FAST MVAE: JOINT SEPARATION AND CLASSIFICATION OF MIXED SOURCES BASED ON MULTICHANNEL VARIATIONAL AUTOENCODER WITH AUXILIARY CLASSIFIER

Li Li[1], *Hirokazu Kameoka*[2], *Shoji Makino*[1]

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

ABSTRACT

This paper proposes an alternative algorithm for the multichannel variational autoencoder (MVAE), a recently proposed multichannel source separation approach. While MVAE is notable for its impressive source separation performance, its convergence-guaranteed optimization algorithm and the fact that it allows us to estimate source-class labels simultaneously with source separation, there are still two major drawbacks, namely, the high computational complexity and the unsatisfactory source classification accuracy. To overcome these drawbacks, the proposed method employs an auxiliary classifier VAE, which is an information-theoretic extension of the conditional VAE, for learning the generative model of the source spectrograms. Furthermore, with the trained auxiliary classifier, we introduce a novel algorithm for the optimization that can both reduce the computational time and improve the source classification performance. We call the proposed method "fast MVAE (fMVAE)". Experimental evaluations revealed that fMVAE achieved source separation performance comparable to that of MVAE and a source classification accuracy rate of about 80% while reducing computational time by about 93%.

Index Terms— Multichannel source separation, multichannel variational autoencoder, auxiliary classifier, source classification

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating out individual source signals from microphone array inputs when both the sources and the mixing methodology are unknown. The frequency-domain BSS approach allows us to perform instantaneous mixture separation and provides the flexibility of utilizing various models for the time-frequency representations of source signals. For example, independent vector analysis (IVA) [1, 2] 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. Multichannel extensions of non-negative matrix factorization (NMF), e.g., multichannel NMF (MNMF) [3, 4] and independent low-rank matrix analysis (ILRMA) [5, 6], provide an alternative solution to jointly solving these two problems by adopting the NMF concept for the source spectrogram modeling. Specifically, the power spectrograms of the underlying source signals are approximated as the linear sum of a limited number of basis spectra scaled by time-varying amplitudes. It is noteworthy that IVA is equivalent to ILRMA in a particular case where only a single basis spectrum consisting of ones is used for each source signal. From this standpoint, ILRMA can be interpreted as a generalized IVA method that incorporates a source model with stronger representation power, which has been shown to significantly improve source separation performance [6].

Motived by this fact and the high capability of deep neural networks (DNNs) as regards spectrogram modeling, some attempts have recently been made to use DNNs as source models instead of the NMF model [7]–[12]. The multichannel variational autoencoder (MVAE) [11] is one such method, and it has achieved great success in multi-speaker separation tasks. MVAE trains a conditional VAE (CVAE) [13, 14] using power spectrograms of clean speech samples and the corresponding speaker ID as auxiliary label inputs so that the trained decoder distribution can be used as a universal generative model of source signals, where the latent space variables and the class labels are the unknown parameters. At the separation phase, MVAE iteratively updates the separation matrix using the iterative projection (IP) method [15] and the unknown parameters of the source generative model with backpropagation. The separated signals are obtained by applying the estimated separation matrix to the observed mixture signals. This optimization algorithm is notable in that convergence to a stationary point is guaranteed and it allows the source-class labels to be estimated simultaneously with source separation. However, there are two major limitations. Firstly, the backpropagation needed for each iteration causes the optimization algorithm to be highly time-consuming, which can be troublesome in practical applications. Secondly, the encoder and decoder in a regular CVAE are free to ignore the class labels by finding networks that can reconstruct any data without using additional information. In such a situation, the additional class labels will have a limited effect on spectrogram generation, which therefore leads to an unsatisfactory source classification result as we show in Section 4.

To address these limitations, this paper proposes "fast MVAE (fMVAE)", which employs an auxiliary classifier VAE (ACVAE) [16] to learn the generative distribution of source spectrograms and adopts a trained auxiliary classifier for optimization at the separation phase.

2. MVAE FOR DETERMINED MULTICHANNEL SOURCE SEPARATION

2.1. Problem formulation

Let us consider a determined situation where I source signals are captured by I microphones. 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. We denote the vectors contain-

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

ing $x_1(f,n),\dots,x_I(f,n)$ and $s_1(f,n),\dots,s_I(f,n)$ by

$$\boldsymbol{x}(f,n) = [x_1(f,n),\dots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{1}$$

$$\boldsymbol{s}(f,n) = [s_1(f,n),\dots,s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I, \tag{2}$$

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. In a determined situation, the relationship between observed signals and source signals can be described as

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

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f),\dots,\boldsymbol{w}_I(f)] \in \mathbb{C}^{I\times I}, \tag{4}$$

where $\boldsymbol{W}^{\mathsf{H}}(f)$ is called the separation matrix. $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose.

Let us assume that source signals follow the local Gaussian model (LGM), i.e., $s_j(f,n)$ independently follows a zero-mean complex proper Gaussian distribution with variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{5}$$

When $s_j(f,n)$ and $s_{j'}(f,n)(j \neq j')$ are independent, $\boldsymbol{s}(f,n)$ follows

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|\boldsymbol{0}, \boldsymbol{V}(f,n)), \tag{6}$$

where $\boldsymbol{V}(f,n) = \mathrm{diag}[v_1(f,n),\dots,v_I(f,n)]$. From (3) and (5), we can show that $\boldsymbol{x}(f,n)$ follows

$$\boldsymbol{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{x}(f,n)|\boldsymbol{0}, (\boldsymbol{W}^{\mathsf{H}}(f))^{-1}\boldsymbol{V}(f,n)\boldsymbol{W}(f)^{-1}). \tag{7}$$

Hence, the log-likelihood of the separation matrices $\mathcal{W} = \{\boldsymbol{W}(f)\}_f$ and source model parameters $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$ given the observed mixture signals $\mathcal{X} = \{\boldsymbol{x}(f,n)\}_{f,n}$ is given by

$$\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) \overset{c}{=} 2N\sum_f \log|\det \boldsymbol{W}^{\mathsf{H}}(f)| - \sum_{f,n,j}\Big(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)|^2}{v_j(f,n)}\Big), \tag{8}$$

where $=^c$ denotes equality up to constant terms. (8) will be split into frequency-wise source separation problems if there is no additional constraint imposed on $v_j(f,n)$. This indicates that there is a permutation ambiguity in the separated components for each frequency.

2.2. Multichannel VAE

To eliminate the permutation ambiguity during the estimation of $\mathcal{W}$, MVAE trains a conditional VAE (CVAE) to model the complex spectrograms $\boldsymbol{S} = \{s(f,n)\}_{f,n}$ of source signals so that the spectral structures can be captured. CVAE consists of an encoder network $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ and a decoder network $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$, where the network parameters ϕ and θ are trained jointly using a set of labeled training samples $\{\boldsymbol{S}_m, c_m\}_{m=1}^M$. Here, $c = \{1,2,\dots,C\}$ denotes the corresponding class label indicating to which class the spectrogram $\boldsymbol{S}$ belongs. For example, if we consider speaker identities as the class category, c will be associated with a different speaker. CVAE is trained by maximizing the following variational lower bound

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\log p_\theta(\boldsymbol{S}|\boldsymbol{z},c)] - KL[q_\phi(\boldsymbol{z}|\boldsymbol{S},c)||p(\boldsymbol{z})]], \tag{9}$$

where $\mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\cdot]$ denotes the sample mean over the training examples $\{\boldsymbol{S}_m, c_m\}_{m=1}^M$ and $KL[\cdot||\cdot]$ denotes Kullback–Leibler divergence. Here, the encoder distribution $q_\phi(\boldsymbol{z}|\boldsymbol{S},c)$ and the prior distribution of the latent space variable $p(\boldsymbol{z})$ are expressed as Gaussian distributions

$$q_\phi(\boldsymbol{z}|\boldsymbol{S},c) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{S},c), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c))), \tag{10}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \tag{11}$$

where $\boldsymbol{\mu}_\phi(\boldsymbol{S},c)$, $\boldsymbol{\sigma}_\phi^2(\boldsymbol{S},c)$ are the encoder network outputs. The decoder distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$ is defined as a zero-mean complex proper Gaussian distribution and a scale parameter g is incorporated to eliminate the energy difference between the normalized training data and test data. Hence, the decoder distribution is expressed as

$$p_\theta(\boldsymbol{S}|\boldsymbol{z},c,g) = \prod_{f,n}\mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \tag{12}$$

$$v(f,n) = g\cdot\sigma_\theta^2(f,n;\boldsymbol{z},c), \tag{13}$$

where $\sigma_\theta^2(f,n;\boldsymbol{z},c)$ denotes the (f,n)-th element of the decoder output. It is worth mentioning that the decoder distribution (12) is given in the same form as the LGM (5) so that the trained decoder distribution can be used as a universal generative model with the ability to generate complex spectrograms belonging to all the source classes involved in the training examples. If we use $p_\theta(\boldsymbol{S}_j|\boldsymbol{z}_j,c_j,g_j)$ to express the generative model of the complex spectrogram of the source j, a stationary point of the log-likelihood (8) that we want to maximize can be searched by iteratively updating (A) the separation matrices $\mathcal{W}$ using the iterative projection (IP) method [15], (B) the CVAE source model parameters $\Psi = \{\boldsymbol{z}_j, c_j\}_j$ with backpropagation, and (C) the global scale parameter $\mathcal{G} = \{g_j\}_j$ with the following update rule

$$g_j \leftarrow \frac{1}{FN}\sum_{f,n}\frac{|y_j(f,n)|^2}{\sigma_\theta^2(f,n;\boldsymbol{z}_j,c_j)}, \tag{14}$$

where $y_j(f,n) = \boldsymbol{w}_j^{\mathsf{H}}(f)\boldsymbol{x}(f,n)$. Note that since the class labels are the model parameters estimated during the optimization, MVAE is also able to perform source classification.

3. PROPOSED METHOD: FAST MVAE

While MVAE is noteworthy in that it works reasonably well for source separation and has the capability to perform source classification simultaneously, there is still huge room for improvement in the source classification performance. With a regular CVAE, which imposes no restrictions on the way in which the encoder and decoder may use the class labels, the encoder and decoder are free to ignore c by finding a distribution that satisfies $q_\phi(\boldsymbol{z}|\boldsymbol{S},c) = q_\phi(\boldsymbol{z}|\boldsymbol{S})$ and $p_\theta(\boldsymbol{S}|\boldsymbol{z},c) = p_\theta(\boldsymbol{S}|\boldsymbol{z})$. As a result, c will have little effect on the generation of source spectrograms and that will limit source classification performance. To avoid such situations, this paper proposes using an auxiliary classifier VAE [16] for learning the generative distribution $p_\theta(\boldsymbol{S}|\boldsymbol{z},c)$.

3.1. Auxiliary classifier VAE

Auxiliary classifier VAE (ACVAE) [16] is a variant of CVAE that incorporates information-theoretic regularization [17] to assist the decoder outputs to be correlated as far as possible with the class labels c by maximizing the mutual information between c and $\boldsymbol{S} \sim p_\theta(\boldsymbol{S}|\boldsymbol{z}, c)$ conditioned on $\boldsymbol{z}$. The mutual information is expressed as

$$\begin{aligned} & I(c, \boldsymbol{S}|\boldsymbol{z}) \\ & = \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(c|\boldsymbol{S})}[\log p(c'|\boldsymbol{S})] + H(c), \end{aligned} \quad (15)$$

where $H(c)$ represents the entropy of c that can be considered as a constant term. However, it is difficult to optimize $I(c, \boldsymbol{S}|\boldsymbol{z})$ directly since it requires access to the posterior $p(c|\boldsymbol{S})$. Fortunately, we can obtain a variational lower bound of the first term of $I(c, \boldsymbol{S}|\boldsymbol{z})$ by using a variational distribution $r(c|\boldsymbol{S})$ to approximate $p(c|\boldsymbol{S})$:

$$\begin{aligned} & \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(\boldsymbol{S}|c)}[\log p(c'|\boldsymbol{S})] \\ =& \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(\boldsymbol{S}|c)}[\log \frac{r(c'|\boldsymbol{S})p(c'|\boldsymbol{S})}{r(c'|\boldsymbol{S})}] \\ \geq& \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c), c'\sim p(\boldsymbol{S}|c)}[\log r(c'|\boldsymbol{S})] \\ =& \mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r(c|\boldsymbol{S})], \end{aligned} \quad (16)$$

the equality of which holds if $r(c|\boldsymbol{S}) = p(c|\boldsymbol{S})$. Therefore, we can indirectly maximize $I(c, \boldsymbol{S}|\boldsymbol{z})$ by increasing the lower bound with respect to $p_\theta(\boldsymbol{S}|\boldsymbol{z}, c)$ and $r(c|\boldsymbol{S})$. One way to realize this involves expressing the variational distribution as a neural network $r_\psi(c|\boldsymbol{S})$ and training it along with $q_\phi(\boldsymbol{z}|\boldsymbol{S}, c)$ and $p_\theta(\boldsymbol{S}|\boldsymbol{z}, c)$. $r_\psi(c|\boldsymbol{S})$ is called an auxiliary classifier. Therefore, the regularization term that we would like to maximize with respect to ϕ, θ, ψ becomes

$$\begin{aligned} & \mathcal{L}(\phi, \theta, \psi) \quad (17) \\ & = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c), q_\phi(\boldsymbol{z}|\boldsymbol{S},c)}[\mathbb{E}_{c\sim p(c), \boldsymbol{S}\sim p_\theta(\boldsymbol{S}|\boldsymbol{z},c)}[\log r_\psi(c|\boldsymbol{S})]], \end{aligned}$$

where $\sum_{k=1}^{C} r_\psi(c = k|S) = 1$. In the regularization term (17), the auxiliary classifier is trained only using reconstructed spectrograms, which is undesirable since it may limit the capability of the trained classifier of classifying the original spectrograms. To remedy this, ACVAE also includes the cross-entropy

$$\mathcal{I}(\psi) = \mathbb{E}_{(\boldsymbol{S},c)\sim p_D(\boldsymbol{S},c)}[\log r_\psi(c|\boldsymbol{S})] \quad (18)$$

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

$$\mathcal{J}(\phi, \theta) + \lambda_{\mathcal{L}}\mathcal{L}(\phi, \theta, \psi) + \lambda_{\mathcal{I}}\mathcal{I}(\psi), \quad (19)$$

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

3.2. Fast algorithm

Note that the auxiliary classifier $r_\psi(c|\boldsymbol{S})$ both assists the encoder and decoder to learn a more disentangled representation, and provides an alternative to the backpropagation process in the original MVAE optimization, which significantly reduces the computational time. Specifically, since the maximum of the distribution $p(\boldsymbol{z}_j, c_j|\boldsymbol{S}_j) = p(\boldsymbol{z}_j|\boldsymbol{S}_j, c_j)p(c_j|\boldsymbol{S}_j)$ searched with backpropagation in the step (B) can be approximately obtained with the trained auxiliary classifier distribution and the encoder distribution $p(\boldsymbol{z}_j, c_j|\boldsymbol{S}_j) \approx p_\theta(\boldsymbol{z}|\boldsymbol{S}_j, c_j)r_\psi(c_j|\boldsymbol{S}_j)$, we can replace the step (B) with the forward calculation of the encoder and the auxiliary classifier. The proposed fast algorithm is summaried as follows:

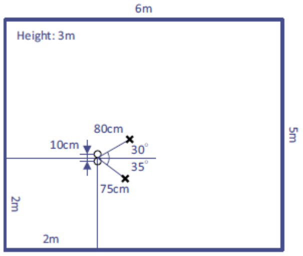

Fig. 1. Configuration of the room where ○ and × represent the position of microphones and sources respectively.

1. Train ϕ, θ and ψ using (19).
2. Initialize $\mathcal{W}$.
3. Iterate the following steps for each j:
   (a) Calculate the temporarily signals $\boldsymbol{S}_j = \{\boldsymbol{w}(f)_j^{\mathsf{H}}\boldsymbol{x}(f, n)\}_f$.
   (b) Update $c_j \leftarrow \mathrm{argmax}_{c_j\in\{1,2,\ldots,C\}}\, r_\psi(c_j|\boldsymbol{S}_j)$.
   (c) Update $\boldsymbol{z}_j \leftarrow \boldsymbol{\mu}_\phi(\boldsymbol{S}_j, c_j)$.
   (d) Update g_j using (14).
   (e) Update $\boldsymbol{w}_j(0), \ldots, \boldsymbol{w}_j(F)$ using IP.

4. EXPERIMENTS

To evaluate the effect of incorporating an auxiliary classifier into both the source model training and the optimization process, we conducted experiments designed to compare the multi-speaker separation performance, source classification accuracies and computational times of fMVAE and the conventional methods, i.e., ILRMA [5, 6] and MVAE [11].

4.1. Experimental conditions

We excerpted speech utterances from two male speakers (‘SM1’, ‘SM2’) and two female speakers (‘SF1’, ‘SF2’) from the Voice Conversion Challenge (VCC) 2018 dataset [18]. The audio files for each speaker were about 7 minutes long and manually segmented into 116 short sentences, where 81 and 35 sentences (about 5 and 2 minutes long, respectively) were used as training and test sets, respectively. The mixture signals were created by using simulated two-channel recordings of two sources where the room impulse responses were synthesized using the image method. We tested two different reverberant conditions where the reverberation times (RT_{60}) were set at 78 and 351 ms, respectively. Fig. 1 shows the configuration of the room. We generated test data involving 4 speaker pairs and 10 sentences for each pair, namely there were a total of 40 test signals, each of which was about 4 to 7 seconds long. All the speech signals were resampled at 16 kHz. The STFT was computed using a Hamming window that was 256 ms long and the window shift was128 ms.

ILRMA was run for 100 iterations and both the proposed method and MVAE were run for 40 iterations. To initialize $\mathcal{W}$

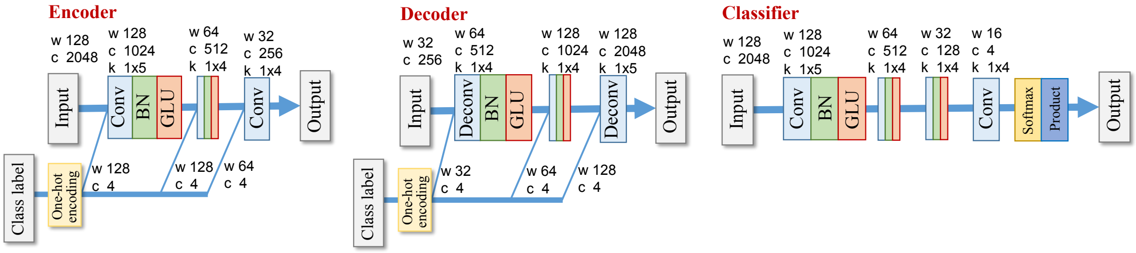

Fig. 2. Network architectures of the encoder and decoder used for MVAE and fMVAE and the classifier used for fMVAE. The inputs and outputs are 1-dimensional data, where the frequency dimension of spectrograms is regarded as the channel dimension. “w”, “c” and “k” denote the width, channel number and kernel size, respectively. “Conv”, “Deconv”, “BN” and “GLU” denote 1-dimensional convolution and deconvolution, batch normalization, gated linear unit, respectively.

Table 1. Average SDR, SIR and SAR scores of ILRMA, MVAE and fMVAE. The bold font shows the highest scores.

method	RT_{60} = 78 ms		
	SDR [dB]	SIR [dB]	SAR [dB]
ILRMA	14.8997	21.3277	18.0584
MVAE	21.5912	27.2663	**25.1616**
fMVAE	**22.5976**	**29.8476**	24.8967
method	RT_{60} = 351 ms		
	SDR [dB]	SIR [dB]	SAR [dB]
ILRMA	4.6840	11.6284	7.2364
MVAE	**8.3157**	**18.0834**	**9.2206**
fMVAE	6.7814	15.7728	7.7883

Table 2. Computational times of MVAE, fMVAE and ILRMA. MVAE and fMVAE were initialized with run ILRMA algorithm for 30 iterations in CPU and run 40 iterations of the optimization algorithms in CPU or GPU. ILRMA runs 100 iterations in CPU.

	rumtime/iteration[sec]	total [sec]
MVAE (GPU)	6.071632	260.5953
fMVAE (CPU)	0.389762	21.54129
fMVAE (GPU)	**0.097272**	**17.56694**
ILRMA (CPU)	0.113571	18.38221

for MVAE and fMVAE, we used ILRMA, which we ran for 30 iterations. Adam [19] was used for training CVAE and ACVAE, and estimating the latent variables in MVAE. The network architectures used for CVAE and ACVAE are shown in Fig. 2. Note that we used the same network architectures for CVAE and ACVAE as we used for the encoder and decoder. All the networks were designed to be fully convolutional with gated linear units [20] so that the inputs were allowed to have arbitrary lengths. The programs were run using an Intel (R) Core i7-6800K CPU@3.40 GHz and a GeForce GTX 1080Ti GPU.

4.2. Results

We calculated the average signal-to-distortion ratios (SDR), signal-to-interference ratios (SIR) and signal-to-artifact ratios (SAR) [21] over the 40 test signals to evaluate the source separation performance and measured the computational times. Table 1 and Table 2 show the source separation results obtained under different RT_{60} conditions and the computational times of ILRMA, MVAE and fMVAE, respectively. fMVAE was about 15 times faster than the original MVAE and even lightly faster than ILRMA in GPU machines. Furthermore, it is noteworthy that fMVAE achieved source separation performance comparable to that of the original MVAE. As the results show, MVAE and fMVAE significantly outperformed ILRMA in terms of all the source separation performance criteria, which confirmed the effect of the incorporation of the CVAE source model. fMVAE obtained a higher SDR and SIR than MVAE in a slightly reverberant environment, but the performance deteriorated slightly when RT_{60} became long. It is interesting to further compare fMVAE with MVAE in highly reverberant environments, which is one direction for our future work.

Table 3. Accuracy rates of source classification obtained with MVAE and fMVAE.

	all iterations	final estimation
MVAE	27.91%	37.50%
fMVAE	**78.63%**	**80.00%**

To evaluate the performance of source classification, we computed the classification accuracy rates over the results estimated in each iteration and in the final estimation alone. Table 3 provides results showing that fMVAE significantly improved source classification accuracy achieving an 80% accuracy rate. Our future work will also include further improving the source classification accuracy.

5. CONCLUSIONS

This paper proposed a fMVAE method that (i) uses an auxiliary classifier VAE instead a regular CVAE to learn the generative distribution of source signals; (ii) employs a trained auxiliary classifier and encoder for optimization. fMVAE allows us to significantly reduce the computational time and improve the source classification performance. The results revealed that fMVAE was about 15 times faster than the original MVAE and achieved a source classification accuracy rate of about 80% with notable source separation performance.

6. REFERENCES

[1] 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.

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

[3] 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, 2010.

[4] 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.

[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*, pp. 245–253, 2010.

[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. 1622–1637, 2016.

[7] 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.

[8] 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.

[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*, pp. 1571–1575, 2018.

[10] S. Leglaive, L. Girin and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. MLSP*, 2018.

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

[12] S. Seki, H. Kameoka, L. Li, T. Toda and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," *eprint arXiv: 1810.00223*, Oct. 2018.

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

[14] K. Sohn, H. Lee and X. Yan, "Learning structured output representation using deep conditional generative models," in *Adv. Neural Information Processing Systems (NIPS)*, pp. 3483–3491, 2015.

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

[16] H. Kameoka, T. Kaneko, K. Tanaka and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *eprint arXiv: 1808.05092*, Aug. 2018.

[17] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Adv. Neural Information Processing Systems (NIPS)*, pp. 2172–2180, 2016.

[18] 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," *eprint arXiv: 1804.04262*, Apr. 2018.

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

[20] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, pp. 933–941, 2017.

[21] 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.