Bayesian Multichannel Speech Enhancement with a Deep Speech Prior

Kouhei Sekiguchi*†, Yoshiaki Bando‡, Kazuyoshi Yoshii*†, and Tatsuya Kawahara†
*RIKEN AIP †Kyoto University ‡National Institute of Advanced Industrial Science and Technology (AIST), Japan

***Abstract*—This paper describes statistical multichannel speech enhancement based on a deep generative model of speech spectra. Recently, deep neural networks (DNNs) have widely been used for converting noisy speech spectra to clean speech spectra or estimating time-frequency masks. Such a supervised approach, however, requires a sufficient amount of training data (pairs of noisy speech data and clean speech data) and often fails in an unseen noisy environment. This calls for a blind source separation method called multichannel nonnegative matrix factorization (MNMF) that can jointly estimate low-rank source spectra and spatial covariances on the fly. However, the assumption of low-rankness does not hold true for speech spectra. To solve these problems, we propose a semi-supervised method based on an extension of MNMF that consists of a deep generative model for speech spectra and a standard low-rank model for noise spectra. The speech model can be trained in advance with auto-encoding variational Bayes (AEVB) by using only clean speech data and is used as a prior of clean speech spectra for speech enhancement. Given noisy speech spectrogram, we estimate the posterior of clean speech spectra while estimating the noise model on the fly. Such adaptive estimation is achieved by using Gibbs sampling in a unified Bayesian framework. The experimental results showed the potential of the proposed method.**

I. Introduction

Speech enhancement forms the basis of automatic speech recognition in a noisy environment. Several methods have been proposed for single-channel speech enhancement. Robust principal component analysis (RPCA), for example, is used for decomposing the spectrogram of an input noisy speech signal into a sparse spectrogram corresponding to speech and a low-rank spectrogram corresponding to noise in an unsupervised manner [1]. Nonnegative matrix factorization (NMF) can be used for supervised speech enhancement [2]. In the training phase, typical spectra of speech and/or noise are learned, and in the denoising phase, an observed spectrogram is approximated by the weighted sums of the learned spectra.

Multichannel extensions of NMF (MNMF) have been developed to deal with spatial information related to a sound propagation process [3]–[6]. The power spectrogram of each source signal is given by the sum of products of basis spectra and their activations. The complex spectrograms of observed multichannel signals are given by the sum of spectrograms of propagated source signals. Ozerov *et al.* [3] pioneered the use of NMF for multichannel source separation, where the spatial covariance matrices were restricted to rank-1 matrices and the EM or multiplicative update algorithm was used for minimizing the cost function based on the Itakura-Saito (IS) divergence. This model was extended to have full-rank spatial covariance matrices [7]. Sawada *et al.* [4] introduced partitioning parameters to have a set of basis spectra shared by all sources and derived a majorization-minimization (MM) algorithm. Nikunen and Virtanen [8] proposed a similar model that represents the spatial covariance matrix of each source as the weighted sum of all possible direction-dependent covariance matrices and used the MM algorithm for minimizing the cost function based on the Euclidean distance. Kitamura *et al.* [5] modified the model in [4] by restricting spatial covariance matrices to rank-1 matrices, resulting in a unified model of NMF and independent vector analysis (IVA). Itakura *et al.* [6] proposed a Bayesian extension of MNMF. In application for speech enhancement, however, a potential problem common in these approaches is that the basic assumption of low-rankness does not hold true for speech spectra.

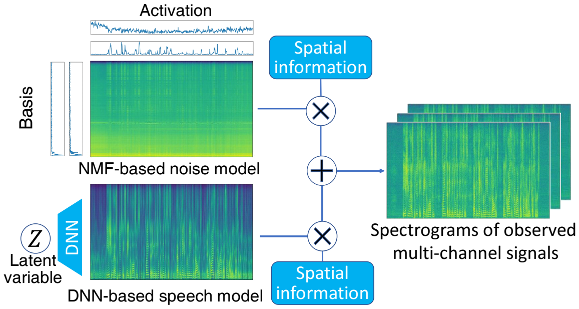


Fig. 1: A generative model of multichannel noisy spectra with a deep speech prior.

Deep neural networks (DNNs) have widely been used for supervised speech enhancement. In a single-channel scenario, one can use a denoising autoencoder (DAE) that takes as input noisy speech spectra and directly outputs clean speech spectra, which is trained with pair data [9]. Alternatively, one can train a DNN that outputs a time-frequency mask [10]. Multichannel extensions have been investigated [11]–[14]. For example, a time-frequency mask is estimated using a long short-term memory (LSTM) and then the estimated mask is used for calculating the steering vectors and spatial covariance matrices of speech and noise used for beamforming [11], [12].

Recently, DAEs have been used for improving multichannel source separation methods that iteratively and alternately optimize the power spectrum densities and spatial covariance matrices of individual sound sources [13], [14]. In the optimization process, the current estimates of the power spectrum densities are refined by using a DAE. Although this approach is experimentally found to work well, it requires pairs of noisy and clean speech for training DAEs and can be unstable in an unseen noisy environment. In addition, the denoising step is not properly derived based on statistical inference.

In this paper we propose a multichannel speech enhancement method that integrates a deep generative model of speech spectra and an NMF-based generative model of noise spectra to formulate a unified probabilistic model of multichannel noisy spectra (Fig. 1). The speech model is a deep latent variable model, which can be *pre-trained* by using only clean speech data and is expected to learn speech characteristics such as fundamental frequencies (F0s) and spectral envelopes in a latent space. Since noise varies depending on the environment, on the other hand, the noise model is *learned on the fly* in an unsupervised manner. Using both models, we can estimate the power spectrum densities and spatial covariance matrices of speech and noise by using Gibbs sampling. This model is a multichannel extension of [15]. The use of spatial information is expected to improve the performance.

The main contribution of the paper is that we first propose a genuine probabilistic generative model of multichannel spectra that involves a pre-trained deep generative model as a speech prior. This enables us to infer all random variables in a Bayesian manner. Since the proposed method uses only clean speech data, it is robust to an unseen noisy environment.

II. Variational Autoencoder

We here review variational autoencoder (VAE) [16]. A VAE is used to estimate a generative process, which generates data from a latent variable, and to estimate the variational posterior of the latent variable, which approximates the true posterior. The VAE is based on the assumption that the data $\boldsymbol{x}$ is generated from a distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, where θ represents the parameters of a DNN and $\boldsymbol{z} \in \mathbb{R}^D$ is a latent variable that is the input to the DNN. The latent variable is often assumed to be generated from the standard Gaussian distribution.

Since in reality the latent variable is unknown, it is impossible to directly estimate the parameter θ. It is also difficult to calculate the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ because the marginal likelihood $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ is intractable. The VAE solves these problems by approximating the true posterior by a variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ given by a DNN with parameters ϕ and input $\boldsymbol{x}$. In the VAE framework, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are called an encoder and a decoder, respectively, and the parameters θ and ϕ are optimized together.

The VAE estimates θ and ϕ so that the log marginal likelihood $\log p_\theta(\boldsymbol{x})$ is maximized. The log marginal likelihood can be rearranged as follows:

$$\log p_\theta(\boldsymbol{x}) \geq -\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) + \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \quad (1)$$

$$\overset{\text{def}}{=} \mathcal{L}(\theta,\phi), \quad (2)$$

where $\mathrm{KL}(q\|p)$ (≥ 0) indicates the Kullback-Leibler (KL) divergence. The VAE tries to find a local maximum of $\log p_\theta(\boldsymbol{x})$ by maximizing the lower bound $\mathcal{L}(\theta,\phi)$. The partial derivatives of $\mathcal{L}(\theta,\phi)$ w.r.t. θ and ϕ are thus necessary. Since the partial derivatives of $\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ cannot be calculated analytically, the VAE approximates $\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ by using a reparameterization trick as follows:

$$\mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] = \mathbb{E}_{p(\epsilon)}[\log p_\theta(\boldsymbol{x}|\tilde{\boldsymbol{z}})] \quad (3)$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}|\tilde{\boldsymbol{z}}^{(l)}), \quad (4)$$

where $\tilde{\boldsymbol{z}}^{(l)} = g_\phi(\boldsymbol{\epsilon}^{(l)})$, $\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$, and L is the number of samples. The $p(\boldsymbol{\epsilon})$ and $g_\phi(\boldsymbol{\epsilon})$ are chosen in accordance with the variational posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. When $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_{d=1}^{D} \mathcal{N}(z_d \mid \mu_{\phi,d}(\boldsymbol{x}), \sigma^2_{\phi,d}(\boldsymbol{x}))$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ and $g_\phi(\boldsymbol{\epsilon}) = \boldsymbol{\mu}_\phi(\boldsymbol{x}) + \boldsymbol{\epsilon} \odot \sqrt{\boldsymbol{\sigma}^2_\phi(\boldsymbol{x})}$, where $\mathcal{N}(\mu, \sigma^2)$ indicates a Gaussian distribution with mean μ and variance σ^2, and $\odot$ indicates an element-wise product. Thus, the derivative of $\tilde{\mathcal{L}}(\theta,\phi) = -\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) + 1/L\sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}|\tilde{\boldsymbol{z}}^{(l)})$ is tractable, and the VAE uses $-\tilde{\mathcal{L}}(\theta,\phi)$ as the cost function for training the DNNs.

III. Proposed Method

We explain the proposed method that integrates a DNN-based generative model of speech spectra and an NMF-based generative model of noise spectra in a unified probabilisic model. First, we formulate the generative process of multichannel observed signals, which consists of a spatial model and speech and noise models. Next, we explain how to estimate the speech and noise spectra in a Bayesian manner.

A. Problem Specification

In this paper, we assume that the observed spectra $\boldsymbol{X}$ contain dominant target speech and additional noise. Let F, T, and M be the number of frequency bins, time frames, and microphones, respectively. The index of a sound source is denoted by i. The observed spectra $\boldsymbol{x}_{ft}$ and the source signals $\boldsymbol{s}_{ft}$ at time frame t and frequency bin f are defined as follows:

$$\boldsymbol{x}_{ft} = [x_{ft1}, \cdots, x_{ftM}] \in \mathbb{C}^M, \quad (5)$$

$$\boldsymbol{s}_{ft} = [s_{ft1}, s_{ft2}] \in \mathbb{C}^2, \quad (6)$$

where the first sound source s_{ft1} is the target speech and the second source s_{ft2} is the noise.

B. Spatial Modeling

A spatial model represents the sound propagation process between sound sources and microphones. First, we assume that each time-frequency bin of the source i, s_{fti}, follows a complex Gaussian distribution as follows:

$$s_{fti} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{fti}), \quad (7)$$

where $\mathcal{N}_{\mathbb{C}}(\mu, \sigma^2)$ indicates a complex Gaussian distribution with mean μ and variance σ^2, and λ_{fti} indicates the power of the sound source i at time frame t and frequency bin f. Next, under the assumption of a time-invariant linear system, the relationship between the observed signals and source signals is given by

$$\boldsymbol{x}^{(i)}_{ft} = \boldsymbol{a}_{fi} s_{fti}, \quad (8)$$

where $\boldsymbol{a}_{fi} \in \mathbb{C}^M$ is a steering vector of the source i at frequency bin f. Since Eq. (8) is a linear transformation of the source signals, each time-frequency bin of the observed signals $\boldsymbol{x}_{ft}$ also follows a complex Gaussian distribution as

$$\boldsymbol{x}_{ft}^{(i)} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \lambda_{fti}\boldsymbol{A}_{fi}\right), \tag{9}$$

where $\boldsymbol{A}_{fi} = \boldsymbol{a}_{fi}\boldsymbol{a}_{fi}^H \in \mathbb{C}^{M\times M}$ is a spatial covariance matrix. In this case, there are two sound sources, and the observed signals are given by

$$\boldsymbol{x}_{ft} = \sum_{i=1}^{2} \boldsymbol{x}_{ft}^{(i)} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sum_{i=1}^{2} \lambda_{fti}\boldsymbol{A}_{fi}\right). \tag{10}$$

In a real noisy environment, the spatial covariance matrix can be a full-rank matrix due to reverberation etc. We represent a full-rank spatial matrix as $\boldsymbol{G}_{fi}$ to distinguish it from the rank-1 matrix $\boldsymbol{A}_{fi}$.

C. Source Modeling

A source model represents the generative process of the power spectrogram of each source. We use different kinds of source models depending on the source properties. Assuming that noise has low-rankness, we use NMF as a noise model. Since an assumption of low-rankness is not suitable for speech, we assume that the power spectrogram of speech is generated from a DNN.

1) DNN-based generative model of speech: We assume that the power spectrum densities of the target speech at time frame t is generated from a DNN, the input of which is a latent variable $\boldsymbol{z}_t \in \mathbb{R}^D$. The latent variable is assumed to be generated from a standard Gaussian distribution. The distribution of the speech spectra is represented as follows:

$$p_\theta(\boldsymbol{s}_{t1}|\boldsymbol{z}_t) = \prod_{f=1}^{F} \mathcal{N}_{\mathbb{C}}(s_{ft1}\,|\,0, \lambda_{ft1}), \tag{11}$$

$$\lambda_{ft1} = \sigma_{\theta,f}^2(\boldsymbol{z}_t), \tag{12}$$

$$p(\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{0}, I_D), \tag{13}$$

where $\boldsymbol{s}_{t1} = [s_{1t1}, \cdots, s_{Ft1}] \in \mathbb{R}^F$, $\boldsymbol{\sigma}_\theta^2(\,\cdot\,) : \mathbb{R}^D \to \mathbb{R}_+^F$ is a non-linear function given by a DNN, and $\sigma_{\theta,f}^2(\,\cdot\,)$ is a f-th element of the $\boldsymbol{\sigma}_\theta^2(\,\cdot\,)$. Since all the frequency bins of the speech spectrum at time frame t are generated from the same latent variable, this model can be considered to capture the dependency between frequency bins. Since it is impossible to obtain the ground-truth values of latent variables $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_T] \in \mathbb{R}^{D\times T}$, the VAE framework described in Section II is used to train the DNN.

2) NMF-based generative model of noise: NMF approximates the power spectrogram with the product of a basis matrix $\boldsymbol{W} = (w_{fk}) \in \mathbb{R}^{F\times K}$ and an activation matrix $\boldsymbol{H} = (h_{kt}) \in \mathbb{R}^{K\times T}$, where K denotes the number of bases. A component w_{fk} indicates the magnitude of the k-th basis at frequency bin f, and h_{kt} indicates the activation of the k-th basis at time frame t.

The noise model is based on the assumption that there are K bases and K activations, and the noise power spectrogram is generated from the product of the bases and the activations. Then, the distribution of the noise is given by

$$p(s_{ft2}|\boldsymbol{W}, \boldsymbol{H}) = \mathcal{N}_{\mathbb{C}}(s_{ft2}\,|\,0, \lambda_{ft2}) \tag{14}$$

$$= \mathcal{N}_{\mathbb{C}}\left(s_{ft2}\,|\,0, \sum_{k=1}^{K} w_{fk}h_{kt}\right). \tag{15}$$

D. Model Formulation

The proposed model represents the generative process of observed multichannel spectrograms by integrating the source models and the spatial model as follows:

$$\log p(\boldsymbol{x}_{ft}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{G}, \boldsymbol{Z}) = \log \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{x}_{ft}|\boldsymbol{0}, \boldsymbol{Y}_{ft}\right)$$

$$= -\mathrm{tr}\left(\boldsymbol{X}_{ft}\boldsymbol{Y}_{ft}^{-1}\right) - \log|\boldsymbol{Y}_{ft}| + \mathrm{const}, \tag{16}$$

where $\boldsymbol{X}_{ft} = \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H$, and $\boldsymbol{Y}_{fti}$ and $\boldsymbol{Y}_{ft}$ are given by

$$\boldsymbol{Y}_{fti} = \lambda_{fti}\boldsymbol{G}_{fi}, \tag{17}$$

$$\boldsymbol{Y}_{ft} = \sum_{i=1}^{2} \boldsymbol{Y}_{fti} = \sum_{i=1}^{2} \lambda_{fti}\boldsymbol{G}_{fi}. \tag{18}$$

Since λ_{ft1} is a power spectrum of the speech, and λ_{ft2} is that of the noise, λ_{fti} is given by

$$\lambda_{ft1} = \sigma_{\theta,f}^2(\boldsymbol{z}_t), \tag{19}$$

$$\lambda_{ft2} = \sum_{k=1}^{K} w_{fk}h_{kt}. \tag{20}$$

To complete a Bayesian formulation, we put a conjugate prior to each parameter as follows:

$$w_{fk} \sim \mathrm{Gamma}(a_w, b_w), \tag{21}$$

$$h_{kt} \sim \mathrm{Gamma}(a_h, b_h), \tag{22}$$

$$\boldsymbol{G}_{fi} \sim \mathrm{Wishart}_{\mathbb{C}}(\nu, \boldsymbol{G}_{fi}^0), \tag{23}$$

where $a_* > 0$, $b_* > 0$, $\nu \geq M$, and $\boldsymbol{G}_{fi}^0$ is positive definite matrix. $\mathrm{Gamma}(a, b)$ indicates Gamma distribution with a shape parameter a and a rate parameter b, and $\mathrm{Wishart}_{\mathbb{C}}(\nu, \boldsymbol{G}_0)$ is a complex Wishart distribution with degree of freedom ν and scale matrix $\boldsymbol{G}_0$ given by

$$\mathrm{Wishart}_{\mathbb{C}}(\boldsymbol{G}\,|\,\nu, \boldsymbol{G}_0) \propto |\boldsymbol{G}|^{\nu-M}\exp(-\mathrm{tr}(\boldsymbol{G}_0^{-1}\boldsymbol{G})). \tag{24}$$

E. Speech Enhancement

To enhance the target speech, we use a multichannel Wiener filter (MWF) [4]. With the MWF, the enhanced speech spectrum $\tilde{s}_{ft1m}$ is given by

$$\tilde{s}_{ft1m} = (\boldsymbol{Y}_{ft1}\boldsymbol{Y}_{ft}^{-1}\boldsymbol{x}_{ft})_m. \tag{25}$$

F. Unsupervised Pre-training

The parameter θ of the deep speech prior $p_\theta(\boldsymbol{s}_{t1}|\boldsymbol{z}_t)$ is estimated by maximizing the log marginal likelihood $\log p_\theta(\boldsymbol{s}_1)$, where $\boldsymbol{s}_1 = \{s_{ft1}\}_{f,t=1}^{F,T}$ is the speech spectra. Since the log marginal likelihood is analytically intractable, we use the VAE for estimating the parameter ϕ of the variational posterior

$q_\phi(\boldsymbol{Z}|\boldsymbol{s}_1)$ and the parameter θ together. The variational posterior is defined as follows:

$$q_\phi(\boldsymbol{Z}|\boldsymbol{s}_1) = \prod_{t=1}^{T}\prod_{d=1}^{D} q_\phi(z_{td} \mid \boldsymbol{s}_{t1}) \tag{26}$$

$$= \prod_{t=1}^{T}\prod_{d=1}^{D} \mathcal{N}\left(z_{td} \mid \mu_{\phi,d}(|\boldsymbol{s}_{t1}|^2), \sigma^2_{\phi,d}(|\boldsymbol{s}_{t1}|^2)\right), \tag{27}$$

where $|\boldsymbol{s}_{t1}|^2$ is the power spectra of the speech of all frequency bins at time frame t. $\boldsymbol{\mu}_\phi(\,\cdot\,) : \mathbb{R}^F \to \mathbb{R}^D$ and $\boldsymbol{\sigma}^2_\phi(\,\cdot\,) : \mathbb{R}^F \to \mathbb{R}^D_+$ are non-linear functions given by DNNs, and $\mu_{\phi,d}(\,\cdot\,)$ and $\sigma^2_{\phi,d}(\,\cdot\,)$ are the d-th element of $\boldsymbol{\mu}_\phi(\,\cdot\,)$ and $\boldsymbol{\sigma}^2_\phi(\,\cdot\,)$, respectively. As mentioned in Section II, the parameters of $\boldsymbol{\sigma}^2_\theta(\,\cdot\,)$, $\boldsymbol{\mu}_\phi(\,\cdot\,)$, and $\boldsymbol{\sigma}^2_\phi(\,\cdot\,)$ are estimated by maximizing the lower bound $\tilde{\mathcal{L}}(\theta,\phi)$ given by

$$\tilde{\mathcal{L}}(\theta,\phi) = \sum_{t=1}^{T}\sum_{d=1}^{D}\frac{1}{2}\left(\log\left(\sigma^2_{\phi,d}(|\boldsymbol{s}_{t1}|^2)\right) - \mu_{\phi,d}(|\boldsymbol{s}_{t1}|^2)^2 - \sigma^2_{\phi,d}(|\boldsymbol{s}_{t1}|^2)\right) + \sum_{f=1}^{F}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}\left\{-\log\sigma^2_{\theta,f}(\boldsymbol{z}_t^{(l)}) - \frac{|s_{ft1}|^2}{\sigma^2_{\theta,f}(\boldsymbol{z}_t^{(l)})}\right\} + \text{const}, \tag{28}$$

$$\boldsymbol{z}_t^{(l)} = \boldsymbol{\mu}_\phi(|\boldsymbol{s}_{t1}|^2) + \boldsymbol{\epsilon}^{(l)} \odot \sqrt{\boldsymbol{\sigma}^2_\phi(|\boldsymbol{s}_{t1}|^2)}, \tag{29}$$

where $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \boldsymbol{I}_D)$ and L is the number of samples.

G. Mathematical Formulas Used for Inference

To estimate the parameters in a Bayesian manner, we lower-bound the log likelihood (Eq. (16)). Here we summarize three mathematical formulas [17] used for this. First, for a concave function $f_1(\boldsymbol{G}) = -\text{tr}(\boldsymbol{X}\boldsymbol{G}^{-1})$ with any matrix $\boldsymbol{X} \succeq \boldsymbol{0}$, we use an inequality given by

$$-\text{tr}\left(\boldsymbol{X}\left(\sum_{n=1}^{N}\boldsymbol{G}_n\right)^{-1}\right) \geq -\sum_{n=1}^{N}\text{tr}\left(\boldsymbol{G}_n^{-1}\boldsymbol{\Phi}_n\boldsymbol{X}\boldsymbol{\Phi}_n^H\right), \tag{30}$$

where $\{\boldsymbol{G}_n\}_{n=1}^N$ is a set of arbitrary matrices, $\{\boldsymbol{\Phi}_n\}_{n=1}^N$ is a set of auxiliary matrices that sum to the identity matrix ($\sum_{n=1}^N \boldsymbol{\Phi}_n = \boldsymbol{I}$), and the equality holds when $\boldsymbol{\Phi}_n = \boldsymbol{G}_n(\sum_{n'=1}^N \boldsymbol{G}_{n'})^{-1}$.

Second, for a convex function $f_2(\boldsymbol{G}) = -\log|\boldsymbol{G}|$ ($\boldsymbol{G} \in \mathbb{R}^{M\times M}; \boldsymbol{G} \succeq \boldsymbol{0}$), we calculate a tangent plane at arbitrary $\boldsymbol{\Omega} \succeq \boldsymbol{0}$ by using a first-order Taylor expansion as follows:

$$-\log|\boldsymbol{G}| \geq -\log|\boldsymbol{\Omega}| - \text{tr}(\boldsymbol{\Omega}^{-1}\boldsymbol{G}) + M, \tag{31}$$

where the equality holds when $\boldsymbol{\Omega} = \boldsymbol{G}$.

Third, for a convex function $f_3(\boldsymbol{x}) = 1/\sum_n x_n$, we apply Jensen's inequality to obtain

$$\frac{1}{\sum_{n=1}^N x_n} \leq \sum_{n=1}^{N}\psi_n^2\frac{1}{x_n}, \tag{32}$$

where $\{\psi_n\}_{n=1}^N$ is a set of auxiliary variables that $\psi_n \geq 0$ and $\sum_n \psi_n = 1$, and the equality holds when $\psi_n = x_n/(\sum_{n'} x_{n'})$.

H. Bayesian Inference

Given observed data $\boldsymbol{X} = (\boldsymbol{X}_{ft})$, our goal is to calculate the full posterior of all random variables, $p(\boldsymbol{Z},\boldsymbol{W},\boldsymbol{H},\boldsymbol{G}|\boldsymbol{X})$. Since this is analytically intractable, we use a Markov chain Monte Carlo (MCMC) method for generating random samples from the posterior. The main difficulty is that the log likelihood function given by Eq. (16) includes the inverse of a summation of matrices and the logarithm of a matrix determinant. To solve these problems, the log likelihood is lower-bounded by using the inequalities (30), (31), and (32) as follows:

$$\log p(\boldsymbol{x}_{ft}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{G},\boldsymbol{Z}) \geq -\text{tr}\left(\sum_{i=1}^{2}\lambda_{fti}\boldsymbol{\Omega}_{ft}^{-1}\boldsymbol{G}_{fi}\right) - \text{tr}\left(\frac{1}{\lambda_{ft1}}\boldsymbol{\Phi}_{ft1}\boldsymbol{X}_{ft}\boldsymbol{\Phi}_{ft1}^H\boldsymbol{G}_{f1}^{-1}\right) - \text{tr}\left(\left(\sum_{k=1}^{K}\frac{\psi_{ftk}^2}{w_{fk}h_{kt}}\right)\boldsymbol{\Phi}_{ft2}\boldsymbol{X}_{ft}\boldsymbol{\Phi}_{ft2}^H\boldsymbol{G}_{f2}^{-1}\right) + \text{const}, \tag{33}$$

where the equality holds when $\boldsymbol{\Phi}_{fti}$, $\boldsymbol{\Omega}_{ft}$, and ψ_{ftk} satisfy

$$\boldsymbol{\Phi}_{fti} = \lambda_{fti}\boldsymbol{G}_{fi}\left(\sum_{i'=1}^{2}\lambda_{fti'}\boldsymbol{G}_{fi'}\right)^{-1}, \tag{34}$$

$$\boldsymbol{\Omega}_{ft} = \sum_{i=1}^{2}\lambda_{fti}\boldsymbol{G}_{fi}, \tag{35}$$

$$\psi_{ftk} = \frac{w_{fk}h_{kt}}{\sum_{k'=1}^K w_{fk'}h_{k't}}. \tag{36}$$

The conditional posteriors of w_{fk}, h_{kt}, and $\boldsymbol{G}_{fi}$ are given as follows:

$$w_{fk} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\backslash w_{fk}} \sim \text{GIG}(a_w, b^w_{fk}, \rho^w_{fk}), \tag{37}$$

$$h_{kt} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\backslash h_{kt}} \sim \text{GIG}(a_h, b^h_{kt}, \rho^h_{kt}), \tag{38}$$

$$\boldsymbol{G}_{fi} \mid \boldsymbol{X}, \boldsymbol{\Theta}_{\backslash \boldsymbol{G}_{fi}} \sim \text{MGIG}_{\mathbb{C}}(\nu, \boldsymbol{R}_{fi}, \boldsymbol{T}_{fi}), \tag{39}$$

where $\boldsymbol{\Theta}$ means the set of all variables and $\boldsymbol{\Theta}_{\backslash\alpha}$ means the set of all variables except α. $\text{GIG}(a,b,\rho)$ and $\text{MGIG}_{\mathbb{C}}(\nu,\boldsymbol{R},\boldsymbol{T})$ represent a generalized inverse Gaussian (GIG) distribution and a complex matrix GIG distribution given by

$$\text{GIG}(x \mid a,b,\rho) \propto x^{a-1}\text{exp}(-(bx+\rho/x)), \tag{40}$$

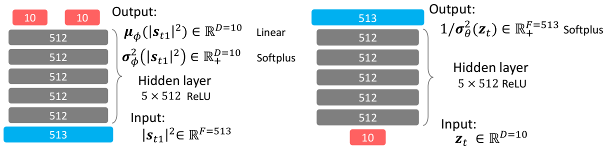
$$\text{MGIG}_{\mathbb{C}}(\boldsymbol{G} \mid \nu,\boldsymbol{R},\boldsymbol{T}) \propto |\boldsymbol{G}|^{\nu-M}\text{exp}(-\text{tr}(\boldsymbol{R}\boldsymbol{G}+\boldsymbol{T}\boldsymbol{G}^{-1})). \tag{41}$$

The parameters of the conditional posteriors b^w_{fk}, ρ^w_{fk}, b^h_{kt}, ρ^h_{kt}, $\boldsymbol{R}_{fi}$, and $\boldsymbol{T}_{fi}$ are given as follows:

$$b^w_{fk} = b_w + \sum_{t=1}^{T}h_{kt}\text{tr}(\boldsymbol{\Omega}_{ft}^{-1}\boldsymbol{G}_{f2}), \tag{42}$$

$$\rho^w_{fk} = \sum_{t=1}^{T}\frac{1}{h_{kt}}\psi^2_{ftk}\text{tr}(\boldsymbol{\Phi}_{ft2}\boldsymbol{X}_{ft}\boldsymbol{\Phi}_{ft2}^H\boldsymbol{G}_{f2}^{-1}), \tag{43}$$

$$b^h_{kt} = b_h + \sum_{f=1}^{F}w_{fk}\text{tr}(\boldsymbol{\Omega}_{ft}^{-1}\boldsymbol{G}_{f2}), \tag{44}$$


10
10
Output:
$\mu_\phi(|s_{t1}|^2) \in \mathbb{R}^{D=10}$
Linear
$\sigma^2_\phi(|s_{t1}|^2) \in \mathbb{R}_+^{D=10}$
Softplus
512
512
512
512
512
Hidden layer
5 × 512 ReLU
Input:
$|s_{t1}|^2 \in \mathbb{R}^{F=513}$
513
513
Output:
$1/\sigma^2_\theta(z_t) \in \mathbb{R}_+^{F=513}$ Softplus
Hidden layer
5 × 512 ReLU
Input:
$z_t \in \mathbb{R}^{D=10}$
10

(a) Encoder $q_\phi(\boldsymbol{z}_t \mid \boldsymbol{s}_{t1})$ outputs the mean and variance of $\boldsymbol{z}_t$

(b) Decoder $p_\theta(\boldsymbol{s}_{t1} \mid \boldsymbol{z}_t)$ outputs the power spectra of speech $\boldsymbol{s}_{t1}$

Fig. 2: Architectures of the DNNs.

$$\rho^h_{kt} = \sum_{f=1}^{F} \frac{1}{w_{fk}} \psi^2_{ftk} \mathrm{tr}(\boldsymbol{\Phi}_{ft2} \boldsymbol{X}_{ft} \boldsymbol{\Phi}^H_{ft2} \boldsymbol{G}^{-1}_{f2}), \tag{45}$$

$$\boldsymbol{R}_{fi} = (\boldsymbol{G}^0_{fi})^{-1} + \sum_{t=1}^{T} \lambda_{fti} \boldsymbol{\Omega}^{-1}_{ft}, \tag{46}$$

$$\boldsymbol{T}_{fi} = \sum_{t=1}^{T} \frac{1}{\lambda_{fti}} \boldsymbol{\Phi}_{fti} \boldsymbol{X}_{ft} \boldsymbol{\Phi}^H_{fti}. \tag{47}$$

As the conditional posterior of $\boldsymbol{z}_t$ is intractable, we use the Metropolis method, which draws a sample $\boldsymbol{z}_t^{\mathrm{new}}$ from a proposal distribution that depends on the previous sample $\boldsymbol{z}_t^{\mathrm{old}}$ and determines whether or not the sample is accepted on the basis of the acceptance rate. The proposal distribution $q(\boldsymbol{z}_t | \boldsymbol{z}_t^{\mathrm{old}})$ and the acceptance rate β are given as follows:

$$q(\boldsymbol{z}_t | \boldsymbol{z}_t^{\mathrm{old}}) = \mathcal{N}(\boldsymbol{z}_t^{\mathrm{old}}, \xi \boldsymbol{I}_D), \tag{48}$$

$$\beta_{\boldsymbol{z}_t^{\mathrm{new}}, \boldsymbol{z}_t^{\mathrm{old}}} = \min\left(1, \frac{p(\boldsymbol{x}_t | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{G}, \boldsymbol{z}_t^{\mathrm{new}}) p(\boldsymbol{z}_t^{\mathrm{new}})}{p(\boldsymbol{x}_t | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{G}, \boldsymbol{z}_t^{\mathrm{old}}) p(\boldsymbol{z}_t^{\mathrm{old}})}\right), \tag{49}$$

where ξ is a variance parameter.

IV. Evaluation

We conducted two types of experiments. In the first experiment, we compared the proposed method with a method that does not use spatial information to confirm the effectiveness of the spatial information. In the second experiment, we compared the proposed method with a state-of-the-art unsupervised multichannel source separation method.

A. Confirmation of Effectiveness of Spatial Information

1) Experimental conditions: To confirm the effectiveness of using spatial information, we compare the case of updating only source models with the case of updating both spatial and source models. First, the spatial covariance matrices $\{\boldsymbol{G}_{fi}\}_{f,i=1}^{F,2}$ were set to be identity matrices, and we updated only the source models ($\boldsymbol{Z}$, $\boldsymbol{W}$, and $\boldsymbol{H}$) 30 times. This meant that we used observed signals of all channels but not spatial information. Next, we updated both the source and spatial models 30 times. We calculated the source separation performance for the enhanced signals that are estimated with the samples of each iteration. We used the signal-to-distortion ratio (SDR) [18], [19] as evaluation metrics.

We used the simulated utterances in the development dataset of CHiME3 [20] for evaluation. The dataset consists of 1640 simulated utterances in four types of noisy environment: on a bus (BUS), in a cafe (CAF), in a pedestrian area (PED), and on a street junction (STR). We randomly chose 25 utterances for each environment. Although these utterances were assumed to be recorded using a tablet with 6 microphones, we selected five channels ($M = 5$), channel 1, 3, 4, 5, and 6. This was because the second microphone was set on the back side and the others were on the front side.

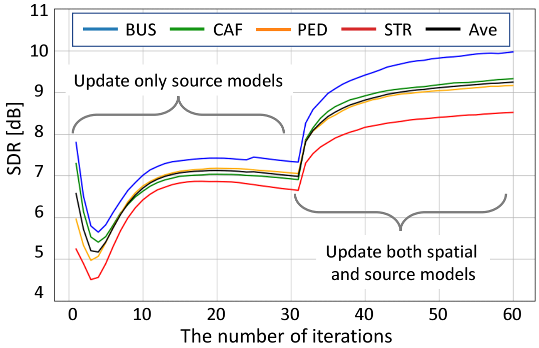


Fig. 3: Source separation performance after each iteration.

To estimate the parameters of the speech prior $p(\boldsymbol{s}_{t1} | \boldsymbol{z}_t)$, we used the VAE. The parameters of the VAE were the same as in [15]. The dimension of the latent variables D was 10. The architectures of the DNNs are shown in Fig. 2. To obtain the spectrograms, we use a short time Fourier transform with a shifting interval of 256 samples and a window length of 1024 samples; the number of frequency bins F was 513. We used the WSJ-0 corpus [21] as a training dataset, which consists of clean speech signals of about 15-hour length. The other parameters of the proposed method were set as follows. The number of the bases $K = 5$, and the parameter of the proposal distribution $\xi = 0.1$. The parameters of the prior distributions were $a_w = b_w = a_h = b_h = 1$, $\nu = M = 5$, and $\boldsymbol{G}^0_{fi} = \frac{1}{\nu} \boldsymbol{I}_M$. The initial values of $\boldsymbol{W}$ and $\boldsymbol{H}$ were sampled from the prior distributions. The initial values of $\boldsymbol{Z}$ were sampled from the Gaussian distribution with mean $\mu_{\phi,d}(|\boldsymbol{x}|^2)$ and variance $\sigma^2_{\phi,d}(|\boldsymbol{x}|^2)$, where the inputs of $\mu_{\phi,d}(|\boldsymbol{x}|^2)$ and $\sigma^2_{\phi,d}(|\boldsymbol{x}|^2)$ were the power spectrograms of the observed signals. To obtain the enhanced speech, the multichannel Wiener filter was used.

2) Experimental result: Fig. 3 shows the average SDRs of the enhanced signals calculated from the samples of each iteration. In all situations, the separation performance after we updated spatial information was significantly improved. Comparing the SDRs of 30th iteration and 60th iteration, they were improved by 2.3 dB on average, and the effectiveness of the spatial information was confirmed.

We find that the separation performance of the first iteration was higher than that after a few iterations. At the first iteration, the power spectrogram of the speech calculated from the initial values of $\boldsymbol{Z}$ was close to the ground truth, whereas $\boldsymbol{W}$ and $\boldsymbol{H}$ do not reflect the observed signals as they were randomly sampled from the prior distributions. The deterioration of the separation performance during the first few iterations is probably because $\boldsymbol{Z}$ was adversely affected from $\boldsymbol{W}$ and $\boldsymbol{H}$ that did not yet approach appropriate values. It is interesting to

TABLE I: The average SDRs for each situation.

Method	Average	BUS	CAF	PED	STR
Proposed	10.6	9.8	10.8	11.7	10.1
ILRMA	12.3	11.6	12.6	13.9	11.1
Input	5.8	2.8	7.3	8.1	5.1

note that the encoder outputs reasonable values of $\boldsymbol{Z}$ from the noisy speech spectra although it was trained with only clean speech spectra.

B. Comparison with a State-of-the-art Unsupervised Multichannel Source Separation Method

1) Experimental conditions: We compared the source separation performance of the proposed method with that of a state-of-the-art unsupervised multichannel source separation method, independent low-rank matrix analysis (ILRMA) [5], using the same data in terms of SDR. According to [5], the ILRMA achieved the best score in a speech separation task compared to other multichannel source separation methods such as independent vector analysis (IVA) [22] and multichannel NMF (MNMF) [3], [4]. The parameters of the ILRMA were set similarly as in the experimental section in [5]. The number of bases for each source was set to be 2, and the number of iterations was set to be 200.

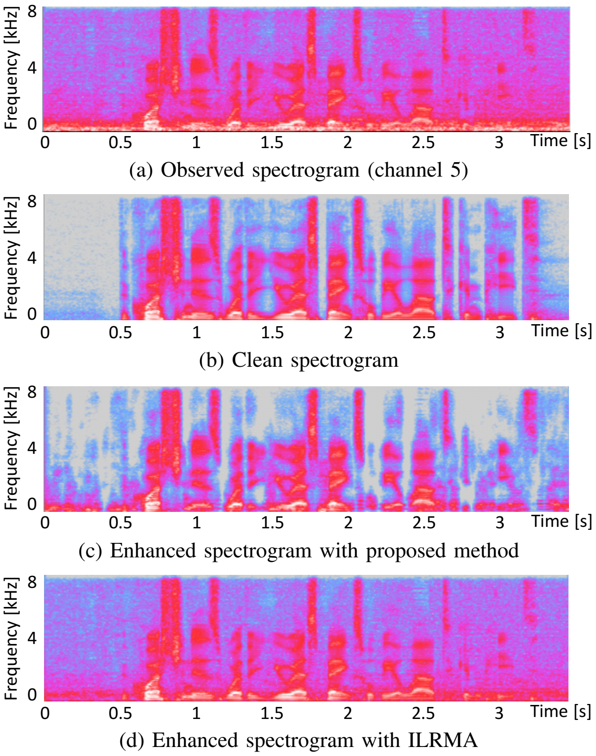
The parameters of the proposed method were the same except the parameter $\boldsymbol{G}^0_{fi}$. We first used VAE-NMF reported in [15], which is a single-channel version of the proposed method. The parameters of the VAE-NMF is the same as those of the proposed method except the parameters of the spatial covariance matrix. After drawing 100 samples for burn-in, we drew 50 samples and calculated the mean of the power spectrograms. The parameter $\boldsymbol{G}^0_{fi}$ is calculated as follows.

$$\boldsymbol{G}^0_{fi} = \frac{1}{\nu} \frac{1}{\sum_{t=1}^{T} r_{fti}} \sum_{t=1}^{T} r_{fti} \frac{\boldsymbol{X}_{ft}}{\bar{\lambda}_{fti}}, \tag{50}$$

$$r_{ft1} = \begin{cases} 1 & (\lambda_{ft1} \geq \lambda_{ft2}) \\ 0 & (\lambda_{ft1} < \lambda_{ft2}) \end{cases}, \tag{51}$$

where $\bar{\lambda}_{fti}$ is the sample average of the power estimated by the VAE-NMF. We updated only the source models 10 times and updated both the source and spatial models 30 times.

2) Experimental results: Table I shows the average SDRs in each environment for the proposed method and the ILRMA. The proposed method had an SDR equivalent to the ILRMA in the situation STR. In the other situations, on the other hand, the average SDRs of the proposed method were lower than those of the ILRMA. Fig. 4 shows the power spectrograms of (a) an observed signal in the STR, (b) the clean signal, (c) the enhanced signal with the proposed method, and (d) the enhanced signal with the ILRMA. In this case, the SDRs of the observed signal, the enhanced signal with the proposed method, and the enhanced signal with the ILRMA are 0.3, 7.0, and 14.7, respectively. As we see, the proposed method suppressed noise largely. However, noise that sounded like speech still remained in the enhanced speech with the proposed method as we find by listening to it. This was partly because when the background sound contains speech, the proposed method sometimes erroneously included the noise in the target speech. Moreover, it sometimes suppressed the target speech, for example, around 2.7 second and 3.0 second. A reason is that for speech sections with heavy noise, many iterations are required to guide the latent variables to appropriate values since the initial value output from the encoder is not close to the ground truth due to the heavy noise. Therefore, if a speech section has low-rankness, the noise model ($\boldsymbol{W}$ and $\boldsymbol{H}$) absorbs the speech before the latent variables settle down to appropriate values.



(a) Observed spectrogram (channel 5)

(b) Clean spectrogram

(c) Enhanced spectrogram with proposed method

(d) Enhanced spectrogram with ILRMA

Fig. 4: Results of speech enhancement by the proposed method and the ILRMA. Power spectrograms are shown.

A promising solution to the problem of noise sounding like speech is to incorporate time dependence between latent variables $\boldsymbol{z}_t$ (e.g. smoothness) and train the VAE so that speech uttered by each speaker is described by a local region in the latent space. If the noise sounding like speech is described by latent variables away from such a local region for the target speech, then the noise can be suppressed according to the time dependence. A solution to the speech suppression is to change the order of updating the variables in addition to the above extension. By updating only the latent variables before updating the noise model, it can be avoided that the noise model absorbs the speech.

V. Conclusion

This paper presented an innovative multichannel speech enhancement method that integrates a DNN-based generative model of speech spectra and NMF-based generative model of noise spectra. The advantage of the proposed method over other DNN-based multichannel source separation methods is that it uses only clean speech signals for training and the noise model is estimated on the fly. In the experiment, we confirmed the effectiveness of using the spatial information. The SDR of the enhanced signal estimated using both the source models and spatial model was superior to that of the enhanced signal estimated using only source models. We compared the proposed method with the ILRMA using the simulated data of the CHiME3 development set. Although the separation performance was lower than that of ILRMA, we were able to find two issues for improving the proposed method. One is the treatment for noise sounding like speech and the other is suppression of speech in sections with heavy noise.

We plan to extend the proposed method by incorporating time dependence between the latent variables. This would improve the speech enhancement performance as discussed in the last section. Dynamical systems with a deep generative model architecture would be suitable for this extension [23]. Moreover, we plan to extend the proposed method to deal with multiple speakers.

VI. Acknowledgment

This study was partially supported by KAKENHI No. 16H01744.

References

[1] Chengli Sun, Qin Zhang, Jian Wang, and Jianxiao Xie. Noise reduction based on robust principal component analysis*. *Journal of Computational Information Systems*, 10(10):4403–4410, 2014.

[2] Kevin Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *IEEE ICASSP*, pages 4029–4032, 2008.

[3] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE TASLP*, 18(3):550–563, 2010.

[4] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE TASLP*, 21(5):971–982, 2013.

[5] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE TASLP*, 24(9):1626–1641, 2016.

[6] Kousuke Itakura, Yoshiaki Bando, Eita Nakamura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Bayesian multichannel audio source separation based on integrated source and spatial models. *IEEE TASLP*, 26(4):831–846, 2018.

[7] Simon Arberet, Alexey Ozerov, Ngoc Duong, Emmanuel Vincent, Rémi Gribonval, Frédéric Bimbot, and Pierre Vandergheynst. Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *ISSPA*, pages 1–4, 2010.

[8] Joonas Nikunen and Tuomas Virtanen. Multichannel audio separation by direction of arrival based spatial covariance model and non-negative matrix factorization. In *IEEE ICASSP*, pages 6677–6681, 2014.

[9] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[10] Xu Li, Junfeng Li, and Yonghong Yan. Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions. In *Interspeech*, pages 1203–1207, 2017.

[11] Hakan Erdogan, John Hershey, Shinji Watanabe, Michael Mandel, Jonathan Le Roux, Hakan Erdogan, John Hershey, Shinji Watanabe, Michael Mandel, and Jonathan Le Roux. Improved MVDR beamforming using single-channel mask prediction networks. In *Interspeech*, pages 1981–1985, 2016.

[12] Tomohiro Nakatani, Nobutaka Ito, Takuya Higuchi, Shoko Araki, and Keisuke Kinoshita. Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In *IEEE ICASSP*, pages 286–290, 2017.

[13] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE TASLP*, 24(9):1652–1664, 2016.

[14] Shinichi Mogami, Hayato Sumino, Daichi Kitamura, Norihiro Takamune, Shinnosuke Takamichi, Hiroshi Saruwatari, and Nobutaka Ono. Independent deeply learned matrix analysis for multichannel audio source separation. *arXiv preprint arXiv:1806:10307*, 2018.

[15] Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. *arXiv preprint arXiv:1710.11439*, 2017.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[17] Kazuyoshi Yoshii, Katsutoshi Itoyama, and Masataka Goto. Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation. In *IEEE ICASSP*, pages 51–55, 2016.

[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE TASLP*, 14(4):1462–1469, 2006.

[19] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *ISMIR*, pages 367–372, 2014.

[20] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third ‘ CHiME’ speech separation and recognition challenge: Analysis and outcomes. *Computer Speech & Language*, 46:605–626, 2017.

[21] John Garofalo, David Graff, Doug Paul, and David Pallett. CSR-I (WSJ0) complete. *Linguistic Data Consortium, Philadelphia*, 2007.

[22] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis: An extension of ica to multivariate components. In *Independent Component Analysis and Blind Signal Separation*, pages 165–172. Springer Berlin Heidelberg, 2006.

[23] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511:05121*, 2015.