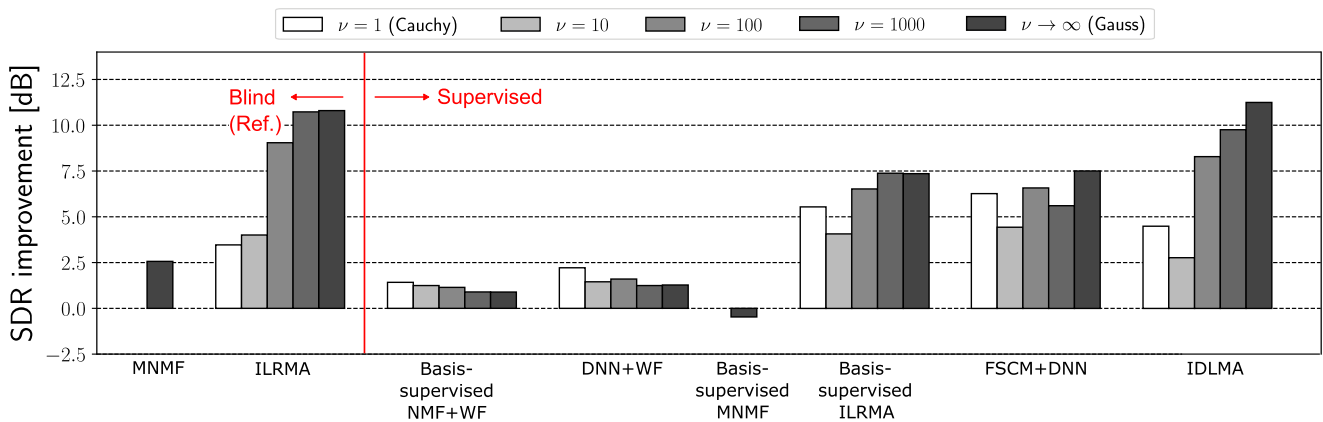


Fig. 13. Average SDR improvements of 25 Vo./other songs.

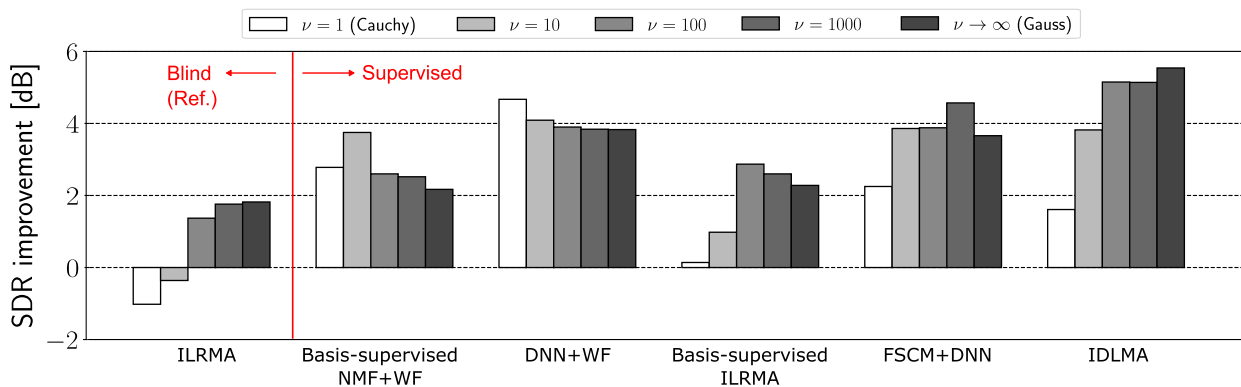


Fig. 14. Average SDR improvements of 25 Ba./Vo./Dr. songs for recording environment (c).

the proposed IDLMA achieves much higher separation performance.

In Vo./other separation, it is difficult for the DNN to learn the structure of the source “other” because it contains pianos, guitars, a mixture of them, and so forth. Indeed, the results for DNN+WF in Fig. 13 indicate that the net separation performance of the DNN is rather low. Nevertheless, IDLMA achieves the best separation performance because it combines DNN inference of the source model and blind estimation of the spatial model. Blind estimation of the spatial model separates the mixtures well by

utilizing the independence between the sources and makes the DNN estimation of the source model easier. Therefore, IDLMA, which iteratively updates the spatial model and the DNN source model, achieves high separation performance.

Figure 14 shows the average SDR improvements of the 25 test songs for Ba./Vo./Dr. (three-source mixing case). We can confirm that the proposed IDLMA outperforms the other methods even in three-source separation.

Note that the separation performance of the basis-supervised methods did not always outperform the blind methods. This is

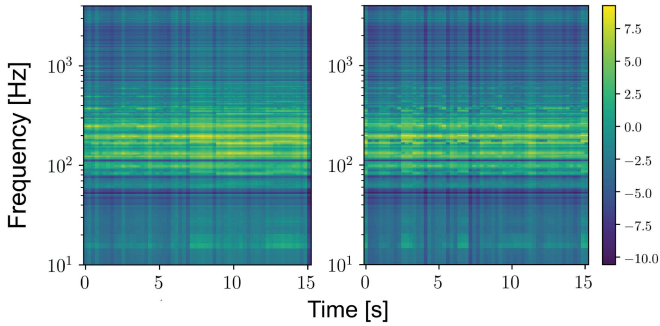


Fig. 15. (a) Spectrogram of real bass represented by NMF. (b) Pseudo-spectrogram of bass augmented by proposed acoustical GMMN.

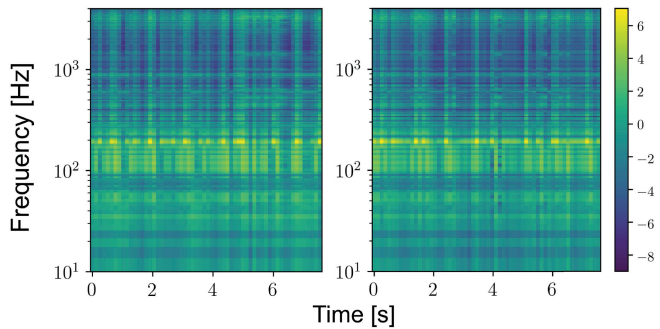


Fig. 16. (a) Spectrogram of real drums represented by NMF. (b) Pseudo-spectrogram of drums augmented by proposed acoustical GMMN.

because the generalizing capability of the basis trained in NMF is poor and the trained basis cannot express the test data well. On the other hand, the DNN can represent the test data relatively well, and this is the key factor in the successful separation in IDLMA.

3) *Computational Time*: To show the efficiency of the proposed approach, we compared the computational times of ILRMA, FSCM+DNN, and IDLMA for 100 iterations of spatial optimization. We used Python 3.5.2 (64-bit) and Chainer 2.1.0 with an Intel Core i7-6850K (3.60 GHz, 6 Cores) CPU. To calculate the DNN outputs, a GeForce GTX 1080Ti GPU was utilized. Examples of computational times were 23.3 s for ILRMA, 287.1 s for FSCM+DNN, and 26.6 s for IDLMA. These results reveal that the proposed method is as fast as conventional ILRMA and more than 10 times faster than FSCM+DNN.

B. Semi-Supervised Case

1) *Task, Dataset, and Conditions*: To confirm the validity of the proposed semi-IDLMA with DNN model adaptation, we conducted a music source separation with a limited dataset. In this experiment, we assume vocals as the supervised source. We used the dataset described in Section V-A except that only vocals and the mixture of the other sources (Ba., Dr., and “other”) in the dev data were used to train DNN_1 and \overline{DNN}_1 in advance. We compared three methods: Gauss-ILRMA (blind, $K = 5$), semi-IDLMA without DNN model adaptation, and the proposed semi-IDLMA with DNN model adaptation. We added

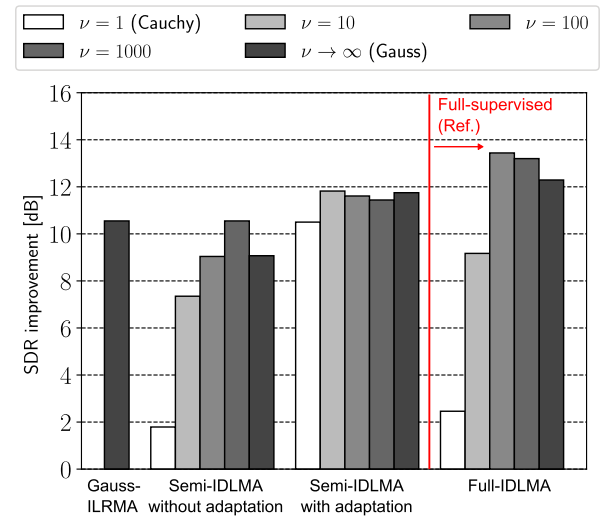


Fig. 17. Average SDR improvements of 25 Ba./Vo. songs for recording environment (a) in semi-supervised case.

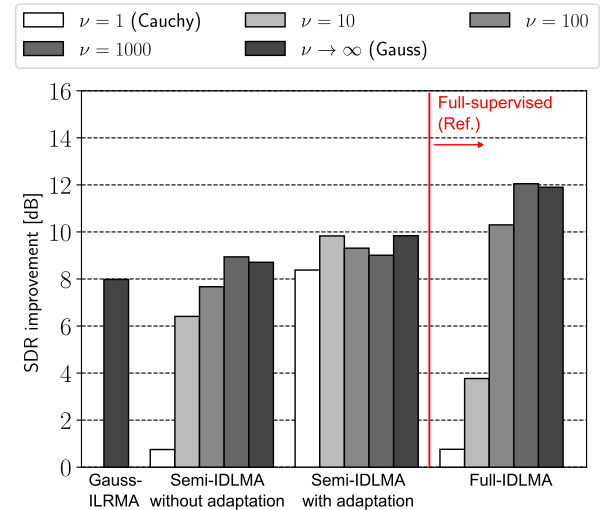


Fig. 18. Average SDR improvements of 25 Ba./Vo. songs for recording environment (b) in semi-supervised case.

semi-IDLMA without DNN model adaptation to confirm the validity of data augmentation and DNN model adaptation. In the proposed semi-IDLMA with DNN model adaptation, since we require well-separated signals, we conducted source separation by Gauss-ILRMA for the first 100 spatial model updates. For IDLMA and semi-IDLMA, the scale parameter matrix Σ_n was updated by DNN_n after every 10 iterations of the spatial optimization. To simulate reverberant mixtures, we produced the two-channel observed signals by convoluting the impulse response E2A ($T_{60} = 300$ ms) as in Section V-A, and the mixtures of Ba. and Vo. (Ba./Vo.) and Dr. and Vo. (Dr./Vo.) were created. All the signals were downsampled to 8 kHz. An STFT was performed using a 512-ms-long Hamming window with a 256-ms-long shift in Ba./Vo. separation and a 256-ms-long Hamming window with a 128-ms-long shift in Dr./Vo. separation.

We constructed a fully connected DNN with three hidden layers for the GMMN. Each layer had 512 units, and a gated

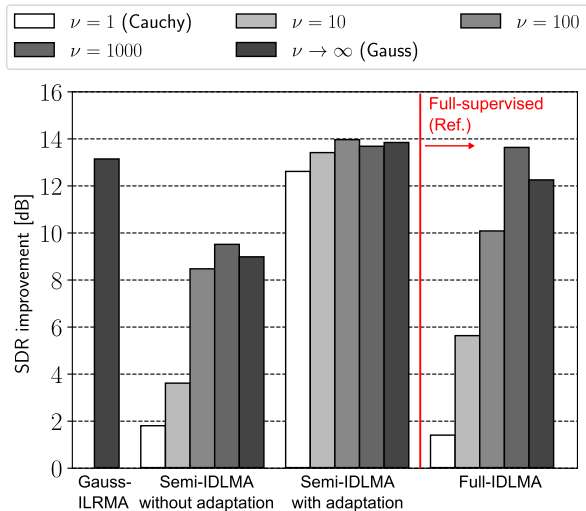


Fig. 19. Average SDR improvements of 25 Dr./Vo. songs for recording environment (a) in semi-supervised case.

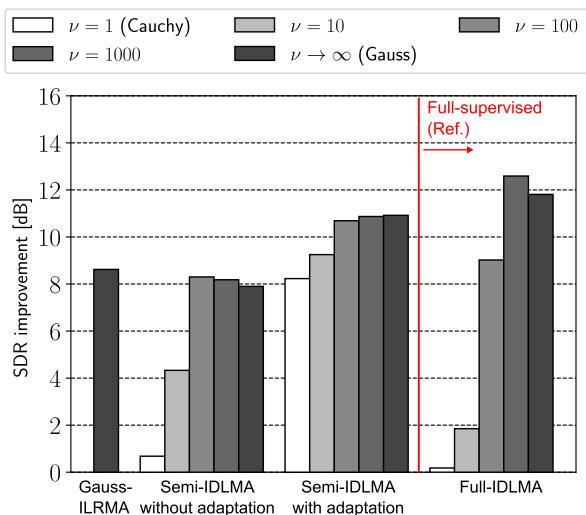


Fig. 20. Average SDR improvements of 25 Dr./Vo. songs for recording environment (b) in semi-supervised case.

activation unit [40] was used for the output of each layer. To optimize the GMMN, ADAGRAD [41] with a 1024-size minibatch was performed for 50 epochs. Adaptation of the DNN was performed by updating the weights and biases between the second hidden layer and the third hidden layer. We set the training epoch to 15 and ρ^2 to 10.0. The other training settings were the same as those in Section III-C.

2) *Example of Data Augmentation*: Figs. 15 and 16 show examples of spectrograms obtained by NMF and spectrograms augmented by the proposed GMMN for the bass and drums, respectively. Note that the frequency scale is logarithmic to show characteristics in detail in the low-frequency range. We observe that the GMMN generates a different spectrogram from the original one while preserving the acoustic structure of each instrument, e.g., most of the power is concentrated at specific

frequencies in the bass (see Fig. 15), and the temporal dynamics is emphasized in the drums (see Fig. 16).

3) *Comparison of Separation Performance*: Figs. 17 to 20 show the average SDR improvements of the 25 test songs for Ba./Vo. and Dr./Vo. in the semi-supervised case. We confirm that the proposed semi-IDLMA with DNN model adaptation achieves the highest separation performance under the limited dataset. In particular, Fig. 19 indicates that the proposed semi-IDLMA with DNN adaptation achieves more accurate separation than full-IDLMA. This is because the proposed semi-IDLMA adapts DNN source models to the specific song to be separated, whereas full-IDLMA uses DNNs trained with various songs that do not include the test data song to be separated. From the results, the effectiveness of the proposed semi-IDLMA with DNN model adaptation is validated.

VI. CONCLUSION

In this paper, we proposed a new determined source separation method that unifies ICA-based blind spatial optimization and DNN-based supervised source model estimation. The proposed method employs the complex Student's t distribution as the source generative model, which is an extension of the LGM. Moreover, we addressed the semi-supervised situation and proposed an appropriate data augmentation scheme for DNN adaptation. From the experimental results, we confirmed that the proposed algorithms outperform other blind and supervised source separation methods in both full- and semi-supervised cases in a music source separation task.

Some improvements remain as a future work. For example, although our spatial model cannot deal with the underdetermined case, the proposed data augmentation scheme can still be utilized in such a case. Moreover, the future work includes extending IDLMA to dynamic acoustic conditions where the source or microphone positions are not fixed. To address this problem, a blockwise batch technique could be utilized as in [42]. The introduction of other generative models such as a complex sub-Gaussian distribution [43], a symmetric alpha-stable distribution [44], and an anisotropic complex distribution [36] also constitutes a future work.

ACKNOWLEDGMENT

The authors would like to thank Dr. K. Kondo and Dr. Y. Takahashi of Yamaha Corporation for their fruitful suggestions and discussions regarding this work.

REFERENCES

- [1] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
- [2] 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, Jan. 2007.
- [3] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Statist.*, vol. 9, no. 1, pp. 60–77, 2000.
- [4] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 165–172.

- [5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2011, pp. 189–192.
- [6] 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. 14, no. 9, pp. 1626–1641, Sep. 2016.
- [7] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed. Cham, Switzerland: Springer, 2018, ch. 6, pp. 125–155.
- [8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [9] 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, Mar. 2006.
- [10] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, 2004.
- [11] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura–Saito divergence. With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.
- [12] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, May 2012.
- [13] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
- [14] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.
- [15] 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, May 2013.
- [16] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex Student's t -distribution for blind audio source separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.
- [17] D. Kitamura *et al.*, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 28, pp. 1–25, 2018.
- [18] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 469–473.
- [19] A. Liutkus, R. Badeau, and G. Richard, "Informed source separation using latent components," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010, pp. 498–505.
- [20] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3734–3738.
- [21] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2135–2139.
- [22] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. 18th Int. Soc. Music Inf. Retrieval*, 2017, pp. 745–751.
- [23] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.
- [24] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 116–120.
- [25] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, "LSTM-based iterative mask estimation and post-processing for multi-channel speech enhancement," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 488–491.
- [26] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep learning based speech beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5389–5393.
- [27] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.
- [28] S. Mogami *et al.*, "Independent deeply learned matrix analysis for multi-channel audio source separation," in *Proc. 26th Eur. Signal Process. Conf.*, 2018, pp. 1571–1575.
- [29] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds, 2014, pp. 2672–2680. [Online]. Available: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
- [30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, p. 14.
- [31] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.
- [32] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1718–1727.
- [33] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 51–55.
- [34] D. Yu and L. Deng, *Automatic Speech Recognition*. London, U.K.: Springer, 2014.
- [35] A. Liutkus *et al.*, "The 2016 signal separation evaluation campaign," in *Proc. Latent Variable Anal. Signal Separation*, 2012, pp. 323–332.
- [36] P. Magron and T. Virtanen, "Complex ISNMF: A phase-aware model for monaural audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 20–31, Jan. 2019.
- [37] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. 2nd Int. Conf. Lang. Resour. Eval.*, 2000, pp. 965–968.
- [38] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
- [39] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, arXiv:1212.5701.
- [40] A. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4797–4805.
- [41] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.
- [42] Y. Mori *et al.*, "Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–17, 2006.
- [43] S. Mogami *et al.*, "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2018, pp. 1684–1691.
- [44] U. Simsekli, A. Liutkus, and A. T. Cemgil, "Alpha-stable matrix factorization," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2289–2293, Dec. 2015.



Naoki Makishima received the B.E. degree in engineering in 2018 from The University of Tokyo, Tokyo, Japan, where he is currently working toward the M.S. degree in information physics and computing. His research interests include audio source separation, signal processing, and machine learning. He is a member of the Acoustical Society of Japan (ASJ). He was the recipient of the 18th best presentation award of ASJ and a Young Scientist Conference Attendance Grant for International Congress on Acoustics 2019.



Shinichi Mogami received the B.E. degree in engineering in 2017 from The University of Tokyo, Tokyo, Japan, where he is currently working toward the M.S. degree in information physics and computing. His research interests include audio signal processing, statistical signal processing, source separation, and machine learning. He is a Member of the IEEE Signal Processing Society and the Acoustical Society of Japan.



Norihiro Takamune received the B.E. degree in engineering and the M.S. degree in information science and technology from The University of Tokyo, Tokyo, Japan, in 2012 and 2015, respectively. He is currently a Researcher with The University of Tokyo. His research interests include music information analysis audio source separation and machine learning.



Daichi Kitamura received the Ph.D. degree from SOKENDAI, Hayama, Japan. He joined The University of Tokyo in 2017 as a Research Associate, and he moved to National Institute of Technology, Kagawa Collage as an Assistant Professor in 2018. His research interests include audio source separation, statistical signal processing, and machine learning. He was the recipient of the Awaya Prize Young Researcher Award from The Acoustical Society of Japan (ASJ) in 2015, Ikushi Prize from Japan Society for the Promotion of Science in 2017, Best Paper

Award from IEEE Signal Processing Society Japan in 2017, and Itakura Prize Innovative Young Researcher Award from ASJ in 2018.



Hayato Sumino received the B.E. degrees from The University of Tokyo, Tokyo, Japan, where he is currently working toward the M.E. degrees. From September 2018 to January 2019, he did research in machine learning and music with the Institute of Research and Coordination Acoustique/Musique, Paris, France. He is also working as a Pianist. He was the recipient of the Gold Medal at the 2017 Chopin International Piano Competition in Asia, in the “Over 18” category, and the Grand Prix of PTNA piano competition in Japan in 2018. He has performed with various

orchestras including the Japan Philharmonic Orchestra, the Chiba Philharmonic Orchestra, and the Brasov Philharmonic Orchestra. His CD, titled “Passion,” has been released in 2019 from Warner Classics Japan.



Shinnosuke Takamichi received the B.E. degree from the Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2013 and 2016, respectively. He is currently an Assistant Professor with The University of Tokyo, Tokyo, Japan. He was the recipient of more than ten paper/achievement awards including the 3rd IEEE Signal Processing Society Japan Young Author Best Paper Award.



Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM Intelligent Systems Laboratory, Tokyo, Japan, in 1993, and the Nara Institute of Science and Technology, Ikoma, Japan, in 2000. Since 2014, he has been a Professor with The University of Tokyo, Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation (BSS), and speech enhancement. He has put his research into the world’s first commercially available

Independent-Component-Analysis-based BSS microphone in 2007. He was the recipient of the paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018. He was also the recipient of the DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. He is an APSIPA Distinguished Lecturer from 2018.



Nobutaka Ono received the B.E., M.S., and Ph.D. degrees from The University of Tokyo, Tokyo, Japan, in 1996, 1998, and 2001, respectively. He became a Research Associate in 2001 and a Lecturer in 2005 with The University of Tokyo. He moved to the National Institute of Informatics in 2011 as an Associate Professor, and moved to Tokyo Metropolitan University in 2017 as a Full Professor. His research interests include acoustic signal processing, machine learning, and optimization algorithms. He was a Chair for the Signal Separation Evaluation Campaign Evaluation Committee in 2013 and 2015, and an Associate Editor for the IEEE

TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2012 to 2015. He is a Senior Member of the IEEE Signal Processing Society and a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee from 2014. He was the recipient of the unsupervised learning ICA pioneer award from SPIE.DSS in 2015.