A deep representation learning speech enhancement method using β-VAE

Yang Xiang*†, Jesper Lisby Højvang†, Morten Højfeldt Rasmussen†, Mads Græsbøll Christensen*
* Audio Analysis Lab, CREATE, Aalborg University, Aalbory, Denmark {yaxi,mgc}@create.aau.dk
† Capturi A/S, Aarhus, Denmark {jlh,mhr}@capturi.com

***Abstract*—In previous work, we proposed a variational autoencoder-based (VAE) Bayesian permutation training speech enhancement (SE) method (PVAE) which indicated that the SE performance of the traditional deep neural network-based (DNN) method could be improved by deep representation learning (DRL). Based on our previous work, we in this paper propose to use β-VAE to further improve PVAE's ability of representation learning. More specifically, our β-VAE can improve PVAE's capacity of disentangling different latent variables from the observed signal without the trade-off problem between disentanglement and signal reconstruction. This trade-off problem widely exists in previous β-VAE algorithms. Unlike the previous β-VAE algorithms, the proposed β-VAE strategy can also be used to optimize the DNN's structure. This means that the proposed method can not only improve PVAE's SE performance but also reduce the number of PVAE training parameters. The experimental results show that the proposed method can acquire better speech and noise latent representation than PVAE. Meanwhile, it also obtains a higher scale-invariant signal-to-distortion ratio, speech quality, and speech intelligibility.**

***Index Terms*—deep representation learning, speech enhancement, variational autoencoder, β-VAE**

I. Introduction

The aim of speech enhancement (SE) is to remove background noise from the observed speech signal. In general, SE is mainly used to reduce the word error rate of the automatic speech recognition system [1] or improve speech quality and intelligibility for human listening [2]. Recently, with the wide application of online meeting systems, SE is required to reduce the WER for accurate live caption when providing high-quality speech audio under various complex noise conditions [3]. Thus, SE research is becoming more and more challenging.

During the past decades, many single-channel SE algorithms have been developed, including signal subspace methods [4], non-negative matrix factorization methods [5], [6], and codebook-based methods [7]. In recent years, deep neural networks (DNN) have shown great potential for SE [2], [8]–[14] because DNNs can use a non-linear process to model complex high-dimensional signals, which is more reasonable in practical applications [15]. Thus, DNN-based methods usually have a better SE performance than these previous linear models [4]–[7].

However, most of the regression-based SE algorithms [2], [8]–[10] do not consider applying DNNs to obtain better speech representations when conducting SE. Instead, they usually use DNNs to directly predict pre-defined targets for SE [2]. Although this approach can avoid inaccurate assumptions [8], it cannot ensure that these methods always work in environments with complex noise [2]. In general, deep representation learning (DRL) is important for DNN because DRL can obtain good signal representations in an unsupervised way and can, potentially, improve DNN's ability to extract useful information in complex environments [15], [16]. Additionally, a better signal representation usually leads to better predictions for DNNs [15]. Thus, DRL has a huge potential for DNN-based SE algorithms and makes them more robust. Moreover, the lack of a good DRL strategy may cause poor generalization of DNN-based SE algorithms [2], [15]. A good DRL algorithm can also disentangle various latent representations [15] of speech signals (e.g., speaker and phoneme information), which can also help DNNs achieve a better SE performance.

Recently, to improve traditional DNN's generalization ability, DRL-based SE algorithms are proposed [17]–[22]. The basic idea of these methods is that they use a variational autoencoder (VAE) [23] to learn speech representations when modeling speech, and apply a non-negative matrix factorization (NMF) to model noise. VAE is a DRL model and can perform efficient approximate posterior inference. Additionally, VAE can also learn the probability distribution of complex data. Thus, VAE is suitable for various speech generative tasks [23]–[25]. These VAE-based algorithms can effectively improve DNN's generalization ability, but they have difficulty obtaining good speech representations from the observed signal because they cannot disentangle speech representations from other latent representations [15], [17]–[22]. This causes the need to use a linear NMF to model noise, so their noise modeling ability is limited compared with these non-linear DNN-based methods [23]. And their SE performance is not always satisfactory in a complex noisy environment [18].

To obtain a better speech representation from the observed signal, a novel VAE-based SE method (named PVAE) is proposed [26]. This method applies an unsupervised method to learn signal representations and derives a novel VAE lower bound, which ensures that VAE can disentangle different latent variables from the observed signal. Compared to the previous VAE-based SE algorithms, PVAE can use non-linear DNNs to model noise, which improves the noise modeling ability. Additionally, this method can adopt various DNN structures [2], so the DNN-based SE algorithms [2] can be directly

This work was partly supported by Innovation Fund Denmark (Grant No.9065-00046).

optimized by PVAE. This is not achieved by VAE-NMF-based algorithms [17]–[22]. The experimental results [26] indicate that the SE performance of the traditional DNN-based methods can be improved by introducing this PVAE-based DRL algorithm.

Inspired by previous works, in this paper we propose a novel β-VAE strategy to improve PVAE's representation learning and disentangling performance [15] with fewer DNN parameters. β-VAE [27], [28] is originally designed to push VAE to learn a more efficient latent representation of the data, which is disentangled if the data contains at least some underlying factors of variation [27]. However, in general, β-VAE has a trade-off problem [28]. A better disentanglement within the latent representations usually causes worse signal reconstruction. In this work, based on the VAE's application in SE [26], we propose a strategy to address this trade-off problem to obtain better speech and noise representation. As a result, our β-VAE can improve disentangling and representation performance without signal reconstruction loss. Moreover, the proposed β-VAE can also optimize the neural network structure of the original PVAE. This means that the proposed β-VAE (named β-PVAE) can possibly achieve a better SE performance with fewer training parameters compared to PVAE.

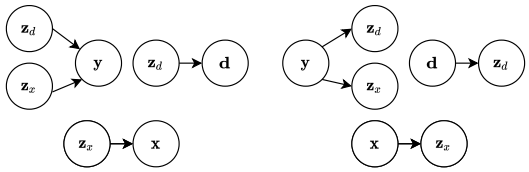
II. Related work

Signal Model: in an additive noisy environment, using the short-time Fourier transform, the observed signal $y_{f,n} \in \mathbb{C}$, speech signal $x_{f,n} \in \mathbb{C}$, and noise $d_{f,n} \in \mathbb{C}$ can be written as

$$y_{f,n} = x_{f,n} + d_{f,n}, \tag{1}$$

where frequency bin $f \in [1, F]$ and time frame index $n \in [1, N]$. N and F denote the number of time frames and frequency bins, respectively. Their log-power spectrum (LPS) vector [8] at each frame can be represented as $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{d}$, respectively, where we omit the frequency and time frame index for simplicity. In [26], we assume that $\mathbf{y}$ is generated from a random process involving the speech latent variables $\mathbf{z}_x \in \mathbb{R}^L$ and the noise latent variables $\mathbf{z}_d \in \mathbb{R}^L$. L is the dimension of latent variables. The latent variables $\mathbf{z}_x$ and $\mathbf{z}_d$ are independent. Similarly, $\mathbf{x}$ and $\mathbf{d}$ are independently generated by $\mathbf{z}_x$ and $\mathbf{z}_d$, respectively. Fig. 1(a) shows the generative process. In [26], it is assumed that $\mathbf{z}_x$ and $\mathbf{z}_d$ can be estimated from speech and noise posterior distributions $p(\mathbf{z}_x|\mathbf{x})$ and $p(\mathbf{z}_d|\mathbf{d})$, respectively, and that they can also be estimated from the noisy speech posterior distributions $p(\mathbf{z}_x|\mathbf{y})$ and $p(\mathbf{z}_d|\mathbf{y})$. To disentangle latent variables, we assume that $p(\mathbf{z}_x, \mathbf{z}_d|\mathbf{y}) = p(\mathbf{z}_x|\mathbf{y})p(\mathbf{z}_d|\mathbf{y})$. Although this assumption is not always accurate in practical environments, it simplifies derivations, and helps us obtain a better signal model. Additionally, its effect towards signal estimation is not significant [26] (related analysis will be also given in Section IV). Fig. 1(b) shows the recognition process.

VAE and β-VAE: the original VAE [23] defines a probabilistic generative process between the observed signal and its latent variables, and provides a principled method to jointly learn latent variables, generative and recognition models. The



(a) Generative model (b) Recognition model

Fig. 1: Graphical illustration of the proposed signal model.

generative and recognition models are jointly trained by maximizing the evidence lower bound [23]

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}[\log q(\mathbf{y})] \geq -\mathcal{L}_n, \\ &\mathcal{L}_n = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[D_{KL}\left(p(\mathbf{z}_y|\mathbf{y})||q(\mathbf{z}_y)\right)\right] \\ &\qquad - \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathbb{E}_{\mathbf{z}_y\sim p(\mathbf{z}_y|\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{z}_y)\right]\right], \end{aligned} \tag{2}$$

where $D_{KL}(||)$ denotes the Kullback-Leibler (KL) divergence. $\mathbf{z}_y \in \mathbb{R}^L$ is the noisy latent variable. Maximizing this lower bound is equivalent to minimizing $\mathcal{L}_n$.

β-VAE [27] is a modification of the original VAE framework, which introduces an adjustable hyperparameter β in the KL divergence term:

$$\begin{aligned} \mathcal{L}_n = &\beta\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[D_{KL}\left(p(\mathbf{z}_y|\mathbf{y})||q(\mathbf{z}_y)\right)\right] \\ &- \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathbb{E}_{\mathbf{z}_y\sim p(\mathbf{z}_y|\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{z}_y)\right]\right]. \end{aligned} \tag{3}$$

In general, $\beta > 1$ results in more disentangled latent representations [27]. Higher values of β can encourage learning a more disentangled representation. However, β-VAE usually has a trade-off problem between the latent representation disentanglement and signal reconstruction.

Bayesian permutation training VAE (PVAE) for SE: Although the VAE-based algorithms [23], [27] can learn signal representations and disentangle latent representations in a self-supervised way, their performance is limited when disentangling desired latent representations for SE application. Therefore, a Bayesian permutation training VAE (PVAE) [26] is proposed for SE. PVAE is a semi-supervised DRL method, which introduces multiple latent variables in VAE and disentangles them in a semi-supervised way. Fig. 2 shows the PVAE framework. It can be seen that PVAE includes three VAE structures: clean speech VAE (C-VAE), noise VAE (N-VAE), and noisy VAE (NS-VAE). C-VAE and N-VAE are trained without supervision to obtain speech and noise latent representations and their posterior estimates $p(\mathbf{z}_x|\mathbf{x})$, $p(\mathbf{z}_d|\mathbf{d})$, respectively. This is achieved by minimizing the following VAE loss function:

$$\begin{aligned} \mathcal{L}_c(\theta_x, \varphi_x; \mathbf{x}) = &\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\{D_{KL}\left(p(\mathbf{z}_x|\mathbf{x})||q(\mathbf{z}_x)\right) \\ &- \mathbb{E}_{\mathbf{z}_x\sim p(\mathbf{z}_x|\mathbf{x})}[\log q(\mathbf{x}|\mathbf{z}_x)]\}, \end{aligned} \tag{4}$$

$$\begin{aligned} \mathcal{L}_d(\theta_d, \varphi_d; \mathbf{d}) = &\mathbb{E}_{\mathbf{d}\sim p(\mathbf{d})}\{D_{KL}\left(p(\mathbf{z}_d|\mathbf{d})||q(\mathbf{z}_d)\right) \\ &- \mathbb{E}_{\mathbf{z}_d\sim p(\mathbf{z}_d|\mathbf{d})}[\log q(\mathbf{d}|\mathbf{z}_d)]\}, \end{aligned} \tag{5}$$

where $\theta_x, \varphi_x, \theta_d, \varphi_d$ are the DNN parameters for the related probability estimation [26]. Additionally, NS-VAE is trained under the supervision of C-VAE and N-VAE's encoders. Based

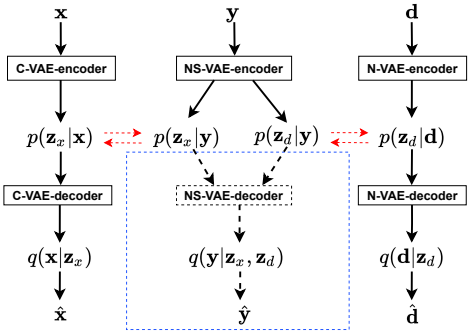


Fig. 2: Model illustration for PVAE and β-PVAE.

on the derivation in [26], the NS-VAE's training loss function can be written as

$$\begin{aligned}
&\mathcal{L}_p(\theta_y,\varphi_y;\mathbf{y})\\
&=\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{x}\sim\mathbf{p}(\mathbf{x})}\{D_{KL}\left(p(\mathbf{z}_x|\mathbf{y})||p(\mathbf{z}_x|\mathbf{x})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_x\sim p(\mathbf{z}_x|\mathbf{y})}[\log\frac{p(\mathbf{z}_x|\mathbf{x})}{q(\mathbf{z}_x)}]\}\\
&\quad+\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{d}\sim p(\mathbf{d})}\{D_{KL}\left(p(\mathbf{z}_d|\mathbf{y})||p(\mathbf{z}_d|\mathbf{d})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_d\sim p(\mathbf{z}_d|\mathbf{y})}[\log\frac{p(\mathbf{z}_d|\mathbf{d})}{q(\mathbf{z}_d)}]\}\\
&\quad-\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathbb{E}_{\mathbf{z}_d,\mathbf{z}_x\sim p(\mathbf{z}_d,\mathbf{z}_x|\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{z}_x,\mathbf{z}_d)\right]\right],
\end{aligned}\tag{6}$$

where θ_y,φ_y are the NS-VAE's network parameters.

In the online SE stage, we assume that the $\mathbf{z}_x,\mathbf{z}_d$ sampled from $p(\mathbf{z}_x|\mathbf{x})$ and $p(\mathbf{z}_d|\mathbf{d})$ are approximately equal to the sample $\mathbf{z}_x$, $\mathbf{z}_d$ sampled from $p(\mathbf{z}_x|\mathbf{y})$, $p(\mathbf{z}_d|\mathbf{y})$, respectively. So, we separately use the NS-VAE encoder's two outputs as input of C-VAE and N-VAE to estimate related signals for SE.

III. β-VAE-based speech enhancement

Inspired by β-VAE, we propose a novel β-VAE strategy (named β-PVAE) to further improve PVAE's SE performance. More specifically, β-VAE is used to improve PVAE's representation learning ability that can better disentangle speech and noise latent variables from the observed signal, which can help PVAE obtain better SE performance. In PVAE, all the PVAE's decoders are trained in an unsupervised way [26]. The accuracy of the restored signal depends on the quality of latent representations. This means that the SE performance in PVAE is determined by the quality of speech and noise latent variables.

In [26], we derived a novel evidence lower bound (ELBO) ($\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}[\log q(\mathbf{y})]\geq-\mathcal{L}_p$). Additionally, β-VAE [27] applies an adjustable hyperparameter β in original VAE's [23] KL divergence term. Following β-VAE's property and PVAE's derivation [26], we apply this hyperparameter in the derived ELBO [26], the (6) can be written as

$$\begin{aligned}
&\mathcal{L}_p(\theta_y,\varphi_y;\mathbf{y})\\
&=\beta\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{x}\sim\mathbf{p}(\mathbf{x})}\{D_{KL}\left(p(\mathbf{z}_x|\mathbf{y})||p(\mathbf{z}_x|\mathbf{x})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_x\sim p(\mathbf{z}_x|\mathbf{y})}[\log\frac{p(\mathbf{z}_x|\mathbf{x})}{q(\mathbf{z}_x)}]\}\\
&\quad+\beta\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{d}\sim p(\mathbf{d})}\{D_{KL}\left(p(\mathbf{z}_d|\mathbf{y})||p(\mathbf{z}_d|\mathbf{d})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_d\sim p(\mathbf{z}_d|\mathbf{y})}[\log\frac{p(\mathbf{z}_d|\mathbf{d})}{q(\mathbf{z}_d)}]\}\\
&\quad-\alpha\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})}\left[\mathbb{E}_{\mathbf{z}_d,\mathbf{z}_x\sim p(\mathbf{z}_d,\mathbf{z}_x|\mathbf{y})}\left[\log q(\mathbf{y}|\mathbf{z}_x,\mathbf{z}_d)\right]\right].
\end{aligned}\tag{7}$$

In (7), we introduce a hyperparameter α in the restoration term. The purpose is to better analyze β-VAE [27] in PVAE. Note, α will not generate any effects for the original β-VAE's property because what is important in (7) is the weight ratio $\beta:\alpha$. This weight ratio can also be written as: $\gamma=\beta:\alpha=(\beta/\alpha):1$, which is equal to the original β-VAE's loss function in (3). β-VAE [27] indicates that a higher value of β encourages VAE learning a more disentangled representation. Thus, we hypothesize that a higher value of $\beta:\alpha$ in (7) can cause a better disentangling performance for speech and noise latent variables. This point will be verified by later experiments.

β-VAE usually has a trade-off problem between the disentanglement and signal reconstruction [27], which means that a good disentangled representation usually leads to poor signal reconstruction performance. In NS-VAE (as shown in Fig. 2), this trade-off is between the quality of observed signal reconstruction and the disentanglement of speech and noise latent variables. In SE application, we only need NS-VAE's disentanglement function, observed signal reconstruction is not useful (dashed part in Fig. 2). This means that we should set a very high weight ratio γ to obtain a better disentanglement performance [27]. Ideally, $\gamma\to+\infty$. One strategy to achieve this purpose is to set $\alpha=0$, so the loss function (7) can be rewritten as

$$\begin{aligned}
\mathcal{L}_\beta(\theta_y;\mathbf{y})&=\beta\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{x}\sim\mathbf{p}(\mathbf{x})}\{D_{KL}\left(p(\mathbf{z}_x|\mathbf{y})||p(\mathbf{z}_x|\mathbf{x})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_x\sim p(\mathbf{z}_x|\mathbf{y})}[\log\frac{p(\mathbf{z}_x|\mathbf{x})}{q(\mathbf{z}_x)}]\}\\
&\quad+\beta\mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}),\mathbf{d}\sim p(\mathbf{d})}\{D_{KL}\left(p(\mathbf{z}_d|\mathbf{y})||p(\mathbf{z}_d|\mathbf{d})\right)\\
&\quad+\mathbb{E}_{\mathbf{z}_d\sim p(\mathbf{z}_d|\mathbf{y})}[\log\frac{p(\mathbf{z}_d|\mathbf{d})}{q(\mathbf{z}_d)}]\}.
\end{aligned}\tag{8}$$

In (8), it can be found that there is no reconstruction term. This means that we do not need to train the NS-VAE's decoder, which reduces the PVAE's training parameters. The dashed part in Fig. 2 is removed in the proposed β-PVAE framework. Comparing the PVAE and proposed β-PVAE, we can find that the β-VAE can be used to optimize the PVAE's network structure and β-PVAE also addresses the β-VAE's trade-off problem for SE application. All in all, the combination of β-VAE and PVAE can not only improve PVAE's disentanglement performance, but also simplify its framework.

To summarize, the proposed β-PVAE includes a training and an enhancement stage for the SE application, which is similar to PVAE [26]. In the training stage, C-VAE and N-VAE are separately pre-trained by self-supervision using (4) and (5). After that, we apply (8) to train NS-VAE. In the enhancement stage, we can separately use the NS-VAE encoder's two outputs as input of C-VAE and N-VAE to obtain the prior distributions $q(\mathbf{x}|\mathbf{z}_x)$ and $q(\mathbf{d}|\mathbf{z}_d)$ for SE. Moreover, to calculate (8), related prior and posterior distributions need to be determined. Here, all the estimations of these distributions are the same as PVAE. More details can be found in [26].

IV. Experiments

In this section, we report two experiments. First, we will investigate the disentanglement ability of the latent variables

TABLE I: Average STOI, PESQ, and SI-SDR comparison for β-PVAE under different γ with a 95% confidence interval (β-PVAE is equal to PVAE when $\gamma = 1$)

Method	STOI	PESQ	SI-SDR
Noisy	88.94($\pm$ 1.77)	2.29($\pm$ 0.02)	8.36($\pm$ 1.13)
Oracle	98.12($\pm$ 0.35)	4.19($\pm$ 0.00)	19.84($\pm$ 0.92)
PVAE ($\gamma = 1$)	89.33($\pm$ 1.72)	2.59($\pm$ 0.03)	10.31($\pm$ 1.03)
$\gamma = 2$	89.81($\pm$ 1.67)	2.69($\pm$ 0.02)	11.84($\pm$ 0.97)
$\gamma = 5$	89.76($\pm$ 1.64)	2.70($\pm$ 0.02)	12.23($\pm$ 0.93)
$\gamma = 10$	89.94($\pm$ 1.70)	2.71($\pm$ 0.02)	12.31($\pm$ 0.94)
$\gamma = 100$	89.98($\pm$ 1.70)	2.72($\pm$ 0.02)	12.45($\pm$ 0.94)
$\gamma = 1000$	90.02($\pm$ 1.71)	2.74($\pm$ 0.01)	12.55($\pm$ 0.94)
$\gamma = +\infty$	90.05($\pm$ 1.71)	2.75($\pm$ 0.01)	13.20($\pm$ 0.95)

in the proposed algorithm. In addition, β-PVAE's SE performance will be indicated.

Datasets: In this work, we use the DNS challenge 2021 corpus [29] to evaluate the performance of the proposed algorithm. We select English speakers and randomly split 70% of speakers for training, 20% for validation, and 10% for evaluation. Then, all the noise from the DNS noise corpus are randomly divided into training, validation, and test noise in a proportion similar to that used for speech utterances. Next, the corresponding training, validation, and test corpus for speech and noise are randomly mixed using DNS script [29] with random signal-to-noise ratio (SNR) levels (SNR range is from –10dB to 15dB). Other parameters of signal mixing are the default values in the DNS script [29]. Finally, we randomly choose 20 hours mixed training utterances, 5 hours mixed validation utterances, and 1 hour mixed test utterances to build experimental dataset. All signals are down-sampled to 16 kHz.

Experimental settings: In the experiments, the neural structures for C-VAE and N-VAE are the same. Their encoders include four hidden 1D convolutional layers [11]. The number of channels in each layer is 32, 64, 128, and 256. The size of each convolving kernel is 3. The two output layers of the encoders are fully connected layers with 128 nodes. Their decoders consist of four hidden 1D convolutional layers (the channel number of each layer is 256, 128, 64, and 32 with 3 kernel) and two fully connected output layers with 257 nodes. For NS-VAE, its encoder's hidden layer setting is the same as C-VAE. NS-VAE's encoder has four output layers with 128 nodes. For C-VAE, N-VAE, and NS-VAE, their activation functions in the hidden and output layer are ReLU and linear activation function, respectively. All networks are trained by the Adam algorithm with a 128 mini-batch size.

Experimental results: To evaluate the SE performance of various algorithms, we will use scale-invariant signal-to-distortion ratio (SI-SDR) in decibel (dB) [30], short-time objective intelligibility (STOI) [31], and perceptual evaluation of speech quality (PESQ) [32] as evaluation metrics.

First, we will investigate β-PVAE's performance in disentangling speech and noise latent variables. Based on our previous derivation and analysis [26], β-PVAE's SE performance is determined by disentanglement performance. Table. I 'Oracle' shows the SE performance with a 95% confidence interval if latent variables are completely disentangled. Here, the signals are reconstructed by mask estimation [9]. The complete disentanglement means that they have the same posterior forms: $p(\mathbf{z}_x|\mathbf{x}) = p(\mathbf{z}_x|\mathbf{y})$ and $p(\mathbf{z}_d|\mathbf{d}) = p(\mathbf{z}_d|\mathbf{y})$. This is because $p(\mathbf{z}_x|\mathbf{x})$ and $p(\mathbf{z}_d|\mathbf{d})$ are learned in an unsupervised way with speech or noise only, which ensures that their latent representation only contains speech or noise representation. 'Oracle' results indicate that β-PVAE achieves a very satisfactory SE performance in SI-SDR, STOI, and PESQ, which shows the importance of disentangling latent variable for achieving excellent SE performance. The NS-VAE's purpose is to disentangle different representations from the observed signal and obtain the closest possible speech and noise posterior. Next, we use KL divergence to evaluate the practical disentanglement performance in latent space. A better disentanglement can lead to a lower KL divergence ($D_{KL}\left(p(\mathbf{z}_d|\mathbf{y})||p(\mathbf{z}_d|\mathbf{d})\right)$ and $D_{KL}\left(p(\mathbf{z}_x|\mathbf{y})||p(\mathbf{z}_x|\mathbf{x})\right)$). Fig. 3 shows the average KL divergence comparison of validation samples for using different ratios $\gamma = \beta : \alpha$ in loss function (7) to train NS-VAE. In (7), we keep $\alpha = 1$ and change different β to determine ratio γ, and $\gamma = +\infty$ means that $\alpha = 0, \beta = 1$, which is equal to (8). In Fig. 3, we see that the KL divergence decreases with the increase of γ for both speech and noise latent variables, which means that the disentangled posteriors get closer to the true posteriors and the NS-VAE achieves a better disentanglement performance. When NS-VAE's decoder is removed ($\gamma = +\infty, \alpha = 0, \beta = 1$), NS-VAE can acquire the best posterior estimation. This verifies our hypothesis and deduction in Section 3. Additionally, although we have an inaccurate posterior conditional assumption $p(\mathbf{z}_x, \mathbf{z}_d|\mathbf{y}) = p(\mathbf{z}_x|\mathbf{y})p(\mathbf{z}_d|\mathbf{y})$, Fig. 3 shows that NS-VAE can still estimate a satisfactory posterior with a low KL divergence. However, this inaccurate assumption may hinder NS-VAE from obtaining a lower KL divergence when $\gamma = +\infty$.

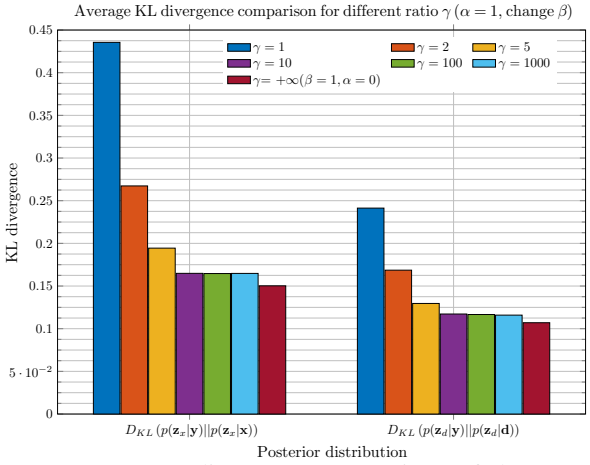


Fig. 3: Average KL divergence comparison of the posterior distribution for different ratio γ.

Next, we will evaluate the SE performance of the proposed β-PVAE. We use basic PVAE [26] as the reference method, which can be more direct to find the effects of β-VAE for the previous PVAE. The enhanced speech is obtained by mask

estimation [9]. Table. I shows the experimental results. We find that β-PVAE achieves a very significant STOI, PESQ, and SI-SDR improvement over PVAE (from $\beta = 1$ to $\beta = 2$). This indicates that good disentanglement performance in latent space can directly lead to an improvement in speech quality and intelligibility. In addition, β-PVAE achieves the best SE performance when $\beta = +\infty$. This illustrates that the proposed β-PVAE can effectively improve PVAE's SE performance with a simpler network structure.

V. Conclusions

In this paper, a β-PVAE-based SE method is proposed to improve previous PVAE's SE performance. More specifically, β-PVAE can improve PVAE's ability to disentangle speech and noise latent variables from the observed signal. In addition, based on VAE's application in SE, the proposed β-PVAE addresses the trade-off problem between disentanglement and signal reconstruction, which widely exists in β-VAE. Compared with the previous PVAE algorithm, β-PVAE also simplifies its neural network and reduces the number of training parameters when improving the SE performance. Experimental results indicate that a good signal representation can achieve a very satisfactory SE performance. Moreover, β-PVAE obtains a better disentanglement performance and achieves higher SI-SDR, PESQ, and STOI scores than PVAE. In future work, we believe that β-PVAE can achieve better SE performance by improving the latent space disentanglement performance or the decoder's signal reconstruction ability.

References

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 745–777, 2014.

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

[3] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human Listening and Live Captioning: Multi-Task Training for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 2686–2690.

[4] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 631–644, 2015.

[5] Y. Xiang, L. Shi, J. L. Højvang, M. H. Rasmussen, and M. G. Christensen, "A novel NMF-HMM speech enhancement algorithm based on poisson mixture model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2021, pp. 721–725.

[6] ——, "An NMF-HMM speech enhancement method based on kullback-leibler divergence," in *Proc. Interspeech*, 2020, pp. 2667–2671.

[7] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 1, pp. 99–113, 2018.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2014.

[9] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Processing Letters*, vol. 28, pp. 2018–2022, 2021.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 1826–1838, 2020.

[13] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[14] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2021, pp. 6628–6632.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] Y. Xie, T. Arildsen, and Z.-H. Tan, "Disentangled speech representation learning based on factorized hierarchical variational autoencoder with self-supervised objective," in *proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2021, pp. 1–6.

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

[18] 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 Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 716–720.

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

[20] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 681–685.

[21] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 676–680.

[22] G. Carbajal, J. Richter, and T. Gerkmann, "Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE, 2021, pp. 126–130.

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

[24] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *arXiv preprint arXiv:2106.06103*, 2021.

[25] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi, "Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis," *arXiv preprint arXiv:2110.08813*, 2021.

[26] Y. Xiang, J. L. Højvang, M. H. Rasmussen, and M. G. Christensen, "A bayesian permutation training deep representation learning method for speech enhancement with variational autoencoder," *Accepted by ICASSP2022 (arXiv preprint arXiv:2201.09875)*, 2022.

[27] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," *International Conference on Learning Representations*, 2017.

[28] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[29] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *Proc. Interspeech*, 2021.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[32] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752.