PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Joint separation, dereverberation and classification of multiple sources using multichannel variational autoencoder with auxiliary classifier

Shota INOUE[(1)], Hirokazu KAMEOKA[(2)], Li LI[(3)], Shoji MAKINO[(4)]

[(1)]University of Tsukuba, Japan, s.inoue@mmlab.cs.tsukuba.ac.jp
[(2)]NTT Communication Science Laboratories, NTT Corporation, Japan, hirokazu.kameoka.uh@hco.ntt.co.jp
[(3)]University of Tsukuba, Japan, lili@mmlab.cs.tsukuba.ac.jp
[(4)]University of Tsukuba, Japan, maki@tara.tsukuba.ac.jp

Abstract

This paper proposes a unified approach to jointly solving the separation, dereverberation, and classification of mixed sound sources from microphone array observations. The proposed method uses a frequency-wise convolutive mixture model to express the mixing process under highly reverberant environments and the auxiliary classifier conditional variational autoencoder (ACVAE) to model the complex spectrograms of underlying sources. Using an ACVAE as the source generative model allows us to estimate the latent vectors and the class index of each source in a test mixture by computing the outputs of the pretrained approximate posterior inference networks without using backpropagation. We experimentally confirmed that the proposed method outperformed conventional methods in terms of both computation time and source classification.

Keywords: Source separation, blind dereverberation, multichannel audio signal processing, multichannel variational autoencoder (MVAE)

1 INTRODUCTION

Source separation is a technique of separating individual source signals from observed mixture signals. In particular, blind source separation (BSS) achieves the separation of source signals without any prior information about sources and spatial transfer characteristics between microphones and sources. One of the most generally used approaches to solving determined BSS problems, with equal numbers of microphones and sources, is independent component analysis (ICA) [1], which achieves source separation by assuming the statistical independence between sources.

A frequency-domain formulation of ICA provides the flexibility of utilizing various models for the time-frequency representation of source signals. For example, determined multichannel non-negative matrix factorization (DMNMF) [2], which was later called “independent low-rank matrix analysis" (ILRMA) [3], adopts the NMF concept to source spectrogram modeling which approximates source power spectrograms as linear combinations of spectral templates scaled by time-varying amplitudes. One drawback as regards ILRMA is that it can fail to work for sources with spectrograms that do not comply with the NMF model. In stead of the NMF model, the multichannel variational autoencoder (MVAE) method [4] uses the pretrained decoder of a conditional VAE (CVAE) as a generative model of the complex spectrograms of each source. This method allows us to perform source separation and source class identification simultaneously by optimizing the decoder inputs, consisting of the latent space variables and the source class index, so that the log-likelihood is maximized. Thanks to the strong representation power of neural networks for source spectrogram modeling, MVAE was shown to produce good source separation performances. However, MVAE still suffers from two main issues, namely, performance degradations caused by long reverberation and expensive computational costs.

To address the first drawback, we have previously proposed MVAE+ [5], which employs a frequency-wise convolutive mixture model instead of the instantaneous mixture model so that it can handle long reverberation. Under this model, we have derived a convergence-guaranteed algorithm for joint separation and dereverberation, consisting of alternately updating the dereverberation filter, the demixing matrix and the source model parameters. To overcome the second drawback, we have recently introduced a fast algorithm called the “fast MVAE” (fMVAE) [6], where the idea is to replace the process of optimizing the decoder inputs using backpropagation with the forward computations of the two pretrained approximate posterior inference networks.

To combine the advantages of MVAE+ and fMVAE, this paper proposes a fast algorithm for MVAE+ that

performs source separation, dereverberation and source classification in a joint manner. The rest of this paper is structured as follows: Section 2 formulates a multichannel BSS problem with frequency-domain convolutive mixture models and reviews MVAE+. Section 3 presents the proposed fast algorithm that not only reduces the computational time but also improves source classification performance. The experimental results of speech separation and source classification under highly reverberant environments are presented in Section 4.

2 UNIFIED APPROACH FOR SOURCE SEPARATION AND DEREVERBERATION

2.1 Problem formulation

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 formulate the separation system as a frequency-domain convolutive mixture model in order to handle highly reverberant environments where the length of the room impulse responses (RIRs) can be longer than STFT frame length [2, 5, 7, 8]. With this model, we can write the relationship between the observed signals $\boldsymbol{x}(f,n) = [x_1(f,n),\ldots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ and sources $\boldsymbol{s}(f,n) = [s_1(f,n),\ldots,s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ in the following multichannel finite-impulse-response:

$$\boldsymbol{s}(f,n) = \sum_{n'=0}^{N'} \boldsymbol{W}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'). \tag{1}$$

Here, $\boldsymbol{W}(f,0)$ corresponds to the separation matrix and $(\cdot)^{\mathsf{H}}$ denotes Hermitian transpose. When $\boldsymbol{W}^{\mathsf{H}}(f,0)$ is invertible, (1) can be rewritten equivalently as follows:

$$\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n) - \sum_{n'=1}^{N'} \boldsymbol{D}^{\mathsf{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{2}$$

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\mathsf{H}}(f,0)\boldsymbol{y}(f,n), \tag{3}$$

where $\boldsymbol{D}^{\mathsf{H}}(f,n') = -(\boldsymbol{W}^{\mathsf{H}}(f,0))^{-1}\boldsymbol{W}^{\mathsf{H}}(f,n')$, $1 \le n' \le N'$ denotes the dereverberation filter, $\boldsymbol{y}(f,n) = [y_1(f,n),\ldots, y_I(f,n)]^{\mathsf{T}}$ denotes a dereverberated version of the mixture signals and $\boldsymbol{s}(f,n)$ denotes the source signals. Note that (2) can be seen as a dereverberation process applied to the observed mixture signal $\boldsymbol{x}(f,n)$, whereas (3) can be seen as an instantaneous demixing process applied to the dereverberated version of the mixture signal $\boldsymbol{y}(f,n)$.

Let us now assume the local Gaussian model (LGM) [9, 10] in which $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with power spectral density $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$,

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{4}$$

We further assume that $s_j(f,n)$ are statistically independent to each other. $\boldsymbol{s}(f,n)$ thus follows

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|0, \boldsymbol{V}(f,n)), \tag{5}$$

where $\boldsymbol{V}(f,n)$ is a diagonal matrix with diagonal entries $v_1(f,n),\ldots,v_I(f,n)$. From (1) and (5), we can show that $\boldsymbol{y}(f,n)$ follows

$$\boldsymbol{y}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{y}(f,n)|0, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}). \tag{6}$$

Hence, the negative log-likelihood of the spectral parameters $\mathscr{V} = \{v_j(f,n)\}_{f,n,j}$, the dereverberation filter $\mathscr{D} =$

$\{\boldsymbol{D}^{\mathsf{H}}(f,n')\}_{f,n'}$, separation matrices $\mathscr{W}$ given the observed signal $\mathscr{X}$ is given as

$$\mathscr{J}(\mathscr{D},\mathscr{W},\mathscr{V}|\mathscr{X}) \stackrel{c}{=} -2N\log\left|\det \boldsymbol{W}^{\mathsf{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\mathbf{w}_j^{\mathsf{H}}\mathbf{y}(f,n)|^2}{v_j(f,n)}\right), \tag{7}$$

where $=^c$ denotes equality up to constant terms. It is important to note that if we individually treat each element of $v_j(f,n)$ as a free parameter to optimize, the negative log-likelihood will be split into frequency-wise source separation and dereverberation problems. This results in the problem that permutations of the separated components of each frequency cannot be uniquely determined. Thus, we typically need to apply permutation alignment to group the separated components of different frequency bins that originate from the same source signal after we obtain $\mathscr{W}$ or apply some constraints to $v_j(f,n)$ to eliminate the permutation ambiguity during the estimation of $\mathscr{W}$.

2.2 MVAE+: a unified approach using multichannel variational autoencoder

The MVAE method [4] uses a conditional VAE (CVAE) [11] to model and estimate the spectrograms of the sources $s_j(f,n)$. A reverberation-aware extension of the MVAE method, which we call MVAE+, employs a separation system given by (2) and (3).

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 whose form is a one-hot vector. Given a set of labeled training samples $\{\boldsymbol{S}_m, c_m\}_{m=1}^M$, a CVAE, consisting of an encoder $q_\phi(\mathbf{z}|\boldsymbol{S},c)$ and a decoder $p_\theta(\boldsymbol{S}|\mathbf{z},c)$, can be trained by maximizing

$$\mathscr{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\mathbf{z},c)] - KL[q_\phi(\mathbf{z}|\boldsymbol{S},c)||p(\mathbf{z})]], \tag{8}$$

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, we define the decoder distribution as a zero-mean complex Gaussian distribution as follows so that it has the same form as the LGM (4):

$$p_\theta(\boldsymbol{S}|\mathbf{z},c,g) = \prod_{f,n}\mathscr{N}_{\mathbb{C}}(s(f,n)|0,v(f,n)), \tag{9}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\mathbf{z},c), \tag{10}$$

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. As regards the encoder distribution $q_\phi(\mathbf{z}|\boldsymbol{S},c)$, we assume a regular Gaussian distribution

$$q_\phi(\mathbf{z}|\boldsymbol{S},c) = \prod_k \mathscr{N}(z(k)|\mu_\phi(k;\boldsymbol{S},c),\sigma_\phi^2(k;\boldsymbol{S},c)), \tag{11}$$

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 $\mu_\phi(\boldsymbol{S},c)$ and $\sigma_\phi^2(\boldsymbol{S},c)$, 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. Note that since c_j denotes the class index of source j, estimating c_j corresponds to identifying the class of source j in a test mixture. The optimization algorithm of MVAE+ consists of iteratively updating the separation matrices $\mathscr{W}$ using the iterative projection (IP) method [12], the source model parameters $\Psi = \{\mathbf{z}_j,c_j\}_j$ using backpropagation, the dereverberation filter $\mathscr{D} = \{\boldsymbol{D}^{\mathsf{H}}(f,n')\}_{f,n'}$ using the multichannel linear prediction method and the global scale $\mathscr{G} = \{g_j\}_j$ using the following update rule:

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{\sigma_\theta^2(f,n;\mathbf{z}_j,c_j)}. \tag{12}$$

MVAE+ is notable in that (i) it takes full advantage of the strong representation power of DNNs for source power spectrogram modeling, (ii) the convergence of the source separation algorithm is guaranteed, and (iii) it

is capable of removing the reverberant component in the observed mixture signals. We experimentally confirmed in [5] that MVAE+ successfully improved on both MVAE and ILRMA+ [8], namely a reverberation-aware extension of ILRMA, showing the effects of the CVAE source model and the frequency-wise convolutive mixture model.

However, MVAE+ does not address the remaining two drawbacks of MVAE, namely the high computational cost and the limited source classification accuracy.

3 PROPOSED METHOD

3.1 Auxiliary classifier VAE

The auxiliary classifier VAE (ACVAE) [13] is a variant of CVAE that incorporates an information-theoretic regularization [14] to enhance the effect of the class label on the decoder output by maximizing the mutual information between c and $S \sim p_{\theta}(S|z,c)$ conditioned on z. As shown in [6, 13], the regularization term that we would like to maximize with respect to ϕ, θ, ψ becomes

$$\mathscr{L}(\phi,\theta,\psi) = \mathbb{E}_{(S,c)\sim p_D(S,c),q_{\phi}(z|S,c)}[\mathbb{E}_{c\sim p(c),S\sim p_{\theta}(S|z,c)}[\log(c|S)]].$$

Since the labeled training samples can also be used to train the auxiliary classifier $(c|S)$, ACVAE also includes the cross-entropy

$$\mathscr{I}(\psi) = \mathbb{E}_{(S,c)\sim p_D(S,c)}[\log(c|S)] \tag{13}$$

in the training criterion. The entire training criterion is thus given by

$$\mathscr{J}(\phi,\theta) + \lambda_{\mathscr{L}}\mathscr{L}(\phi,\theta,\psi) + \lambda_{\mathscr{I}}\mathscr{I}(\psi), \tag{14}$$

where $\lambda_{\mathscr{L}} \geq 0$ and $\lambda_{\mathscr{I}} \geq 0$ are weight parameters.

Note that the auxiliary classifier $(c|S)$ both assists the encoder and decoder to learn a more disentangled representation.

3.2 optimization process for proposed method

In this subsection, we describe an optimization algorithm for obtaining $\mathscr{D}$, $\mathscr{W}$, and Ψ. The MVAE+ algorithm consists of iteratively maximizing the log-likelihood (7) with respect to $\mathscr{D}$, $\mathscr{W}$, and Ψ in turn. One drawback as regards MVAE+ is that the process of optimizing Ψ has particularly been computationally expensive. Note that when $\mathscr{D}$ and $\mathscr{W}$ are fixed, maximizing the log-likelihood (7) with respect to Ψ is equivalent to finding the maximum point of the distribution $p(z_j,c_j|S_j)$ for each j. While the MVAE and MVAE+ method used backpropagation for finding z_j and c_j, fMVAE took a different approach to reduce the computational effort.

By using the fact that the trained encoder $q_{\phi}(z_j|S_j,c_j)$ of the CVAE model and the trained auxiliary classifier $(c_j|S_j)$ are approximations to the exact posteriors $p(z_j|S_j,c_j)$ and $p(c_j|S_j)$, the distribution $p(z_j,c_j|S_j)$ can be approximated as the product of $q_{\phi}(z_j|S_j,c_j)$ and $(c_j|S_j)$. Thus, we can search for the points that approximately maximize $p(z_j,c_j|S_j)$ by computing the output of $(c_j|S_j)$ followed by computing the mean of $q_{\phi}(z_j|S_j,c_j)$. These values can be obtained simply via forward computations of the two networks. Here, we use this idea of fMVAE to update Ψ.

The update rules for $\mathscr{D}$ can be derived as in [5, 8]. Let us vectorize $\{D(f,n')\}$ by

$$\boldsymbol{d}^*(f) = [\boldsymbol{d}_1^{\mathsf{T}}(f,1),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,1),\boldsymbol{d}_1^{\mathsf{T}}(f,2),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,2),\ldots,\boldsymbol{d}_1^{\mathsf{T}}(f,N'),\ldots,\boldsymbol{d}_I^{\mathsf{T}}(f,N')]^{\mathsf{H}} \in \mathbb{C}^{I^2N'}, \tag{15}$$

where $\boldsymbol{d}_i(f,n')$ is the i-th column of $\boldsymbol{D}(f,n')$ and $(\cdot)^*$ represents the complex conjugate. Then, the update rule for each $\boldsymbol{d}^*(f)$ can be derived as a closed form:

$$\boldsymbol{d}^*(f) \leftarrow \left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\Sigma_{w/v(f,n)}\boldsymbol{X}(f,n)\right)^{-1}\left(\sum_n \boldsymbol{X}^{\mathsf{H}}(f,n)\Sigma_{w/v(f,n)}\boldsymbol{x}(f,n)\right), \tag{16}$$

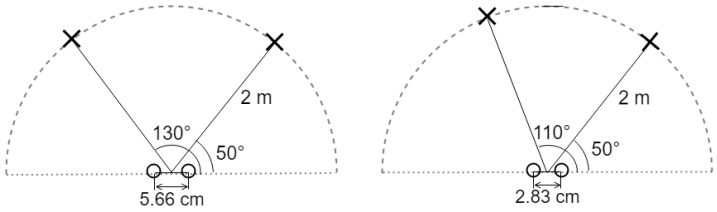

Figure 1. Microphone and source positions, where ∘ and × represent the positions of microphones and sources respectively.

with

$$\boldsymbol{X}(f,n) = [I \otimes \boldsymbol{x}^{\mathsf{T}}(f,n-1), I \otimes \boldsymbol{x}^{\mathsf{T}}(f,n-2), \ldots, I \otimes \boldsymbol{x}^{\mathsf{T}}(f,n-N')] \in \mathbb{C}^{I \times I^2 N'}, \tag{17}$$

and $\Sigma_{w/v(f,n)} = \sum_j \frac{\boldsymbol{w}_j(f)\boldsymbol{w}_j^{\mathsf{H}}(f)}{v_j(f,n)}$, which is assumed to be positive definite. Here, $\otimes$ stands for the Kronecker product.

We employ the following update rules derived on the basis of the IP method [12] to update $\mathscr{W}$:

$$\boldsymbol{w}_j(f) \leftarrow (\boldsymbol{W}^{\mathsf{H}}(f,0)\Sigma_{y/v_j(f)})^{-1}\boldsymbol{e}_j, \tag{18}$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\mathsf{H}}(f)\Sigma_{y/v_j}(f)\boldsymbol{w}_j(f)}}, \tag{19}$$

where $\Sigma_{y/v_j}(f) = (1/N)\sum_n \boldsymbol{y}(f,n)\boldsymbol{y}^{\mathsf{H}}(f,n)/v_j(f,n)$ and $\boldsymbol{e}_j$ denotes the j-th column of the $I \times I$ identity matrix. Therefore, the proposed algorithm is summarized as follows:

1. Train θ, ϕ and ψ using (14).
2. Initialize $\mathbf{z}_j$, c_j, $\mathscr{G}$, $\mathscr{W}$ and $\mathscr{D}$.
3. Repeat the following update steps for each j.
   (a) Update $\boldsymbol{w}_j(f)$ using (18) and (19).
   (b) Update $c_j \leftarrow argmax_{c_j \in \{1,2,\ldots,C\}}(c_j|S_j)$
   (c) Update $\mathbf{z}_j \leftarrow \mu_\phi(\boldsymbol{S}_j, c_j)$
   (d) Update g_j using (12).
   (e) Update $\boldsymbol{d}^*(f)$ using (16).

4 EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed method under highly reverberant environments, we conducted multi-speaker separation experiments to compare the source separation performance, computational time and source classification accuracy of the proposed approach with those of ILRMA+ [8] and MVAE+ [5].

4.1 Experimental conditions

For learning the parameters of CVAE for MVAE+, we used the same clean speech samples as in [4, 5]. For learning the parameters of ACVAE, we used reverberant speech samples that were generated from the same samples mentioned above and the measured RIRs. Specifically, we used six RIRs involved in the CENSREC-4 database for training networks of ACVAE and used two RIRs excerpted from the RWCP database to generate

Table 1. Conditions of RIRs used for training CVAE source model of the proposed method and generating observed mixture signals in tests.

Database	Recording environment	Reverberation time (RT_{60}) [ms]
CENSREC-4 (Used in training phase)	In-car	55 ms
	Office	250
	Lounge	500
	Japanese style bath	600
	Living room	650
	Meeting room	650
RWCP (Used in test phase)	Japanese style room	600 ms
	Meeting room	780

the multichannel mixture signals. Table 1 and Fig. 1 respectively show the conditions of RIRs and 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 [15] to compose the training and evaluation sets. The audio files for each speaker were manually segmented into 116 short sentences (about 7 minutes totally), where 81 and 35 sentences (about 5 and 2 minutes) 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 80 test signals in total under each reverberant environment. The length of each signal was about 4 to 7 seconds long. The class label c was a four-dimensional one-hot vector that indicates the speaker identity.

We resampled all mixture signals at 16 kHz. The STFT was computed using a 256 ms long Hamming window with 64 ms window shift. For 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+, 60 iterations for MVAE+ and 40 iterations for the proposed method. To initialize $\boldsymbol{W}^{\mathsf{H}}(f)$ and $\boldsymbol{D}^{\mathsf{H}}(f,n')$ for MVAE+ and the proposed method, we ran ILRMA+ for 30 iterations. We used the same network architectures for the encoder and decoder in this paper as those used in [4, 5], i.e., a three-layer fully-convolutional network with gated linear units (GLUs) [16] and a three-layer fully-deconvolutional network with GLUs. We also used the same network architectures for ACVAE as those used in [6]. Specifically, architectures of the encoder and decoder were the same as those described above, and the classifier network consisted of a four-layer fully-convolutional network with GLUs and a softmax-layer as the output layer. Adam optimization [17] was used for training networks and estimating Ψ during the source separation. Note that we must take account of the sum-to-one constraints in MVAE+ when updating c_j, which can be easily implemented by inserting appropriately designed softmax layer outputs

$$c_j = \mathrm{Softmax}(u_j), \tag{20}$$

and treat u_j as the parameter to be estimated instead.

The optimization algorithms were run using an Intel(R) Core(TM) i7-7800X CPU@3.50 GHz and a TITAN V GPU.

4.2 Results

We took the average of the signal-to-distortion ratios (SDRs), signal-to-interference ratios (SIRs) and signal-to-artifact ratios (SARs) [18] over the 80 test signals under each condition to evaluate the source separation performance, and we measured the computational times of ILRMA+, MVAE+ and the proposed method.

Tab. 2–3 respectively show the separation performances and computational times of ILRMA+, MVAE+ and the proposed method under the two reverberant conditions. The proposed method was 7.5 times faster than MVAE+ each iteration and slightly faster than ILRMA+ in terms of total computational time when using GPU. We confirmed that the proposed method achieved a comparable performance in terms of SIR improvement compared with MVAE+. The proposed method also outperformed ILRMA+ in terms of SDR and SIR improvements while SAR was slightly worse than ILRMA+. Further investigation on the SAR degradation is one direction of our

Table 2. Average SDR, SIR and SAR improvements of ILRMA+, MVAE+ and the proposed method. The values in bold are the highest scores.

RIRs	Methods	Improvement [dB]		
		SDR	SIR	SAR
JR1 T_{60} = 600 ms	ILRMA+	5.06	11.19	1.15
	MVAE+	**6.66**	**14.73**	**2.22**
	Proposed	5.50	14.34	0.90
OFC T_{60} = 780 ms	ILRMA+	5.46	11.56	1.62
	MVAE+	**6.89**	**14.89**	**2.64**
	Proposed	5.99	14.60	1.49

Table 3. Computational times of MVAE+, the proposed method and ILRMA+. MVAE+ and the proposed method were initialized with by running the ILRMA+ algorithm for 30 iterations in CPU. MVAE+ ran 60 iterations of the optimization algorithm in GPU and the proposed method ran 40 iterations of the optimization algorithm in CPU or GPU. ILRMA ran 100 iterations in CPU.

Method	JR1 (RT_{60} = 600 ms)		OFC (RT_{60} = 780 ms)	
	Runtime / Iteration [sec]	Total [sec]	Runtime / Iteration [sec]	Total [sec]
MVAE+ (GPU)	5.935058	389.018216	6.156569	409.731895
Proposed (CPU)	0.935716	63.547227	1.189260	81.256035
Proposed (GPU)	0.689945	54.147118	0.935897	71.375679
ILRMA+ (CPU)	0.694537	70.446281	0.951857	96.179699

Table 4. Accuracy rates of source classifications obtained with MVAE+ and the proposed method.

	all iterations [%]	final estimation [%]
MVAE+	50.51	74.38
Proposed	**78.98**	**80.00**

future work.

To evaluate the performances of source classification, we computed the classification accuracy rates over all the results estimated at each iteration and those estimated at the final iteration. Table 4 shows the obtained classification accuracy rates. We confirmed that the proposed method could estimate the attribute class of each source signal more accurately by utilizing the classifier in the optimization process for class label estimation.

5 CONCLUSIONS

This paper proposed a unified approach to simultaneously solving source separation, dereverberation, and source classification problems. In the proposed method, we used the frequency-domain convolutive mixture model to model the separation system, and VAEs with an auxiliary classifier (ACVAE) to model and estimate the power spectrograms and speaker identity of the source signals. The optimization process of the proposed method consists of iteratively updating (i) the spectral parameters of each source by the forward calculation of the auxiliary classifier VAE, (ii) the separation matrices using the IP method and (iii) the dereverberation filters using multichannel linear prediction. The experimental results showed that the proposed method achieved comparable performance to MVAE+ in terms of SIR improvement with a source classification rate of 80 % and a reduction of about 83 % in the computational time in each iteration.

ACKNOWLEDGEMENTS

This work was supported by SECOM Science and Technology Foundation, JSPS KAKENHI Grant Number 19H04131, 17H01763 and 18J20059.

REFERENCES

[1] Hyvärinen, A.; Oja, E. Independent component analysis: algorithms and applications, in *Neural networks*, Vol 13, 2000, pp 411–430.

[2] Kameoka, H.; Yoshioka, T.; Hamamura, M.; Le Roux, J.; Kashino, K. Statistical model of speech signals based on composite autoregressive system with application to blind source separation, in *Proc. LVA/ICA*, 2010, pp 245–253.

[3] Kitamura, D.; Ono, N.; Sawada, H.; Kameoka, H.; Saruwatari, H. Determined blind source separation with independent low-rank matrix analysis, in *Audio Source Separation*, Springer, Mar. 2018, pp 125–155.

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

[5] Inoue, S.; Kameoka, H.; Li, L.; Seki, S.; Makino, S. Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder, in *Proc. ICASSP*, 2019, pp 96–100.

[6] Li, L.; Kameoka, H.; Makino, S. Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier, in *Proc. ICASSP*, 2019, pp 546–550.

[7] Yoshioka, T.; Nakatani, T.; Miyoshi, M.; Okuno, HG. Blind separation and dereverberation of speech mixtures by joint optimization, IEEE Trans. ASLP, Vol 19 (1), 2011, pp 69–84.

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

[9] Fev́otte, C.; Bertin, N.; Durrieu, J.-L. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models, in *Proc. WASPAA*, 2005, pp 78–81.

[10] Vincent, E.; Arberet, S.; Gribonval, R. Underdetermined instantaneous audio source separation via local Gaussian modeling, in *Proc. ICA*, 2009, pp 775–782.

[11] Kingma, DP.; Mohamedy, S.; Rezendey, DJ.; Welling, M. Semi-supervised learning with deep generative models, in *Adv. Neural Information Processing Systems (NIPS)*, 2014, pp 3581–3589.

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

[13] Kameoka, H.; Kaneko, T.; Tanaka, K.; Hojo, N. ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder, arXiv preprint arXiv:1808.05092, 2018.

[14] Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in *Adv. NIPS*, 2016, pp 2172–2180.

[15] Lorenzo-Trueba, J.; Yamagishi, J.; Toda, T.; Saito, D.; Villavicencio, F.; Kinnunen, T.; Ling, Z. The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods, arXiv preprint arXiv:1804.04262, Apr. 2018.

[16] Dauphin, YN.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks, arXiv preprint arXiv:1612.08083, 2016.

[17] Kingma, DP.; Ba, J.; Adam: A method for stochastic optimization, in *Proc. ICLR*, 2015.

[18] Vincent E.; Gribonval, R.; Févotte, C. Performance measurement in blind audio source separation, IEEE Trans. ASLP, Vol 14 (4), 2006, pp 1462–1469.