STATISTICAL SPEECH ENHANCEMENT BASED ON PROBABILISTIC INTEGRATION OF VARIATIONAL AUTOENCODER AND NON-NEGATIVE MATRIX FACTORIZATION

Yoshiaki Bando[1], *Masato Mimura*[1], *Katsutoshi Itoyama*[1], *Kazuyoshi Yoshii*[1,2], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2]Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo 103-0027, Japan

ABSTRACT

This paper presents a statistical method of single-channel speech enhancement that uses a variational autoencoder (VAE) as a prior distribution on clean speech. A standard approach to speech enhancement is to train a deep neural network (DNN) to take noisy speech as input and output clean speech. Although this supervised approach requires a very large amount of pair data for training, it is not robust against unknown environments. Another approach is to use non-negative matrix factorization (NMF) based on basis spectra trained on clean speech in advance and those adapted to noise on the fly. This semi-supervised approach, however, causes considerable signal distortion in enhanced speech due to the unrealistic assumption that speech spectrograms are linear combinations of the basis spectra. Replacing the poor linear generative model of clean speech in NMF with a VAE—a powerful nonlinear deep generative model—trained on clean speech, we formulate a unified probabilistic generative model of noisy speech. Given noisy speech as observed data, we can sample clean speech from its posterior distribution. The proposed method outperformed the conventional DNN-based method in unseen noisy environments.

***Index Terms*—** Single-channel speech enhancement, variational autoencoder, Bayesian signal processing

1. INTRODUCTION

Deep neural networks (DNNs) have demonstrated excellent performance in single-channel speech enhancement [1–6]. The denoising autoencoder (DAE) [1], for example, is a typical variant of such networks, which is trained to directly convert a noisy speech spectrogram to a clean speech spectrogram with a supervised training. Alternatively, a DNN can be trained to predict time-frequency (TF) masks called ideal ratio masks (IRMs) that represent ratios of speech to input signals and are used for obtaining a speech spectrogram from a noisy spectrogram [4]. Although it is necessary to prepare as training data a large amount of pairs of clean speech signals and their noisy versions, these supervised methods often deteriorate in unknown noisy environments. This calls for semi-supervised methods that are trained by using only clean speech data in advance and then adapt to unseen noisy environments.

Statistical source separation methods based on the additivity of speech and noise spectrograms have also been used for speech enhancement [7–11]. Non-negative matrix factorization (NMF) [9, 12], for example, regards a noisy speech spectrogram as a non-negative matrix and approximates it as the product of two non-negative matrices (a set of basis spectra and a set of the corresponding activations). If a partial set of basis spectra is trained in advance from clean speech spectrograms, the noisy spectrogram is decomposed into the sum of speech and noise spectrograms in a semi-supervised manner. Robust principal component analysis (RPCA) [13, 14] is another promising method that can decompose a noisy spectrogram into a sparse speech spectrogram and a low-rank noise spectrogram in an unsupervised manner. These conventional statistical methods, however, have a common problem that the linear representation or the sparseness assumption of speech spectrograms is not satisfied in reality and results in considerable signal distortion.

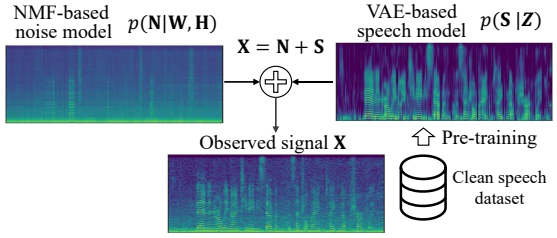


Fig. 1. Overview of our speech enhancement model.

Recently, deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have gained a lot of attention for learning a probability distribution over complex data (e.g., images and audio signals) that cannot be represented by conventional linear models [15–19]. GANs and VAEs are both based on two kinds of DNNs having different roles. In GANs [15], a *generator* is trained to synthesize data that fool a *discriminator* from a latent space while the discriminator is trained to detect synthesized data in a minimax-game fashion. In VAEs [16, 17], on the other hand, an *encoder* that embeds observed data into a latent space and a *decoder* that generates data from the latent space are trained jointly such that the lower bound of the log marginal likelihood for the observed data is maximized. Although in general GANs can generate more realistic data, VAEs provide a principled scheme of inferring the latent representations of both given and new data.

In this paper we propose a unified probabilistic generative model of noisy speech spectra by combining a VAE-based generative model of speech spectra with an NMF-based generative model of noise spectra (Fig. 1). The VAE is trained in advance from a sufficient amount of clean speech spectra and its decoder is used as a prior distribution on clean speech spectra included in noisy speech spectra. Given observed data, we can estimate both the latent representations of speech spectra as well as the basis spectra and their activations of noise spectra through Bayesian inference based on a Markov chain Monte Carlo (MCMC) initialized by the encoder of the VAE. Our Bayesian approach can adapt to both unseen speech and noise spectra by using prior knowledge of clean speech and the low-rankness assumption on noise instead of fixing all the parameters in advance.

Thanks to JSPS KAKENHI No. 15J08765 for funding.

*Demo page: http://sap.ist.i.kyoto-u.ac.jp/members/yoshiaki/demo/vae-nmf/

2. RELATED WORK

This section overviews DNN-based speech enhancement and introduces the variational autoencoder (VAE).

2.1. DNN-based speech enhancement

Various network architectures and cost functions for enhancing speech signals have been reported [1–6]. The popular approach of DNN-based speech enhancement is to train a DNN to directly represent clean speech [6]. The DNN is trained using simulated noisy data constructed by adding noise to speech as input and clean speech as the target. There are several methods that combine a supervised NMF and a DNN [20, 21]. A DNN is trained to estimate activation vectors of the pre-trained basis vectors corresponding to speech and noise. Bayesian WaveNet [22] uses two networks: one, called a prior network, represents how likely a signal is speech and the other, called a likelihood network, represents how likely a signal is included in the observation. These two networks enhance the noisy speech signal with a maximum a posteriori (MAP) estimation. Another reported method uses two networks that are trained to represent how likely the input signal is speech or noise, respectively [23]. The speech signal is enhanced by optimizing a cost function so that the estimated speech maximizes the speech-likelihood network and minimizes the noise-likelihood network. All the above mentioned methods are trained with datasets of both speech and noise signals. A DNN-based method using only training data of speech signals was reported [24]. This method represents speech and noise spectra with two autoencoders (AEs). The AE for speech is pre-trained, whereas that for noise is trained at the inference for adapting to the observed noise signal. Since the inference of this framework is under-determined, the estimated speech is constrained to be represented by a pre-trained NMF model. It, thus, might have the same problem as the semi-supervised NMF.

2.2. Variational autoencoder

A VAE [16] is a framework for learning the probability distribution of a dataset. In this subsection, we denote by $\mathbf{X}$ a dataset that contains F-dimensional samples $\boldsymbol{x}_t \in \mathbb{R}^F$ $(t = 1, \ldots, T)$. The VAE assumes that a D-dimensional latent variable (denoted by $\boldsymbol{z}_t \in \mathbb{R}^D$) follows a standard Gaussian distribution and each sample $\boldsymbol{x}_t$ is stochastically generated from a conditional distribution $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$:

$$\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)\,, \tag{1}$$

$$\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{z}_t)\,, \tag{2}$$

where $\mathcal{N}(\mu, \sigma)$ represents a Gaussian distribution with mean parameter μ and variance parameter σ. $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ is called a decoder and parameterized as a well-known probability density function whose parameters are given by nonlinear functions represented as neural networks. For example, Kingma et al. [16] reported a VAE model that has the following Gaussian likelihood function:

$$\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{z}_t) = \prod_f p(x_{ft}|\boldsymbol{z}_t) = \prod_f \mathcal{N}\left(\mu_f^x(\boldsymbol{z}_t), \sigma_f^x(\boldsymbol{z}_t)\right), \tag{3}$$

where $\mu_f^x : \mathbb{R}^D \to \mathbb{R}$ and $\sigma_f^x : \mathbb{R}^D \to \mathbb{R}_+$ are neural networks representing the mean and variance parameters, respectively.

The objective of VAE training is to find a likelihood function $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ that maximizes the log marginal likelihood:

$$\underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} \log p(\mathbf{X}) = \underset{p(\boldsymbol{x}_t|\boldsymbol{z}_t)}{\operatorname{argmax}} \prod_t \int p(\boldsymbol{x}_t|\boldsymbol{z}_t)\, p(\boldsymbol{z}_t)\, d\boldsymbol{z}_t. \tag{4}$$

Since calculating this marginal likelihood is intractable, it is approximated with a variational Bayesian (VB) framework. The VAE first approximates the posterior distribution of $\boldsymbol{z}_t$ with the following variational posterior distribution $q(\boldsymbol{z}_t)$ called an encoder:

$$p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T|\mathbf{X}) \approx \prod_t q(\boldsymbol{z}_t) = \prod_{d,t} q(z_{dt}) \tag{5}$$

$$= \prod_{d,t} \mathcal{N}(\mu_d^z(\boldsymbol{x}_t), \sigma_d^z(\boldsymbol{x}_t))\,, \tag{6}$$

where $\mu_d^z : \mathbb{R}^F \to \mathbb{R}$ and $\sigma_d^z : \mathbb{R}^F \to \mathbb{R}_+$ are nonlinear functions representing the mean and variance parameters, respectively. These functions are formulated with DNNs. By using the variational posterior, the log marginal likelihood is lower-bounded as follows:

$$\log p(\mathbf{X}) = \sum_t \log \int p(\boldsymbol{x}_t|\boldsymbol{z}_t)\, p(\boldsymbol{z}_t)\, d\boldsymbol{z}_t \tag{7}$$

$$\geq \sum_t \int q(\boldsymbol{z}_t) \log \frac{p(\boldsymbol{x}_t|\boldsymbol{z}_t)\, p(\boldsymbol{z}_t)}{q(\boldsymbol{z}_t)} d\boldsymbol{z}_t \tag{8}$$

$$= -\sum_t \mathbb{KL}\left[q(\boldsymbol{z}_t)\,|p(\boldsymbol{z}_t)\right] + \sum_k \mathbb{E}_q\left[\log p(\boldsymbol{x}_t|\boldsymbol{z}_t)\right], \tag{9}$$

where $\mathbb{KL}[\cdot\,|\cdot]$ represents the Kullback-Leibler divergence. The VAE is trained so that $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ and $q(\boldsymbol{z}_t)$ maximize this variational lower bound. The first term of Eq. (9) is analytically tractable and the second term can be approximated with a Monte-Carlo algorithm. The lower bound can be maximized by using a stochastic gradient descent (SGD) [25].

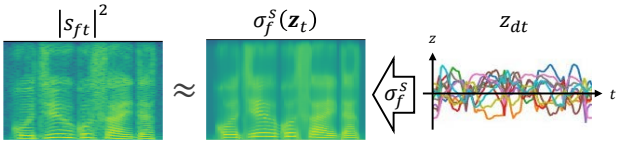


Fig. 2. VAE representation of a speech spectrogram.

3. STATISTICAL SPEECH ENHANCEMENT BASED ON COMBINATION OF VAE AND NMF

This section describes the proposed probabilistic generative model called VAE-NMF, that combines a VAE-based speech model and a NMF-based noise model. We formulate the generative process of an observed complex spectrogram $\mathbf{X} \in \mathbb{C}^{F\times T}$ by formulating the process of a speech spectrogram $\mathbf{S} \in \mathbb{C}^{F\times T}$ and a noise spectrogram $\mathbf{N} \in \mathbb{C}^{F\times T}$. The characteristics of speech and noise signals are represented by their priors based on VAE and NMF, respectively.

3.1. VAE-based speech model

In our speech model we assume a frame-wise D-dimensional latent variable $\mathbf{Z} \in \mathbb{R}^{D\times T}$. Each time-frame of the latent variable $\boldsymbol{z}_t$ is supposed to represent the characteristics of a speech spectrum such as fundamental frequency, spectral envelope, and type of phoneme. The specific representation of $\boldsymbol{z}_t$ is obtained automatically by conducting the VAE training with a dataset of clean speech spectra. As in the conventional VAEs, we put the standard Gaussian prior on each element of $\mathbf{Z}$:

$$z_{dt} \sim \mathcal{N}(0, 1)\,. \tag{10}$$

Since the speech spectra are primarily characterized by its power spectral density (PSD), it follows a zero-mean complex Gaussian distribution whose variance parameter is formulated with $\mathbf{Z}$ (Fig. 2):

$$s_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_f^s(\boldsymbol{z}_t)\right), \tag{11}$$

where $\mathcal{N}_{\mathbb{C}}(\mu, \sigma)$ is a complex Gaussian distribution with mean parameter μ and variance parameter σ. $\sigma_f^s(\boldsymbol{z}_t) : \mathbb{R}^D \to \mathbb{R}_+$ is a

nonlinear function representing the relationship between $\mathbf{Z}$ and the speech signal $\mathbf{S}$. This function is formulated by using a DNN and obtained by the VAE training.

3.2. Generative model of mixture signals

In our Bayesian generative model, the input complex spectrogram $\mathbf{X} \in \mathbb{C}^{F\times T}$ is represented as the sum of a speech spectrogram $\mathbf{S}$ and a noise spectrogram $\mathbf{N}$:

$$x_{ft} = s_{ft} + n_{ft}. \tag{12}$$

We put the VAE-based hierarchical prior model (Eqs. (10) and (11)) on the speech spectrogram $\mathbf{S}$. On the other hand, we assume that the PSD of the noise spectrogram is low-rank and put an NMF-based prior model on it. More specifically, the PSD of a noise spectrogram can be represented as the product of K spectral basis vectors $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}_+^{F\times K}$ and their activation vectors $\mathbf{H} \in \mathbb{R}_+^{K\times T}$. The zero-mean complex Gaussian distribution is put on each TF bin of the noise spectrogram $\mathbf{N}$ as follows:

$$n_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk}h_{kt}\right). \tag{13}$$

For mathematical convenience, we put conjugate prior distributions on $\mathbf{W}$ and $\mathbf{H}$ as follows:

$$w_{fk} \sim \mathcal{G}(a_0, b_0), \qquad h_{kt} \sim \mathcal{G}(a_1, b_1), \tag{14}$$

where $\mathcal{G}(\alpha, \beta)$ is a gamma distribution with the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$; a_0, b_0, a_1, and b_1 are hyperparameters that should be set in advance.

By marginalizing out the speech and noise complex spectrograms $\mathbf{S}$ and $\mathbf{N}$, we obtain the following Gaussian likelihood:

$$x_{ft}|\,\mathbf{W}, \mathbf{H}, \mathbf{Z} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t)\right). \tag{15}$$

Since this likelihood function is independent of the phase term of the input spectrogram $\mathbf{X}$, it is equivalent to the following exponential likelihood:

$$|x_{ft}|^2\big|\,\mathbf{W}, \mathbf{H}, \mathbf{Z} \sim \mathrm{Exp}\left(\sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t)\right), \tag{16}$$

where $|x_{ft}|^2$ is the power spectrogram of $\mathbf{X}$ and $\mathrm{Exp}(\lambda)$ is the exponential distribution with a mean parameter λ. Maximization of the exponential likelihood on a power spectrogram corresponds to minimization of Itakura-Saito divergence, which is widely used in audio source separation [12, 26].

3.3. Pre-training of VAE-based speech model

The goal of the pre-training of the VAE-based speech model is to find $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ that maximizes the following marginal likelihood $p(\mathbf{S})$ from the dataset of clean speech signal (denoted by $\mathbf{S} \in \mathbb{C}^{F\times T}$ in this subsection):

$$p(\mathbf{S}) = \prod_t \int p(\boldsymbol{s}_t|\boldsymbol{z}_t)\, p(\boldsymbol{z}_t)\, p\boldsymbol{z}_t. \tag{17}$$

As stated in Sec. 2.2, it is difficult to analytically calculate this marginal likelihood. We approximate it by using the Variational mean-field approximation. Let $q(\mathbf{Z})$ be the variational posterior distribution of $\mathbf{Z}$. Since $p(\mathbf{S}|\mathbf{Z})$ is independent from the phase term of the speech spectrogram $\mathbf{S}$, the variational posterior $q(\mathbf{Z})$ is defined by ignoring the phase term as follows:

$$q(\mathbf{Z}) = \prod_{d,t} q(z_{dt}) = \prod_{d,t} \mathcal{N}\left(\mu_d^z\left(|\boldsymbol{s}_t|^2\right), \sigma_d^z\left(|\boldsymbol{s}_t|^2\right)\right), \tag{18}$$

where $|\boldsymbol{s}_t|^2$ is the power spectrum of $\boldsymbol{s}_t$ and $\mu_d^z : \mathbb{R}_+^F \to \mathbb{R}$ and $\sigma_d^z : \mathbb{R}_+^F \to \mathbb{R}_+$ are nonlinear functions representing the mean and variance parameters of the Gaussian distribution. These two functions are defined with DNNs. The marginal likelihood is approximately calculated as follows:

$$\log p(\mathbf{S}) \geq -\mathbb{KL}\left[q(\mathbf{Z})\,|p(\mathbf{Z})\right] + \mathbb{E}_q\left[\log p(\mathbf{S}|\mathbf{Z})\right] \tag{19}$$

$$= -\sum_{d,t} \frac{1}{2}\left\{\left(\mu_d^z(|\boldsymbol{s}_t|^2)\right)^2 + \sigma_d^z(|\boldsymbol{s}_t|^2) - \log\sigma_d^z(|\boldsymbol{s}_t|^2)\right\} + \sum_{f,t} \mathbb{E}_q\left[-\log\sigma_f^s(\boldsymbol{z}_t) - \frac{|s_{ft}|^2}{\sigma_f^s(\boldsymbol{z}_t)}\right] + \mathrm{const.} \tag{20}$$

The DNNs for σ_f^s, μ_d^z, and σ_d^z are optimized by using SGD so that this variational lower bound is maximized.

3.4. Bayesian inference of VAE-NMF

To enhance the speech signal in a noisy observed signal, we calculate the full posterior distribution of our model: $p(\mathbf{W}, \mathbf{H}, \mathbf{Z}|\mathbf{X})$. Since the true posterior is analytically intractable, we approximate it with a finite number of random samples by using a Markov chain Monte Carlo (MCMC) algorithm [27]. MCMC alternatively and iteratively samples one of the latent variables ($\mathbf{W}$, $\mathbf{H}$, and $\mathbf{Z}$) according to their conditional posterior distributions.

By fixing the speech parameter $\mathbf{Z}$, the conditional posterior distributions $p(\mathbf{W}|\mathbf{X}, \mathbf{H}, \mathbf{Z})$ and $p(\mathbf{H}|\mathbf{X}, \mathbf{W}, \mathbf{Z})$ can be derived with a variational approximation [26, 27] as follows:

$$w_{fk}|\mathbf{X}, \mathbf{H}, \mathbf{Z} \sim \mathcal{GIG}\left(a_0, b_0 + \sum_t \frac{h_{kt}}{\lambda_{ft}}, \sum_t |x_{ft}|^2 \frac{\phi_{ftk}^2}{h_{kt}}\right), \tag{21}$$

$$h_{kt}|\mathbf{X}, \mathbf{W}, \mathbf{Z} \sim \mathcal{GIG}\left(a_1, b_1 + \sum_f \frac{w_{fk}}{\lambda_{ft}}, \sum_f |x_{ft}|^2 \frac{\phi_{ftk}^2}{w_{fk}}\right), \tag{22}$$

$$\lambda_{ft} = \sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t),\ \phi_{ftk} = \frac{w_{fk}h_{kt}}{\sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t)}, \tag{23}$$

where $\mathcal{GIG}(\gamma, \rho, \tau) \propto x^{\gamma-1}\exp(-\rho x - \tau/x)$ is the generalized inverse Gaussian distribution and λ_{ft} and ϕ_{ftk} are auxiliary variables.

The latent variable of speech $\mathbf{Z}$ is updated by using a Metropolis method [27] because it is hard to analytically derive the conditional posterior $p(\mathbf{Z}|\mathbf{X}, \mathbf{W}, \mathbf{H})$. The latent variable is sampled at each time frame by using the following Gaussian proposal distribution $q(\boldsymbol{z}_t^*\,|\boldsymbol{z}_t)$ whose mean is the previous sample $\boldsymbol{z}_t$:

$$\boldsymbol{z}_t^* \sim q(\boldsymbol{z}_t^*\,|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_t, \sigma\mathbf{I}), \tag{24}$$

where σ is a variance parameter of the proposal distribution. This candidate $\boldsymbol{z}_t^*$ is randomly accepted with the following probability:

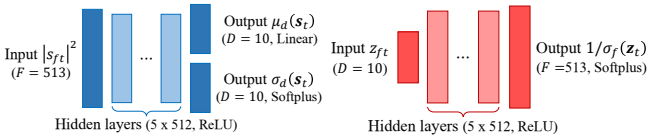
$$a_{\boldsymbol{z}_t^*|\boldsymbol{z}_t} = \min\left(1, \frac{p(\boldsymbol{x}_t|\mathbf{W}, \mathbf{H}, \boldsymbol{z}_t^*)\, p(\boldsymbol{z}_t^*)}{p(\boldsymbol{x}_t|\mathbf{W}, \mathbf{H}, \boldsymbol{z}_t)\, p(\boldsymbol{z}_t)}\right). \tag{25}$$

3.5. Reconstruction of complex speech spectrogram

In this paper we obtain the enhanced speech with Wiener filtering by maximizing the conditional posterior $p(\mathbf{S}|\mathbf{X}, \mathbf{W}, \mathbf{H}, \mathbf{Z})$. Let $\hat{\mathbf{S}} \in \mathbb{C}^{F\times T}$ be the speech spectrogram that maximizes the conditional posterior. It is given by the following equation:

$$\hat{s}_{ft} = \frac{\sigma_f^s(\boldsymbol{z}_t)}{\sum_k w_{fk}h_{kt} + \sigma_f^s(\boldsymbol{z}_t)} x_{ft}. \tag{26}$$

We simply use the mean values of the sampled latent variables as $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{Z}$ in Eq. (26).



(a) Encoder: $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ (b) Decoder: $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$

Fig. 3. Configuration of the VAE used in the Sec. 4.

4. EXPERIMENTAL EVALUATION

This section reports experimental results with noisy speech signals whose noise signals were captured in actual environments.

4.1. Experimental settings

To compare VAE-NMF with a DNN-based supervised method, we used CHiME-3 dataset [28] and DEMAND noise database[1]. The CHiME-3 dataset was used for both the training and evaluation. The DEMAND database was used for constructing another evaluation dataset for unseen noise conditions. The evaluation with the CHiME-3 was conducted by using its development set, which consists of 410 simulated noisy utterances in each of four different noisy environments: on a bus (BUS), in a cafe (CAF), in a pedestrian area (PED) and on a street junction (STR). The average signal-to-noise ratio (SNR) of the noisy speech signals was 5.8 dB. The evaluation with the DEMAND was conducted by using 20 simulated noisy speech signals in each of four different noisy environments: on a subway (SUB), in a cafe (CAF), at a town square (SQU), and in a living room (LIV). We generated these signals by mixing the clean speech signals of the CHiME-3 development set with the noise signals in the DEMAND database. The SNR of these noisy speech signals was set to be 5.0 dB. The sampling rate of these signals was 16 kHz. The enhancement performance was evaluated by using the signal-to-distortion ratio (SDR) [29].

To obtain the prior distribution of speech signals $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$, we trained a VAE that had two networks of $p(\boldsymbol{s}_t|\boldsymbol{z}_t)$ and $q(\boldsymbol{z}_t)$ as shown in Fig. 3. The dimension of the latent variables D was set to be 10. The training data were about 15 hours of clean speech signals in the WSJ-0 corpus [30]. Their spectrograms were obtained with a short-time Fourier transform (STFT) with a window length of 1024 samples and a shifting interval of 256 samples. To make the prior distribution robust against a scale of the speech power, we randomly changed the average power of the spectrogram between 0.0 and 10.0 at each parameter update.

The parameters for VAE-NMF were as follows. The number of bases K was set to be 5. The hyperparameters a_0, b_0, a_1, b_1, and σ were set to be 1.0, 1.0, and 1.0, $K/scale$, and 0.01, respectively. The $scale$ represents the empirical average power of the input noisy spectrogram. After drawing 100 samples for burn-in, we drew 50 samples to estimate the latent variables. These parameters had been determined empirically. The latent variables of noise $\mathbf{W}$ and $\mathbf{H}$ were randomly initialized. Since the latent variable of speech $\mathbf{Z}$ depends on the initial state, the initial sample was drawn from $q(\boldsymbol{z}_t|\boldsymbol{s}_t)$ by setting the observation $\boldsymbol{x}_t$ as the speech signal $\boldsymbol{s}_t$.

We compared VAE-NMF with a DNN-based supervised method and the unsupervised RPCA. We implemented a DNN that outputs IRMs (DNN-IRM). It had five hidden layers with ReLU activation functions. It takes as an input 11 frames of noisy 100-channel log-Mel-scale filterbank features and predicts one frame of IRMs[2]. We trained DNN-IRM with the training dataset of CHiME-3, which was generated by using the WSJ-0 speech utterances and noise signals. The noise signals were recorded in the same environments as those in the evaluated data.

[1] http://parole.loria.fr/DEMAND/

[2] SDRs were evaluated by dropping 2048 samples (5 frames) at both ends.

Table 1. Enhancement performance in SDR for CHiME-3 dataset

Method	Average	BUS	CAF	PED	STR
VAE-NMF	10.10	**9.47**	10.62	10.93	9.39
DNN-IRM	**10.93**	8.92	**11.92**	**12.92**	**9.95**
RPCA	7.53	6.13	8.10	9.13	6.77
Input	6.02	3.26	7.21	8.83	4.78

Table 2. Enhancement performance in SDR for DEMAND dataset

Method	Average	SUB	CAF	SQU	LIV
VAE-NMF	**11.17**	**10.56**	**9.57**	**12.38**	**12.16**
DNN-IRM	9.85	9.13	9.15	10.69	10.42
RPCA	7.03	6.48	6.37	6.99	8.28
Input	5.21	5.25	5.24	5.19	5.16

4.2. Experimental results

The enhancement performance is shown in Tables 1 and 2. In the experiments using the CHiME-3 test set (Table 1), DNN-IRM, which was trained using the noisy data recorded in the same environments at the test data, yielded the highest average SDR. The proposed VAE-NMF achieved higher SDRs than RPCA in all conditions and even outperformed the supervised DNN-IRM in BUS condition without any prior training of noise signals. From the results obtained using the test set constructed with the DEMAND noise data, we can see that VAE-NMF outperformed the other methods in all the conditions. The noise data in DEMAND is unknown to DNN-IRM trained using the CHiME-3 training set, and its enhancement performance deteriorated significantly. These results clearly show the robustness of the proposed VAE-NMF against various types of noise conditions.

The SDR performance of VAE-NMF for the CAF condition in the DEMAND test set was lower than those for the other conditions. In this condition, the background noise contained conversational speech. Since VAE-NMF estimates speech component independently at each time frame, the background conversations were enhanced at the time frames where the power of the target speech was relatively small. This problem would be solved by making the VAE-based speech model to maintain time dependencies of a speech signal. The variational recurrent neural network [31] would be useful for this extension.

5. CONCLUSION

We presented a semi-supervised speech enhancement method, called VAE-NMF, that involves a probabilistic generative model of speech based on a VAE and that of noise based on NMF. Only the speech model is trained in advance by using a sufficient amount of clean speech. Using the speech model as a prior distribution, we can obtain posterior estimates of clean speech by using an MCMC sampler while adapting the noise model to noisy environments. We experimentally confirmed that VAE-NMF outperformed the conventional supervised DNN-based method in unseen noisy environments.

One interesting future direction is to extend VAE-NMF to the multichannel scenario. Since complicated speech signals and a spatial mixing process can be represented by a VAE and a well-studied phase-aware linear model (e.g., [2, 3, 32]), respectively, it would be effective to integrate these models in a unified probabilistic framework. We also investigate GAN-based training of the speech model to accurately learn a probability distribution of speech.

6. REFERENCES

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech*, 2013, pp. 436–440.

[2] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 196–200.

[3] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7092–7096.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Interspeech*, pp. 3642–3646, 2017.

[6] Z.-Q. Wang and D. Wang, "Recurrent deep stacking networks for supervised speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 71–75.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[9] S. Mohammed and I. Tashev, "A statistical approach to semi-supervised speech enhancement with low-order non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 546–550.

[10] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[11] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

[12] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[13] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *Journal of Computational Information Systems*, vol. 10, no. 10, pp. 4403–4410, 2014.

[14] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 57–60.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] O. Fabius and J. R. van Amersfoort, "Variational recurrent auto-encoders," *arXiv preprint arXiv:1412.6581*, 2014.

[18] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Interspeech*, 2017, pp. 1273–1277.

[19] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," in *Interspeech*, 2016, pp. 1770–1774.

[20] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[21] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 499–503.

[22] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," *Interspeech*, pp. 2013–2017, 2017.

[23] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3734–3738.

[24] M. Sun, X. Zhang, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 93–104, 2016.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. 785152, pp. 1–17, 2009.

[27] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[28] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 504–511.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[31] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, 2015, pp. 2980–2988.

[32] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.