Non-Parallel Voice Conversion with Cyclic Variational Autoencoder

Patrick Lumban Tobing[1], *Yi-Chiao Wu*[1], *Tomoki Hayashi*[1], *Kazuhiro Kobayashi*[2], *Tomoki Toda* [2]

[1]Graduate School of Information Science, Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan
{patrick.lumbantobing, yichiao.wu, hayashi.tomoki}@g.sp.m.is.nagoya-u.ac.jp,
kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

In this paper, we present a novel technique for a non-parallel voice conversion (VC) with the use of cyclic variational autoencoder (CycleVAE)-based spectral modeling. In a variational autoencoder (VAE) framework, a latent space, usually with a Gaussian prior, is used to encode a set of input features. In a VAE-based VC, the encoded latent features are fed into a decoder, along with speaker-coding features, to generate estimated spectra with either the original speaker identity (reconstructed) or another speaker identity (converted). Due to the non-parallel modeling condition, the converted spectra can not be directly optimized, which heavily degrades the performance of a VAE-based VC. In this work, to overcome this problem, we propose to use CycleVAE-based spectral model that indirectly optimizes the conversion flow by recycling the converted features back into the system to obtain corresponding cyclic reconstructed spectra that can be directly optimized. The cyclic flow can be continued by using the cyclic reconstructed features as input for the next cycle. The experimental results demonstrate the effectiveness of the proposed CycleVAE-based VC, which yields higher accuracy of converted spectra, generates latent features with higher correlation degree, and significantly improves the quality and conversion accuracy of the converted speech.

Index Terms: voice conversion, non-parallel, spectral modeling, variational autoencoder, cyclic mapping flow

1. Introduction

Using a voice conversion (VC) system, voice characteristics of a source speaker can be transformed into that of a desired target speaker, while keeping the linguistic contents intact. Such transformation can be achieved, for example, by performing statistical conversion of spectral envelope parameters of the vocal tract spectrum, and a proper alteration of excitation features, such as fundamental frequency (F_0). Within two decades, many speech applications have been realized by employing the VC framework, such as creation of speech database with various voice characteristics [1], singing voice conversion [2], recovery of impaired speech signal [3, 4], expressive speech synthesis [5, 6], body-conducted speech processing [7, 8], and articulatory controllable speech modification [9]. For flexible development of related applications, it is important to consider a VC technique that can be realized using easily available speech data.

There are two main VC frameworks, non-parallel VC and parallel VC. In the non-parallel VC, it is not straightforward to measure the correspondence between source spectral features and the target spectral features, due to the non-existence of paired utterances. On the other hand, in a parallel VC [10, 11], because of the availability of the paired utterances, their correspondence can be directly achieved by performing time-alignment, such as with dynamic-time-warping (DTW) algorithm. However, not all of the time a proper parallel dataset, i.e., where the source and the target speakers utter the same set of sentences, can be collected for the development of a VC system. Consequently, as our main focus in this work, a consideration for a reliable non-parallel VC using data-driven statistical modeling would be highly beneficial for real-life applications.

Indeed, the challenge in developing the non-parallel spectral conversion model has attracted many works within the recent years, such as: with the use of clustered spectral matching algorithms [12, 13]; with adaptation/alignment of speaker model parameters [14, 15]; with restricted Boltzmann machine [16]; with generative adversarial networks (GAN)-based methods [17, 18]; and with variational autoencoder (VAE)-based frameworks [19, 20, 21, 22]. In this work, we focus on the use of VAE-based system, due to its potential in employing latent space to represent common hidden aspects of speech signal, between different speakers, e.g., phonetical attributes. Further, its implementation can be flexibly realized through any network architectures, such as with convolutional or recurrent models.

In a VAE framework [23], a latent space, usually with a Gaussian prior, is used for encoding a set of input features. In a VAE-based VC [19], additional speaker-coding features are used, alongside the encoded latent features, to reconstruct the spectral features in the generation phase. Speaker-code associated with the source (original) speaker is used to estimate the reconstructed spectra, while speaker-code associated with a desired target speaker is used to estimate converted spectra. However, due to the non-parallel condition, the spectral model parameters are optimized with respect only to the reconstructed spectra. Hence, because of the only reliance in speaker-code capability to disentangle speaker identity, the performance of a conventional VAE-based VC is still insufficient.

In this paper, to improve VAE-based VC, we propose to use cycle-consistent mapping flow [24], i.e., CycleVAE-based VC, that indirectly optimizes the conversion flow by recycling the converted spectral features. Specifically, in the proposed CycleVAE, the converted features are fed-back into the system to generate corresponding cyclic reconstructed spectra that can be directly optimized. The cyclic flow can, then, be continued by feeding the cyclic reconstructed features back into the system. Therefore, the conversion flow, i.e., the estimation of converted spectra, is indirectly considered in the computation of both the reconstruction losses and the regularizations of latent space. In the experiments, it has been demonstrated that the proposed CycleVAE-based VC shows higher correlation degree of latent features, i.e., more similar latent attributes between different speakers (possibly within phonetical space), and higher accuracy of converted spectra. Perceptual evaluation also shows significant improvements in both quality and accuracy of converted speech, especially when the speaker identities are considerably distant, such as in cross-gender conversions.

2. Conventional VAE-based VC

The flow of conventional VAE-based VC is illustrated by the upper part of Fig. 1. Let $\boldsymbol{X}_t = [\boldsymbol{e}_t^{(x)\top}, \boldsymbol{s}_t^{(x)\top}]^\top$, $\boldsymbol{e}_t^{(x)} = [e_t^{(x)}(1), \ldots, e_t^{(x)}(D_e)]^\top$, and $\boldsymbol{s}_t^{(x)} = [s_t^{(x)}(1), \ldots, s_t^{(x)}(D_s)]^\top$ be the $D_e + D_s$, D_e, and D_s-dimensional feature vectors of the input, the excitation, and the spectra, respectively, at frame t. In the training phase, given a set of network parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, a sequence of input features $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ and time-invariant D_c-dimensional source speaker-code features $\boldsymbol{c}^{(x)}$ [19], a set of updated network parameters $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\}$ is estimated by maximizing the variational lower bound function [23] as follows:

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\} = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}), \tag{1}$$

where

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{X}_t)||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) + \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\boldsymbol{X}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)})], \tag{2}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{X}_t) = \mathcal{N}(\boldsymbol{z}_t; f_{\boldsymbol{\phi}}^{(\mu)}(\boldsymbol{X}_t), \operatorname{diag}(f_{\boldsymbol{\phi}}^{(\sigma)}(\boldsymbol{X}_t)^2)), \tag{3}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)}) \approx \mathcal{N}(\boldsymbol{s}_t^{(x)}; g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(x)}), \boldsymbol{I}), \tag{4}$$

$$\hat{\boldsymbol{z}}_t^{(x)} = f_{\boldsymbol{\phi}}^{(\mu)}(\boldsymbol{X}_t) + f_{\boldsymbol{\phi}}^{(\sigma)}(\boldsymbol{X}_t) \odot \boldsymbol{\epsilon} \ \text{ s. t. } \ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{5}$$

$\boldsymbol{z}_t$ denotes a D_z-dimensional latent feature vector, $f_{\boldsymbol{\phi}}(\cdot)$ denotes an encoder network, $g_{\boldsymbol{\theta}}(\cdot)$ denotes a decoder network, $\odot$ denotes an element-wise product, and $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is for a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Therefore, the reconstructed source spectra feature vector $\hat{\boldsymbol{s}}_t^{(x)}$, i.e., estimated spectra with the same speaker characteristics as the input source speaker, is given by

$$\hat{\boldsymbol{s}}_t^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(x)}). \tag{6}$$

On the other hand, the converted source-to-target spectra $\hat{\boldsymbol{s}}_t^{(y|x)}$, i.e., estimated spectra with the voice characteristics of a desired target speaker, is given by

$$\hat{\boldsymbol{s}}_t^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(y)}), \tag{7}$$

where $\boldsymbol{c}^{(y)}$ denotes the time-invariant D_c-dimensional target speaker-code features [19]. In this paper, we use not only source, but also target speakers as input in training. In order to use the corresponding target speaker as the input speaker, i.e., optimization of reconstructed target spectra and/or performing target-to-source conversion, the notations of x and y, in Eqs. (1)–(7), are swapped with each other. Though, the performance of VAE-based VC is noticeably insufficient because the conversion flow is not considered in the parameter optimization.

3. Proposed CycleVAE-based VC

In this paper, to improve the VAE-based VC, as illustrated in Fig. 1, we propose CycleVAE, which is capable of recycling the converted spectra back into the system, so that the conversion flow is indirectly considered in the parameter optimization. A similar idea has also been proposed as a cycle-consistent flow in a self-supervised method for visual correspondence [24].

In the proposed CycleVAE-based VC, the parameter optimization is defined as follows:

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}\} = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(y)}), \tag{8}$$

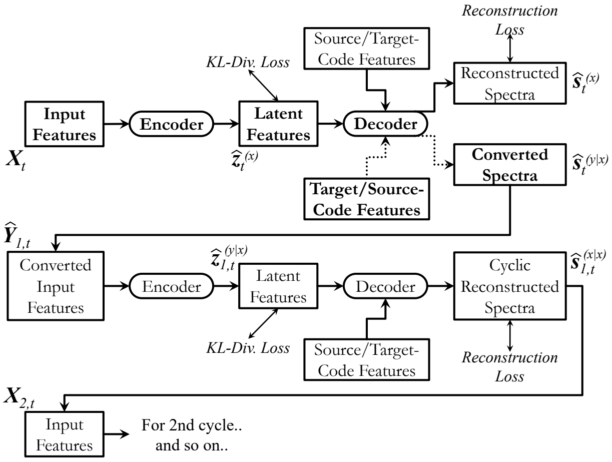
Reconstruction Loss
KL-Div. Loss
Source/Target-Code Features
Reconstructed Spectra
$\hat{\boldsymbol{s}}_t^{(x)}$
Input Features
Encoder
Latent Features
Decoder
$\boldsymbol{X}_t$
$\hat{\boldsymbol{z}}_t^{(x)}$
Converted Spectra
$\hat{\boldsymbol{s}}_t^{(y|x)}$
Target/Source-Code Features
$\hat{\boldsymbol{Y}}_{1,t}$
$\hat{\boldsymbol{z}}_{1,t}^{(y|x)}$
$\hat{\boldsymbol{s}}_{1,t}^{(x|x)}$
Converted Input Features
Encoder
Latent Features
Decoder
Cyclic Reconstructed Spectra
KL-Div. Loss
Source/Target-Code Features
Reconstruction Loss
$\boldsymbol{X}_{2,t}$
Input Features
For 2nd cycle.. and so on..

Figure 1: *Flow of the conventional VAE-based (upper-part) and the proposed CycleVAE-based (whole diagram) VC. Converted input features include converted excitation features, such as linearly transformed F_0 values. One full-cycle includes the estimation of both reconstructed and cyclic reconstructed spectra. Each of encoder and decoder networks are shared for all cycles.*

where

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(y)}) = \sum_{n=1}^{N} -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{X}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) - D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) + \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\boldsymbol{X}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})] + \mathbb{E}_{q_{\boldsymbol{\phi}(\boldsymbol{z}_t|\hat{\boldsymbol{Y}}_t)}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})], \tag{9}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t}) = \mathcal{N}(\boldsymbol{z}_{n,t}; f_{\boldsymbol{\phi}}^{(\mu)}(\hat{\boldsymbol{Y}}_{n,t}) \operatorname{diag}(f_{\boldsymbol{\phi}}^{(\sigma)}(\hat{\boldsymbol{Y}}_{n,t})^2)), \tag{10}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)}) \approx \mathcal{N}(\boldsymbol{s}_t^{(x)}; g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(y|x)}, \boldsymbol{c}^{(x)}), \boldsymbol{I}), \tag{11}$$

$$\hat{\boldsymbol{z}}_{n,t}^{(y|x)} = f_{\boldsymbol{\phi}}^{(\mu)}(\hat{\boldsymbol{Y}}_{n,t}) + f_{\boldsymbol{\phi}}^{(\sigma)}(\hat{\boldsymbol{Y}}_{n,t}) \odot \boldsymbol{\epsilon} \text{ s. t. } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{12}$$

where $\boldsymbol{s}_{n,t}^{(x)}$ and $\boldsymbol{s}_{n,t}^{(x|x)}$ are random variables, $\boldsymbol{s}_t^{(x)}$ is an observed value, and

$$\hat{\boldsymbol{Y}}_{n,t} = [\hat{\boldsymbol{e}}_t^{(y|x)\top}, \hat{\boldsymbol{s}}_{n,t}^{(y|x)\top}]^\top, \tag{13}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(y)}), \tag{14}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(x)}), \tag{15}$$

$$\boldsymbol{X}_{n,t} = [\boldsymbol{e}_t^{(x)\top}, \hat{\boldsymbol{s}}_{n-1,t}^{(x|x)\top}]^\top, \tag{16}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(y|x)}, \boldsymbol{c}^{(x)}). \tag{17}$$

The index of the n-th cycle is denoted as n. The total number of cycle is N. $\hat{\boldsymbol{Y}}_{n,t}$ denotes the converted input features at n-th cycle, $\hat{\boldsymbol{e}}_t^{(y|x)}$ denotes the converted source-to-target excitation features, e.g., linearly transformed F_0, $\hat{\boldsymbol{s}}_{n,t}^{(x|x)}$ denotes the cyclic reconstructed spectra at n-th cycle, and at $n = 1$, $\hat{\boldsymbol{s}}_{1,t}^{(y|x)} = \hat{\boldsymbol{s}}_t^{(y|x)}$, $\hat{\boldsymbol{s}}_{1,t}^{(x)} = \hat{\boldsymbol{s}}_t^{(x)}$, $\hat{\boldsymbol{z}}_{1,t}^{(x)} = \hat{\boldsymbol{z}}_t^{(x)}$ and $\boldsymbol{X}_{1,t} = \boldsymbol{X}_t$. Hence, in the proposed CycleVAE-based VC, the conversion flow is indirectly optimized through the consideration of the converted spectra $\hat{\boldsymbol{s}}_{n,t}^{(y|x)}$ in each n-th cycle.

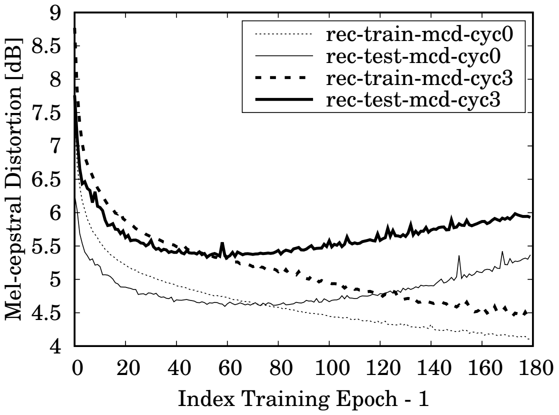

Figure 2: *Mel-cepstral distortion (mcd) of reconstructed (rec) spectra, estimated using the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. mcds were computed with only the speech frames of the input speech.*

4. Experimental Evaluation

4.1. Experimental conditions

We used a subset of the Voice Conversion Challenge (VCC) 2018 [25] dataset, which included four speakers, i.e., SF1, SM1, TF1, and TM1. The speaker notations are as follows: S denotes source speaker, T denotes target speaker, F denotes female speaker, and M denotes male speaker. The total number of utterances in the training and the testing sets were 81 and 35, respectively. The average length per one audio sample is about 3.5 seconds. To develop a non-parallel training dataset, the first 40 utterances were used for corresponding source speaker, while the last 41 were for the target speaker.

WORLD [26] package was used to perform speech analysis. As the spectral envelope parameters, we used the zeroth through 34th mel-cepstrum coeficients converted from the spectral envelope, which was extracted frame-by-frame. As the excitation features, we used log-scaled of continuous F_0 also including an unvoiced/voiced binary decision feature, and 2-dimensional aperiodicity coding coefficients. To perform excitation conversion, mean and variance transformation [11] was performed with respect to the log-scaled F_0 values. The sampling rate of the speech signal was 22,050 kHz. The number of FFT points was 1024. The frame shift was set to 5 ms.

To develop the spectral networks, we used a recurrent neural network (RNN)-based model, which was as follows: dilated convolutional layers were used, to capture the context of -4/+4 input frames, with a kernel size of 3 and 2 layers of 1 and 3 dilation, respectively; gated recurrent unit (GRU) [27] was used with 1024 hidden units and 1 hidden layer; a linear output layer was used; output frame was also fed-back into GRU. Fixed normalization and denormalization layers were used before convolutional and after output layers, respectively, that were set with the statistics of training data. Dropout [28] layers were used with 0.5 probability after convolutional and GRU layers. Network parameters are initialized with Glorot [29] method, and optimized using Adam [30] with 0.0001 learning rate. A batch-frame size of 80 was used.

Four one-to-one spectral models were developed for each of the conventional VAE- and the proposed CycleVAE-based VC, with respect to the four corresponding speaker pairs, i.e., SF1-TF1, SF1-TM1, SM1-TF1, and SM1-TM1. To code the speaker identity, a binary decision value was used. Search of hyperparameters was conducted by varying the number of latent dimensions to 8, 16, 32, 50, and 64, and the number of cycles N, in Eq. (12), to 1, 2, 3, 4, and 5. The optimum number of latent dimensions for both VAE and CycleVAE was 16. The optimum number of cycles for CycleVAE was 3. Objective evaluation was performed to measure the accuracy of the reconstructed and the converted spectra, and the degree of latent features correlation. Another RNN-based parallel spectral conversion models were developed as the upper bound in measuring conversion accuracy. Subjective evaluation was performed to perceptually measure the quality and the accuracy of converted speech between conventional VAE and proposed CycleVAE [1].

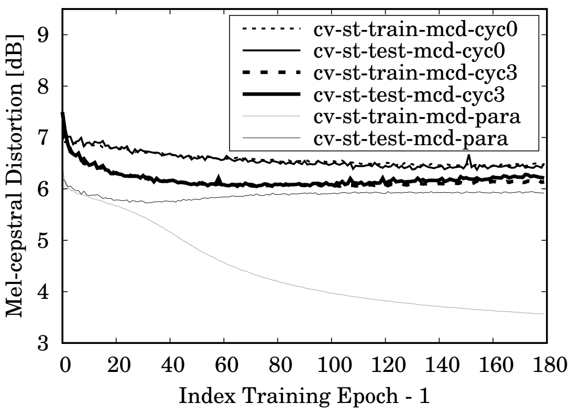

Figure 3: *Mel-cepstral distortion (mcd) of converted source-to-target (cv-st) spectra, estimated using the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. mcds were computed, through DTW alignment, with only the speech frames of corresponding source and target speech.*

Table 1: *Mel-cepstral distortion (MCD) of converted spectra (Cv) and GV-postfiltered [11] converted spectra (PF) with the conventional VAE, the proposed CycleVAE, and parallel spectral modeling as the lower bound, for each speaker-pair conversions. (S: source speaker; T: target speaker; F: female speaker; M: male speaker; Init. denotes the initial MCD values.)*

MCD [dB]	**Init.**	**VAE**		**CycleVAE**		**Parallel**	
		Cv	**PF**	**Cv**	**PF**	**Cv**	**PF**
SF1-TF1	8.18	6.41	6.95	**6.24**	**6.78**	5.92	6.42
SF1-TM1	8.73	6.49	7.03	**5.97**	**6.49**	5.60	6.03
SM1-TF1	9.06	6.83	7.42	**6.29**	**6.78**	6.00	6.43
SM1-TM1	7.68	5.74	6.15	**5.71**	**6.10**	5.36	5.72

4.2. Objective evaluation

Mel-cepstral distortion (MCD) [11] was used to measure the accuracy of both the reconstructed and the converted spectra. Their values are respectively charted, during 180 training epochs, in Figs. 2 and 3. It can be observed that the proposed CycleVAE-based VC yields higher accuracy of converted spectra and lower accuracy of reconstructed spectra compared to the conventional VAE. This trend is somewhat inline with [31],

[1] Implementation is being made available at `https://github.com/patrickltobing/cyclevae-vc`

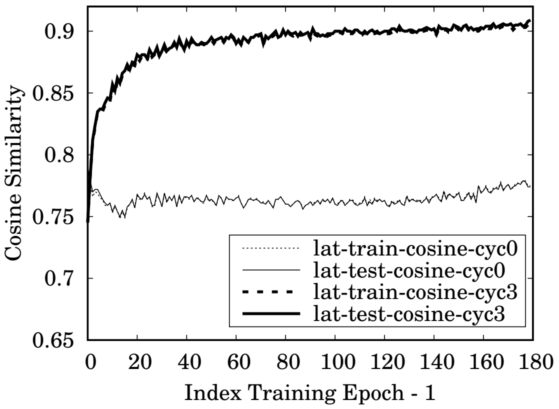

Figure 4: *Cosine similarity (cosine) between latent features of corresponding source and target speech, encoded with the conventional VAE-based (cyc0) and the proposed CycleVAE-based (cyc3) VC, during 180 training epochs, for training (train) and testing (test) sets. cosines were computed, through DTW alignment, with only the speech frames of source and target speech.*

where reconstruction performance is not a proper measure for a better disentanglement of speaker identity (or for better conversion performance). Moreover, MCD values of converted spectra were also computed after applying global variance (GV)-postfilter [11], as given in Table 1. The result shows that the proposed CycleVAE is more suited to additional postfiltering phase compared to the conventional VAE, especially when the speaker identities are considerably distant.

To measure the condition of the latent features, we computed the cosine similarities between the latent features of the source and of the target speaker within the same utterances, which were charted during 180 training epochs, as in Fig. 4. It can be clearly seen that the proposed CycleVAE-based VC generates latent features with higher correlation degree compared to conventional VAE. As studied in [32], higher cosine similarities would be produced by latent attributes that represent either equal phonetic space or equal speaker identities. Hence, CycleVAE is more likely to give latent representations that are closer to phonetic domain due to different speaker identities.

4.3. Subjective evaluation

Perceptual evaluation was performed to compare the quality and the conversion accuracy of converted speech, between the conventional VAE- and the proposed CycleVAE-based VC, by conducting two forced-choice preference tests. In the quality preference test, each listener was presented with two audio stimuli at a time, and was asked to choose a prefered audio by considering both speech naturalness and intelligibility. In the similarity preference test, i.e., to measure the conversion accuracy, each listener was given two audio stimulis, and a reference audio with different utterance, then, was asked to choose a prefered audio that has the closer speaker characteristics to the reference speaker. The numbers of distinct utterances in quality and similarity tests were 6 and 5, respectively, which were randomly chosen from the testing set. Converted speech using parallel spectral models were also included. GV-postfiltered converted spectra was used. The number of listeners was 10.

The results of quality and similarity preference tests are given in Tables 2 and 3, respectively. These results show that the proposed CycleVAE-based VC significantly improves the overall quality and accuracy of converted speech, especially for cross-gender (SF1-TM1, SM1-TF1) conversions, compared to conventional VAE. Their performances for same-gender conversions are statistically similar. This tendency is inline with the objective measurements shown in Table 1, where the conventional VAE-based VC suffers from degradation in cross-gender conversions and the CycleVAE significantly improves them. All audio samples and complete perceptual results can be accessed at `http://bit.ly/2Wg3oIt`.

Table 2: *Result of preference test on speech quality for all, same-gender (S-Gender), and cross-gender (X-Gender) conversion categories using the conventional VAE and the proposed CycleVAE-based VC. CI denotes the 95% confidence interval of the sample mean. p-values were computed using the two-tailed Mann–Whitney U-test with $\alpha < 0.05$. Bold indicates statistically significant better scores.*

Quality Preference	VAE	CycleVAE	CI	p-value
All	40.83%	**59.17%**	±6.27%	6.01e-05
S-Gender	52.50%	47.50%	±9.07%	4.40e-01
X-Gender	29.17%	**70.83%**	±8.25%	1.18e-10

Table 3: *Result of preference test on speaker similarity (Spk. Sim.) for all, same-gender (S-Gender), and cross-gender (X-Gender) conversion categories using the conventional VAE and the proposed CycleVAE-based VC. CI denotes the 95% confidence interval of the sample mean. p-values were computed using the two-tailed Mann–Whitney U-test with $\alpha < 0.05$. Bold indicates statistically significant better scores.*

Spk. Sim. Preference	VAE	CycleVAE	CI	p-value
All	39.00%	**61.00%**	±6.82%	1.11e-05
S-Gender	46.00%	54.00%	±9.94%	2.59e-01
X-Gender	32.00%	**68.00%**	±9.30%	3.81e-07

5. Conclusions

We have presented a novel framework to improve conventional VAE, for a non-parallel VC, by using a cycle-consistent flow, i.e., the proposed CycleVAE. Specifically, the converted spectra, which is not directly optimized, is recycled back into the system, to generate cyclic reconstructed spectra that can be directly optimized. The cyclic flow can be continued by feeding the cyclic reconstructed features back into the system. The experimental results demonstrate that the proposed CycleVAE-based VC yields higher correlation degree of latent features and more accurate converted spectra, while significantly improves the quality and conversion accuracy of the converted speech. Future work includes development of many-to-many VC, and incorporates the use of discrete latent space [33], better prior [34], i-vector [35], additional classifier network [22], and neural waveform generator [36] to produce naturaly sounding converted speech [37] with the proposed CycleVAE.

6. Acknowledgements

This work was partly supported by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant Number JP17H06101.

7. References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, Seatle, Washington, USA, May 1998, pp. 285–288.

[2] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Commun.*, vol. 99, pp. 211–220, 2018.

[3] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, 2007.

[4] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3067–3071.

[5] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Commun.*, vol. 51, no. 3, pp. 268–283, 2009.

[6] O. Türk and M. Schröder, "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 18, no. 5, pp. 965–973, 2010.

[7] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling," *Speech Commun.*, vol. 50, no. 3, pp. 228–243, 2008.

[8] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, 2012.

[9] P. L. Tobing, K. Kobayashi, and T. Toda, "Articulatory controllable speech modification based on statistical inversion and production mappings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2337–2350, 2017.

[10] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[11] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[12] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[13] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 7909–7913.

[14] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. ICASSP*, Las Vegas, USA, Mar. 2008, pp. 4605–4608.

[15] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 6905–6909.

[16] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2032–2045, 2016.

[17] F. Fang, J. Yamagishi, I. Echizen, , and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5279–5283.

[18] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," in *Proc. SLT*, Athens, Greece, Dec. 2018, pp. 266–273.

[19] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, Jeju, South Korea, Dec. 2016, pp. 1–6.

[20] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, , and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3364–3368.

[21] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5274–5278.

[22] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *CoRR arXiv preprint arXiv:1808.05092*, 2018.

[23] D. P. Kingma and J. Ba, "Auto-encoding variational bayes," *CoRR arXiv preprint arXiv:1312.6114*, 2013.

[24] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," *CoRR arXiv preprint arXiv:1903.07593*, 2019.

[25] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," *Corr arXiv preprint arXiv:1804.04262*, 2018.

[26] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[27] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR arXiv preprint arXiv:1406.1078*, 2014.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learning Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, vol. 9, Sardinia, Italy, May 2010, pp. 249–256.

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

[31] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *CoRR arXiv preprint arXiv:1901.08810*, 2019.

[32] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *CoRR arXiv preprint arXiv:1704.04222*, 2017.

[33] A. van den Oord and O. Vinyals, "Neural discrete representation learning," in *Adv. NIPS*, Long Beach, USA, Dec. 2017, pp. 6306–6315.

[34] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Adv. NIPS*, Long Beach, USA, Dec. 2017, pp. 1878–1889.

[35] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation," in *Proc. ICASSP*, New Orleans, USA, Mar. 2017, pp. 5535–5539.

[36] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR arXiv preprint arXiv:1609.03499*, 2016.

[37] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder," in *Proc. ICASSP*, Brighton, UK, May 2019, pp. 6815–6819.