Generalized Multichannel Variational Autoencoder for Underdetermined Source Separation

Shogo Seki[1], Hirokazu Kameoka[2], Li Li[3], Tomoki Toda[1], Kazuya Takeda[1]

[1]*Nagoya University, Nagoya, Japan*
[2]*NTT Communication Science Laboratories, Atsugi, Japan*
[3]*University of Tsukuba, Tsukuba, Japan*
seki.shogo@g.sp.m.is.nagoya-u.ac.jp, kameoka.hirokazu@lab.ntt.co.jp

***Abstract*—This paper deals with a multichannel audio source separation problem under underdetermined conditions. Multichannel Non-negative Matrix Factorization (MNMF) is one of the powerful approaches, which adopts the NMF concept for source power spectrogram modeling. It works reasonably well for particular types of sound sources, however, one limitation is that it can fail to work for sources with spectrograms that do not comply with the NMF model. To address this limitation, a novel technique called the Multichannel Variational Autoencoder (MVAE) method was recently proposed, where a Conditional VAE (CVAE) is used instead of the NMF model for source power spectrogram modeling. This approach has shown to perform impressively in determined source separation tasks thanks to the representation power of DNNs. This paper generalizes MVAE originally formulated under determined mixing conditions so that it can also deal with underdetermined cases. The proposed method was evaluated on an underdetermined source separation task of separating out three sources from two microphone inputs. Experimental results revealed that the generalized MVAE method achieved better performance than the conventional MNMF method.**

***Index Terms*—Underdetermined source separation, Variational audoencoder, Non-negative matrix factorization**

I. Introduction

Blind source separation (BSS) refers to a problem of separating out individual source signals from microphone array inputs where the transfer functions between the sources and microphones are unknown. The frequency-domain BSS approach allows the utilization of various models for the time-frequency representations of source signals and/or array responses. For example, Independent Vector Analysis (IVA) [1], [2] offers a way of jointly solving frequency-wise source separation and permutation alignment under the assumption that the magnitudes of the frequency components originating from the same source are likely to vary coherently over time.

Other approaches involve multichannel extensions of Non-negative Matrix Factorization (NMF) [3]–[6]. NMF was originally applied to music transcription and monaural source separation tasks [7], [8] where the idea is to interpret the power spectrogram of a mixture signal and approximate it as the product of two non-negative matrices. This can be viewed as approximating the power spectrum of a mixture signal observed at each time frame by the sum of basis spectra scaled by time-varying magnitudes. Multichannel NMF (MNMF) is an extension of this approach to a multichannel case that allows for the use of spatial information. It can also be seen as an approach to frequency-domain BSS using spectral templates as a clue for jointly solving frequency-wise source separation and permutation alignment.

TABLE I: Comparison with the conventional methods

Method	Separation	Source model
ILRMA [4], [6]	Determined	NMF
MNMF [3], [5]	Underdetermined	NMF
MVAE [10]	Determined	VAE
Proposed	Underdetermined	VAE

The original MNMF [3] was formulated under a general problem setting where sources can outnumber microphones and a determined version of MNMF was subsequently proposed in [4]. While the determined version is applicable only to determined cases, it allows an implementation of a significantly faster algorithm than the general version. The determined MNMF framework was later called Independent Low-Rank Matrix Analysis (ILRMA) [9]. The MNMF framework including ILRMA is notable in that the optimization algorithm is guaranteed to converge, however, one limitation is that it can fail to work for sources with spectrograms that do not comply with the NMF model.

To address this limitation, a technique called the Multichannel Variational Autoencoder (MVAE) method was recently proposed in [10]. It is an extension of ILRMA in which a Variational Autoencoder (VAE) [11] is used instead of the NMF model to estimate the power spectrograms of the sources in a mixture. Several studies [12], [13] have demonstrated that the VAE-based power spectrogram modeling is effective in a speech enhancement task. MVAE, which focuses on a source separation task, allows the estimation of the separation matrices by employing a single Conditional VAE (CVAE) [14], trained using the spectrograms of speech samples with speaker ID labels, as a generative model of the speech spectrograms of multiple speakers. This approach is noteworthy in that it can exploit the benefits of the representation power of neural networks for source power spectrogram modeling and has shown to outperform ILRMA on a determined source separation task.

While the original MVAE method was formulated under determined mixing conditions, this paper generalizes it so that it can also deal with underdetermined cases (TABLE I).

978-9-0827-9703-9/19/$31.00

From the formulation assuming a mixing system, we derive a separation algorithm. Experimental results demonstrate that the proposed method outperforms the conventional MNMF method.

II. Problem Formulation

We consider a situation where J source signals are observed by I microphones. Let $s_j(f,n)$ and $x_i(f,n)$ be the Short-Time Fourier Transform (STFT) coefficient of the j-th source signal and the i-th observed signal, where f and n are the frequency and time indices, respectively. We denote the vectors containing $s_1(f,n),\cdots,s_J(f,n)$ and $x_1(f,n),\cdots,x_I(f,n)$ by

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

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

where $(\cdot)^{\mathsf{T}}$ denotes the transpose. Now, we use a mixing system of the form

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

$$\mathbf{A}(f) = [\mathbf{a}_1(f),\cdots,\mathbf{a}_J(f)] \in \mathbb{C}^{I\times J}, \tag{4}$$

to describe the relationship between $\mathbf{s}(f,n)$ and $\mathbf{x}(f,n)$ where $\mathbf{A}(f)$ is called the mixing matrix.

Here, we assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n)$

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

(5) is called the Local Gaussian Model (LGM) [15]. When $s_j(f,n)$ and $s_{j'}(f,n)$ are independent for $j \neq j'$, $\mathbf{s}(f,n)$ follows

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

where $\mathbf{V}(f,n)$ is a diagonal matrix with diagonal entries $v_1(f,n),\cdots,v_J(f,n)$. From (3) and (6), $\mathbf{x}(f,n)$ is shown to follow

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, \mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)), \tag{7}$$

where $(\cdot)^{\mathsf{H}}$ denotes the conjugate transpose. Thus, given the observed mixture signals $\mathcal{X} = \{\mathbf{x}(f,n)\}_{f,n}$, using the mixing matrices $\mathcal{A} = \{\mathbf{A}(f)\}_f$ and the variances of source signals $\mathcal{V} = \{v_j(f,n)\}_{j,f,n}$, the log-likelihood is given by

$$\begin{aligned}\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) &\stackrel{c}{=} \\ -\sum\nolimits_{f,n}&\Big[\mathrm{tr}(\mathbf{x}^{\mathsf{H}}(f,n)\left(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)\right)^{-1}\mathbf{x}(f,n)) \\ &+\log\det\left(\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f)\right)\Big],\end{aligned} \tag{8}$$

where $\stackrel{c}{=}$ denotes the equality up to constant terms. If there is no constraint imposed on $v_j(f,n)$, (8) will be split into frequency-wise source separation problems. This indicates that there is a permutation ambiguity in the separated components for each frequency since permutation of j does not affect the value of the log-likelihood. Thus, we usually need to perform permutation alignment after $\mathcal{A}$ is obtained.

III. Related Work

A. MNMF

The spatial covariance of the observed mixture signal can be rewritten as the linear sum of the outer products of $\mathbf{a}_j(f)$ multiplied by $v_j(f,n)$:

$$\begin{aligned}\mathbf{A}(f)\mathbf{V}(f,n)\mathbf{A}^{\mathsf{H}}(f) &= \sum\nolimits_j \mathbf{a}_j(f)v_j(f,n)\mathbf{a}_j^{\mathsf{H}}(f) \\ &= \sum\nolimits_j v_j(f,n)\mathbf{R}_j(f)\left(=\hat{\mathbf{X}}(f,n)\right),\end{aligned} \tag{9}$$

where $\mathbf{R}_j(f)$ represents the spatial covariance of source j. As with IVA, MNMF makes it possible to jointly solve frequency-wise source separation and permutation alignment by imposing a constraint on $v_j(f,n)$. Specifically, $v_j(f,n)$ is modeled as the linear sum of K_j spectral templates $h_{j,1}(f),\cdots,h_{j,K_j}(f) \geq 0$ scaled by time-varying activations $u_{j,1}(n),\cdots,u_{j,K_j}(n) \geq 0$:

$$v_j(f,n) = \sum\nolimits_{k=1}^{K_j} h_{j,k}(f)u_{j,k}(n). \tag{10}$$

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

$$v_j(f,n) = \sum\nolimits_{k=1}^{K} b_{j,k}h_k(f)u_k(n), \tag{11}$$

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

The optimization algorithm of MNMF consists of iteratively updating the spatial covariance matrices $\mathcal{R} = \{\mathbf{R}_j(f)\}_{j,f}$, and the source models $\mathcal{H}_1 = \{h_{j,k}(f)\}_{j,k,f}$, $\mathcal{U}_1 = \{u_{j,k}(n)\}_{j,k,n}$ or $\mathcal{B} = \{b_{j,k}\}_{j,k}$, $\mathcal{H}_2 = \{h_k(f)\}_{k,f}$, $\mathcal{U}_2 = \{u_k(n)\}_{k,n}$. We can derive update equations using the principle of the Majorization-Minimization (MM) algorithm [16].

B. ILRMA

ILRMA is a special class of MNMF designed to solve determined source separation problems. Unlike MNMF, which uses the mixing system (3), ILRMA uses a separation system of the form

$$\mathbf{s}(f,n) = \mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n), \tag{12}$$

$$\mathbf{W}(f) = [\mathbf{w}_1(f),\cdots,\mathbf{w}_J(f)] \in \mathbb{C}^{I\times J}, \tag{13}$$

assuming the mixing matrix is invertible. The inverse matrix $\mathbf{W}^{\mathsf{H}}(f)$ is called the separation matrix. From (6) and (12), $\mathbf{x}(f,n)$ is shown to follow

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}(f,n)|\mathbf{0}, (\mathbf{W}^{\mathsf{H}}(f))^{-1}\mathbf{V}(f,n)(\mathbf{W}(f))^{-1}). \tag{14}$$

Given the observed signals $\mathcal{X}$, using the separation matrices $\mathcal{W} = \{\mathbf{W}(f)\}_f$ and $\mathcal{V}$, the log-likelihood is given by

$$\begin{aligned}\log p(\mathcal{X}|\mathcal{W},\mathcal{V}) &\stackrel{c}{=} 2N\sum\nolimits_f \log|\det\mathbf{W}^{\mathsf{H}}(f)| \\ &- \sum\nolimits_{f,n,j}\left[\log v_j(f,n) + \frac{|\mathbf{w}_j^{\mathsf{H}}(f)\mathbf{x}(f,n)|^2}{v_j(f,n)}\right],\end{aligned} \tag{15}$$

where $v_j(f,n)$ is modeled as (10) or (11) as with MNMF.

As with MNMF, we can derive the MM-based update equations for $\mathcal{H}_1$, $\mathcal{U}_1$ or $\mathcal{B}$, $\mathcal{H}_2$, $\mathcal{U}_2$. Since ILRMA is a natural extension of IVA, we can use a fast update rule called the Iterative Projection (IP) [17] for the separation matrices, originally developed for IVA.

C. MVAE

One limitation of the MNMF framework including ILRMA is that since $v_j(f,n)$ is restricted to (10) or (11), it can fail to work for sources with spectrograms that do not actually follow this form. The MVAE method is an extension of ILRMA that replaces (10) with a pretrained CVAE. Let $\tilde{\mathbf{S}} = \{s(f,n)\}_{f,n}$ be the complex spectrogram of a particular sound source. MVAE models the generative model of $\tilde{\mathbf{S}}$ using a CVAE with an auxiliary input c. Here, we assume that c is represented as a one-hot vector, indicating the class of a source. Thus, the elements of c must sum to unity. For example, if we consider speaker identities as the source class, each element of c will be associated with a different speaker.

The CVAE consists of an encoder network and a decoder network, which are assumed to be trained using labeled training examples $\{\tilde{\mathbf{S}}_m, c_m\}_{m=1}^M$ prior to separation. The encoder distribution $q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c)$ is expressed as a Gaussian distribution:

$$q_\phi(\mathbf{z}|\tilde{\mathbf{S}}, c) = \prod\nolimits_k \mathcal{N}(z(k)|\mu_\phi(k;\tilde{\mathbf{S}}, c), \sigma^2_\phi(k;\tilde{\mathbf{S}}, c)), \tag{16}$$

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

$$p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g) = \prod\nolimits_{f,n} \mathcal{N}_{\mathbb{C}}(s(f,n)|0, v(f,n)), \tag{17}$$

$$v(f,n) = g \cdot \sigma^2_\theta(f,n;\mathbf{z},c), \tag{18}$$

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

$$\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_{\mathrm{D}}(\tilde{\mathbf{S}},c)} \Big[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\tilde{\mathbf{S}},c)}[\log p(\tilde{\mathbf{S}}|\mathbf{z},c)] - \mathrm{KL}[q(\mathbf{z}|\tilde{\mathbf{S}},c)||p(\mathbf{z})]\Big], \tag{19}$$

where $\mathbb{E}_{(\tilde{\mathbf{S}},c)\sim p_{\mathrm{D}}(\tilde{\mathbf{S}},c)}[\cdot]$ denotes the sample mean over the training examples and $\mathrm{KL}[\cdot||\cdot]$ is the Kullback-Leibler divergence.

The trained decoder distribution $p_\theta(\tilde{\mathbf{S}}|\mathbf{z}, c, g)$ is considered as a universal generative model that is capable of generating spectrograms of all the sources involved in the training examples. MVAE employs the decoder part of the CVAE as the source model $v_j(f,n)$ in (15) and treats the input $\mathbf{z}$ and c to the decoder as the model parameters to be estimated. The optimization algorithm of MVAE consists of updating the separation matrices using IP, the global scale using the MM algorithm, and the input to the pretrained decoder using backpropagation. The advantage of the MVAE is that it can leverage the strong representation power of VAE for source power spectrogram modeling.

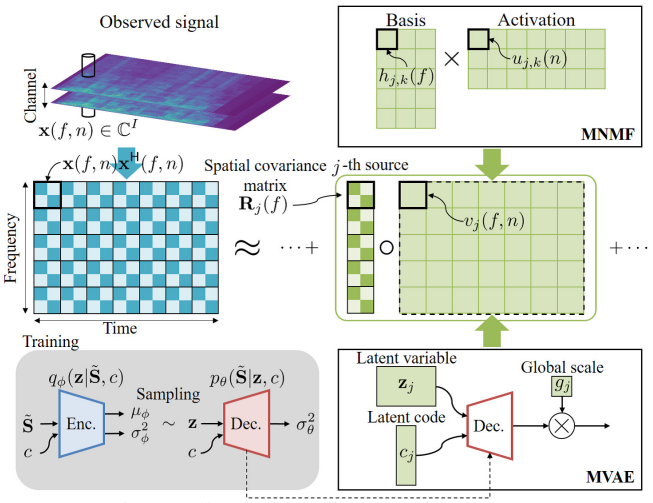


Fig. 1: Illustration of generalized MVAE

IV. Generalized MVAE

Fig. 1 shows an illustration of generalized MVAE and MNMF with the source model given by (10). As with the original MVAE method, we use the decoder network of the pretrained CVAE as the generative model of source power spectrograms.

Since the decoder distribution is given in the same form as the LGM, we can use $p_\theta(\tilde{\mathbf{S}}_j|\mathbf{z}_j, c_j, g_j)$ to develop the log-likelihood of the form (8). Hence, we can derive an iterative algorithm for estimating $\mathcal{R}$, $\mathcal{G} = \{g_j\}_j$, and $\Psi = \{\mathbf{z}_j, c_j\}_j$ in the same way as the derivation of the MM-based algorithm for MNMF. From [16], we can show the following inequality:

$$\begin{aligned}\mathcal{L} &= -\log p(\mathcal{X}|\mathcal{A},\mathcal{V}) \\ &\overset{c}{\leq} \sum\nolimits_j \sum\nolimits_{f,n} \Big[\frac{\mathrm{tr}(\mathsf{X}(f,n)\mathbf{P}_j(f,n)\mathbf{R}_j^{-1}(f)\mathbf{P}_j(f,n))}{v_j(f,n)} \\ &\qquad + v_j(f,n)\mathrm{tr}(\mathbf{K}^{-1}(f,n)\mathbf{R}_j(f))\Big],\end{aligned} \tag{20}$$

where the equality holds when

$$\mathbf{P}_j(f,n) = v_j(f,n)\mathbf{R}_j(f)\Big(\sum\nolimits_j v_j(f,n)\mathbf{R}_j(f)\Big)^{-1}, \tag{21}$$

$$\mathbf{K}(f,n) = \mathsf{X}(f,n) = \mathbf{x}(f,n)\mathbf{x}^{\mathsf{H}}(f,n). \tag{22}$$

Thus, we can use the right-hand side of (20) as a majorizer of $\mathcal{L}$ where $\mathcal{P} = \{\mathbf{P}_j(f,n)\}_{j,f,n}$ and $\mathcal{K} = \{\mathbf{K}(f,n)\}_{f,n}$ are auxiliary variables. An iterative algorithm consists of minimizing this majorizer with respect to $\mathcal{R}$, $\mathcal{G}$, and Ψ and updating $\mathcal{P}$ and $\mathcal{K}$ at (21) and (22). The optimal update of $\mathcal{R}$ is analytically obtained as

$$\mathbf{R}_j(f) \leftarrow \mathbf{\Lambda}_j^{-1}(f)\#(\mathbf{R}_j(f)\mathbf{\Omega}_j(f)\mathbf{R}_j(f)), \tag{23}$$

where $\#$ denotes the geometric mean of two positive semidefinite matrices [18]:

$$\mathbf{G}\#\mathbf{H} = \mathbf{G}^{\frac{1}{2}}(\mathbf{G}^{-\frac{1}{2}}\mathbf{H}\mathbf{G}^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{G}^{\frac{1}{2}}. \tag{24}$$

$\mathbf{\Lambda}_j(f)$, $\mathbf{\Omega}_j(f)$ are given as follows:

$$\mathbf{\Lambda}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathsf{X}}^{-1}(f,n), \tag{25}$$

$$\mathbf{\Omega}_j(f) = \sum\nolimits_n v_j(f,n)\hat{\mathsf{X}}^{-1}(f,n)\mathsf{X}(f,n)\hat{\mathsf{X}}^{-1}(f,n). \tag{26}$$

Algorithm 1 MVAE algorithm

Train ϕ and θ with (19)
Initialize $\mathcal{R}$, Ψ, and $\mathcal{G}$
repeat
  for each j **do**
    Update $\mathcal{R}_j = \{\mathbf{R}_j(f)\}_f$ using (24)
    Update $\psi_j = \{\mathbf{z}_j, c_j\}$ with (20) using backpropagation
    Update g_j using (28)
  end for
until converge

Since the majorizer is split into source-wise terms, Ψ can be updated parallelly using backpropagation. Note that we must take account of the sum-to-one constraints when updating c_j. This can be easily implemented by inserting an appropriately designed softmax layer that outputs c_j

$$c_j = \mathrm{softmax}(d_j), \tag{27}$$

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

$$g_j \leftarrow g_j \times \sqrt{\frac{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathsf{X}}^{-1}(f,n)\mathsf{X}(f,n)\hat{\mathsf{X}}^{-1}(f,n)\mathbf{R}_j(f))}{\sum_{f,n} \sigma_\theta^2(f,n;\mathbf{z}_j,c_j)\mathrm{tr}(\hat{\mathsf{X}}^{-1}(f,n)\mathbf{R}_j(f))}}, \tag{28}$$

The source separation algorithm of the generalized MVAE is summarized as Algorithm 1.

V. Experimental Evaluations

A. Settings

The proposed method was evaluated on an underdetermined source separation task of separating out three sources from two microphone inputs. As the experimental data, we used speech samples of the Voice Conversion Challenge (VCC) 2018 dataset [19], which contains recordings of 6 female and 6 male US English speakers. From the dataset, we used utterances of 2 female and 2 male speakers, 'SF1', 'SF2', 'SM1', and 'SM2'. 81 utterances and 35 utterances of each speaker were used for training and evaluation, respectively.

Fig. 2 shows the room configuration of the evaluation, where ○ and × shows the locations of microphones and sources, respectively. Using the evaluation data, all the 3-speaker patterns were prepared: 'SF1+SF2+SM2', SF1+SM1+SF2', 'SF1+SM1+SM2', and 'SM1+SF2+SM2'. For each speaker pattern, 10 speech mixtures were generated by randomly choosing utterances of individual speakers and randomly placing them at × in Fig. 2. All the speech mixtures were generated at two different reverberant conditions: $T_{60} = 78$ ms and $T_{60} = 351$ ms.

All the speech signals were resampled at 16 kHz and STFT analysis was conducted with 256 ms frame length and 128 ms hop length. We designed the encoder and decoder networks of the CVAE as in Fig. 3. In this experiment, speaker identities are considered as the source class category: latent code c_j shown in Fig. 1 is represented as a four-dimensional one-hot vector. The Adam [20] algorithm with learning rate 0.0002 was used to train the CVAE and the SGD algorithm with learning rate 0.0005 was used to update the VAE source model Ψ.

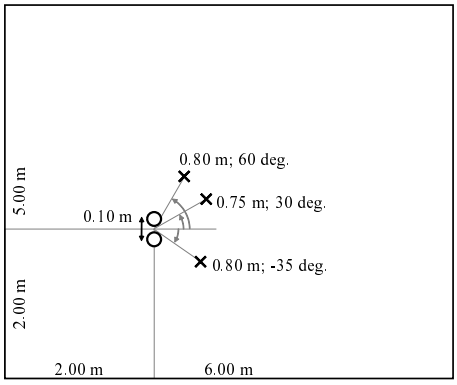


Fig. 2: Room configuration

Following methods including the proposed method were evaluated for comparison.

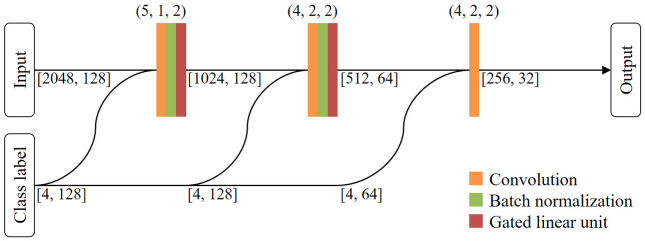
- MNMF1: MNMF with source model given by (10)
- MVAE1: MVAE initialized by MNMF1
- MNMF2: MNMF with source model given by (11)
- MVAE2: MVAE initialized by MNMF2
- SB-MNMF: MNMF2 with spectral dictionaries
- SB-MVAE: MVAE initialized by SB-MNMF

The source separation algorithms were run for 300 iterations for the conventional methods and 100 iterations for the proposed methods. The parameters of the proposed methods were initialized using the MNMF methods run with 200 iterations through the encoder network. The numbers of basis spectra in the all MNMF algorithms were set to 10 per speaker. We also evaluated the MNMF in semi-blind condition, where the spectral dictionaries are trained with the same dataset as the CVAE. The spectral dictionary of each speaker were obtained by Itakura-Saito NMF (IS-NMF) [8] with 1000 iterations.

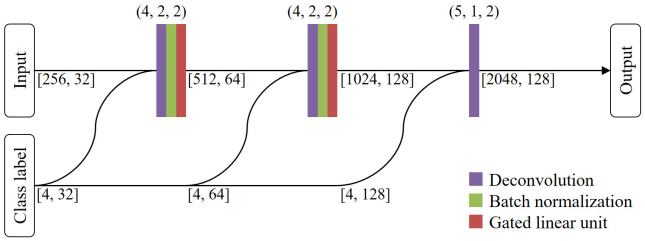
As the evaluation metrics, the Signal-to-Distortion Ratio (SDR), the source Image-to-Spatial distortion Ratio (ISR), the Signal-to-Inference Ratio (SIR), and the Signal-to-Artifact Ratio (SAR) [21] between the reference signals and the separated signals were calculated for each mixture and averaged.

B. Results

The separation performance under each reverberant condition is shown in Fig. 4. We can see that the proposed methods consistently achieves better performances than the baseline methods (MNMF1, MNMF2). Since the difference between the baseline methods and the proposed methods is simply types of source model, these results imply that the use of VAE source model successfully contributed to improving the separation performance. In the comparison with the MNMF in semi-blind condition, the proposed method achieves better SDR performances. Furthermore, the proposed method significantly improves the SIR performances. This results show us that the proposed method using VAE as source model can accurately estimate the spatial covariance and improves suppression performance for the interference.



(a) Encoder network



(b) Decoder network

Fig. 3: Network configurations of (a) encoder and (b) decoder, where $[\mathtt{c}, \mathtt{l}]$ denotes the input channel and frame length. Both convolution and deconvolution represents 1-dimensional operation. $(\mathtt{k}, \mathtt{s}, \mathtt{p})$ represents the kernel size, the stride size along frame, and the zero padding size at both ends, respectively.

VI. Conclusion

This paper proposed generalization of the MVAE method originally formulated under determined conditions so that it can also be applied to underdetermined source separation tasks. The separation algorithm of the proposed method was derived from the formulation assuming a mixing system. Experimental results revealed that the generalized MVAE method achieved better performance than the conventional methods and demonstrated that VAE source model successfully contributed to improving the separation performance.

We plan to compare the proposed method with other DNN-based method [22] and investigate the performance in music source separation.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 17H01763.

References

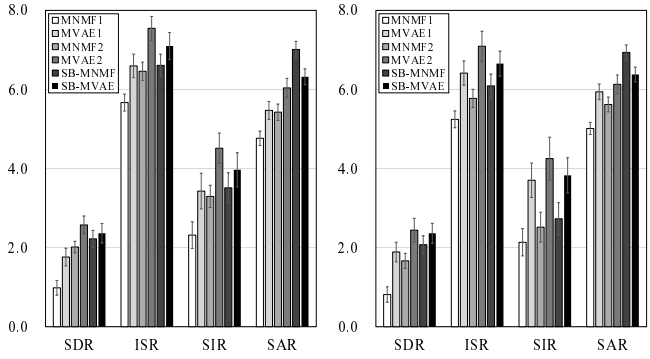
[1] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *Proc. ICA*, pp. 165–172, 2006.

[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. 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.

[5] 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.



(a) $T_{60} = 78$ ms (b) $T_{60} = 351$ ms

Fig. 4: Averaged separation performances [dB] at reverberant conditions of (a) $T_{60} = 78$ ms and (b) $T_{60} = 351$ ms

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

[8] 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.

[9] 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*, Springer, 2018.

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

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.

[12] 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.

[13] 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.

[14] 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.

[15] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, IGI global, 2011.

[16] H. Kameoka, H. Sawada, and T. Higuchi, "General formulation of multichannel extensions of nmf variants," in *Audio Source Separation*, Springer, 2018.

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

[18] K. Yoshii, "Correlated tensor factorization for audio source separation," in *Proc. ICASSP*, pp. 731–735, 2018.

[19] 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*, 2018.

[20] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization," in *Proc. ICLR*, 2015.

[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.

[22] 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.