Conditional Deep Hierarchical Variational Autoencoder for Voice Conversion

Kei Akuzawa*† and Kotaro Onishi*‡ and Keisuke Takiguchi* and Kohki Mametani* and Koichiro Mori*
* DeNA Co., Ltd. Tokyo, Japan
† University of Tokyo, Tokyo, Japan
‡ The University of Electro-Communications, Tokyo, Japan
E-mail: akuzawa-kei@weblab.t.u-tokyo.ac.jp, {kotaro.onishi, keisuke.takiguchi, koki.mametani, koichiro.mori}@dena.com

***Abstract*—Variational autoencoder-based voice conversion (VAE-VC) has the advantage of requiring only pairs of speeches and speaker labels for training. Unlike the majority of the research in VAE-VC which focuses on utilizing auxiliary losses or discretizing latent variables, this paper investigates how an increasing model expressiveness has benefits and impacts on the VAE-VC. Specifically, we first analyze VAE-VC from a rate-distortion perspective, and point out that model expressiveness is significant for VAE-VC because rate and distortion reflect similarity and naturalness of converted speeches. Based on the analysis, we propose a novel VC method using a deep hierarchical VAE, which has high model expressiveness as well as having fast conversion speed thanks to its non-autoregressive decoder. Also, our analysis reveals another problem that similarity can be degraded when the latent variable of VAEs has redundant information. We address the problem by controlling the information contained in the latent variable using β-VAE objective. In the experiment using VCTK corpus, the proposed method achieved mean opinion scores higher than 3.5 on both naturalness and similarity in inter-gender settings, which are higher than the scores of existing autoencoder-based VC methods.**

I. Introduction

Voice conversion (VC) [1] is a technique for modifying a speech from a source speaker to match the vocal characteristics of a target speaker while keeping its phonetic content. VC can be employed in many practical applications such as speaking aids [2] and entertainment (e.g., singing VC [3]).

Although continued research efforts have improved overall quality of VC, there are still problems that prohibits the technique from being used in production. For example, many conventional approaches [4], [5] rely on parallel corpora in which the linguistic contents of speech data of multiple speakers are aligned. It is known that collecting such data is expensive even for one-to-one conversion, let alone more practical situation like many-to-one and many-to-many conversion. Also, recent studies [6], [7] proposed utilizing pre-trained speech recognition and synthesis models for VC. This approach, which is referred to as Phonetic PosteriorGram-VC (PPG-VC), has several advantages: it does not require parallel corpora and can be applied to a many-to-one VC task. However, pre-training the speech recognition model needs speech transcription. Moreover, pre-training the speech synthesis model with expressive or noisy speech corpora remains difficult although some recent studies have tackled the problem [8], [9].

Compared to the methods above, Variational Autoencoder-based VC (VAE-VC) [10] does not require parallel corpora nor text transcriptions. Instead, it can train on pairs of speeches and speaker labels. In addition, since VAE-VC learns by reconstructing speech rather than mapping from text (or PPG) to speech, it is possible in principle to incorporate speeches into training regardless of the quality of them. However, the conversion quality of VAE-VC is still limited compared to PPG-VC as shown in voice conversion challenge (VCC) 2020 [11], even though many researches have improved the quality by utilizing auxiliary losses [12], [13], [14] or discretizing latent variables [15], [16].

Unlike the existing research, this paper investigates how an increasing model expressiveness has benefits and impacts on the VAE-VC. Specifically, we first conduct a rate-distortion (RD) analysis and show that the model expressiveness is significant for getting both good naturalness and similarity of converted speeches. Based on the finding, we propose a novel VC method utilizing deep hierarchical VAEs (DHVAEs) [17], [18], [19], [20], which have high model expressiveness thanks to their hierarchical latent representations. In addition, the conversion is relatively fast thanks to the absence of auto-gressive decoder which is often used in neural VC approaches [15], [7]. However, DHVAEs cannot be used for VC in the same way as conventional VAE because they are unconditional models and have hierarchical latent variables. Therefore, we propose a novel model called conditional deep hierarchical VAE (CDHVAE), which can perform VC by splitting the latent variables into speaker-dependent and invariant variables and inferring to the speaker-invariant latent variables. Also, our RD analysis reveals that the mere use of high model expressiveness is insufficient because similarity can be degraded when the latent variable has redundant speaker information. We address the problem by controlling the information contained in the latent variable using β-VAE objective.

The contributions of this paper are summarized as follows.

- With an analysis from the RD perspective, we show that an increasing expressiveness of the model and β-VAE objective are significant for getting good naturalness and similarity in VAE-VC.
- We propose CDHVAE as one of the instances of VAEs with high model expressiveness. CDHVAE achieved mean opinion scores (MOSs) higher than 3.5 on both

naturalness and similarity in inter-gender settings, which outperforms existing autoencoder-based VC methods.

II. Preliminaries

A. Problem Statement

In this paper, mel-spectrogram is used as an acoustic feature. Let $x \in X$ be a segment of mel-spectrogram where X is the $(80 \times T)$-dimensional Euclidean Space. Here, 80 is the dimension of the features. T is the sequence length of a segment, and we set $T = 40$, which corresponds to 0.5 seconds, in our experiments. Note that an utterance would be split into segments with a sequence length $T = 40$ without overlapping at a conversion step. Also, let $y \in Y$ be the speaker label where Y is the group of all speakers. The training set $D = \{\{x_1, y_1\}, ..., \{x_m, y_m\}\}$ contains m pairs of $(x_i, y_i) \in (X, Y)$, where x_i is produced by the speaker y_i. Given a tuple of a source speech, its speaker label, and target speaker label (x_s, y_s, y_t), the goal of VC is to obtain an acoustic feature $x_{s \to t}$ that contains speaker characteristics of y_t while preserving the linguistic content of x_s.

B. VAE-VC

Here we present a brief overview of VAE-VC [10], [12]. VAE-VC approaches are aimed at obtaining speaker-invariant latent variables that contain only linguistic information by reconstructing speech from the latent variables and speaker labels. Specifically, the methods consider a data generating process where an acoustic feature x is generated from the latent variable z and speaker label y, and parameterize the process with a deep neural network (DNN) decoder $p(x|z, y)$. Here, z and y are defined to be independent, i.e., $p(z, y) = p(z)p(y)$. The independence encourages z to model information invariant to the speaker, i.e., linguistic information. The objective of the VAE is given as follows:

$$\begin{aligned}\max \mathcal{L} &= -D_{KL}[q(z|x, y)||p(z)] + \mathbb{E}_{q(z|x,y)}[\log p(x|z, y)] \\ &=: -\mathrm{KL} - \mathrm{Rec}, \end{aligned} \tag{1}$$

where the first term is the Kullback-Leibler (KL) divergence between the encoder $q(z|x, y)$ and the prior $p(z)$, and the second term is a negative reconstruction error.

Using the trained VAE, the conversion step is performed by decoding the speech from the speaker-invariant latent variable and the label of the target speaker. Specifically, first, the encoder $q(z_s|x_s, y_s)$ is used to extract speaker-invariant latent variables z_s from the source speech x_s. Then, VC can be performed by inputting z_s and target speaker y_t to the decoder $p(x|z_s, y_t)$.

C. Deep Hierarchical Variational autoencoders

DHVAEs, which are also called as bidirectional-inference autoencoder [21], are the family of VAE variants with high model expressiveness [17], [18], [19], [20]. They share the same data generating process while their architectural details are different for each study. It has three components parameterized by DNNs: an encoder $\Pi_{l=1}^{L} q(z_l|x, z_{<l})$, prior $\Pi_{l=1}^{L} p(z_l|z_{<l})$, and decoder $p(x|z)$. Here, contrary to standard Gaussian prior for vanilla VAEs, the prior of DHVAEs enhances the expressiveness by hierarchically modeling L latent variables $z = \{z_1, ..., z_L\}$, where L is a hyperparameter. The objective is given as follows:

$$\begin{aligned}\max \mathcal{V} = &-\sum_{l=1}^{L} \mathbb{E}_{q(z_{<l}|x)} D_{KL}[q(z_l|x, z_{<l})||p(z_l|z_{<l})] \\ &+ \mathbb{E}_{q(z|x)}[\log p(x|z)], \end{aligned} \tag{2}$$

which also consists of the KL and reconstruction terms.

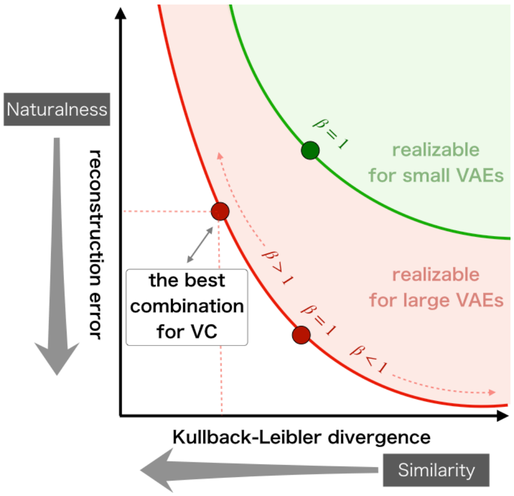
Naturalness
reconstruction error
β = 1
realizable for small VAEs
β > 1
realizable for large VAEs
β = 1
β < 1
the best combination for VC
Kullback-Leibler divergence
Similarity

Fig. 1. Schematic representation of the rate-distortion analysis on VAE-VC. y-axis measures the reconstruction error, which affects naturalness. x-axis measures the KL divergence between the encoder and the prior, which affects similarity.

III. Proposed Approach

A. Rate-distortion perspective on VAE-VC

In this section, we adapt the RD analysis on VAE from [22] to voice conversion setting. Based on the analysis, we will point out that it is necessary to use VAE with high model expressiveness for getting both good naturalness and similarity. Speech naturalness and similarity are subjective measures. The former qualifies how natural (human-like) the converted speech $x_{s \to t}$ is. The latter qualifies how similar the converted speech $x_{s \to t}$ is to that of the target speaker y_t. Here, the *rate* equals to the KL divergence term (KL), which measures the degree of compression. Also, the *distortion* equals to the reconstruction error term (Rec).

The RD analysis givin in [22] suggests that there is a tradeoff between KL and Rec, given a model architecture and the accompanied expressiveness. Specifically, the phase diagram in Figure 1 shows the realizable combination of KL and Rec for a given model (either large or small). Also, the Pareto frontier, i.e., the optimal combination of KL and Rec, is called a RD curve. The RD curve consists of a set of models trained by the following *β-VAE objective*, in which the weighting

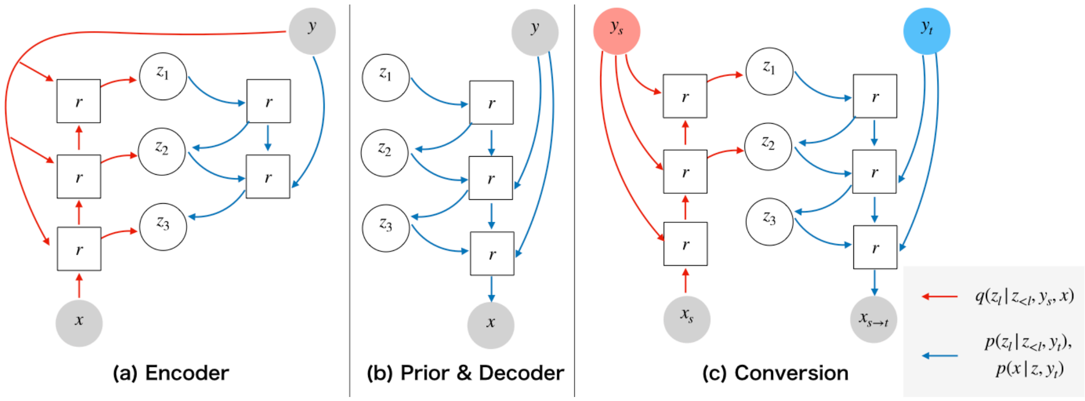

Fig. 2. Schematic diagrams of CDHVAE with $L \coloneqq 3$ and $K \coloneqq 2$. (a) Inference process of the encoder $\Pi_{l=1}^{L} q(z_l|x, y, z_{<l})$. (b) Generating process of the decoder $p(x|z, y)$ and the prior $\Pi_{l \leq K} p(z_l|z_{<l}) \Pi_{l>K} p(z_l|z_{<l}, y)$. (c) Conversion process. Note that there is no red arrow from r to z_3 in (c) because z_3 is sampled from prior at conversion step. Here, r denotes residual unit that supports hierarchical modeling (further information can be found in [17], [19]).

parameter $\beta \geq 0$ is introduced into Eq. 1:

$$\max \mathcal{L}_\beta = -\beta \mathrm{KL} - \mathrm{Rec}. \tag{3}$$

When $\beta = 1$, the objective targets a single point on the RD curve with slope 1 because Rec (y-axis) can be expressed as a linear function $\mathrm{Rec} = -\mathcal{L} - \mathrm{KL}$. Setting $\beta < 1$ leads the situation where Rec is low but KL is high while $\beta > 1$ leads the opposite case (high Rec and low KL), all without having to change the model architecture.

Next, we summarize the relationship between the RD curve and VC performance. (i) First, low Rec is a necessary condition for high naturalness. This is because high Rec disables the decoder to construct natural speech, and degrades the naturalness of the converted speech sampled from the decoder. (ii) Second, low KL is a necessary condition for high similarity. This is because high KL suggests that z contains much information about x, which contradicts the intention of obtaining speaker-invariant latent variables. On the other hand, low KL indicates that z is well-compressed so redundant speaker information is lost from z (note that the speaker information is inherently redundant for z because the decoder is conditioned on y). Based on (i) and (ii), we conclude that Rec and KL affect naturalness and similarity, respectively. Note that, however, low KL and low Rec are necessary conditions but not sufficient on their own: to obtain high naturalness and similarity both KL and Rec need to be low. For example, even if Rec is low, the naturalness can be low when KL is so large that the source and target speaker information is entangled in the inferred latent variables.

Based on the analysis, we can take two approaches to getting good naturalness and similarity. One is using a hyperparameter β in Eq. 3 to achieve the best combination of KL and Rec for a given model (note that there is no guarantee that $\beta = 1$ achieves the best combination for VC). The other is using a VAE with higher model expressivess to expand the realizable combinations, i.e., to move the RD curve to the left.

B. Conditional Deep Hierarchical VAE

Based on the analysis in Section III-A, we propose using a VAE with high model expressiveness, as well as using the β-VAE objective. Namely, this paper adapts DHVAEs (see, Section II-C) to VC tasks because they have high model expressiveness thanks to the hierarchical architecture. However, they cannot be naively applied to VC tasks because they are unconditional models and have hierarchical latent variables. Then, we propose CDHVAE, which separates speaker-dependent and invariant latent variables to perform VC.

First, we present a data generating process of CDHVAE, where the conditioning with speaker label y is added to DHVAEs. Namely, as well as DHVAE, CDHVAE has encoder, decoder, and prior. However, the three components are conditioned with y as follows:

$$\begin{aligned}
\text{Encoder} &= q(z_l|x, z_{<l}, y) \ \ \forall l \leq L, \\
\text{Decoder} &= p(x|z, y), \\
\text{Prior} &= \begin{cases} p(z_l|z_{<l}) & \text{if } l \leq K, \\ p(z_l|z_{<l}, y) & \text{else.} \end{cases}
\end{aligned}$$

The data generating process is illustrated in Figure 2-(a, b). Here, $p(z_{\leq K})$ and $p(z_{>K}, y)$ are the speaker-invariant and dependent latent variables, respectively. Also, K is a hyperparameter that controls the expressiveness of the speaker-invariant and dependent latent variables. It is because $p(z_{\leq K})$ becomes more expressive when K becomes large, and vice versa ($p(z_{>K}, y)$ becomes more expressive when K becomes small). More specifically, as K increases, the hierarchy of $p(z_{\leq K})$ increases, so the model expressiveness of $p(z_{\leq K})$ increases. Then, a significantly small K makes it difficult for speaker-invariant variables $z_{\leq K}$ to model linguistic information, which contradicts the intention of VAE-VC. Also, a significantly large K makes it difficult for speaker-dependent variables $z_{>K}$ to model speeches, which can degrades output speech quality.

Using the components, the objective of CDHVAE is given as follows:

$$\begin{aligned}\max \mathcal{J}_\beta = &-\beta \sum_{l=1}^{K} \mathbb{E}_{q(z_{<l}|x,y)} D_{KL}[q(z_l|x, z_{<l}, y)||p(z_l|z_{<l})] \\ &-\beta \sum_{l=K+1}^{L} \mathbb{E}_{q(z_{<l}|x,y)} D_{KL}[q(z_l|x, z_{<l}, y)||p(z_l|z_{<l}, y)] \\ &+ \mathbb{E}_{q(z|x)}[\log p(x|z)]. \qquad (4)\end{aligned}$$

Here, we introduce a weighting parameter β to balance the KL and reconstruction terms, as was done in Eq. 3.

Using the trained CDHVAE, VC can be performed with the following procedure. First, the speaker-invariant latent variable $z_{\leq K}$ is obtained from the encoder $q(z_{\leq K}|x_s, y_s)$, given x_s and y_s. Then, the converted speech $x_{s\to t}$ can be obtained by inputting $z_{\leq K}$ and the target speaker label y_t to the prior and the decoder. This data generating process is illustrated in Figure 2-(c), and can be expressed as follows:

$$x_{s\to t} \sim \mathbb{E}_{q(z_{\leq K}|x_s,y_s)} \mathbb{E}_{\Pi_{l=K+1}^{L} p(z_l|z_{<l},y_t)} p(x|z, y_t). \qquad (5)$$

In practice, we approximate expectations with the mean of $q(z_{\leq K}|x_s, y_s)$ and $\Pi_{l=K+1}^{L} p(z_l|z_{<l}, y_t)$. Also, we split an utterance x_s into segments with a sequence length $T = 40$ and perform segment-wise conversion to adopt the same settings as training, but in principle it is possible to perform utterance-wise conversion.

The remaining challenge is what neural network architecture should be used to parameterize the data generating process of CDHVAE and to incorporate speaker labels. We choose to adapt the model architecture of Nouveau VAE (NVAE) [19] for CDHVAE, which is mainly composed of depthwise convolutions [23], [24] and recently achieved state-of-the-art results in image generation. However, because the original NVAE is an unconditional model, we add a simple modification for enabling conditioning: replacing all Batch Normalization (BN) layers with Conditional Instance Normalization (CIN) [25], [26]. Specifically, the original NVAE have BN layers in components called residual cells which correspond to r in Figure 2. Then, we obtain speaker embedding from one-hot label y using linear transformation, and use them as scale and location parameters in CIN. Also, since NVAE has been proposed in the literature on image generation, the input should be (C, H, W)-dimensional tensor, where C is the channel size, H is the image height, and W is the image width. Therefore, we set $C = 1$, $H = 80$, and $W = T = 40$ such that the mel-spectrogram can be used instead of an image.

IV. Related Works

There are two VC approaches that utilize pure unparallel corpus and require neither parallel corpus nor text transcription: generative adversarial network (GAN) [27], [28] and autoencoder-based methods including VAE-VC. VAE-VC was initially proposed by [10]. Subsequent research proposed auxiliary tasks to improve the naturalness [12] and similarity [13], [14] of the converted speech. Also, some studies utilized vanilla autoencoder with some techniques (adversarial training [29] or architectural constraint [30]) instead of VAE. These methods are common to VAE-VC in that they attempted to obtain speaker-invariant latent variables; therefore, these methods also may benefit from increasing model expressiveness as discussed in Section III-A. In addition, similar to our method, some VAE-VC methods [31], [32] adopt U-Net-like architectures [33] and create hierarchical latent embeddings. However, our method is different in that it adopts a very large hierarchy (e.g., $L = 35$ in our experiments); moreover, we show that the performance of VC significantly benefits from increasing model expressiveness, even without vector quantization which they employed.

From a technical perspective, our study adapts DHVAEs for voice conversion. In the literature on image generation, DHVAEs recently achieved competitive performance compared to other deep generative models (e.g., GANs and autoregressive models) [19], [20]. While these DHVAEs are very recently applied to text-to-speech (TTS) [21], [34], they have not been applied to VC so far. In addition, applying DHVAEs to VC requires unique techniques such as performing inference to speaker-invariant latent variables and controlling the balance of rate and distortion.

V. Experiment

A. Settings

In this section, we will evaluate CDHVAE on many-to-many VC tasks. We strongly encourage readers to listen to the demos that can be found in https://dena.github.io/CDHVAE/. The evaluation is performed on the VCTK corpus [35]. While the original corpus includes utterances from 109 speakers, we trained the models on 20 speakers (10 females and 10 males), as performed in [29], [30]. The acoustic features were mel-spectrograms extracted from 48kHz audio. As the vocoder, we used MelGAN [36], which was trained on the same corpus.

CDHVAE is trained with a batch size of 8 for 200 epochs. Regarding the model architecture, we adapt *CelebA model* in Table 6 of [19] with two modifications: replacing BN with CIN for conditioning, and removing normalizing flows for faster training. The other training settings are the same with the official NVAE implementation [1]. In the CelebA model, the number of latent variables is set to $L = 35$. Also, based on our informal preliminary experiments, the hyperparameter in Eq. 4 is set to $K = 10$. While this paper does not include hyperparameter study regarding K due to computational resource limitation, we recommend not to use very small values nor very large values since K balances the expressiveness of the speaker-invariant and speaker-dependent latent variables. CDHVAE is compared to the recent autoencoder-based method proposed by [29], whose pretrained-model is publicly available [2]. Also, CDHVAE was trained with various values of $\beta \in \{1, 10, 50\}$. Here, because we observed that CDHVAE with $\beta = 1$ resulted in high naturalness but low similarity in our preliminary experiments, we did not test $\beta < 1$.

[1] https://github.com/NVlabs/NVAE

[2] https://github.com/jjery2243542/voice_conversion

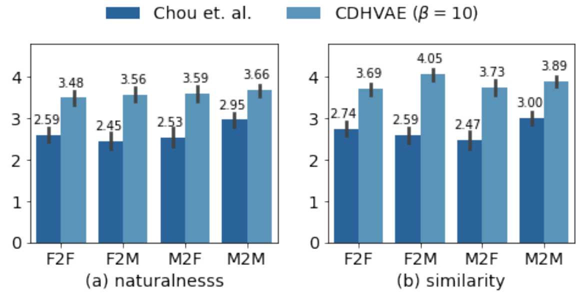

Fig. 3. Comparing the baseline and CDHVAE using MOS. The error bars show 95% confidence intervals.

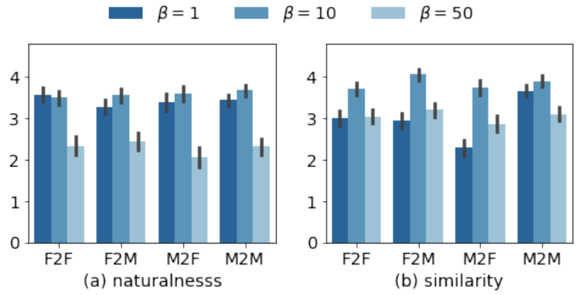

Fig. 4. Hyperparameter study on β using MOS. The error bars show 95% confidence intervals.

We conducted naturalness and similarity tests using MOSs in the same manner as in previous studies [30], [11]. We first selected one sentence for each of randomly selected 10 speakers (5 males and 5 females), and converted it to one randomly selected male speaker and one randomly selected female speaker, resulting in 20 utterances composed of 5 male-to-male (M2M), 5 male-to-female (M2F), 5 F2M, and 5 F2F samples. Then, subjects in Amazon Mechanical Turk were asked to evaluate 80 utterances (20-utterance times 4-method). After applying post-screening [37], the answers for 29 and 45 subjects, who rated more than 70 utterances, were collected for naturalness and similarity tests, respectively.

B. Results

From the MOS results, we can make the following observations. **(i) Using the VAE with high model expressiveness improved naturalness and similarity:** the results in Figure 3 shows that CDHVAE with $\beta = 10$ achieved higher performance than the baseline method [29]. Moreover, it achieved MOSs higher than 3.5 on both naturalness and similarity in inter-gender settings. To give readers a better idea of what this means, notice that none of the existing autoencoder-based VC methods (e.g., AutoVC [30] nor the methods in VCC 2020) reached 3.5 on naturalness and similarity at the same time. Even though these are not fair comparisons due to the difference in experimental settings (e.g., training data or Hz of speeches), it is encouraging to see that even with the simple β-VAE objective, the increasing expressiveness of VAEs enabled competitive or higher quality compared to the existing autoencoder-based VC methods.

(ii) β is an important factor for balancing the tradeoff between naturalness and similarity: the results in Figure 4 shows that CDHVAE with low β value ($\beta = 1$) achieved lower similarity scores than those with $\beta = 50$ in inter-gender settings. This indicates that a small KL value is a necessary condition for good similarity. On the other hand, CDHVAE with high β value ($\beta = 50$) achieved the lowest naturalness scores because the very large β makes reconstruction difficult. This indicates that a small Rec value is a necessary condition for good naturalness. Here, note that small KL and Rec values are merely the necessary conditions as noted in Section III-A, and properly selected $\beta = 10$ achieved the highest MOSs for both naturalness and similarity.

Conversion speed: On a 16-GB Tesla T4 GPU, we can convert a segment of mel-spectrogram of the size 80 × 40 in 0.172 seconds (344 ms / 1-second speech).

VI. Discussions

In the experiment using VCTK corpus, the proposed model called CDHVAE achieved MOSs higher than 3.5 on both naturalness and similarity in inter-gender settings. While we used simple β-VAE objective for CDHVAE, the performance could be improved by combining with the existing auxiliary losses for VAE-VC. Other future work may be to train CDHVAE with expressive or noisy speech corpora because the possible merit of VAE-VC is not requiring a pre-trained speech synthesis model or mapping from PPG to speech.

VII. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20J11448.

References

[1] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

[2] Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, Vol. 49, No. 9, pp. 743–759, 2007.

[3] Hironori Doi, Tomoki Toda, Tomoyasu Nakano, Masataka Goto, and Satoshi Nakamura. Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–6. IEEE, 2012.

[4] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[5] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6805–6809, 2019.

[6] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016.

[7] Li-Juan Liu, Yan-Nian Chen, Jing-Xuan Zhang, Yuan Jiang, Ya-Jun Hu, Zhen-Hua Ling, and Li-Rong Dai. Non-Parallel Voice Conversion with Autoregressive Conversion Model and Duration Adjustment. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 126–130, 2020.

[8] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*, 2018.

[9] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5901–5905. IEEE, 2019.

[10] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6. IEEE, 2016.

[11] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98, 2020.

[12] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. In *Proc. Interspeech 2017*, pp. 3364–3368, 2017.

[13] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder. *arXiv preprint arXiv:1808.05092*, 2018.

[14] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. Non-parallel voice conversion with cyclic variational autoencoder. *Proc. Interspeech 2019*, pp. 674–678, 2019.

[15] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6309–6318, 2017.

[16] Da-Yi Wu and Hung-yi Lee. One-shot voice conversion by vector quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7734–7738. IEEE, 2020.

[17] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc., 2016.

[18] Lars Maaløe, Marco Fraccaro, Valentin Lievin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. In *33rd Conference on Neural Information Processing Systems*, p. 8882. Neural Information Processing Systems Foundation, 2019.

[19] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 19667–19679. Curran Associates, Inc., 2020.

[20] Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

[21] Yoonhyung Lee, Joongbo Shin, and Kyomin Jung. Bidirectional variational inference for non-autoregressive text-to-speech. In *International Conference on Learning Representations*, 2021.

[22] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

[23] Vincent Vanhoucke. Learning visual representations at scale. *ICLR invited talk*, Vol. 1, p. 2, 2014.

[24] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

[26] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

[27] T. Kaneko and H. Kameoka. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104, 2018.

[28] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273. IEEE, 2018.

[29] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. *Proc. Interspeech 2018*, pp. 501–505, 2018.

[30] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pp. 5210–5219. PMLR, 2019.

[31] Tuan Vu Ho and Masato Akagi. Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder. 2020.

[32] Da-Yi Wu, Yen-Hao Chen, and Hung-Yi Lee. Vqvc+: One-shot voice conversion by vector quantization and u-net architecture. *arXiv preprint arXiv:2006.04154*, 2020.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[34] Peng Liu, Yuewen Cao, Songxiang Liu, Na Hu, Guangzhi Li, Chao Weng, and Dan Su. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. *arXiv preprint arXiv:2102.06431*, 2021.

[35] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.

[36] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 2019.

[37] Flavio Protasio Ribeiro, Dinei Florencio, Cha Zhang, and Mike Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *ICASSP*. IEEE, May 2011.