INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Adaptive Neural Speech Enhancement with a Denoising Variational Autoencoder

Yoshiaki Bando[1,2], *Kouhei Sekiguchi*[2,3], *Kazuyoshi Yoshii*[2,3]

[1]National Institute of Advanced Industrial Science and Technology, Japan
[2]RIKEN AIP, Japan
[3]Graduate School of Informatics, Kyoto University, Japan

y.bando@aist.go.jp, {sekiguch, yoshii}@kuis.kyoto-u.ac.jp

Abstract

This paper presents a neural speech enhancement method that has a statistical feedback mechanism based on a denoising variational autoencoder (VAE). Deep generative models of speech signals have been combined with unsupervised noise models for enhancing speech robustly regardless of the condition mismatch from the training data. This approach, however, often yields unnatural speech-like noise due to the unsuitable prior distribution on the latent speech representations. To mitigate this problem, we use a denoising VAE whose encoder estimates the latent vectors of clean speech from an input mixture signal. This encoder network is utilized as a prior distribution of the probabilistic generative model of the input mixture, and its condition mismatch is handled in a Bayesian manner. The speech signal is estimated by updating the latent vectors to fit the input mixture while noise is estimated by a nonnegative matrix factorization model. To efficiently train the encoder network, we also propose a multi-task learning of the denoising VAE with the standard mask-based enhancement. The experimental results show that our method outperforms the existing mask-based and generative enhancement methods in unknown conditions.

Index Terms: speech enhancement, deep speech prior, denoising variational autoencoder, nonnegative matrix factorization

1. Introduction

Speech enhancement is an essential function for various applications such as automatic speech recognition and speech telecommunication [1–6]. A standard approach to monaural speech enhancement is to train a deep neural network (DNN) in a supervised manner to discriminatively estimate the time-frequency (TF) mask from a mixture spectrogram [4, 7]. While such a supervised method has demonstrated excellent performance in the known conditions, which are included in the training data, they often deteriorate with unknown conditions such as unknown noise environments and speakers.

To obtain the robustness against various environments, a generative approach has been studied by utilizing the additivity of audio spectrograms [8–12]. A popular generative model is a nonnegative matrix factorization (NMF) model [13, 14], which represents an observed spectrum as the weighted sum of basis spectra. Since the NMF models are limited by its linearity, a generative speech model based on a variational autoencoder (VAE) [15] has been recently gained attention for combining with the NMF-based noise model [10, 11]. A clean speech signal is obtained by estimating the latent variables of the combined model to fit the input mixture signal. While this method precisely estimates the speech spectrogram by using the deep generative model trained from clean speech signals, the noise spectrogram is adaptively estimated without any training data.

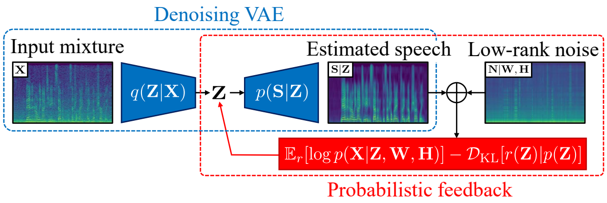
Denoising VAE
Input mixture
$\mathbf{X}$
$q(\mathbf{Z}|\mathbf{X})$
$\mathbf{Z}$
$p(\mathbf{S}|\mathbf{Z})$
Estimated speech
$\mathbf{S}|\mathbf{Z}$
Low-rank noise
$\mathbf{N}|\mathbf{W}, \mathbf{H}$
$\mathbb{E}_r[\log p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \mathbf{H})] - \mathcal{D}_{\mathrm{KL}}[r(\mathbf{Z})|p(\mathbf{Z})]$
Probabilistic feedback

Figure 1: *Overview of our adaptive neural speech enhancement*

A common problem of the VAE-based enhancement methods is that the enhanced speech often includes unnatural speech-like noise, which is caused by an unsuitable prior distribution. The latent vectors generating a speech spectrogram are typically assumed to follow the standard Gaussian distribution [10–12, 16]. While this prior is put on the entire training data, the latent vector for a time frame of each speech utterance has a bias originating from the speech characteristics (e.g., its phone, pitch, and envelope). The existing enhancement methods ignore this bias and unintentionally try to make the estimated speech into an "average" speech signal. In practice, since the latent space has a too high degree of freedom, the estimated signal, including the silent part, is encouraged to be a speech-like signal resulting in the unnatural noise.

In this paper, we combine the discriminative and generative approaches for achieving both the accurate speech reconstruction and high robustness against unknown environments (Fig. 1). We train a denoising VAE that consists of two networks: one is the denoising encoder to estimate the latent vectors of clean speech from a noisy mixture, and the other is the generative decoder to generate a speech spectrogram from the latent variable. This training is efficiently conducted by a multi-task learning with the standard mask-based enhancement. In the test time, we use the output of the encoder as a prior distribution of clean speech and update the speech latent vectors to fit the input mixture signal while the noise signal is estimated by an NMF-based noise model. The inference algorithm is derived as a variational Bayesian inference.

The main contribution of this paper is to efficiently combine discriminative and generative approaches for neural speech enhancement. Since the results of the denoising encoder network are used as prior information, its estimation errors are resolved based on the likelihood function for an input mixture signal. The proposed method thus can be considered as an adaptive speech enhancement method that has the feedback mechanism based on the generative model. We experimentally show that the proposed method outperforms the VAE-based semi-supervised enhancement method. In addition, our method outperforms discriminative enhancement methods in unknown conditions.

*Demo page: https://ybando.jp/demo/is2020/

http://dx.doi.org/10.21437/Interspeech.2020-2291

2. Related Work

This section reviews existing adaptive speech enhancement methods and introduces deep generative models.

2.1. Adaptive neural speech enhancement

Adaptive speech enhancement has been studied by using the additional classifiers or transfer learning [17–23]. Speech enhancement can adapt to a specific speaker by taking as input a speaker embedding estimated from a clean speech [17, 18]. To obtain the generalizability of unknown speakers without using clean speech signals, a multi-task learning of enhancement and speaker embedding has been proposed [19]. A pre-trained enhancement DNN can be adapted to a test noise environment based on transfer learning [21, 22]. This approach uses a set of noisy speech signals of the target domain and adapts the network in an unsupervised manner by conducting an adversarial training [22] or using an additional senone classifier of speech signals [21]. A recent study [23] shows that noise classification with noise embedding can improve the generalization capability of unseen noise conditions. While these methods have been studied to predict a precise enhancement result, the generative approach aims at maintaining the consistency between the estimated results and the observed signal.

2.2. Deep generative models for speech enhancement

The deep spectral modeling has been investigated to improve conventional linear modelings such as NMF [8, 9, 24, 25]. Smaragdis et al. [25] proposed a nonlinear extension of NMF called a nonnegative autoencoder (NAE). They regard the decoder of an autoencoder as an alternative of spectral basis vectors and the latent variables as their activations. As in supervised NMF, an NAE is trained on speech signals, and its decoder is used to separate speech mixtures. It was reported that NAE can efficiently represent speech spectrograms, which have been worsely approximated by NMFs. NAE has been extended with convolutional networks [9] for handling the temporal dependencies of spectrograms, and time-domain networks [24] for representing source signals precisely.

Hybrid models of NMF and deep spectral models have been proposed for speech enhancement [10–12, 16]. The speech enhancement using an NMF-based noise model and deep speech model can work in unknown conditions by estimating the latent variables to fit the input mixture. In this approach, a VAE instead of the normal autoencoder is utilized to regularize the speech model for preventing the overfitting. More precisely, the speech signals are represented by the decoder of a VAE whose latent variable is assumed to follow a standard Gaussian distribution. The original method has been proposed with a Markov-chain Monte-Carlo algorithm [11]. Fast inference algorithms such as Markov-chain expectation-maximization (MCEM) [16] and approximated variational EM (VEM) [10, 12] algorithms have been proposed.

3. Adaptive Neural Speech Enhancement

Our method utilizes a denoising VAE whose encoder estimates the posterior distribution of the speech latent vectors given a mixture signal. We use this posterior distribution as a prior distribution of the probabilistic generative model of the input mixture. Although the output of the encoder network often includes estimation errors due to the condition mismatch, the latent vectors are jointly updated with the NMF-based noise model such that the estimated spectrogram fits the input mixture signal. The enhancement algorithm is formulated as a VEM algorithm.

3.1. Denoising variational autoencoder

We represent a speech spectrogram $\mathbf{S} \in \mathbb{C}^{T \times F}$ based on a deep spectral model [11]. The speech spectrogram is assumed to follow a zero-mean complex Gaussian distribution characterized by D-dimensional latent vectors $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]^\mathsf{T} \in \mathbb{R}^{T \times D}$:

$$s_{tf} \mid \mathbf{Z} \sim \mathcal{N}_\mathbb{C}\left(0, \sigma_{tf}^2(\mathbf{Z})\right), \tag{1}$$

where $\sigma_{tf}^2 : \mathbb{R}^{T \times D} \to \mathbb{R}_+$ is a nonlinear function (i.e., a decoder network) associating the latent representation $\mathbf{Z}$ and the power spectral density (PSD) of the speech spectrogram. In the training, the latent variable $\mathbf{Z}$ is assumed to follow a standard Gaussian distribution:

$$z_{td} \sim \mathcal{N}(0, 1). \tag{2}$$

The denoising VAE consists of a denoising encoder network that predicts the posterior distribution $q(\mathbf{Z}|\mathbf{X})$ given a noisy mixture $\mathbf{X} \in \mathbb{C}^{T \times F}$, and the generative decoder network σ_{tf}^2. More specifically, the encoder network approximates the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ with Gaussian distributions:

$$q(\mathbf{Z}|\mathbf{X}) = \prod_{t,d} \mathcal{N}\left(\mu_{td}(\mathbf{X}), \phi_{td}^2(\mathbf{X})\right) \tag{3}$$

where $\mu_{td}(\mathbf{X}) \in \mathbb{R}$ and $\phi_{td}^2(\mathbf{X}) \in \mathbb{R}_+$ are the outputs of the encoder network.

The encoder and decoder networks are jointly trained by maximizing a lower bound of the log marginal likelihood $\log p(\mathbf{S})$ [26]. This lower bound is called an evidence lower bound (ELBO) whose maximization corresponds to the minimization of the expected Kullback-Leibler (KL) divergence $\mathbb{E}_{p(\mathbf{X}|\mathbf{S})}[\mathcal{D}_\mathrm{KL}[q(\mathbf{Z}|\mathbf{X})|p(\mathbf{Z}|\mathbf{S})]]$. The ELBO for a speech signal $\mathbf{S}$ in the training dataset is defined as follows:

$$\begin{aligned}\mathcal{L}_\mathrm{DnVAE} = \mathbb{E}_{p(\mathbf{X}|\mathbf{S})} \big[\mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}[\log p(\mathbf{S}|\mathbf{Z})] \\ - \mathcal{D}_\mathrm{KL}[q(\mathbf{Z}|\mathbf{X})|p(\mathbf{Z})]\big].\end{aligned} \tag{4}$$

The expectations by $p(\mathbf{X}|\mathbf{S})$ are difficult to calculate because the generative model of a mixture signal $p(\mathbf{X}|\mathbf{S})$ is implicitly assumed in the training stage. They are approximated by the training samples of the clean speech $\mathbf{S}$ and noisy mixture $\mathbf{X}$:

$$\begin{aligned}\mathcal{L}_\mathrm{DnVAE} \approx -\sum_{t,f} \left\{ \mathbb{E}_q[\log \sigma_{tf}^2(\mathbf{Z})] + \mathbb{E}_q\left[\frac{|s_{tf}|^2}{\sigma_{tf}^2(\mathbf{Z})}\right] \right\} \\ + \sum_{u,d} \left\{ \frac{1}{2} + \log \phi_{td}(\mathbf{X}) - \frac{\mu_{td}^2(\mathbf{X}) + \phi_{td}^2(\mathbf{X})}{2} \right\}.\end{aligned} \tag{5}$$

The remaining expectations are approximately calculated with Monte-Carlo sampling from the variational posterior $q(\mathbf{Z}|\mathbf{X})$. The denoising VAE is trained such that the encoder and decoder networks maximize this approximated ELBO by using a stochastic gradient descent (SGD) method [27].

3.2. Multi-task learning

Since the gradient for the denoising encoder is propagated through the decoder, it is difficult to efficiently train the denoising task. To overcome this limitation, we jointly train a mask-based speech enhancement as a multi-task learning. This subtask is called the phase sensitive approximation (PSA) [7], which is often used for speech enhancement:

$$\mathcal{L}_{\mathrm{PSA}} = \sum_{t,f} \left(m_{tf}(\mathbf{X})|x_{tf}| - \cos(\angle x_{tf} - \angle s_{tf})|s_{tf}| \right)^2, \quad (6)$$

where $m_{tf}(\mathbf{X}) \in [0,1]$ is the speech mask estimated by a network that shares the most of the parameters with the encoder network (as detailed in Sec. 4 with Fig. 2). The entire objective function to be maximized is defined as follows:

$$\mathcal{L}_{\mathrm{MTL}} = \mathcal{L}_{\mathrm{DnVAE}} - \alpha \mathcal{L}_{\mathrm{PSA}}, \quad (7)$$

where $\alpha \in \mathbb{R}_+$ is a scaling parameter that controls the effect of the subtask.

3.3. Generative model of mixture signals

To refine the estimated latent vector $\mathbf{Z}$, we utilize a probabilistic generative model of a noisy speech signal [11, 16]. In this model, the input complex spectrogram $\mathbf{X} \in \mathbb{C}^{T \times F}$ is represented by the sum of a speech spectrogram $\mathbf{S} \in \mathbb{C}^{T \times F}$ and a noise spectrogram $\mathbf{N} \in \mathbb{C}^{T \times F}$ as follows:

$$x_{tf} = s_{tf} + n_{tf}. \quad (8)$$

We assume the speech generative model of Eq. (1) as in the denoising VAE. In contrast, we replace the standard Gaussian prior of Eq. (2) with the prior distribution utilizing the outputs of the denoising encoder network:

$$p(\mathbf{Z}) = \prod_{t,d} \mathcal{N}\left(\mu_{td}(\mathbf{X}), \phi_{td}^2(\mathbf{X}) + \sigma_z^2\right), \quad (9)$$

where $\sigma_z^2 \in \mathbb{R}_+$ represents a variance parameter that controls how $\mathbf{Z}$ differs from the output of the encoder network.

On the other hand, the PSD of the noise spectrogram is assumed to be low-rank and represented by an NMF model. More specifically, the noise PSD is represented by the product of K spectral basis vectors $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}_+^{F \times K}$ and their activation vectors $\mathbf{H} \in \mathbb{R}_+^{K \times T}$:

$$n_{tf} \mid \mathbf{W}, \mathbf{H} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk} h_{kt}\right). \quad (10)$$

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

$$x_{tf} \mid \mathbf{W}, \mathbf{H}, \mathbf{Z} \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_k w_{fk} h_{kt} + \sigma_{tf}^2(\mathbf{Z})\right). \quad (11)$$

Maximization of this likelihood is equivalent to minimization of the Itakura-Saito divergence between the observation $|x_{tf}|^2$ and the estimated PSD $\sum_k w_{fk} h_{kt} + \sigma_{tf}^2(\mathbf{Z})$.

3.4. Statistical Inference

The purpose of this inference is to update the latent vector $\mathbf{Z}$ by estimating its posterior $p(\mathbf{Z}|\mathbf{X}, \mathbf{W}, \mathbf{H})$. This posterior is approximately estimated by a variational distribution $r(\mathbf{Z})$:

$$r(\mathbf{Z}) = \prod_{t,d} \mathcal{N}\left(z_{td} \mid a_{td}, \exp(b_{td})\right), \quad (12)$$

where $a_{td} \in \mathbb{R}$ and $b_{td} \in \mathbb{R}$ represent the mean and log-variance parameters of the Gaussian posterior, respectively. These two parameters are initialized by the outputs of the denoising encoder network ($\mu_{td}(\mathbf{X})$ and $\log \phi_{td}^2(\mathbf{X})$) and updated such that the KL divergence $\mathcal{D}_{\mathrm{KL}}[r(\mathbf{Z})|p(\mathbf{Z}|\mathbf{X}, \mathbf{W}, \mathbf{H})]$ is minimized. As in the training of the denoising VAE, this minimization is conducted by maximizing an ELBO:

$$\mathcal{L} = \mathbb{E}_r[\log p(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{Z})] - \mathcal{D}_{\mathrm{KL}}[r(\mathbf{Z})|p(\mathbf{Z})] \quad (13)$$

$$= -\sum_{t,f} \left\{ \mathbb{E}_r[\log y_{tf}] + |x_{tf}|^2 \mathbb{E}_r\left[y_{tf}^{-1}\right] \right\}$$

$$+ \sum_{t,d} \left\{ \frac{b_{td}}{2} - \frac{\exp(b_{td}) + (a_{td} - \mu_{td}(\mathbf{X}))^2}{2(\phi_{td}^2(\mathbf{X}) + \sigma_z^2)} \right\} + \text{const.} \quad (14)$$

where $y_{tf} = \sum_k w_{fk} h_{kt} + \sigma_{tf}^2(\mathbf{Z})$ is the estimated PSD of x_{tf}.The expectations by $r(\mathbf{Z})$ are approximately calculated by the Monte-Carlo sampling. The parameters of the variational posterior a_{td} and b_{td} are updated by an SGD method.

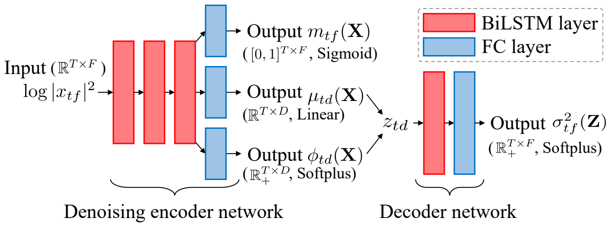
Input $(\mathbb{R}^{T \times F})$
$\log |x_{tf}|^2$
Output $m_{tf}(\mathbf{X})$
$([0,1]^{T \times F}$, Sigmoid)
Output $\mu_{td}(\mathbf{X})$
$(\mathbb{R}^{T \times D}$, Linear)
Output $\phi_{td}(\mathbf{X})$
$(\mathbb{R}_+^{T \times D}$, Softplus)
z_{td}
Output $\sigma_{tf}^2(\mathbf{Z})$
$(\mathbb{R}_+^{T \times F}$, Softplus)
BiLSTM layer
FC layer
Denoising encoder network
Decoder network

Figure 2: *Network architectures of encoder and decoder*

The noise parameters $\mathbf{W}$ and $\mathbf{H}$ are alternatively and iteratively updated with the variational posterior such that the log marginal likelihood $\log p(\mathbf{X}|\mathbf{W}, \mathbf{H})$ is maximized as in [16]:

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_t |x_{tf}|^2 h_{kt} \mathbb{E}_r[y_{tf}^{-2}]}{\sum_t h_{kt} \mathbb{E}_r[y_{tf}^{-1}]}}, \quad (15)$$

$$h_{kt} \leftarrow h_{kt} \sqrt{\frac{\sum_f |x_{tf}|^2 w_{fk} \mathbb{E}_r[y_{tf}^{-2}]}{\sum_f w_{fk} \mathbb{E}_r[y_{tf}^{-1}]}}. \quad (16)$$

In this paper, we calculate the expectations $\mathbb{E}[y_{tf}^{-1}]$ and $\mathbb{E}[y_{tf}^{-2}]$ with 10 samples from $r(\mathbf{Z})$. The estimated clean speech signal is finally obtained by a TF-mask estimated with Wiener filtering from the estimated speech and noise PSDs.

4. Experimental Evaluation

Our speech enhancement is evaluated with simulated noisy speech signals whose noise is recorded in real environments.

4.1. Dataset

The networks used in this evaluation was trained with the training set of the CHiME-4 dataset [28]. This subset has 7138 simulated noisy utterances with the corresponding clean speech and noise signals. The speech signals were provided by the WSJ0 English speech corpus [29]. The noise signals were recorded at the four different environments: on a bus, in a cafeteria, in a pedestrian area, and on a street junction. The sampling rate of these signals was 16 kHz.

The simulated test set of the CHiME-4 dataset was used for evaluating the proposed method in known conditions. This test set has 1320 noisy utterances generated in the same way as those in the training set. Note that the speakers of this test set are independent from the training set. While the noise signals in this set were captured in the same environments of those in the training set, the instances used for each subset are isolated. The average signal-to-noise ratio (SNR) of this subset was 7.51 dB.

To evaluate the performance in unseen conditions, we generated another test set called the TIMIT+ROUEN test set. In this test set, the speech signals were provided by the TIMIT En-

Table 1: *Enhancement results for known (CHiME-4) data*

Method	$p(\mathbf{Z})$	SDR [dB]	PESQ	STOI
BiLSTM-MSA	–	14.22	**2.74**	**0.93**
BiLSTM-PSA	–	**14.62**	2.67	0.92
VAE-NMF	Eq. (2)	12.68	2.56	0.91
DnVAE-NMF w/ MTL	Fixed	13.75	2.67	0.92
DnVAE-NMF w/ MTL	Eq. (2)	13.04	2.59	0.91
DnVAE-NMF	Eq. (9)	13.88	2.68	0.92
DnVAE-NMF w/ MTL	Eq. (9)	14.20	2.70	**0.93**
Noisy mixture	–	7.54	2.18	0.87

Table 2: *Enhancement results for unseen (TIMIT+ROUEN) data*

Method	$p(\mathbf{Z})$	SDR [dB]	PESQ	STOI
BiLSTM-MSA	–	12.14	2.64	0.87
BiLSTM-PSA	–	12.40	2.48	0.87
VAE-NMF	Eq. (2)	12.01	2.56	0.88
DnVAE-NMF w/ MTL	Fixed	11.72	2.55	0.87
DnVAE-NMF w/ MTL	Eq. (2)	12.45	2.57	0.88
DnVAE-NMF	Eq. (9)	12.44	2.59	0.88
DnVAE-NMF w/ MTL	Eq. (9)	**12.75**	**2.65**	**0.89**
Noisy mixture	–	7.57	2.20	0.83

glish speech corpus [30], and the noise signals were provided by the LITIS ROUEN Audio scene dataset [31]. We generated 1320 noisy speech signals by randomly selecting speech and noise signals from these datasets. The speech and noise signals were mixed in SNRs randomly sampled from those of the CHiME-4 test set such that the average SNRs of the two test sets were equivalent.

4.2. Experimental conditions

The architectures of the encoder and decoder networks of the proposed method is summarized in Fig. 2. The encoder network consisted of three bidirectional long-short term memory (BiLSTM) layers followed by three fully-connected (FC) layers that respectively output μ_{td}, ϕ_{td}, and m_{tf}. The decoder consisted of a BiLSTM layer followed by a fully-connected layer that outputs σ^2_{tf}. Each of the BiLSTMs had 512-dimensional hidden units trained with a dropout rate of 0.2.

The hyperparameters of the proposed method were empirically determined as follows. The denoising VAE was trained by an Adam optimizer [27] for 200 epochs at the learning rate of 1.0×10^{-3}. The dimension of the latent variable D and the number of bases K were set to 20 and 5, respectively. The scaling parameter of the multi-task learning α was set to 1.0. The variational posterior $r(\mathbf{Z})$ and the model parameters $\mathbf{W}$ and $\mathbf{H}$ were alternatively and iteratively updated for 200 times. The variational posterior $r(\mathbf{Z})$ was updated by the Adam optimizer at the learning rate of 0.2. The scaling parameter σ_z was set to 0.1. The noise parameters $\mathbf{W}$ and $\mathbf{H}$ were randomly initialized. The input spectrograms were obtained by the short time Fourier transform with the window size of 1024 samples and the hop length of 256 samples.

The proposed method (DnVAE-NMF) was evaluated in terms of the three criteria: source-to-distortion ratio (SDR) in dB [32], perceptual evaluation of speech quality (PESQ) ranging from -0.5 to 4.5 [33], and short-time objective intelligibility (STOI) ranging from 0 to 1 [34]. As baseline methods, we evaluated two supervised mask-based methods called BiLSTM-MSA and -PSA. The BiLSTM-MSA and -PSA train enhancement networks with a magnitude spectrum approximation (MSA) criterion and a PSA criterion [7], respectively. The network architecture of them was the same as that of the denoising encoder network. We also evaluated the semi-supervised version of the proposed method (VAE-NMF) that is a VEM version of [11]. We trained a VAE that has the same architecture as the proposed denoising VAE by using only clean speech signals.

4.3. Experimental results

The enhancement performance is shown in Tables 1 and 2. In the experiments using the CHiME-4 test set (Table 1), the highest performance was achieved by BiLSTM-MSA and -PSA, which are discriminative methods. The proposed DnVAE-NMF with the multi-task learning (MTL) performed comparable to them in STOI and significantly outperformed the semi-supervised VAE-NMF in all of the three criteria. In contrast, in the experiments using the TIMIT+ROUEN test set (Table 2), which is the unknown condition for the discriminative methods, BiLSTM-MSA and -PSA significantly deteriorated. Although the proposed DnVAE-NMF was also degraded in this condition, its performance was improved from those of the mask-based methods in all of the SDR, PESQ, and STOI. These results show that the proposed DnVAE-NMF successfully combined the discriminative and generative methods in a unified framework.

The proposed DnVAE-NMF was also compared with its three variants. The first variant fixed the posterior $r(\mathbf{Z})$ to the output of the denoising encoder $q(\mathbf{Z})$, the second one used the standard Gaussian prior (Eq. (2)) on $\mathbf{Z}$ instead of using the encoder output (Eq. (9)), and the third was trained without multi-task learning. The proposed DnVAE-NMF outperformed all of the three variants, and we can see that updating the posterior $r(\mathbf{Z})$ with a modified prior (Eq. (9)) and conducting the multi-task learning were effective for improving the performance.

5. Conclusion

This paper presented an adaptive neural speech enhancement based on a denoising VAE that consists of a denoising encoder network and a generative decoder network. The output of the denoising encoder network is used to construct a prior distribution of the latent variable, and the latent variable is updated to fit the input mixture by using the speech generative model (decoder) and the NMF-based noise model. We also proposed a multi-task learning of the denoising VAE and mask-based enhancement such that the denoising encoder is efficiently trained. Experimental results showed that our method outperformed mask-based methods in unknown environments.

One interesting future direction is to extent the proposed method to a time-domain method. Recent speech enhancement (and separation) methods improve their performance by directly enhancing time domain signals [35, 36]. The proposed generative framework would be applied to the time-domain methods to obtain robustness against unknown environments. Since our method ignores the probabilistic dependencies of the latent variables among time frames, we will also investigate more precise prior models of the latent variable with a recurrent VAE [10,37].

6. Acknowledgements

This study was partially supported by JSPS KAKENHI No. 20H01159, No. 19K21553, and No. 19H04137.

7. 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," in *Interspeech*, 2017, pp. 3642–3646.

[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] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 708–712.

[8] K. Osako, Y. Mitsufuji, R. Singh, and B. Raj, "Supervised monaural source separation based on autoencoders," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 11–15.

[9] S. Venkataramani, C. Subakan, and P. Smaragdis, "Neural network alternatives to convolutive audio models for source separation," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2017, pp. 1–6.

[10] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 371–375.

[11] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 716–720.

[12] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *Interspeech*, 2019, pp. 3158–3162.

[13] 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.

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

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

[16] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2018, pp. 1–6.

[17] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5554–5558.

[18] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[19] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 181–185.

[20] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 86–90.

[21] S. Wang, W. Li, S. M. Siniscalchi, and C. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6219–6223.

[22] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Interspeech*, 2019, pp. 3148–3152.

[23] H. Li and J. Yamagishi, "Noise tokens: Learning neural noise templates for environment-aware speech enhancement," *arXiv preprint arXiv:2004.04001*, 2020.

[24] S. Venkataramani, E. Tzinis, and P. Smaragdis, "End-to-end non-negative autoencoders for sound source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 116–120.

[25] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 86–90.

[26] D. I. J. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 2059–2065.

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

[28] E. Vincent, S. Watanabe, J. Barker, and R. Marxer, "The 4th CHiME speech separation and recognition challenge," 2016.

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

[30] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[31] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014.

[32] 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.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 749–752.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[35] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

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

[37] 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.