Independent Deeply Learned Matrix Analysis for Multichannel Audio Source Separation

Shinichi Mogami, Hayato Sumino, Daichi Kitamura,
Norihiro Takamune, Shinnosuke Takamichi, Hiroshi Saruwatari
The University of Tokyo

Nobutaka Ono
Tokyo Metropolitan University

Abstract—In this paper, we address a multichannel audio source separation task and propose a new efficient method called independent deeply learned matrix analysis (IDLMA). IDLMA estimates the demixing matrix in a blind manner and updates the time-frequency structures of each source using a pretrained deep neural network (DNN). Also, we introduce a complex Student's t-distribution as a generalized source generative model including both complex Gaussian and Cauchy distributions. Experiments are conducted using music signals with a training dataset, and the results show the validity of the proposed method in terms of separation accuracy and computational cost.

Index Terms—multichannel audio source separation, independent component analysis, deep neural networks

I. INTRODUCTION

Blind source separation (BSS) is a technique for extracting specific sources from an observed multichannel mixture signal without knowing a priori information about the mixing system. The most commonly used algorithm for BSS in the (over)determined case (number of microphones $\geq$ number of sources) is independent component analysis (ICA) [1]. Recently, *independent low-rank matrix analysis (ILRMA)* [2], [3], which is a unification of independent vector analysis (IVA) [4] and non-negative matrix factorization (NMF) [5], was 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 permutation problem. The source generative model assumed in ILRMA was generalized from a complex Gaussian distribution [2] to complex Student's t-distribution (t-ILRMA) [6] for more robust BSS. As a more general framework, in [7], demixing matrix optimization based on a given spectrogram estimate for the source was proposed, showing that the precise source spectrogram model enables accurate spatial model estimation.

In the underdetermined case (number of microphones $<$ number of sources), the Duong model [8] is a commonly used framework. In the Duong model, frequency-wise spatial covariances, which encode source locations and their spatial spreads, are estimated by an expectation-maximization (EM) algorithm, where the permutation problem must be solved after the optimization. Similarly to ILRMA, an NMF-based low-rank assumption is employed in the Duong model to automatically solve the permutation problem, resulting in multichannel NMF (MNMF) [9], [10]. Note that these algorithms 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 [2].

In supervised (informed) source separation, deep neural network (DNN) has shown promising performance in both single-channel [11] and multichannel source separation [12]. In fact, when sufficient data of the audio sources are available, DNN can effectively model their time-frequency structures. However, it is almost impossible to compose an appropriate and generalized spatial model with DNN from training data observed in a multichannel format. This is because the spatial model depends on many factors, including source and microphone locations, the recording room, and reverberation. Therefore, it is reasonable to combine a pretrained DNN source model and a blind estimation of the spatial model. Nugraha et al. proposed a DNN-based multichannel source separation framework [13] using the Duong model (hereafter referred to as *Duong+DNN*). Although this is a convincing approach, a large computational cost is required to estimate the spatial covariance (the EM algorithm in the Duong model) and the performance is not satisfactory owing to the difficulty of parameter optimization.

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, and the demixing matrix (spatial model) is efficiently optimized using a majorization-minimization (MM) algorithm [14]. Since the proposed method utilizes a time-frequency spectrogram matrix estimated by DNN to optimize the spatial model, we call this method *independent deeply learned matrix analysis (IDLMA)*. Table I shows the relationship between the existing and proposed methods. The spatial model is blindly estimated in all the methods, while the source spectrogram model is estimated by DNN in Duong+DNN and the proposed IDLMA.

II. CONVENTIONAL METHOD

A. Formulation

Let N and M be the numbers of sources and channels, respectively. The short-time Fourier transform (STFT) of the

TABLE I
CLASSIFICATION OF MULTICHANNEL SOURCE SEPARATION METHODS

	Source spectrogram model	
	Blind	Supervised
Mixing model	MNMF [10], [15]	Duong+DNN [13]
Demixing model	ILRMA [2], [6]	**Proposed IDLMA**

multichannel source, observed, and estimated signals are defined as $\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijN})^{\top}, \boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijM})^{\top}$, and $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijN})^{\top}$, where $i = 1, \ldots, I; j = 1, \ldots, J; n = 1, \ldots, N$; and $m = 1, \ldots, M$ are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, and $^{\top}$ denotes the transpose. We also denote these spectrograms as $\boldsymbol{S}_n \in \mathbb{C}^{I \times J}, \boldsymbol{X}_m \in \mathbb{C}^{I \times J}$, and $\boldsymbol{Y}_n \in \mathbb{C}^{I \times J}$, whose elements are s_{ijn}, x_{ijn}, and y_{ijn}, respectively. In ILRMA, the following mixing system is assumed:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \qquad (1)$$

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{iN}) \in \mathbb{C}^{M \times N}$ is a frequency-wise mixing matrix and $\boldsymbol{a}_{in}$ is the steering vector for the nth source. The assumption of the mixing system (1) corresponds to restricting the spatial covariance in the Duong model to a rank-1 matrix [8]. When $M = N$ and $\boldsymbol{A}_i$ is not a singular matrix, the estimated signal $\boldsymbol{y}_{ij}$ can be represented as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \qquad (2)$$

where $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1} = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathsf{H}}$ is the demixing matrix, $\boldsymbol{w}_{in}$ is the demixing filter for the nth source, and $^{\mathsf{H}}$ denotes the Hermitian transpose. ILRMA estimates both $\boldsymbol{W}_i$ and $\boldsymbol{y}_{ij}$ from only the observation $\boldsymbol{x}_{ij}$ assuming statistical independence between s_{ijn} and $s_{ijn'}$, where $n \neq n'$.

B. ILRMA and Its Generalization with Student's t-distribution

In [2], [3], the following time-frequency-varying complex Gaussian source generative model is assumed (hereafter referred to as Gauss-ILRMA):

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{1}{\pi \sigma_{ijn}^2} \exp\left(-\frac{|y_{ijn}|^2}{\sigma_{ijn}^2}\right), \qquad (3)$$

$$\sigma_{ijn}^2 = \sum_k t_{ikn} v_{kjn}, \qquad (4)$$

where σ_{ijn} is the variance (source spectrogram model), $k = 1, \ldots, K$ is the index of the bases, and t_{ikn} and v_{kjn} are the parameters in the NMF-based low-rank model. We also denote the variance matrix as $\boldsymbol{\Sigma}_n \in \mathbb{R}_{\geq 0}^{I \times J}$, whose elements are σ_{ijn}. In t-ILRMA [6], (3) is generalized to a complex Student's t-distribution as follows:

$$\prod_{i,j} p(y_{ijn}) = \prod_{i,j} \frac{1}{\pi \sigma_{ijn}^2} \left(1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{\sigma_{ijn}^2}\right)^{-\frac{2+\nu}{2}}, \quad (5)$$

$$\sigma_{ijn}^p = \sum_k t_{ikn} v_{kjn}, \qquad (6)$$

where ν is the degree-of-freedom parameter and p is the domain parameter. When $\nu \to \infty$ and $p = 2$, (5) and (6) become (3) and (4), respectively. Also, (5) with $\nu = 1$ represents the Cauchy-distribution likelihood. The demixing matrix $\boldsymbol{W}_i$ and NMF source model $t_{ikn} v_{kjn}$ can be optimized in the maximum-likelihood (ML) sense on the basis of (3) or

(5). Since the low-rank structure of $|\boldsymbol{Y}_n|^{\cdot 2}$ is ensured by the NMF source model, the permutation problem can be avoided, where $|\cdot|^{\cdot p}$ for matrices denotes the element-wise absolute and pth-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 spectrogram 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 spectrogram model by employing DNN [11]. On the other hand, since the spatial parameters depend on many factors, it is simply impractical to train a general spatial model with 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 demixing matrix $\boldsymbol{W}_i$ and the supervised learning of variance matrix $\boldsymbol{\Sigma}_n$ based on DNN, where the loss function in DNN is designed to maximize the likelihood of the source generative model. In addition, similarly to t-ILRMA, we use a generalized model based on a complex Student's t-distribution including both Gaussian and Cauchy distributions. Duong+DNN also employs DNN that maximizes the likelihood of the Gaussian or Cauchy distribution. However, since the mixing model (spatial covariance) in Duong+DNN is defined by only the Gaussian model, the estimations of the spectral and spatial parameters are inconsistent. In the proposed method, this conflict is resolved by modeling the spatial parameters with the Student's t-distribution model and deriving its optimization algorithm fully consistently in the ML sense.
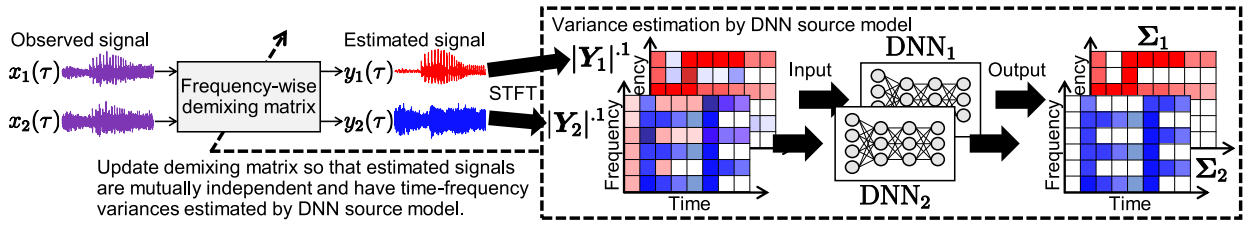
B. Cost Function in IDLMA

Let DNN_n be the DNN source model that enhances the nth source component from a mixture signal, namely, the variance matrix $\boldsymbol{\Sigma}_n$ is estimated by DNN_n, and these DNN source models are trained in advance. Fig. 1 shows the principle of the separation mechanism in the proposed IDLMA.

On the basis of (3), the cost function (negative log-likelihood of $\boldsymbol{x}_{ij} = \boldsymbol{W}_i^{-1} \boldsymbol{y}_{ij}$) in IDLMA with the complex Gaussian distribution (Gauss-IDLMA) is obtained as

$$\mathcal{L}_{\text{Gauss}} = \sum_{i,j,n} \left[\frac{|y_{ijn}|^2}{\sigma_{ijn}^2} + 2\log \sigma_{ijn}\right] - 2J \sum_i \log |\det \boldsymbol{W}_i|, \tag{7}$$

and (7) can be generalized with (5) (t-IDLMA) as

$$\mathcal{L}_t = \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2}\right)\log\left(1 + \frac{2}{\nu}\frac{|y_{ijn}|^2}{\sigma_{ijn}^2}\right) + 2\log \sigma_{ijn}\right]$$
$$- 2J \sum_i \log |\det \boldsymbol{W}_i|, \tag{8}$$

Fig. 1. Principle of source separation based on IDLMA in case of $N = M = 2$.

where $y_{ijn} = \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}$. Note that $\mathcal{L}_t$ converges to $\mathcal{L}_{\text{Gauss}}$ when $\nu \to \infty$.

C. Update Rule of Source Spectrogram Model Based on DNN

DNN_n is trained so that the source spectrogram $|\tilde{\boldsymbol{S}}_n|^{\cdot 1}$ is predicted from an input mixture spectrogram $|\tilde{\boldsymbol{X}}|^{\cdot 1}$, where $\tilde{\boldsymbol{S}}_n \in \mathbb{C}^{I \times J}$ and $\tilde{\boldsymbol{X}} \in \mathbb{C}^{I \times J}$ are source and mixture spectrograms in the training data, respectively. When we define the output spectrogram as $\boldsymbol{D}_n = \text{DNN}_n(|\tilde{\boldsymbol{X}}|^{\cdot 1}) \in \mathbb{R}_{\geq 0}^{I \times J}$, the loss function of DNN_n for Gauss-IDLMA is defined as

$$\text{L}_{\text{Gauss}}(\boldsymbol{D}_n) = \sum_{i,j} \left(\frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1} - \log \frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1} - 1 \right), \tag{9}$$

where $\tilde{s}_{ijn}$ and d_{ijn} are the elements of $\tilde{\boldsymbol{S}}_n$ and $\boldsymbol{D}_n$, respectively, and δ_1 is a small value to avoid division by zero [13]. Also, the loss function of DNN_n for t-IDLMA is defined as

$$\text{L}_t(\boldsymbol{D}_n) = \sum_{i,j} \left[\left(1 + \frac{\nu}{2} \right) \log \left(1 + \frac{2}{\nu} \frac{|\tilde{s}_{ijn}|^2 + \delta_1}{d_{ijn}^2 + \delta_1} \right) + \log(d_{ijn}^2 + \delta_1) \right]. \tag{10}$$

Since minimizing (9) or (10) is equivalent to the ML estimation of σ_{ijn} in (7) or (8), DNN_n can be interpreted as the proper source generative model based on (3) or (5), respectively. Similarly to (8), $\text{L}_t(\boldsymbol{D}_n)$ converges to $\text{L}_{\text{Gauss}}(\boldsymbol{D}_n)$ up to a constant when $\nu \to \infty$.

The variance matrix is updated by the trained DNN_n as

$$|\boldsymbol{\Sigma}_n|^{\cdot 1} \leftarrow \text{DNN}_n(|\boldsymbol{Y}_n|^{\cdot 1}), \tag{11}$$

$$\sigma_{ijn} \leftarrow \max(\sigma_{ijn}, \varepsilon), \tag{12}$$

where ε is a small value to increase the numerical stability of the spatial update described in Sect. III-D. The DNN architectures used in this paper are described in detail in Sect. IV-B.

D. Update Rule of Demixing Matrix

The demixing matrix $\boldsymbol{W}_i$ can be optimized while taking the statistical independence between sources and the variance matrix $\boldsymbol{\Sigma}_n$ into account on the basis of (3) or (5). In Gauss-IDLMA, $\boldsymbol{W}_i$ can be updated by applying iterative projection (IP) [16] to (7), where IP is a fast and stable optimization algorithm that can be applied to the sum of $|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2$ and $-\log |\det \boldsymbol{W}_i|$. In t-IDLMA, IP cannot be applied to (8) because $|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2$ is intrinsic in the logarithm function.

Therefore, we apply an MM algorithm [14] to derive the update rule of $\boldsymbol{w}_{in}$.

To design a majorization function for (8), we apply the tangent line inequality

$$\log z \leq \frac{1}{\alpha}(z - \alpha) + \log \alpha \tag{13}$$

to the logarithm term in (8), where $z > 0$ is the original variable and $\alpha > 0$ is an auxiliary variable. The majorization function can be designed as

$$\mathcal{L}_t \leq \sum_{i,j,n} \left[\left(1 + \frac{\nu}{2} \right) \frac{1}{\alpha_{ijn}} \left(1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{\sigma_{ijn}^2} - \alpha_{ijn} \right) \right.$$
$$\left. + \left(1 + \frac{\nu}{2} \right) \log \alpha_{ijn} + 2 \log \sigma_{ijn} \right]$$
$$- 2J \sum_i \log |\det \boldsymbol{W}_i|$$
$$=: \mathcal{L}_t^+, \tag{14}$$

where α_{ijn} is the auxiliary variable, and $\mathcal{L}_t$ and $\mathcal{L}_t^+$ become equal only when

$$\alpha_{ijn} = 1 + \frac{2}{\nu} \frac{|y_{ijn}|^2}{\sigma_{ijn}^2}. \tag{15}$$

We can apply IP in analogy with the derivation in Gauss-ILRMA. The majorization function (14) is reformulated as

$$\mathcal{L}_t^+ = J \sum_{i,n} \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{U}_{in} \boldsymbol{w}_{in} - 2J \sum_i \log |\det \boldsymbol{W}_i| + \text{const.}, \tag{16}$$

$$\boldsymbol{U}_{in} = \frac{1}{J} \left(1 + \frac{2}{\nu} \right) \sum_j \frac{1}{\alpha_{ijn} \sigma_{ijn}^2} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}. \tag{17}$$

By applying IP and substituting (15), the demixing filter $\boldsymbol{w}_{in}$ can be updated as follows:

$$\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{in})^{-1} \boldsymbol{e}_n, \tag{18}$$

$$\boldsymbol{w}_{in} \leftarrow \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{U}_{in} \boldsymbol{w}_{in}}}, \tag{19}$$

where

$$\boldsymbol{U}_{in} = \frac{1}{J} \sum_j \frac{1}{c_{ijn}} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}, \tag{20}$$

$$c_{ijn} = \frac{\nu}{\nu + 2} \sigma_{ijn}^2 + \frac{2}{\nu + 2} |y_{ijn}|^2, \tag{21}$$

and $\boldsymbol{e}_n$ is an N-dimensional vector whose nth element is one and whose other elements are zero. After calculating (18) and (19), we update the separated signal by $y_{ijn} \leftarrow \boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}$. In particular, when $\nu \to \infty$, the majorization function (14)

converges to the original cost function (7), and (20) converges to

$$\boldsymbol{U}_{in} = \frac{1}{J} \sum_j \frac{1}{\sigma_{ijn}{}^2} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}}. \qquad (22)$$

The update rule (18)–(21) is equal to that in t-ILRMA.

To fix the scales of y_{ijn} among the frequency bins, the following back-projection technique is applied before updating $\boldsymbol{\Sigma}_n$ by (11) and (12):

$$y_{ijn} \leftarrow [\boldsymbol{W}_i^{-1}(\boldsymbol{e}_n \circ \boldsymbol{y}_{ij})]_{m_{\text{ref}}}, \qquad (23)$$

where y_{ijn} is an element of $\boldsymbol{Y}_n$, $\circ$ is the Hadamard product, $[\cdot]_n$ is the nth value of the vector, and m_{ref} is the index of the reference channel.

E. Relation between Parameter ν and Numerical Stability

In Gauss-IDLMA, $\boldsymbol{U}_{in}$ defined by (22) can be interpreted as the spatial covariance matrix $\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathsf{H}}$ weighted by $\sigma_{ijn}{}^{-2}$. 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 $\sigma_{ijn}{}^{-2}$ may be an excessively large value, reducing the numerical stability of Gauss-IDLMA in IP. In t-IDLMA, on the other hand, c_{ijn} in (20) is the point internally dividing $\sigma_{ijn}{}^2$ and $|y_{ijn}|^2$ with a ratio of $\nu : 2$. Since y_{ijn} is the output of a linear filter, $|y_{ijn}|^2$ contains fewer chasms than $\sigma_{ijn}{}^2$; 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 strong inference of DNN is discounted. Thus, there is a tradeoff when setting ν. The appropriate selection of ν will be discussed in the next section.

IV. EXPERIMENTAL EVALUATION

A. Task, Dataset, and Conditions

We confirmed the validity of the proposed method by conducting a music source separation task. We compared four methods: ILRMA (blind, $K = 20$), DNN+WF, Duong+DNN, and proposed IDLMA, where DNN+WF applies a Wiener filter constructed using all the outputs of the DNN source models to the observed monaural signal [17]. Note that MNMF was not included in this experiment because its performance is almost always inferior to that of ILRMA [18]. For Duong+DNN and IDLMA, the variance matrix $\boldsymbol{\Sigma}_n$ was updated by DNN_n after every 10 iterations of the spatial optimization.

We used the DSD100 dataset of SiSEC2016 [19] as the dry sources and the training datasets of DNN, where only bass (Ba.), drums (Dr.), and vocals (Vo.) were used in this experiment. The 50 songs in the dev data were used to train DNN_n and the top 25 songs in alphabetical order in the test data were used for performance evaluation. The test songs were trimmed only in the interval of 30 to 60 s. To simulate a reverberant mixture, we produced the two-channel observed signals by convoluting the impulse response E2A ($T_{60} = 300\,\text{ms}$) obtained from the RWCP database [20] with each source, and the mixture of Ba. and Vo. (Ba./Vo.) or Dr. and Vo. (Dr./Vo.) was separated. The recording condition of
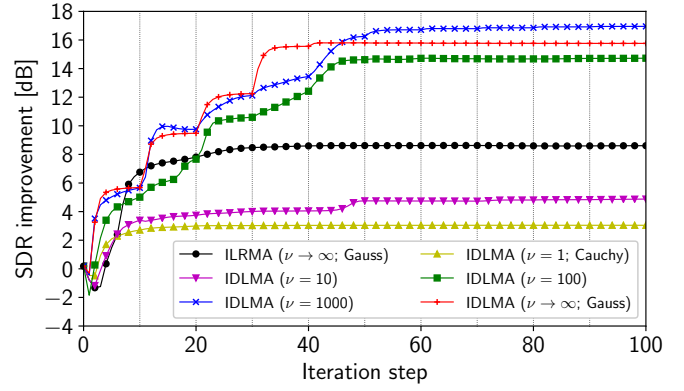

Fig. 2. Example of SDR improvements for each method for Ba./Vo.

E2A is given in [6]. All the signals were downsampled to 8 kHz. STFT was performed using a 512-ms-long Hamming window with a 256-ms-long shift in the Ba./Vo. case and a 256-ms-long Hamming window with a 128-ms-long shift in the Dr./Vo. case. We used the signal-to-distortion ratio (SDR) [21] as the total separation performance.

B. Architecture and Training of DNN Source Model

We constructed a fully connected DNN with four hidden layers. Each layer had 1024 units, and a rectified linear unit was used for the output of each layer. To prepare the training data of mixture signals, we defined the following vectors:

$$\vec{\boldsymbol{s}}_{jn} = (\tilde{\boldsymbol{s}}_{(j-2c)n}^{\top}, \tilde{\boldsymbol{s}}_{(j-2c+2)n}^{\top}, \cdots, \tilde{\boldsymbol{s}}_{(j+2c)n}^{\top})^{\top} \in \mathbb{C}^{I(2c+1)}, \qquad (24)$$

$$\vec{\boldsymbol{x}}_j = \left(\textstyle\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\right)\left(\|\textstyle\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_2\right)^{-1} \in \mathbb{C}^{I(2c+1)}, \qquad (25)$$

$$\bar{\boldsymbol{s}}_{jn} = (\alpha_{jn}\tilde{\boldsymbol{s}}_{jn})\left(\|\textstyle\sum_n \alpha_{jn}\vec{\boldsymbol{s}}_{jn}\|_2 + \delta_2\right)^{-1} \in \mathbb{C}^{I}, \qquad (26)$$

where $\tilde{\boldsymbol{s}}_{jn} \in \mathbb{C}^{I}$ is the STFT of the nth source at j (the column vector of $\tilde{\boldsymbol{S}}_n$), $\vec{\boldsymbol{x}}_j$ and $\bar{\boldsymbol{s}}_{jn}$ are the mixture and source vectors, respectively, α_{jn} is a random variable in the range $[0.05, 1]$, which controls the signal-to-noise ratio in $\vec{\boldsymbol{x}}_j$, and δ_2 is a small value to avoid division by zero. The input and output vectors of DNN_n are $|\vec{\boldsymbol{x}}_j|^{\cdot 1}$ and $|\bar{\boldsymbol{s}}_{jn}|^{\cdot 1}$, respectively.

To optimize DNN, we added the term $(\lambda/2)\sum_q g_q{}^2$ to (9) or (10) for regularization, where g_q is the weight coefficient in DNN, and ADADELTA [22] with a 128-size mini-batch was performed for 200 epochs. The parameter ε was experimentally optimized and set to $0.1 \times (IJ)^{-1} \sum_{i,j} \hat{r}_{ijn}$. The other parameters were set to $\delta_1 = \delta_2 = 10^{-5}$, $c = 3$, and $\lambda = 10^{-5}$.

C. Comparison of Separation Performance

Fig. 2 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 the appropriate ν (=1000) converges to a higher SDR than Gauss-IDLMA ($\nu = \infty$), as mentioned in Sect. III-E.

Figs. 3 and 4 show the average SDR improvements of 25 test songs for Ba./Vo. and Dr./Vo., respectively. We can confirm
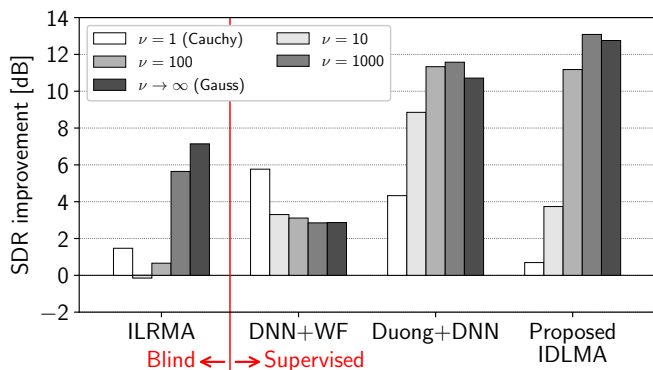
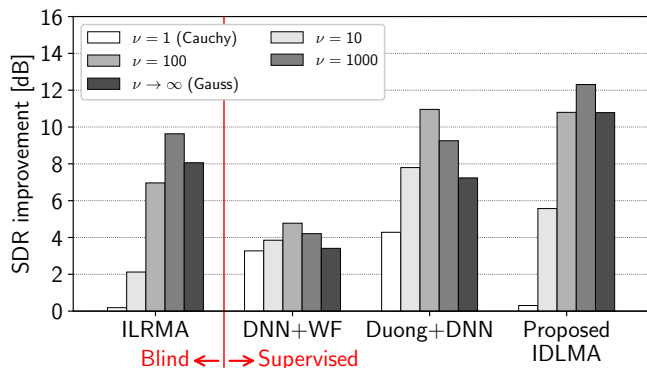Fig. 3. Average SDR improvements of 25 Ba./Vo. songs.

Fig. 4. Average SDR improvements of 25 Dr./Vo. songs.

that the proposed IDLMA outperforms the other methods for both mixtures of instruments. In particular, t-IDLMA with $\nu = 1000$ achieves the highest separation accuracy.

D. Computational Times

To show the efficiency of the proposed approach, we compared the computational times of ILRMA, Duong+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 Duong+DNN, and 26.6 s for IDLMA. These results confirm that the proposed method is as fast as conventional ILRMA and more than 10 times faster than Duong+DNN.

V. CONCLUSION

In this paper, we proposed a new determined source separation method that unifies ICA-based blind spatial optimization and the DNN-based supervised source spectrogram model. The proposed method employs a complex Student's t-distribution as the source generative model. An experimental comparison showed the efficacy of the proposed method in terms of both the separation accuracy and the computational cost.

ACKNOWLEDGMENT

REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[2] 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. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[3] 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. Springer, 2018 (in press), ch. 6, 31 pages.

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

[5] 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.

[6] 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. MLSP*, 2017.

[7] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *Proc. ICASSP*, 2015, pp. 469–473.

[8] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[9] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. ASLP*, vol. 20, no. 4, pp. 1118–1133, 2012.

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

[11] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. ICASSP*, 2014, pp. 3734–3738.

[12] 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. ICASSP*, 2015, pp. 116–120.

[13] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, Sept 2016.

[14] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 60–77, 2000.

[15] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[16] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[17] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, 2015, pp. 2135–2139.

[18] S. Mogami, D. Kitamura, N. Takamune, Y. Mitsui, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Experimental evaluation of independent low-rank matrix analysis based on complex Student's t-distribution," in *Proc. 2017 Autumn Meeting of Acoustical Society of Japan*, 2017, pp. 515–518 (in Japanese).

[19] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 Signal Separation Evaluation Campaign," in *Proc. LVA/ICA*, 2017, pp. 323–332.

[20] 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. LREC*, 2000, pp. 965–968.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.