IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 8, 2019, accepted November 4, 2019, date of publication November 19, 2019, date of current version December 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954120

Underdetermined Source Separation Based on Generalized Multichannel Variational Autoencoder

SHOGO SEKI[1], HIROKAZU KAMEOKA[2], (Senior Member, IEEE), LI LI[3], TOMOKI TODA[4], (Senior Member, IEEE), AND KAZUYA TAKEDA[5], (Senior Member, IEEE)

[1]Graduate School of Informatics, Nagoya University, Nagoya 464–0861, Japan
[2]Nippon Telegraph and Telephone Corporation, Atsugi 243–0198, Japan
[3]Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba 305–8573, Japan
[4]Information Technology Center, Nagoya University, Nagoya 464–0861, Japan
[5]Institutes of Innovation for Future Society, Nagoya University, Nagoya 464–8603, Japan

Corresponding author: Shogo Seki (seki.shogo@g.sp.m.is.nagoya-u.ac.jp)

This work was supported by JSPS KAKENHI Grant Number JP17H01763 and JST CREST Grant Number JPMJCR19A3, Japan.

ABSTRACT This paper deals with a multichannel audio source separation problem under underdetermined conditions. Multichannel non-negative matrix factorization (MNMF) is a powerful method for underdetermined audio source separation, which adopts the NMF concept to model and estimate the power spectrograms of the sound sources in a mixture signal. This concept is also used in independent low-rank matrix analysis (ILRMA), a special class of the MNMF formulated under determined conditions. While these methods work reasonably well for particular types of sound sources, one limitation is that they can fail to work for sources with spectrograms that do not comply with the NMF model. To address this limitation, an extension of ILRMA called the multichannel variational autoencoder (MVAE) method was recently proposed, where a conditional VAE (CVAE) is used instead of the NMF model for expressing source power spectrograms. This approach has performed impressively in determined source separation tasks thanks to the representation power of deep neural networks. While the original MVAE method was formulated under determined mixing conditions, this paper proposes a generalized version of it by combining the ideas of MNMF and MVAE so that it can also deal with underdetermined cases. We call this method the generalized MVAE (GMVAE) method. In underdetermined source separation and speech enhancement experiments, the proposed method performed better than baseline methods.

INDEX TERMS Underdetermined source separation, variational audoencoder, non-negative matrix factorization.

I. INTRODUCTION

Blind source separation (BSS) refers to the problem of separating out underlying source signals present in observed mixture signals received by a microphone array. A frequency-domain method is typically used to tackle BSS problems for convolutive mixtures by using various models for source signals and/or array responses. For example, an extension of independent component analysis (ICA) [1] called independent vector analysis (IVA) [2], [3] makes it possible to jointly perform frequency-wise source separation and permutation alignment by assuming that the magnitudes of the frequency components originating from the same source are likely to vary coherently over time.

Other methods involve multichannel extensions of non-negative matrix factorization (NMF) [4]–[9]. NMF is a dimension reduction method for matrices consisting of only non-negative entries. In audio signal processing, NMF was originally applied for music transcription and monaural source separation tasks [10], [11], where the power spectrogram (or the magnitude spectrogram) of a mixture signal is regarded as a non-negative matrix to be approximated as the product of two non-negative matrices. This can be viewed as approximating the power spectrum (or the magnitude spectrum) of a mixture signal observed at each time frame by the sum of a fixed number of basis spectra scaled by time-varying magnitudes.

Multichannel NMF (MNMF) is a method that extends the NMF so that it can additionally use spatial information for source separation. It can also be seen as a frequency-domain BSS method that uses spectral templates as clues

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

TABLE 1. Categorization of proposed and conventional methods.

Method	Mixing condition	Source model
ILRMA [5], [7], [9]	Determined	NMF
MNMF [4], [6], [8]	Underdetermined	NMF
MVAE [14], [15]	Determined	VAE
GMVAE (Proposed)	Underdetermined	VAE

for jointly performing frequency-wise source separation and permutation alignment. MNMF was originally formulated as a method [4] for handling underdetermined as well as determined scenarios in which sources can outnumber microphones. A determined version of MNMF, focused on solving BSS problems in determined settings, was subsequently proposed [5]. While the determined version of MNMF is applicable only to determined cases, it provides a significantly faster algorithm than the general version. This determined MNMF framework was later called independent low-Rank matrix analysis (ILRMA) [12]. It is worthwhile to note that the optimization algorithms for MNMF and ILRMA are guaranteed to converge to a stationary point, and work reasonably well for some types of sound sources. However, they can fail to work when encountering sound sources with spectrograms that do not follow the NMF model, resulting in performance limitations.

To address these limitations, new methods using variational autoencoders (VAEs) [13] have been proposed as alternatives to NMF-based source modeling [14]–[19]. A VAE is a type of generative neural network capable of modeling high-dimensional data such as images. The idea of these methods is to use a VAE to model the spectra of source signals. Some of these methods [16], [17] were designed to deal with speech enhancement tasks by modeling the spectrogram of a particular source to be enhanced using a regular VAE and expressing the spectrograms of the other sources using the NMF model. This allows these methods to handle semi-supervised scenarios in which interference sources are unseen in the training set. We hereafter refer to this type of method as "VAE-NMF". Another VAE-based method worth noting is the multichannel VAE (MVAE) method [14], [15]. This method is an extension of ILRMA with the difference being that a conditional VAE (CVAE) [20] instead of the NMF model is used as a generative model of source spectrograms. By training the CVAE using the spectrograms of class-labeled speech samples, the resulting decoder can be used as a generative model of the speech spectrograms of multiple speakers where its inputs are interpreted as the model parameters to be optimized. Thanks to the ability of a VAE to accurately represent spectrograms, the MVAE method consistently performed better than ILRMA in determined source separation tasks.

While the original MVAE method was formulated under determined mixing conditions, we propose a generalized version of the original MVAE method by combining the ideas of MNMF and the MVAE method so that it can also deal with underdetermined cases. We call this method the generalized MVAE (GMVAE) method to distinguish it from the MVAE method (Table 1). The remainder of this paper is organized as follows. In Section 2, we begin by formulating the BSS problem and state the motivation for introducing VAE-based source models. In Section 3, we review related work including those on MNMF, ILRMA and the MVAE method and show that the relationship between MNMF and the GMVAE method corresponds to that between ILRMA and the MVAE method. In Section 4, we discuss the development of a convergence-guaranteed parameter optimization algorithm for the GMVAE method by combining the ideas for the parameter optimization processes introduced in MNMF and the MVAE method. In Section 5, we experimentally show the superiority of the GMVAE method over MNMF in underdetermined source separation tasks and over VAE-NMF in semi-supervised speech enhancement tasks. Note that this paper is an extended journal version of our preprint paper [18] and conference paper [19].

II. PROBLEM FORMULATION

Suppose that there are J source signals and that a mixed signal from these sound sources is captured by I microphones. Let $s_j(f,n)$ and $x_i(f,n)$ respectively be the short-time Fourier transform (STFT) coefficient of the j-th source signal and that of the i-th observed signal, where f and n are the frequency and time indices, respectively. We denote the vectors containing the STFT coefficients of all the sources $s_1(f,n),\ldots,s_J(f,n)$ and the observed signals $x_1(f,n),\ldots,x_I(f,n)$ as

$$\mathbf{s}(f,n) = [s_1(f,n),\ldots,s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J, \tag{1}$$

$$\mathbf{x}(f,n) = [x_1(f,n),\ldots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ represents the transpose and $\mathbb{C}$ denotes complex numbers. We assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n)$:

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{3}$$

Equation (3) is usually called the local Gaussian model (LGM) [21]–[23]. When $s_j(f,n)$ and $s_{j'}(f,n)$ are mutually independent for $j \neq j'$, $\mathbf{s}(f,n)$ follows a complex Gaussian distribution

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{s}(f,n)|\mathbf{0}, \mathbf{V}(f,n)), \tag{4}$$

where $\mathbf{V}(f,n)$ is a diagonal covariance matrix with diagonal entries $v_1(f,n),\ldots,v_J(f,n)$.

In a general situation in which the sources can outnumber the microphones, a mixing system is given as follows:

$$\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n), \tag{5}$$

which describes the relationship between $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$, where $\mathbf{A}(f) = [\mathbf{a}_1(f),\ldots,\mathbf{a}_J(f)] \in \mathbb{C}^{I\times J}$ is referred to as a mixing matrix. From (4) and (5), $\mathbf{x}(f,n)$ is shown to follow

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)), \tag{6}$$

where $(\cdot)^{\mathsf{H}}$ represents the conjugate transpose. Thus, given an observed mixed signal $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$, using the mixing matrices $\mathcal{A} = \{\mathbf{A}(f)\}_f$ and variance in source signals $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$, the log-likelihood is given as

$$\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \stackrel{c}{=} -\sum\nolimits_{f,n}\left[\mathrm{tr}(\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)) + \mathrm{logdet}\hat{\mathbf{X}}(f,n)\right], \quad (7)$$

where $\stackrel{c}{=}$ denotes the equality up to constant terms and

$$\mathbf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n), \quad (8)$$
$$\hat{\mathbf{X}}(f,n) = \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f). \quad (9)$$

If there is no constraint imposed on $v_j(f,n)$, (7) will be split into multiple frequency-wise source separation problems, which indicates that there is a permutation ambiguity in the separated components for each frequency since permutation of j does not affect the value of the log-likelihood. Thus, permutation alignment is generally required after $\mathcal{A}$ is obtained.

III. RELATED WORK

A. MULTICHANNEL NON-NEGATIVE MATRIX FACTORIZATION (MNMF)

The covariance matrix of $\mathbf{x}(f,n)$ can be written as the linear sum of the outer products of a steering vector $\mathbf{a}_j(f)$ multiplied by source variances $v_j(f,n)$. With MNMF, the outer product of $\mathbf{a}_j(f)$, namely the spatial covariance of the j-th source denoted by $\mathbf{R}_j(f)$, is treated as a full-rank matrix:

$$\hat{\mathbf{X}}(f,n) = \sum\nolimits_j v_j(f,n)\mathbf{a}_j(f)\mathbf{a}_j^{\mathsf{H}}(f)$$
$$= \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f). \quad (10)$$

As with IVA, imposing a constraint on $v_j(f,n)$ allows us to jointly carry out frequency-wise source separation and permutation alignment. With MNMF, $v_j(f,n)$ is modeled as the sum of K_j spectral templates $h_{j,1}(f), \ldots, h_{j,K_j}(f) \geq 0$ scaled by time-varying activations $u_{j,1}(n), \ldots, u_{j,K_j}(n) \geq 0$:

$$v_j(f,n) = \sum\nolimits_{k=1}^{K_j} h_{j,k}(f)u_{j,k}(n). \quad (11)$$

It is also possible to share all the spectral templates of every source and let the contribution of the k-th spectral template to source j be determined in a data-driven manner. Thus, $v_j(f,n)$ can also be expressed as

$$v_j(f,n) = \sum\nolimits_{k=1}^{K} b_{j,k}h_k(f)u_k(n), \quad (12)$$

where $b_{j,k} \in [0,1]$ is a continuous indicator variable satisfying $\sum_k b_{j,k} = 1$. Here $b_{j,k}$ can be interpreted as the expectation of a binary indicator variable that describes the index of the source to which the k-th template is assigned.

The separation algorithm of MNMF consists of iteratively updating the spatial covariance matrices $\mathcal{R} = \{\mathbf{R}_j(f)\}_{j,f}$, and the source model parameters $\mathcal{H}_1 = \{h_{j,k}(f)\}_{j,k,f}$, $\mathcal{U}_1 = \{u_{j,k}(n)\}_{j,k,n}$ or $\mathcal{B} = \{b_{j,k}\}_{j,k}$, $\mathcal{H}_2 = \{h_k(f)\}_{k,f}$, $\mathcal{U}_2 = \{u_k(n)\}_{k,n}$. By using the principle of the majorization-minimization (MM) algorithm [24], [25], we can derive update equations [26].

B. INDEPENDENT LOW-RANK MATRIX ANALYSIS (ILRMA)

ILRMA is a special class of MNMF designed to solve determined source separation problems. Unlike MNMF, which uses the mixing system shown in (5), ILRMA uses the following separation system:

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n), \quad (13)$$

assuming the mixing matrix is invertible. The inverse matrix $\mathbf{W}^{\mathsf{H}}(f) = \left[\mathbf{w}_1^{\mathsf{H}}(f), \ldots, \mathbf{w}_J^{\mathsf{H}}(f)\right]^{\mathsf{H}} \in \mathbb{C}^{J\times I}$ is called the separation matrix.

As with MNMF, the MM-based update equations for $\mathcal{H}_1$ and $\mathcal{U}_1$ or for $\mathcal{B}$, $\mathcal{H}_2$ and $\mathcal{U}_2$ are obtained as closed-form expressions. The separation matrix $\mathbf{W}^{\mathsf{H}}(f)$ can be updated using a fast update rule called iterative projection (IP) [27], originally developed for IVA.

C. DEEP NEURAL NETWORK APPROACH

Instead of using the NMF model, algorithms for the LGM-based multichannel source separation framework, where $v_j(f,n)$ is updated with the output of pretrained deep neural networks at each iteration, have been proposed [9], [28]. One drawback of these algorithms is that updating $v_j(f,n)$ in this way does not guarantee an increase in the log-likelihood.

D. MULTICHANNEL VARIATIONAL AUTOENCODER (MVAE) METHOD

One limitation of the MNMF framework including ILRMA is that it can fail to work for sources with spectrograms that are difficult to express using the NMF model given by (11) or (12). The MVAE method is an improved variant of ILRMA that replaces (11) with a CVAE [14], [15]. The MVAE method models the generative model of the complex spectrogram of a particular sound source using a CVAE with an auxiliary input, indicating the classes of a source, which is represented as a one-hot vector.

The optimization algorithm of the MVAE method consists of updating the separation matrices using IP, the global scale using the MM algorithm and the inputs to the pretrained decoder using backpropagation. The advantage of using the MVAE method is that it can leverage the strong representational power of a VAE for modeling the power spectrogram of sources.

E. VAE-NMF

After our preprint paper on this work [18] was first made public, methods have been made to model sources using NMF and a VAE for multichannel speech enhancement [16], [17]. These methods are designed to model the spectrogram of a particular source to be enhanced using a regular VAE and express the spectrograms of the other sources using the NMF model. This allows these methods to handle semi-supervised scenarios in which interference sources are unseen in the training set.

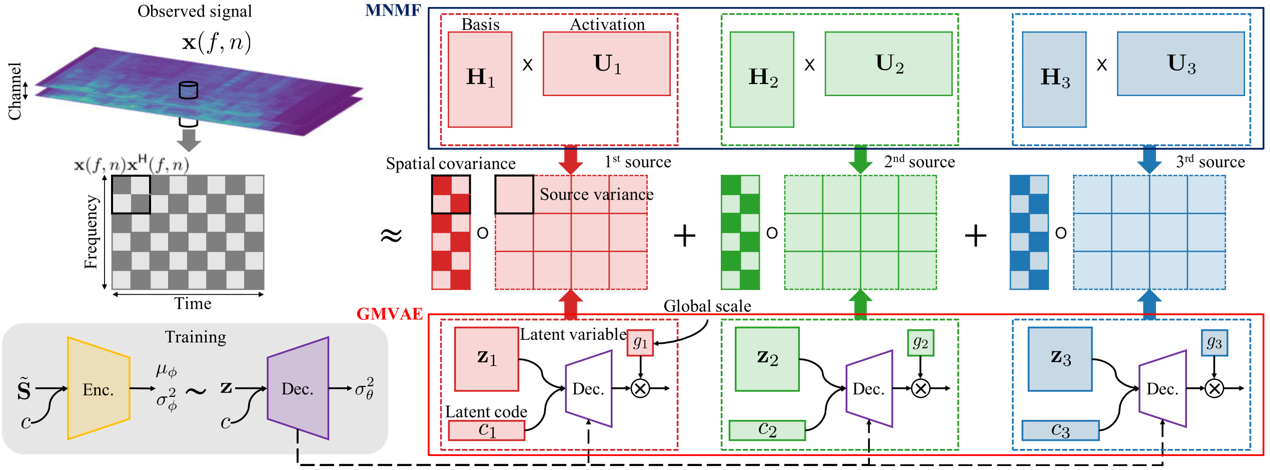


FIGURE 1. Illustration of the modeling concepts of MNMF and GMVAE method. Network parameters to be optimized at training time and parameters of the NMF and CVAE source models to be optimized at separation (inference) time are in colored blocks.

IV. GENERALIZED MVAE METHOD (GMVAE)

A. OVERVIEW

Figure 1 illustrates the modeling concepts of MNMF and the GMVAE method. These methods share the same log-likelihood (7) to maximize, which can be interpreted as the similarity between the outer product of each observed signal vector (8) and the sum of full-rank spatial covariances scaled by source variances (10). As this figure shows, while MNMF represents source spectrograms using the NMF model, the GMVAE method represents them using a trained CVAE decoder network. Note that with the GMVAE method, we treat the spatial covariance $\mathbf{R}_j(f)$ in the same manner as MNMF. At separation (inference) time, the network parameters are fixed at the pretrained values for all the assumed sources and decoder inputs, namely the latent variable $\mathbf{z}_j$, latent code c_j, and global scale g_j become the parameters to be estimated.

B. CVAE PRETRAINING

Our CVAE consists of an encoder network and decoder network, which we train using class-labeled training examples prior to separation. Given a source spectrogram $\tilde{\mathbf{S}}$ with the one-hot encoded class label c, the encoder distribution $q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)$ is expressed as a Gaussian distribution:

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c) = \prod\nolimits_d \mathcal{N}(z(d)|\mu_\phi(d; \tilde{\mathbf{S}}, c), \sigma_\phi^2(d; \tilde{\mathbf{S}}, c)), \quad (14)$$

where $\mathbf{z}$ denotes a latent variable, and $z(d)$, $\mu_\phi(d; \tilde{\mathbf{S}}, c)$, and $\sigma_\phi^2(d; \tilde{\mathbf{S}}, c)$ represent the d–th elements of $\mathbf{z}$, $\mu_\phi(\tilde{\mathbf{S}}, c)$, and $\sigma_\phi^2(\tilde{\mathbf{S}}, c)$, respectively. The decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ is expressed as a zero-mean complex Gaussian distribution (i.e., the LGM):

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g) = \prod\nolimits_{f,n} \mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \quad (15)$$

$$v(f,n) = g\sigma_\theta^2(f,n; \mathbf{z}, c), \quad (16)$$

where $\sigma_\theta^2(f,n; \mathbf{z}, c)$ represents the (f,n)–th element of the decoder output $\sigma_\theta^2(\mathbf{z}, c)$ and g is the global scale of the generated spectrogram. During CVAE training, both the encoder and decoder network parameters ϕ and θ are trained using the following objective function:

$$\begin{aligned}\mathcal{J}(\phi, \theta; \tilde{\mathbf{S}}, c) = {} & \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\tilde{\mathbf{S}},c)}[\log p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c)] \\ & - \mathrm{KL}[q_\theta(\mathbf{z}|\tilde{\mathbf{S}}, c)||p(\mathbf{z})], \quad (17)\end{aligned}$$

where $p(\mathbf{z})$ is a standard Gaussian distribution and $\mathrm{KL}[\cdot||\cdot]$ is the Kullback-Leibler divergence.

The trained decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ can be used as a generative model capable of generating spectrograms of all the sources involved in the training examples.

C. PARAMETER ESTIMATION

Since the decoder distribution is designed to be of the same form as the LGM, using $p_\theta(\tilde{\mathbf{S}}_j|\mathbf{z}_j, c_j, g_j)$ leads to the same log-likelihood as (7). Thus, we can derive an iterative algorithm for estimating $\mathcal{Z} = \{\mathbf{z}_j\}_j$, $\mathcal{C} = \{c_j\}_j$, $\mathcal{G} = \{g_j\}_j$ and $\mathcal{R}$ in the same manner as the derivation of an MM algorithm for MNMF.

The MM algorithm is an iterative algorithm that searches for a stationary point of a cost function by iteratively minimizing an auxiliary function called a "majorizer" that is guaranteed to never go below the objective function. When constructing an MM algorithm for a certain minimization problem, the main issue is to design the majorizer. If a majorizer is properly designed, the algorithm is guaranteed to converge to a stationary point of the cost function. If we can build a tight majorizer/minorizer that is easy to optimize, we can generally expect to obtain a fast-converging algorithm.

As shown in a previous study [26], we can build a majorizer $\mathcal{L}^+$ for the negative log-likelihood function $\mathcal{L} = -\log p(\mathcal{X}|\mathcal{A}, \mathcal{V})$ using the right side of the following inequality:

$$\mathcal{L} = -\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \overset{c}{\leq} \sum\nolimits_j \sum\nolimits_{f,n} \Bigg[\frac{\mathrm{tr}(\mathbf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f)\mathbf{P}_j(f,n))}{g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)} + g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\mathbf{Q}^{-1}(f,n)\mathbf{R}_j(f))\Bigg], \tag{18}$$

where $\overset{c}{\leq}$ denotes the inequality that holds when constant terms are ignored. The equality holds when the auxiliary variables $\mathcal{P} = \{\mathbf{P}_j(f,n)\}_{j,f,n}$ and $\mathcal{Q} = \{\mathbf{Q}(f,n)\}_{f,n}$ are given by

$$\mathbf{P}_j(f,n) = g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathbf{R}_j(f) \times \left(\sum\nolimits_j g_j\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathbf{R}_j(f)\right)^{-1}, \tag{19}$$

$$\mathbf{Q}(f,n) = \hat{\mathbf{X}}(f,n). \tag{20}$$

An iterative algorithm that consists of minimizing this majorizer with respect to $\mathcal{Z}$, $\mathcal{C}$, $\mathcal{G}$, and $\mathcal{R}$ and updating $\mathcal{P}$ and $\mathcal{Q}$ using (19) and (20) is guaranteed to not increase the negative log-likelihood $\mathcal{L}$. The optimal update of $\mathcal{R}$ is analytically obtained as

$$\mathbf{R}_j(f) \leftarrow \Lambda_j^{-1}(f)\#(\mathbf{R}_j(f)\Omega_j(f)\mathbf{R}_j(f)), \tag{21}$$

where # denotes the geometric mean of two positive definite matrices [29]:

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}. \tag{22}$$

$\Lambda_j(f)$, $\Omega_j(f)$ are given as follows:

$$\Lambda_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathbf{X}}^{-1}(f,n), \tag{23}$$

$$\Omega_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n). \tag{24}$$

Since the majorizer is split into source-wise terms, $\mathcal{Z}$ and $\mathcal{C}$ can be updated in parallel using backpropagation. Since the sum-to-one constraints for c_j must be taken into account, this can be easily implemented by inserting an appropriately designed softmax layer that outputs c_j:

$$c_j = \mathrm{softmax}(e_j), \tag{25}$$

and treating e_j as the parameter to be estimated instead. The optimal update of $\mathcal{G}$ is obtained as follows:

$$g_j \leftarrow g_j \times \sqrt{\frac{\sum_{f,n}\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{f,n}\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathbf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}. \tag{26}$$

D. REGULARIZATION OF z AND c

In CVAE pretraining, the encoder is trained so that the distribution of the latent variable $\mathbf{z}$ becomes close to a standard Gaussian distribution. Thus, to let the trained decoder produce spectrograms that resemble those seen in the training data, $\mathbf{z}$ must not deviate from the assumed distribution. To prevent $\mathbf{z}$ from deviating from a standard Gaussian distribution, we consider introducing regularization for $\mathbf{z}_j$ given by

$$\mathcal{L}_{\mathcal{Z}} = -\sum\nolimits_j \log p(\mathbf{z}_j), \tag{27}$$

where $p(\mathbf{z}_j) = \mathcal{N}(\mathbf{z}_j; \mathbf{0}, \mathbf{I})$.

For the optimization of the latent code c, the resulting $c_1, \ldots, c_J$ must be disjoint since the class of each source is usually different. To promote the orthogonality between $c_1, \ldots, c_J$, we use the following regularization term:

$$\mathcal{L}_{\mathcal{C}} = \|\mathbf{C}\mathbf{C}^{\mathsf{T}} - \mathbf{I}\|_1, \tag{28}$$

where $\mathbf{C} \in [0,1]^{J\times L}$ is a matrix composed of J latent codes (L-dimensional vectors) and $\mathbf{I} \in \mathbb{R}^{J\times J}$ is an identity matrix. This regularization term plays the role of encouraging each latent code c_j to become a different one-hot vector.

Thus, the objective function for $\mathcal{Z}$ and $\mathcal{C}$ is given as

$$\mathcal{I} = \mathcal{L}^{+} + \lambda_{\mathcal{Z}}\mathcal{L}_{\mathcal{Z}} + \lambda_{\mathcal{C}}\mathcal{L}_{\mathcal{C}}, \tag{29}$$

where $\lambda_{\mathcal{Z}} \geq 0$ and $\lambda_{\mathcal{C}} \geq 0$ are weight parameters.

E. SEPARATION PROCESS

After convergence, we can obtain separated source signals by applying a multichannel Wiener filter:

$$\hat{\mathbf{s}}_j(f,n) = v_j(f,n)\mathbf{R}_j(f)\left(\sum\nolimits_j v_j(f,n)\mathbf{R}_j(f)\right)^{-1}\mathbf{x}(f,n), \tag{30}$$

followed by applying the inverse STFT.

F. ADVANTAGES OVER RELATED WORK

The GMVAE method has several important advantages. First, it provides the flexibility of allowing it to adapt to different scenarios. A typical case is that in which we know which sources are present in a mixture. In this case, we can simply fix c_j at the corresponding one-hot vector and run the iteration (Algorithm 1). Another case is that in which we are given no information about the sources. It may appear that the GMVAE method works only in supervised and informed scenarios where audio samples of all the sources in a test mixture are included in the training set. However, thanks to the CVAE-based source modeling, if the training set contains a wide enough variety of sources, the GMVAE

Algorithm 1 Fully informed GMVAE

```
Train $\phi$ and $\theta$ with (17)
for each $j$ do
    Fix $c_j$ at a specific one-hot vector
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ with (18) using backpropagation
    Update $\mathcal{G}$ using (26)
    Update $\mathcal{R}$ using (21)
until converge
```

method can work in nearly blind settings where there is no information about which of the sources are present in a test mixture and can even handle sources that are unseen in the training set. For such cases, one simple way would be to treat c_j as a free parameter, initialized for example at a uniform distribution (i.e., $[1/L, \ldots, 1/L]$), and run the iteration until convergence (Algorithm 2). For semi-supervised speech enhancement scenarios where only the source to be enhanced is known, we can simply specify (instead of having it estimate) one of the latent codes (Algorithm 3).

Algorithm 2 Uninformed GMVAE

```
Train $\phi$ and $\theta$ with (17)
for each $j$ do
    Initialize $c_j$ at a uniform distribution
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $\mathcal{C}$ with (18) using backpropagation
    Update $\mathcal{G}$ using (26)
    Update $\mathcal{R}$ using (21)
until converge
```

Algorithm 3 Partially informed GMVAE

```
Train $\phi$ and $\theta$ with (17)
Initialize $c^{\mathrm{Target}}$ at a specific one-hot vector
Initialize $c^{\mathrm{Non-target}}$ at a uniform distribution
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $c^{\mathrm{Non-target}}$ with (18) using backpropagation
    Update $\mathcal{G}$ using (26)
    Update $\mathcal{R}$ using (21)
until converge
```

Second, the CVAE modeling can potentially have a certain effect in avoiding local optima problems in supervised and semi-supervised scenarios. One possible situation in these scenarios that can lead to poor local optima is when the source index pre-assigned to each $v_j(f, n)$ is different from the source to which the estimate of $\mathbf{R}_j(f)$ corresponds most closely. Once this kind of mismatch occurs, it usually becomes difficult to avoid getting stuck in incorrect local optima. This is one of telling examples of the problem that is very likely to occur when the source index is pre-specified for each j. It should be noted that supervised MNMF and VAE-NMF fall into this type of method. With the GMVAE method, however, we can take a soft-decision approach by treating c_j as a free parameter (instead of specifying it), initialized as a uniform distribution, and let the algorithm find the best c_j so that the distribution of the source to which $\mathbf{R}_j(f)$ is likely to correspond can be estimated along with $\mathbf{R}_j(f)$. We can then determine the index $\hat{j}$ that corresponds to the source of interest from inspection of $c_1, \ldots, c_J$ and forcing $c_{\hat{j}}$ to the corresponding one-hot vector during the iteration (Algorithm 4).

Algorithm 4 GMVAE with one-hot enforcement

```
Train $\phi$ and $\theta$ with (17)
for each $j$ do
    Initialize $c_j$ with a specific one-hot vector
end for
Initialize $\mathcal{Z}$, $\mathcal{G}$, and $\mathcal{R}$
repeat
    Update $\mathcal{Z}$ and $\mathcal{C}$ with (18) using backpropagation
    Update $\mathcal{G}$ using (26)
    Update $\mathcal{R}$ using (21)
until converge
Determine $c_j$ which is the most similar to the target as $c^{\mathrm{Target}}$
Determine $c_{j'} (j \neq j')$ as $c^{\mathrm{Non-target}}$
Update $c^{\mathrm{Target}}$ with a specific one-hot vector
repeat
    Update $\mathcal{Z}$ and $c^{\mathrm{Non-target}}$ with (18) using backpropagation
    Update $\mathcal{G}$ using (26)
    Update $\mathcal{R}$ using (21)
until converge
```

V. EXPERIMENTAL EVALUATION

A. EXPERIMENTAL SETTINGS

We conducted three experiments to evaluate the GMVAE method. The first two are speaker-closed and speaker-open underdetermined source separation experiments where the task is to separate out three sources from their mixtures captured by two microphones. The other is a semi-supervised speech enhancement experiment where the task is to extract a known source from noisy observations contaminated by unknown sources. As the experimental data, we used audio samples from the Voice Conversion Challenge (VCC) 2018 dataset [30], which contains recordings of 6 female and 6 male U.S. English speakers. The average duration of each utterance is 3.5 seconds, and the dataset includes 81 utterances of individual speakers for training and 35 utterances for evaluation. For these experiments, we used the utterances of four female and four male speakers, 'SF1', 'SF2', 'SF3', 'SF4', 'SM1', 'SM2', 'SM3', and 'SM4'. For training, we used 100 utterances of 'SF1', 'SF2', 'SM1', and 'SM2'. Another 10 utterances of 'SF1', 'SF2', 'SM1', and 'SM2' were used for evaluation under speaker-closed conditions in the source separation task and treated as the target sources in the speech enhancement task. Similarly, 10 utterances of 'SF3', 'SF4', 'SM3', and 'SM4' were used for evaluation under speaker-open conditions in the source separation task and treated as the interference sources in the speech enhancement task.

Figure 2 shows the configuration of the room used for the experiments. Reverberation time T_{60} was set to

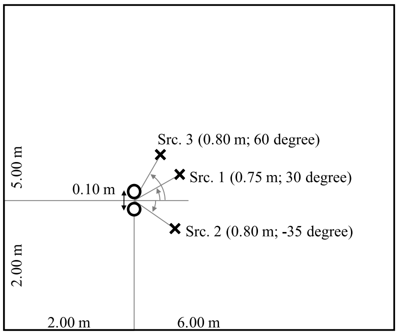

Src. 3 (0.80 m; 60 degree)
Src. 1 (0.75 m; 30 degree)
Src. 2 (0.80 m; -35 degree)
0.10 m
5.00 m
2.00 m
2.00 m
6.00 m

FIGURE 2. Configuration of room used for our experiments, where ○ and × are locations of microphones and sound sources, respectively.

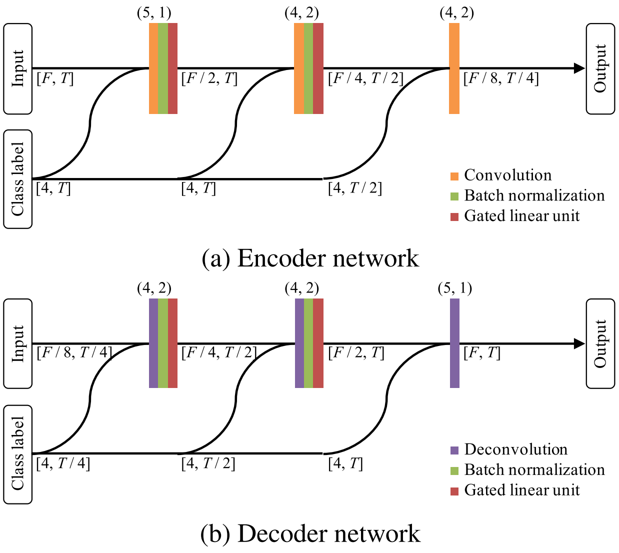
TABLE 2. Methods for comparison.

(a) Source separation task

Notation	Method	Initialization
Baseline1	Unsupervised uninformed MNMF [6] (Equation (11))	–
Baseline2	Unsupervised uninformed MNMF [6] (Equation (12))	–
Baseline3	Fully-supervised uninformed MNMF [6]	–
Baseline4	Fully-supervised fully-informed MNMF [6]	–
Proposed1	Uninformed GMVAE (Algorithm 2)	Baseline1
Proposed2	Uninformed GMVAE (Algorithm 2)	Baseline2
Proposed3	Uninformed GMVAE (Algorithm 2)	Baseline3
Proposed4	Fully informed GMVAE (Algorithm 1)	Baseline4

(b) Speech enhancement task

Notation	Method	Initialization
Baseline5	Semi-supervised partially-informed MNMF [6]	–
Baseline6	VAE-NMF [17]	Baseline5
Proposed5	Partially informed GMVAE (Algorithm 3)	Baseline5
Proposed6	GMVAE with one-hot enforcement (Algorithm 4)	Baseline5



(a) Encoder network

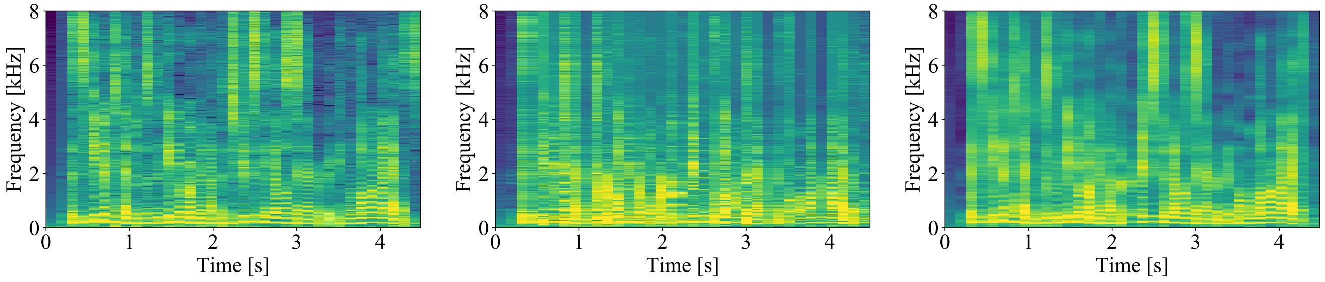
(b) Decoder network

FIGURE 3. Network configurations of (a) encoder and (b) decoder, where [c, t] denotes input channel and input length. Both convolution and deconvolution represent 1-dimensional operation. (k, s) represent kernel size and stride size along frame, respectively.

78 and 351 ms. In the source separation task, we created test data using all possible combinations of three speakers for both the speaker-closed and speaker-open conditions. For each set of speakers, 10 speech mixtures were generated by randomly choosing the utterances and randomly allocating them at locations indicated in Figure 2. In the speech enhancement task, 40 speech mixtures were generated by randomly choosing the utterances of the target and interference speakers where target and interference sources are located at the Src. 1 and Src. 2, respectively.

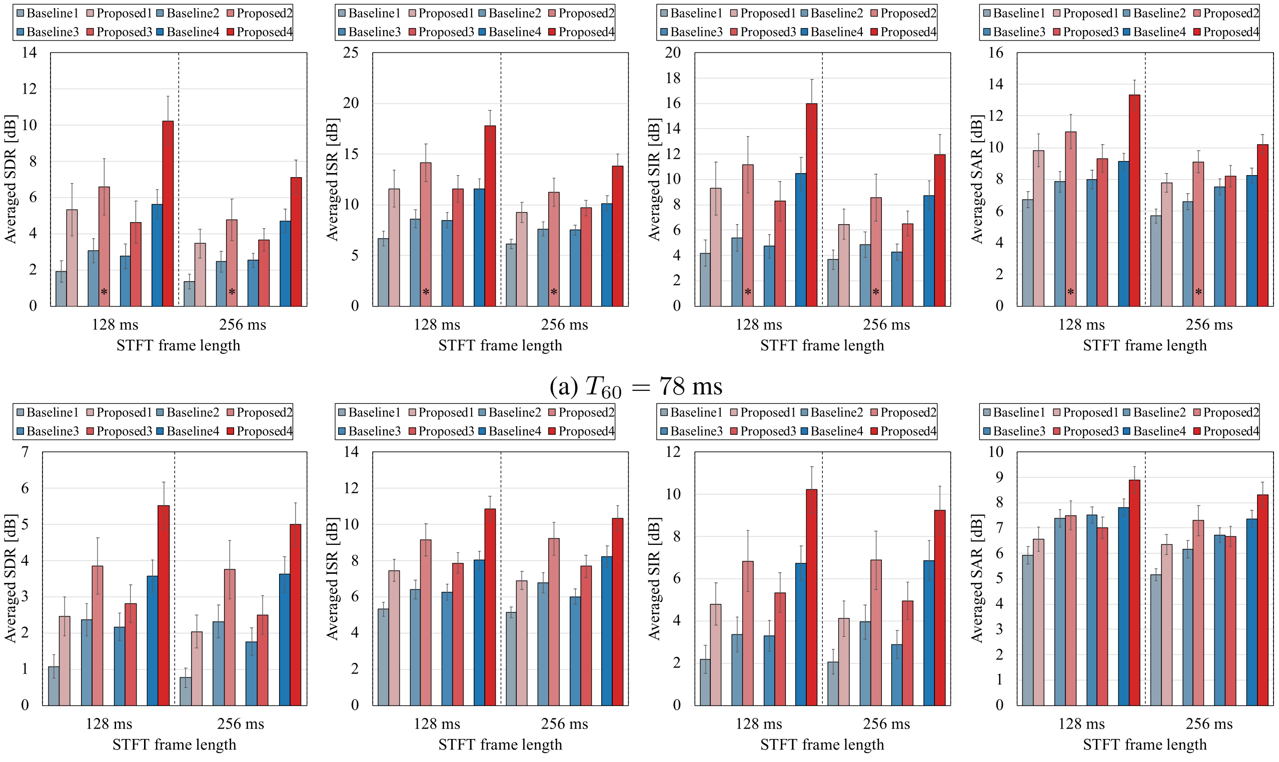
We tested several different versions of the proposed and baseline methods for comparison. We use the terms "fully supervised/semi-supervised/unsupervised" and "fully informed/partially informed/uninformed" to properly categorize each version of the methods. Fully supervised, semi-supervised, and unsupervised refer to whether a method requires training examples and fully informed, partially informed, and uninformed refer to how much information about which sources are present in a test mixture is given to a method. All versions of the GMVAE method are fully supervised since they all require training examples to train the CVAE. Thus, we omit "fully supervised" when referring to this method. At separation time, the GMVAE method can be implemented in either fully informed, uninformed, or partially informed manners. Hence, we refer to these versions as fully informed GMVAE, uninformed GMVAE, and partially informed GMVAE. MNMF can perform in either unsupervised, semi-supervised, or fully supervised manners. We implemented unsupervised uninformed, fully supervised uninformed, and fully supervised fully informed MNMFs for comparison. VAE-NMF falls into the semi-supervised partially informed category. Categorization of each version is summarized in Table 2.

All the speech signals were resampled at 16 kHz. We tested two different STFT configurations, i.e., a 128-ms window length with a 64-ms shift length and a 256-ms window length with a 128-ms shift length. The numbers of basis spectra for these baseline versions were set to 10 per speaker, as in a previous study [4]. The spectral dictionaries used for the fully/semi-supervised MNMF versions were trained for each speaker using the same dataset used for the CVAE training and obtained using an Itakura-Saito NMF (IS-NMF) [11] with 1000 iterations. For a fair comparison, MNMF was run for 200 iterations for the initialization of each method. All the versions, including the baseline ones, were then run for 100 iterations. For the speech enhancement task, we implemented Algorithm 4, which consists of updating $c_1, \ldots, c_J$ freely during the first 50 iterations, then searching for the index $\hat{j}$ that corresponds to the target speaker, and finally running the last 50 iterations while fixing $c_{\hat{j}}$ at the corresponding one-hot vector. We refer to this algorithm as "GMAVE with one-hot enforcement". The encoder and decoder networks of our CVAE are shown in Figure 3. At training time, the batch size and length were set to 9 and 128, respectively. The Adam algorithm [31] with a learning rate of 0.0001 was used for the CVAE pretraining. The number of training epochs was set to 1000. The VAE used with VAE-NMF was trained for each



(a) Reference (b) MNMF (c) GMVAE

FIGURE 4. Spectrograms of (a) reference source and estimated sources by using (b) MNMF and (c) GMVAE.



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

FIGURE 5. Source separation performances under speaker-closed conditions.

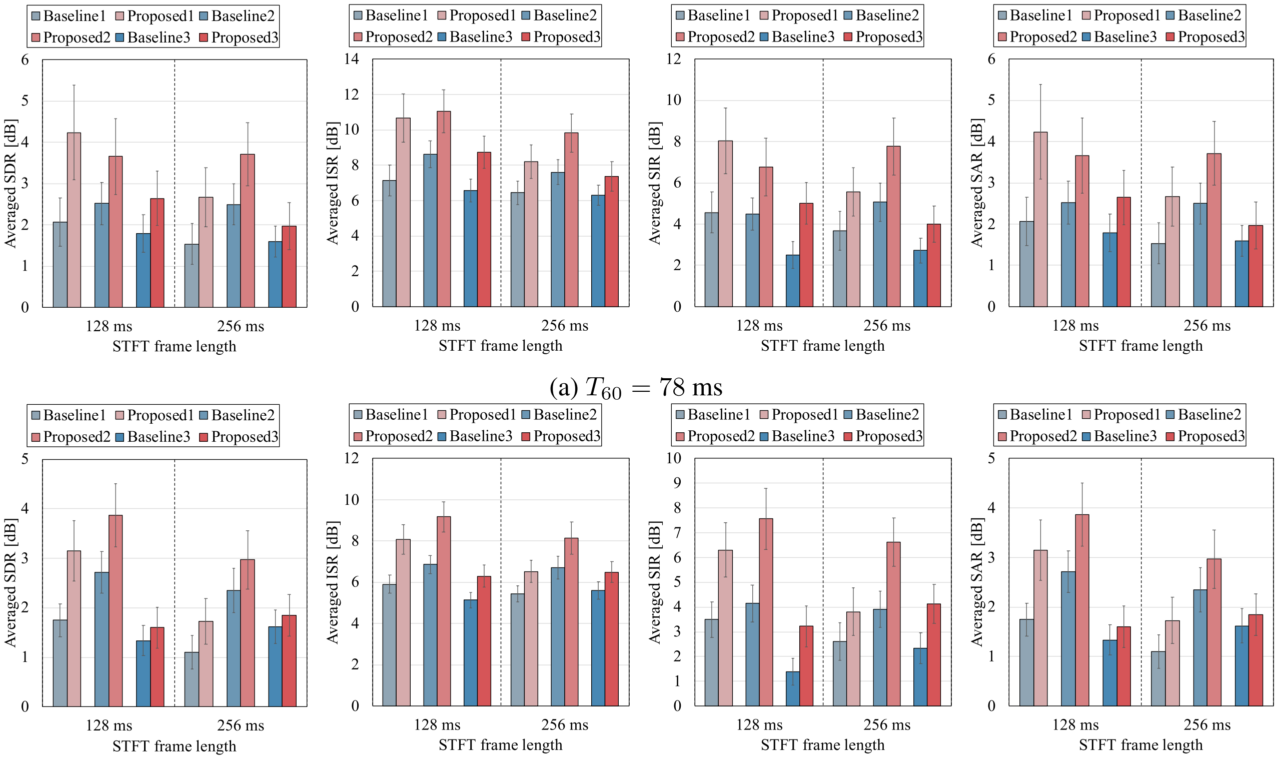
speaker using the same training dataset and training configuration, where the same network architectures as our CVAE except for the conditioning part were used. At separation time, the Adam algorithm with a learning rate of 0.01 was used for updating $\mathcal{Z}$ and $\mathcal{C}$. The number of training epochs per iteration was set to 10.

As the evaluation metrics, we used the averages of the signal-to-distortion ratio (SDR), source image-to-spatial distortion ratio (ISR), signal-to-inference ratio (SIR), and signal-to-artifact ratio (SAR) [32] between the reference signals and separated signals. Note that, in the speech enhancement task, separation performances of both the target source and interference source were evaluated and permutation of estimated sources was not considered in the evaluation.

B. EXPERIMENTAL RESULTS

Figures 4 (b) and (c) show examples of the NMF- and CVAE-based source models fitted to the speech spectrogram shown in Figure 4 (a). As these examples show, the CVAE source model was able to express harmonic structures and higher-frequency components better than the NMF model.

We next show the performances in the source separation task. A comparison of the separation performance of each version under speaker-closed conditions is shown in Figure 5, where error bars show the 95 % confidence intervals. When comparing the performance of the uninformed versions (Baseline1 to Baseline3 and Proposed1 to Proposed3) at $T_{60} = 78$ ms, the proposed versions outperformed the baseline ones for both STFT configurations. The comparison of Baseline3 and Proposed3 directly reflects



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

FIGURE 6. Source separation performances under speaker-open conditions.

the difference in ability between the NMF- and CVAE-based source models. The results thus indicate the superiority of our CVAE source model over the NMF counterpart. The comparison between Proposed1, 2 &, 3 indicates that initialization can affect separation performances. It indicates that using Baseline2 for initialization worked better than using Baseline1 & 3. Focusing on the comparison of the fully informed versions (Baseline4 and Proposed4), Proposed4 significantly outperformed Baseline4 and achieved the best performance. This indicates that the prior information for the sources in a target mixture can contribute to improving performance. Although the performances of all of the versions degraded for the longer reverberant condition ($T_{60} = 351$ ms), the proposed versions still performed better than the baseline ones. The comparisons between the performances obtained with the two STFT configurations showed that using a 128-ms frame length worked better, especially for a shorter reverberant condition.

A comparison of the separation performance of each method under speaker-open conditions is shown in Figure 6, where the fully informed versions (Baseline4 and Proposed4) are omitted since all the sources in the mixture are unseen in the training data. We can confirm from the comparisons between Baseline1 and Proposed1, Baseline2 and Proposed2, and Baseline3 and Proposed3 that the proposed versions consistently performed better than the baseline ones, especially in terms of the SIR metric. This may imply the ability of the GMVAE method to estimate the spectrogram of each source accurately, leading to an accurate estimation of its spatial covariance. Another interesting finding from these results is that the GMVAE method can perform reasonably well under speaker-open conditions even though it is a method that requires supervisions. We also confirmed that unlike under the speaker-closed conditions, using a 128-ms STFT frame length was more robust against varying reverberation conditions than using longer frame lengths.

Table 3 shows an ablation study on Proposed2, where the best performances are denoted in bold font and the last columns correspond to the separation performances denoted as $*$ in Figure 5. These results indicate that each regularization technique improved the separation performance, and Proposed2 using both regularizations achieved the best performance. These results also indicate that the regularizations were effective, especially when the STFT frame length was 128 ms. Figure 7 shows examples of the estimated $\mathcal{Z}$ and $\mathcal{C}$ without and with the regularizations, where the histograms represent $\mathcal{Z}$ at initialization step and separation step. Estimated $\mathcal{C}$ is also shown in the figure. We can confirm that the regularization for $\mathcal{Z}$ prevented $\mathcal{Z}$ from deviating from a standard Gaussian distribution, and the regularization for $\mathcal{C}$ promoted the orthogonality of $\mathcal{C}$.

We finally show the performances of the speech enhancement task. A comparison of the enhancement performances

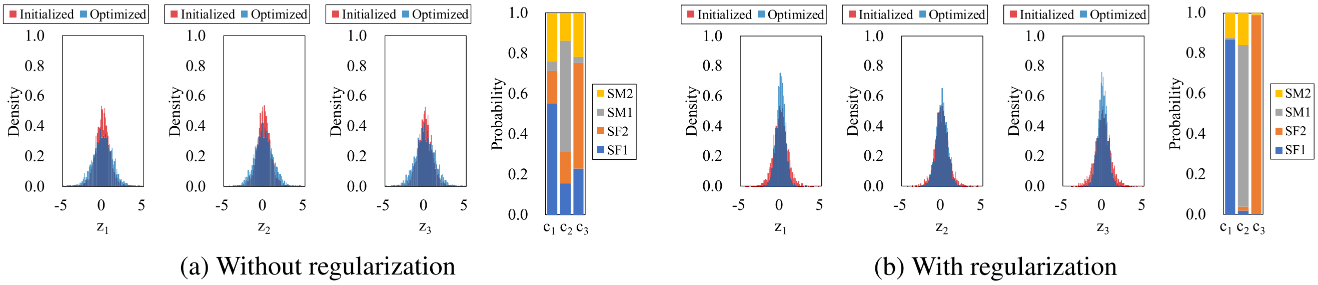
TABLE 3. Ablation study on Proposed2 under speaker-closed conditions at T_{60} = 78 [ms].

(a) 128-ms STFT frame length

$\mathcal{L}_{\mathcal{Z}}$	$\mathcal{L}_{\mathcal{C}}$	Avg. SDR	Avg. ISR	Avg. SIR	Avg. SAR
✗	✗	5.93	13.45	10.34	10.45
✗	✓	6.13	13.69	10.50	10.62
✓	✗	6.43	14.02	10.99	10.78
✓	✓	**6.59**	**14.16**	**11.14**	**11.02**

(b) 256-ms STFT frame length

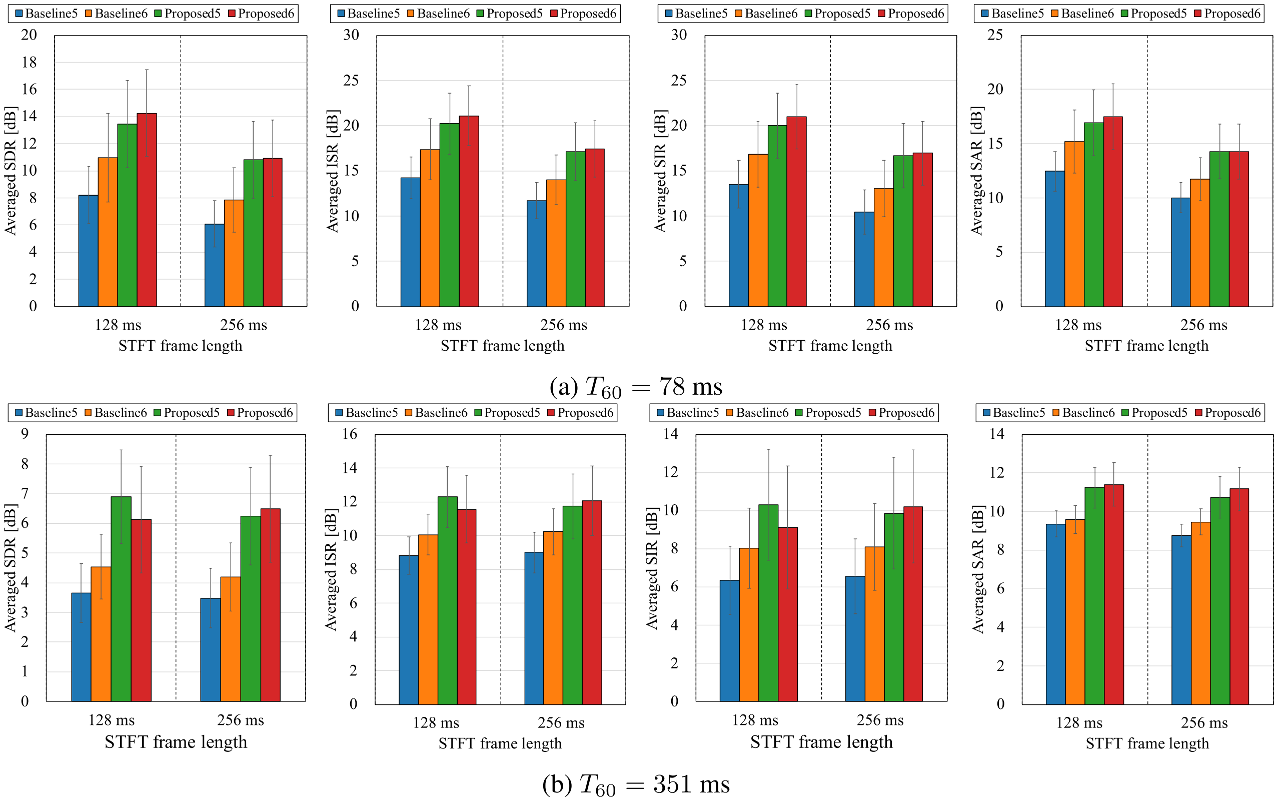
$\mathcal{L}_{\mathcal{Z}}$	$\mathcal{L}_{\mathcal{C}}$	Avg. SDR	Avg. ISR	Avg. SIR	Avg. SAR
✗	✗	4.46	10.80	8.12	8.73
✗	✓	4.56	10.92	8.24	8.84
✓	✗	4.63	11.08	8.41	8.95
✓	✓	**4.78**	**11.24**	**8.56**	**9.11**



(a) Without regularization

(b) With regularization

FIGURE 7. Regularization effects on $\mathcal{Z}$ and $\mathcal{C}$.



(a) $T_{60} = 78$ ms

(b) $T_{60} = 351$ ms

FIGURE 8. Speech enhancement performances.

of Baseline5, Baseline6, Proposed5, and Proposed6 is shown in Figure 8. Comparisons among Baseline5, Baseline6, and Proposed5 revealed that Proposed5 outperformed the baseline versions and performed better than VAE-NMF. Moreover, Proposed6 performed better than the other versions particularly under the small reverberant condition. This shows a certain effect of the one-hot enforcement process adopted in Algorithm 4.

VI. CONCLUSION

We proposed the GMVAE method, a generalized version of the MVAE method that can also deal with underdetermined cases. We developed a convergence-guaranteed parameter optimization algorithm for the GMVAE method by combining the ideas for the parameter optimization processes introduced in MNMF and the MVAE method. We further introduced two regularization techniques for avoiding undesirable solutions and presented several algorithms designed for fully informed, partially informed, and uninformed source separation and speech enhancement tasks. Our experimental results revealed that the proposed GMVAE method outperformed MNMF in source separation tasks and VAE-NMF in speech enhancement tasks, demonstrating the advantage of the CVAE source model. The results also indicate that the GMVAE method can perform reasonably well even under speaker-open conditions.

REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, vol. 46. Hoboken, NJ, USA: Wiley, 2004.

[2] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.*, 2006, pp. 165–172.

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

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

[5] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, 2010, pp. 245–253.

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

[7] 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, Language Process.*, vol. 24, no. 9, pp. 1622–1637, Sep. 2016.

[8] K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii, "Student's t multichannel nonnegative matrix factorization for blind source separation," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement*, Sep. 2016, pp. 1–5.

[9] S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, and N. Ono, "Independent low-rank matrix analysis based on complex student's t-distribution for blind audio source separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, Sep. 2017, pp. 1–6.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, no. 3, Oct. 2003, pp. 177–180.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*. Cham, Switzerland: Springer, 2018, pp. 125–155.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[14] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," 2018, *arXiv:1808.00892*. [Online]. Available: https://arxiv.org/abs/1808.00892

[15] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.

[16] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Nov. 2018, pp. 1233–1239.

[17] 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.*, Feb. 2019, pp. 101–105.

[18] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," 2018, *arXiv:1810.00223*. [Online]. Available: https://arxiv.org/abs/1810.00223

[19] K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1973–1977.

[20] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[21] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2005, pp. 78–81.

[22] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.*, 2009, pp. 775–782.

[23] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, 2011, pp. 162–185.

[24] J. De Leeuw and W. J. Heiser, "Convergence of correction matrix algorithms for multidimensional scaling," in *Geometric Representations of Relational Data*. Ann Arbor, MI, USA: Mathesis Press, 1977, pp. 735–752.

[25] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[26] H. Kameoka, H. Sawada, and T. Higuchi, "General formulation of multichannel extensions of NMF variants," in *Audio Source Separation*. Cham, Switzerland: Springer, 2018, pp. 95–124.

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

[28] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.

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

[30] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," 2018, *arXiv:1804.04262*. [Online]. Available: https://arxiv.org/abs/1804.04262

[31] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

SHOGO SEKI received the B.E. degree in engineering and the M.E. degree in information science from Nagoya University, Nagoya, Japan, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree. His research interests include audio, speech, and music signal processing. He is a Student Member of the Acoustical Society of Japan and the Institute of Electronics, Information and Communication Engineers. He received the Acoustical Society of Japan 2016 Student Presentation Award.

HIROKAZU KAMEOKA received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 2002, 2004, and 2007, respectively. He is currently a Research Scientist with the NTT Communication Science Laboratories and a Visiting Associate Professor with The University of Tokyo. His research interests include computational auditory scene analysis, statistical signal processing, speech and music processing, and machine learning. He is a member of the Information Processing Society of Japan (IPSJ) and the Acoustical Society of Japan (ASJ). He received 13 awards over the past ten years, including the Yamashita Memorial Research Award, in 2005, from IPSJ, the Itakura Prize Innovative Young Researcher Award, in 2007, and the Awaya Prize Young Researcher Award, in 2008, from ASJ, the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award, in 2009, and the IEICE ISS Young Researcher's Award in Speech Field, in 2011.

LI LI received the B.E. degree from the Shanghai University of Finance and Economics, China, in 2014, and the M.S. degree from the University of Tsukuba, Japan, in 2018. She is currently pursuing the Ph.D. degree with the Graduate School, University of Tsukuba. She has been a Research Fellow of the Japan Society of Promotion of Science, since 2018. Her research interests include audio and speech signal processing, source separation, and machine learning. She received the 13th Student Presentation Award from the Acoustical Society of Japan and the second IEEE Signal Processing Society Tokyo Joint Chapter Student Award.

TOMOKI TODA received the B.E. degree from Nagoya University, Japan, in 1999, and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science, from 2003 to 2005. He was an Assistant Professor, from 2005 to 2011, and an Associate Professor, from 2011 to 2015, with NAIST. Since 2015, he has been a Professor with the Information Technology Center, Nagoya University. His research interests include statistical approaches to speech and audio processing. He received more than ten paper/achievement awards, including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (*Speech Communication* Journal).

KAZUYA TAKEDA received the B.E. and M.E. degrees in electrical engineering and the D.Eng. degree from Nagoya University, Nagoya, Japan, in 1983, 1985, and 1994, respectively. From 1986 to 1989, he was with the Advanced Telecommunication Research Laboratories (ATR), Osaka, Japan. He was a Visiting Scientist with MIT, from November 1987 to April 1988. From 1989 to 1995, he was a Researcher and a Research Supervisor with the KDD Research and Development Laboratories, Kamifukuoka, Japan. From 1995 to 2003, he was an Associate Professor of the Faculty of Engineering, Nagoya University. Since 2003, he has been a Professor with the Department of Media Science, Graduate School of Information Science, Nagoya University. His main research interest is at ATR was corpus-based speech synthesis. His current research interests include media signal processing and its applications, which include spatial audio, robust speech recognition, and driving behavior modeling.