Semi-supervised Multichannel Speech Separation Based on a Phone- and Speaker-Aware Deep Generative Model of Speech Spectrograms

Yicheng Du*, Kouhei Sekiguchi†*, Yoshiaki Bando‡, Aditya Arie Nugraha†,
Mathieu Fontaine†, Kazuyoshi Yoshii*†, Tatsuya Kawahara*

*Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
Email: du@sap.ist.i.kyoto-u.ac.jp, yoshii@kuis.kyoto-u.ac.jp
†Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan
‡National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

***Abstract*—This paper describes a semi-supervised multichannel speech separation method that uses clean speech signals with frame-wise phonetic labels and sample-level speaker labels for pre-training. A standard approach to statistical source separation is to formulate a probabilistic model of multichannel mixture spectrograms that combines source models representing the time-frequency characteristics of sources with spatial models representing the covariance structure between channels. For speech separation and enhancement, deep generative models with latent variables have successfully been used as source models. The parameters of such a speech model can be trained beforehand from clean speech signals with a variational autoencoder (VAE) or its conditional variant (CVAE) that takes speaker labels as auxiliary inputs. Because human speech is characterized by both phonetic features and speaker identities, we propose a probabilistic model that combines a phone- and speaker-aware deep speech model with a full-rank spatial model. Our speech model is trained with a CVAE taking both phone and speaker labels as conditions. Given speech mixtures, the spatial covariance matrices, latent variables of sources, and phone and speaker labels of sources are jointly estimated. Comparative experimental results showed that the performance of speech separation can be improved by explicitly considering phonetic features and/or speaker identities.**

***Index Terms*—multichannel source separation, speech separation, variational autoencoder**

I. Introduction

Multichannel source separation aims to reconstruct source signals from observed mixture signals obtained by a microphone array. It is a fundamental technique for automatic speech recognition (ASR) [1] since most ASR systems require the speech signals to be separated from mixture signals that contain multiple speakers and noise.

One approach to multichannel source separation is to use a unified probabilistic model based on source models representing the power spectral densities (PSDs) of sources and a spatial model representing the sound propagation process. Nonnegative matrix factorization (NMF) [2] has often been used as a source model. NMF approximates the PSDs of each source spectrogram as the product of two low-rank matrices corresponding to a set of basis spectra and a set of their activations, respectively. Multichannel NMF (MNMF) [3]–[5] was proposed by integrating an NMF-based source model with a full-rank spatial model [6]. A determined version of MNMF called independent low-rank matrix analysis (ILRMA) [7] was then derived by restricting the mixing system to a determined rank-1 spatial model. The demixing system can be estimated with a stable and fast update rule called iterative projection [8]. A drawback of the NMF-based source model, however, is that its low-rank assumption is incompatible with speech spectrograms.

One promising approach to avoid the low-rank assumption is to use the decoder of a variational autoencoder (VAE) [9] trained on clean speech signals as a deep speech model for speech enhancement [10]–[13]. In a semi-supervised speech separation method called multichannel VAE (MVAE) [14]–[16], a rank-1 or full-rank spatial model is integrated with a *speaker-aware* deep speech model trained with a conditional VAE (CVAE) [17] that takes speaker labels as auxiliary inputs and learns speaker-independent latent features. In speech separation, the speaker labels and the latent features are estimated from the current estimate of speech signals, and the deep speech model with the estimated speaker labels and latent features are then used for separating mixture signals into speaker-coherent speech signals. These two mutually-dependent steps are iterated.

Because the selective listening ability of humans is considered to make effective use of not only speaker identity features but also phonetic features, we propose a semi-supervised speech separation method that integrates a *phone- and speaker-aware* deep speech model with a full-rank spatial model. In fact, phonetic features have been proven to improve the performances of speech separation and ASR [18]–[20]. Our deep speech model is trained with a CVAE that takes both a sample-level speaker label and frame-wise phonetic labels as auxiliary inputs. In addition, phone and speaker classifiers are trained in a supervised manner by using annotated speech signals. Given speech mixtures, the full-rank spatial covariance matrices (SCMs) and phone and speaker labels of sources are jointly estimated by using the trained speech model and classifiers.

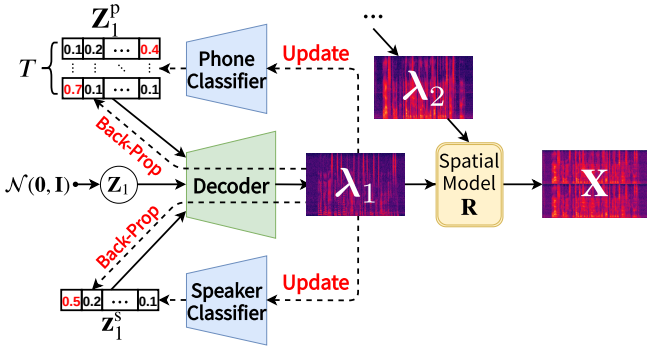


Fig. 1: An overview of the proposed method. The phone and speaker labels $\mathbf{Z}_n^{\mathrm{p}}$ and $\mathbf{z}_n^{\mathrm{s}}$ of each source n are updated by backpropagation or classifiers.

The major contribution of this paper is to show that both phonetic features and speaker identities are important clues for computational speech separation.

II. Related Work

This section reviews speaker-aware and phone-aware speech separation and speech enhancement methods, especially the ones that utilize speech generative models based on a deep neural network (DNN).

A. Deep Speech Models

Deep generative models have been used as speech models in semi-supervised speech enhancement [10]–[13] and speech separation [14]–[16]. Those models are typically trained in the VAE framework [9]. A VAE consists of a DNN-based encoder that estimates the distribution of latent variables given the observed variables, and a DNN-based decoder that estimates the distribution of observed variables given the latent variables. As a deep speech model, the decoder usually represents the speech PSDs. In a CVAE [17], both encoder and decoder are conditioned by auxiliary variables.

B. Speaker-Aware Speech Separation

The deep speech model in the MVAE [14], [15] is trained using a CVAE on clean speech signals with speaker labels. The speaker label of each sample is given to both the encoder and decoder. In the separation phase, the speaker labels are estimated and kept coherent over time. Although the MVAE achieves good separation performance, the speaker labels has limited effect because both encoder and decoder tend to ignore the labels by estimating latent features that can reconstruct speech spectrograms without utilizing the speaker labels [16].

C. Phone-Aware Speech Separation and Enhancement

Phonetic features has scarcely been dealt with for speech separation and speech enhancement. Wang *et al.* [19] unified an HMM-DNN-based ASR system and phone-specific DNN models for speech enhancement. In the test phase, the ASR system provides the phone label of each frame and the phone-specific models are then used to perform speech enhancement. Takahashi *et al.* [20] proposed a transfer learning approach that incorporates phonetic and linguistic information. In the test phase, a DNN-based separation model iteratively takes as inputs features extracted using an end-to-end ASR model.

III. Proposed Method

This section describes the proposed multichannel speech separation method that integrates a phone- and speaker-aware deep speech model with a full-rank spatial model. Fig. 1 shows an overview of the proposed method.

A. Problem Specification

Suppose that there are N sources (speakers) and M microphones. Let $\mathbf{s}_{ft} = [s_{1ft}, \ldots, s_{Nft}]^{\mathsf{T}} \in \mathbb{C}^N$ and $\mathbf{c}_{nft} = [c_{nft1}, \ldots, c_{nftM}]^{\mathsf{T}} \in \mathbb{C}^M$ be the short-time Fourier transform (STFT) coefficients of the sources and those of the image of source n, respectively, at a time-frequency (TF) bin of frequency f and time t. The mixture $\mathbf{x}_{ft}$ is given by

$$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{c}_{nft}. \tag{1}$$

Given mixtures $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T}$ as observed data, our goal is to estimate source images $\mathbf{C} = \{\mathbf{c}_{nft}\}_{n=1,f=1,t=1}^{N,F,T}$, where F and T are the number of frequency bins and that of time frames, respectively.

B. Source Modeling

We formulate a phone- and speaker-aware deep speech model that represents the generative process of a complex speech spectrogram under a condition that frame-wise phonetic labels and a sample-level speaker label are given. The TF bins of each source are assumed to follow circularly-symmetric complex Gaussian distributions as follows:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}(0, \lambda_{nft}), \tag{2}$$

where $\boldsymbol{\lambda}_n = \{\lambda_{nft}\}_{f=1,t=1}^{F,T}$ represents the PSDs of source n determined by a DNN with parameter $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{\lambda}_n = g_n \cdot \mathrm{DNN}_{\boldsymbol{\theta}}\big(\mathbf{Z}_n, \mathbf{Z}_n^{\mathrm{p}}, \mathbf{z}_n^{\mathrm{s}}\big), \tag{3}$$

where $g_n \in \mathbb{R}_+$ represents the overall gain of source n, $\mathbf{Z}_n = \{\mathbf{z}_{nt}\}_{t=1}^{T} \in \mathbb{R}^{D\times T}$ is a set of frame-wise latent variables whose prior distribution is the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{Z}_n^{\mathrm{p}} = \{\mathbf{z}_{nt}^{\mathrm{p}}\}_{t=1}^{T} \in \mathbb{R}^{P\times T}$ is a sequence of one-hot vectors representing frame-wise phonetic labels (P is the number of kinds of phones), and $\mathbf{z}_n^{\mathrm{s}} \in \mathbb{R}^S$ is a one-hot vector indicating a speaker label (S is the number of known speaker identities). $\mathbf{Z}_n$ is supposed to represent the acoustic characteristics other than phonetic features and speaker identities such as fundamental frequencies.

C. Spatial Modeling

To represent relatively long reverberation, we use a full-rank spatial model [6] as follows:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\Big(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{R}_{nf}\Big), \tag{4}$$

where $\mathbf{R}_{nf} \in \mathbb{S}_+^M$ is a full-rank SCM and $\mathbb{S}_+^M$ denotes the set of complex positive semidefinite matrices of size M. The generative model of $\mathbf{X}$, *i.e.*, the log-likelihood function of $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_n\}_{n=1}^N$ and $\mathbf{R} = \{\mathbf{R}_{nf}\}_{n=1,f=1}^{N,F}$, is given by

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\lambda}, \mathbf{R}) &= \sum_{f=1}^{F}\sum_{t=1}^{T} \log \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{ft}|\mathbf{0}, \hat{\mathbf{X}}_{ft}) \\ &= -\sum_{f=1}^{F}\sum_{t=1}^{T} \big(\mathrm{Tr}(\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}) + \log|\hat{\mathbf{X}}_{ft}|\big) + \mathrm{const}, \quad (5)\end{aligned}$$

where $\mathbf{X}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}} \in \mathbb{S}_+^M$ and $\hat{\mathbf{X}}_{ft} = \sum_{n=1}^N \lambda_{nft}\mathbf{R}_{nf} \in \mathbb{S}_+^M$ are observed and reconstructed matrices, and $\cdot^{\mathsf{H}}$ is the Hermitian transposition.

D. Pre-training of Source Model

The CVAE [17] framework is used to train the deep speech model given by (3). Suppose we have the PSDs of clean speech signals $\{\mathbf{S}_i\}_{i=1}^I \in \mathbb{R}_+^{I\times F\times T}$ with frame-wise phonetic labels $\{\mathbf{Z}_i^{\mathrm{p}}\}_{i=1}^I \in \{0,1\}^{I\times P\times T}$ and speaker labels $\{\mathbf{z}_i^{\mathrm{s}}\}_{i=1}^I \in \{0,1\}^{I\times S}$ as training data, where I is the number of samples. Let $\{\mathbf{Z}_i\}_{i=1}^I \in \mathbb{R}^{I\times D\times T}$ be the corresponding latent variables. We aim to train a probabilistic decoder $p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ as the source model by maximizing the marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$. Because $p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ and the true posterior density $p_{\boldsymbol{\theta}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ are intractable, we introduce a variational posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ to approximate the true posterior. We here aim to maximize a variational lower bound $\mathcal{L}_{\mathrm{CVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ derived as follows:

$$\begin{aligned}&\log p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}) \\ &\geq \int q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})p(\mathbf{Z}_i)}{q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})} d\mathbf{Z}_i \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})}[\log p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})] \\ &\quad - \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})||p(\mathbf{Z}_i)) := \mathcal{L}_{\mathrm{CVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}), \quad (6)\end{aligned}$$

where $\mathrm{KL}(q||p)$ denotes the Kullback-Leibler (KL) divergence between two probability distributions q and p, and the equality holds if and only if $p_{\boldsymbol{\theta}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}) = q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$. We build a CVAE that consists of a decoder $p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ and an encoder $q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})$ formulated as follows:

$$p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}) = \prod_{f=1}^{F}\prod_{t=1}^{T} \mathcal{N}_{\mathbb{C}}([\mathbf{S}_i]_{ft} \mid \mathbf{0}, [\hat{\boldsymbol{\lambda}}_i]_{ft}), \quad (7)$$

$$q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}) = \prod_{d=1}^{D}\prod_{t=1}^{T} \mathcal{N}([\mathbf{Z}_i]_{dt} \mid [\boldsymbol{\mu}_i]_{dt}, [\boldsymbol{\sigma}_i^2]_{dt}), \quad (8)$$

where $\hat{\boldsymbol{\lambda}} \in \mathbb{R}_+^{F\times T}$ are the output of the decoder, and $\boldsymbol{\mu}_i \in \mathbb{R}^{D\times T}$ and $\boldsymbol{\sigma}_i^2 \in \mathbb{R}_+^{D\times T}$ are the output of the encoder. The first term of (6) can be approximated by using a Monte Carlo integration method as follows:

$$\begin{aligned}&\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})}[\log p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})] \\ &\qquad \simeq \frac{1}{L}\sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\mathbf{S}_i|\mathbf{Z}_i^{(l)}, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}}), \quad (9)\end{aligned}$$

where L is the number of samples. The second term of (6) can be calculated analytically as follows:

$$\begin{aligned}&-\mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{Z}_i|\mathbf{S}_i, \mathbf{Z}_i^{\mathrm{p}}, \mathbf{z}_i^{\mathrm{s}})||p(\mathbf{Z}_i)) \\ &\quad = \frac{1}{2}\sum_{d=1}^{D}\sum_{t=1}^{T}\big(1 + \log([\boldsymbol{\sigma_i}^2]_{dt}) - [\boldsymbol{\mu}_i]_{dt}^2 - [\boldsymbol{\sigma}_i^2]_{dt}\big). \quad (10)\end{aligned}$$

The lower bound $\mathcal{L}_{\mathrm{CVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ given by (6) can be approximately calculated by (7)–(10). Using this lower bound, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be optimized jointly by a stochastic gradient descent (SGD) method. The decoder is then used as the deep speech model in (3).

We also train phone and speaker classifiers. The encoder and the classifiers are used to initialize $\mathbf{Z}$, $\mathbf{Z}^{\mathrm{p}}$, and $\mathbf{z}_i^{\mathrm{s}}$ before source separation. As we will describe in Section III-E, the classifiers can also be used to update these parameters.

E. Source Separation

Given $\mathbf{X}$, we aim to estimate $\mathbf{g} = \{g_n\}_{n=1}^N$, $\mathbf{Z} = \{\mathbf{Z}_n\}_{n=1}^N$, $\mathbf{Z}^{\mathrm{p}} = \{\mathbf{Z}_n^{\mathrm{p}}\}_{n=1}^N$, $\mathbf{Z}^{\mathrm{s}} = \{\mathbf{z}_n^{\mathrm{s}}\}_{n=1}^N$, and $\mathbf{R}$ that maximize the log-likelihood function given by (5). Because (5) is hard to directly maximize, we use a minorization-maximization algorithm that maximizes a lower bound $\mathcal{L}(\boldsymbol{\lambda}, \mathbf{R})$ of (5) given by

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\lambda}, \mathbf{R}) \geq &-\sum_{n,f,t}\frac{\mathrm{Tr}(\mathbf{X}_{ft}\boldsymbol{\Phi}_{nft}^{\mathsf{H}}\mathbf{R}_{nf}^{-1}\boldsymbol{\Phi}_{nft})}{\lambda_{nft}} \\ &- \sum_{n,f,t}\lambda_{nft}\mathrm{Tr}(\boldsymbol{\Omega}_{ft}^{-1}\mathbf{R}_{nf}) + \mathrm{const} := \mathcal{L}(\boldsymbol{\lambda}, \mathbf{R}), \quad (11)\end{aligned}$$

where $\boldsymbol{\Phi}_{nft}$ and $\boldsymbol{\Omega}_{ft}$ are auxiliary variables and the equality holds if and only if $\boldsymbol{\Phi}_{nft} = \lambda_{nft}\mathbf{R}_{nf}\big(\sum_n \lambda_{nft}\mathbf{R}_{nf}\big)^{-1}$ and $\boldsymbol{\Omega}_{ft} = \hat{\mathbf{X}}_{ft}$. The update rule of $\mathbf{g}$ is given by

$$g_n \leftarrow g_n \times \sqrt{\frac{\sum_{f,t}[\mathrm{DNN}_{\boldsymbol{\theta}}(\mathbf{Z}_n, \mathbf{Z}_n^{\mathrm{p}}, \mathbf{z}_n^{\mathrm{s}})]_{ft}\mathrm{Tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{R}_{nf})}{\sum_{f,t}[\mathrm{DNN}_{\boldsymbol{\theta}}(\mathbf{Z}_n, \mathbf{Z}_n^{\mathrm{p}}, \mathbf{z}_n^{\mathrm{s}})]_{ft}\mathrm{Tr}(\hat{\mathbf{X}}_{ft}^{-1}\mathbf{R}_{nf})}}. \quad (12)$$

The update rule of $\mathbf{R}$ is given by

$$\mathbf{R}_{nf} \leftarrow (\mathbf{R}_{nf}\mathbf{A}_{nf}\mathbf{R}_{nf})\#\mathbf{B}_{nf}^{-1}, \quad (13)$$

where $\mathbf{A}_{nf} = \sum_t \lambda_{nft}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{X}_{ft}\hat{\mathbf{X}}_{ft}^{-1}$, $\mathbf{B}_{nf} = \sum_t \lambda_{nft}\hat{\mathbf{X}}_{ft}^{-1}$. $\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$ denotes the geometric mean of two symmetric positive definite matrices [21].

$\mathbf{Z}$, $\mathbf{Z}^{\mathrm{p}}$, and $\mathbf{Z}^{\mathrm{s}}$ can be updated by backpropagation. Since each element of $\mathbf{z}_{nt}^{\mathrm{p}}$ and $\mathbf{z}_n^{\mathrm{s}}$ should be positive and their summation should be one, we define $\mathbf{a}_{nt}^{\mathrm{p}} \in \mathbb{R}^P$ and $\mathbf{a}_n^{\mathrm{s}} \in \mathbb{R}^S$ as parameters to be updated and obtain $\mathbf{z}_{nt}^{\mathrm{p}}$ and $\mathbf{z}_n^{\mathrm{s}}$ by applying the softmax function to $\mathbf{a}_{nt}^{\mathrm{p}}$ and $\mathbf{a}_n^{\mathrm{s}}$. Alternatively, the phone and speaker classifiers can be used to update $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{s}}$. Specifically, separated signals are fed into the classifiers to derive new $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{s}}$ every few iterations.

Source separation is performed with a multichannel Wiener filter after the lower bound is converged as follows:

$$\hat{\mathbf{c}}_{nft} = \lambda_{nft}\mathbf{R}_{nf}\hat{\mathbf{X}}_{ft}^{-1}\mathbf{x}_{ft}. \quad (14)$$

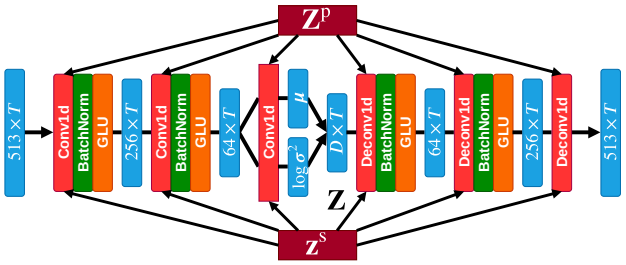


Fig. 2: Network architecture of the CVAE.

IV. Evaluation

This section reports our comparative experiment conducted to validate the effectiveness of using phonetic features and/or speaker identities for speech separation.

A. Experimental Conditions

We used the WSJ0 [22] training set (83 speakers, 15.15 hours) for evaluation. We randomly selected two utterances per speaker and synthesized 83 two-speaker mixtures using Pyroomacoustics [23] ($N = 2$). The reverberation time RT_{60} was 0 or 129 ms and the number of microphones was $M = 2$. The rest of the utterances were used for the pre-training of deep speech models. The phonetic labels of the utterances were obtained by performing forced alignment with a pre-trained GMM-HMM-based acoustic model used for ASR. The number of kinds of phones was $P = 72$.

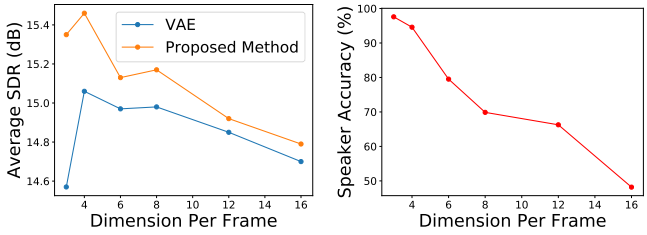
The deep speech model was trained with a CVAE as described in Section III-D. In the same way as previous studies [14], [15], we used gated convolutional networks [24] to build the CVAE and the classifiers as shown in Fig. 2. The power spectrograms were obtained by STFT with a window length of 1024 samples (64 ms) and a shift interval of 256 samples (16 ms). We used AdamW optimizer [25] with a learning rate of 0.001 and L_2 regularization of 0.01 to train the CVAE and classifiers. In addition, gradient clipping [26] with threshold 2.0 and learning rate warmup were applied for the stability of training. We experimented with different settings of dimension of the latent variable $\mathbf{Z}$.

In the separation phase, the parameters of the source models and the spatial model were updated iteratively for 150 iterations. The backpropagation was performed 40 times in one iteration, and $\boldsymbol{\Phi}$ and $\boldsymbol{\Omega}$ were updated every 4 times. We used AdamW optimizer [25] with a learning rate of 0.002 to update the parameters by backpropagation. Additionally, we applied a weight decay of 0.2 to $\mathbf{Z}$ in order to force $\mathbf{Z}$ to fit its prior distribution. $\mathbf{Z}$ was always updated by backpropagation, and $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{s}}$ were updated by backpropagation or using the classifiers. The parameters were initialized with the separated signals and the separation matrix obtained by ILRMA [7] run for 100 iterations. $\mathbf{Z}$, $\mathbf{Z}^{\mathrm{p}}$, and $\mathbf{Z}^{\mathrm{s}}$ were initialized by the outputs of the encoder and the classifiers given the separated signals.

For comparison, we tested a basic speech model trained with a vanilla VAE [12], a speaker-aware model trained with a CVAE conditioned by a speaker label only, and a phone-aware model trained with a CVAE conditioned by phonetic labels only. All these models as well as the proposed phone- and speaker-aware speech model were trained by using the same data. The source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) [27] were used as evaluation measures. We also tested an oracle setting that ground-truth labels are given to the source model and an oracle setting that ground-truth source images are directly used as $\boldsymbol{\lambda}$ in (3).

TABLE I: Average SDRs and phone or speaker classification accuracies of the phone-aware method and the speaker-aware method ($\mathrm{RT}_{60} = 129$ ms).

Model	Method for Update	SDR (dB)	Accuracy
CVAE (only phones)	Backpropagation	14.74 ± 3.89	3.17%
	Classifier	14.92 ± 3.73	55.58%
CVAE (only speaker)	Backpropagation	15.18 ± 3.25	92.17%
	Classifier	14.80 ± 3.69	78.92%



(a) SDR (b) Speaker Accuracy

Fig. 3: Average SDRs and speaker classification accuracies with respect to the dimension of $\mathbf{Z}$ ($\mathrm{RT}_{60} = 129$ ms)

B. Experimental Results

Table I shows the SDRs and the classification accuracies of the phone-aware method and the speaker-aware method. While the backpropagation failed to estimate phonetic labels $\mathbf{Z}^{\mathrm{p}}$, it successfully estimated speaker labels $\mathbf{Z}^{\mathrm{s}}$. We thus report the results obtained by updating $\mathbf{Z}^{\mathrm{p}}$ with the phone classifier and/or updating $\mathbf{Z}^{\mathrm{s}}$ by backpropagation.

Fig. 3 shows the average SDRs of the proposed method and the baseline method using the basic speech model, and the speaker classification accuracies of the proposed method with different dimensions of $\mathbf{Z}$. We found that the separation performance was sensitive to the dimension of $\mathbf{Z}$. Although a larger dimension led to precise speech modeling, the separation performance and the speaker classification accuracy were degraded. This would be because the estimation of $\mathbf{Z}$ easily gets stuck in bad local optima and $\mathbf{Z}$ tends to represent phonetic features and speaker identities, limiting the effect of $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{s}}$ for speech modeling. The dimension per frame of $\mathbf{Z}$ was set to 4 in the following results because it achieved the best separation performance in all methods.

The evaluation results are shown in Table II. The proposed phone- and speaker-aware method achieved the best separation performance. We found that the soft representation of speaker labels estimated by backpropagation gives better performance than using one-hot vector representation. As a result, the speaker-aware methods overperformed the oracle setting using ground-truth labels. We also found that the proposed method

TABLE II: Speech separation performances of the proposed and compared methods.

(a) $\mathrm{RT}_{60} = 0$ ms

Method	$\mathbf{Z}^{\mathrm{p}}$	$\mathbf{Z}^{\mathrm{s}}$	SDR (dB)	SIR (dB)	SAR (dB)	$\mathbf{Z}^{\mathrm{p}}$ Accu.	$\mathbf{Z}^{\mathrm{s}}$ Accu.	Oracle SDR[1] (dB)
VAE	–	–	22.87 ± 6.64	27.70 ± 8.98	27.37 ± 3.46	–	–	–
CVAE (only phones)	✓	–	23.30 ± 6.46	28.11 ± 8.86	27.89 ± 2.73	63.78%	–	23.37 ± 6.40
CVAE (only speaker)	–	✓	23.66 ± 6.55	28.56 ± 8.98	28.36 ± 3.05	–	86.14%	23.40 ± 6.58
Proposed Method	✓	✓	23.76 ± 6.54	28.53 ± 8.93	28.51 ± 2.72	64.08%	87.35%	23.50 ± 6.15
Oracle Setting[2]	–	–	28.40 ± 2.17	35.15 ± 2.90	29.92 ± 2.07	–	–	–

(b) $\mathrm{RT}_{60} = 129$ ms

Method	$\mathbf{Z}^{\mathrm{p}}$	$\mathbf{Z}^{\mathrm{s}}$	SDR (dB)	SIR (dB)	SAR (dB)	$\mathbf{Z}^{\mathrm{p}}$ Accu.	$\mathbf{Z}^{\mathrm{s}}$ Accu.	Oracle SDR[1] (dB)
VAE	–	–	15.06 ± 3.31	21.79 ± 3.31	17.08 ± 1.63	–	–	–
CVAE (only phones)	✓	–	14.92 ± 3.73	21.73 ± 6.01	16.99 ± 1.88	55.58%	–	15.19 ± 3.06
CVAE (only speaker)	–	✓	15.18 ± 3.25	22.04 ± 5.59	17.15 ± 1.64	–	92.17%	15.18 ± 2.96
Proposed Method	✓	✓	15.46 ± 2.78	22.45 ± 5.04	17.31 ± 1.47	56.74%	95.18%	15.38 ± 2.73
Oracle Setting[2]	–	–	16.79 ± 0.90	27.49 ± 1.98	17.24 ± 0.87	–	–	–

[1]Ground-truth labels are given to the source model of $\boldsymbol{\lambda}$ in (3)

[2]Ground-truth source images are directly used as $\boldsymbol{\lambda}$ in (3)

could help to mitigate the block permutation, resulting in a less standard deviation of the separation performance under the condition of $\mathrm{RT}_{60} = 129$ ms. This indicates that the proposed speech model was effectively trained with the help of phonetic and speaker labels such that only frequency-coherent speech spectrograms are allowed to be generated.

V. Conclusion

This paper presented a semi-supervised speech separation method that integrates a phone- and speaker-aware deep speech generative model and a full-rank spatial model into a unified probabilistic model. The deep speech model is pre-trained on clean speech signals with frame-wise phonetic labels and sample-level speaker labels. In the test phase, the latent features, the phonetic labels, and the speaker labels are jointly inferred by backpropagation or classifiers trained in a supervised manner. We experimentally showed that both phonetic and speaker features improved the performance of speech separation.

Acknowledgment: This study was partially supported by JSPS KAKENHI No. 19H04137 and NII CRIS Collaborative Research Program operated by NII CRIS and LINE Corporation.

References

[1] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015, pp. 504–511.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[4] S. Arberet *et al.*, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Proc. ISSPA*, 2010, pp. 1–4.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE TASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

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

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

[10] 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. IEEE ICASSP*, 2018, pp. 716–720.

[11] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE MLSP*, 2018, pp. 1–6.

[12] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2197–2212, 2019.

[13] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE ICASSP*, 2019, pp. 101–105.

[14] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[15] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Underdetermined source separation based on generalized multichannel variational autoencoder," *IEEE Access*, vol. 7, pp. 168 104–168 115, 2019.

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

[17] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. NIPS*, 2014, pp. 3581–3589.

[18] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, p. 62, 2015.

[19] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in *Proc. IEEE ICASSP*, 2016, pp. 146–150.

[20] N. Takahashi *et al.*, "Improving voice separation by incorporating end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2020, pp. 41–45.

[21] M. Congedo, B. Afsari, A. Barachant, and M. Moakher, "Approximate joint diagonalization and geometric mean of symmetric positive definite matrices," *PLOS ONE*, vol. 10, no. 4, pp. 1–25, 2015.

[22] J. Garofolo, D. Graff, P. Paul, and D. Pallett, "CSR-I (WSJ0) complete," in *Philadelphia: Linguistic Data Consortium*, 2007.

[23] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE ICASSP*, 2018, pp. 351–355.

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

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[26] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, 2013, pp. 1310–1318.

[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.