Semi-Supervised Multichannel Speech Enhancement With a Deep Speech Prior

Kouhei Sekiguchi, *Member, IEEE*, Yoshiaki Bando, *Member, IEEE*, Aditya Arie Nugraha, *Member, IEEE*, Kazuyoshi Yoshii, *Member, IEEE*, and Tatsuya Kawahara, *Fellow, IEEE*

***Abstract*—This paper describes a semi-supervised multichannel speech enhancement method that uses clean speech data for prior training. Although multichannel nonnegative matrix factorization (MNMF) and its constrained variant called independent low-rank matrix analysis (ILRMA) have successfully been used for unsupervised speech enhancement, the low-rank assumption on the power spectral densities (PSDs) of all sources (speech and noise) does not hold in reality. To solve this problem, we replace a low-rank speech model with a deep generative speech model, i.e., formulate a probabilistic model of noisy speech by integrating a deep speech model, a low-rank noise model, and a full-rank or rank-1 model of spatial characteristics of speech and noise. The deep speech model is trained from clean speech data in an unsupervised auto-encoding variational Bayesian manner. Given multichannel noisy speech spectra, the full-rank or rank-1 spatial covariance matrices and PSDs of speech and noise are estimated in an unsupervised maximum-likelihood manner. Experimental results showed that the full-rank version of the proposed method was significantly better than MNMF, ILRMA, and the rank-1 version. We confirmed that the initialization-sensitivity and local-optimum problems of MNMF with many spatial parameters can be solved by incorporating the precise speech model.**

***Index Terms*—Multichannel speech enhancement, deep learning, variational autoencoder, nonnegative matrix factorization.**

I. Introduction

SPEECH enhancement plays a vital role for automatic speech recognition (ASR) in noisy environments. Although the performance and robustness of ASR have been drastically improved thanks to the development of deep learning techniques, ASR in unseen noisy environments that are not covered by training data is still an open problem. Many methods have thus been proposed for single-channel or multichannel speech enhancement. These methods can be categorized into supervised, semi-supervised, and unsupervised methods.

A popular approach to supervised speech enhancement is to train deep neural networks (DNNs) by using pairs of noisy and clean speech signals. In single-channel speech enhancement, one can use denoising autoencoders (DAEs) that take noisy speech spectra as input, and output clean speech spectra [1]. Alternatively, DNNs can be trained to estimate time-frequency masks, i.e., classify each time-frequency bin into speech or noise [2], [3]. In multichannel speech enhancement using phase information, the estimated masks are used for calculating the spatial covariance matrices (SCMs) of speech and noise. This allows one to use beamforming methods [4], [5]. Although this approach has successfully been used as a front end of ASR, the performance of speech enhancement is often considerably degraded in unseen noisy environments due to the nature of supervised mask estimation [6].

To mitigate the sensitivity to acoustic characteristics of noisy environments, one may use unsupervised methods such as multichannel extensions of nonnegative matrix factorization (NMF) [7]–[12]. Each variant of multichannel NMF (MNMF) can be interpreted as maximum-likelihood or Bayesian estimation of a probabilistic model representing the generative process of the complex spectrograms of mixture signals (e.g., speech + noise) and is used for general blind sound separation (BSS). The key assumption underlying the family of MNMF is that the power spectral densities (PSDs) of all sound sources have low-rank

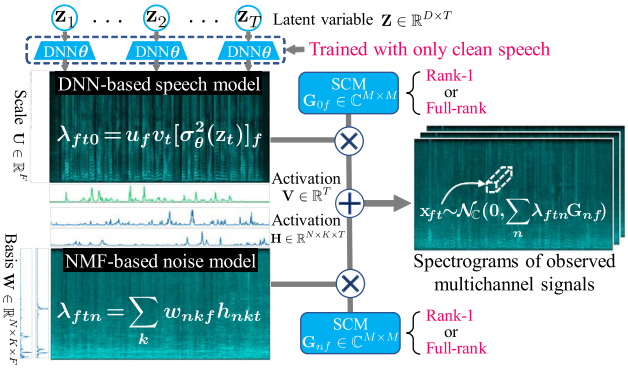


Fig. 1. A probabilistic generative model of multichannel noisy speech spectra with a deep speech prior.

Manuscript received January 31, 2019; revised May 28, 2019 and August 28, 2019; accepted September 11, 2019. Date of publication October 7, 2019; date of current version November 26, 2019. This work was supported by JST ERATO No. JPMJER1401 and JSPS KAKENHI No. 19H04137. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maria de Diego. *(Corresponding author: Kouhei Sekiguchi.)*

K. Sekiguchi and K. Yoshii are with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan, and also with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (e-mail: kouhei.sekiguchi@riken.jp; yoshii@kuis.kyoto-u.ac.jp).

Y. Bando is with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan (e-mail: y.bando@aist.go.jp).

A. A. Nugraha is with the Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan (e-mail: adityaarie.nugraha@riken.jp).

T. Kawahara is with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: kawahara@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2019.2944348

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

structure. In speech enhancement, however, the performance of MNMF is limited because the low-rank assumption does not hold for the PSDs of speech. Several studies thus integrated a DAE into an optimization step of MNMF which estimates the PSDs of speech [13], [14]. Although such integration of a powerful DNN and a physically founded statistical model is promising, supervised learning of DAEs causes sensitivity to noisy environments again.

To solve the problems of the DNN- and MNMF-based conventional methods, we propose a semi-supervised method that uses only clean speech data for prior training. More specifically, we formulate a DNN-based speech model that represents the generative process of the complicated PSDs of clean speech and an NMF-based noise model that represents the generative process of the low-rank PSDs of noise. A unified generative model of observed noisy speech is then obtained by integrating those source models with a full-rank or rank-1 spatial model as in MNMF [9] or its constrained version called independent low-rank matrix analysis (ILRMA) [11], respectively. A key feature of our method is that the deep speech model is a latent variable model that implicitly, but precisely represents the acoustic characteristics of speech spectra such as fundamental frequencies (F0s), harmonic structures, and spectral envelopes. To achieve this, the speech model is trained from clean speech data in an unsupervised variational auto-encoding manner. The noise model, on the other hand, is learned on-the-fly without pre-training. Given noisy speech as observed data, the latent variables of the speech model, the full-rank or rank-1 SCMs and PSDs of speech and noise can be estimated in an unsupervised maximum-likelihood manner by combining a majorization-minimization algorithm with Metropolis sampling or backpropagation.

A main contribution of this paper is to propose a new statistical framework that integrates a physically-founded linear model (multichannel spatial model) with a powerful deep generative model (single-channel source model) in a principled manner. Another important contribution is to experimentally show that the full-rank spatial model can outperform the rank-1 spatial model thanks to the deep speech prior. In our previous work [15], we developed the MNMF-based full-rank model with the deep speech prior, called MNMF-DP. In this paper, we propose the ILRMA-based rank-1 model with the deep speech prior, called ILRMA-DP, and investigate the configurations of these models. Note that MNMF [9] with richer expressive power often underperforms ILRMA [11] because MNMF is sensitive to the initialization of SCMs and tends to get stuck at local optima. Interestingly, our full-rank model outperforms the rank-1 version even when the SCMs are initialized randomly. This indicates that the precise source modeling helps the estimation of SCMs and alleviates the initialization sensitivity. We confirm the superiority of the proposed semi-supervised method over the semi-supervised versions of MNMF and ILRMA in which the basis spectra of the NMF-based speech model are trained from clean speech data as in [16], [17].

The rest of the paper is organized as follows. Section II reviews related work on NMF-based and DNN-based speech enhancement. Section III explains the proposed methods based on full-rank and rank-1 spatial models and Section IV describes parameter estimation. Section V reports comparative experiments. Finally, Section VI concludes this paper.

II. Related Work

This section reviews existing NMF- and DNN-based speech enhancement methods in comparison with the proposed method consisting of a DNN-based speech model and an NMF-based noise model.

A. NMF-Based Speech Enhancement

Multichannel extensions of NMF have been developed for using the spatial information of sound propagation processes [7]–[12]. The PSDs of each source signal is given by the sum of the products of basis spectra and their activations. The complex spectrograms of observed multichannel signals are then given by the sum of the complex spectrograms (called *images*) of the propagated source signals. The first formulation of MNMF was proposed by Ozerov *et al.* [7], where the SCMs are restricted to rank-1 matrices and the cost function based on the Itakura-Saito (IS) divergence is minimized by using a multiplicative update or expectation-maximization (EM) algorithm. This method was extended to deal with full-rank SCMs [8]. Sawada *et al.* [9] introduced a partitioning function to share a set of basis spectra by all sources and derived a majorization-minimization (MM) algorithm. The full-rank version of our model can be regarded as an extension of [8] and uses the MM algorithm of [9]. Nikunen and Virtanen [10] proposed a model similar to [9] which represents the SCM of each source as the weighted sum of all possible direction-dependent SCMs. Kitamura *et al.* [11] proposed a method called independent low-rank matrix analysis (ILRMA) by restricting the SCMs of [9] to rank-1 matrices, resulting in a unified model of NMF and independent vector analysis (IVA). The rank-1 version of our model can be regarded as an extension of ILRMA without a partitioning function.

The common feature of those MNMF variants is that the PSDs of each source are assumed to have a low-rank structure given by nonnegative matrix factorization (NMF) [18]. While it is difficult to use NMF in an unsupervised manner for single-channel speech enhancement, MNMF can work well in an unsupervised manner because the spatial information plays a central role in multi-channel speech enhancement. NMF has been used in a semi-supervised manner by training the basis spectra of speech from clean speech data in advance [16], [17]. In this paper, we evaluate a semi-supervised version of MNMF with pretrained basis spectra for fair comparison with the proposed semi-supervised method.

B. DNN-Based Speech Enhancement

Deep neural networks (DNNs) have been widely used for supervised speech enhancement. A typical approach to single-channel speech enhancement is to train a denoising autoencoder (DAE) that takes noisy speech spectra as input and outputs clean speech spectra by using paired data [1]. Alternatively, one can train a DNN that outputs time-frequency masks [2], [3]. The use of spatial information has recently been investigated for

multi-channel speech enhancement. In [4], [5], time-frequency masks are estimated using a DNN and then used for calculating the steering vectors and/or SCMs of speech and noise used for beamforming. In [13], [14], a DAE is integrated into a process of SCM-based multichannel source separation; (1) the observed mixture spectra are separated into speech and noise by using the current estimate of the speech and noise SCMs, (2) the PSDs of the enhanced speech are further refined by using the DAE, and (3) the speech and noise SCMs are updated by using the current estimate of the PSDs of the speech and noise. Generative adversarial networks (GANs) are effective for supervised single-channel speech enhancement [19]–[21], but adaptation to unseen noisy environments is an open problem.

Recently, deep generative models of speech spectra based on variational autoencoders (VAEs) have been used for semi-supervised speech enhancement. Bando *et al.* [22] first proposed a unified model that consists of an NMF-based noise model and a DNN-based speech model with latent variables for single-channel speech enhancement. The speech model is given as the decoder of a VAE trained beforehand from clean speech data in an unsupervised manner. On the other hand, the noise model is optimized on-the-fly for observed noisy speech data. This approach mitigates the sensitivity to the acoustic characteristics of noisy environments. Leglaive *et al.* [23] proposed a similar model for maximum likelihood estimation. In our work [15], we developed a multichannel extension of [22] to treat the SCMs of speech and noise. This method can also be considered as an extension of Bayesian MNMF [12] because an NMF-based model corresponding to speech was replaced with a DNN-based model. Leglaive *et al.* [24] also proposed a similar model for maximum likelihood estimation. In this paper, we propose a rank-1 variant of [15], investigate parameter configuration, initialization, and latent variable estimation for both full-rank and rank-1 models, and compare them with the state-of-the-art methods.

III. Probabilistic Modeling

We explain the proposed multichannel speech enhancement methods that integrate two kinds of source models—a DNN-based generative model of speech spectra and an NMF-based generative model of noise spectra—with a full-rank or rank-1 spatial model in a unified probabilistic model.

A. Problem Specification

Let T, F, M, and N be the number of time frames, frequency bins, microphones, and noise sources, respectively. Let $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ be the observed multichannel complex spectra of noisy speech including a single speech source and N noise sources. There are $N+1$ sources in total. Let $\mathbf{s}_{ft} = [s_{ft0},\ldots,s_{ftN}]^{\mathrm{T}} \in \mathbb{C}^{N+1}$ be the source spectrum at frequency f and time t, where s_{ft0} and the others correspond to the speech and the noise, respectively. Let $\mathbf{x}_{ftn} = [x_{ftn1},\ldots,x_{ftnM}]^{\mathrm{T}} \in \mathbb{C}^M$ be the image of source n. Assuming the additivity of complex spectra, the observed spectrum

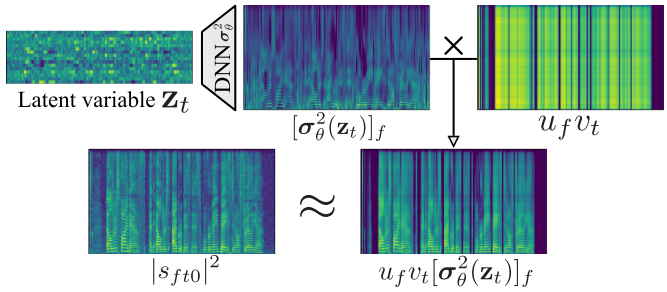


Fig. 2. The proposed DNN-based speech model. The PSDs $\{\lambda_{ft0}\}_{f=1}^F$ of a speech spectrum $\{s_{ft0}\}_{f=1}^F$ at time t are obtained by feeding the latent variable $\mathbf{z}_t$ following the standard Gaussian distribution into a DNN $\boldsymbol{\sigma}_{\theta}^2$ with parameters θ and then scaling the output $\boldsymbol{\sigma}_{\theta}^2(\mathbf{z}_t)$ according to u_f and v_t.

$\mathbf{x}_{ft} = [x_{ft1},\ldots,x_{ftM}]^{\mathrm{T}} \in \mathbb{C}^M$ is given by

$$\mathbf{x}_{ft} = \sum_{n=0}^{N} \mathbf{x}_{ftn}. \tag{1}$$

Given $\mathbf{X}$ as observed data, the goal of speech enhancement is to estimate the speech image $\mathbf{x}_{ft0}$.

B. Source Modeling

We formulate a source model that represents the generative process of the complex spectrum s_{ftn} of each source n. In this paper s_{ftn} is assumed to be circularly-symmetric complex Gaussian distributed as follows:

$$s_{ftn} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{ftn}), \tag{2}$$

where $\mathcal{N}_{\mathbb{C}}(\mu,\sigma^2)$ indicates a univariate circularly-symmetric complex Gaussian distribution with mean μ and variance σ^2, and λ_{ftn} indicates the PSD of source n at frequency f and time t. As in MNMF [7]–[12], a noise model is based on NMF assuming that the PSDs of noise have a low-rank structure. Since the low-rank assumption is not suitable for speech, a speech model is based on a DNN [22].

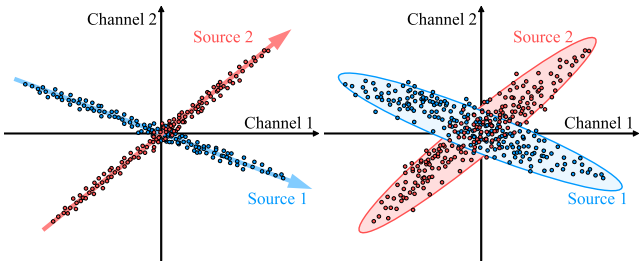
1) DNN-Based Speech Model: As in Fig. 2, the PSD of the speech at frequency f and time t is determined by a DNN as follows:

$$\lambda_{ft0} = u_f v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_t)]_f, \tag{3}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\cdot)$ is a nonlinear function (DNN) with parameters $\boldsymbol{\theta}$ that maps a D-dimensional real vector $\mathbf{z}_t \in \mathbb{R}^D$ to an F-dimensional nonnegative vector $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z}_t) \in \mathbb{R}_+^F$, $[\cdot]_f$ indicates the f-th element of a vector, $u_f \geq 0$ is a scaling factor at frequency f, and $v_t \geq 0$ is an activation at time t. $\mathbf{z}_t$ implicitly represents the characteristics (e.g., fundamental frequencies (F0s), harmonic structures, and formants) of the PSDs $\{\lambda_{ft0}\}_{f=1}^F$ of the speech at time t. Note that $\{\lambda_{ft0}\}_{f=1}^F$ are jointly determined by $\mathbf{z}_t$ in an interdependent manner. We put a standard Gaussian prior on $\mathbf{z}_t$ as follows:

$$\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D), \tag{4}$$

where $\mathbf{0}_D$ and $\mathbf{I}_D$ are the all-zero vector and the identity matrix of size D, respectively. While the DNN specified by $\boldsymbol{\theta}$



(a) Rank-1 model (b) Full-rank model

Fig. 3. Two variants of spatial models. Blue and red dots indicate source images $\{\mathbf{x}_{ft1}\}_{t=1}^{T}$ and $\{\mathbf{x}_{ft2}\}_{t=1}^{T}$ in frequency f, respectively. In the rank-1 model, dots are distributed on steering vectors $\mathbf{a}_{1f}$ and $\mathbf{a}_{2f}$. In the full-rank model, dots are widely and elliptically distributed.

is trained from clean speech data (Section IV-A), the latent variables $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1}^{T}$ are estimated on-the-fly. The scaling factors $\mathbf{U} = \{u_f\}_{f=1}^{F}$ and the activations $\mathbf{V} = \{v_t\}_{t=1}^{T}$ are introduced for dissolving the scale ambiguity of model parameters (Section III-D).

2) NMF-Based Noise Model: The PSD of each source of the noise ($n \geq 1$) at frequency f and time t is represented in the framework of NMF as follows:

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf} h_{nkt}, \tag{5}$$

where K denotes the number of bases, $w_{nkf} \geq 0$ indicates the magnitude of basis k of source n at frequency f, and $h_{nkt} \geq 0$ indicates the activation of basis k of source n at time t. $\mathbf{W} = \{w_{nkf}\}_{n=1,k=1,f=1}^{N,K,F}$ and $\mathbf{H} = \{h_{nkt}\}_{n=1,k=1,t=1}^{N,K,T}$ are estimated on-the-fly in speech enhancement for $\mathbf{X}$.

C. Spatial Modeling

We formulate a spatial model that represents the sound propagation process between each source n and the M microphones. In this paper, we use two variants of spatial models, a full-rank model with full-rank SCMs and a rank-1 model with rank-1 SCMs, as shown in Fig. 3. In the rank-1 model, we assume a time-invariant linear mixing system and its corresponding demixing system as follows:

$$\mathbf{x}_{ftn} = \mathbf{a}_{nf} s_{ftn}, \tag{6}$$

$$s_{ftn} = \mathbf{d}_{nf}^{\mathrm{H}} \mathbf{x}_{ft}, \tag{7}$$

where $\mathbf{a}_{nf} \in \mathbb{C}^M$ and $\mathbf{d}_{nf} \in \mathbb{C}^M$ are the steering vector and demixing filter of source n at frequency f, respectively. Using Eq. (2) and Eq. (6), we say

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_M, \lambda_{ftn} \mathbf{G}_{nf} \right), \tag{8}$$

where $\mathbf{G}_{nf} = \mathbf{a}_{nf} \mathbf{a}_{nf}^{\mathrm{H}} \in \mathbb{S}_{+}^{M}$ is the rank-1 SCM of source n at frequency f and $\mathbb{S}_{+}^{M}$ indicates the set of positive definite matrices of size M. Using Eq. (1), Eq. (8), and the reproductive property of the Gaussian distribution, we say

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}_M, \sum_{n=0}^{N} \lambda_{ftn} \mathbf{G}_{nf} \right). \tag{9}$$

As shown in Eq. (8), $\mathbf{G}_{nf}$ is a rank-1 matrix in an idealized situation. However, in a real noisy environment, $\mathbf{G}_{nf}$ can be a full-rank matrix due to reverberation and reflection. Therefore, in the full-rank model, we assume $\mathbf{G}_{nf}$ is a full-rank matrix. The number of parameters of a full-rank SCM is $M(M+1)/2$ and that of a rank-1 SCM is only M. While the rank-1 model is a restricted version of the full-rank model, ILRMA based on the rank-1 model [11] is empirically known to work better than MNMF based on the full-rank model [9] because the rank-1 model is less sensitive to parameter initialization. Note that this has not been confirmed in the proposed model with the deep speech prior (interestingly, the opposite results were obtained as shown in Section V).

The rank-1 model is valid only in a determined condition in which the number of sources is equal to that of microphones, i.e., $N + 1 = M$. Substituting Eq. (6) into Eq. (1), we get

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft}, \tag{10}$$

where $\mathbf{A}_f = [\mathbf{a}_{0f}, \ldots, \mathbf{a}_{Nf}] \in \mathbb{C}^{M \times (N+1)}$ is a non-singular square matrix called a mixing matrix. If $\mathbf{A}_f$ is given, the source spectrum $\mathbf{s}_{ft}$ can be estimated as follows:

$$\mathbf{s}_{ft} = \mathbf{D}_f \mathbf{x}_{ft}, \tag{11}$$

where $\mathbf{D}_f = \mathbf{A}_f^{-1} = [\mathbf{d}_{0f}, \ldots, \mathbf{d}_{Nf}]^{\mathrm{H}} \in \mathbb{C}^{(N+1) \times M}$ is a demixing matrix.

D. Unified Source and Spatial Modeling

We formulate a unified probabilistic model that represents the generative process of the observed data $\mathbf{X}$ by integrating the source models described in Section III-B with the spatial models described in Section III-C

1) MNMF with a Deep Speech Prior (MNMF-DP): Substituting Eq. (3) and Eq. (5) into Eq. (9), we obtain the likelihood function of unknown variables $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ for $\mathbf{X}$ as follows:

$$\begin{aligned} &\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}) \\ &= \sum_{f=1}^{F} \sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}} \left(\mathbf{x}_{ft} | \mathbf{0}_M, \mathbf{Y}_{ft} \right) \\ &= \sum_{f=1}^{F} \sum_{t=1}^{T} \left(-\mathrm{tr}\left(\mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \right) - \log |\mathbf{Y}_{ft}| \right) + \mathrm{const}, \end{aligned} \tag{12}$$

where $\mathbf{X}_{ft} \in \mathbb{S}_{+}^{M}$ and $\mathbf{Y}_{ft} \in \mathbb{S}_{+}^{M}$ are observed and reconstructed matrices given by

$$\mathbf{X}_{ft} = \mathbf{x}_{ft} \mathbf{x}_{ft}^{\mathrm{H}}, \tag{13}$$

$$\mathbf{Y}_{ft} = \sum_{n=0}^{N} \lambda_{ftn} \mathbf{G}_{nf}. \tag{14}$$

λ_{ft0} and $\lambda_{ftn}(n \geq 1)$ are the PSD of the speech and that of the noise, respectively, which are given by

$$\lambda_{ftn} = \begin{cases} u_f v_t [\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z}_t)]_f & (n = 0), \\ \sum_{k=1}^{K} w_{nkf} h_{nkt} & (n \geq 1). \end{cases} \tag{15}$$

We define $\mathbf{Y}_{ftn} = \lambda_{ftn} \mathbf{G}_{nf}$ and $\mathbf{Y}_{ftnk} = w_{nkf} h_{nkt} \mathbf{G}_{nf}$.

Our goal is to estimate $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ such that the log-likelihood $\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G})$ given by Eq. (12) is maximized. To avoid the scale ambiguity of the parameters, we put normalization constraints on $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{G}$ as follows:

$$\sum_{f=1}^{F} u_f = 1, \tag{16}$$

$$\sum_{f=1}^{F} w_{nkf} = 1, \tag{17}$$

$$\mathrm{tr}(\mathbf{G}_{nf}) = 1. \tag{18}$$

2) ILRMA with a Deep Speech Prior (ILRMA-DP): When $\mathbf{G}_{nf}$ is a rank-1 matrix given by $\mathbf{G}_{nf} = \mathbf{a}_{nf} \mathbf{a}^{\mathrm{H}}_{nf}$, $\mathbf{Y}_{ft}$ is given as follows:

$$\begin{aligned} \mathbf{Y}_{ft} &= \sum_{n=0}^{N} \lambda_{ftn} \mathbf{a}_{nf} \mathbf{a}^{\mathrm{H}}_{nf} \\ &= \mathbf{A}_f \boldsymbol{\Lambda}_{ft} \mathbf{A}^{\mathrm{H}}_f \\ &= \mathbf{D}^{-1}_f \boldsymbol{\Lambda}_{ft} \mathbf{D}^{-\mathrm{H}}_f, \end{aligned} \tag{19}$$

where $\boldsymbol{\Lambda}_{ft} = \mathrm{Diag}(\lambda_{ft0}, \ldots, \lambda_{ftN})$ is a diagonal matrix. Substituting Eq. (11) and Eq. (19) into Eq. (12), we get

$$\begin{aligned} &\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{D}) \\ &= -\sum_{f=1}^{F} \sum_{t=1}^{T} \mathrm{tr}\left(\mathbf{s}^{\mathrm{H}}_{ft} \mathbf{D}^{-\mathrm{H}}_f \left(\mathbf{D}^{\mathrm{H}}_f \boldsymbol{\Lambda}^{-1}_{ft} \mathbf{D}_f \right) \mathbf{D}^{-1}_f \mathbf{s}_{ft} \right) \\ &\quad - \sum_{f=1}^{F} \sum_{t=1}^{T} \log \left| \mathbf{D}^{-1}_f \boldsymbol{\Lambda}_{ft} \mathbf{D}^{-\mathrm{H}}_f \right| + \mathrm{const} \\ &= -\sum_{f=1}^{F} \sum_{t=1}^{T} \mathrm{tr}\left(\mathbf{s}^{\mathrm{H}}_{ft} \boldsymbol{\Lambda}^{-1}_{ft} \mathbf{s}_{ft} \right) - \sum_{f=1}^{F} \sum_{t=1}^{T} \log |\boldsymbol{\Lambda}_{ft}| \\ &\quad + T \sum_{f=1}^{F} \log |\mathbf{D}_f \mathbf{D}^{\mathrm{H}}_f| + \mathrm{const} \\ &= -\sum_{f=1}^{F} \sum_{t=1}^{T} \sum_{n=0}^{N} \left(\frac{|s_{ftn}|^2}{\lambda_{ftn}} + \log \lambda_{ftn} \right) \\ &\quad + T \sum_{f=1}^{F} \log |\mathbf{D}_f \mathbf{D}^{\mathrm{H}}_f| + \mathrm{const}. \end{aligned} \tag{20}$$

Our goal is to estimate the demixing matrices $\mathbf{D}$ instead of the mixing matrices $\mathbf{A}$ and to estimate $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$ such that the log-likelihood $\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{D})$ given by Eq. (20) is maximized. To avoid the scale ambiguity of the parameters, we put the normalization constraints on $\mathbf{U}$ and $\mathbf{W}$ given by Eq. (16) and Eq. (17) and that on $\mathbf{D}$ given by

$$\mathrm{tr}\left(\mathbf{d}_{nf} \mathbf{d}^{\mathrm{H}}_{nf} \right) = \mathbf{d}^{\mathrm{H}}_{nf} \mathbf{d}_{nf} = 1. \tag{21}$$

E. Speech Enhancement

Using the estimated parameters, we can perform statistical speech enhancement.

1) Full-Rank Model: To estimate the enhanced speech spectrum $\mathbf{x}^{\mathrm{FR}}_{ft0} \in \mathbb{C}^M$, we use a multichannel Wiener filter (MWF). Using Eq. (8) and Eq. (9), the posterior expectation of the speech image $\mathbf{x}^{\mathrm{FR}}_{ft0} \in \mathbb{C}^M$ is given as follows:

$$\mathbf{x}^{\mathrm{FR}}_{ft0} = \mathbb{E}[\mathbf{x}_{ft0}|\mathbf{x}_{ft}] = \mathbf{Y}_{ft0} \mathbf{Y}^{-1}_{ft} \mathbf{x}_{ft}. \tag{22}$$

2) Rank-1 Model: To estimate the enhanced speech spectrum $s^{\mathrm{R1}}_{ft0} \in \mathbb{C}$, we use a linear demixing filter as follows:

$$s^{\mathrm{R1}}_{ft0} = \mathbf{d}^{\mathrm{H}}_{0f} \mathbf{x}_{ft}. \tag{23}$$

To solve the scale ambiguity of $\{s_{ft0}\}_{f=1}^{F}$ over frequency bins, we use a projection back technique [25] for estimating the enhanced speech image $\mathbf{x}^{\mathrm{R1}}_{ft0} \in \mathbb{C}^M$ as follows:

$$\mathbf{x}^{\mathrm{R1}}_{ft0} = \mathbf{a}_{0f} s_{ft0} = \mathbf{a}_{0f} \mathbf{d}^{\mathrm{H}}_{0f} \mathbf{x}_{ft}. \tag{24}$$

Substituting Eq. (19) into Eq. (22), we can easily prove that Eq. (24) can also be obtained by the MWF as follows:

$$\begin{aligned} \mathbf{x}^{\mathrm{R1}}_{ft0} &= \left(\lambda_{ft0} \mathbf{a}_{0f} \mathbf{a}^{\mathrm{H}}_{0f} \right) \left(\mathbf{D}^{\mathrm{H}}_f \boldsymbol{\Lambda}^{-1}_{ft} \mathbf{D}_f \right) \mathbf{x}_{ft} \\ &= \lambda_{ft0} \mathbf{a}_{0f} \mathbf{e}^{\mathrm{T}}_1 \boldsymbol{\Lambda}^{-1}_{ft} \mathbf{D}_f \mathbf{x}_{ft} \\ &= \mathbf{a}_{0f} \mathbf{d}^{\mathrm{H}}_{0f} \mathbf{x}_{ft}. \end{aligned} \tag{25}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^{\mathrm{T}}$ is a one-hot vector.

IV. Parameter Estimation

We explain how to train the DNN-based speech model (Section III-B1) from clean speech data in an unsupervised manner. We then explain how to optimize the parameters of MNMF-DP (Section III-D1) and those of ILRMA-DP (Section III-D2) for semi-supervised speech enhancement using the trained deep speech prior.

A. Pretraining of Deep Speech Prior

The nonlinear mapping function $\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\cdot)$ given by Eq. (3) is optimized in the framework of a VAE. Suppose that we have training data $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{I}$, where I is the number of frames and $\tilde{\mathbf{x}}_i \in \mathbb{C}^F$ is a complex spectrum of clean speech. Let $\tilde{\mathbf{Z}} = \{\tilde{\mathbf{z}}_i\}_{i=1}^{I}$ be the corresponding latent variables. We formulate the hierarchical generative process of $\tilde{\mathbf{X}}$ as follows:

$$\begin{aligned} \tilde{\mathbf{z}}_i &\sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D), \\ \tilde{\mathbf{x}}_i &\sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_F, \mathrm{Diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_i))), \end{aligned} \tag{26}$$

where $\mathrm{Diag}(\cdot)$ indicates a diagonal matrix.

Our goal is to estimate $\boldsymbol{\theta}$ such that the likelihood $p(\tilde{\mathbf{X}}|\boldsymbol{\theta})$ is maximized. Since $\log p(\tilde{\mathbf{X}}|\boldsymbol{\theta})$ is analytically intractable and is hard to directly maximize, we derive a lower bound $\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ of $\log p(\tilde{\mathbf{X}}|\boldsymbol{\theta})$ by introducing a variational posterior distribution $q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i)$ with parameters $\boldsymbol{\phi}$ as follows:

$$\begin{aligned}\log p(\tilde{\mathbf{X}}|\boldsymbol{\theta}) &= \sum_{i=1}^{I} \log \int p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i) p(\tilde{\mathbf{z}}_i) d\tilde{\mathbf{z}}_i \\ &= \sum_{i=1}^{I} \log \int \frac{q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i)}{q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i)} p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i) p(\tilde{\mathbf{z}}_i) d\tilde{\mathbf{z}}_i \\ &\geq \sum_{i=1}^{I} \int q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i) \log \frac{p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i) p(\tilde{\mathbf{z}}_i)}{q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i)} d\tilde{\mathbf{z}}_i \\ &= \sum_{i=1}^{I} \left(\mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i)] - \mathrm{KL}(q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i) \| p(\tilde{\mathbf{z}}_i)) \right) \\ &\stackrel{\mathrm{def}}{=} \mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}), \end{aligned} \tag{27}$$

where $\mathrm{KL}(q\|p)$ indicates the Kullback-Leibler (KL) divergence between two probability distributions q and p. Our goal is to maximize $\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

For mathematical convenience, in this paper $q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i)$ is set to a Gaussian distribution as follows:

$$q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i) = \mathcal{N}(\tilde{\mathbf{z}}_i|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i), \mathrm{Diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i))), \tag{28}$$

where $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\cdot)$ and $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\cdot)$ are the D-dimensional output vectors of a DNN with parameters $\boldsymbol{\phi}$. The first term of Eq. (27) is approximated via Monte Carlo integration as follows:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i)] \approx \frac{1}{L} \sum_{l=1}^{L} \log p_{\boldsymbol{\theta}} \left(\tilde{\mathbf{x}}_i|\tilde{\mathbf{z}}_i^{(l)} \right), \tag{29}$$

where L is the number of samples and $\tilde{\mathbf{z}}_i^{(l)}$ is obtained by using the reparametrization trick [26] as follows:

$$\tilde{\boldsymbol{\epsilon}}_i^{(l)} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D), \tag{30}$$

$$\tilde{\mathbf{z}}_i^{(l)} = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i) + \tilde{\boldsymbol{\epsilon}}_i^{(l)} \odot \sqrt{\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i)}, \tag{31}$$

where $\odot$ indicates the element-wise product. The second term of Eq. (27) can be analytically calculated as follows:

$$\begin{aligned}&\mathrm{KL}(q_{\boldsymbol{\phi}}(\tilde{\mathbf{z}}_i|\tilde{\mathbf{x}}_i) \| p(\tilde{\mathbf{z}}_i)) \\ &= \frac{1}{2} \sum_{d=1}^{D} \left([\boldsymbol{\mu}_{\phi}(\tilde{\mathbf{x}}_i)]_d^2 + [\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i)]_d - \log[\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i)]_d - 1 \right). \end{aligned} \tag{32}$$

The lower bound $\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ given by Eq. (27) can be approximately calculated by using Eq. (29), Eq. (30), Eq. (31), and Eq. (32). The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ of the two DNNs are jointly optimized by using a stochastic gradient method such that $\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is maximized.

The generation parameters $\boldsymbol{\theta}$ are used for formulating the generative model of $\mathbf{X}$ described in Section III. The inference parameters $\boldsymbol{\phi}$ are used for initializing $\mathbf{Z}$, i.e., $\mathbf{z}_t \leftarrow \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_t)$ as described in Section IV-D, where $\mathbf{x}_t$ is any complex spectrum whose PSDs are the same as the average PSDs of noisy speech over all channels at frame t.

B. Optimization of MNMF-DP

We aim to estimate the parameters $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ that maximize $\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G})$ given by Eq. (12). Since it is hard to directly maximize the log-likelihood with respect to each of these parameters, we use an MM algorithm that iteratively maximizes lower bounds of the log-likelihood as in MNMF [9].

1) Matrix Inequalities: To derive the lower bounds, we use two matrix inequalities on positive semidefinite matrices [12]. For a convex function $f_1(\mathbf{S}) = -\log|\mathbf{S}|$ with respect to $\mathbf{S} \in \mathbb{S}_+^M$, we calculate a tangent plane at an arbitrary point $\boldsymbol{\Omega} \in \mathbb{S}_+^M$ by using a first-order Taylor expansion as follows:

$$-\log|\mathbf{S}| \geq -\log|\boldsymbol{\Omega}| - \mathrm{tr}\left(\boldsymbol{\Omega}^{-1}\mathbf{S}\right) + M, \tag{33}$$

where the equality holds if and only if $\boldsymbol{\Omega} = \mathbf{S}$. For a concave function $f_2(\mathbf{S}) = -\mathrm{tr}(\mathbf{S}^{-1}\mathbf{R})$ with any matrix $\mathbf{R} \in \mathbb{S}_+^M$ with respect to $\mathbf{S} \in \mathbb{S}_+^M$, we have

$$-\mathrm{tr}\left(\left(\sum_{k=1}^{K} \mathbf{S}_k \right)^{-1} \mathbf{R} \right) \geq -\sum_{k=1}^{K} \mathrm{tr}\left(\mathbf{S}_k^{-1} \boldsymbol{\Phi}_k \mathbf{R} \boldsymbol{\Phi}_k^{\mathrm{H}} \right), \tag{34}$$

where $\{\mathbf{S}_k\}_{k=1}^K$ ($\mathbf{S}_k \in \mathbb{S}_+^M$) is a set of positive semidefinite matrices, $\{\boldsymbol{\Phi}_k\}_{k=1}^K$ is a set of auxiliary matrices that sum to the identity matrix, i.e., $\sum_{k=1}^K \boldsymbol{\Phi}_k = \mathbf{I}_M$, and the equality holds if and only if $\boldsymbol{\Phi}_k = \mathbf{S}_k(\sum_{k'=1}^K \mathbf{S}_{k'})^{-1}$.

2) Deriving Lower Bounds: Using Eq. (33) and Eq. (34) and introducing auxiliary matrices $\boldsymbol{\Omega} = \{\boldsymbol{\Omega}_{ft}\}_{f,t=1}^{F,T}$ and $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{ft0}\}_{f,t=1}^{F,T} \cup \{\boldsymbol{\Phi}_{ftn}\}_{f,t,n=1}^{F,T,N}$, we can derive a lower bound $\mathcal{L}_{\mathrm{FR}}^1(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \boldsymbol{\Omega}, \boldsymbol{\Phi})$ of Eq. (12) as follows:

$$\begin{aligned}&\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}) \\ &\geq -\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{n=0}^{N} \lambda_{ftn}^{-1} \mathrm{tr}\left(\mathbf{G}_{nf}^{-1} \boldsymbol{\Phi}_{ftn} \mathbf{X}_{ft} \boldsymbol{\Phi}_{ftn}^{\mathrm{H}} \right) \\ &\quad - \sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{n=0}^{N} \lambda_{ftn} \mathrm{tr}\left(\mathbf{G}_{nf} \boldsymbol{\Omega}_{ft}^{-1} \right) \\ &\quad - \sum_{f=1}^{F}\sum_{t=1}^{T} \log|\boldsymbol{\Omega}_{ft}| + \mathrm{const} \end{aligned} \tag{35}$$

$$\stackrel{\mathrm{def}}{=} \mathcal{L}_{\mathrm{FR}}^1(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \boldsymbol{\Omega}, \boldsymbol{\Phi}), \tag{36}$$

where the equality holds, i.e., the lower bound is maximized, if and only if

$$\boldsymbol{\Omega}_{ft} = \mathbf{Y}_{ft}, \tag{37}$$

$$\boldsymbol{\Phi}_{ftn} = \mathbf{Y}_{ftn} \mathbf{Y}_{ft}^{-1}. \tag{38}$$

Note that $\mathbf{Y}_{ft} = \sum_{n=0}^{N} \mathbf{Y}_{ftn}$ and $\mathbf{Y}_{ftn} = \lambda_{ftn} \mathbf{G}_{nf}$.

Using $\mathbf{\Omega}$ and $\mathbf{\Phi}$ and introducing additional auxiliary matrices $\mathbf{\Psi} = \{\mathbf{\Psi}_{ftnk}\}_{f,t,n,k=1}^{F,T,N,K}$, we can derive another lower bound, $\mathcal{L}_{\mathrm{FR}}^{2}(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \mathbf{\Omega}, \mathbf{\Phi}, \mathbf{\Psi})$ of Eq. (12) as follows:

$$\begin{aligned}
&\log p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}) \\
&\geq -\sum_{f=1}^{F}\sum_{t=1}^{T} u_f^{-1} v_t^{-1} [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f^{-1} \mathrm{tr}\Big(\mathbf{G}_{0f}^{-1}\mathbf{\Phi}_{ft0}\mathbf{X}_{ft}\mathbf{\Phi}_{ft0}^{\mathrm{H}}\Big) \\
&\quad -\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{k=1}^{K} w_{nkf}^{-1} h_{nkt}^{-1} \mathrm{tr}\Big(\mathbf{G}_{nf}^{-1}\mathbf{\Psi}_{ftnk}\mathbf{X}_{ft}\mathbf{\Psi}_{ftnk}^{\mathrm{H}}\Big) \\
&\quad -\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{n=0}^{N} \lambda_{ftn} \mathrm{tr}\Big(\mathbf{G}_{nf}\mathbf{\Omega}_{ft}^{-1}\Big) \\
&\quad -\sum_{f=1}^{F}\sum_{t=1}^{T} \log|\mathbf{\Omega}_{ft}| + \mathrm{const} \\
&\stackrel{\mathrm{def}}{=} \mathcal{L}_{\mathrm{FR}}^{2}(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \mathbf{\Omega}, \mathbf{\Phi}, \mathbf{\Psi}),
\end{aligned} \tag{39}$$

where the equality conditions are Eq. (37), Eq. (38), and

$$\mathbf{\Psi}_{ftnk} = \mathbf{Y}_{ftnk}\mathbf{Y}_{ft}^{-1}. \tag{40}$$

Note that $\mathbf{Y}_{ft} = \sum_{n=0}^{N}\mathbf{Y}_{ftn}$ and $\mathbf{Y}_{ftnk} = w_{nkf}h_{nkt}\mathbf{G}_{nf}$. Since $\mathcal{L}_{\mathrm{FR}}^{1}$ is tighter than $\mathcal{L}_{\mathrm{FR}}^{2}$, it is better to use $\mathcal{L}_{\mathrm{FR}}^{1}$ for parameter estimation if possible. However, maximization of $\mathcal{L}_{\mathrm{FR}}^{1}$ with respect to $\mathbf{W}$ and $\mathbf{H}$ has no closed-form solution due to the existence of $\lambda_{ftn}^{-1} = (\sum_k w_{nkf}h_{nkt})^{-1}$ $(n \geq 2)$ in the first term of Eq. (35). We thus use $\mathcal{L}_{\mathrm{FR}}^{1}$ for estimating $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{G}$, and use $\mathcal{L}_{\mathrm{FR}}^{2}$ for $\mathbf{W}$ and $\mathbf{H}$.

3) Updating Speech Model: To update the latent variables $\mathbf{Z}$, we use the Metropolis sampling [27] or the backpropagation [28]. In the sampling, a proposal $\mathbf{z}_t^{\mathrm{new}} \sim \mathcal{N}(\mathbf{z}_t^{\mathrm{old}}, \xi\mathbf{I}_D)$ with a small number ξ is accepted as a next sample of $\mathbf{z}_t$ with probability $\beta_t = \min(1, \gamma_t)$, where γ_t is given by

$$\begin{aligned}
\log\gamma_t &= \mathcal{L}_{\mathrm{FR}}^{1}(\mathbf{z}_t^{\mathrm{new}}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \mathbf{\Omega}, \mathbf{\Phi}) + \log p(\mathbf{z}_t^{\mathrm{new}}) \\
&\quad - \mathcal{L}_{\mathrm{FR}}^{1}(\mathbf{z}_t^{\mathrm{old}}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{H}, \mathbf{G}, \mathbf{\Omega}, \mathbf{\Phi}) - \log p(\mathbf{z}_t^{\mathrm{old}}) \\
&= -\sum_{f=1}^{F}\left(\frac{1}{\lambda_{ft0}^{\mathrm{new}}} - \frac{1}{\lambda_{ft0}^{\mathrm{old}}}\right)\mathrm{tr}\Big(\mathbf{G}_{0f}^{-1}\mathbf{\Phi}_{ft0}\mathbf{X}_{ft}\mathbf{\Phi}_{ft0}^{\mathrm{H}}\Big) \\
&\quad -\sum_{f=1}^{F}\left(\lambda_{ft0}^{\mathrm{new}} - \lambda_{ft0}^{\mathrm{old}}\right)\mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{\Omega}_{ft}^{-1}\Big) \\
&\quad -\frac{1}{2}\sum_{d=1}^{D}\left((z_{td}^{\mathrm{new}})^2 - (z_{td}^{\mathrm{old}})^2\right),
\end{aligned} \tag{41}$$

where $\lambda_{ft0}^{\mathrm{new}} = u_f v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t^{\mathrm{new}})]_f$ and $\lambda_{ft0}^{\mathrm{old}} = u_f v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t^{\mathrm{old}})]_f$. In the backpropagation, the lower bound $\mathcal{L}_{\mathrm{FR}}^{1}$ given by Eq. (36) is regarded as an objective function of $\mathbf{Z}$. It is maximized with respect to $\mathbf{z}_t$ by using a stochastic gradient descent method. Both sampling and backpropagation algorithms update $\mathbf{Z}$ several times in one iteration. In practice, we update $\mathbf{Z}$ several times without updating $\mathbf{Y}_{ft}$ to reduce the computational cost of calculating $\mathbf{Y}_{ft}^{-1}$ in $\mathbf{\Phi}_{ft0}$ and $\mathbf{\Omega}_{ft}^{-1}$.

To derive the multiplicative updating (MU) rule of the scaling factors $\mathbf{U}$, we let the partial derivative of $\mathcal{L}_{\mathrm{FR}}^{1}$ given by Eq. (36) with respect to u_f equal to zero as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} u_f^{-2} v_t^{-1} [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f^{-1} \mathrm{tr}\Big(\mathbf{G}_{0f}^{-1}\mathbf{\Phi}_{ft0}\mathbf{X}_{ft}\mathbf{\Phi}_{ft0}^{\mathrm{H}}\Big) \\
&\quad - \sum_{t=1}^{T} v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f \mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{\Omega}_{ft}^{-1}\Big) = 0.
\end{aligned} \tag{42}$$

Substituting Eq. (37) and Eq. (38) including the current estimate of u_f denoted by u_f^{old} into Eq. (42), we have

$$u_f^{\mathrm{old}} a_f^u u_f^{\mathrm{old}} = u_f b_f^u u_f, \tag{43}$$

where a_f^u and b_f^u are given by

$$a_f^u = \sum_{t=1}^{T} v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f \mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\Big), \tag{44}$$

$$b_f^u = \sum_{t=1}^{T} v_t [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f \mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{Y}_{ft}^{-1}\Big). \tag{45}$$

Solving Eq. (43), we have the MU rule of u_f given by

$$u_f \leftarrow u_f \sqrt{\frac{a_f^u}{b_f^u}}. \tag{46}$$

Similarly, the MU rule of the activations $\mathbf{V}$ can be obtained as follows:

$$a_t^v = \sum_{f=1}^{F} u_f [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f \mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\Big), \tag{47}$$

$$b_t^v = \sum_{f=1}^{F} u_f [\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{2}(\mathbf{z}_t)]_f \mathrm{tr}\Big(\mathbf{G}_{0f}\mathbf{Y}_{ft}^{-1}\Big), \tag{48}$$

$$v_t \leftarrow v_t \sqrt{\frac{a_t^v}{b_t^v}}. \tag{49}$$

4) Updating Noise Models: Letting the partial derivatives of $\mathcal{L}_{\mathrm{FR}}^{2}$ given by Eq. (39) with respect to $\mathbf{W}$ and $\mathbf{H}$ equal to zero, the closed-form MU rules of $\mathbf{W}$ and $\mathbf{H}$ are obtained as follows:

$$a_{nkf}^w = \sum_{t=1}^{T} h_{nkt} \mathrm{tr}\Big(\mathbf{G}_{nf}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\Big), \tag{50}$$

$$b_{nkf}^w = \sum_{t=1}^{T} h_{nkt} \mathrm{tr}\Big(\mathbf{G}_{nf}\mathbf{Y}_{ft}^{-1}\Big), \tag{51}$$

$$w_{nkf} \leftarrow w_{nkf} \sqrt{\frac{a_{nkf}^w}{b_{nkf}^w}}, \tag{52}$$

$$a_{nkt}^h = \sum_{f=1}^{F} w_{nkf} \mathrm{tr}\Big(\mathbf{G}_{nf}\mathbf{Y}_{ft}^{-1}\mathbf{X}_{ft}\mathbf{Y}_{ft}^{-1}\Big), \tag{53}$$

$$b^h_{nkt} = \sum_{f=1}^{F} w_{nkf} \operatorname{tr}\left(\mathbf{G}_{nf}\mathbf{Y}^{-1}_{ft}\right), \tag{54}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{a^h_{nkt}}{b^h_{nkt}}}. \tag{55}$$

5) Updating Spatial Models: To derive the update rule of the spatial covariance matrices $\mathbf{G}$, we let the partial derivative of $\mathcal{L}^1_{\mathrm{FR}}$ with respect to $\mathbf{G}_{nf}$ equal to zero as follows:

$$\sum_{t=1}^{T} \lambda^{-1}_{ftn}\mathbf{G}^{-1}_{nf}\mathbf{\Phi}_{ftn}\mathbf{X}_{ft}\mathbf{\Phi}^{\mathrm{H}}_{ftn}\mathbf{G}^{-1}_{nf} - \sum_{t=1}^{T}\lambda_{ftn}\mathbf{\Omega}^{-1}_{ft} = \mathbf{0}_{M\times M}, \tag{56}$$

where $\mathbf{0}_{M\times M}$ is the all-zero matrix of size $M \times M$. Substituting Eq. (37) and Eq. (38) including the current estimate of $\mathbf{G}_{nf}$ denoted by $\mathbf{G}^{\mathrm{old}}_{nf}$ into Eq. (56), we have

$$\mathbf{G}^{\mathrm{old}}_{nf}\mathbf{A}^{\mathbf{G}}_{nf}\mathbf{G}^{\mathrm{old}}_{nf} = \mathbf{G}_{nf}\mathbf{B}^{\mathbf{G}}_{nf}\mathbf{G}_{nf}, \tag{57}$$

where $\mathbf{A}^{\mathbf{G}}_{nf} \in \mathbb{S}^M_+$ and $\mathbf{B}^{\mathbf{G}}_{nf} \in \mathbb{S}^M_+$ are given by

$$\mathbf{A}^{\mathbf{G}}_{nf} = \sum_{t=1}^{T}\lambda_{ftn}\mathbf{Y}^{-1}_{ft}\mathbf{X}_{ft}\mathbf{Y}^{-1}_{ft}, \tag{58}$$

$$\mathbf{B}^{\mathbf{G}}_{nf} = \sum_{t=1}^{T}\lambda_{ftn}\mathbf{Y}^{-1}_{ft}. \tag{59}$$

Solving Eq. (57) as in [29], [30], we have the closed-form update rule of $\mathbf{G}_{nf}$ as follows:

$$\mathbf{G}_{nf} \leftarrow \left(\mathbf{G}_{nf}\mathbf{A}^{\mathbf{G}}_{nf}\mathbf{G}_{nf}\right)\#(\mathbf{B}^{\mathbf{G}}_{nf})^{-1}, \tag{60}$$

where $\mathbf{A}\#\mathbf{B}$ indicates the geometric mean of two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$ [31], [32] as follows:

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}}\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{B})^{\frac{1}{2}}. \tag{61}$$

6) Normalizing Parameters: To meet the normalization constraints given by Eq. (16), Eq. (17), and Eq. (18), we adjust the scales of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ in each iteration as follows:

$$\mu_{nf} = \operatorname{tr}(\mathbf{G}_{nf}), \quad \begin{cases} \mathbf{G}_{nf} \leftarrow \mu^{-1}_{nf}\mathbf{G}_{nf}, \\ u_f \leftarrow \mu_{0f}u_f, \\ w_{nkf} \leftarrow \mu_{nf}w_{nkf}\ (n \geq 1), \end{cases} \tag{62}$$

$$\nu_0 = \sum_{f=1}^{F} u_f, \quad \begin{cases} u_f \leftarrow \nu_0^{-1}u_f, \\ v_t \leftarrow \nu_0 v_t, \end{cases} \tag{63}$$

$$\nu_{nk} = \sum_{f=1}^{F} w_{nkf}, \quad \begin{cases} w_{nkf} \leftarrow \nu^{-1}_{nk}w_{nkf}, \\ h_{nkt} \leftarrow \nu_{nk}h_{nkt}. \end{cases} \tag{64}$$

C. Optimization of ILRMA-DP

We aim to estimate the parameters $\mathbf{Z}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{D}$ that maximize $\log p(\mathbf{X}|\mathbf{Z},\mathbf{U},\mathbf{V},\mathbf{W},\mathbf{H},\mathbf{D})$ given by Eq. (20) by using an MM algorithm as in ILRMA [11].

1) Updating Speech Model: The latent variables $\mathbf{Z}$ are updated with Metropolis sampling or backpropagation as in the full-rank model (Section IV-B3). In the sampling, instead of Eq. (41), γ_t is given by

$$\log\gamma_t = \sum_{f=1}^{F}\left(\frac{|s_{ft0}|^2}{\lambda^{\mathrm{old}}_{ft0}} - \frac{|s_{ft0}|^2}{\lambda^{\mathrm{new}}_{ft0}} + \log\frac{\lambda^{\mathrm{old}}_{ft0}}{\lambda^{\mathrm{new}}_{ft0}}\right) - \frac{1}{2}\sum_{d=1}^{D}\left((z^{\mathrm{new}}_{td})^2 - (z^{\mathrm{old}}_{td})^2\right). \tag{65}$$

In the backpropagation, the likelihood given by Eq. (20) is regarded as a negative cost function.

The update rules of $\mathbf{U}$ and $\mathbf{V}$ can be obtained directly by letting the partial derivatives of $\log p(\mathbf{X}|\mathbf{Z},\mathbf{U},\mathbf{V},\mathbf{W},\mathbf{H},\mathbf{D})$ equal to zero as follows:

$$u_f \leftarrow \frac{1}{T}\sum_{t=1}^{T}\frac{|s_{ft0}|^2}{v_t[\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z}_t)]_f}, \tag{66}$$

$$v_t \leftarrow \frac{1}{F}\sum_{f=1}^{F}\frac{|s_{ft0}|^2}{u_f[\boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z}_t)]_f}. \tag{67}$$

2) Updating Noise Models: The closed-form MU rules of $\mathbf{W}$ and $\mathbf{H}$ are obtained as follows:

$$a^w_{nkf} = \sum_{t=1}^{T} h_{nkt}|s_{ftn}|^2\lambda^{-2}_{ftn}, \tag{68}$$

$$b^w_{nkf} = \sum_{t=1}^{T} h_{nkt}\lambda^{-1}_{ftn}, \tag{69}$$

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{a^w_{nkf}}{b^w_{nkf}}}, \tag{70}$$

$$a^h_{nkt} = \sum_{f=1}^{F} w_{nkf}|s_{ftn}|^2\lambda^{-2}_{ftn}, \tag{71}$$

$$b^h_{nkt} = \sum_{f=1}^{F} w_{nkf}\lambda^{-1}_{ftn}, \tag{72}$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{a^h_{nkt}}{b^h_{nkt}}}. \tag{73}$$

3) Updating Spatial Models: The update rule of $\mathbf{D}$ is obtained in the same way to [11], [33] as follows:

$$\mathbf{\Upsilon}_{nf} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbf{X}_{ft}}{\lambda_{ftn}}, \tag{74}$$

$$\mathbf{d}_{nf} \leftarrow (\mathbf{D}_f\mathbf{\Upsilon}_{nf})^{-1}\mathbf{e}_n, \tag{75}$$

$$\mathbf{d}_{nf} \leftarrow \left(\mathbf{d}^{\mathrm{H}}_{nf}\mathbf{\Upsilon}_{nf}\mathbf{d}_{nf}\right)^{-\frac{1}{2}}\mathbf{d}_{nf}, \tag{76}$$

where $\mathbf{e}_n = [0,\ldots,1,\ldots,0]^{\mathrm{T}}$ indicates a unit vector with the n-th element equal to 1.

Algorithm 1: Speech Enhancement Based on MNMF-DP.

for iteration $= 1$ to MaxIteration **do**
  Update $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ by Eqs. (46), (49), (52), (55), and (60)
  Compute $\mathbf{\Omega}$ and $\{\mathbf{\Phi}_{ft0}\}_{f,t=1}^{F,T}$ by Eqs. (37) and (38)
  if Sampling **then**
    for Z_iteration $= 1$ to Z_MaxIteration **do**
      for $t = 1$ to T **do**
        Sample $\mathbf{z}_t^{\mathrm{new}}$ from $\mathcal{N}_{\mathbb{C}}(\mathbf{z}_t, \xi\mathbf{I}_D)$
        Compute γ_t by Eq. (41)
        Sample q from $\mathrm{Uniform}(0,1)$
        if $\gamma_t > q$ **then** $\mathbf{z}_t \leftarrow \mathbf{z}_t^{\mathrm{new}}$
      end for
    end for
  end if
  if Backpropagation **then**
    for Z_iteration $= 1$ to Z_MaxIteration **do**
      Compute $\mathcal{L}_{\mathrm{FR}}^1$ by Eq. (35)
      Update $\mathbf{Z}$ by Adam with $\mathcal{L}_{\mathrm{FR}}^1$
    end for
  end if
  Normalize parameters by Eqs. (62), (63), and (64)
end for
Compute $\mathbf{x}_{ft0}^{\mathrm{FR}}$ by Eq. (22)

Algorithm 2: Speech Enhancement Based on ILRMA-DP.

for iteration $= 1$ to MaxIteration **do**
  Update $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$ by Eqs. (66), (67), (70), and (73)
  if Sampling **then**
    for Z_iteration $= 1$ to Z_MaxIteration **do**
      for $t = 1$ to T **do**
        Sample $\mathbf{z}_t^{\mathrm{new}}$ from $\mathcal{N}_{\mathbb{C}}(\mathbf{z}_t, \xi\mathbf{I}_D)$
        Compute γ_t by Eq. (65)
        Sample q from $\mathrm{Uniform}(0,1)$
        if $\gamma_t > q$ **then** $\mathbf{z}_t \leftarrow \mathbf{z}_t^{\mathrm{new}}$
      end for
    end for
  end if
  if Backpropagation **then**
    for Z_iteration $= 1$ to Z_MaxIteration **do**
      Compute the log likelihood by Eq. (20)
      Update $\mathbf{Z}$ by Adam with the log likelihood
    end for
  end if
  Update $\mathbf{D}$ by Eq. (75)
  Normalize parameters by Eqs. (63), (64), and (77)
end for
Compute $\mathbf{x}_{ft0}^{\mathrm{R1}}$ by Eq. (24)

4) Normalizing Parameters: To meet the normalization constraints given by Eq. (16), Eq. (17), and Eq. (21), we normalize $\mathbf{D}$ as follows:

$$\mu_{nf} = \mathbf{d}_{nf}^{\mathrm{H}}\mathbf{d}_{nf}, \quad \begin{cases} \mathbf{d}_{nf} \leftarrow \mu_{nf}^{-\frac{1}{2}}\mathbf{d}_{nf}, \\ u_f \leftarrow \mu_{0f}^{-1}u_f, \\ w_{nkf} \leftarrow \mu_{nf}^{-1}w_{nkf}\ (n \geq 1), \end{cases} \tag{77}$$

We then normalize $\mathbf{U}$ and $\mathbf{W}$ by using Eq. (63) and Eq. (64).

D. Initialization of MNMF-DP and ILRMA-DP

It is crucial to appropriately initialize the scaling factors $\mathbf{U}$, the speech activations $\mathbf{V}$, the speech latent variables $\mathbf{Z}$, the basis spectra $\mathbf{W}$, the noise activations $\mathbf{H}$, and the SCMs $\mathbf{G}$ or the demixing matrices $\mathbf{D}$. We use the inference model of the VAE specified by $\boldsymbol{\phi}$ for initializing $\mathbf{Z}$ as $\mathbf{z}_t \leftarrow \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}_t)$. $\mathbf{U}$ and $\mathbf{V}$ are initialized as $\mathbf{u} = \frac{1}{F}\mathbf{1}_F$ and $\mathbf{v} = \mathbf{1}_T$.

Considering Eq. (17), the initial values of $\mathbf{W}$ are sampled from a Dirichlet distribution as follows:

$$\mathbf{w}_{nk} \sim \mathrm{Dirichlet}(\alpha_0\mathbf{1}_F), \tag{78}$$

where $\mathbb{E}_{\mathrm{init}}[w_{nkf}] = \frac{1}{F}$ and α_0 is a concentration parameter ($\alpha_0 = 2$ in our experiments). Considering Eq. (17), Eq. (18), and the scale of the observed PSDs, the initial values of $\mathbf{H}$ are sampled from gamma distributions as follows:

$$h_{nkt} \sim \mathrm{Gamma}\left(\alpha_0, \frac{\alpha_0}{\mathbb{E}_{\mathrm{emp}}[|x|^2]}\frac{NK}{FM}\right), \tag{79}$$

where $\mathbb{E}_{\mathrm{init}}[h_{nkt}] = \frac{FM}{NK}\mathbb{E}_{\mathrm{emp}}[|x|^2]$ and $\mathbb{E}_{\mathrm{emp}}[|x|^2]$ indicates the empirical mean of the observed PSDs given by

$$\mathbb{E}_{\mathrm{emp}}\left[|x|^2\right] = \frac{1}{FTM}\sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{m=1}^{M}|x_{ftm}|^2. \tag{80}$$

Since the initialization of $\mathbf{G}$ or $\mathbf{D}$ is considered to have a strong impact on the performance of speech enhancement, we propose and compare several initialization methods.

1) MNMF-DP: $\mathbf{G}$ can be initialized without using the observed data $\mathbf{X}$. The most naive way of initialization is to set $\mathbf{G}_{nf}$ to the identity matrix as follows:

$$\mathbf{G}_{nf} \leftarrow \frac{1}{M}\mathbf{I}_M. \tag{81}$$

Note that under a determined condition with $M = N + 1$ only, one can directly associate $N + 1$ sources with M channels one by one as follows:

$$\mathbf{G}_{nf} \leftarrow \mathrm{Diag}(\mathbf{e}_n), \ \text{i.e.,}\ \mathbf{a}_{nf} \leftarrow \mathbf{e}_{n+1}, \tag{82}$$

Alternatively, $\mathbf{G}$ can be initialized in an adaptive manner by using the observed data $\mathbf{X}$. Assuming that the target speech is predominant in $\mathbf{X}$, one may set the speech SCM $\mathbf{G}_{0f}$ to the average of the observed SCMs and the noise SCMs to the identity matrix as follows:

$$\begin{cases} \mathbf{G}_{0f} \leftarrow \dfrac{\sum_{t=1}^{T}\mathbf{X}_{ft}}{\sum_{t=1}^{T}\mathrm{tr}(\mathbf{X}_{ft})}, \\ \mathbf{G}_{nf} \leftarrow \dfrac{1}{M}\mathbf{I}_M\ \ (n \geq 1). \end{cases} \tag{83}$$

A more sophisticated way of initialization is to use a fast speech enhancement method based on a complex Gaussian mixture model (cGMM) [34] that classifies each time-frequency bin into speech or noise. In this paper we initialize the cGMM with Eq. (83). Using the estimated posterior probability ω_{ft} that the bin at frequency f and time t was generated from the speech, we have

$$\begin{cases} \mathbf{G}_{0f} \leftarrow \dfrac{\sum_{t=1}^{T} \omega_{ft} \mathbf{X}_{ft}}{\sum_{t=1}^{T} \omega_{ft} \mathrm{tr}(\mathbf{X}_{ft})}, \\ \mathbf{G}_{nf} \leftarrow \dfrac{\sum_{t=1}^{T} (1-\omega_{ft}) \mathbf{X}_{ft}}{\sum_{t=1}^{T} (1-\omega_{ft}) \mathrm{tr}(\mathbf{X}_{ft})} \quad (n \geq 1). \end{cases} \tag{84}$$

2) ILRMA-DP: In the determined condition of the rank-1 model, $\mathbf{D}$ cannot be initialized in a way corresponding to Eq. (81) because the identity matrix is a full-rank matrix. A naive way of initialization that corresponds to Eq. (82) is to set $\mathbf{D}_f$ to the identity matrix as follows:

$$\mathbf{D}_f \leftarrow \mathbf{I}_{N+1}, \ \text{i.e.,} \ \mathbf{d}_{nf} \leftarrow \mathbf{e}_{n+1}. \tag{85}$$

The demixing matrices $\mathbf{D}$ can be initialized in an adaptive manner by using the observed data $\mathbf{X}$. If the mixing matrix $\mathbf{A}_f = [\mathbf{a}_{0f}, \mathbf{a}_{1f}, \ldots, \mathbf{a}_{Nf}]$ is given, $\mathbf{D}_f$ is given by

$$\mathbf{D}_f \leftarrow \mathbf{A}_f^{-1}, \tag{86}$$

where $\mathbf{A}_f$ can be estimated from the full-rank SCMs $\mathbf{G}$. Using $\mathbf{G}_{0f}$ in Eq. (83) and $\{\mathbf{G}_{nf}\}_{n=1}^{N}$ in Eq. (81), we have

$$\begin{cases} \mathbf{a}_{0f} = \mathrm{PE}\left(\sum_{t=1}^{T} \mathbf{X}_{ft}\right), \\ \mathbf{a}_{nf} = \mathbf{e}_{n+1} \quad (n \geq 1), \end{cases} \tag{87}$$

where $\mathrm{PE}(\cdot)$ indicates a normalized eigenvector that corresponds to the first principal component of a matrix. Alternatively, using Eq. (84), we have

$$\begin{cases} \mathbf{a}_{0f} = \mathrm{PE}\left(\sum_{t=1}^{T} \omega_{ft} \mathbf{X}_{ft}\right), \\ \mathbf{a}_{nf} = \mathrm{PE}\left(\sum_{t=1}^{T} (1-\omega_{ft}) \mathbf{X}_{ft}\right) \quad (n \geq 1). \end{cases} \tag{88}$$

V. Evaluation

This section reports experiments conducted for investigating the performance of our semi-supervised speech enhancement methods based on the MNMF-DP or ILRMA-DP with different configurations. First, we investigate the impacts of the model complexities (i.e., the number of noise sources N and the number of noise bases K) and verify the effectiveness of the low-rank noise model. We then evaluate the two methods used for optimizing the latent variables $\mathbf{Z}$ (i.e., Metropolis sampling and backpropagation methods described in Section IV-B3 and Section IV-C1) and the three methods used for initializing the spatial parameters $\mathbf{G}$ or $\mathbf{D}$ (i.e., identity-, observation-, and cGMM-based methods described in Section IV-D). Finally, we compare our method with the state-of-the-art unsupervised, semi-supervised, and supervised methods.[1,2]

[1]Demo: http://sap.ist.i.kyoto-u.ac.jp/members/sekiguch/demo/TASLP2019
[2]Code: https://github.com/sekiguchi92/TASLP2019

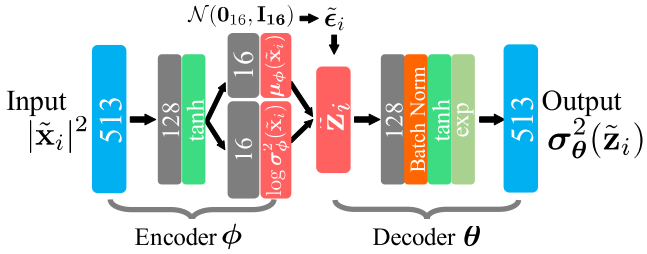


Fig. 4. The VAE for clean speech spectra.

A. Configurations

1) Test Data: The simulated data sampled at 16 kHz in the evaluation dataset of CHiME3 [35] were used for evaluation. This dataset contains 1320 noisy speech signals emulated to be uttered in four types of noisy environments: bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). We randomly chose 25 utterances for each environment (100 utterances in total). These simulated utterances were emulated to be recorded by using a tablet with 6 microphones. We selected five channels ($M = 5$) excluding the second channel because of its orientation on the back side of the tablet, in contrast to the other five microphones placed on the front side. We used short-time Fourier transform (STFT) with a shifting interval of 256 points and a window length of 1024 points ($F = 513$). The average number of time frames was $T = 379$.

2) Performance Measures: The performance of speech enhancement was measured in terms of the signal-to-distortion ratio (SDR) [36], [37]. For comparison with conventional methods, the perceptual evaluation of speech quality (PESQ) [38] and the short-time objective intelligibility (STOI) [39] were also calculated. The fifth channel of the enhanced speech spectra $\{\mathbf{x}_{ft0}^{\mathrm{FR/R1}}\}_{f=1,t=1}^{F,T}$ was compared with the ground-truth clean speech spectra because the fifth microphone was considered to be the closest to the mouth of a speaker.

3) Pretraining Configurations: The deep speech prior described in Section III-B1 was trained in advance from clean speech data in a variational autoencoding manner as described in Section IV-A. The VAE had an inference network (encoder) parameterized by $\boldsymbol{\phi}$ and a generation network (decoder) parameterized by $\boldsymbol{\theta}$, as shown in Fig. 4. The architecture of the VAE was similar to that proposed in [23]. The dimensions of the observed and latent spaces were $F = 513$ and $D = 16$, respectively. We used the WSJ-0 corpus [40] containing clean speech signals of about 15 hours. The speakers of the WSJ-0 corpus were disjoint with those of the test data. The power spectrogram of each utterance was scaled such that the average power was equal to a random number $\rho \sim \mathrm{Gamma}(2, 2)$.

4) Optimization Configurations: The number of iterations was set to 100. For MNMF-DP, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$ were updated simultaneously and $\mathbf{Z}$ was then updated in each iteration. For the ILRMA-DP, $\mathbf{W}$, $\mathbf{H}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{Z}$, and $\mathbf{D}$ were updated in this order. When the sampling method was used for optimizing $\mathbf{Z}$, the variance of the proposal distribution was set to $\xi = 10^{-4}$ and $\mathbf{Z}$ was sampled 50 times per iteration. When the backpropagation

TABLE I
THE AVERAGE SDRS [dB] FOR 100 NOISY SPEECH SIGNALS IN FOUR DIFFERENT ENVIRONMENTS

(a) MNMF-DP

# of noise sources N	Number of noise bases K										
	1	2	4	8	16	32	64	128	256	512	1024
1	16.4	17.1	17.5	17.9	18.1	18.4	18.6	18.7	18.7	18.7	18.7
2	17.3	17.7	18.0	18.2	18.6	18.7	18.8	18.8	18.8	18.8	18.8
3	17.6	18.0	18.3	18.5	18.6	18.7	18.7	18.7	18.8	18.7	18.7
4	17.9	18.1	18.2	18.5	18.6	18.6	18.6	18.6	18.7	18.6	18.6

(b) ILRMA-DP

# of noise sources N	Number of noise bases K										
	1	2	4	8	16	32	64	128	256	512	1024
4	16.2	16.3	16.2	16.2	16.0	15.8	15.7	15.6	15.5	15.3	15.1

TABLE II
THE ELAPSED TIMES [s] PER ITERATION FOR PROCESSING MULTICHANNEL NOISY SPEECH SIGNALS OF 2 [s]

(a) MNMF-DP

# of noise sources N	Number of noise bases K										
	1	2	4	8	16	32	64	128	256	512	1024
1	0.97	0.98	0.99	0.97	0.98	0.99	1.00	1.09	1.22	1.49	2.02
2	1.15	1.15	1.13	1.11	1.15	1.14	1.30	1.41	1.70	2.18	3.27
3	1.34	1.34	1.34	1.35	1.36	1.43	1.53	1.72	2.08	2.87	4.45
4	1.50	1.51	1.50	1.49	1.51	1.63	1.73	2.00	2.51	3.57	5.74

(b) ILRMA-DP

# of noise sources N	Number of noise bases K										
	1	2	4	8	16	32	64	128	256	512	1024
4	0.28	0.28	0.28	0.29	0.30	0.33	0.40	0.54	0.85	1.44	2.73

method was used, $\mathbf{Z}$ was updated 50 times per iteration by using the Adam optimizer [41] with a learning rate of 0.001.

B. Evaluation of Model Complexities

We investigated the best model complexities of MNMF-DP and ILRMA-DP by changing the number of noise sources N and the number of noise basis spectra K.

1) Experimental Conditions: For MNMF-DP, we tested all possible combinations of $K = 2^l$ $(l = 0, \ldots, 10)$ and $N = 1, 2, 3, 4$. For ILRMA-DP, we changed only K because $N = 4$ must hold under a determined condition with $M = 5$ (one speech source and four noise sources). We used the sampling method for optimizing the latent variables $\mathbf{Z}$ and the observation-based method given by Eq. (83) or Eq. (87) for initializing the spatial parameters $\mathbf{G}$ or $\mathbf{D}$, respectively.

2) Experimental Results: Table I-(a) shows the average SDRs over the 100 utterances obtained by MNMF-DP. The average SDR of the input noisy signals (the fifth channel) was 7.5 dB. Regardless of the choice of N, the performance converged to around 18.7 dB as K increased. This might be because most noise sources in the test dataset were diffusive. If there are multiple directional noise sources, it would be necessary to carefully choose N. Note that when $K \geq T$, the low-rank assumption on the PSDs of noise is considered to have no effect in theory because the noise model is capable of perfectly fitting any PSDs. In reality, the performance was not degraded even when $K = 1024$. This result raised a question whether the low-rank assumption, which is useful in MNMF, is still necessary in the proposed model. To answer this question, the effectiveness of the low-rank assumption was verified in Section V-C. Table II-(a) shows the elapsed times per iteration for processing noisy speech signals of 2 [s] on a workstation with Intel Xeon W-2145 (3.70 GHz). Considering both the performance and the computational cost, the combination of $N = 1$ and $K = 64$ can be regarded as best.

Table I-(b) shows the average SDRs obtained by ILRMA-DP. The performance was maximized when $K = 2$ and it monotonically decreased as K increased. Because the rank-1 spatial model is incapable of precisely representing realistic sound propagation processes, the source models (speech and noise models) play an influential role for speech enhancement. In each iteration, the noise model fits the current estimate of the noise spectra $\{|s_{ft1}|^2\}_{f=1,t=1}^{F,T}$ given by Eq. (11) using the demixing matrices $\mathbf{D}$. The noise model based on NMF with large K overfit the imperfect estimate of the noise spectra in a few iterations before $\mathbf{D}$ was fully optimized. When the noise model with $K = 256$ was updated once per four iterations, the average SDR was improved to 16.1 dB.

C. Evaluation of Low-Rank Modeling

We investigated the effectiveness of the low-rank assumption on the noise PSDs. The sampling method was used for optimizing the latent variables $\mathbf{Z}$.

1) Experimental Conditions: We tested three variants of the noise model in MNMF-DP with $N = 1$.

1) High-rank model: $K = T$. $\mathbf{W}$ and $\mathbf{H}$ were initialized by using Eq. (78) and Eq. (79) and then iteratively updated by using Eq. (52) and Eq. (55).
2) 1-on-1 model: This model was the same as the high-rank model except that $\mathbf{H}$ was initialized as follows:

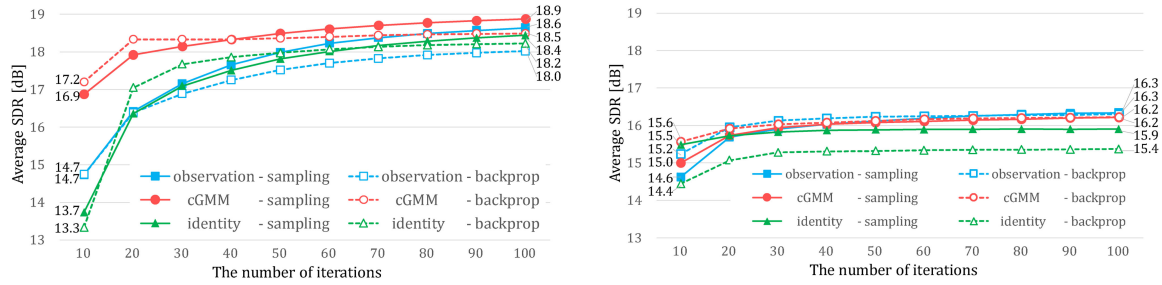
$$\begin{cases} h_{1kt} \sim \mathrm{Gamma}\left(\alpha_0, \frac{\alpha_0}{\mathbb{E}_{\mathrm{emp}}[|x|^2]} \frac{1}{FM}\right) & (k = t), \\ h_{1kt} = 0 & (k \neq t). \end{cases} \tag{89}$$

   Since $h_{1kt} = 0$ $(k \neq t)$ was kept in Eq. (55), the K bases correspond to the T frames one by one.
3) Non-factorized model: The NMF-based noise model was removed from the proposed model, i.e., the noise PSDs $\{\lambda_{ft1}\}_{f=1,t=1}^{F,T}$ in Eq. (15) were directly estimated. An updating rule can be obtained as follows:

$$\lambda_{ft1} \leftarrow \lambda_{ft1} \sqrt{\frac{\mathrm{tr}\left(\mathbf{G}_{1f} \mathbf{Y}_{ft}^{-1} \mathbf{X}_{ft} \mathbf{Y}_{ft}^{-1}\right)}{\mathrm{tr}\left(\mathbf{G}_{1f} \mathbf{Y}_{ft}^{-1}\right)}}. \tag{90}$$

   λ_{ft1} was initialized as $\lambda_{ft1} = w_{1tf} h_{1tt}$, where $\mathbf{W}$ and $\mathbf{H}$ were initialized as in the 1-on-1 model.

2) Experimental Results: Table III shows the average SDRs and log-likelihoods obtained by the three models. While the 1-on-1 model and the non-factorized model were better than



(a) MNMF-DP (b) ILRMA-DP

Fig. 5. The evolutions of average SDRs [dB] over iterations. The dotted lines indicate the SDRs obtained by the backpropagation method and the solid lines indicate the SDRs obtained by the sampling method.

TABLE III
THE AVERAGE SDRS [dB] AND LOG-LIKELIHOODS OBTAINED BY THE THREE VARIANTS OF MNMF-DP

Noise model	High-rank	1-on-1	Non-factorized
SDR [dB]	18.7	15.8	16.2
Log-likelihood	1.64×10^6	1.67×10^6	1.67×10^6

the high-rank model in terms of the log-likelihood, the high-rank model attained the best SDR. Since the architecture of the high-rank model was the same as that of the 1-on-1 model, the high-rank model was considered to get stuck in local optima in which the noise PSDs were approximated as low-rank matrices consisting of a fewer number of bases. This indicates that when $K \geq T$ in Table I-(a), the low-rank constraint on the noise PSDs was still effective. Comparing the SDR (16.2 dB) obtained by the non-factorized model with that (18.6 dB) obtained by the best MNMF-DP with $N = 1$ and $K = 64$, the low-rank modeling can be said to be effective.

D. Evaluation of Optimization and Initialization Methods

We investigated the initialization sensitivity and optimization difficulty of MNMF-DP and ILRMA-DP.

1) Experimental Conditions: Considering Tables I and II, we used MNMF-DP with $N = 1$ and $K = 64$ and ILRMA-DP with $K = 2$ as the best performing models with the reasonable computational costs. To optimize $\mathbf{Z}$ in MNMF-DP, the sampling or backpropagation method was used (Section IV-B3). To initialize $\mathbf{G}$, the identity-, observation-, or cGMM-based method given by Eqs. (81), (83), or (84), respectively, was used (Section IV-D1). To optimize $\mathbf{Z}$ in ILRMA-DP, on the other hand, the sampling or backpropagation method was used (Section IV-C1). To initialize $\mathbf{D}$, the identity-, observation-, or cGMM-based method given by Eqs. (86), (87), or (88) was used (Section IV-D2). In total, we tested six configurations for each model.

2) Experimental Results: Fig. 5-(a) shows the SDR evolutions over iterations obtained by the six configurations of MNMF-DP. The combination of the sampling-based optimization and the cGMM-based initialization attained the best SDR of 18.9 dB. Regardless of an initialization method, the sampling method was slightly better than the backpropagation method in terms of the performance obtained after sufficiently many iterations. The backpropagation method converged faster to the affordable performance than the sampling method. When the same optimization method was used, the performance difference between the initialization methods was smaller than 0.5 dB. This indicates that our MNMF-DP is insensitive to the initialization of $\mathbf{G}$ because the deep speech prior plays an influential role even before $\mathbf{G}$ is not fully optimized. In fact, our model can work even in a single-channel scenario without spatial information [22]. This is a noticeable advantage of the proposed method over MNMF that heavily relies on $\mathbf{G}$ for speech enhancement.

Fig. 5-(b) shows the SDR evolutions over iterations obtained by the six configurations of ILRMA-DP. The use of the observation- or cGMM- based initialization method reached the SDR of 16.3 dB or 16.2 dB, respectively, regardless of the optimization strategy. The backpropagation method tended to converge faster than the sampling method. When the identity-based initialization method was used, the backpropagation method underperformed the sampling method. The initial values of $\mathbf{Z}$ given by the encoder ϕ of the VAE were considered to be close to optimal values. In the backpropagation method, however, $\mathbf{Z}$ was quickly adapted to the inaccurate estimate of the speech spectra $\mathbf{s}_{ft}$ given by Eq. (11) before the demixing matrices $\mathbf{D}$ were fully optimized.

E. Key Findings

Considering the experimental results shown in Section V-B and Section V-D, we summarize recommended configurations. In general, it is recommended to use the full-rank model with $N = 1$, $K = 64$, the observation-based initialization method given by Eq. (83), and the sampling-based optimization method. To squeeze the performance and accelerate the convergence in exchange of the additional implementation cost, one can use the cGMM-based initialization method given by Eq. (84) instead of the observation-based initialization method. If the computational cost is a main concern, one may use ILRMA-DP with $K = 2$, the observation-based initialization method given by Eq. (87), and the backpropagation-based optimization method.

F. Comparison with the State-of-the-Art Methods

We compared the proposed semi-supervised method with the state-of-the-art unsupervised, semi-supervised, and supervised methods in terms of the SDR, PESQ, and STOI.

1) Experimental Conditions: We used MNMF-DP with $N = 1$ and $K = 64$ initialized by the cGMM-based method given by Eq. (84) and ILRMA-DP with $N = 4$ and $K = 2$ initialized by the observation-based method given by Eq. (87). The sampling method was used for estimating $\mathbf{Z}$.

- Unsupervised methods: We tested MNMF [9], ILRMA [11], and cGMM [34]. MNMF and ILRMA had the same architectures as the proposed MNMF-DP and ILRMA-DP, respectively, except that an NMF-based low-rank model was used for speech instead of the DNN-based model. The number of noise sources N, that of speech bases K_s, that of noise bases K_n, and the initialization strategy were experimentally optimized. We used MNMF with $N = 1$, $K_s = 8$, and $K_n = 256$ initialized by the cGMM-based method and ILRMA with $N = 4$, $K_s = 8$, and $K_n = 1$ initialized by the observation-based method. We also tested a weighted delay-and-sum (DS) beamforming called *beamformit* [42] and found that the average SDR of the enhanced speech was 6.3 dB.
- Semi-supervised methods: For fair comparison with the proposed semi-supervised method, we also tested semi-supervised versions of MNMF and ILRMA. K_s speech bases were estimated in advance by using NMF [43] or vector quantization (VQ) [44] based on the Itakura-Saito (IS) divergence (called IS-NMF and IS-VQ, respectively) for the clean speech data of the WSJ-0 corpus [40]. IS-VQ iterated two steps; 1) given codebooks (bases), each speech spectrum in the training dataset were clustered into the nearest codebook based on the IS divergence, and 2) each codebook was updated to the average of the spectra assigned to the codebook. In the speech enhancement phase, while the speech bases were fixed, the other parameters were updated as in the proposed method. We conducted a preliminary experiment using $K_s = 2^l$ $(l = 0, \ldots, 8)$ and decided to use MNMF based on IS-NMF (MNMF-NMF) with $N = 1$, $K_s = 4$, and $K_n = 256$ and MNMF based on IS-VQ (MNMF-VQ) with $N = 1$, $K_s = 8$, and $K_n = 256$ initialized by the cGMM-based method, and ILRMA based on IS-NMF (ILRMA-NMF) with $N = 4$, $K_s = 16$, and $K_n = 1$ and ILRMA based on IS-VQ (ILRMA-VQ) with $N = 4$, $K_s = 256$, and $K_n = 2$ initialized by the observation-based method.
- Supervised method: We tested a DNN-based beamforming method. To estimate speech masks $\{\omega_{ft}\}_{f=1,t=1}^{F,T}$, a feed-forward DNN was trained by using the training dataset of CHiME3 that contains pairs of multichannel noisy speech signals and ground-truth clean speech signals. We extracted three kinds of acoustic features as the input to the DNN. At each time t, the log of the outputs of 100-channel mel-scale filter banks (LMFBs) was computed from the magnitude spectrogram of the fifth channel and LMFBs were stacked over 11 frames from time $t-5$ to $t+5$. The $(M-1)$-dimensional inter-channel level and phase differences (ILDs and IPDs) were also calculated at each time t as proposed in [45]. The DNN was trained such that the cross-entropy loss between ideal binary masks and estimated masks was minimized. To use the minimum

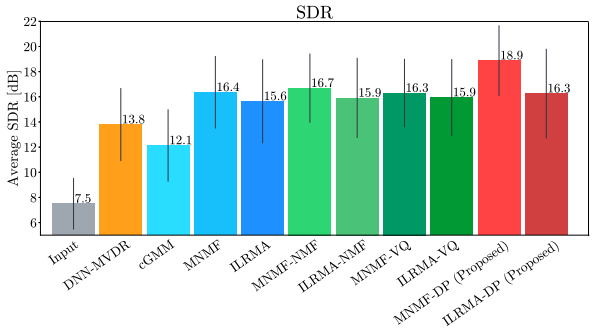


Fig. 6. The average SDRs obtained by the 11 methods.

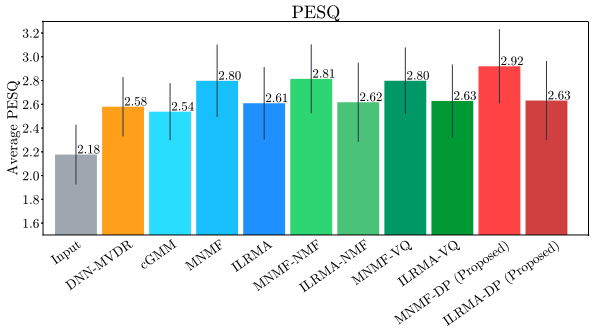


Fig. 7. The average PESQs obtained by the 11 methods.

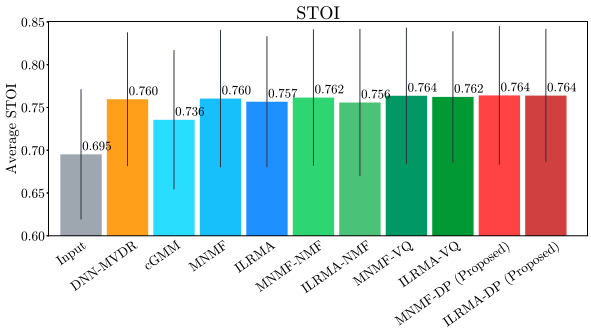


Fig. 8. The average STOIs obtained by the 11 methods.

variance distortionless response (MVDR) beamforming [46], the steering vector of speech $\mathbf{a}_{0f}$ and the SCM of noise $\mathbf{G}_{1f}$ at frequency f are given by

$$\begin{cases} \mathbf{a}_{0f} = \mathrm{PE}\left(\sum_{t=1}^{T} \omega_{ft}\mathbf{X}_{ft}\right), \\ \mathbf{G}_{1f} = \sum_{t=1}^{T}(1-\omega_{ft})\mathbf{X}_{ft}, \end{cases} \tag{91}$$

where $\mathrm{PE}(\cdot)$ indicates a normalized eigenvector that corresponds to the first principal component of a matrix. The demixing filter $\mathbf{d}_{0f}$ at frequency f is given by

$$\mathbf{d}_{0f} = \frac{\mathbf{G}_{1f}^{-1}\mathbf{a}_{0f}}{\mathbf{a}_{0f}^{\mathrm{H}}\mathbf{G}_{1f}^{-1}\mathbf{a}_{0f}}. \tag{92}$$

The image of clean speech was estimated as follows:

$$\mathbf{x}_{ft0}^{\mathrm{MVDR}} = \mathbf{a}_{0f}s_{ft0} = \mathbf{a}_{0f}\mathbf{d}_{0f}^{\mathrm{H}}\mathbf{x}_{ft}. \tag{93}$$

2) Experimental Results: Figs. 6, 7, and 8 show the average SDRs, PESQs, and STOIs, respectively. MNMF-DP

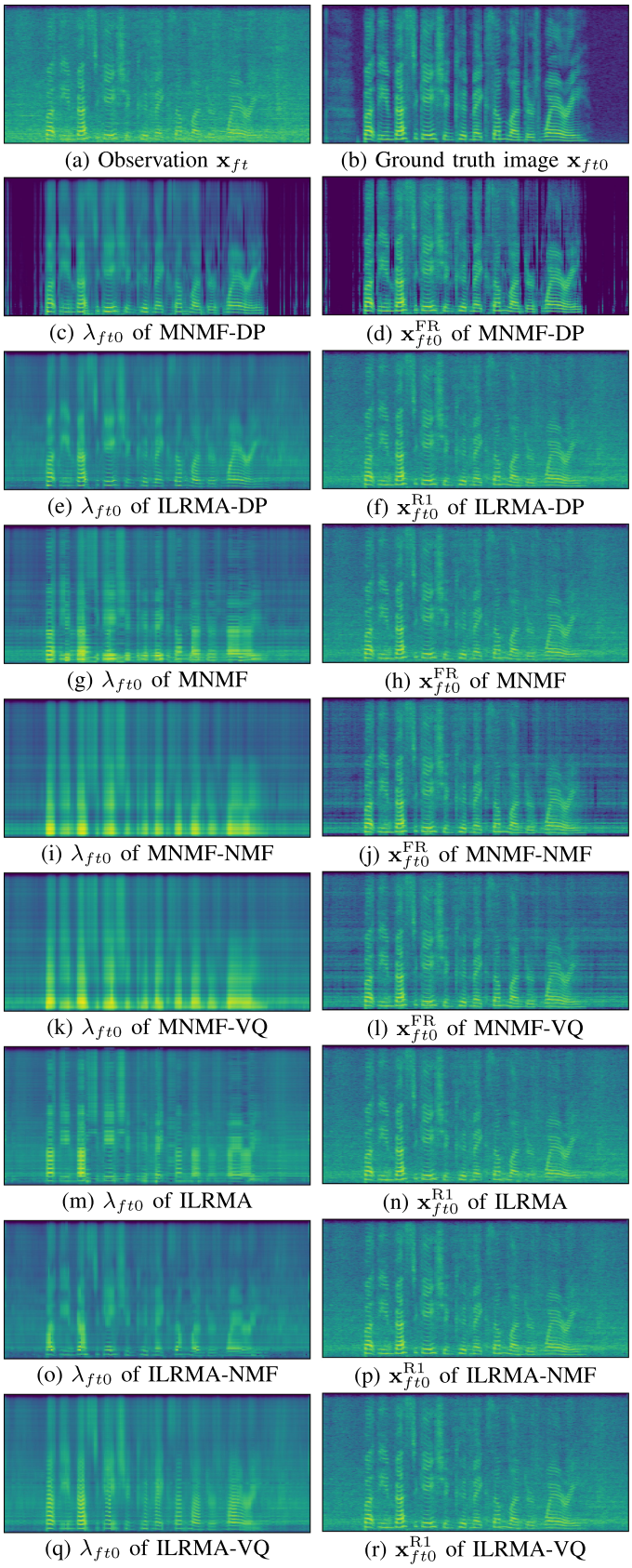


Fig. 9. Comparison of speech enhancement methods. The observation, ground truth, and separated speech show the 5th channel only.

performed best in all measures. Although, MNMF is generally known to often underperform ILRMA because of the strong initialization sensitivity [11], in this experiment MNMF (16.4 dB) outperformed ILRMA (15.6 dB) because the cGMM-based method given by Eq. (84) provided a good initial estimate of $\mathbf{G}$. When the observation-based method given by Eq. (83) was used for MNMF as in ILRMA, the SDR was drastically degraded (12.6 dB). Fig. 9 shows examples of the noisy spectra $\{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T}$ in the BUS environment, the ground-truth speech image $\{\mathbf{x}_{ft0}\}_{f=1,t=1}^{F,T}$, the estimated speech PSDs $\{\lambda_{ft0}\}_{f=1,t=1}^{F,T}$ and the separated speech spectra $\{\mathbf{x}_{ft0}^{\mathrm{FR/R1}}\}_{f=1,t=1}^{F,T}$ obtained by MNMF-DP (19.2 dB), ILRMA-DP (14.9 dB), MNMF (15.3 dB), ILRMA (13.4 dB), MNMF-NMF (16.1 dB), ILRMA-NMF (14.7 dB), MNMF-VQ (15.8 dB), and ILRMA-VQ (14.6 dB). This clearly showed that the deep speech prior is better at representing the characteristic structures of speech PSDs than the NMF-based low-rank model. In semi-supervised MNMF and ILRMA, the numbers of speech bases were determined to maximize the SDRs, but were too low to precisely represent speech PSDs.

We also compared the proposed method with its original single-channel version [22]. The SDR of the optimally-tuned single-channel method was 11.9 dB. This indicates that the proposed MNMF-DP and ILRMA-DP successfully utilize the spatial information. Comparing ILRMA-DP with MNMF-DP, however, while MNMF-DP successfully suppressed the noise components, the enhanced speech spectra obtained by ILRMA-DP as well as those obtained by ILRMA were still noisy. This indicates that the idealized rank-1 spatial model based on the time-invariant demixing matrices has a performance limitation in speech enhancement.

VI. Conclusion

This paper presented a semi-supervised multichannel speech enhancement method that integrates a DNN-based generative model of speech spectra, an NMF-based generative model of noise spectra, and a full-rank or rank-1 spatial model in a unified probabilistic model. The full-rank and rank-1 versions of the proposed method, called MNMF-DP and ILRMA-DP, are extensions of MNMF [9] and ILRMA [11], respectively, i.e., an NMF-based model for one of sources is replaced with the deep speech prior capable of precisely representing the PSDs of clean speech. An advantage of our method is that only clean speech data are used for training the deep speech prior. The speech prior can generalize well to unseen speech spectra and the low-rank noise model and the spatial model can adapt to unseen acoustic environments. We showed that MNMF-DP significantly outperformed the rank-1 counterpart, the unsupervised and semi-supervised versions of MNMF and ILRMA, and the supervised DNN-based beamforming method in terms of the SDR, PESQ, and STOI. We also showed that MNMF-DP is less sensitive to initialization and is less likely to get stuck in local optima than MNMF.

Future work includes the online extension of the full-rank model for real-time speech enhancement. We also plan to extend the proposed method to jointly perform dereverberation and source localization.

REFERENCES

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[2] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7092–7096.

[3] X. Li, J. Li, and Y. Yan, "Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions," in *Proc. Interspeech*, 2017, pp. 1203–1207.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.

[5] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[6] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 5, pp. 960–971, May 2019.

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

[8] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. Int. Conf. Inf. Sci., Signal Process. and their Applicat.*, 2010, pp. 1–4.

[9] 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.

[10] J. Nikunen and T. Virtanen, "Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6677–6681.

[11] 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.

[12] K. Itakura, Y. Bando, E. Nakamura, K. Itoyama, K. Yoshii, and T. Kawahara, "Bayesian multichannel audio source separation based on integrated source and spatial models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 831–846, Apr. 2018.

[13] 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.

[14] S. Mogami *et al.*, "Independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1557–1561.

[15] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. APSIPA*, 2018, pp. 1233–1239.

[16] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[17] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4029–4032.

[18] D. D. Lee and S. H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[20] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, 2017, pp. 2008–2012.

[21] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5024–5028.

[22] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 716–720.

[23] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE Int. Workshop Machine Learn. Signal Process.*, 2018, pp. 1–6.

[24] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 101–105.

[25] 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.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[27] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. New York, NY, USA: Springer, 2005.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[29] K. Yoshii, K. Kitamura, Y. Bando, E. Nakamura, and T. Kawahara, "Independent low-rank tensor analysis for audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1671–1675.

[30] K. Yoshii, "Correlated tensor factorization for audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 731–735.

[31] T. Ando, C.-K. Li, and R. Mathias, "Geometric means," *Linear Algebra Appl.*, vol. 385, pp. 305–334, 2004.

[32] W.-H. Chen, "A review of geometric mean of positive definite matrices," *Brit. J. Math. Comput. Sci.*, vol. 5, no. 1, pp. 1–12, 2015.

[33] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2011, pp. 189–192.

[34] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2014, pp. 268–272.

[35] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Comput. Speech Lang.*, vol. 46, pp. 605–626, 2017.

[36] 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.

[37] C. Raffel *et al.*, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 367–372.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[40] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[41] D. P. Kingma and J. Lei Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[42] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[43] 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.

[44] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[45] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

[46] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

Kouhei Sekiguchi received the B.E. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2015 and 2017, respectively. He is currently a researcher at the Center for Advanced Intelligence Project (AIP), RIKEN, Japan, and working toward the Ph.D. degree in Kyoto University. His research interests include microphone array signal processing and machine learning. He is a member of IEEE and IPSJ.

Yoshiaki Bando received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2015 and 2018, respectively. He is currently a Researcher at Artificial Intelligence Research Center (AIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. His research interests include microphone array signal processing, deep Bayesian learning, and robot audition. He is a member of IEEE, RSJ, and IPSJ.

Aditya Arie Nugraha received the B.S. and M.S. degrees in electrical engineering from Institut Teknologi Bandung, Indonesia, in 2008 and 2011, respectively, the M.E. degree in computer science and engineering from Toyohashi University of Technology, Japan, in 2013, and the Ph.D. degree in informatics from Université de Lorraine and Inria Nancy–Grand-Est, France, in 2017. He was a research engineer at Inria Nancy–Grand-Est in 2018. He is currently a Postdoctoral Researcher at the Center for Advanced Intelligence Project (AIP), RIKEN, Japan. His research interests include source separation and machine learning.

Kazuyoshi Yoshii received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He is an Associate Professor at the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan. His research interests include music informatics, audio signal processing, and statistical machine learning.

Tatsuya Kawahara received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Japan, in 1987, 1989, and 1995, respectively.

He is currently a Professor at the School of Informatics, Kyoto University. He has published more than 400 papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet). He received the Commendation for Science and Technology by MEXT, Japan, in 2012. He was an editorial board member of *Elsevier Journal of Computer Speech and Language* and IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is the Editor-in-Chief of *APSIPA Transactions on Signal and Information Processing*. He is a board member of APSIPA and ISCA, and a Fellow of IEEE.