ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746980

INVESTIGATION AND COMPARISON OF OPTIMIZATION METHODS FOR VARIATIONAL AUTOENCODER-BASED UNDERDETERMINED MULTICHANNEL SOURCE SEPARATION

Shogo Seki[1], *Hirokazu Kameoka*[1], *Li Li*[1,2]

[1]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan
[2]Information Technology Center, Nagoya University, Japan

ABSTRACT

In this paper, we investigate two algorithms for variational autoencoder (VAE)-based underdetermined multichannel source separation. We previously extended the multichannel VAE (MVAE) method for determined multichannel source separation and proposed the generalized MVAE (GMVAE) method for underdetermined multichannel source separation. The GMVAE method employs a conditional VAE (CVAE) as the source model representing the power spectrograms of the underlying sources present in a mixture. While we developed a convergence-guaranteed parameter estimation algorithm using a majorization-minimization/minorization-maximization (MM) algorithm, an expectation-maximization (EM) algorithm also allows us to design another algorithm with the same property. However, a comparison of the MM-based and EM-based algorithms has not yet been revealed. To elucidate this, we investigate the MM-based and EM-based algorithms for the GMVAE method, using an improved CVAE variant called auxiliary classifier VAE (ACVAE). The experimental results suggest that the EM-based algorithm takes less computational cost, achieving comparable separation performance with the MM-based algorithm.

Index Terms— Underdetermined multichannel source separation, variational autoencoder, convergence-guaranteed algorithm

1. INTRODUCTION

Source separation refers to the problem of separating underlying source signals present in a mixture signal observed by a microphone array. Source separation can contribute to helping other tasks, e.g., automatic speech recognition under a multi-speaker conversation situation and speaker diarization of an over-lapped speech.

Source separation is typically tackled using a frequency-domain approach with the various source signal and/or array responses models. As one of the popular approaches for source separation problems, independent component analysis (ICA) [1] is a well-established method where the separation system is assumed to exist. While ICA employs an assumption that a mixing process generating a multichannel mixture signal from source signals is invertible, full-rank spatial covariance analysis (FCA) [2] is known as one of the more flexible methods. Unlike ICA, FCA assumes the mixing system instead of the separation system and can deal with an underdetermined case where the sources outnumber the microphones. Since both ICA and FCA perform frequency-wise source separation and all the model parameters are independent among frequencies, it is necessary to solve permutation indeterminacy that occurred in frequency-domain approaches [3].

This work was partly supported by JST, CREST Grant Number JPMJCR19A3, Japan.

To address the permutation indeterminacy, several methods attempting to have part of model parameters shared among frequencies have been developed [4, 5], and non-negative matrix factorization (NMF) [6] is regarded as the generalization. NMF was originally applied for a music transcription task [7], where the power spectrum of a mixture signal observed at each time frame is approximated by the sum of a fixed number of basis spectra scaled by time-varying magnitudes. Multichannel NMF (MNMF) is an FCA variant that incorporates the NMF concept into the power spectrogram modeling of each source [8–10], and independent low-rank matrix analysis (ILRMA) is the determined version that introduces the NMF source model on ICA [11, 12]. Using spectral templates as acoustical clues, MNMF and ILRMA jointly perform frequency-wise source separation and permutation alignment. However, they can fail to work when encountering sound sources with spectrograms that do not follow the NMF source model, resulting in performance limitations.

Recently, generative approaches using deep neural networks (DNNs) have been proposed to model source spectrograms more flexibly than the NMF source model [13–20], where a variational autoencoder (VAE) [21] plays a central role. For determined multichannel source separation, a method called the multichannel VAE (MVAE) method using a conditional VAE (CVAE) [22] as the source model has been proposed [18]. The MVAE method demonstrated that the CVAE source model is better than the NMF source model at expressing the spectrogram of each source and correctly discriminating the spectrogram of one source from that of another. Motivated by the great success of the MVAE method, underdetermined counterpart, the generalized MVAE (GMVAE) method, was subsequently proposed [19].

Similar to the MVAE method, a convergence-guaranteed parameter estimation algorithm of the GMVAE method was developed, using an majorization-minimization/minorization-maximization (MM) algorithm [23]. On the other hand, an expectation-maximization (EM) algorithm [24] is also known as an iterative algorithm that keeps increasing a log-likelihood function. Hence, an EM algorithm allows us to design another convergence-guaranteed parameter estimation algorithm. A comparison of different algorithms for a same objective typically help in choosing the suitable approach and in considering further developments. Although comparisons of MM-based and EM-based algorithms for FCA and MNMF have been reported in [10] and in [25], such comparison for the GMVAE method has not yet been revealed.

To elucidate this, we study the MM-based and EM-based convergence-guaranteed parameter estimation algorithms for the GMVAE method. These algorithms are investigated from the perspectives of computational cost and source separation performance. Through the investigation, we employs an improved CVAE variant called auxiliary classifier VAE (ACVAE), which has achieved success in several tasks [20, 26].

2. PROBLEM FORMULATION

Suppose that there are J source signals and that I microphones receive a mixture signal. Let $s_j(f,n)$ and $x_i(f,n)$ be the short-time Fourier transform (STFT) coefficients of the j-th source signal and the mixture signal at i-th microphone, where f and n are the frequency and time indices, respectively. We denote the vectors composed of the STFT coefficients of all the sources and the microphones as $\mathbf{s}(f,n) = [s_1(f,n), s_2(f,n), \ldots, s_J(f,n)]^{\mathsf{T}} \in \mathbb{C}^J$ and $\mathbf{x}(f,n) = [x_1(f,n), x_2(f,n), \ldots, x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$, where $(\cdot)^{\mathsf{T}}$ represents the transpose and $\mathbb{C}$ denotes the set of complex numbers. We begin by employing the local Gaussian modeling (LGM) [27], which assumes that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance, i.e., power spectral density (PSD), $v_j(f,n)(= \mathbb{E}[|s_j(f,n)|^2])$:

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(0, v_j(f,n)). \tag{1}$$

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:

$$\mathbf{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{V}(f,n)), \tag{2}$$

where $\mathbf{V}(f,n)$ is a diagonal matrix whose diagonal entries are $v_1(f,n), v_2(f,n), \ldots, v_J(f,n)$. In a general situation where J can outnumber I, a mixing system is given by:

$$\mathbf{x}(f,n) = \mathbf{A}(f)\mathbf{s}(f,n), \tag{3}$$

where $\mathbf{A}(f) = [\mathbf{a}_1(f), \mathbf{a}_2(f), \ldots, \mathbf{a}_J(f)] \in \mathbb{C}^{I \times J}$ is referred to as a mixing matrix. From Eqs. (2) and (3), $\mathbf{x}(f,n)$ is shown to follow the complex Gaussian distribution with a zero-mean vector $\mathbf{0}$ and a covariance $\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f,n)$, where $(\cdot)^{\mathsf{H}}$ represents the conjugate transpose. We further employ a full-rank spatial covariance matrix (SCM) [2] on the outer product of a steering vector $\mathbf{a}_j(f)$:

$$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^{\mathsf{H}}(f). \tag{4}$$

$\mathbf{x}(f,n)$ can be rewritten as follows:

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \textstyle\sum_j v_j(f,n)\mathbf{R}_j(f)). \tag{5}$$

Thus, the log-likelihood function $\mathcal{L}$ is given by:

$$\mathcal{L} \stackrel{c}{=} -\textstyle\sum_{f,n} \left[\mathrm{tr}\left(\mathbf{X}(f,n)\hat{\mathbf{X}}^{-1}(f,n) \right) + \log \det \hat{\mathbf{X}}(f,n) \right], \tag{6}$$

where $\stackrel{c}{=}$ represents the equality up to constant, and $\mathbf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)$ and $\hat{\mathbf{X}}(f,n) = \sum_j v_j(f,n)\mathbf{R}_j(f)$.

Once PSDs $\{v_j(f,n)\}_{j,f,n}$ and SCMs $\{\mathbf{R}_j(f)\}_{j,f}$ are estimated, the j-th separated signal is obtained by applying a multichannel Wiener filter $\mathbf{M}_j(f,n)$:

$$\hat{\mathbf{s}}_j(f,n) = \underbrace{v_j(f,n)\mathbf{R}_j(f)\left(\textstyle\sum_j v_j(f,n)\mathbf{R}_j(f)\right)^{-1}}_{\mathbf{M}_j(f,n)} \mathbf{x}(f,n), \tag{7}$$

followed by applying the inverse STFT.

3. THE GMVAE METHOD

3.1. CVAE and ACVAE Source Model

The original GMVAE method represents the source power spectrograms as the decoder outputs of the CVAE that is trained in advance using labeled training examples. Given a normalized source spectrogram $\tilde{\mathbf{S}}$ and the one-hot encoded label $\mathbf{c}$, the encoder and decoder distributions are assumed to follow a Gaussian distribution and a zero-mean complex Gaussian distribution:

$$q_\phi(\mathbf{Z}|\tilde{\mathbf{S}}, \mathbf{c}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\tilde{\mathbf{S}}, \mathbf{c}), \mathrm{diag}\boldsymbol{\sigma}^2_\phi(\tilde{\mathbf{S}}, \mathbf{c})), \tag{8}$$

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{Z}, \mathbf{c}) = \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathrm{diag}\boldsymbol{\sigma}^2_\theta(\mathbf{Z}, \mathbf{c})), \tag{9}$$

where $\boldsymbol{\mu}_\phi(\tilde{\mathbf{S}}, \mathbf{c})$, $\boldsymbol{\sigma}^2_\phi(\tilde{\mathbf{S}}, \mathbf{c})$, and $\boldsymbol{\sigma}^2_\theta(\tilde{\mathbf{S}}, \mathbf{c})$ denote the encoder and decoder outputs. For CVAE, both encoder and decoder network parameters ϕ and θ are trained by maximizing the following training criterion:

$$\begin{aligned}\mathcal{I}(\phi,\theta) = \mathbb{E}_{(\tilde{\mathbf{S}},\mathbf{c})\sim p_{\mathcal{D}}(\tilde{\mathbf{S}},\mathbf{c})}[\mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\tilde{\mathbf{S}},\mathbf{c})}[\log p_\theta(\tilde{\mathbf{S}}|\mathbf{Z},\mathbf{c})] \\ - \mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\tilde{\mathbf{S}},\mathbf{c})||p(\mathbf{Z})],\end{aligned} \tag{10}$$

where $\mathbb{E}_{(\tilde{\mathbf{S}},\mathbf{c})\sim p_{\mathcal{D}}(\tilde{\mathbf{S}},\mathbf{c})}[\cdot]$ denotes the sample mean over a dataset, $\mathcal{D}_{\mathrm{KL}}[\cdot||\cdot]$ is the Kullback-Leivler (KL) divergence, and $p(\mathbf{Z})$ is a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

ACVAE is a CVAE variant that incorporates the expectation of the mutual information $I(\mathbf{c}; \tilde{\mathbf{S}}|\mathbf{Z})$ into the training criterion with the aim of making the decoder output $\tilde{\mathbf{S}} \sim p_\theta(\tilde{\mathbf{S}}|\mathbf{Z}, \mathbf{c})$ as correlated as possible with the class label $\mathbf{c}$. Since it is difficult to use the mutual information directly, ACVAE uses the following variational lower bound instead:

$$\begin{aligned}\mathcal{J}(\phi,\theta,\psi) =& \mathbb{E}_{(\tilde{\mathbf{S}},\mathbf{c})\sim p_{\mathcal{D}}(\tilde{\mathbf{S}},\mathbf{c}),\mathbf{Z}\sim q_\phi(\mathbf{Z}|\tilde{\mathbf{S}},\mathbf{c})}[\\ & \mathbb{E}_{\mathbf{c}'\sim p_{\mathcal{D}}(\mathbf{c}),\tilde{\mathbf{S}}\sim p_\theta(\tilde{\mathbf{S}}|\mathbf{Z},\mathbf{c}')}[\log r_\psi(\mathbf{c}'|\tilde{\mathbf{S}})]],\end{aligned} \tag{11}$$

where $r_\psi(\mathbf{c}|\tilde{\mathbf{S}})$ is an auxiliary classifier distribution with the network parameter ψ. ACVAE also incorporates the cross-entropy:

$$\mathcal{K}(\psi) = \mathbb{E}_{(\tilde{\mathbf{S}},\mathbf{c})\sim p_{\mathcal{D}}(\tilde{\mathbf{S}},\mathbf{c})}[\log r_\psi(\mathbf{c}|\tilde{\mathbf{S}})]. \tag{12}$$

Thus, the whole training criterion of ACVAE is given by:

$$\mathcal{I}(\phi,\theta) + \lambda_{\mathcal{J}}\mathcal{J}(\phi,\theta,\psi) + \lambda_{\mathcal{K}}\mathcal{K}(\psi), \tag{13}$$

where $\lambda_{\mathcal{J}} \geq 0$ and $\lambda_{\mathcal{K}} \geq 0$ are weight parameters.

Since the decoder distribution is given in the same form as the LGM, using the trained decoder, Eq. (1) is reformulated as follows:

$$v_j(f,n) = g_j\boldsymbol{\sigma}^2_\theta(f,n;\mathbf{Z}_j,\mathbf{c}_j), \tag{14}$$

where $\boldsymbol{\sigma}^2_\theta(f,n;\mathbf{Z},\mathbf{c})$ represents the (f,n)-th element of the decoder output $\boldsymbol{\sigma}^2_\theta(\mathbf{Z},\mathbf{c})$. g_j is the global scale of j-th source signal, which compensates for the energy gap between the training and test time.

3.2. MM-based Parameter Estimation Algorithm

An MM algorithm refers to an iterative algorithm that searches for a stationary point of an objective function by iteratively maximizing an auxiliary function called "minorizer" that is guaranteed to never become above the objective function. Applying an MM algorithm

to the log-likelihood function in Eq. (6), the following minorizer $\mathcal{L}_{\mathrm{MM}}(\leq \mathcal{L})$ can be constructed [19]:

$$\mathcal{L}_{\mathrm{MM}} \stackrel{c}{=} -\sum\nolimits_{f,n,j}\Bigg[\frac{\mathrm{tr}(\mathbf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f)\mathbf{P}_j(f,n))}{v_j(f,n)} + v_j(f,n)\mathrm{tr}(\mathbf{K}^{-1}(f,n)\mathbf{R}_j(f))\Bigg], \quad (15)$$

where $\mathbf{P}_j(f,n)$ and $\mathbf{K}(f,n)$ are auxiliary variables defined by:

$$\mathbf{P}_j(f,n) \leftarrow \mathbf{M}_j(f,n), \quad (16)$$

$$\mathbf{K}(f,n) \leftarrow \hat{\mathbf{X}}(f,n). \quad (17)$$

An MM-based parameter estimation algorithm consists of updating auxiliary variables $\{\mathbf{P}_j(f,n)\}_{j,f,n}$ and $\{\mathbf{K}(f,n)\}_{f,n}$, and maximizing the minorizer with respect to $\{g_j\}_j$, $\{\mathbf{Z}_j\}_j$, $\{\mathbf{c}_j\}_j$, and $\{\mathbf{R}_j(f)\}_{j,f}$. The decoder inputs $\{\mathbf{Z}_j\}_j$ and $\{\mathbf{c}_j\}_j$ can be updated by backpropagation:

$$\{\mathbf{Z}_j, \mathbf{c}_j\}_j \leftarrow \{\mathbf{Z}_j, \mathbf{c}_j\}_j - \eta\nabla_{\{\mathbf{Z}_j,\mathbf{c}_j\}_j}\mathcal{L}_{\mathrm{MM}}, \quad (18)$$

where η represents a learning rate. Note that, to take a sum-to-one constraint on $\mathbf{c}_j$ into account, we design a softmax layer that output $\mathbf{c}_j$, and the layer input is treated as the parameter to be estimated instead. The optimal updates of $\{g_j\}_j$ $\{\mathbf{R}_j(f)\}_{j,f}$ are obtained as:

$$g_j \leftarrow g_j \times \sqrt{\frac{\sum_{f,n}\frac{1}{\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)}\mathrm{tr}(\mathbf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f)\mathbf{P}_j(f,n))}{\sum_{f,n}\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)\mathrm{tr}(\mathbf{K}^{-1}(f,n)\mathbf{R}_j(f))}}, \quad (19)$$

$$\mathbf{R}_j(f) \leftarrow \mathbf{\Psi}_j^{-1}(f)\#\mathbf{\Omega}_j(f), \quad (20)$$

where $\#$ denotes the geometric mean of two positive semidefinite matrices [28], and $\mathbf{\Psi}_j(f) = \sum_n g_j\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)\mathbf{K}^{-1}(f,n)$ and $\mathbf{\Omega}_j(f) = \sum_n \frac{\mathbf{P}_j(f,n)\mathbf{X}(f,n)\mathbf{P}_j(f,n)}{g_j\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)}$.

3.3. EM-based Parameter Estimation Algorithm

An EM algorithm maximizes a log-likelihood function by iteratively maximizing the conditional expectation of the log-likelihood function for complete data called "Q-function" through iterative updates called E- and M-steps. Regarding the mixture and source spectrograms $\{\mathbf{x}(f,n)\}_{f,n}$ and $\{\mathbf{s}(f,n)\}_{f,n}$ as observed and unobserved data, the following Q-function $\mathcal{L}_{\mathrm{EM}}(\leq \mathcal{L})$ can be obtained:

$$\mathcal{L}_{\mathrm{EM}} \stackrel{c}{=} -\sum\nolimits_{f,n,j}\left[\frac{\mathrm{tr}(\mathbf{R}_j^{-1}(f)\mathbf{\Lambda}_j(f,n))}{v_j(f,n)} + \log\det v_j(f,n)\mathbf{R}_j(f)\right]. \quad (21)$$

$\mathbf{\Lambda}_j(f,n)$ is the conditional expectation of the outer product of $\mathbf{s}(f,n)$ defined by:

$$\mathbf{\Lambda}_j(f,n) \leftarrow \mathbf{M}_j(f,n)\mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n)\mathbf{M}_j^{\mathsf{H}}(f,n) + (\mathbf{I} - \mathbf{M}_j(f,n))v_j(f,n)\mathbf{R}_j(f), \quad (22)$$

which amounts to conducting the E-step. The M-step consists of maximizing the Q-function with respect to $\{g_j\}_j$, $\{\mathbf{Z}_j\}_j$, $\{\mathbf{c}_j\}_j$, and $\{\mathbf{R}_j(f)\}_{j,f}$. Similar to the MM-based algorithm, we can update the decoder inputs $\{\mathbf{Z}_j\}_j$ and $\{\mathbf{c}_j\}_j$ by backpropagation:

$$\{\mathbf{Z}_j, \mathbf{c}_j\}_j \leftarrow \{\mathbf{Z}_j, \mathbf{c}_j\}_j - \eta\nabla_{\{\mathbf{Z}_j,\mathbf{c}_j\}_j}\mathcal{L}_{\mathrm{EM}}. \quad (23)$$

The optimal updates of $\{g_j\}_j$ and $\{\mathbf{R}_j(f)\}_{j,f}$ are obtained as:

$$g_j \leftarrow \frac{1}{FNI}\sum\nolimits_{f,n}\frac{1}{\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)}\mathrm{tr}(\mathbf{R}_j^{-1}(f)\mathbf{\Lambda}_j(f,n)), \quad (24)$$

$$\mathbf{R}_j(f) \leftarrow \frac{1}{N}\sum\nolimits_{n}\frac{1}{g_j\sigma_\theta^2(f,n;\mathbf{Z}_j,\mathbf{c}_j)}\mathbf{\Lambda}_j(f,n). \quad (25)$$

3.4. Discussion

We analyze the MM-based and EM-based algorithms in terms of matrix inversion and multiplication, ignoring the common backpropagation part. At each iteration, the MM-based algorithm is required to repeat updating auxiliary variables $\mathbf{P}_j(f,n)$ and $\mathbf{K}_j(f,n)$ for updating each parameter, and multiple matrix inversions and multiplications are also required for updating $\{\mathbf{R}_j(f)\}_{j,f}$, which amounts $(3N+2J)F$ matrix inversions and $(5N+2)JF$ matrix multiplications. On the other hand, the EM-based algorithm updates an auxiliary variable $\mathbf{\Lambda}_j(f,n)$ once at each iteration, and $\{\mathbf{R}_j(f)\}_{j,f}$ is simply updated by weighted sum, resulting in $(N+J)F$ matrix inversions and $2NJF$ matrix multiplications at each iteration. Thus, the EM-based parameter estimation algorithm takes less computational cost and it is expected to reduce computational time.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Settings

The proposed method was experimentally evaluated under an underdetermined multichannel speech separation scenario where the task is to separate out three sources from their mixtures captured by two microphones.

We used audio samples from the Voice Conversion Challenge (VCC) 2018 dataset [29], which contains recordings of six female and six male U.S. English speakers and includes 116 utterances of individual speakers. We used 100 utterances of two female and two male speakers for training and another 10 utterances of different two female and two male speakers for the test. Similar to [19], we generated 40 mixtures of three speakers by randomly choosing the utterances from the test dataset, where a reverberation time T_{60} was set to 78 ms. All the speech signals were sampled at 16 kHz, and STFT analysis was conducted with a 128-ms window length with a 64-ms shift length.

We used MNMF methods with the EM-based and MM-based optimizations (EM-MNMF, MM-MNMF) [9, 10], and the ConvTasNet [30] as the baseline methods, which are also used for initializing the GMVAE method. The number of NMF bases was set to 10 for each speaker, and the spectral dictionaries of each speaker were obtained using Itakura-Saito NMF (IS-NMF) [31] with 1000 iterations. At the test time, both MNMFs were run for 300 iterations, where the intermideate separated signals at 200-th iteration were used for initializing the GMVAE method. We used an asteroid [32] recipe and trained a ConvTasNet model, using a single-channel three-speaker source separation dataset generated from LibriSpeech [33], i.e., Libri3Mix [34]. After the training, we independently fed the individual channels of a multichannel mixture signal to the ConvTasNet, and the separated signals at each channel were then concatenated to construct multichannel separated signals.

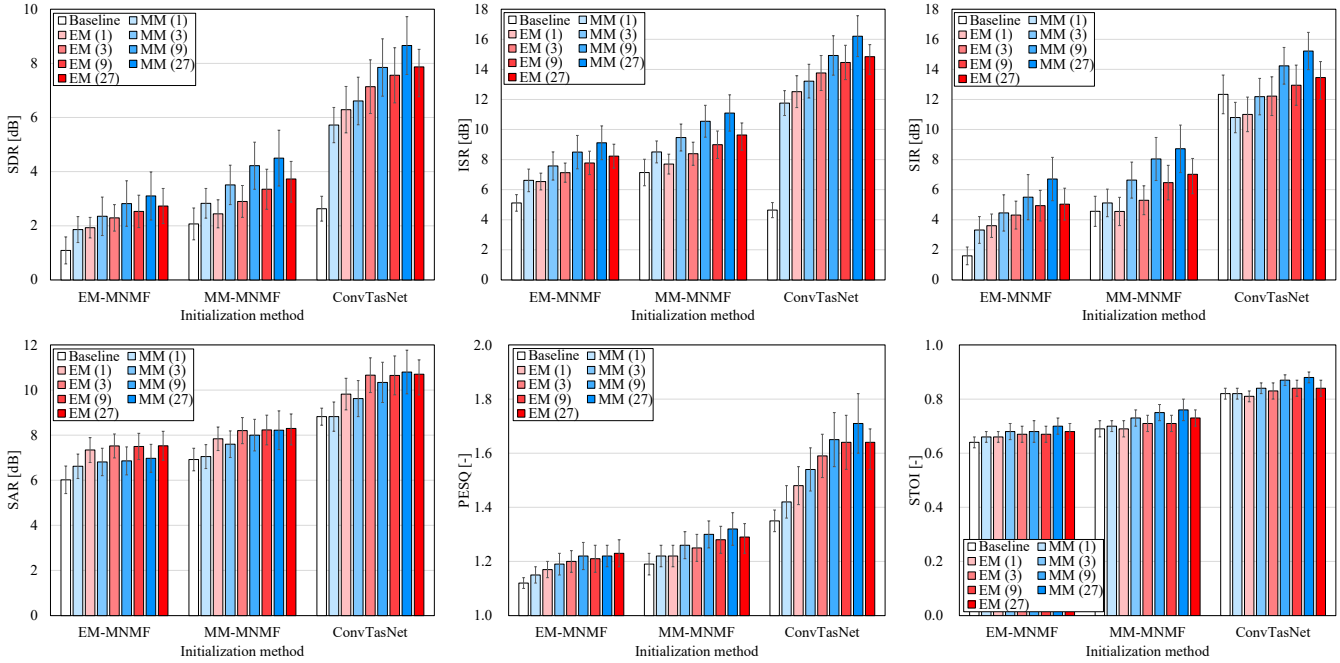

Fig. 1: Source separation performances under a speaker-open condition, where the error bars and the numbers in parentheses show the 95 % confidence intervals and the number of backpropagation optimizations, respectively.

Table 1: Computational times per iteration for a 6.89-second speech mixture [s], where the numbers in parentheses show the number of backpropagation optimizations.

Optimization approach	MNMF	The GMVAE method			
		(1)	(3)	(9)	(27)
MM (CPU)	**0.95**	1.41	2.07	4.16	10.15
EM (CPU)	3.34	**0.89**	**1.55**	**3.64**	**9.46**
MM (GPU)	–	1.05	1.15	1.47	2.43
EM (GPU)	–	**0.59**	**0.71**	**1.00**	**1.92**

These separated signals were also used for initializing the GMVAE method.

We used three-layer convolutional, three-layer deconvolutional, and four-layer convolutional neural networks with gated linear units (GLUs) for the encoder, decoder, and auxiliary classifier, respectively. All the weight parameters $\lambda_{\mathcal{J}}$ and $\lambda_{\mathcal{K}}$ were set to 1. The number of training epochs was set to 1000. We used the Adam algorithm [35] with a learning rate of 0.0001 and 0.01 for the training and test, respectively.

As the evaluation metrics, we used the signal-to-distortion ratio (SDR), the source image-to-spatial distortion ratio (ISR), the signal-to-inference ratio (SIR), the signal-to-artifact ratio (SAR) [36], the perceptual evaluation of speech quality (PESQ) [37], and the short-time objective intelligibility (STOI) [38].

4.2. Experimental Results

Table 1 shows a comparison of the computational time of each method, where faster performances are denoted in bold fonts. We used an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and a NVIDIA Tesla K80 for the comparison. Unlike MNMF, in the GMVAE method, the EM-based algorithm is consistently faster than the MM-based one. We can see that, when using a GPU, the EM-based GMVAE methods with less than 10 backpropagations are faster than or as fast as the MM-based MNMF. Furthermore, even when using a CPU, the EM-based GMVAE method with one backpropagation is as fast as the MM-based MNMF.

Fig. 1 shows a comparison of the separation performance of each method with different initialization. We can see that, when increasing the number of backpropagations, the GMVAE method consistently achieves better performance and that the EM-based GMVAE method achieves comparable performances with the MM-based GMVAE method. Comparing the initialization methods, we can see that, the better initialization method is used, i.e., ConvTasNet, the larger performance improvements the GMVAE method can gain.

These results demonstrate that 1) the EM-based GMVAE method is as fast as the conventional MM-based MNMF while achieving better performance, 2) the EM-based GMVAE method is faster than the MM-based GMVAE method while achieving comparable performance, and 3) better initialization methods can help the GMVAE method achieve higher separation performances.

5. CONCLUSION

This paper developed the GMVAE method for underdetermined multichannel source separation and investigated the MM-based and EM-based convergence-guaranteed parameter estimation algorithms. We analyzed and compared the MM-based and EM-based algorithms in terms of computational cost and separation performance, using an ACVAE instead of a CVAE as the source model of the GMVAE method. The experimental results demonstrated that the EM-based GMVAE method consistently outperformed a conventional MNMF and that the EM-based GMVAE was faster than the MM-based GMVAE method achieving comparable performance.

Future work includes use of inferences from encoder and auxiliary classifier [20] and jointly-diagonalizability constraint on a full-rank SCM [39, 40] for speeding up and improving the algorithm.

6. REFERENCES

[1] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[2] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, 2010.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components," in *Proc. ICA*, pp. 165–172, 2006.

[5] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions," in *Proc. ICA*, pp. 601–608, 2006.

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

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA*, pp. 177–180, 2003.

[8] A. Ozerov and C. Févotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2009.

[9] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. ISSPA*, pp. 1–4, 2010.

[10] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[11] 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. LVA/ICA*, pp. 245–253, 2010.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[13] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel Audio Source Separation With Deep Neural Networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[14] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent Deeply Learned Matrix Analysis for Multichannel Audio Source Separation," in *Proc. EUSIPCO*, pp. 1557–1561, 2018.

[15] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization," in *Proc. ICASSP*, pp. 716–720, 2018.

[16] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-Supervised Multichannel Speech Enhancement With a Deep Speech Prior," *IEEE/ACM Trans. ASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.

[17] S. Leglaive, L. Girin, and R. Horaud, "Semi-Supervised Multichannel Speech Enhancement with Variational Autoencoders and Non-negative Matrix Factorization," in *Proc. ICASSP*, pp. 101–105, 2019.

[18] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised Determined Source Separation with Multichannel Variational Autoencoder," *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[19] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Underdetermined Source Separation Based on Generalized Multichannel Variational Autoencoder," *IEEE Access*, vol. 7, pp. 168104–168115, 2019.

[20] L. Li, H. Kameoka, S. Inoue, and S. Makino, "FastMVAE: A Fast Optimization Algorithm for the Multichannel Variational Autoencoder Method," *IEEE Access*, vol. 8, pp. 228740–228753, 2020.

[21] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. ICLR*, 2014.

[22] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-Supervised Learning with Deep Generative Models," in *Proc. NIPS*, pp. 3581–3589, 2014.

[23] D. R. Hunter and K. Lange, "A Tutorial on MM Algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[25] H. Sawada, R. Ikeshita, and T. Nakatani, "Experimental Analysis of EM and MU Algorithms for Optimizing Full-rank Spatial Covariance Model," in *Proc. EUSIPCO*, pp. 885–889, 2021.

[26] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-Parallel Voice Conversion With Auxiliary Classifier Variational Autoencoder," *IEEE/ACM Trans. ASLP*, vol. 27, no. 9, pp. 1432–1443, 2019.

[27] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. WASPAA*, pp. 78–81, 2005.

[28] K. Yoshii, "Correlated Tensor Factorization for Audio Source Separation," in *Proc. ICASSP*, pp. 731–735, 2018.

[29] 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 Non-parallel Methods," in *Proc. Odyssey*, pp. 195–202, 2018.

[30] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[31] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[32] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers," in *Proc. Interspeech*, pp. 2637–2641, 2020.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206–5210, 2015.

[34] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," 2020.

[35] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.

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

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

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2125–2136, 2011.

[39] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast Multichannel Nonnegative Matrix Factorization with Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 2610–2625, 2020.

[40] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A Joint Diagonalization Based Efficient Approach to Underdetermined Blind Audio Source Separation Using the Multichannel Wiener Filter," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 1950–1965, 2021.