JOINT SEPARATION AND DEREVERBERATION OF REVERBERANT MIXTURES WITH MULTICHANNEL VARIATIONAL AUTOENCODER

Shota Inoue[1], *Hirokazu Kameoka*[2], *Li Li*[1], *Shogo Seki*[3], *Shoji Makino*[1]

[1]University of Tsukuba, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]University of Nagoya, Japan

s.inoue@mmlab.cs.tsukuba.ac.jp, kameoka.hirokazu@lab.ntt.co.jp

ABSTRACT

In this paper, we deal with a multichannel source separation problem under a highly reverberant condition. The multichannel variational autoencoder (MVAE) is a recently proposed source separation method that employs the decoder distribution of a conditional VAE (CVAE) as the generative model for the complex spectrograms of the underlying source signals. Although MVAE is notable in that it can significantly improve the source separation performance compared with conventional methods, its capability to separate highly reverberant mixtures is still limited since MVAE uses an instantaneous mixture model. To overcome this limitation, in this paper we propose extending MVAE to simultaneously solve source separation and dereverberation problems by formulating the separation system as a frequency-domain convolutive mixture model. A convergence-guaranteed algorithm based on the coordinate descent method is derived for the optimization. Experimental results revealed that the proposed method outperformed the conventional methods in terms of all the source separation criteria in highly reverberant environments.

***Index Terms*—** Blind source separation, blind dereverberation, multichannel audio signal processing, multichannel variational autoencoder (MVAE)

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating individual source signals from recorded microphone array inputs without any prior information about source signals and the transfer characteristics between sources and microphones. The most commonly used approach for determined BSS problems is independent component analysis (ICA) [1], which achieves source separation by assuming the statistical independence between the sources. Among the ICA-based methods, those methods performing separation in the frequency domain provide the flexibility of utilizing various models for the time–frequency representations of source signals and array responses, which play critical roles in BSS. For example, independent vector analysis (IVA) [2–4] solves frequency-wise source separation and permutation alignment simultaneously by assuming that the magnitudes of the frequency components originating from the same source tend to vary coherently over time. Determined multichannel non-negative matrix factorization (DMNMF) [5], which was later named as independent low-rank matrix analysis (ILRMA) [6, 7], adopts the NMF concept to source spectrogram modeling that approximates each source power spectrogram as the linear combination of a limited set of spectral templates scaled by magnitudes varying with time.

Recently, to achieve further improvements, some attempts have been made to combine deep neural networks (DNNs) with the ICA-based multichannel source separation framework [8–10]. One of them is the multichannel variational autoencoder (MVAE) [10]. MVAE trains a conditional VAE using spectrograms of clean source signals and the corres ponding attribute class labels so that the trained decoder distribution can be used as a generative model of the underlying source signals in a mixture, which is called the CVAE source model. MVAE has shown to significantly outperform conventional methods, which indicates that the CVAE source model has the capability to improve the source separation performance. However, one drawback is that the source separation performance degrades in highly reverberant environments since MVAE assumes an instantaneous mixture model.

To address this drawback, this study proposes an extension of MVAE that allows us to simultaneously perform source separation and dereverberation. Specifically, we formulate the separation system of mixture signals as a frequency-domain convolutive mixture model, which has been shown to be effective for separating highly reverberant mixtures in many previous studies [5, 11, 12].

The rest of this paper is structured as follows. In Section 2, we formulate a multichannel BSS problem with instantaneous mixture models and review MVAE. In Section 3, we present the MVAE using frequency-domain convolutive mixture models and derive a convergence-guaranteed algorithm for the optimization. The results of highly reverberant source separation experiments are presented in Section 4.

2. MULTICHANNEL VARIATIONAL AUTOENCODER

We consider a determined situation where J source signals are observed by I microphones ($J = I$). Let $x_i(f, n)$ and $s_j(f, n)$ denote the short-time Fourier transform (STFT) coefficients of the signal observed at the i-th microphone and the j-th source signal, where f and n are the frequency and time indices, respectively. Now, we use a separation system

This work was supported by JSPS KAKENHI Grant Number 17H01763 and 18J20059, and SECOM Science and Technology Foundation.

of the form

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f)\boldsymbol{x}(f,n), \tag{1}$$

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f), \dots, \boldsymbol{w}_I(f)] \in \mathbb{C}^{I\times I} \tag{2}$$

to describe the relationship between the observed signals $\boldsymbol{x}(f,n) = [x_1(f,n), \dots, x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ and sources $\boldsymbol{s}(f,n) = [s_1(f,n), \dots, s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$, where $\boldsymbol{W}^{\mathsf{H}}(f)$ is called the separation matrix and $(\cdot)^{\mathsf{H}}$ denotes Hermitian transpose.

Let us assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$

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

We call (3) the local Gaussian model (LGM). We further assume $s_j(f,n)$ to be independent from one source to the others. $\boldsymbol{s}(f,n)$ thus follows

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

where $\boldsymbol{V}(f,n)$ is a diagonal matrix with diagonal entries $v_1(f,n), \dots, v_I(f,n)$. From (1) and (4), we can show that $\boldsymbol{x}(f,n)$ follows

$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|0, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}). \tag{5}$$

Hence, the negative log-likelihood of the parameters $\mathcal{V} = \{v_j(f,n)\}_{f,n,j}$ and $\mathcal{W} = \{\boldsymbol{W}^{\mathsf{H}}(f)\}_f$ given the observed mixture signals $\mathcal{X} = \{x_i(f,n)\}_{i,f,n}$ is given as

$$\begin{aligned}\mathcal{L}(\mathcal{V}, \mathcal{W}|\mathcal{X}) \overset{c}{=} & -2N\log\left|\det \boldsymbol{W}^{\mathsf{H}}(f)\right| \\ & + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|s_j(f,n)|^2}{v_j(f,n)}\right),\end{aligned} \tag{6}$$

where $=^c$ denotes equality up to constant terms. It is important to note that if we individually treat $v_j(f,n)$ as a free parameter indexed by frequency f, the negative log-likelihood will be split into frequency-wise source separation problems. This means that the permutation of the separated components in each frequency is not uniquely determined. Permutation alignment is thus needed after $\mathcal{W}$ has been obtained.

To solve this problem, for MVAE [10], a conditional VAE (CVAE) [13] is used to model and estimate the spectrograms of the sources $s_j(f,n)$. Let $\boldsymbol{S} = \{s(f,n)\}_{f,n}$ be the complex spectrogram of a particular sound source and c be the corresponding attribute class label, which is represented as a one hot vector. With a set of labeled training samples $\{\boldsymbol{S}_m, c_m\}_{m=1}^M$, a CVAE network consisting of an encoder network $q_\phi(\mathbf{z}|\boldsymbol{S}, c)$ and a decoder network $p_\theta(\boldsymbol{S}|\mathbf{z}, c)$ is trained jointly by maximizing

$$\begin{aligned}\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[&\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\mathbf{z},c)] \\ & - KL[q_\phi(\mathbf{z}|\boldsymbol{S},c)||p(\mathbf{z})]],\end{aligned} \tag{7}$$

where $\mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\cdot]$ denotes the sample mean over the training examples and $KL[\cdot||\cdot]$ is the Kullback–Leibler divergence. Here, MVAE defines the decoder distribution as a zero-mean complex Gaussian distribution as follows so that it has the same form as the LGM (3).

$$p_\theta(\boldsymbol{S}|\mathbf{z}, c, g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \tag{8}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\mathbf{z},c), \tag{9}$$

where $\sigma_\theta^2(f,n;\mathbf{z},c)$ denotes the (f,n)-th element of the decoder output and g represents the global scale of the generated spectrogram. Regarding the encoder distribution $q_\phi(\mathbf{z}|\boldsymbol{S},c)$, a regular Gaussian distribution is adopted

$$q_\phi(\mathbf{z}|\boldsymbol{S},c) = \prod_k \mathcal{N}(z(k)|\mu_\phi(k;\boldsymbol{S},c), \sigma_\phi^2(k;\boldsymbol{S},c)), \tag{10}$$

where $z(k)$, $\mu_\phi(k;\boldsymbol{S},c)$ and $\sigma_\phi^2(k;\boldsymbol{S},c)$ represent the k-th element of the latent space variable $\mathbf{z}$ and the encoder outputs $\boldsymbol{\mu}_\phi(\boldsymbol{S},c)$ and $\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c)$, respectively. The trained decoder distribution can then be used as a generative model of the complex spectrogram of the j-th source $p_\theta(\boldsymbol{S}_j|\mathbf{z}_j, c_j, g_j)$, where $\mathbf{z}_j$, c_j and g_j are the unknown parameters of the model. This generative model is called the CVAE source model. The optimization algorithm of MVAE consists of iteratively updating the separation matrices $\mathcal{W}$ using the iterative projection (IP) method [4], the source model parameters $\Psi = \{\mathbf{z}_j, c_j\}_j$ using backpropagation and the global scale $\mathcal{G} = \{g_j\}_j$ using the following update rule:

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)}. \tag{11}$$

MVAE is notable in that (i) it takes full advantage of the strong representation power of DNNs for source power spectrogram modeling, and (ii) the convergence of the source separation algorithm is guaranteed. However, similar to the other instantaneous mixture model-based methods, the source separation capability of MVAE is limited in a highly reverberant environment where the length of the room impulse responses (RIRs) can be large than the STFT frame length.

3. PROPOSED METHOD

3.1. Formulation

In this section, we describe the proposed method that extends MVAE to simultaneously solve source separation and dereverberation problems. Specifically, instead of the instantaneous mixture model (1), we formulate the separation system as a frequency-domain convolutive mixture model, which has been shown to be effective for separating highly reverberant mixtures [5, 11, 12]. With the frequency-domain convolutive mixture model that has a multichannel finite-impulse-response form, the relationship between the observed signals $\boldsymbol{x}(f,n)$ and sources $\boldsymbol{s}(f,n)$ is written as

$$\boldsymbol{s}(f,n) = \sum_{n'=0}^{N'}\boldsymbol{W}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'). \tag{12}$$

Here, $\boldsymbol{W}^{\mathsf{H}}(f,n')$, $0 \le n' \le N'$ are the coefficient matrices of size $I\times I$ and $\boldsymbol{W}^{\mathsf{H}}(f,0)$ is equivalent to $\boldsymbol{W}^{\mathsf{H}}(f)$ in (1).

When $\boldsymbol{W}^{\mathsf{H}}(f,0)$ is invertible, the dereverberated mixture signal $\boldsymbol{y}(f,n) = [y_i(f,n),\dots,y_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ and the source signal $\boldsymbol{s}(f,n)$ can be written as

$$\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n) - \sum_{n'=1}^{N'} \boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{13}$$

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{y}(f,n), \tag{14}$$

where $\boldsymbol{D}^{\mathsf{H}}(f,n') = -(\boldsymbol{W}^{\mathsf{H}}(f,0))^{-1}\boldsymbol{W}^{\mathsf{H}}(f,n')$, $1 \leq n' \leq N'$. Note that (13) can be seen as a dereverberation process of the observed mixture signal $\boldsymbol{x}(f,n)$, whereas (14) can be seen as an instantaneous demixing processing of the dereverberated mixture signal $\boldsymbol{y}(f,n)$. Therefore, the negative log-likelihood of interest is a function of the dereverberation filter $\mathcal{D} = \{\boldsymbol{D}^{\mathsf{H}}(f,n')\}_{f,n'}$, separation matrices $\mathcal{W}$, spectral parameters Ψ and scale parameter $\mathcal{G}$:

$$\mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}) \overset{c}{=} -2N\log\left|\det \boldsymbol{W}^{\mathsf{H}}(f)\right| + \sum_{f,n,j}\Bigg(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)(\boldsymbol{x}(f,n) - \sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'))|^2}{v_j(f,n)}\Bigg). \tag{15}$$

3.2. Related work

In [12], the idea of formulating the separation system of highly reverberant mixture signals using a frequency-domain convolutive mixture model has been adopted to ILRMA [6], which allows the method to solve source separation and dereverberation simultaneously. We refer to this method as ILRMA+ hereafter. The proposed method is different from ILRMA+ in the way of modeling the sources $s_j(f,n)$, where the proposed method uses the CVAE source model whereas ILRMA+ employs a non-negative matrix factorization model. Specifically, ILRMA+ models the variance $v_j(f,n) = \sum_{k=1}^{K_j} b_{j,k}(f)h_{j,k}(n)$, which amounts to assuming that the power spectrograms of source signals can be approximated by the linear combination of a small number of spectral templates $b_{j,1}(f),\dots,b_{j,K_j}(f) \geq 0$ scaled by magnitudes varying with time $h_{j,1}(n),\dots,h_{j,K_j}(n) \geq 0$. From this viewpoint, the proposed method can be seen as an extension of ILRMA+ that replaces the NMF model with the CVAE source model to achieve better source separation performance by enhancing the representation power of the source model.

3.3. Optimization process

We describe the optimization algorithm in this subsection, in which the objective function (15) is iteratively decreased using a coordinate descent method in which each iteration comprises the following four minimization steps:

$$\hat{\mathcal{D}} \leftarrow \underset{\mathcal{D}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{16}$$

$$\hat{\mathcal{W}} \leftarrow \underset{\mathcal{W}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{17}$$

$$\hat{\Psi} \leftarrow \underset{\Psi}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}), \tag{18}$$

$$\hat{\mathcal{G}} \leftarrow \underset{\mathcal{G}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\Psi,\mathcal{G}|\mathcal{X}). \tag{19}$$

By dropping the constant terms with respect to $\mathcal{D}$ from (15), we obtain

$$\mathcal{I}(\mathcal{D}) = \sum_{f,n}\left|\boldsymbol{x}(f,n) - \sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n')\right|^2_{\boldsymbol{\Sigma}_{w/v(f,n)}}, \tag{20}$$

where $|x|_{\boldsymbol{\Sigma}_{w/v(f,n)}} = \sqrt{\boldsymbol{x}^{\mathsf{H}}\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}}$ with $\boldsymbol{\Sigma}_{w/v(f,n)} = \sum_j \frac{\boldsymbol{w}_j(f)\boldsymbol{w}_j^{\mathsf{H}}(f)}{v_j(f,n)}$, which is assumed to be positive definite. To obtain independent updating rules for each f, we vectorize $\{\boldsymbol{D}(f,n')\}_{n'}$ as

$$\begin{aligned}\boldsymbol{d}(f) = [&\boldsymbol{d}_1^{\mathsf{T}}(f,1),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,1),\boldsymbol{d}_1^{\mathsf{T}}(f,2),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,2),\dots,\\ &\boldsymbol{d}_1^{\mathsf{T}}(f,N'),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,N')]^{\mathsf{T}} \in \mathbb{C}^{I^2N'},\end{aligned} \tag{21}$$

where $\boldsymbol{d}_i(f,n')$ is the i-th column of $\boldsymbol{D}(f,n')$. Thus, the term $\sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n')$ in (20) can be rewritten as

$$\sum_{n'=1}^{N'}\boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n') = \boldsymbol{X}(f,n)\boldsymbol{d}^*(f), \tag{22}$$

where $\boldsymbol{d}^*(f)$ represents the complex conjugate of $\boldsymbol{d}(f)$ and

$$\begin{aligned}\boldsymbol{X}(f,n) = [&\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-1),\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-2),\dots,\\ &\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-N')] \in \mathbb{C}^{I\times I^2N'}.\end{aligned} \tag{23}$$

Here, $\otimes$ stands for the Kronecker product. By substituting (22) into (20), we obtain

$$\begin{aligned}\mathcal{I}(\mathcal{D}) = &\sum_{f,n}\big(\boldsymbol{x}(f,n) - \boldsymbol{X}(f,n)\boldsymbol{d}^*(f)\big)^H\\ &\times\boldsymbol{\Sigma}_{w/v_{(}f,n)}\big(\boldsymbol{x}(f,n) - \boldsymbol{X}(f,n)\boldsymbol{d}^*(f)\big).\end{aligned} \tag{24}$$

Since (24) is a quadratic equation with respect to $\boldsymbol{d}^*(f)$, this function can be readily minimized by calculating the partial derivative of $\mathcal{I}(\mathcal{D})$ to be zero. The update rules for each $\boldsymbol{d}^*(f)$ is given as a closed form:

$$\begin{aligned}\boldsymbol{d}^*(f) \leftarrow &\left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{X}(f,n)\right)^{-1}\\ &\times\left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}(f,n)\right).\end{aligned} \tag{25}$$

We employ the following update rules derived on the basis of the IP method [4] to update $\mathcal{W}$:

$$\boldsymbol{w}_j(f) \leftarrow (\boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{\Sigma}_{y/v_j(f)})^{-1}\boldsymbol{e}_j, \tag{26}$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{\Sigma}_{y/v_j}(f)\boldsymbol{w}_j(f)}}, \tag{27}$$

where $\boldsymbol{\Sigma}_{y/v_j}(f) = (1/N)\sum_n \boldsymbol{y}(f,n)\boldsymbol{y}^{\mathsf{H}}(f,n)/v_j(f,n)$ and $\boldsymbol{e}_j$ denotes the j-th column of the $I \times I$ identity matrix.

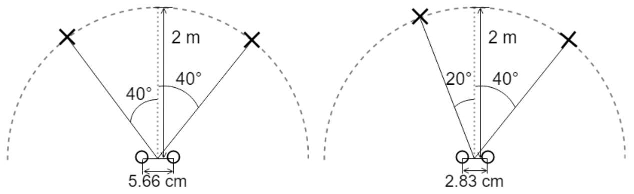

Fig. 1. Microphone and source positions, where ○ and × represent the positions of microphones and sources, respectively.

To update Ψ and $\mathcal{G}$, we apply backpropagation and (11). The parameters of encoder and decoder network are trained by maximizing (7) with labeled clean audio samples. These training and optimization procedures are the same as those of MVAE.

Therefore, the proposed algorithm is summarized as follows:

1. Train θ and ϕ using (7).
2. Initialize Ψ, $\mathcal{G}$, $\mathcal{W}$ and $\mathcal{D}$.
3. Repeat the following updates until convergence.
   (a) Update $\boldsymbol{w}_j(f)$ for each j using (26) and (27).
   (b) Update $\mathbf{z}_j, c_j$ for each j using backpropagation.
   (c) Update g_j for each j using (11).
   (d) Update $\boldsymbol{d}^*(f)$ for each f using (25).

4. EXPERIMENTS

To evaluate the effectiveness of the proposed method in highly reverberant environments, we conducted experiments in which we compared the source separation performance of the proposed approach with those of ILRMA [6], MVAE [10] and ILRMA+ [12]. Specifically, we used RIRs measured in a Japanese-style room (JR1) and an office room (OFC), the reverberation times (T_{60}) of which were 0.60 s and 0.78 s, respectively. Fig. 1 shows the two configurations of the microphones and sources we tested. We used utterances of two female speakers "SF1" and "SF2", and two male speakers "SM1" and "SM2" excerpted from the Voice Conversion Challenge (VCC) 2018 dataset [14] for composing the training and evaluation sets. The audio files for each speaker were manually segmented into 116 short sentences (about 7 min) where 81 and 35 sentences (about 5 and 2 min) were provided as training and evaluation sets, respectively. We generated 10 speech combinations for each speaker pair, namely, SF1+SF2, SF1+SM1, SF2+SM2 and SM1+SM2. Hence, there were in total 80 test signal in each reverberant environment. The length of each signal was about 4 to 7 s long. Speaker identities were considered as the only class category. Thus, the class label c was a four-dimensional one hot vector.

All mixture signals were resampled at 16 kHz. The STFT was computed using the Hamming window with 256 ms long and 64 ms window shift. For ILRMA and ILRMA+, the basis number K was set at 5. The dereverberation filter length N' was set at 3 for JR1 and 4 for OFC, respectively. We run 100 iterations for ILRMA and ILRMA+, and 60 iterations for MVAE and the proposed method. To initialize $\boldsymbol{W}^{\mathsf{H}}(f)$ and $\boldsymbol{D}^{\mathsf{H}}(f, n')$ of the proposed method, we run ILRMA+ for 30 iterations. For the encoder and decoder networks, we employed the same architectures as those used in [10], i.e., a three-layer fully convolutional network with gated linear units (GLUs) [15] and a three-layer fully deconvolutional network with GLUs. Adam optimization [16] was used for training CVAE and estimating Ψ during the source separation. Note that we must take into account the sum-to-one constraints when updating c_j. This can be easily implemented by inserting an appropriately designed softmax layer

$$c_j = \text{Softmax}(u_j), \tag{28}$$

and treat u_j as the parameter to be estimated instead.

Table 1. The average SDR, SIR and SAR improvements achieved by each method. The bold font shows the top scores.

RIRs	Methods	Improvement (dB)		
		SDR	SIR	SAR
JR1 $T_{60} = 0.60$ (s)	ILRMA	2.57	7.60	-0.94
	ILRMA+	5.06	11.20	1.15
	MVAE	3.68	10.67	-0.42
	MVAE+	**6.66**	**14.74**	**2.22**
OFC $T_{60} = 0.78$ (s)	ILRMA	2.43	7.48	-1.04
	ILRMA+	5.43	11.48	1.63
	MVAE	3.53	10.43	-0.50
	MVAE+	**6.89**	**14.90**	**2.64**

We took the average of the signal-to-distortion ratios (SDR), signal-to-interference ratios (SIR) and signal-to-artifact ratios (SAR) [17] as the evaluation criteria. Table 1 shows the separation performances under the two reverberant conditions. The proposed method is shown to outperform all the conventional methods in terms of SDR, SIR and SAR. By comparing the results of ILRMA+ and the proposed method with those of ILRMA and MVAE, it is confirmed that the frequency-domain convolutive mixture models are effective for improving the source separation performances under highly reverberant conditions.

5. CONCLUSIONS

In this paper, we proposed an extension of MVAE that is capable of solving source separation and dereverberation problems simultaneously by formulating the separation system as a frequency-domain convolutive mixture model. A convergence-guaranteed optimization process was derived, which consists of iteratively updating (i) the spectral parameters of each source by applying backpropagation using the CVAE source model, (ii) the separation matrices using the IP method and (iii) the dereverberation filters using multichannel linear prediction. The experimental results showed that the combination of the CVAE source model and the frequency-domain convolutive mixture model was able to improve the source separation performances in highly reverberant environments.

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] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

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

[4] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[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. LVA/ICA*, 2010, pp. 245–253.

[6] 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.

[7] 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*, S. Makino, Ed. Mar. 2018, pp. 125–155, Springer.

[8] 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.

[9] 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*, 2018.

[10] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv preprint arXiv:1808.00892*, Aug. 2018.

[11] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. ASLP*, vol. 19, no. 1, pp. 69–84, 2011.

[12] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *Proc. ICASSP*, 2018, pp. 31–35.

[13] D. P. Kingma, S. Mohamedy, D. J. Rezendey, and M. Welling, "Semi-supervised learning with deep generative models," in *Adv. Neural Information Processing Systems (NIPS)*, 2014, pp. 3581–3589.

[14] 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," *arXiv preprint arXiv:1804.04262*, Apr. 2018.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[17] 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.