The Speaker and Language Recognition Workshop (Odyssey 2020)
1-5 November 2020, Tokyo, Japan

MANY-TO-MANY VOICE CONVERSION USING CYCLE-CONSISTENT VARIATIONAL AUTOENCODER WITH MULTIPLE DECODERS

Dongsuk Yook, Seong-Gyun Leem, Keonnyeong Lee, and In-Chul Yoo

Artificial Intelligence Laboratory, Department of Computer Science and Engineering,
Korea University, Republic of Korea

yook@korea.ac.kr, {sgleem, gnl0813, icyoo}@ai.korea.ac.kr

ABSTRACT

One of the obstacles in many-to-many voice conversion is the requirement of the parallel training data, which contain pairs of utterances with the same linguistic content spoken by different speakers. Since collecting such parallel data is a highly expensive task, many works attempted to use non-parallel training data for many-to-many voice conversion. One of such approaches is using the variational autoencoder (VAE). Though it can handle many-to-many voice conversion without the parallel training, the VAE based voice conversion methods suffer from low sound qualities of the converted speech. One of the major reasons is because the VAE learns only the self-reconstruction path. The conversion path is not trained at all. In this paper, we propose a cycle consistency loss for the VAE to explicitly learn the conversion path. In addition, we propose to use multiple decoders to further improve the sound qualities of the conventional VAE based voice conversion methods. The effectiveness of the proposed method is validated using objective and the subjective evaluations.

1. INTRODUCTION

Voice conversion (VC) is a task of converting the speaker-related voice characteristics in an utterance while maintaining the linguistic information. Conventional VC methods require parallel speech data for the model training. The parallel speech data contain pairs of utterances that have the same linguistic contents spoken by different speakers. However, such parallel speech data are highly expensive that they restrict the use of VC in many applications. Therefore, many recent VC approaches attempted to use non-parallel training data. Early works using non-parallel training data adopt Gaussian mixture models (GMM) [1, 2, 3]. Recently, deep learning based VC approaches that have shown promising results use cycle-consistent adversarial networks (CycleGAN) [4, 5, 6, 7, 8, 9], variational autoencoders (VAE) [10, 11, 12, 13], and VAE with generative adversarial networks (GAN) [14, 15].

In the CycleGAN [16] based VC approaches, the speech features of a source speaker are converted to match the characteristics of a target speaker using a GAN [17], and the converted speech features are again converted back through another GAN to match the original speech features from the source speaker. By using the cycle-consistency loss [18], the linguistic contents are forced to be retained in the converted speech. However, the CycleGAN can learn only one-to-one mapping between two speakers. To achieve complete mapping among n speakers, $n(n-1)/2$ CycleGAN models must be trained separately, which increases the training time and the memory space prohibitively. Though the extensions of the CycleGAN for many-to-many VC have been proposed [6, 7, 8, 9], they do not scale well as the number of speakers increases. For example, the number of speakers used in the experiments [6, 7, 8] were at most 4.

The VAE based VC approaches, on the other hand, can perform many-to-many VC for hundreds of speakers using non-parallel training data. A VAE [19] is composed of an encoder and a decoder. In the VC task, the encoder transforms the input speech features into the latent vectors containing the linguistic information of the input speech. Then, the latent vectors together with a target speaker identity vector are fed into the decoder to generate the converted speech features of the target speaker. Since the decoder is conditioned on a target speaker identity vector, it is sometimes called the conditional VAE.

Though the VAE models can be trained quickly, the sound qualities of the converted speech are usually low. To improve the sound quality, a VAE and Wasserstein generative adversarial network (WGAN) [20] hybrid called variational autoencoding Wasserstein generative adversarial networks (VAEWGAN) [14] was proposed. In this method, the decoder of the VAE is considered as the generator of the WGAN in order to train the decoder better. Though VAEWGAN based VC reduces some muffled sound, the qualities of the converted speech are still unsatisfactory.

One of the major drawbacks of the VAE based VC approaches is that the VAE models are not explicitly trained to convert the speech from a source speaker to a target speaker. Rather, they are trained to recover the same input speech from the source speaker using the latent vectors and the source speaker identity vector. In this paper, we propose to utilize a cycle consistency loss for the VAE to explicitly learn the mapping from a source speaker to a target speaker. To improve the sound quality further, we also propose a multi-decoder VAE which has a separate decoder for each target speaker. The cycle consistency loss and the multiple decoders can be incorporated into the VAEWGAN as well [21].

The rest of the paper is organized as follows. In Section 2, we describe the proposed methods in detail. Section 3 analyzes the experimental results, and Section 4 concludes the paper.

2. CYCLE-CONSISTENT VAE AND VAEWGAN

2.1. Variational Autoencoder

The loss function of the VAE is defined as follows:

$$\mathcal{L}_{\mathrm{VAE}}(\phi,\theta;x,X)=\mathbb{D}_{\mathrm{KL}}\left(q_\phi(z|x)\parallel p(z)\right)-$$

10.21437/Odyssey.2020-31

$$\mathbb{E}_{z\sim q_{\phi}(z|x)}[\log p_{\theta}(x|z,X)]\,, \qquad (1)$$

where $\mathbb{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence, q_{ϕ} is an encoding model with parameter ϕ that infers the linguistic information of input speech x, $p(z)$ is a prior distribution for latent vector z, and p_{θ} is a decoding model with parameter θ that generates the reconstructed speech using z and source speaker identity vector X which is typically represented as a one-hot vector.

By minimizing equation (1), the VAE is trained to reconstruct the input speech from the latent vector z and the source speaker identity vector X. To convert the speech from a source speaker to a target speaker, the source speaker identity vector X is replaced with the target speaker identity vector Y. Due to the absence of explicit model training for the conversion between the source speaker and the target speaker (i.e., only self-reconstruction training), the VAE based VC methods generally produce the converted speech with low sound quality.

2.2. Variational Autoencoder with Wasserstein Generative Adversarial Network

The VAEWGAN has been proposed to improve the sound quality of the VAE based VC method. In this approach, the decoder of the VAE is the generator of the WGAN. The loss function of the WGAN is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{WGAN}}(\theta,\psi;\phi,x,Y) = \mathbb{E}_{y|Y}\big[D_{\psi}(y)\big] - \\ \mathbb{E}_{z\sim q_{\phi}(z|x)}\Big[D_{\psi}\big(G_{\theta}(z,Y)\big)\Big]\,, \qquad (2)\end{aligned}$$

where G_{θ} is a generator with parameter θ, D_{ψ} is a discriminator with parameter ψ, and y is the speech from the target speaker represented by speaker identity vector Y. Since the decoder of the VAE is the generator of the WGAN, G_{θ} is p_{θ}.

Now, the loss function of the VAEWGAN that converts speaker X's voice to speaker Y's voice is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{VAEWGAN}}(\phi,\theta,\psi;x,y,X,Y) = \\ \mathcal{L}_{\mathrm{VAE}}(\phi,\theta;x,X) + \\ \mathcal{L}_{\mathrm{VAE}}(\phi,\theta;y,Y) + \\ \lambda_1\mathcal{L}_{\mathrm{WGAN}}(\theta,\psi;\phi,x,Y)\,, \qquad (3)\end{aligned}$$

where λ_1 is the weight of the WGAN loss. Equation (3) is minimized for the VAE and the generator, and it is maximized for the discriminator. First, the VAE is trained in the same way as in Section 2.1. Second, the VAE and the WGAN are jointly trained such that the VAE gets an additional error signal from the discriminator of the WGAN.

Though the VAEWGAN produces somewhat higher sound quality than the VAE, it can handle only one-to-one voice conversion. In the next sections, we propose the extensions of the VAE and the VAEWGAN, called cycle-consistent VAE (CycleVAE) and cycle-consistent VAEWGAN (CycleVAEWGAN), respectively, which can improve the performance for many-to-many voice conversion by using multiple decoders and explicitly learning many-to-many mapping functions.

2.3. Cycle-Consistent Variational Autoencoder (CycleVAE)

In order to improve the sound quality of the VAE based VC, we propose to use a separate decoder for each speaker instead of a single decoder for all speakers. We also propose to use the cycle consistency loss for explicit conversion path training. The speaker identity vectors are not needed for the multi-decoder VAE since each speaker has an independent decoder. It can be expected that the sound quality can be improved since each decoder learns its corresponding speaker's voice characteristics by the additional conversion path training while the conventional VAE must handle multiple speakers with only a single decoder by self-reconstruction training.

Fig. 1 shows the concept of the CycleVAE for two speakers. When the speech x from speaker X is fed into the network, it passes through the encoder and is compressed into the latent vector z. The reconstruction error is computed using the reconstructed speech $x'_{X\to X}$ by the speaker X decoder model p_{θ_X}. Up to this point, the loss function is similar to the vanilla VAE except that it does not require the speaker identity vectors, which is as follows:

$$\begin{aligned}\mathcal{L}'_{\mathrm{VAE}}(\phi,\theta;x,X) = \mathbb{D}_{\mathrm{KL}}\Big(q_{\phi}(z|x) \parallel p(z)\Big) - \\ \mathbb{E}_{z\sim q_{\phi}(z|x)}\big[\log p_{\theta_X}(x|z)\big]\,. \qquad (4)\end{aligned}$$

The same input speech x from speaker X goes through the encoder and the speaker Y decoder model p_{θ_Y} as well to generate the converted speech $x'_{X\to Y}$ which has the same linguistic contents as x but in speaker Y's voice. Then, the converted speech $x'_{X\to Y}$ goes through the encoder and the speaker X decoder model p_{θ_X} to generate the converted back speech $x''_{X\to Y\to X}$ which should recover the original speech x. This cyclic conversion encourages the explicit training of voice conversion path from Y to X without parallel data. Similarly, y and $y'_{Y\to X}$ are used to train the conversion path from X to Y explicitly. The cycle consistency loss of the multi-decoder VAE

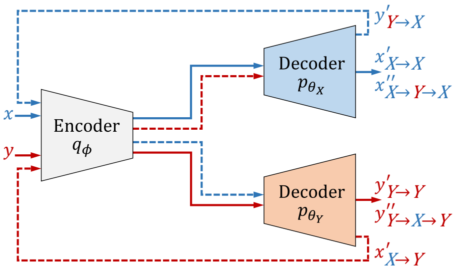

Figure 1. *The multi-decoder CycleVAE for two speakers. A separate decoder is used for each speaker. The solid lines indicate the usual self-reconstruction paths. The dashed lines correspond to the proposed explicit conversion paths. The multi-decoder CycleVAE learns both paths. The same color lines represent the same speaker's voice.* $x'_{X\to Y}$ *is used to train the conversion path from speaker* Y *to speaker* X*, while* $y'_{Y\to X}$ *is used to train the conversion path from speaker* X *to speaker* Y*. See Fig. 3 for the details of the encoder and the decoder.*

for two speakers given the input speech x from speaker X is defined as follows:

$$\mathcal{L}_{\text{Cycle}}(\phi,\theta;x,X,Y)=\mathbb{D}_{\text{KL}}\left(q_{\phi}(z|x'_{X\to Y})\parallel p(z)\right)-\mathbb{E}_{z\sim q_{\phi}(z|x'_{X\to Y})}\left[\log p_{\theta_X}(x|z)\right]. \quad (5)$$

Now, given the input speech x from speaker X, the loss function of the CycleVAE for two speakers can be defined as follows:

$$\mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,X,Y)=\mathcal{L}'_{\text{VAE}}(\phi,\theta;x,X)+\lambda_2\mathcal{L}_{\text{Cycle}}(\phi,\theta;x,X,Y)\,, \quad (6)$$

where λ_2 is the weight of the cycle consistency loss.

It can be easily extended for more than two speakers by summing over all pairs of the training speakers. The loss function of the CycleVAE for more than two speakers can be computed as follows:

$$\textstyle\sum_{X,Y}\sum_{x|X}\mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,X,Y)\,, \quad (7)$$

where the second summation is usually over a mini-batch.

2.4. Cycle-Consistent Variational Autoencoder with Wasserstein Generative Adversarial Network (CycleVAEWGAN)

The CycleVAE can be extended to utilize the WGAN as in the VAEWGAN case. In the CycleVAEWGAN, the decoders of the CycleVAE are shared with the generators of the WGANs. Each decoder has its own WGAN. Fig. 2 shows the concept of the CycleVAEWGAN for two speakers. Since there are multiple WGANs, equation (2) is modified as follows:

$$\mathcal{L}'_{\text{WGAN}}(\theta,\psi;\phi,x,Y)=\mathbb{E}_{y|Y}\left[D_{\psi_Y}(y)\right]-\mathbb{E}_{z\sim q_{\phi}(z|x)}\left[D_{\psi_Y}\left(G_{\theta_Y}(z)\right)\right], \quad (8)$$

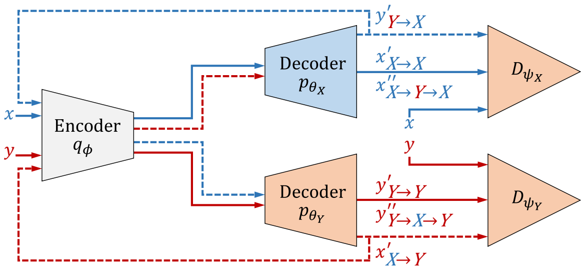

Figure 2. *The multi-decoder CycleVAEWGAN for two speakers. D_{ψ} represents a discriminator with parameter ψ. Each speaker has one's own decoder and discriminator. The solid lines indicate the usual self-reconstruction paths. The dashed lines correspond to the proposed explicit conversion paths. The multi-decoder CycleVAEWGAN learns both paths. The same color lines represent the same speaker's voice. See Fig. 3 for the details of the encoder, the decoder, and the discriminator.*

where G_{θ_Y} is a generator with parameter θ_Y for speaker Y, D_{ψ_Y} is a discriminator with parameter ψ_Y for speaker Y, and y is the speech from target speaker Y. Since the decoders of the CycleVAE are the generators of the WGANs, G_{θ_Y} is p_{θ_Y}.

Now, given the input speech x from speaker X, the loss function of the CycleVAEWGAN for two speakers is defined as follows:

$$\mathcal{L}_{\text{CycleVAEWGAN}}(\phi,\theta,\psi;x,X,Y)=\mathcal{L}_{\text{CycleVAE}}(\phi,\theta;x,X,Y)+\lambda_1\mathcal{L}'_{\text{WGAN}}(\theta,\psi;\phi,x,X)+\lambda_1\mathcal{L}'_{\text{WGAN}}(\theta,\psi;\phi,x'_{X\to Y},X)\,. \quad (9)$$

Note that $\mathcal{L}'_{\text{WGAN}}$ is used twice in the equation, i.e., one for the self-reconstruction path and the other for the conversion path. Equation (9) is minimized for the CycleVAE and the generators, and is maximized for the discriminators. The first stage of the CycleVAEWGAN training is identical to the training procedure of the CycleVAE. In the second stage of the training, the CycleVAE and the WGANs are jointly optimized where the CycleVAE receives the additional error signals from the WGANs. While the VAEWGAN performs only one-to-one voice conversion from a source speaker to a target speaker, the multi-decoder CycleVAEWGAN can perform many-to-many voice conversion among multiple speakers since each decoder has its own WGAN.

It also can be easily extended to more than two speakers by summing over all pairs of the training speakers. The loss function of the CycleVAEWGAN for more than two speakers can be computed as follows:

$$\textstyle\sum_{X,Y}\sum_{x|X}\mathcal{L}_{\text{CycleVAEWGAN}}(\phi,\theta,\psi;x,X,Y)\,, \quad (10)$$

where the second summation is usually over a mini-batch. The proposed CycleVAEWGAN is different from [13] in that it can utilize multiple decoders and WGANs.

3. EXPERIMENTS

We compared the performance of the voice conversion methods explained in the previous section. For the experiments, 2 male speakers and 2 female speakers, namely SF1, SF2, TM1 and TM2, from VCC2018 corpus [22] were used. The numbers of the training, validation, and the testing utterances per speaker were 72, 9, and 35, respectively. 36-dimensional Mel-cepstral coefficients (MCCs), aperiodicities (AP), and fundamental frequency (F0) were extracted every 5 ms from the speech waveforms of which sampling rate was 22.05 kHz. In all methods, the decoders converted the MCCs only. The F0's were converted by the logarithm Gaussian normalized transformation [23], and the APs were used without any modification to synthesize the waveforms of the converted speech using the WORLD vocoder [24].

We designed the architecture of our models based on [12]. Fig. 3 shows the details of the encoder, the decoder, and the discriminator used in the experiments. All models used the gated linear units (GLU) [25]. The batch normalization [26] was applied to each convolutional neural network (CNN) [27] layers. We used the Adam optimizer [28] with a mini-batch size of 16 randomly selected 128-frame segments to train the models. Each mini-batch from a speaker was used to train all conversion paths as well as the self-reconstruction path. One epoch of training consisted of 4 mini-batches which correspond

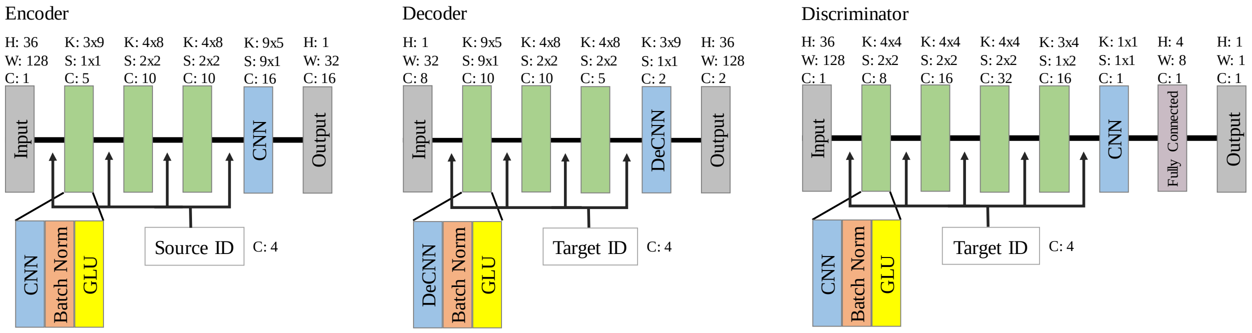

Figure 3. *The architectures of the encoder, the decoder, and the discriminator used in the experiments. Since we assumed Gaussian distributions with diagonal covariance matrices for the encoder and the decoder, the outputs of the encoder and the decoder are pairs of mean and variance. The target speaker identity vectors (Target ID) are not used for the multi-decoder CycleVAE and CycleVAEWGAN.*

Table 1. *λ_1 and λ_2 values used for the experiments.*

	λ_1	λ_2
VAEWGAN	1.0	0.0
CycleVAE	0.0	1.0
CycleVAEWGAN	0.5	1.0

to 4 times 3 conversion path training and 4 self-reconstruction path training. The learning rate was set to 0.0008, and the number of epochs was set to 500. Each model was bootstrapped from a VAE model that was trained for 500 epochs. Table 1 shows the λ_1 and λ_2 values used for the experiments. All experiments were repeated 5 times starting with randomly initialized weights. Table 2 summarizes the memory requirement and training time on a GeForce RTX 2080 GPU machine for each method.

3.1. Objective Evaluations

One of drawbacks of the VAE based approaches is the over-smoothing of the generated data [2]. The global variance (GV) of MCCs can be used to measure the degree of over-smoothing as the high GV values correlate with the sharpness of the spectra. We computed the GV for each of the MCC indices. Fig. 4 shows the average GV over all evaluation utterances for the real speech and the converted speech by the conventional VAE and the proposed CycleVAE. The average GVs over all indices and all evaluation utterances were 0.082, 0.065, and 0.066 for the real speech and the converted speech by the VAE and CycleVAE, respectively.

For the case of the original and the converted speech utterances containing the same linguistic information, the difference between the MCCs of the two speech utterances should be small. We used two metrics to measure this difference, i.e., the Mel-cepstral distortion (MCD) [2] and the modulation spectral distance (MSD) [29]. Tables 3 and 4 show MCD and MSD of 420 utterances (35 testing utterances for each of the 12 conversion directions), respectively, for various VC methods. Firstly, by comparing the baseline VAE and VAEWGAN columns in the tables, we confirmed that the VAEWGAN outperforms the VAE [14]. Secondly, to measure the effectiveness of the cycle consistency loss alone, we built the CycleVAE and the CycleVAEWGAN that use a single common decoder (i.e., p_{θ_X} and p_{θ_Y} are shared in Fig. 1). Since there is only one WGAN in the single-decoder

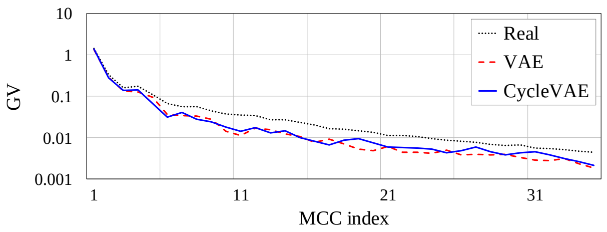

Figure 4. *The average global variances of MCCs for real speech utterances and the converted utterances by the VAE and the CycleVAE.*

Table 2. *Time complexity (average training time per epoch in seconds) and space complexity (number of model parameters).*

	VAE	VAEWGAN	CycleVAE (single-decoder)	CycleVAEWGAN (single-decoder)	CycleVAE (multi-decoder)	CycleVAEWGAN (multi-decoder)
Time	0.673	1.860	2.732	3.831	2.650	3.694
Space	51,194	77,510	51,194	77,510	95,204	194,476

Table 3. *Mean and 95% confidence interval of the MCDs for various VC methods.*

	VAE	VAEWGAN	CycleVAE (single-decoder)	CycleVAEWGAN (single-decoder)	CycleVAE (multi-decoder)	CycleVAEWGAN (multi-decoder)
F to F	7.315 ± 0.132	7.292 ± 0.127	7.261 ± 0.141	7.239 ± 0.115	**6.994** ± 0.123	7.126 ± 0.111
M to F	7.576 ± 0.095	7.499 ± 0.102	7.538 ± 0.096	7.439 ± 0.092	**7.204** ± 0.088	**7.204** ± 0.086
F to M	7.177 ± 0.070	7.268 ± 0.076	7.222 ± 0.074	7.203 ± 0.080	**6.972** ± 0.070	7.076 ± 0.074
M to M	7.108 ± 0.069	7.091 ± 0.074	7.114 ± 0.068	7.065 ± 0.069	**6.983** ± 0.074	6.999 ± 0.074
Average	7.335 ± 0.051	7.319 ± 0.052	7.334 ± 0.051	7.265 ± 0.048	**7.064** ± 0.046	7.126 ± 0.045

Table 4. *Mean and 95% confidence interval of the MSDs for various VC methods.*

	VAE	VAEWGAN	CycleVAE (single-decoder)	CycleVAEWGAN (single-decoder)	CycleVAE (multi-decoder)	CycleVAEWGAN (multi-decoder)
F to F	1.913 ± 0.012	1.928 ± 0.013	1.918 ± 0.013	1.919 ± 0.012	1.910 ± 0.012	**1.909** ± 0.012
M to F	1.899 ± 0.009	1.918 ± 0.010	1.902 ± 0.009	1.905 ± 0.009	**1.893** ± 0.008	1.898 ± 0.008
F to M	1.924 ± 0.009	1.933 ± 0.011	1.923 ± 0.009	1.922 ± 0.009	**1.913** ± 0.009	1.929 ± 0.010
M to M	1.915 ± 0.017	1.930 ± 0.018	1.918 ± 0.016	1.918 ± 0.016	**1.902** ± 0.017	1.918 ± 0.017
Average	1.912 ± 0.006	1.928 ± 0.006	1.914 ± 0.005	1.915 ± 0.005	**1.904** ± 0.005	1.913 ± 0.005

CycleVAEWGAN, the WGAN was modified to be conditioned on the source and the target speaker identity vectors as in the conditional CycleGAN (CC-GAN) [9]. The fourth and fifth columns of the tables shows these results. By comparing the second and the fourth columns (or the third and the fifth columns) of the tables, we confirmed the effectiveness of the cycle consistency loss [13, 21]. Finally, the results of the proposed multi-decoder approaches with the cycle consistency loss are shown in the last two columns of the tables. It can be seen that the multi-decoder approaches improved the performances further. It is interesting to note that unlike the VAE or the CycleVAE having a single common decoder, adding the WGANs to the CycleVAE with multiple decoders does not improve the performance further. It is suspected that since the multi-decoder cycle consistency loss is effective enough in learning the conversion path explicitly, the additional WGANs for conversion path learning may not be necessary.

3.2. Subjective Evaluations

We also conducted two subjective evaluations, i.e., naturalness test and similarity test. A set of 16 utterances was selected randomly such that 4 utterances were assigned to each pair of F to F, M to F, F to M, and M to M conversions, where F and M represent female and male, respectively. A total of 48 utterances (16 target speakers' utterances, 16 converted utterances by the conventional VAE, and 16 converted utterances by the proposed CycleVAE with multiple decoders) were played to 10 listeners participated in the subjective evaluations.

The mean opinion score (MOS) was used for the naturalness test. The listeners evaluated the naturalness of the speech in the scales of 1 (bad) to 5 (excellent) when the utterances were played in random order. Table 5 shows that the proposed CycleVAE based VC generally exhibits slightly higher naturalness scores than the conventional VAE based VC.

In the similarity test, a target speaker's utterance was played first, then a pair of two converted utterances by the two methods were played in random order. The listeners were asked to select the more similar utterance to the target speaker's speech or 'fair' if they could not tell the difference. Table 6 shows that the proposed CycleVAE based VC outperforms the conventional VAE based VC significantly.

4. CONCLUSION

In this paper, we proposed the new many-to-many voice conversion methods based on the VAE. The proposed methods

Table 5. *Sound quality test* (*MOS and 95% confidence interval*).

	VAE	CycleVAE	Target Voice
F to F	2.93 ± 0.38	2.70 ± 0.56	4.73 ± 0.40
M to F	2.33 ± 0.31	2.65 ± 0.42	
F to M	2.85 ± 0.43	2.93 ± 0.49	4.76 ± 0.42
M to M	3.08 ± 0.50	2.98 ± 0.54	
Average	2.79 ± 0.32	2.81 ± 0.34	4.74 ± 0.41

Table 6. *Similarity test* (*%*).

	VAE	Fair	CycleVAE
F to F	10.0	65.0	25.0
M to F	2.5	25.0	72.5
F to M	7.5	72.5	20.0
M to M	22.5	52.5	25.0
Average	10.6	53.8	35.6

use multiple decoders and explicitly learn the conversion path for many-to-many voice conversion without parallel training data. The effectiveness of the proposed methods was validated by the objective evaluations and the subjective evaluations using VCC2018 corpus.

We are currently running the experiments using a larger corpus consisting of more than 100 speakers to find out how the proposed methods scale for a larger number of speakers. The proposed methods can be further extended by utilizing multiple encoders, i.e., one encoder for each source speaker. Also, replacing the vocoder with powerful neural vocoders such as the WaveNet [30] or the WaveRNN [31] can be another future research direction.

5. ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2017R1E1A1A01078157). Also, it was partly supported by the MSIT (Ministry of Science and ICT) under the ITRC (Information Technology Research Center) support program (IITP-2018-0-01405) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and IITP grant funded by the Korean government (MSIT) (No. 2018-0-00269).

6. REFERENCES

1. Yannis Stylianou, Olivier Cappé, and Eric Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.
2. Tomoki Toda, Alan Black, and Keiichi 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.
3. Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, “Voice conversion using partial least squares regression,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 912–921, 2010.
4. Takuhiro Kaneko and Hirokazu Kameoka, “CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks,” European Signal Processing Conference, pp. 2114–2118, 2018.
5. Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, “CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 6820–6824, 2019.
6. Dongsuk Yook, In-Chul Yoo, and Seungho Yoo, “Voice conversion using conditional CycleGAN,” International Conference on Computational Science and Computational Intelligence, pp. 1460–1461, 2018.
7. Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks,” IEEE Spoken Language Technology Workshop, pp. 266–273, 2018.
8. Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, “StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion,” Interspeech, pp. 679–683, 2019.
9. Shindong Lee, BongGu Ko, Keonnyeong Lee, In-Chul Yoo, and Dongsuk Yook, “Many-to-many voice conversion using conditional cycle-consistent adversarial networks,” IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 6279–6283, 2020.
10. Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, “Voice conversion from non-parallel corpora using variational autoencoder,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–6, 2016.
11. Aaron van den Oord and Oriol Vinyals, “Neural discrete representation learning,” Neural Information Processing Systems, pp. 6309–6318, 2017.
12. Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, “ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 9, pp. 1432–1443, 2019.
13. Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda, “Non-parallel voice conversion with cyclic variational autoencoder,” Interspeech, pp. 674–678, 2019.
14. 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,” Interspeech, pp. 3364–3368, 2017.
15. 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,” Interspeech, pp. 501–505, 2018.
16. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” IEEE International Conference on Computer Vision, pp. 2223–2232, 2017.
17. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” Neural Information Processing Systems, pp. 2672–2680, 2014.
18. Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei Efros, “Learning dense correspondence via 3D-guided cycle consistency,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 117–126, 2016.
19. Diederik Kingma and Max Welling, “Auto-encoding variational Bayes,” arXiv:1312.6114, 2013.
20. Martin Arjovsky, Soumith Chintala, and Leon Bottou, “Wasserstein generative adversarial networks,” International Conference on Machine Learning, pp. 214–223, 2017.

21. Keonnyeong Lee, In-Chul Yoo, and Dongsuk Yook, “Voice conversion using cycle-consistent variational autoencoder,” arXiv:1909.06805, 2019.

22. Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” The Speaker and Language Recognition Workshop, pp. 195–202, 2018.

23. Kun Liu, Jianping Zhang, and Yonghong Yan, “High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin,” *International Conference on Fuzzy Systems and Knowledge Discovery*, 2017.

24. Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Transactions on Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.

25. Yann Dauphin, Angela Fan, Michael Auli, and David Grangier, “Language modeling with gated convolutional networks,” International Conference on Machine Learning, pp. 933–941, 2017.

26. Sergey Ioffe and Christian Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” International Conference on Machine Learning, pp. 448–456, 2015.

27. Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, “Imagenet classification with deep convolutional neural networks,” Neural Information Processing Systems, pp. 1097–1105, 2012.

28. Diederik Kingma and Jimmy Lei Ba, “Adam: A method for stochastic optimization,” International Conference on Learning Representations, 2015.

29. Shinnosuke Takamichi, Tomoki Toda, Alan Black, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura, “Postfilters to modify the modulation spectrum for statistical parametric speech synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 755–767, 2016.

30. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv:1609.03499, 2016.

31. Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, “Efficient neural audio synthesis,” arXiv:1802.08435, 2018.