PoP-IDLMA: Product-of-Prior Independent Deeply Learned Matrix Analysis for Multichannel Music Source Separation

Takuya Hasumi, Tomohiko Nakamura, *Member, IEEE*, Norihiro Takamune, Hiroshi Saruwatari, *Member, IEEE*, Daichi Kitamura, *Senior Member, IEEE*, Yu Takahashi, *Member, IEEE*, and Kazunobu Kondo

***Abstract*—Independent deeply learned matrix analysis (IDLMA) is a state-of-the-art determined audio source separation method based on pretrained deep neural networks (DNNs). Owing to the excellent expression power of DNNs, IDLMA can handle a wider range of sources than conventional source models such as nonnegative matrix factorization (NMF). However, owing to its supervised nature, the separation performance of IDLMA often degrades in the presence of timbral mismatches between the training data and the to-be-separated data. In this paper, we propose two source models that encompass the NMF- and DNN-based source models by constructing a prior distribution of the source power spectrogram (product of priors: PoP) on the basis of the product-of-expert concept. Since the NMF-based source model works well for a fully blind situation, the proposed models can handle the timbral mismatch without losing the expression power of DNNs. By introducing the PoP-based source models into IDLMA, we propose IDLMA extensions (PoP-IDLMAs) and derive their efficient parameter estimation algorithms on the basis of the majorization–minimization algorithm. Experimental results demonstrated the effectiveness of the proposed PoP-IDLMAs and that the proposed models greatly improve the source power estimation in frequency bands above 500 Hz.**

***Index Terms*—Independent deeply learned matrix analysis, independent low-rank matrix analysis, multichannel music source separation, product of experts.**

I. Introduction

BLIND source separation (BSS) is a technique of extracting source signals from their mixture without knowing any information about sources or a mixing process. It plays an important role in multichannel audio source separation and has thus far been well studied [1]. The BSS problem is divided into two situations: undetermined (the number of microphones M is smaller than that of sources N) and (over-)determined ($M \geq N$) situations. In this article, we focus on the determined situation.

For the determined situation, a typical BSS approach is to assume the statistical independence of sources, for example, frequency-domain independent component analysis [2], [3], [4], [5], independent vector analysis (IVA) [6], [7], and independent low-rank matrix analysis (ILRMA) [8]. In this approach, the determined BSS problem is formulated as the problem of finding a demixing filter (the inverse system of the mixing process) simultaneously with the estimation of source power spectrograms. For example, ILRMA uses a source model based on a nonnegative matrix factorization (NMF) [9], [10]. The NMF represents each slice of a source power spectrogram by a sum of common spectral templates weighted by their activations, i.e., it approximates a source power spectrogram using a low-rank nonnegative matrix. This representation is suited for capturing recurring spectral patterns, and ILRMA achieves the state-of-the-art performance in the determined BSS methods.

Alongside with extensions for the fully blind situation [11], [12], [13], ILRMA has been extended for a spatially blind but source-supervised situation, where a mixing system is still unknown but training data of each source are available. This extension is named independent deeply learned matrix analysis (IDLMA) [14]. It is constructed by replacing the NMF-based source model with a source model based on a pretrained deep neural network (DNN) in the ILRMA framework. Owing to the flexible expression power of a DNN, IDLMA works well even for sources that the NMF assumption is not suited for (e.g., a singing voice).

Although a DNN-based source model can handle a wider range of sources, its performance is often degraded by a timbral mismatch between the training data and an observed signal. One cause of this performance degradation is the supervised nature of the DNN-based source model. For example, the higher-frequency components greatly fluctuate owing to musical instrument types and performers' skills, which make the DNN training difficult. Indeed, we experimentally observed such a performance degradation of the DNN-based source model particularly in higher frequency bands, as we will show in Section V-F.

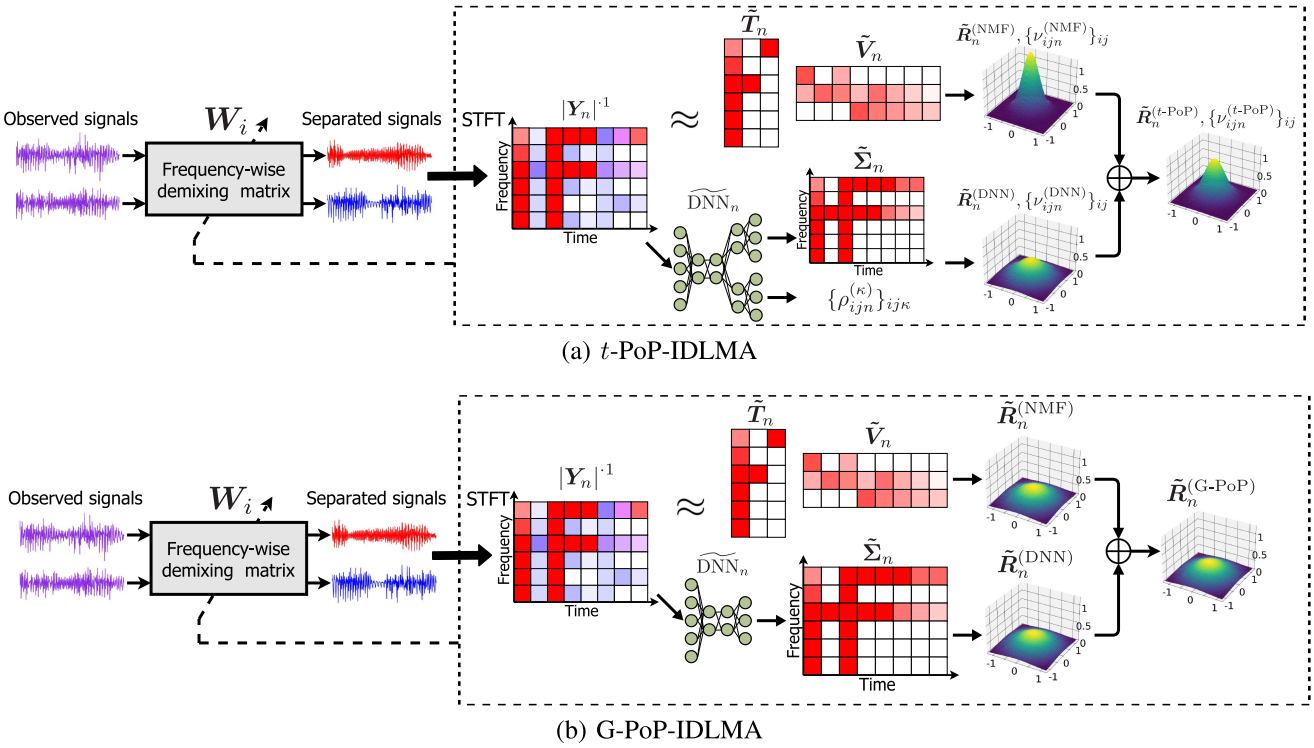
Manuscript received 29 November 2022; revised 20 May 2023; accepted 28 June 2023. Date of publication 6 July 2023; date of current version 19 July 2023. This work was supported in part by JSPS-CAS Joint Research Program under Grant JPJSBP120197203 and in part by JSPS KAKENHI under Grants JP19K20306, JP19H01116, and JP21H05054. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jong Won Shin. *(Corresponding author: Tomohiko Nakamura.)*

Takuya Hasumi, Tomohiko Nakamura, Norihiro Takamune, and Hiroshi Saruwatari are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: takuya_hasumi@ipc.i.u-tokyo.ac.jp; tomohiko.nakamura.jp@ieee.org; norihiro_takamune@ipc.i.u-tokyo.ac.jp; hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp).

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

Yu Takahashi and Kazunobu Kondo are with the Yamaha Corporation, Shizuoka 430-8650, Japan (e-mail: yu.takahashi@music.yamaha.com; kazunobu.kondo@music.yamaha.com).

Digital Object Identifier 10.1109/TASLP.2023.3293044



(a) t-PoP-IDLMA

(b) G-PoP-IDLMA

Fig. 1. Overview of proposed t-PoP-IDLMA and G-PoP-IDLMA. $|\cdot|^{\cdot\tau}$ returns the τth power of the absolute value of each element. See Section III for variables of t-PoP-IDLMA and Section IV for variables of G-PoP-IDLMA.

To alleviate this problem while maintaining the capability of handling various sources, we should extend the DNN-based source model to include an adaptive mechanism against timbral mismatches.

In this article, we propose a source model capable of handling timbral mismatches by unifying the NMF- and DNN-based source models. The idea of developing the proposed model is to combine unsupervised and supervised source models. Unlike the DNN-based source model, the NMF-based source model can work well in an unsupervised manner. We pretrain only the DNN part and use the NMF part in an unsupervised manner. Hence, the NMF part accounts for the time–frequency components that are difficult for the DNN part to represent. The NMF and DNN parts are described with probability distributions of a source power spectrogram. To combine the two distributions in a Bayesian manner, we use a product-of-expert (PoE) technique [15]. PoE represents a probability distribution as a product of multiple probability distributions called experts. By associating the distributions of the NMF and DNN parts with the experts, we can construct a prior distribution of the source power spectrogram as their product. Each of the two distributions can be seen as a prior distribution of the source power spectrogram in the ILRMA/IDLMA framework. Named after this aspect, we call the proposed prior distribution *product of priors (PoP)*.

By replacing the DNN-based source model with the PoP-based source model, we propose an IDLMA extension named *PoP-IDLMA* (see Fig. 1). Furthermore, we propose a variant of PoP-IDLMA by taking the limit of one of the hyperparameters of the PoP-based source model under a certain condition. To distinguish them, we call the former *t-PoP-IDLMA* and the latter *G-PoP-IDLMA*. For both PoP-IDLMAs, we derive efficient parameter estimation algorithms based on the majorization–minimization (MM) algorithm [16]. We conducted experiments on determined source separation and showed the effectiveness of the proposed methods.

While we focus on the IDLMA and ILRMA families throughout this article, the idea of PoP can be extended for underdetermined source separation methods such as multichannel NMF (MNMF) [17], [18] and fast MNMF [19], [20] because they use generative models of a source power spectrogram similarly to the determined source separation methods. We leave such extensions as our future work.

The remainder of this article is organized as follows. In Section II, we briefly describe ILRMA and IDLMA. In Section III, we propose the PoP-based source model and introduce it to IDLMA for constructing t-PoP-IDLMA. We also derive its parameter estimation algorithm on the basis of the MM algorithm. In Section IV, we present G-PoP-IDLMA and derive its parameter estimation algorithm similarly to t-PoP-IDLMA. In Section V, we show the effectiveness of the proposed methods through multichannel music source separation experiments. In Section VI, we conclude this article.

This article is partially based on our previous conference article [21], with the following five contributions. (i) We propose a PoP-based source model by combining the prior distributions of the NMF- and DNN-based source models. Note that the method presented in [21] is used for the DNN part. (ii) We extend the PoP-based source model so that it can avoid the DNN training

cost caused by changing hyperparameter. (iii) We introduce these source models into the IDLMA framework and propose efficient parameter estimation algorithms for t- and G-PoP-IDLMAs. (iv) Through music source separation experiments, we demonstrated the effectiveness of t- and G-PoP-IDLMAs and (v) that the NMF part improves the source power estimation in the frequency band where the DNN part failed to estimate.

II. Conventional Methods

A. Formulation of Determined Audio Source Separation

In this section, we formulate a determined audio source separation problem with M microphones and N sources ($M \geq N$). The short-time Fourier transforms (STFTs) of source, observed, and separated signals are defined as

$$\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijN})^{\mathsf{T}} \in \mathbb{C}^N, \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijM})^{\mathsf{T}} \in \mathbb{C}^M, \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijN})^{\mathsf{T}} \in \mathbb{C}^N, \tag{3}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$ are the indices of frequency bins, time frames, sources, and channels, respectively. The superscript $^{\mathsf{T}}$ denotes the transpose operator.

When the mixing system is time-invariant and an analysis window is sufficiently longer than the reverberation time, $\boldsymbol{x}_{ij}$ is represented as an instantaneous mixture:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{4}$$

where $\boldsymbol{A}_i \in \mathbb{C}^{M \times N}$ is the mixing matrix. If $M = N$ and A_i is nonsingular, we can write $\boldsymbol{y}_{ij}$ as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{5}$$

where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathsf{H}} \in \mathbb{C}^{N \times M}$ is the demixing matrix and the superscript $^{\mathsf{H}}$ is the Hermite transpose operator.

In ILRMA and IDLMA, y_{ijn} is assumed to follow an isotropic complex Gaussian distribution with zero mean and variance $r_{ijn} \in \mathbb{R}_{\geq 0}$:

$$p(y_{ijn}; r_{ijn}) = \frac{1}{\pi r_{ijn}} \exp\left(-\frac{|y_{ijn}|^2}{r_{ijn}}\right). \tag{6}$$

The variance r_{ijn} corresponds to the (i, j)th entry of the power spectrogram of source n, and we call it the source power spectrogram. With this assumption, the source separation problem is formulated as a maximum likelihood estimation problem with respect to r_{ijn} and $\boldsymbol{W}_i$ for a given x_{ijm}. Let $\boldsymbol{X}_m$ and $\boldsymbol{Y}_n$ be $I \times J$ complex matrices consisting of $\{x_{ijm}\}_{i=1,j=1}^{I,J}$ and $\{y_{ijn}\}_{i=1,j=1}^{I,J}$, respectively. By taking the negative of the log-likelihood function, we obtain a cost function as

$$\begin{aligned}\mathcal{L} &= -\log p(\{\boldsymbol{X}_m\}_{m=1}^M) \\ &= -\log p(\{\boldsymbol{Y}_n\}_{n=1}^N) - \sum_i \log|\det \boldsymbol{W}_i|^{2J}\end{aligned}$$

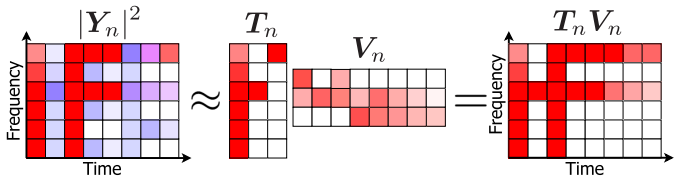


Fig. 2. Source model of ILRMA.

$$\overset{c}{=} \sum_{i,j,n}\left(\log r_{ijn} + \frac{|\boldsymbol{w}_{in}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{r_{ijn}}\right) - 2J\sum_i \log|\det \boldsymbol{W}_i|, \tag{7}$$

where $\overset{c}{=}$ denotes the equality up to constants. The second equation of (7) comes from (5) and the change of variables formula. For brevity, we represent an $I \times J$ nonnegative matrix consisting of $\{r_{ijn}\}_{i=1,j=1}^{I,J}$ as $\boldsymbol{R}_n \in \mathbb{R}_{\geq 0}^{I \times J}$.

ILRMA and IDLMA represent $\boldsymbol{R}_n$ with an NMF and a DNN, respectively. To distinguish them, we hereafter add superscripts $^{(\mathrm{NMF})}$ and $^{(\mathrm{DNN})}$ to $\boldsymbol{R}_n$ for ILRMA and IDLMA, respectively.

B. ILRMA [8]

1) Representation of $\boldsymbol{R}_n^{(NMF)}$: Fig. 2 shows the source model of ILRMA. In this model, the source power spectrogram $\boldsymbol{R}_n^{(\mathrm{NMF})}$ is represented as a product of two nonnegative matrices with rank K:

$$\boldsymbol{R}_n^{(\mathrm{NMF})} = \boldsymbol{T}_n \boldsymbol{V}_n, \tag{8}$$

or equivalently,

$$r_{ijn}^{(\mathrm{NMF})} = \sum_{k=1}^K t_{ikn} v_{kjn}, \tag{9}$$

where $k = 1, \ldots, K$ is the index of the NMF bases. The matrices $\boldsymbol{T}_n \in \mathbb{R}_{\geq 0}^{I \times K}$ and $\boldsymbol{V}_n \in \mathbb{R}_{\geq 0}^{K \times J}$ are the basis and activation matrices consisting of $\{t_{ikn}\}_{i=1,k=1}^{I,K}$ and $\{v_{kjn}\}_{k=1,j=1}^{K,J}$, respectively. The column vectors of $\boldsymbol{T}_n$ represent the spectral patterns of source n and the row vectors of $\boldsymbol{V}_n$ are the energies of the corresponding bases.

2) Parameter Estimation Algorithm: By substituting (9) into (7), we obtain the cost function of ILRMA as

$$\begin{aligned}\mathcal{L}_{\mathrm{ILRMA}} \overset{c}{=} &\sum_{i,j,n}\left(\log\sum_k t_{ikn}v_{kjn} + \frac{|\boldsymbol{w}_{in}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_k t_{ikn}v_{kjn}}\right) \\ &- 2J\sum_i \log|\det \boldsymbol{W}_i|.\end{aligned} \tag{10}$$

The minimization of (10) can be performed by iteratively updating the parameters of the NMF-based source model (t_{ikn} and v_{kjn}) and those of the spatial model ($\boldsymbol{W}_i$) [8].

The parameter estimation algorithm of ILRMA is based on the MM algorithm [16], which offers the guarantee that (10) does not increase at each update. In the MM algorithm, we design an auxiliary function that is tangent to an original cost function. By

Algorithm 1: IP Algorithm.

function $\mathsf{IP}(\{\boldsymbol{X}_m\}_m, \{\boldsymbol{W}_i\}_i, \{\boldsymbol{R}_n\}_n)$
  for $i = 1, \ldots, I$ **do**
    for $n = 1, \ldots, N$ **do**
      $\boldsymbol{U}_{in} = (1/J) \sum_j \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathsf{H}} / r_{ijn}$
      $\boldsymbol{w}_{in} \leftarrow (\boldsymbol{W}_i \boldsymbol{U}_{in})^{-1} \boldsymbol{e}_n$
      $\boldsymbol{w}_{in} \leftarrow \boldsymbol{w}_{in} / \sqrt{\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{U}_{in} \boldsymbol{w}_{in}}$
    end for
  end for
  return$\{\boldsymbol{W}_i\}_i$
end Function

using the auxiliary function, we can derive update rules that do not increase the original cost function:

Theorem 1: Let $f(\boldsymbol{\theta})$ be a cost function and $f^+(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$ be its auxiliary function that satisfies $f(\boldsymbol{\theta}) = \min_{\bar{\boldsymbol{\theta}}} f^+(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}})$. The cost function $f(\boldsymbol{\theta})$ is not increased by iteratively performing the following updates.

$$\bar{\boldsymbol{\theta}} \leftarrow \arg\min_{\bar{\boldsymbol{\theta}}} f^+(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}), \quad \boldsymbol{\theta} \leftarrow \arg\min_{\boldsymbol{\theta}} f^+(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}). \tag{11}$$

By adequately designing an auxiliary function of (10), we can obtain update rules for t_{ikn} and v_{kjn} [8]:

$$t_{ikn} \leftarrow t_{ikn} \sqrt{\frac{\sum_j \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-2} v_{kjn} |y_{ijn}|^2}{\sum_j \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-1} v_{kjn}}}, \tag{12}$$

$$v_{kjn} \leftarrow v_{kjn} \sqrt{\frac{\sum_i \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-2} t_{ikn} |y_{ijn}|^2}{\sum_i \left(\sum_{k'} t_{ik'n} v_{k'jn}\right)^{-1} t_{ikn}}}. \tag{13}$$

Since the terms in the outer parentheses in (12) and (13) are nonnegative, the nonnegativity of t_{ikn} and v_{kjn} always holds once their initial values are nonnegative.

The demixing matrix $\boldsymbol{W}_i$ is updated by the iterative projection (IP) algorithm [22]:

$$\{\boldsymbol{W}_i\}_i \leftarrow \mathsf{IP}(\{\boldsymbol{X}_m\}_m, \{\boldsymbol{W}_i\}_i, \{\boldsymbol{R}_n^{(\mathrm{NMF})}\}_n), \tag{14}$$

where $\mathsf{IP}(\cdot,\cdot,\cdot)$ is defined in Algorithm 1. Here $\boldsymbol{e}_n$ is the N-dimensional unit vector whose nth element is one. This algorithm guarantees that (10) does not increase at each update [23]. It is also used in IDLMA and our proposed methods, as we will show in Sections II-C, III-C, and IV-B.

After the parameter estimation, the projection back (PB) [5] is applied to $\boldsymbol{y}_{ij}$ to resolve the scale uncertainty between $\boldsymbol{w}_{in}$ and $r_{ijn}^{(\mathrm{NMF})}$:

$$\boldsymbol{y}_{ij} \leftarrow \mathrm{diag}(\boldsymbol{d}_i) \boldsymbol{y}_{ij}, \tag{15}$$

where $\mathrm{diag}(\boldsymbol{d}_i) \in \mathbb{C}^{N \times N}$ is a matrix that has elements of $\boldsymbol{d}_i \in \mathbb{C}^N$ on the main diagonal and zero elsewhere. The N-dimensional vector $\boldsymbol{d}_i$ is computed as $\boldsymbol{d}_i = (\boldsymbol{W}_i^{\mathsf{T}})^{-1} \boldsymbol{e}_{m_{\mathrm{ref}}}$, where m_{ref} denotes a reference channel index.

C. IDLMA [14]

1) Representation of $\boldsymbol{R}_n^{(DNN)}$: Fig. 3 shows the source model of IDLMA. In this model, $\boldsymbol{R}_n^{(\mathrm{DNN})}$ is obtained with the pretrained

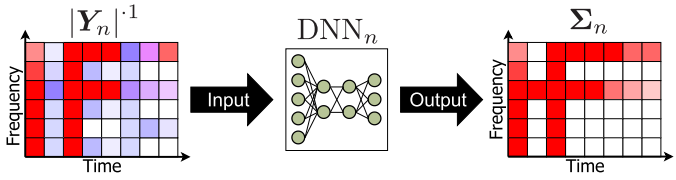


Fig. 3. Source model of IDLMA.

DNN DNN_n. Let $|\cdot|^{\cdot\tau}$ denote the elementwise τth power of absolute values of a matrix. It converts $|\boldsymbol{Y}_n|^{\cdot 1}$ into the source magnitude spectrogram $\boldsymbol{\Sigma}_n \in \mathbb{R}_{\geq 0}^{I \times J}$:

$$\boldsymbol{\Sigma}_n = \mathrm{DNN}_n(|\boldsymbol{Y}_n|^{\cdot 1}). \tag{16}$$

Let σ_{ijn} be the (i,j)th entry of $\boldsymbol{\Sigma}_n$. We obtain r_{ijn} as

$$r_{ijn}^{(\mathrm{DNN})} = \max(\sigma_{ijn}^2, \varepsilon_1), \tag{17}$$

where $\max(\cdot,\cdot)$ returns a maximum value of two inputs and ε_1 is a small value used to prevent numerical instability.

2) Parameter Estimation Algorithm: The parameter estimation algorithm of IDLMA consists of two stages: separation and DNN training stages. The separation stage is performed after the DNN training stage. We describe the separation stage in this section and the DNN training stage in Section II-C3.

In the separation stage, the parameters of the source and spatial models are estimated from observed signals $\boldsymbol{X}_m$. The cost function of IDLMA is defined by replacing r_{ijn} with $r_{ijn}^{(\mathrm{DNN})}$ in (7):

$$\mathcal{L}_{\mathrm{IDLMA}} \overset{c}{=} \sum_{i,j,n} \left(\log r_{ijn}^{(\mathrm{DNN})} + \frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{r_{ijn}^{(\mathrm{DNN})}} \right) - 2J \sum_i \log |\det \boldsymbol{W}_i|. \tag{18}$$

As in ILRMA, the parameter estimation algorithm of IDLMA consists of iterative updates of the source model and demixing matrices.

The source power spectrogram $\boldsymbol{R}_n^{(\mathrm{DNN})}$ is updated in accordance with (16) and (17), where $\boldsymbol{Y}_n$ is obtained with the current estimates of $\boldsymbol{W}_i$. For the update of $\boldsymbol{W}_i$, we can use the IP algorithm because the terms of (18) involved in $\boldsymbol{W}_i$ have the same form as those of ILRMA:

$$\{\boldsymbol{W}_i\}_i \leftarrow \mathsf{IP}(\{\boldsymbol{X}_m\}_m, \{\boldsymbol{W}_i\}_i, \{\boldsymbol{R}_n^{(\mathrm{DNN})}\}_n). \tag{19}$$

The PB technique is applied to $\boldsymbol{y}_{ij}$ after every update of $\boldsymbol{W}_i$, which can reduce linear distortion.

3) DNN Training: In the DNN training stage, we train DNN_n so that it can estimate a clean magnitude spectrogram from a noisy magnitude spectrogram. The point of IDLMA is that a cost function for the DNN training is consistent with the cost function (18) used in the separation stage in a maximum likelihood sense.

Let $\breve{s}_{ijn}$ be the (i,j)th element of a clean complex spectrogram of source n. The cost function for the DNN training is derived by respectively replacing $\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}$ and $r_{ijn}^{(\mathrm{DNN})}$ with $\breve{s}_{ijn}$

and σ_{ijn}^2 in $\mathcal{L}_{\text{IDLMA}}$:

$$\mathcal{C}_{\text{IDLMA}}^{(n)} = \sum_{i,j} \left(\frac{|\breve{s}_{ijn}|^2 + \varepsilon_2}{\sigma_{ijn}^2 + \varepsilon_2} - \log \frac{|\breve{s}_{ijn}|^2 + \varepsilon_2}{\sigma_{ijn}^2 + \varepsilon_2} - 1 \right), \quad (20)$$

where ε_2 is a small value used to prevent numerical instability. The right-hand side of (20) is the Itakura–Saito divergence between $|\breve{s}_{ijn}|^2 + \varepsilon_2$ and $\sigma_{ijn}^2 + \varepsilon_2$. When ε_2 is negligibly small, the minimization of (20) with respect to σ_{ijn}^2 is equivalent to the maximum likelihood estimation of (18) with respect to σ_{ijn}^2. Hence, the DNN training with $\mathcal{C}_{\text{IDLMA}}^{(n)}$ corresponds to the emulation of the maximum likelihood estimation with respect to σ_{ijn} in the separation stage.

III. Proposed t-PoP-IDLMA

A. t-PoP-Based Source Model

1) PoP: In this section, we propose the PoP-based source model by unifying the NMF- and DNN-based source models on the basis of PoE [15]. PoE designs a probability distribution of a random variable by multiplying multiple probability distributions of the variable. The multiplication is analogous to an "and" operation of multiple conditions, and the designed distribution has high values at events where such conditions tend to be satisfied simultaneously. In the proposed model, we treat the source power spectrogram r_{ijn} as a latent variable and construct its prior distribution by multiplying the NMF- and DNN-based probability distributions of r_{ijn}. By using the prior distribution (i.e., PoP), we define the source model through the marginalization of r_{ijn}.

As in ILRMA and IDLMA, y_{ijn} is assumed to obey an isotropic Gaussian distribution with zero mean and variance r_{ijn}. To clarify that r_{ijn} is a latent variable, we rewrite (6) as

$$p(y_{ijn}|r_{ijn}) = \frac{1}{\pi r_{ijn}} \exp\left(-\frac{|y_{ijn}|^2}{r_{ijn}} \right). \quad (21)$$

Following PoE, we can define a prior distribution of r_{ijn} with a set of hyperparameters $\theta_{ijn}^{(t\text{-PoP})}$ as

$$p\left(r_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) \propto q\left(r_{ijn}; \theta_{ijn}^{(\text{NMF})}\right) q\left(r_{ijn}; \theta_{ijn}^{(\text{DNN})}\right), \quad (22)$$

where $q(r_{ijn}; \theta_{ijn}^{(\text{NMF})})$ and $q(r_{ijn}; \theta_{ijn}^{(\text{DNN})})$ are the NMF- and DNN-based probability distributions with sets of parameters $\theta_{ijn}^{(\text{NMF})}$ and $\theta_{ijn}^{(\text{DNN})}$, respectively.

For the right-hand side of (22) to be a probability distribution, a normalization constant should exist. Unfortunately, it is not always described in an explicit form. However, by adequately choosing probability distributions for $q(r_{ijn}; \theta_{ijn}^{(\text{NMF})})$ and $q(r_{ijn}; \theta_{ijn}^{(\text{DNN})})$, we can write the right-hand side of (22) in a closed form, which helps the derivation of a parameter estimation algorithm.

Let us choose an inverse gamma distribution for $q(r_{ijn}; \theta_{ijn}^{(l)})$ for part label $l \in \{\text{NMF}, \text{DNN}\}$:

$$q\left(r_{ijn}; \theta_{ijn}^{(l)}\right) = \mathcal{IG}\left(r_{ijn}; \alpha_{ijn}^{(l)}, \beta_{ijn}^{(l)}\right), \quad (23)$$

$$\mathcal{IG}\left(r_{ijn}; \alpha_{ijn}^{(l)}, \beta_{ijn}^{(l)}\right) := \frac{(\beta_{ijn}^{(l)})^{\alpha_{ijn}^{(l)}}}{\Gamma(\alpha_{ijn}^{(l)})} r_{ijn}^{-\alpha_{ijn}^{(l)}-1} e^{-\beta_{ijn}^{(l)}/r_{ijn}}, \quad (24)$$

where $\theta_{ijn}^{(l)} = \{\alpha_{ijn}^{(l)}, \beta_{ijn}^{(l)}\}$, $\alpha_{ijn}^{(l)} > 0$ is the shape parameter, $\beta_{ijn}^{(l)} > 0$ is the scale parameter, and $\Gamma(\cdot)$ is the gamma function. Since a product of two inverse gamma distributions is also an inverse gamma distribution, we can explicitly write the proposed PoP as

$$p\left(r_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) = \mathcal{IG}\left(r_{ijn}; \alpha_{ijn}^{(t\text{-PoP})}, \beta_{ijn}^{(t\text{-PoP})}\right), \quad (25)$$

where $\theta_{ijn}^{(t\text{-PoP})} = \theta_{ijn}^{(\text{NMF})} \cup \theta_{ijn}^{(\text{DNN})}$ and

$$\alpha_{ijn}^{(t\text{-PoP})} = \alpha_{ijn}^{(\text{NMF})} + \alpha_{ijn}^{(\text{DNN})} + 1, \quad (26)$$

$$\beta_{ijn}^{(t\text{-PoP})} = \beta_{ijn}^{(\text{NMF})} + \beta_{ijn}^{(\text{DNN})}. \quad (27)$$

It should be noted that we can combine more than two probability distributions in the same manner.

2) Source Model: The proposed source model is defined as a marginalization distribution $p(y_{ijn}; \theta_{ijn}^{(t\text{-PoP})})$:

$$p\left(y_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) = \int_0^\infty p(y_{ijn}|r_{ijn}) p\left(r_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) \mathrm{d}r_{ijn}. \quad (28)$$

An inverse gamma distribution is a conjugate prior distribution of a normal distribution (see [24] for example). Thus, we can compute the marginal distribution in a closed form:

$$p\left(y_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) = \frac{1}{\pi \tilde{r}_{ijn}^{(t\text{-PoP})}} \left(1 + \frac{2|y_{ijn}|^2}{\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})}} \right)^{-1-\nu_{ijn}^{(t\text{-PoP})}/2}, \quad (29)$$

where

$$\tilde{r}_{ijn}^{(t\text{-PoP})} = \frac{\beta_{ijn}^{(t\text{-PoP})}}{\alpha_{ijn}^{(t\text{-PoP})}}, \quad \nu_{ijn}^{(t\text{-PoP})} = 2\alpha_{ijn}^{(t\text{-PoP})}. \quad (30)$$

The resulting distribution is identical to a complex isotropic Student's-t distribution with the degree-of-freedom (DoF) parameter $\nu_{ijn}^{(t\text{-PoP})}$ and scale parameter $\tilde{r}_{ijn}^{(t\text{-PoP})}$. Decreasing $\nu_{ijn}^{(t\text{-PoP})}$ leads to a more heavy-tailed probability distribution, i.e., it controls the Gaussianity of the distribution. We call the source model (29) *the t-PoP-based source model*.

B. Interpretation of t-PoP-Based Source Model

We have thus far derived the proposed t-PoP-based source model. In this section, we provide an interpretation of the t-PoP-based source model to bridge it with the source models of ILRMA and IDLMA. On the basis of this interpretation, we parameterize $\theta_{ijn}^{(\text{NMF})}$ and $\theta_{ijn}^{(\text{DNN})}$ with an NMF and a DNN.

Similarly to (30), we define

$$\tilde{r}_{ijn}^{(l)} = \frac{\beta_{ijn}^{(l)}}{\alpha_{ijn}^{(l)}}, \quad \nu_{ijn}^{(l)} = 2\alpha_{ijn}^{(l)}, \quad (31)$$

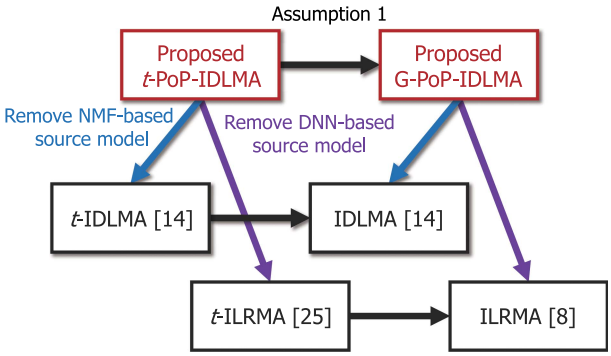


Fig. 4. Relationship between proposed and conventional source models.

for $l \in \{\mathrm{NMF}, \mathrm{DNN}\}$. Since $\alpha_{ijn}^{(l)}$ and $\beta_{ijn}^{(l)}$ can be represented by $\tilde{r}_{ijn}^{(l)}$ and $\nu_{ijn}^{(l)}$, respectively, we hereafter redefine $\theta_{ijn}^{(l)}$ by $\theta_{ijn}^{(l)} := \{\tilde{r}_{ijn}^{(l)}, \nu_{ijn}^{(l)}\}$ for $l \in \{\mathrm{NMF}, \mathrm{DNN}\}$.

Let us consider the case where we choose a uniform distribution for $q(r_{ijn}; \theta_{ijn}^{(\mathrm{DNN})})$ in (22). In this case, since $p(r_{ijn}; \theta_{ijn}^{(t\text{-PoP})})$ equals $q(r_{ijn}; \theta_{ijn}^{(\mathrm{NMF})})$, the resultant source model is given as a complex isotropic Student's-t distribution with the DoF parameter $\nu_{ijn}^{(\mathrm{NMF})}$ and scale parameter $\tilde{r}_{ijn}^{(\mathrm{NMF})}$. It coincides with the source model of a Student's-t-distribution-based extension of ILRMA (t-ILRMA) [25]. Furthermore, by invoking the fact that a complex isotropic Student's-t distribution becomes a complex isotropic Gaussian distribution as the DoF parameter goes to infinity, we obtain a complex isotropic Gaussian distribution with zero mean and variance $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ as $\nu_{ijn}^{(\mathrm{NMF})} \to \infty$. It coincides with the NMF-based source model of ILRMA. Similarly to (9), we can parameterize $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ as

$$\boldsymbol{R}_n^{(\mathrm{NMF})} = \tilde{\boldsymbol{T}}_n \tilde{\boldsymbol{V}}_n, \tag{32}$$

or equivalently

$$\tilde{r}_{ijn}^{(\mathrm{NMF})} = \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn}, \tag{33}$$

where $\tilde{t}_{ijn}$ and $\tilde{v}_{ikn}$ are the (i,j)th entries of the basis and activation matrices $\tilde{\boldsymbol{T}}_n \in \mathbb{R}_{\geq 0}^{I \times J}$ and $\tilde{\boldsymbol{V}}_n \in \mathbb{R}_{\geq 0}^{I \times J}$, respectively. We can provide a similar interpretation for $q(r_{ijn}; \theta_{ijn}^{(\mathrm{DNN})})$. Fig. 4 shows the relationship between the proposed and conventional source models, where t-IDLMA [14] is a Student's-t-distribution-based extension of IDLMA. As in the IDLMA family, $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ is estimated from $|\boldsymbol{Y}_n|^{\cdot 1}$ by using a pretrained DNN $\mathrm{D\tilde{N}N}_n$.

The above interpretations reveal the relationship between the proposed source model and the source models of ILRMA and IDLMA. With the notations (31), we can rewrite the parameters of the t-PoP-based source model $\tilde{r}_{ijn}^{(t\text{-PoP})}$ and $\nu_{ijn}^{(t\text{-PoP})}$ as

$$\tilde{r}_{ijn}^{(t\text{-PoP})} = \frac{\nu_{ijn}^{(t\text{-PoP})} - 2}{\nu_{ijn}^{(t\text{-PoP})}} \left[\eta_{ijn} \tilde{r}_{ijn}^{(\mathrm{NMF})} + (1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\mathrm{DNN})} \right], \tag{34}$$

$$\nu_{ijn}^{(t\text{-PoP})} = \nu_{ijn}^{(\mathrm{NMF})} + \nu_{ijn}^{(\mathrm{DNN})} + 2, \tag{35}$$

where

$$\eta_{ijn} = \frac{\nu_{ijn}^{(\mathrm{NMF})}}{\nu_{ijn}^{(\mathrm{NMF})} + \nu_{ijn}^{(\mathrm{DNN})}}. \tag{36}$$

Since η_{ijn} is in the open set (0,1), the term in parentheses in (34) can be seen as an η_{ijn}-weighted sum of the NMF- and DNN-based source power spectrograms. The DoF parameters $\nu_{ijn}^{(\mathrm{NMF})}$ and $\nu_{ijn}^{(\mathrm{DNN})}$ determine the weighting factor η_{ijn}.

C. Parameter Estimation Algorithm

1) Cost Function: By replacing the DNN-based source model with the t-PoP-based source model, we propose t-PoP-IDLMA. As in ILRMA and IDLMA, the source separation problem can be formulated as a maximum likelihood estimation problem with respect to $\theta_{ijn}^{(t\text{-PoP})}$ and $\boldsymbol{W}_i$. The negative log-likelihood of $\{\boldsymbol{X}_m\}_m$, i.e., the cost function, is given as

$$\mathcal{L}_{t\text{-PoP}} = \sum_{i,j,n} \log p\left(y_{ijn}; \theta_{ijn}^{(t\text{-PoP})}\right) - \sum_i \log |\det \boldsymbol{W}_i|^{2J} \tag{37}$$

$$\overset{c}{=} \sum_{i,j,n} \left(1 + \frac{\nu_{ijn}^{(t\text{-PoP})}}{2}\right) \log \left(1 + \frac{2|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})}}\right) + \sum_{i,j,n} \log \tilde{r}_{ijn}^{(t\text{-PoP})} - 2J \sum_i \log |\det \boldsymbol{W}_i|. \tag{38}$$

The parameter estimation algorithm of t-PoP-IDLMA consists of two stages as in IDLMA. In the DNN training stage, the DNN is trained with the training data of each source, which we will describe in Section III-D. In the separation stage, we estimate $\tilde{t}_{ikn}, \tilde{v}_{kjn}, \tilde{r}_{ijn}^{(\mathrm{DNN})}$, and $\nu_{ijn}^{(\mathrm{DNN})}$ from $\{\boldsymbol{X}_m\}_m$. Note that $\nu_{ijn}^{(\mathrm{NMF})}$ is treated as a hyperparameter. In the following, we derive update rules of $\tilde{t}_{ikn}, \tilde{v}_{kjn}, \tilde{r}_{ijn}^{(\mathrm{DNN})}$, and $\nu_{ijn}^{(\mathrm{DNN})}$ on the basis of the MM algorithm.

2) Update Rule of $\boldsymbol{W}_i$: The cost function $\mathcal{L}_{t\text{-PoP}}$ includes $|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2$ in the logarithm term, which makes it difficult to minimize $\mathcal{L}_{t\text{-PoP}}$. To construct an upper bound of this term, we can use the following lemma [25]:

Lemma 1: For a concave function $f(\theta)$, its tangent line at point θ_o is greater than or equal to $f(\theta)$:

$$f(\theta) \leq f'(\theta_\mathrm{o})(\theta - \theta_\mathrm{o}) + f(\theta_\mathrm{o}). \tag{39}$$

The equality holds if and only if $\theta = \theta_\mathrm{o}$.

Since the logarithmic function is concave, we obtain

$$\log \left(1 + \frac{2|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})}}\right) \leq \frac{1}{\zeta_{ijn}} \left[1 + \frac{2|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})}} - \zeta_{ijn}\right] + \log \zeta_{ijn}, \tag{40}$$

where $\zeta_{ijn} > 0$ is the auxiliary variable. The equality of (40) holds if and only if

$$\zeta_{ijn} = 1 + \frac{2|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})}}. \tag{41}$$

Hence, the auxiliary function of $\mathcal{L}_{t\text{-PoP}}$ is given as

$$\begin{aligned}\mathcal{L}_{t\text{-PoP}}^{+} \overset{c}{=} & \sum_{i,j,n} \frac{|\boldsymbol{w}_{in}^{\mathsf{H}} \boldsymbol{x}_{ij}|^2}{\zeta_{ijn} \nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})} / (2 + \nu_{ijn}^{(t\text{-PoP})})} \\ & + \sum_{i,j,n} \log \tilde{r}_{ijn}^{(t\text{-PoP})} - 2J \sum_{i} \log |\det \boldsymbol{W}_i| \\ & + \sum_{i,j,n} \left(1 + \frac{\nu_{ijn}^{(t\text{-PoP})}}{2}\right) \left(\frac{1}{\zeta_{ijn}} - 1 + \log \zeta_{ijn}\right).\end{aligned} \tag{42}$$

The $\boldsymbol{w}_{in}$-related terms in (42) are only quadratic and log-determinant terms, which fits the requirements for using the IP algorithm [23]. Hence, the update rule of $\boldsymbol{W}_i$ is given as

$$\{\boldsymbol{W}_i\}_i \leftarrow \mathsf{IP}(\{\boldsymbol{X}_m\}_m, \{\boldsymbol{W}_i\}_i, \{\boldsymbol{\Xi}_n\}_n), \tag{43}$$

where $\boldsymbol{\Xi}_n$ is an $I \times J$ matrix consisting of ξ_{ijn} given by

$$\xi_{ijn} = \frac{\nu_{ijn}^{(t\text{-PoP})}}{\nu_{ijn}^{(t\text{-PoP})} + 2} \tilde{r}_{ijn}^{(t\text{-PoP})} + \frac{2}{\nu_{ijn}^{(t\text{-PoP})} + 2} |y_{ijn}|^2. \tag{44}$$

Note that ξ_{ijn} is the denominator of the first term of (42) in which the equality condition (41) is substituted.

3) Update Rules of $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$: By invoking (33) and (34), we find that the first and second terms of (42) include the sums over k in the reciprocal and logarithmic functions, respectively. These terms make it difficult to analytically solve the minimization of (42) with respect to $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$. To overcome this problem, we derive update rules of t_{ikn} and v_{kjn} on the basis of the MM algorithm.

For a reciprocal function, we can use the following lemma:

Lemma 2: For a series of nonnegative values $\{h_k\}_k$,

$$\frac{1}{\sum_k h_k} \leq \sum_k \frac{\lambda_k^2}{h_k}, \tag{45}$$

where $\lambda_k \geq 0$ is the auxiliary variable such that $\sum_k \lambda_k = 1$.

This lemma can be proved by Jensen's inequality [24]. Using Lemma 2, we can obtain the following inequality:

$$\begin{aligned}&\frac{1}{\eta_{ijn} \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn} + (1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}} \\ &\leq \left(\sum_k \frac{(\lambda_{ijkn}^{(\text{NMF})})^2}{\eta_{ijn} \tilde{t}_{ikn} \tilde{v}_{kjn}} + \frac{(\lambda_{ijn}^{(\text{DNN})})^2}{(1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}}\right),\end{aligned} \tag{46}$$

where $\lambda_{ijkn}^{(\text{NMF})} > 0$ and $\lambda_{ijn}^{(\text{DNN})} > 0$ are the auxiliary variables that satisfy $\sum_k \lambda_{ijkn}^{(\text{NMF})} + \lambda_{ijn}^{(\text{DNN})} = 1$. The equality of (46) holds if and only if

$$\lambda_{ijkn}^{(\text{NMF})} = \frac{\eta_{ijn} t_{ikn} v_{kjn}}{\tilde{r}_{ijn}^{(t\text{-PoP})}}, \quad \lambda_{ijn}^{(\text{DNN})} = \frac{(1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}}{\tilde{r}_{ijn}^{(t\text{-PoP})}}. \tag{47}$$

Using Lemma 1, we can derive the following inequality for the second term of (42):

$$\begin{aligned}&\log \left(\eta_{ijn} \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn} + (1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}\right) \\ &\leq \frac{\eta_{ijn} \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn} + (1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}}{\gamma_{ijn}} - 1 + \log \gamma_{ijn}\end{aligned} \tag{48}$$

where $\gamma_{ijn} > 0$ is an auxiliary variable. The equality of (48) holds if and only if

$$\gamma_{ijn} = \eta_{ijn} \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn} + (1 - \eta_{ijn}) \tilde{r}_{ijn}^{(\text{DNN})}. \tag{49}$$

Taken together, the upper bound of (42) is obtained as

$$\begin{aligned}\mathcal{L}_{t\text{-PoP}}^{++} = & \sum_{i,j,n} \frac{\left(\nu_{ijn}^{(t\text{-PoP})} + 2\right) |y_{ijn}|^2}{\left(\nu_{ijn}^{(t\text{-PoP})} - 2\right) \zeta_{ijn}} \sum_k \frac{\left(\lambda_{ijkn}^{(\text{NMF})}\right)^2}{\eta_{ijn} \tilde{t}_{ikn} \tilde{v}_{kjn}} \\ & + \sum_{i,j,n} \frac{\eta_{ijn}}{\gamma_{ijn}} \sum_k \tilde{t}_{ikn} \tilde{v}_{kjn} + \mathcal{D}_{\backslash \tilde{t}, \tilde{v}}^{(t\text{-PoP})},\end{aligned} \tag{50}$$

where $\mathcal{D}_{\backslash \tilde{t}, \tilde{v}}^{(t\text{-PoP})}$ denotes terms that do not include $\tilde{t}_{ikn}$ or $\tilde{v}_{kjn}$. By solving $\partial \mathcal{L}_{t\text{-PoP}}^{++} / \partial \tilde{t}_{ikn} = 0$ and $\partial \mathcal{L}_{t\text{-PoP}}^{++} / \partial \tilde{v}_{ijn} = 0$ and substituting equality conditions (41), (47), and (49) into the solutions, we can derive the following update rules:

$$\tilde{t}_{ikn} \leftarrow \tilde{t}_{ikn} \sqrt{\frac{\sum_j \nu_{ijn}^{(\text{NMF})} \tilde{v}_{kjn} |y_{ijn}|^2 / (\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})} \xi_{ijn})}{\sum_j \nu_{ijn}^{(\text{NMF})} \tilde{v}_{kjn} / (\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})})}}, \tag{51}$$

$$\tilde{v}_{kjn} \leftarrow \tilde{v}_{kjn} \sqrt{\frac{\sum_i \nu_{ijn}^{(\text{NMF})} \tilde{t}_{ikn} |y_{ijn}|^2 / (\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})} \xi_{ijn})}{\sum_i \nu_{ijn}^{(\text{NMF})} \tilde{t}_{ikn} / (\nu_{ijn}^{(t\text{-PoP})} \tilde{r}_{ijn}^{(t\text{-PoP})})}}, \tag{52}$$

where ξ_{ijn} is given as (44).

4) Update Rules of $\tilde{r}_{ijn}^{(DNN)}$ and $\nu_{ijn}^{(DNN)}$: In t-PoP-IDLMA, the DNN of the nth source $\widetilde{\text{DNN}}_n$ estimates the DNN-based source model parameters $\tilde{r}_{ijn}^{(\text{DNN})}$ and $\nu_{ijn}^{(\text{DNN})}$ from $|\boldsymbol{Y}_n|^{\cdot 1}$. Although we can determine $\nu_{ijn}^{(\text{DNN})}$ before training as in t-IDLMA, it requires the retraining of a DNN per $\nu_{ijn}^{(\text{DNN})}$, which leads to a prohibitive computational cost for the hyperparameter search. Hence, we designed $\widetilde{\text{DNN}}_n$ to output both $\tilde{r}_{ijn}^{(\text{DNN})}$ and $\nu_{ijn}^{(\text{DNN})}$.

For $\widetilde{\text{DNN}}_n$, we adopted a DNN proposed in [21], where $\nu_{ijn}^{(\text{DNN})}$ is represented by a weighted sum of anchors:

$$\nu_{ijn}^{(\text{DNN})} = \sum_{\kappa \in \mathcal{K}} \rho_{ijn}^{(\kappa)} \kappa, \tag{53}$$

where $\mathcal{K}$ is a set of anchors and $\rho_{ijn}^{(\kappa)}$ is a weight of anchor κ that satisfies

$$0 \leq \rho_{ijn}^{(\kappa)} \leq 1, \quad \sum_{\kappa \in \mathcal{K}} \rho_{ijn}^{(\kappa)} = 1. \tag{54}$$

Algorithm 2: Parameter Estimation Algorithm of t-PoP-IDLMA.

Input: $\{\boldsymbol{X}_m\}_m$, $\{\widetilde{\mathrm{DNN}}_n\}_n$, $\{\nu_{ijn}^{(\mathrm{NMF})}\}_{ijn}$
Output: $\{\boldsymbol{Y}_n\}_n$

```
for I_(out)^(t-PoP) iterations do
    Update {r̃_ijn^(DNN)}_ijn and {ν_ijn^(DNN)}_ijn by (53), (55), and (56)
    Update {r̃_ijn^(t-PoP)}_ijn and {ν_ijn^(t-PoP)}_ijn by (34) and (35)
    Update {Ξ_n}_n by (44)
    for I_(in)^(t-PoP) iterations do
        Update {t̃_ijn}_ijn and {ṽ_ijn}_ijn by (51) and (52)
        Update {r̃_ijn^(t-PoP)}_ijn by (34)
        Update {Ξ_n}_n by (44)
        {W_i}_i ← IP({X_m}_m, {W_i}_i, {Ξ_n}_n)
        Update {Y_n}_n by (5)
        Update {Y_n}_n by (15)
    end for
end for
```

In this network, the DNN outputs not $\nu_{ijn}^{(\mathrm{DNN})}$ but $\{\rho_{ijn}^{(\kappa)}\}_\kappa$ for each time–frequency slot. Since $\nu_{ijn}^{(\mathrm{DNN})}$ is always between the minimum and maximum values of $\rho_{ijn}^{(\kappa)}$, the sum-of-anchors model can avoid an excessive increase in $\nu_{ijn}^{(\mathrm{DNN})}$, which degrades the separation performance of IDLMA [21]. Hence, we can update $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ and $\nu_{ijn}^{(\mathrm{DNN})}$ by using (53) and the following rules:

$$\{\{\tilde{\sigma}_{ijn}\}_{ij}, \{\rho_{ijn}^{(\kappa)}\}_{ij\kappa}\} \leftarrow \widetilde{\mathrm{DNN}}_n(|\boldsymbol{Y}_n|^{\cdot 1}), \tag{55}$$

$$\tilde{r}_{ijn}^{(\mathrm{DNN})} \leftarrow \max(\tilde{\sigma}_{ijn}^2, \epsilon_1), \tag{56}$$

where $\tilde{\sigma}_{ijn}$ is the (i,j)th entry of the source magnitude spectrogram obtained using $\widetilde{\mathrm{DNN}}_n$.

5) Entire Procedure of Separation Stage: Fig. 1(a) shows the overview of the separation process of t-PoP-IDLMA, where the spatial and source models are iteratively updated. Algorithm 2 shows the entire parameter estimation algorithm of t-PoP-IDLMA in the separation stage, where $\mathcal{I}_{(\mathrm{in})}^{(t\text{-PoP})}$ and $\mathcal{I}_{(\mathrm{out})}^{(t\text{-PoP})}$ denote the numbers of inner and outer iterations, respectively. The inner iteration does not include the update of the DNN part, whereas the outer iteration includes the update.

D. DNN Training

As in IDLMA, the DNN part is trained before the separation stage described in Section III-C. In the DNN training stage, we set $\tilde{r}_{ijn}^{(\mathrm{NMF})} = 0$ because the DNN part is responsible for source components similar to the training data. In the spirit of IDLMA, we design a cost function for the DNN training to be consistent with the cost function (38) of the separation stage:

$$\mathcal{C}_{t\text{-PoP}}^{(n)} = \sum_{i,j} \log\left[\frac{\nu_{ijn}^{(\mathrm{DNN})}}{\nu_{ijn}^{(t\text{-PoP})}}(\tilde{\sigma}_{ijn}^2 + \varepsilon_2)\right] + \sum_{i,j}\left(1 + \frac{\nu_{ijn}^{(t\text{-PoP})}}{2}\right)\log\left[1 + \frac{2(|\check{s}_{ijn}|^2 + \varepsilon_2)}{\nu_{ijn}^{(\mathrm{DNN})}(\tilde{\sigma}_{ijn}^2 + \varepsilon_2)}\right], \tag{57}$$

where $\tilde{\sigma}_{ijn}$ is estimated from a noisy mixture using $\widetilde{\mathrm{DNN}}_n$. The noisy mixture is generated by mixing the clean spectrogram of source n $\check{s}_{ijn}$ and the spectrogram of other sources.

The minimization of $\mathcal{C}_{t\text{-PoP}}^{(n)}$ with respect to $\tilde{\sigma}_{ijn}$ is consistent with the maximum likelihood estimation of $\tilde{\sigma}_{ijn}$ in (38). Thus, the DNN training with $\mathcal{C}_{t\text{-PoP}}^{(n)}$ matches the maximum likelihood estimation as in IDLMA.

IV. Proposed G-PoP-IDLMA

A. G-PoP-Based Source Model

t-PoP-IDLMA successfully combines the NMF- and DNN-based source models. However, the cost function (57) for the DNN training includes $\nu_{ijn}^{(\mathrm{NMF})}$. Thus, we need to train $\widetilde{\mathrm{DNN}}_n$ whenever $\nu_{ijn}^{(\mathrm{NMF})}$ changes. In this section, we propose an extension of t-PoP-IDLMA that can avoid the DNN training caused by the change of $\nu_{ijn}^{(\mathrm{NMF})}$.

On the basis of our interpretation in Section III-B, we introduce the following assumption.

Assumption 1: $\nu_{ijn}^{(\mathrm{NMF})}, \nu_{ijn}^{(\mathrm{DNN})} \to \infty$ (i.e., $\nu^{(t\text{-PoP})} \to \infty$) while keeping η_{ijn} finite.

Since a complex isotropic Student's-t distribution becomes a complex isotropic Gaussian distribution as the DoF parameter goes to infinity, we can convert (29) into

$$p(y_{ijn}; \theta_{ijn}^{(\mathrm{G\text{-}PoP})}) = \frac{1}{\pi \tilde{r}_{ijn}^{(\mathrm{G\text{-}PoP})}} \exp\left(-\frac{|y_{ijn}|^2}{\tilde{r}_{ijn}^{(\mathrm{G\text{-}PoP})}}\right), \tag{58}$$

where $\theta_{ijn}^{(\mathrm{G\text{-}PoP})} := \left\{\eta_{ijn}, \tilde{r}_{ijn}^{(\mathrm{NMF})}, \tilde{r}_{ijn}^{(\mathrm{DNN})}\right\}$ and

$$\tilde{r}_{ijn}^{(\mathrm{G\text{-}PoP})} := \eta_{ijn}\tilde{r}_{ijn}^{(\mathrm{NMF})} + (1 - \eta_{ijn})\tilde{r}_{ijn}^{(\mathrm{DNN})}. \tag{59}$$

Since the source model is based on a complex isotropic Gaussian distribution, we call it *the G-PoP-based source model*.

From (59), the source power spectrogram $\tilde{r}_{ijn}^{(\mathrm{G\text{-}PoP})}$ is the η_{ijn}-weighted sum of the NMF- and DNN-based source models, which clarifies the relationship between the proposed source model and the source models of ILRMA and IDLMA, as shown in Fig. 4. This interpretation is true when $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ and $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ have values of similar magnitude. During training, $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ and $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ change by the parameter updates. The value range of $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ tends to be determined by the DNN training data. However, since $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$ are not normalized and the magnitude of y_{ijn} depends on that of $\boldsymbol{w}_{in}$, the value range of $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ depends on the to-be-separated data, which may result in increasing the difference in magnitude between $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ and $\tilde{r}_{ijn}^{(\mathrm{DNN})}$. Since this difference may change the substantive role of η_{ijn}, we will examine it in Section V-E.

B. Parameter Estimation Algorithm

1) Cost Function: By replacing the IDLMA's source model with the G-PoP-based source model, we can construct G-PoP-IDLMA. Its cost function is given as the negative log-likelihood of $\{\boldsymbol{X}_m\}_m$:

$$\begin{aligned}\mathcal{L}_{\text{G-PoP}} \stackrel{c}{=} &\sum_{i,j,n}\left[\frac{|\boldsymbol{w}_{in}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\eta_{ijn}\tilde{r}_{ijn}^{(\text{NMF})}+(1-\eta_{ijn})\tilde{r}_{ijn}^{(\text{DNN})}}\right]\\ &+\sum_{i,j,n}\log\left[\eta_{ijn}\tilde{r}_{ijn}^{(\text{NMF})}+(1-\eta_{ijn})\tilde{r}_{ijn}^{(\text{DNN})}\right]\\ &-2J\sum_i\log|\det\boldsymbol{W}_i|. \end{aligned} \tag{60}$$

By setting $\eta_{ijn}=1$ and $\eta_{ijn}=0$, $\mathcal{L}_{\text{G-PoP}}$ reduces to the cost functions of ILRMA (10) and IDLMA (18), respectively. This finding clarifies that PoP-IDLMA encompasses the source models of ILRMA and IDLMA.

Similarly to Section III-C, we describe the DNN training stage in Section IV-C and the separation stage in this section. In the following, we derive update rules of $\boldsymbol{W}_i$, $\tilde{t}_{ikn}$, $\tilde{v}_{kjn}$, and $\tilde{r}_{ijn}^{(\text{DNN})}$. Note that η_{ijn} is treated as a hyperparameter.

2) Update Rule of $\boldsymbol{W}_i$: Since the $\boldsymbol{w}_{in}$-related terms of $\mathcal{L}_{\text{G-PoP}}$ are only quadratic and log-determinant terms, we can use the IP algorithm as in t-PoP-IDLMA. Hence, the update rule of $\boldsymbol{W}_i$ is defined as

$$\{\boldsymbol{W}_i\}_i \leftarrow \mathsf{IP}\left(\{\boldsymbol{X}_m\}_m,\{\boldsymbol{W}_i\}_i,\left\{\boldsymbol{R}_n^{(\text{G-PoP})}\right\}_n\right). \tag{61}$$

It is identical to the update rule obtained by applying Assumption 1 to (43) because $\xi_{ijn}\to\tilde{r}_{ijn}^{(\text{G-PoP})}$ as $\nu_{ijn}^{(t\text{-PoP})}\to\infty$.

3) Update Rules of $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$: As in Section III-C3, we construct an auxiliary function of $\mathcal{L}_{\text{G-PoP}}$ and derive update rules of $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$. The difficulty in directly minimizing $\mathcal{L}_{\text{G-PoP}}$ with respect to $\tilde{t}_{ikn}$ and $\tilde{v}_{kjn}$ is that the first and second terms of (60) include the sums over k in the reciprocal and logarithmic functions, respectively. Applying inequalities (46) and (48) to these terms yields the following auxiliary function:

$$\begin{aligned}\mathcal{L}_{\text{G-PoP}}^{+} = &\sum_{ijn}|y_{ijn}|^2\sum_k\frac{\left(\lambda_{ijkn}^{(\text{NMF})}\right)^2}{\eta_{ijn}\tilde{t}_{ikn}\tilde{v}_{jkn}}\\ &+\sum_{i,j,n}\frac{\eta_{ijn}}{\gamma_{ijn}}\sum_k\tilde{t}_{ikn}\tilde{v}_{kjn}+\mathcal{D}_{-\tilde{t},\tilde{v}}^{(\text{G-PoP})},\end{aligned} \tag{62}$$

where $\mathcal{D}_{-\tilde{t},\tilde{v}}^{(\text{G-PoP})}$ denotes terms that do not include $\tilde{t}_{ikn}$ or $\tilde{v}_{kjn}$. By solving $\partial\mathcal{L}_{\text{G-PoP}}^{+}/\partial\tilde{t}_{ikn}=0$ and $\partial\mathcal{L}_{\text{G-PoP}}^{+}/\partial\tilde{v}_{kjn}=0$ and substituting the equality conditions into the solutions, we can obtain

$$\tilde{t}_{ikn}\leftarrow\tilde{t}_{ikn}\sqrt{\frac{\sum_j\eta_{ijn}\tilde{v}_{kjn}|y_{ijn}|^2/(\tilde{r}_{ijn}^{(\text{G-PoP})})^2}{\sum_j\eta_{ijn}\tilde{v}_{kjn}/\tilde{r}_{ijn}^{(\text{G-PoP})}}}, \tag{63}$$

$$\tilde{v}_{kjn}\leftarrow\tilde{v}_{kjn}\sqrt{\frac{\sum_i\eta_{ijn}\tilde{t}_{ikn}|y_{ijn}|^2/(\tilde{r}_{ijn}^{(\text{G-PoP})})^2}{\sum_i\eta_{ijn}\tilde{t}_{ikn}/\tilde{r}_{ijn}^{(\text{G-PoP})}}}. \tag{64}$$

Algorithm 3: Parameter Estimation Algorithm of G-PoP-IDLMA.

Input: $\{\boldsymbol{X}_m\}_m$, $\{\widehat{\text{DNN}}_n\}_n$, $\{\eta_{ijn}\}_{ijn}$
Output: $\{\boldsymbol{Y}_n\}_n$

```
for I_(out)^(G-PoP) iterations do
    Update {r̃_ijn^(DNN)}_ijn by (65) and (66)
    Update {r̃_ijn^(G-PoP)}_ijn by (59)
    for I_(in)^(G-PoP) iterations do
        Update {t̃_ijn}_ijn and {ṽ_ijn}_ijn by (63) and (64)
        Update {r̃_ijn^(G-PoP)}_ijn by (59)
        {W_i}_i ← IP({X_m}_m, {W_i}_i, {R_n^(G-PoP)}_n)
        Update {Y_n}_n by (5)
        Update {Y_n}_n by (15)
    end for
end for
```

Interestingly, these update rules coincide with (51) and (52) under Assumption 1.

4) Update Rule of $\tilde{r}_{ijn}^{(DNN)}$: In G-PoP-IDLMA, the DNN of the nth source $\widehat{\text{DNN}}_n$ estimates a source magnitude spectrogram $\tilde{\sigma}_{ijn}$ from $|\boldsymbol{Y}_n|^{\cdot 1}$. As in IDLMA, we update $\tilde{r}_{ijn}^{(\text{DNN})}$ as

$$\{\tilde{\sigma}_{ijn}\}_{ij}\leftarrow\widehat{\text{DNN}}_n(|\boldsymbol{Y}_n|^{\cdot 1}), \tag{65}$$

$$\tilde{r}_{ijn}^{(\text{DNN})}\leftarrow\max(\tilde{\sigma}_{ijn}^2,\epsilon_1). \tag{66}$$

5) Entire Procedure of Separation Stage: Fig. 1(b) shows the overview of the separation process of G-PoP-IDLMA, where the spatial and source models are iteratively updated. Algorithm 3 shows the entire parameter estimation algorithm of G-PoP-IDLMA. It has the inner and outer iterations to balance the update amount of the DNN and NMF parts similarly to Algorithm 2.

C. DNN Training

In the DNN training stage, we train $\widehat{\text{DNN}}_n$ so that it can estimate a clean magnitude spectrogram $|\breve{s}_{ijn}|$ from a noisy mixture. Since the NMF part is responsible for the components not included in the training data, we can set $\eta_{ijn}=0$ during the DNN training. The resultant cost function is given as

$$\mathcal{C}_{\text{G-PoP}}^{(n)}=\sum_{i,j}\left(\frac{|\breve{s}_{ijn}|^2+\varepsilon_2}{\tilde{\sigma}_{ijn}^2+\varepsilon_2}-\log\frac{|\breve{s}_{ijn}|^2+\varepsilon_2}{\tilde{\sigma}_{ijn}^2+\varepsilon_2}-1\right). \tag{67}$$

It is identical to the cost function for the DNN training in IDLMA (20). Hence, we can use the same DNN training procedure as in IDLMA.

It should be noted that (67) does not include η_{ijn}. Thus, once the DNN is trained, it can be used for any η_{ijn} values, which is the primary advantage of G-PoP-IDLMA compared with t-PoP-IDLMA.

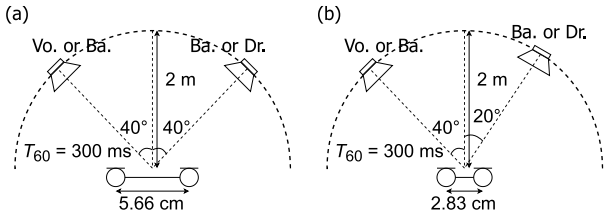


Fig. 5. Recording conditions of stereo mixtures.

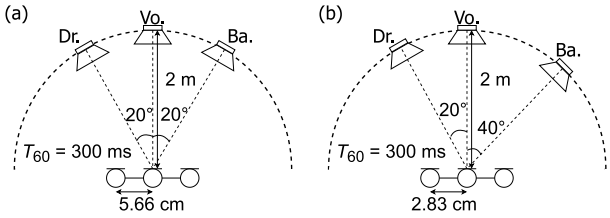
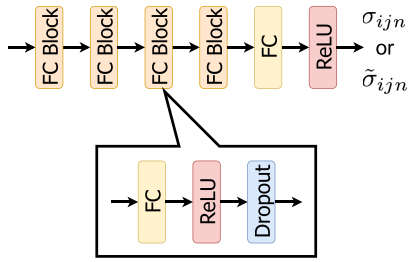
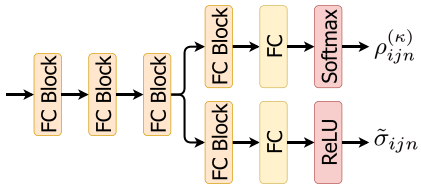


Fig. 6. Recording conditions of three-channel mixtures.



(a) DNN+WF, FSCM+DNN, IDLMA, t-IDLMA, and proposed G-PoP-IDLMA



(b) Proposed t-PoP-IDLMA

Fig. 7. DNN architectures used in experiments.

V. Experimental Evaluation

A. Experimental Settings

1) Common Settings: To evaluate the effectiveness of the proposed PoP-IDLMAs, we conducted experiments on determined multichannel music source separation using the DSD100 dataset [26]. This dataset consists of `dev` and `test` sets (50 songs per set) and separate recordings of vocals (Vo.), bass (Ba.), drums (Dr.), and other instruments. The recordings of Vo., Ba., and Dr. were used as dry sources.

We generated test data by extracting 30- to 60-s segments of the top 25 songs in the `test` set in alphabetical order and convolving them with the E2A impulse response ($T_{60} = 300$ ms) in the RWCP database [27]. The test data were composed of stereo and three-channel mixtures, where the number of channels equals that of sources, i.e., $N = M$. The other settings were as follows:

Stereo mixtures: The stereo mixtures were generated with two recording conditions for each pair of Vo., Ba., and Dr. (i.e., Ba./Dr., Vo./Ba., and Vo./Dr.). The number of mixtures was 50 for each instrument pair. The recording conditions are shown in Fig. 5.

Three-channel mixtures: The three-channel mixtures were also generated with two recording conditions, which are shown in Fig. 6. The sources were Vo., Ba., and Dr. (Vo./Ba./Dr.). The number of mixtures was 50.

The sampling frequency was set at 8 kHz as in [14]. For STFT, we used the hamming window of 512 ms (4096 samples) with a frame shift of 256 ms (2048 samples). The evaluation metric was the source-to-distortion ratio (SDR) improvement computed using the BSSEval toolbox [28].

2) Compared Methods: We compared the proposed PoP-IDLMAs with one BSS method and four source-supervised methods. The BSS method is *ILRMA* [8], which is the NMF-only counterpart of the proposed PoP-IDLMAs. The number of bases was set to $K = 20$. The initial values of t_{ikn} and v_{kjn} were drawn from a uniform distribution over [0,1), and $\boldsymbol{W}_i$ was initialized with an identity matrix. We did not use t-ILRMA for the comparison because it showed a similar performance to ILRMA as shown in [25].

The source-supervised methods were the combination of the DNN and the Wiener filter (*DNN+WF*) [29], the combination of the full-rank spatial covariance model with DNN (*FSCM+DNN*) [30], *IDLMA* [14], and *t-IDLMA* [14]. IDLMA and t-IDLMA are the DNN-only counterparts of the proposed PoP-IDLMAs. For these four methods, we used the same DNN architecture as in [14]. Fig. 7(a) shows this architecture. It consists of four fully connected (FC) blocks, an FC layer, and a rectified linear unit (ReLU) nonlinearity [31]. Each FC block is composed of an FC layer with 2048 hidden units, a ReLU nonlinearity, and a dropout layer with a drop rate of 0.3. For t-IDLMA, we set the DoF parameter $\nu = 500$, which provided the highest separation performance for the stereo and three-channel mixtures on average. For IDLMA and t-IDLMA, the demixing matrix $\boldsymbol{W}_i$ was initialized with an identity matrix.

The proposed methods are t-PoP-IDLMA and G-PoP-IDLMA. We set the number of basis $K = 20$ to match it with that in ILRMA. The initial values of $\tilde{t}_{ikn}, \tilde{v}_{kjn}$, and $\boldsymbol{W}_i$ were set in the same manner as those in ILRMA. The numbers of inner and outer iterations were set to 10: $(\mathcal{I}_{(\text{in})}^{(t\text{-PoP})}, \mathcal{I}_{(\text{out})}^{(t\text{-PoP})}) = (10, 10)$ for t-PoP-IDLMA and $(\mathcal{I}_{(\text{in})}^{(\text{G-PoP})}, \mathcal{I}_{(\text{out})}^{(\text{G-PoP})}) = (10, 10)$ for G-PoP-IDLMA. For t-PoP-IDLMA, we used the same DNN architecture as in [21], which has two heads for $\rho_{ijn}^{(\kappa)}$ and $\tilde{\sigma}_{ijn}$. Fig. 7(b) shows this architecture. We set $\mathcal{K} = \{1, 10, 100, 1000\}$ and varied $\nu_{ijn}^{(\text{NMF})} = 1, 10, 100$, and 1000. For G-PoP-IDLMA, we varied η_{ijn} as $\eta_{ijn} = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$, and 10^{-10} and used the same DNNs as those used in the source-supervised methods. Since the used values of $\nu_{ijn}^{(\text{NMF})}$ and η_{ijn} were independent of i, j, and n, we hereafter drop these indices from the two parameters for the simplicity.

3) DNN Training: For the DNN training, we used all 50 songs in the `dev` set of the DSD100 dataset as training data and the bottom 25 songs in alphabetical order in the `test` set as

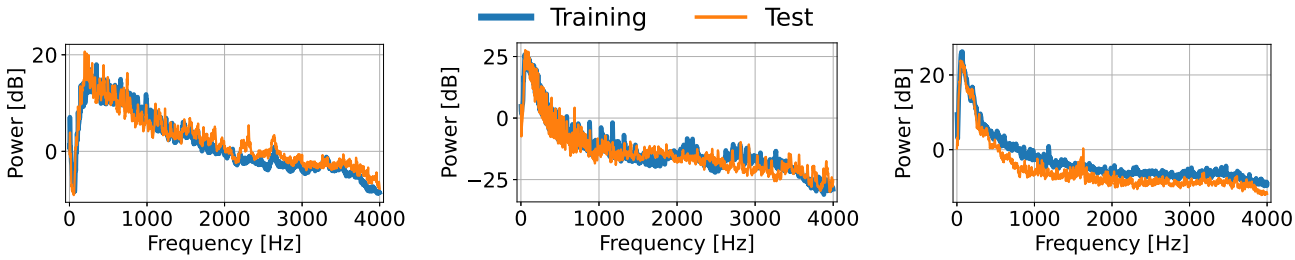


Fig. 8. Average power spectra of training and test data for each musical instrument.

TABLE I
SDR IMPROVEMENTS [DB] OF CONVENTIONAL AND PROPOSED METHODS FOR STEREO MIXTURES

Method	Source model	$\nu^{(\mathrm{NMF})}$	η	Mixture Ba./Dr.	Mixture Vo./Ba.	Mixture Vo./Dr.
ILRMA [8]	Unsupervised NMF	-	-	6.01	8.23	9.93
DNN+WF [29]	Supervised DNN	-	-	4.03	10.15	6.09
FSCM+DNN [30]	Supervised DNN	-	-	2.96	13.12	6.78
IDLMA [14]	Supervised DNN	-	-	6.41	12.52	12.99
t-IDLMA [14]	Supervised DNN	-	-	7.16	12.65	12.93
t-PoP-IDLMA	Proposed t-PoP-based source model (Unsupervised NMF+Supervised DNN)	10^0	-	**8.56**	**13.25**	13.25
		10^1	-	8.21	12.82	12.96
		10^2	-	8.48	12.77	13.24
		10^3	-	8.26	12.68	**13.72**
G-PoP-IDLMA	Proposed G-PoP-based source model (Unsupervised NMF+Supervised DNN)	-	10^{-2}	8.53	10.98	13.44
		-	10^{-4}	8.48	11.92	13.54
		-	10^{-6}	8.39	12.62	13.55
		-	10^{-8}	8.27	12.95	13.51
		-	10^{-10}	8.21	13.18	13.42

validation data. All DNNs were trained for 2000 epochs using an Adadelta [32] optimizer with a batch size of 128. The gradient clipping [33] was applied to the weights of the DNNs so that their l^2 norms were less than or equal to 10. We set $\varepsilon_1 = 10^{-1/2}$ and $\varepsilon_2 = 10^{-5}$ and the other training conditions were the same as those in [14].

B. Comparison of Average Spectra Between Training and Test Data

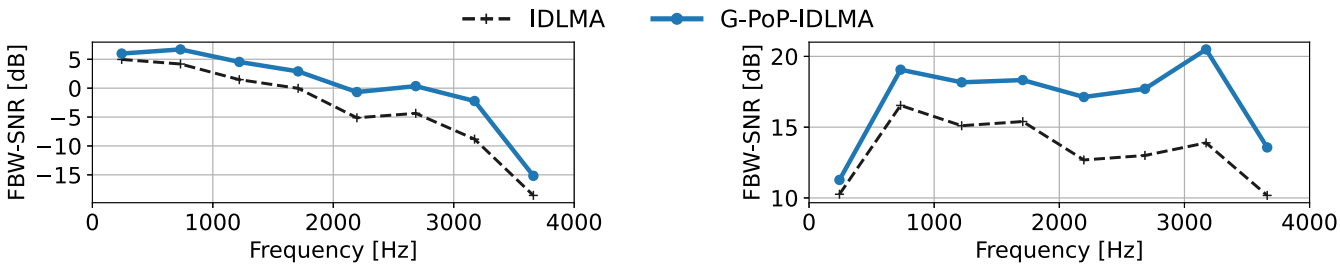
Before discussing the separation results, we examined the average spectra of the training and test data to show the timbral mismatches. Fig. 8 shows the average power spectra of the training and test data for each musical instrument. The spectra labeled as Training and Test were computed from the clean audio signals of the DNN training and the dry sources of the test data, respectively. For vocals and bass, the spectral differences between Training and Test were greater in the frequency band above 2000 Hz. For drums, the average spectrum of Training was apparently different from that of Test in the frequency band above around 500 Hz. These results show that the timbral mismatches were most pronounced in the higher frequency band.

C. Results for Stereo Mixtures

Table I shows average SDR improvements of all methods. The SDR improvements of DNN+WF and FSCM+DNN were greater than those of ILRMA for the Vo./Ba. mixture but smaller for the other stereo mixtures. IDLMA and t-IDLMA consistently provided greater average SDR improvements for all stereo mixtures, showing the stable performance of IDLMA. Although the conventional IDLMAs outperformed ILRMA by a large margin

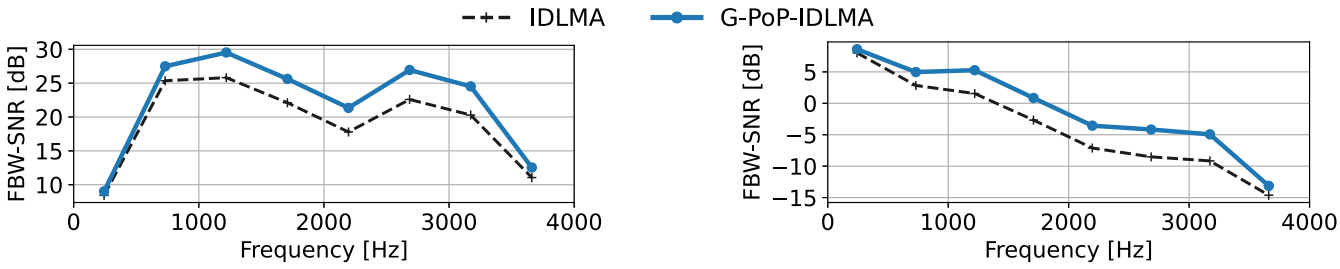
TABLE II
SDR IMPROVEMENTS [DB] OF CONVENTIONAL AND PROPOSED METHODS FOR THREE-CHANNEL MIXTURE

Method	$\nu^{(\mathrm{NMF})}$	η	Vo./Ba./Dr.
ILRMA [8]	-	-	5.75
DNN+WF [29]	-	-	5.70
FSCM+DNN [30]	-	-	6.31
IDLMA [14]	-	-	8.22
t-IDLMA [14]	-	-	8.55
t-PoP-IDLMA	10^0	-	7.30
	10^1	-	8.43
	10^2	-	8.13
	10^3	-	8.79
	10^4	-	9.58
	10^5	-	**9.67**
	10^6	-	9.52
G-PoP-IDLMA	-	10^{-2}	8.13
	-	10^{-4}	8.55
	-	10^{-6}	9.01
	-	10^{-8}	9.19
	-	10^{-10}	9.19



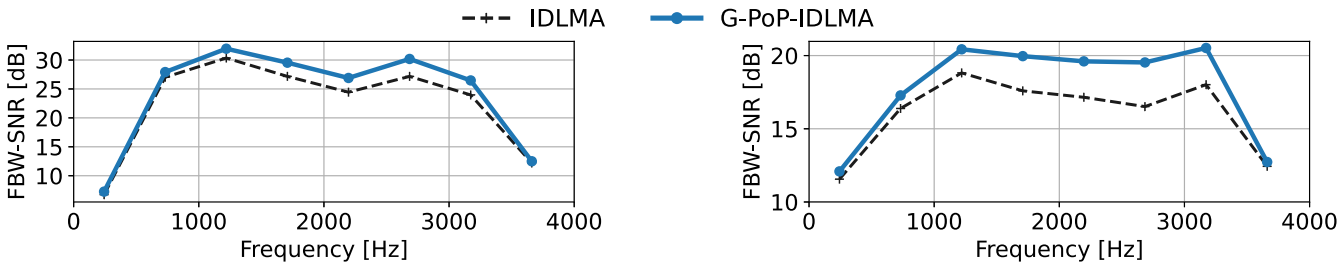
(a) Ba. (b) Dr.

Fig. 9. FBW-SNRs for Ba./Dr. mixtures.



(a) Vo. (b) Ba.

Fig. 10. FBW-SNRs for Vo./Ba. mixtures.



(a) Vo. (b) Dr.

Fig. 11. FBW-SNRs for Vo./Dr. mixtures.

(more than 2 dB) for the Vo./Ba. and Vo./Dr. mixtures, their differences in SDR were moderate (0.4 dB for IDLMA and 1 dB for t-IDLMA) for the Ba./Dr. mixtures. This may be because the spectrograms of bass and drums tend to be of low rank and are easier for NMF to represent.

t-PoP-IDLMA with all $\nu^{(\mathrm{NMF})}$ achieved greater and comparable SDR improvements to the conventional methods. G-PoP-IDLMA with all η achieved greater SDR improvements than the conventional methods for the Ba./Dr. and Vo./Dr. mixtures. For the Vo./Ba. mixture, it had greater SDR improvements with $\eta = 10^{-8}$ and 10^{-10}. Interestingly, the proposed PoP-IDLMAs outperformed the other methods by a large margin for the Ba./Dr. mixtures, where the SDR differences between ILRMA and IDLMAs were moderate. This result shows the effectiveness of unifying the NMF- and DNN-based source models.

t-PoP-IDLMA exhibited a greater separation performance than G-PoP-IDLMA, but their differences in SDR were slight. This result shows that the unification of the NMF- and DNN-based source models has a greater impact on SDR than the difference in probability distribution.

We observed a correlation between η values and the significance of the timbral mismatches. The smaller η provided slightly higher SDR improvements for the Ba./Dr. mixture, whereas the greater η had the higher SDR improvements for the Vo./Ba. mixture. This tendency correlates with the significance of the spectral differences between the training and test data as described in Section V-B. Although a clear tendency was not observed for the Vo./Dr. mixture, this result suggests that the greater η should be used as the timbral mismatches become more significant.

D. Results for Three-Channel Mixtures

Table II shows average SDR improvements for the three-channel mixtures. The SDR improvements of t-PoP-IDLMA monotonically increased in the range of $\nu^{(\mathrm{NMF})}$ used in Section V-C and we increased $\nu^{(\mathrm{NMF})}$ until they started to decrease. The IDLMA family consistently worked well compared with the other methods as in Section V-C. t-PoP-IDLMA ($\nu^{(\mathrm{NMF})} = 10^4, 10^5$, and 10^6) and G-PoP-IDLMA ($\eta =$

10^{-6}, 10^{-8}, and 10^{-10}) outperformed the conventional methods, showing the effectiveness of the proposed methods for more severe situations.

t-PoP-IDLMA achieved the highest SDR improvement with $\nu^{(\mathrm{NMF})} = 10^5$. However, it had lower SDR improvement than conventional IDLMA when $\nu^{(\mathrm{NMF})} = 10^0$, which was the best hyperparameter for the stereo mixtures. By contrast, G-PoP-IDLMA worked stably with $\eta = 10^{-8}$ and 10^{-10} for the stereo and three-channel mixtures. This performance stability is another advantage of G-PoP-IDLMA.

E. Effect of η

As described in Section IV-A, the G-PoP-based source model is identical to the DNN-based source model of IDLMA when $\eta = 0$. However, we experimentally observed that G-PoP-IDLMA behaved differently with IDLMA, although η decreased to a value close to zero 10^{-10}. To examine this phenomenon, we compared $\sum_{i,j,n} \tilde{r}_{ijn}^{(\mathrm{NMF})}$ and $\sum_{i,j,n} \tilde{r}_{ijn}^{(\mathrm{DNN})}$ along with the iterations. We hereafter call the two quantities the energies of the NMF and DNN parts, respectively.

We experimentally found that the energies of the NMF and DNN parts automatically became balanced as the iteration proceeded. At the early iterations, the energy of the NMF part was small and the DNN part dominated the demixing matrix updates. At the late iterations, the energy of the NMF part gradually became the same as that of the DNN part. This observation indicates that η practically determines how confident the NMF part is only at the early iterations. At the early iterations, since the NMF part is still in convergence, $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ is frequently more accurate than $\tilde{r}_{ijn}^{(\mathrm{NMF})}$. By contrast, at the late iterations, the NMF part converges well. Hence, $\tilde{r}_{ijn}^{(\mathrm{NMF})}$ and $\tilde{r}_{ijn}^{(\mathrm{DNN})}$ are equally useful for the demixing matrix estimation. Even when η is small, the NMF part affects the separation performance after a sufficient number of iterations were performed. This result clarifies the role and effectiveness of the NMF part.

If η was affected uniformly in all iterations, we needed to precisely control η along with the iterations. However, owing to the automatic energy balancing, the proposed methods are free from such painstaking tuning. This is another advantage of the PoP-based source model.

F. Effect of PoP-Based Source Model

To assess the effect of using the PoP-based source model, we compared G-PoP-IDLMA with IDLMA in terms of frequency-band-wise source-to-noise ratio (FBW-SNR). The FBW-SNR is defined as

$$\mathrm{SNR}_{\omega,n} = \frac{1}{\#B_\omega} \sum_{i \in B_\omega} 10 \log_{10} \frac{\sum_j |a_{inm_{\mathrm{ref}}} s_{ijn}|^2}{\sum_j |y_{ijn} - a_{inm_{\mathrm{ref}}} s_{ijn}|^2}, \tag{68}$$

where $\omega = 1, \ldots, 7$ is the frequency band index, B_ω is given as $B_\omega = \{250(\omega - 1) + 1, \ldots, 250\omega\}$, $\#B_\omega$ is the number of elements in B_ω, and $a_{inm_{\mathrm{ref}}}$ is the (m_{ref}, n)th entry of the mixing matrix $\boldsymbol{A}_i$.

Fig. 9 shows the average FBW-SNRs over 50 mixtures for the Ba./Dr. mixtures, where G-PoP-IDLMA was with $\eta_{ijn} = 10^{-10}$. The FBW-SNRs of G-PoP-IDLMA were higher than those of IDLMA in all the frequency bands and the improvements from IDLMA were remarkable in the frequency bands above 500 Hz, which is consistent with the average spectral difference shown in Section V-B. In these frequency bands, the DNN outputs had many zeros, whereas the NMF part succeeded in the source power estimation. We observed the same trends for the other stereo mixtures, as shown in Figs. 10 and 11. The FBW-SNR gaps between G-PoP-IDLMA and IDLMA were large particularly for drums and bass. This should be because the spectrograms of these instruments tend to match the low-rank assumption of NMF. These results show that the NMF part can compensate for the source power estimation in the frequency bands where the DNN part failed in power spectrogram estimation.

VI. Conclusion

We proposed two source models that encompass NMF- and DNN-based source models used in ILRMA and IDLMA, respectively. The proposed source models use the PoP, a prior distribution of the source power spectrogram, which is constructed by multiplying the probability distributions based on NMF and DNN in accordance with the PoE concept. Since the PoP can be written as an inverse gamma distribution, we can introduce the PoP-based source models into the IDLMA framework without violating the generative modeling. The resultant IDLMA extensions are t- and G-PoP-IDLMAs. For the proposed PoP-IDLMAs, we derived efficient parameter estimation algorithms on the basis of the MM algorithm. Experimental results showed the effectiveness of the proposed PoP-IDLMAs and the importance of unifying the NMF- and DNN-based source models. Furthermore, the assessment of the results clarified that the NMF part can compensate for the source power estimation in the frequency bands where the DNN part failed in the estimation.

References

[1] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. e12, pp. 1–14, 2019.

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

[3] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Int. Workshop ICA BSS*, pp. 365–371, 1999.

[4] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and K. Shikano, "Blind source separation based on subband ICA and beamforming," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, pp. 94–97.

[5] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[6] 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.

[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Proc. Int. Conf. Independent Compon. Anal. Signal Separation*, 2006, pp. 601–608.

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

[9] 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.
[10] 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, 2009.
[11] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2020, pp. 1–35, 2020.
[12] S. Mogami et al., "Independent low-rank matrix analysis based on generalized Kullback–Leibler divergence," *IEICE Trans. Fundam. Electron. Comm. Comput. Sci.*, vol. E102.A, no. 2, pp. 458–463, 2019.
[13] S. Mogami et al., "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model for determined blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 503–518, 2020.
[14] N. Makishima et al., "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, Oct. 2019.
[15] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[16] D. R. Hunter and K. Lange, "Quantile regression via an MM algorithm," *J. Comput. Graph. Stat.*, vol. 9, no. 1, pp. 60–77, 2000.
[17] A. Ozerov and C. Fevotte, "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.
[18] 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.
[19] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1950–1965, 2021.
[20] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2610–2625, 2020.
[21] T. Hasumi et al., "Empirical bayesian independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. Eur. Signal Process. Conf.*, 2021, pp. 331–335.
[22] 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.
[23] 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.
[24] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[25] 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. 1, pp. 1–25, 2018.
[26] A. Liutkus et al., "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2017, pp. 323–332.
[27] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2000, pp. 965–968.
[28] 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.
[29] 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.
[30] 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.
[31] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2011, pp. 315–323.
[32] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
[33] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

Takuya Hasumi received the B.E. degree from Waseda University, Tokyo, Japan, in 2020 and the M.S. degree from The University of Tokyo, Tokyo, in 2022. His research interests include audio source separation, text-to-speech synthesis, and music information retrieval. He is currently a Member of the Acoustical Society of Japan (ASJ). He was the recipient of the 22nd best student presentation award of ASJ.

Tomohiko Nakamura (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 2011, 2013, and 2016, respectively. He joined SECOM Intelligent Systems Laboratory as a Researcher in 2016 and moved to the University of Tokyo as Project Research Associate in 2019. He is currently a Senior Researcher with the National Institute of Advanced Industrial Science and Technology. His research interests include signal-processing-inspired deep learning, audio signal processing, and music signal processing. He was the recipient of the more than 10 article and achievement awards.

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 multichannel audio source separation and machine learning.

Hiroshi Saruwatari (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and the Nara Institute of Science and Technology, Japan, in 2000. Since 2014, he has been a Professor of The University of Tokyo, Japan. His research interests include statistical speech signal processing, blind source separation (BSS), audio enhancement, and robot audition. He has successfully achieved his carrier, especially on BSS researches, and put his research into the world's first commercially available Independent-Component-Analysis-based BSS microphone in 2007. He was the recipient of the article Awards from IEICE in 2001 and 2006, TAF in 2004, 2009 and 2012, IEEE-IROS2005 in 2006, and 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 Hattori-Hoko Award in 2018, and the first prize in IEEEMLSP2007 BSS Competition. He is professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ, including chair posts of international conferences and Associate Editor of journals.

Daichi Kitamura (Senior Member, IEEE) received the Ph.D. degree from SOKENDAI, Hayama, Japan. He joined The University of Tokyo, Tokyo, Japan, in 2017 as a Research Associate, and moved to the National Institute of Technology, Kagawa Collage, Takamatsu, Japan, 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 article Award from IEEE Signal Processing Society Japan in 2017, Itakura Prize Innovative Young Researcher Award from ASJ in 2018, and IEEE Signal Processing Society 2019 Young Author Best article Award.

Yu Takahashi (Member, IEEE) received the B.E. degree in information engineering from the Himeji Institute of Technology, Himeji, Japan, and the M.E. and Ph.D. degrees in information science from the Nara Institute of Science and Technology, Ikoma, Japan, in 2007 and 2010 respectively. He is currently a Researcher with Yamaha corporation. His research interests include statistical signal processing for audio and music.

Kazunobu Kondo received the B.E., M.E. and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2014, respectively. In 1993, he joined the Electronics Development Center, Yamaha Company, Ltd. He is currently a Principal Engineer and a Group Manager of Yamaha Research and Development Division. His research interests include blind source separation, noise reduction, and dereverbetation. He is a Member of the IEICE and the Acoustical Society of Japan.