NCSP'21
RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing 2021
March 1-2, 2021

Single-Channel Muti-speaker Separation via Discriminative Training of Variational Autoencoder Spectrogram Model

Naoya Murashima[1], Hirokazu Kameoka[2], Li Li[1], Shogo Seki[2], Shoji Makino[1]

[1]University of Tsukuba, Japan, E-mail:s1920710@s.tsukuba.ac.jp
[2]NTT Communication Science Laboratories, Japan, E-mail:kameoka.hirokazu@lab.ntt.co.jp

1. Introduction

This paper deals with single-channel speaker-dependent speech separation. Speech separation is a technique to separate out the signal of each speaker from a mixture signal of multiple speakers and can be used to improve the accuracy of speech recognition and the quality of voice communication. A discriminative approach using deep neural network (DNN) has recently proved powerful in single-channel source separation tasks [1–4]. The general idea is to train a DNN that predicts TF masks or TF embeddings from a given mixture signal based on spectro-temporal features. Recently, methods to train a DNN that directly predicts the waveform of each speaker has also been proposed. Although these methods can achieve reasonably good separation, they can fail to work if there is a large mismatch between training and test conditions caused by, for example, reverberation.

Meanwhile, a generative approach, including the non-negative matrix factorization (NMF) method [5], is attractive in its flexibility in addressing the mismatch between training and test conditions. For example, this can be achieved by explicitly incorporating the generative process that causes the mismatch into the generative model of observed signals, and simultaneously estimating the parameters of the entire model during test time. The idea of the NMF method is to approximate the spectrum of a mixture signal observed at each short-term frame as a linear sum of a limited number of basis spectra scaled by time-varying amplitudes. In the supervised NMF (SNMF) method [6], separation is achieved by fitting the basis spectra, pretrained on each source, to an observed mixture signal and then applying a Wiener filter. However, one problem with the SNMF method is that the training criterion for the basis spectra is inconsistent with the objective function at test time. In other words, the basis spectra are not trained so that the separated signals at test time become optimal. The discriminative NMF (DNMF) method [7] was later proposed to solve this inconsistency. Specifically, the idea is to make the training scenario consistent with the test scenario, and train the basis spectra so that the separated signals (the outputs of the Wiener filters) directly become optimal. While these NMF-based methods work reasonably well for particular types of sound sources, one limitation is that they can fail to work for sources with spectrograms that do not comply with the NMF model.

In recent years, with the aim of modeling source spectrograms more flexibly than the NMF model, generative approach-based methods using DNNs have been proposed [8–16]. For multichannel source separation in a determined condition, a method that uses the conditional variational autoencoder (CVAE) [17] for source spectrogram modeling, called the multichannel VAE (MVAE) method, has been proposed. This method has been shown to significantly outperform independent low-rank matrix analysis [18, 19], which uses the NMF model for spectrogram modeling. This indicates that CVAE is better than the NMF model at expressing the spectrogram of each source and correctly discriminating the spectrogram of one source from that of another. With the same motivation, a monaural speech enhancement method (VAE-NMF method) [13, 14] and its multi-channel extension [15, 16] have also been proposed.

Motivated by the success of the MVAE method, we propose in this paper a VAE-based monaural source separation (VASS) method using a CVAE for source spectrogram modeling. We further propose a discriminative counterpart of the VASS method, called the discriminative VASS (DVASS) method, namely an extension to the VASS method equivalent to the extension from the SNMF method to the DNMF method.

2. Conventional methods

2.1 Problem formulation

We consider a situation where a mixture of the signals of J speakers is observed. Let $\mathbf{Y} = \{y(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$, $\mathbf{S}_j = \{s_j(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$ be the complex spectrograms of the observed signal and the signal of the jth speaker, where f and n are the frequency and time indices, respectively. Let us now assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$

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

(1) is called the local Gaussian model (LGM) [20, 21]. When $\mathbf{S}_j$ and $\mathbf{S}_{j'}$ $(j = j')$ are independent, from $\mathbf{Y} = \sum_j \mathbf{S}_j$ $y(f,n)$, we can show that $y(f,n)$ follows

$$y(f,n) \sim \mathcal{N}_{\mathbb{C}}(y(f,n) \mid 0, v(f,n)), \tag{2}$$

where $v(f,n) = \sum_j v_j(f,n)$. when $\mathbf{V}_j = \{v_j(f,n)\}_{f,n}$, the negative log-likelihood $-\log p(\mathbf{Y}|\mathbf{V})$ of $\mathbf{V} = \{\mathbf{V}_1,\ldots,\mathbf{V}_J\}$ given $\mathbf{Y}$ is equivalent up to a constant term to the IS divergence between $\tilde{y}(f,n) = |y(f,n)|^2$ and $v(f,n)$

$$\mathcal{D}_{\mathrm{IS}}(\tilde{\mathbf{Y}}|\mathbf{V}) = \sum_{f,n}\left(\frac{\tilde{y}(f,n)}{v(f,n)} - \log\frac{\tilde{y}(f,n)}{v(f,n)} - 1\right), \tag{3}$$

where $\tilde{\mathbf{Y}} = \{\tilde{y}(f,n)\}_{f,n}$. Since $\mathbf{Y}$ and $\mathbf{S}_1,\ldots,\mathbf{S}_J$ are jointly Gaussian, the minimum mean square error estimator of $\mathbf{S}_j$ given $\mathbf{Y}$ and $\mathbf{V}_1,\ldots,\mathbf{V}_J$ is given by

$$\mathbb{E}[\mathbf{S}_j|\mathbf{Y}] = \frac{\mathbf{V}_j}{\sum_{j'}\mathbf{V}_{j'}} \odot \mathbf{Y}, \tag{4}$$

where $\frac{\cdot}{\cdot}$ and $\odot$ denote elementwise multiplication and division. Note that the multiplicative factor of (4) is called the Wiener mask. (4) implies that once $\mathbf{V}_1,\ldots,\mathbf{V}_J$ is estimated, we can estimate the signal of each speaker. Thus, the single-channel speech separation problem can be formulated as the problem of estimating $\mathbf{V}_1,\ldots,\mathbf{V}_J$ with (3) as the objective function, under some constraint or assumption imposed on $\mathbf{V}_1,\ldots,\mathbf{V}_J$.

2.2 SNMF method

The SNMF method is a monaural speech separation method that uses the NMF model to express $\mathbf{V}_j$. Namely,

$\mathbf{V}_j$ is represented as the product of two non-negative matrices $\mathbf{W}_j\mathbf{H}_j$, i.e., $v_j(f,n) = \sum_k w_{j,k}(f)$. Here, the basis matrix $\mathbf{W}_j$ is assumed to be trained prior to separation using the training utterances of each speaker. Representing the spectrogram as the product of two non-negative matrices (a low-rank matrix) corresponds to representing the spectra observed at each frame as a non-negative combination of a finite number of basis spectra. Therefore, we can expect to obtain basis spectra unique to each speaker through the pretraining of $\mathbf{W}_j$. At test time, after fitting $\mathbf{WH}$ to the spectrogram of a test mixture signal $\mathbf{Y}$ with $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_J]$ fixed at the pretrained basis spectra, $\mathbf{V}_1, \ldots, \mathbf{V}_J$ can be estimated using the estimate of $\mathbf{H} = [\mathbf{H}_1^\mathsf{T}, \ldots, \mathbf{H}_J^\mathsf{T}]^\mathsf{T}$. The source signals can then be separated out using (4). A common way to train $\mathbf{W}_j$ is to solve

$$\{\hat{\mathbf{W}}_j, \hat{\mathbf{H}}_j\} = \operatorname*{argmin}_{\mathbf{W}_j, \mathbf{H}_j} \mathcal{D}(\tilde{\mathbf{S}}_j' | \mathbf{W}_j\mathbf{H}_j), \tag{5}$$

where $\tilde{\mathbf{S}}_j'$ is a concatenation of the power spectrograms of all training utterances of speaker j. $\mathcal{D}$ is a cost function that measures the dissimilarity of $\tilde{\mathbf{S}}_j'$ and $\mathbf{W}_j\mathbf{H}_j$, such as the IS divergence. At test time, given the power spectrogram $\tilde{\mathbf{Y}}$ of the mixture signal, we must solve

$$\hat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{H}} \mathcal{D}(\tilde{\mathbf{Y}} | \hat{\mathbf{W}}\mathbf{H}), \tag{6}$$

where $\hat{\mathbf{W}} = [\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_J]$ denotes the basis matrix containing the pretrained basis spectra. $\hat{\mathbf{W}}_j\hat{\mathbf{H}}_j$ corresponds to the estimate of the power spectrogram associated with speaker j. The complex spectrogram $\hat{\mathbf{S}}_j$ of speaker j can then be obtained as

$$\hat{\mathbf{S}}_j = \frac{\hat{\mathbf{W}}_j\hat{\mathbf{H}}_j}{\sum_{j'} \hat{\mathbf{W}}_{j'}\hat{\mathbf{H}}_{j'}} \odot \mathbf{Y}. \tag{7}$$

2.3 DNMF method

If we assume using the Wiener filter output (7) to obtain the signal of each speaker, the training and test objectives become inconsistent. Namely, the basis spectra are not necessarily trained in such a way that the separated signals at test time will be optimal. The DNMF method has been developed to address this inconsistency in the SNMF method, based on the idea of training the basis spectra in such a way that the separated signals become optimal at test time.

With the SNMF method, at test time, the basis matrix $\hat{\mathbf{W}}$ is used not only for estimating $\hat{\mathbf{H}}$ from $\tilde{\mathbf{Y}}$ in (6) but also for constructing the Wiener filter in (7). However, the basis matrices used in these steps do not necessarily have to be the same; rather, it would be more advantageous at test time to treat them as different variables and train them separately. We thus use $\mathbf{W}$ and $\mathbf{B}$ to denote the basis matrices at these steps, and discuss what criteria should be used to train them. By using the power spectrogram $\tilde{\mathbf{Y}}' = \{|y'(f,n)|^2\}_{f,n}$ of a random mixture of training utterances as the input and that of each of the utterances as the regression target, we can train $\mathbf{W}$ and $\mathbf{B}$ based on the process that exactly mimics the test scenario. After solving (6) by using the basis matrix $\hat{\mathbf{W}}$ obtained via (5), we can train $\mathbf{B}$ so that the output of (7) matches the regression target as closely as possible. Therefore, in the DNMF method, the training objective for $\mathbf{B}$ can be defined as

$$\hat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}} \sum_j \mathcal{D}\left(|\mathbf{S}_j'| \,\middle|\, \frac{\mathbf{B}_j\hat{\mathbf{H}}_j}{\mathbf{B}\hat{\mathbf{H}}} \odot |\mathbf{Y}'|\right), \tag{8}$$

where $|\mathbf{Y}'|$ and $|\mathbf{S}_1'|, \ldots, |\mathbf{S}_J'|$ denote the magnitude spectrograms of the mixture signal and the source signals, respectively. Note that here $|\cdot|$ is used to denote an operation of taking the elementwise absolute value of a matrix. At test time, the separated signals can be obtained by the same process as (6) and (7) using $\hat{\mathbf{W}}$ trained via (5) and $\hat{\mathbf{B}}$ trained via (8).

Both the SNMF and DNMF methods assume that the spectrogram of each speech spectrogram can be represented by a low-rank matrix. However, this assumption is not always accurate, and it limits the separation performance of both methods.

3. Proposed methods

3.1 CVAE Source Model

Since the matrix product representation $\mathbf{Wh}$ can be regarded as a single-layer linear fully-connected NN with $\mathbf{h}$ as the input, a deeper model with multiple nonlinear layers can be a more powerful alternative to the NMF model. One idea would be to express the variance $v_j(f,n)$ in the LGM (1) as the output of a DNN. As described below, this corresponds to a special case of a VAE. The MVAE method, mentioned earlier, is a multichannel source separation method that uses a CVAE, conditioned on a speaker code, as the source spectrogram model based on this idea. This model is called the CVAE source model. This paper proposes a single-channel speech separation method based on the CVAE source model.

A CVAE is a type of autoencoder consisting of an encoder and decoder. It is unique in that both the encoder and decoder are modeled in the form of parametric probability distributions, and both distributions are conditioned on auxiliary variables. Let $\mathbf{S}$ be the complex spectrogram of a particular speaker's utterance and $\mathbf{c}$ be the speaker code. Here, we assume that the speaker code $\mathbf{c}$ is represented a one-hot vector. Now, we condition the decoder distribution on $\mathbf{c}$ and further define it as a zero-mean complex Gaussian distribution so that it has the same form as the LGM (1):

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

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

where $\sigma_\theta^2(f,n;\mathbf{z},\mathbf{c})$ denotes the (f,n)th element of the decoder network output $\boldsymbol{\sigma}_\theta^2(\mathbf{z},\mathbf{c})$, $\mathbf{z}$ represents a latent variable generated from the encoder distribution, and η is a parameter corresponding to the scale (total energy) of $\mathbf{S}$. Next, we define the encoder distribution as a Gaussian distribution with diagonal covariance:

$$q_\phi(\mathbf{z}|\mathbf{S},\boldsymbol{c}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{S},\boldsymbol{c}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{S},\boldsymbol{c}))), \tag{11}$$

and define the prior distribution $p(\mathbf{z})$ as a standard Gaussian distribution. Here, the mean $\boldsymbol{\mu}_\phi(\mathbf{S},\mathbf{c})$ and variance $\boldsymbol{\sigma}_\phi^2(\mathbf{S},\mathbf{c})$ are assumed to be the encoder network outputs. All the unknown network parameters θ and ϕ are trained using a set of speaker-labeled training samples $\{\mathbf{S}_m, \mathbf{c}_m\}_{m=1}^M$. The goal is to train θ and ϕ so that the encoder distribution $q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ becomes consistent with the posterior $p_\theta(\mathbf{z}|\mathbf{S},\mathbf{c}) \propto p_\theta(\mathbf{S}|\mathbf{z},\mathbf{c})p(\mathbf{z})$. The decoder distribution with the resulting θ is expected to fit the true distribution of the spectrograms of each speaker reasonably well. If we define the training objective as the Kullback-Leibler (KL) divergence between $q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})$ and $p_\theta(\mathbf{z}|\mathbf{S},\mathbf{c})$, the training objective is equal up to a constant term to

$$\begin{aligned}\mathcal{J}(\phi,\theta) = \mathbb{E}_{(\mathbf{S},\mathbf{c})\sim p_D(\mathbf{S},\mathbf{c})}\big[&\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})||p(\mathbf{z})] \\ &- \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{S},\mathbf{c})}[\log p_\theta(\mathbf{S}|\mathbf{z},\mathbf{c})]\big],\end{aligned} \tag{12}$$

where $\mathbb{E}_{(\mathbf{S},\mathbf{c})\sim p_D(\mathbf{S},\mathbf{c})}[\cdot]$ denotes the sample mean over the training examples $\{\mathbf{S}_m, \mathbf{c}_m\}_{m=1}^M$, and $\mathrm{KL}[\cdot||\cdot]$ denotes the KL divergence. Thus, minimizing (12) amounts to distribution fitting. It should be noted that the second term in (12) is equal up to a constant term to the expectation of the IS divergence between $\hat{\mathbf{S}} = \{|s(f,n)|^2\}_{f,n}$ and $v(f,n)$, owing to the decoder distribution defined in the same form as the LGM.

In the CVAE source model, the latent variable $\mathbf{z}$ can be interpreted as context information corresponding to the linguistic content of $\mathbf{S}$, and the decoder NN parameter θ as the quantity that governs the mapping from the context information to the spectrogram. In this respect, $\mathbf{z}$ and θ can be regarded as corresponding to the coefficient (activation) matrix $\mathbf{H}$ and basis matrix $\mathbf{W}$ in the NMF model, respectively.

3.2 Proposed1: VASS Method

The VASS method corresponds to the SNMF method in which the NMF-type source model is replaced by the CVAE source model. Like the SNMF method, the VASS method consists of pretraining the source model (training step), fitting the source model to the spectrogram of an observed mixture signal (test step 1), and extracting source signals using the Wiener mask (test step 2). Thanks to the conditional modeling, the CVAE source model with a single set of parameters can be made to represent the spectrograms of all speakers in the training set by training the parameters using (12) as the objective. Let $\hat{\theta}$ be the parameters of the CVAE source model obtained after the training step. The first step at test time (test step 1) can be formulated as a maximum likelihood estimation problem

$$\{\hat{\mathbf{Z}}, \hat{\mathbf{C}}, \hat{\boldsymbol{\eta}}\} = \underset{\mathbf{Z},\mathbf{C},\boldsymbol{\eta}}{\operatorname{argmax}} \log p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta}), \tag{13}$$

where the likelihood function $p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ can be derived based on the assumption that the complex spectrogram $y(f, n)$ of a mixture signal follows

$$y(f,n) \sim \mathcal{N}_{\mathbb{C}}(y(f,n)|0, v(f,n)) \tag{14}$$

$$v(f,n) = \sum_j \underbrace{v_j(f,n)}_{\eta_j \sigma^2_{\hat{\theta}}(f,n;\mathbf{z}_j,\mathbf{c}_j)} . \tag{15}$$

As mentioned in subsection 2.1, $-\log p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ is equal up to a constant term to the IS divergence between $|y(f, n)|^2$ and $v(f, n)$. Hence, this problem is equivalent to finding $\mathbf{Z}$, $\mathbf{C}$, and $\boldsymbol{\eta}$ that minimize $\mathcal{D}_{\text{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ with $\hat{\theta}$ fixed, where $\tilde{\mathbf{Y}} = \{|y(f, n)|^2\}_{f,n}$. Once $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, and $\hat{\boldsymbol{\eta}}$ are estimated, the signal of each speaker can be obtained by the Wiener filter (test step 2)

$$\mathbf{S}_j = \frac{\eta_j \boldsymbol{\sigma}^2_{\hat{\theta}}(\hat{\mathbf{z}}_j, \hat{\mathbf{c}}_j)}{\sum_{j'} \eta_{j'} \boldsymbol{\sigma}^2_{\hat{\theta}}(\hat{\mathbf{z}}_{j'}, \hat{\mathbf{c}}_{j'})} \odot \mathbf{Y}. \tag{16}$$

Note that there are several possible ways to solve (13). The first is to simply optimize $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ using the gradient method (back propagation for $\mathbf{Z}$ and $\mathbf{C}$) with $\log p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ or $\mathcal{D}_{\text{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ as the criterion. The second method is to optimize them using the Expectation-Maximization (EM) algorithm, treating the complex spectrogram $s_j(f, n)$ of each speaker as the latent variable. We can keep increasing the log-likelihood $\log p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ by iteratively increasing an auxiliary function defined as $\mathbb{E}_{\mathbf{S}\sim p(\mathbf{S}|\mathbf{Y},\mathbf{Z}',\mathbf{C}',\boldsymbol{\eta}')}[\log p(\mathbf{S}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})]$ through iterative updates called the E- and M-steps. The M-step is a process of updating $\mathbf{Z}$, $\mathbf{C}$, $\boldsymbol{\eta}$ so that the auxiliary function increases. $\mathbf{Z}$ and $\mathbf{C}$ can be updated by backpropagation. When $\mathbf{Z}$ and $\mathbf{C}$ are fixed, η that maximizes the auxiliary function can be derived analytically. The E-Step is a process of recomputing the auxiliary function each time $\mathbf{Z}$, $\mathbf{C}$, and $\boldsymbol{\eta}$ are updated, by substituting the updated $\mathbf{Z}$, $\mathbf{C}$, and $\boldsymbol{\eta}$ into $\mathbf{Z}'$, $\mathbf{C}'$, and $\boldsymbol{\eta}'$. Since $\log p(\mathbf{S}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ is split in J individual terms, namely $\sum_j \sum_{f,n} \log p(s_j(f, n)|0, v_j(f, n))$, $(\mathbf{z}_1, \mathbf{c}_1, \boldsymbol{\eta}_1), \ldots, (\mathbf{z}_J, \mathbf{c}_J, \boldsymbol{\eta}_J)$ can be updated in parallel at the M-step. The third method is to increase $\log p(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\eta})$ iteratively by using another form of an auxiliary function as in [12]. Owing to space limitations, the details and derivations of the algorithms for these three methods are omitted. In the experiments described below, we used the method based on the EM algorithm.

3.3 Proposed2: DVASS Method

In the VASS method, as in the SNMF method, the training objective for the parameter θ of the CVAE source model does not make the separated signals (Wiener filter outputs) optimal at test time. To address this mismatch between the training and test objectives, we further propose improving the VASS method by following the idea of the DNMF method. Recall that the idea of the DNMF method was to treat the basis matrix responsible for obtaining the coefficient matrix and that responsible for constructing the Wiener filter as separate variables. In the same manner, we treat the CVAE source model parameters responsible for obtaining (13) and those responsible for constructing the Wiener filter as separate variables, and denote them by θ and ϑ, respectively. As in the DNMF method, we can train these parameters by following the process that exactly mimics the speech separation process at test time. Let $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, and $\hat{\boldsymbol{\eta}}$ represent the values obtained by (13), with $\hat{\theta}$ fixed at the value obtained by (12). By using $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, and $\hat{\boldsymbol{\eta}}$, we can train ϑ so that the output of (16) matches the target signal as closely as possible. The training objective can be defined as

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmin}} \sum_j \mathcal{D}\left(|\mathbf{S}'_j| \,\middle|\, \frac{\eta_j \boldsymbol{\sigma}^2_{\vartheta}(\hat{\mathbf{z}}_j, \hat{\mathbf{c}}_j)}{\sum_{j'} \eta_{j'} \boldsymbol{\sigma}^2_{\vartheta}(\hat{\mathbf{z}}_{j'}, \hat{\mathbf{c}}_{j'})} \odot |\mathbf{Y}'|\right). \tag{17}$$

At test time, the separated signals can be obtained by performing (13) and (16) using the trained θ and ϑ. An overview of the DVASS method is shown in Fig. 1 where the criterion for (17) is denoted as $\mathcal{J}_{re}(\vartheta)$.

4. Experimental evaluations

The proposed method was evaluated on a single-channel speech separation task of separating out two speakers. We chose the SNMF and DC [1] methods as baseline methods for comparison. As the experimental data, we used speech samples of the CMU ARCTIC database [22]. We used a set of the utterances of two female ('clb' and 'slt') and two male ('bdl' and 'rms') speakers. For each speaker, we used 1000 utterances for training, and 132 utterances for testing. We generated 81 speech mixtures for three speaker combinations: bdl+clb, bdl+rms, and clb+slt. Each test mixture signal was generated so that the energy of each speaker is equal. 560 mixture signals $\mathbf{Y}'$ used for training ϑ were generated in the same manner. All the speech signals were resampled at 8 [kHz] and STFT analysis was conducted with 512 [ms] frame length and 256 [ms] hop length. In the VASS method, we used a three-layer fully-convolutional network with gated linear units and a three-layer fully-deconvolutional network with gated linear units as the encoder and decoder networks in the CVAE model, as in [8]. In the DVASS method, we used the same network architectures for the encoder and decoder. We used Adam [23] for NN training and updating the model parameters $\mathbf{z}$ and $\mathbf{c}$. In the VASS and DVASS methods, the initial separated signals were obtained using the SNMF method that was run for 100 iterations, and $\mathbf{z}$ was initialized by feeding the initially separated signals into the encoder. For each paired training sample $(\mathbf{S}, \mathbf{c})$, we fixed $\mathbf{c}$ at an one-hot vector corresponding to the speaker of $\mathbf{S}$. The VASS and DVASS methods were run for two iterations. In the SNMF method, the number of bases was set to 10 for each source, and the KL divergence criterion was used as $\mathcal{D}$. As the evaluation metrics,

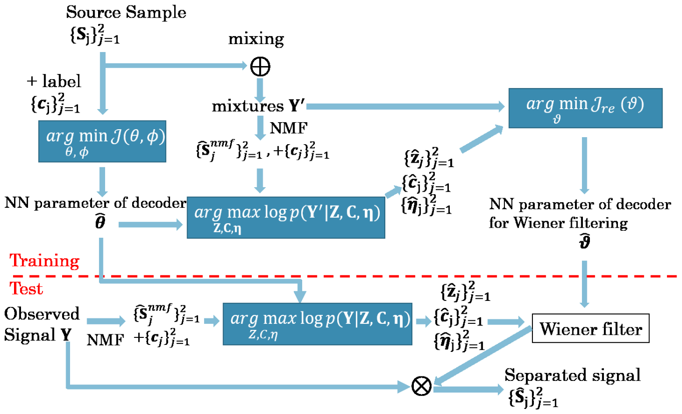
Source Sample
$\{\mathbf{S}_j\}_{j=1}^2$
mixing
+ label
$\{c_j\}_{j=1}^2$
mixtures $\mathbf{Y}'$
$\underset{\theta,\phi}{arg\ \min}\ \mathcal{J}(\theta,\phi)$
NMF
$\{\hat{\mathbf{S}}_j^{nmf}\}_{j=1}^2, +\{c_j\}_{j=1}^2$
$\underset{\vartheta}{arg\ \min}\ \mathcal{J}_{re}(\vartheta)$
$\{\hat{\mathbf{z}}_j\}_{j=1}^2$
$\{\hat{\mathbf{c}}_j\}_{j=1}^2$
$\{\hat{\boldsymbol{\eta}}_j\}_{j=1}^2$
NN parameter of decoder
$\hat{\boldsymbol{\theta}}$
$\underset{\mathrm{Z,C,\eta}}{arg\ max}\log p(\mathbf{Y}'|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$
NN parameter of decoder for Wiener filtering
$\hat{\boldsymbol{\vartheta}}$
Training
Test
Observed Signal $\mathbf{Y}$
NMF
$\{\hat{\mathbf{S}}_j^{nmf}\}_{j=1}^2$
$+\{c_j\}_{j=1}^2$
$\underset{Z,C,\eta}{arg\ max}\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$
$\{\hat{\mathbf{z}}_j\}_{j=1}^2$
$\{\hat{\mathbf{c}}_j\}_{j=1}^2$
$\{\hat{\boldsymbol{\eta}}_j\}_{j=1}^2$
Wiener filter
Separated signal
$\{\hat{\mathbf{S}}_j\}_{j=1}^2$

Figure 1: Schematic overview of DVASS.

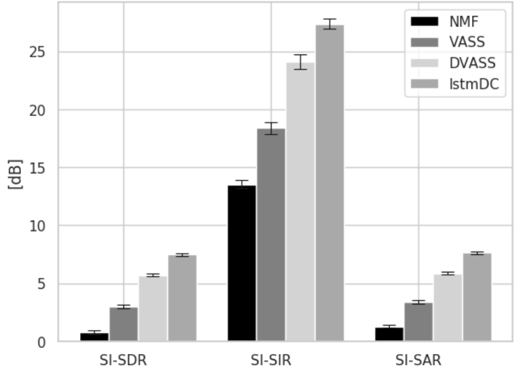

Figure 2: Averaged separation performances [dB].

we used the scale-invariant signal-to-distortion ratio (SDR), the scale-invariant signal-to-inference ratio (SIR), and the scale-invariant signal-to-artifact ratio (SAR) [24] between the reference and separated signals.

The experimental results are shown in Fig.2. Compared with the baseline SNMF method, the high separation performance of the proposed VASS method was confirmed. The performance difference between the SNMF and VASS methods may reflect the difference in the ability of each source model to achieve separation. The DVASS method showed higher separation performance than the VASS method in all metrics. This confirms the effectiveness of discriminative training. However, we also confirmed that the DVASS method still had room for improvement up to the high separation performance of the DC method.

5. Conclusion

In this paper, we proposed the VASS method as a single-channel speech separation method using the CVAE source model, and also proposed the DVASS method which trains the CVAE source model based on a discriminative criterion. The effectiveness of the proposed method was investigated through specific two-speaker separation experiments. The experimental evaluation showed that both the VASS and DVASS methods performed better than the SNMF method, and the DVASS method performed better than the VASS method.

References

[1] J. R. Hershey, *et al.*,"Deep clustering: Discriminative embeddings for segmentation and separation" *ICASSP*, 31–35, 2016.

[2] Y. Liu, *et al.*,"Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation" *IEEE/ACM Trans. ASLP*, 27(12), 2092–2102, 2019.

[3] J. Le Roux, *et al.*, "Phasebook and friends: Leveraging discrete representations for source separation" *IEEE JSTSP*, 13(2), 370–382, 2019.

[4] D. Wang, *et al.*, "Supervised speech separation based on deep learning:An overview" *IEEE/ACM Trans. ASLP*, 26(10), 1702–1726, 2018.

[5] D. D. Lee, *et al.*, "Algorithms for non-negative matrix factorization" *NIPS*, 556–562, 2001.

[6] P. Smaragdis, *et al.*, "Supervised and semi-supervised separation of sounds from single-channel mixtures" *ICA*, 414–421, 2007.

[7] F. Weninger, *et al.*, "Discriminative NMF and its application to single-channel source separation" *Interspeech*, 865–869, 2014.

[8] H. Kameoka, *et al.*, "Supervised determined source separation with multichannel variational autoencoder" *Neural Computation*, 31(9), 1891–1914, 2019.

[9] A. A. Nugraha, *et al.*, "Multichannel audio source separation with deep neural networks" *IEEE/ACM Trans. ASLP*, 24(9), 1652–1664, 2016.

[10] N. Makishima, *et al.*, "Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation" *IEEE/ACM Trans. ASLP*, 27(10), 1601–1615, 2019.

[11] L. Li, *et al.*, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier" *IEEE Access*, 8(1), 228740–228753, 2020.

[12] S. Seki, *et al.*, "Generalized Multichannel Variational Autoencoder for Underdetermined Source Separation" *IEEE Access*, 7(1), 168104–168115, 2019.

[13] Y. Bando, *et al.*, "statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization" *ICASSP*, 716–720, 2018.

[14] S. Leglaive, *et al.*, "A variance modeling framework based on variational autoencoders for speech enhancement" *MLSP*, 2018.

[15] K. Sekiguchi, *et al.*, "Bayesian multichannelspeech enhancement with a deep speech prior" *APSIPA*, 1233–1239, 2018.

[16] S. Leglaive, *et al.*, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization" *ICASSP*, 101–105, 2019.

[17] D. P. Kingma, *et al.*, "Semi-supervised learning with deep generative models" *NIPS*, 2014.

[18] H. Kameoka, *et al.*, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation" *LVA/ICA*, 245–253, 2010.

[19] D. Kitamura, *et al.*, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization" *IEEE/ACM Trans. ASLP*, 24(9), 1626–1641, 2016.

[20] C. Fevotte, *et al.*, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," *WASPAA*, pp. 78–81, 2005.

[21] E. Vincent, *et al.*, "Underdetermined instantaneous audio source separation via local Gaussian modeling," *ICA*, pp. 775–782, 2009.

[22] J. Kominek., *et al.*, "The CMU Arctic speech databases" *WSS*, 2004.

[23] D. P. Kingma, *et al.*, "Adam: A method for stochastic optimization" *ICLR*, 2015.

[24] J. Le Roux, *et al.*, "SDR – Half-baked or Well Done?" *ICASSP*, 626–630, 2019.