Journal of Signal Processing, Vol.25, No.4, pp.145-149, July 2021

SELECTED PAPER AT NCSP'21

Single-Channel Multispeaker Separation with Variational Autoencoder Spectrogram Model

Naoya Murashima[1], Hirokazu Kameoka[2], Li Li[1], Shogo Seki[2] and Shoji Makino[1,3]

[1]Graduate School Science and Technology, University of Tsukuba, Ibaraki 305-8573, Japan
E-mail: naodrrb@gmail.com, {lili@mmlab.cs, maki@tara}.tsukuba.ac.jp
[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa 243-0124, Japan
E-mail: hirokazu.kameoka.uh@hco.ntt.co.jp
[3]Graduate School of Information, Production and Systems, Waseda University, Fukuoka 808-0135, Japan
E-mail: s.makino@waseda.jp

1. Abstract

This paper deals with single-channel speaker-dependent speech separation. While discriminative approaches using deep neural networks (DNNs) have recently proved powerful, generative approaches, including methods based on non-negative matrix factorization (NMF), are still attractive because of their flexibility in handling the mismatch between training and test conditions. Although NMF-based methods work reasonably well for particular sound sources, one limitation is that they can fail to work for sources with spectrograms that do not comply with the NMF model. To address this problem, attempts have recently been made to replace the NMF model with DNNs. With a similar motivation to these attempts, we propose in this paper a variational autoencoder (VAE)-based monaural source separation (VASS) method using a conditional VAE (CVAE) for source spectrogram modeling. We further propose an extension of the VASS method, called the discriminative VASS (DVASS) method, which uses a discriminative criterion for model training so that the separated signals directly become optimal. Experimental results revealed that the VASS method performed better than an NMF-based method, and the DVASS method performed better than the VASS method.

2. Introduction

Speech separation is a technique of separating 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 a 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 Time-Frequency (TF) masks or TF embeddings from a given mixture signal based on spectro-temporal features. Recently, methods of training a DNN that directly predicts the waveform of each speaker have 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.

Generative approaches, including non-negative matrix factorization (NMF) [5], are attractive in the 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 the test time. The idea of NMF is to approximate the spectrum of a mixture signal observed in each short-term frame as a linear sum of a limited number of basis spectra scaled by time-varying amplitudes. In supervised NMF (SNMF) [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 the test time. In other words, the basis spectra are not trained so that the separated signals at the test time become optimal. Discriminative NMF (DNMF) [7] was later proposed to resolve 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 Wiener filters) directly become optimal. Although these NMF-based methods work reasonably well for particular 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 by the NMF model, generative approach-based methods using DNNs have been proposed [8–16]. For multichannel source separation under 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 the 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) [13, 14] and its multichannel 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 SNMF to DNMF.

3. Conventional Methods

3.1 Problem formulation

We consider a situation where a mixture of signals of J speakers is observed. Let $\mathbf{Y} = \{y(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$ and $\mathbf{S}_j = \{s_j(f,n)\}_{f,n} \in \mathbb{C}^{F\times N}$ 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 $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$.

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

Equation (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 $\odot$ respectively denote elementwise division and multiplication. Note that the multiplicative factor of Eq. (4) is called the Wiener mask. Equation (4) implies that once $\mathbf{V}_1, \ldots, \mathbf{V}_J$ are 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 Eq. (3) as the objective function, under some constraint or assumption imposed on $\mathbf{V}_1, \ldots, \mathbf{V}_J$.

3.2 SNMF

SNMF 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 in 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 the 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 Eq. (4). A common way of training $\mathbf{W}_j$ is to solve

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

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 Itakura-Saito (IS) divergence. At the test time, given the power spectrogram $\tilde{\mathbf{Y}}$ of the mixture signal, we must solve

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

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

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

3.3 DNMF

If we assume the use of the Wiener filter output, Eq. (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. DNMF has been developed to address this inconsistency in SNMF, based on the idea of training the basis spectra in such a way that the separated signals become optimal at the test time.

With SNMF, at the test time, the basis matrix $\hat{\mathbf{W}}$ is used not only for estimating $\hat{\mathbf{H}}$ from $\tilde{\mathbf{Y}}$ in Eq. (6) but also for constructing the Wiener filter in Eq. (7). However, the basis matrices used in these steps do not have to be the same; rather, it would be more advantageous at the 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 $\mathbf{B}$ using the process that exactly mimics the test scenario. After solving Eq. (6) by using the basis matrix $\hat{\mathbf{W}}$ obtained via Eq. (5), we can train $\mathbf{B}$ so that the output of Eq. (7) matches the regression target as closely as possible. Therefore, in DNMF, the training objective for $\mathbf{B}$ can be defined as

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

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 the operation of taking the elementwise absolute value of a matrix. At the test time, the separated signals can be obtained by the same process as Eq. (6) and refeqWiener3 using $\hat{\mathbf{W}}$ trained via Eq. (5) and $\hat{\mathbf{B}}$ trained via Eq. (8).

Both SNMF and DNMF assume that each speech spectrogram can be represented by a low-rank matrix. However,

this assumption is not always accurate and limits the separation performance of both methods.

4. Proposed Methods

4.1 CVAE source model

Since the matrix product representation $\mathbf{Wh}$ can be regarded as a single-layer linear fully connected neural network (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, Eq. (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. In this paper, a single-channel speech separation method based on the CVAE source model is proposed.

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 as 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 (Eq. (1)):

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

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

where $\sigma^2_\theta(f,n;\mathbf{z},\mathbf{c})$ denotes the (f,n)th element of the decoder network output $\boldsymbol{\sigma}^2_\theta(\mathbf{z},\mathbf{c})$, $\mathbf{z}$ 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:

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

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 $\boldsymbol{\sigma}^2_\phi(\mathbf{S},\mathbf{c})$ are assumed to be the encoder network outputs. Both 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 Eq. (12) amounts to distribution fitting. Note that the second term in Eq. (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.

4.2 Proposed method 1: VASS method

The VASS method corresponds to SNMF in which the NMF-type source model is replaced by the CVAE source model. Similarly to SNMF, 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). Owing 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 Eq. (12) as the objective. Let $\hat{\theta}$ be the parameters of the CVAE source model obtained after the training step. The first step at the test time (test step 1) can be formulated as a maximum likelihood estimation problem.

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

where the likelihood function $p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ can be derived on the basis of 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 sect. 3.1, $-\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ 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 $\mathcal{D}_{\mathrm{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ 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 using 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 of solving Eq. (13). The first is to simply optimize $\mathbf{Z}$, $\mathbf{C}$, and $\boldsymbol{\eta}$ using the gradient method (backpropagation for $\mathbf{Z}$ and $\mathbf{C}$) with $\log p(\mathbf{Y}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ or $\mathcal{D}_{\mathrm{IS}}(\tilde{\mathbf{Y}}|\mathbf{V})$ 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 the process

of updating $\mathbf{Z}$, $\mathbf{C}$, and $\boldsymbol{\eta}$ to increase the auxiliary function. $\mathbf{Z}$ 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 the 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 $\boldsymbol{\eta}'$ respectively. Since $\log p(\mathbf{S}|\mathbf{Z},\mathbf{C},\boldsymbol{\eta})$ is split into 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 an auxiliary function with another form 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.

4.3 Proposed method 2: DVASS method

In the VASS method, as in SNMF, the training objective for the parameter θ of the CVAE source model does not make the separated signals (Wiener filter outputs) optimal at the test time. To address this mismatch between the training and test objectives, we further propose improving the VASS method by following the idea of DNMF. Recall that the idea of DNMF is 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 Eq. (13) and those responsible for constructing the Wiener filter as separate variables, and denote them by θ and ϑ, respectively. As in DNMF, we can train these parameters by following the process that exactly mimics the speech separation process at the test time. Let $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, and $\hat{\boldsymbol{\eta}}$ represent the values obtained using Eq. (13), with $\hat{\theta}$ fixed at the value obtained using Eq. (12). By using $\hat{\mathbf{Z}}$, $\hat{\mathbf{C}}$, and $\hat{\boldsymbol{\eta}}$, we can train ϑ to match the output of Eq. (16) to 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 the test time, the separated signals can be obtained by evaluating Eq. (13) and Eq. (16) using the trained θ and ϑ. An overview of the DVASS method is shown in Fig. 1, where the criterion for Eq. (17) is denoted as $\mathcal{J}_{re}(\vartheta)$.

5. Experimental Evaluations

The proposed method was evaluated on a single-channel speech separation task of separating 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 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. We generated 560 mixture signals $\mathbf{Y}'$ used for training ϑ in the same manner. All the speech signals were resampled at 8 [kHz] and STFT analysis was conducted with a frame length of 512 [ms] and a hop length of 256 [ms] . 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, respectively, 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 initially separated signals were obtained by SNMF 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 a one-hot vector corresponding to the speaker of $\mathbf{S}$. The VASS and DVASS methods were run for two iterations. In SNMF, 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, we used the scale-invariant signal-to-distortion ratio (SDR), scale-invariant signal-to-inference ratio (SIR), and scale-invariant signal-to-artifact ratio (SAR) [24] between the reference and separated signals.

The experimental results are shown in Figure 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 a 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.

6. 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 SNMF, and the DVASS method performed better than the VASS method.

References

[1] J. R. Hershey, Z.Chen, J. Le Roux and S. Watanabe: Deep clustering: Discriminative embeddings for segmentation and separation, 2016 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 31–35, 2016.

[2] Y. Liu and D. Wang: Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation, IEEE/ACM Trans. Audio Speech Lang. Process, Vol. 27, No. 12, pp. 2092–2102, 2019.

[3] J. Le Roux, G. Wichern, S. Watanabe, A.Sarroff and J. R. Hershey: Phasebook and friends: Leveraging discrete representations for source separation, IEEE J.

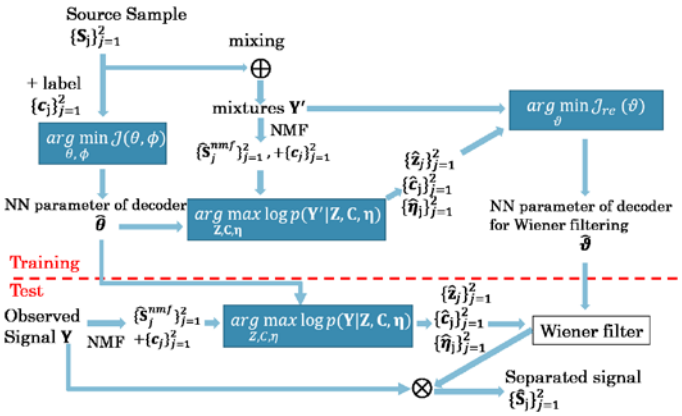

Figure 1: Schematic overview of DVASS

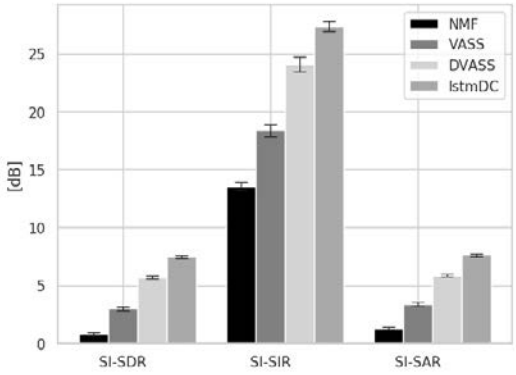

Figure 2: Average separation performance [dB]

Sel. Top. Signal Process., Vol. 13, No. 2, pp. 370–382, 2019.

[4] D. Wang and J. Chen: Supervised speech separation based on deep learning: An overview, IEEE/ACM Trans. Audio Speech Lang. Process., Vol. 26, No. 10, pp. 1702–1726, 2018.

[5] D. D. Lee and H. S. Seung: Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst., pp. 556–562, 2001.

[6] P. Smaragdis, B. Raj and M. Shashanka: Supervised and semi-supervised separation of sounds from single-channel mixtures, 2007 Int. Conf. Independent Compon. Anal. Signal Sep., pp. 414–421, 2007.

[7] F. Weninger, J. Le Roux, J. R. Hershey and S. Watanabe: Discriminative NMF and its application to single-channel source separation, 2014 Annu. Conf. Int. Speech Commun. Assoc., pp. 865–869, 2014.

[8] H. Kameoka, L. Li, S. Inoue and S. Makino: Supervised determined source separation with multichannel variational autoencoder, Neural Comput., Vol. 31, No. 9, pp. 1891–1914, 2019.

[9] A. A. Nugraha, A. Liutkus and E. Vincent: Multichannel audio source separation with deep neural networks, IEEE/ACM Trans. Audio Speech Lang. Process., Vol. 24, No. 9, pp. 1652–1664, 2016.

[10] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari and N. Ono: Independent deeply learned matrix analysis for determined audio source separation, IEEE/ACM Trans. Audio Speech Lang. Process, Vol. 27, No. 10, pp. 1601–1615, 2019.

[11] L. Li, H. Kameoka and S. Makino: FastMVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier, IEEE Access, Vol. 8, No. 1, pp. 228740–228753, 2020.

[12] S. Seki, H. Kameoka, L. Li, T. Toda and K. Takeda: Generalized multichannel variational autoencoder for underdetermined source separation, IEEE Access, Vol. 7, No. 1, pp. 168104–168115, 2019.

[13] 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, 2018 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 716–720, 2018.

[14] S. Leglaive, L. Girin and R. Horaud: A variance modeling framework based on variational autoencoders for speech enhancement, 2018 Int. Workshop Mach. Learn. Signal Process., 2018.

[15] K. Sekiguchi, Y. Bando, K. Yoshii and T. Kawahara: Bayesian multichannel speech enhancement with a deep speech prior, 2018 Asia Pac. Signal Inf. Process. Assoc. Annu. Summit Conf., pp. 1233–1239, 2018.

[16] S. Leglaive, L. Girin and R. Horaud: Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization, 2019 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 101–105, 2019.

[17] D. P. Kingma, D. J. Rezende, S. Mohamed and M. Welling: Semi-supervised learning with deep generative models, Adv. Neural Inf. Process. Syst., 2014.

[18] 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, 2010 Int. Conf. Latent Var. Anal. Signal Sep., pp. 245–253, 2010.

[19] D. Kitamura, N. Ono, H. Sawada, H. Kameoka and H. Saruwatari: Determined blind source separation unifying independent vector analysis and non-negative matrix factorization, IEEE/ACM Trans. Audio Speech Lang. Process., Vol. 24, No. 9, pp. 1626–1641, 2016.

[20] C. Fevotte and J. F. Cardoso: Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models, 2005 IEEE Workshop Appl. Signal Process. Audio Acoust., pp. 78-–81, 2005.

[21] E. Vincent, S. Arberet and R. Gribonval: Underdetermined instantaneous audio source separation via local Gaussian modeling, 2009 Int. Conf. Independent Compon. Anal. Signal Sep., pp. 775-–782, 2009.

[22] J. Kominek and A. W. Black: The CMU arctic speech databases, 2004 ISCA Speech Synth. Workshop, pp. 223–224, 2004.

[23] D. P. Kingma and J. Ba: Adam: A method for stochastic optimization, 2015 Int. Conf. Learn. Represent., 2015.

[24] J. Le Roux, S. Wisdom, H. Erdogan and J. R. Hershey: SDR – Half-baked or well done?, 2019 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 626–630, 2019.