Cross-Lingual Voice Conversion using a Cyclic Variational Auto-encoder and a WaveNet Vocoder

Hikaru Nakatani, Patrick Lumban Tobing, Kazuya Takeda and Tomoki Toda
Nagoya University, Nagoya, Japan
E-mail: nakatani.hikaru@g.sp.m.is.nagoya-u.ac.jp, patrick.lumbantobing@g.sp.m.is.nagoya-u.ac.jp,
takeda@i.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—We propose a novel, cross-lingual voice conversion (VC) method using a cyclic variational auto-encoder (CycleVAE). Voice conversion is the transformation of the voice of one speaker into the voice of another speaker, while cross-lingual VC performs voice conversion between speakers who speak different languages. When using VC methods based on parallel learning, it is necessary to prepare accented speech uttered by the source or target speaker, using the pronunciation system of the speaker's mother tongue. On the other hand, VC methods which use a non-parallel learning approach can utilize the natural speech data of both the source and target speakers, produced in their own native languages. It then becomes necessary, however, to deal with the issues of time-alignment and language mismatches. To address these issues, we apply CycleVAE to cross-lingual VC as a sophisticated, non-parallel method of VC. We also apply the WaveNet vocoder in the waveform generation process of CycleVAE-VC to improve overall conversion quality. Our objective and subjective experimental results when performing cross-lingual VC from a native English speaker to a native Japanese speaker confirm that the proposed method achieves a higher level of naturalness and speaker similarity than a conventional RNN-based parallel VC method using accented speech.**

I. Introduction

Voice conversion (VC) is the process of converting the para-linguistic and non-linguistic information in the source/input speech into those of a target speaker, while retaining the linguistic information contained within the original speech. This makes it possible to convert a speaker's voice into the voice of a famous cartoon character or the voice of another specific person, for example. Medical applications for overcoming a speaker's physical constraints are also possible, such as providing more natural-sounding vocal expression and vocalization support for people with speech impairments such as dysphonia. Parallel VC methods which use the same language and the same utterance data sets have been extensively studied to develop existing VC systems. Proposed methods include using Gaussian Mixture Models (GMM) [1], [2], deep neural networks (DNN) [3], [4], frequency warping [5], [6], case-based modeling [7], [8], etc.

Cross-lingual VC refers to a situation where the source and target speaker are speaking different languages. This technique can be applied to various applications, such as dubbing a foreign language movie using the original actor's voice, personalizing translated speech, and pronunciation practice when learning foreign languages. On the other hand, cross-lingual VC is more of a challenging task compared to parallel VC, since parallel data are usually not available in cross-lingual VC. In parallel VC, spectral mapping can be easily defined through the alignment of speech using dynamic time warping (DTW) [9]. However, in cross-lingual VC, finding frame and segment pairs between speech samples in different languages, especially languages with largely different phoneme set is quite a difficult task. Despite this difficulty, various techniques have been proposed to realize cross-lingual VC until now. For example, a parallel learning method using synthesized speech [10] has been proposed in which the training data is constructed by crudely synthesizing English speech using a Japanese text-to-speech system. Conversion accuracy when using this method is lower than when performing conversion between speakers using the same language, however, because the generated speech is degraded during synthesis. Another method of cross-lingual VC involves the use of bilingual speech [11]. Conversion models are developed separately using a bilingual speaker's English and Japanese speech, then VC is performed using voice data in the language which was not used for training. Although this method is effective, it is not realistic for practical use since the number of bilingual speakers is fairly limited. Another popular approach is the use of whatever prior information is available when developing the conversion model. In eigenvoice conversion [12], a speaker-adaptive conversion model is trained in advance using existing parallel databases consisting of the speech of multiple speakers in a single language, then the data is adapted to the target speaker in another language in an unsupervised manner [13]. Other methods use phoneme information to extract speaker-independent features, such as phoneme posteriorgrams (PPGs), using a pre-trained phoneme recognizer [14]. Cross-lingual VC can also be achieved by developing a target speaker-dependent decoder which converts speaker-independent features into the target speaker's acoustic features. The decoder can be easily trained through a reconstruction process, using only the target speaker's data in an arbitrary language [15]. These methods are effective for cross-lingual VC, but some limitations remain, e.g., the need to use parallel databases, or a pre-trained phoneme recognizer.

Various non-parallel VC techniques have also been proposed, in which fully unsupervised factorization has been applied to non-parallel VC. These approaches include using Generative Adversarial Networks (GAN) [16], [17], Boltzmann Machines (BM) [18], or variational auto-encoders (VAE)

[19], [20], [21]. In the latter methods using VAE, a decoder is trained not only with latent variables, but also using speaker codes that represent speaker labels. After the VAE learns how to extract speaker-independent features from speech spectra at the encoder, VC is then possible by simply inputting speaker independent features and speaker codes into the decoder after training. In a method using a cyclic variational auto-encoder (CycleVAE) [22], the converted spectra are indirectly optimized by re-inputting the converted spectra to the system while maintaining a cycle-consistent mapping flow [23]. CycleVAE has been confirmed to achieve higher conversion accuracy than normal VAE when the source and target speakers speak the same language, by considering the converted features during training.

In this paper, we propose a cross-lingual VC method using CycleVAE. In particular, we use CycleVAE to maintain the pronunciation of a native speaker's English speech, and then convert only the speaker characteristics of the source speaker into those of the target Japanese speaker. In VC frameworks using parallel learning, it is necessary to prepare a parallel data set using the accented, non-native speech of the target speaker. But when training a CycleVAE, parallel data is no longer necessary, and it is possible to train the model using only the speech of each speaker speaking in their native languages, i.e., the English speech of the English speaker and the Japanese speech of the Japanese speaker. In addition, CycleVAE is trained through the simple reconstruction process without pre-trained phoneme recognizer or anything pre-trained in advance, so it is flexible enough for practical use. Furthermore, we also investigate the effect of using a WaveNet vocoder [24] in comparison to using a conventional vocoder for CycleVAE-VC. Conventional vocoders are based only on simplified assumptions that discard phase information, often causing significant degradation in conversion quality. In contrast, the WaveNet vocoder is a deep, auto-regressive, neural vocoder which learns how to directly map acoustic features to a speech waveform in a data-driven manner, using a Convolutional Neural Network (CNN). It has been shown that substituting a WaveNet vocoder for a conventional vocoder improves the quality of synthesized speech in existing speech processing methods, so it has been widely used in research fields related to speech, including for VC. This approach has not been applied within existing CycleVAE-VC frameworks, however, creating an opportunity for the improvement of conversion quality. Hence, in this study, we combine CycleVAE with the WaveNet vocoder to observe how it affects conversion quality in a cross-lingual VC scenario. We also propose a novel fine-tuning approach of the WaveNet vocoder to properly conjugate CycleVAE and WaveNet vocoder. We conduct objective and subjective evaluation experiments to quantitatively evaluate the conversion accuracy of each method.

II. Related Work

A. CycleVAE-based non-parallel VC

Let $\boldsymbol{X}_t = [\boldsymbol{e}_t^{(x)\top}, \boldsymbol{s}_t^{(x)\top}]^\top$, $\boldsymbol{e}_t^{(x)} = [e_t^{(x)}(1), \ldots, e_t^{(x)}(D_e)]^\top$, $\boldsymbol{s}_t^{(x)} = [s_t^{(x)}(1), \ldots, s_t^{(x)}(D_s)]^\top$, $\boldsymbol{c}^{(x)} = [c^{(x)}(1), \ldots, c^{(x)}(D_c)]^\top$, be the $D_e + D_s$, D_e, D_s, and D_c-dimensional feature vectors of the input, the excitation, the spectra, at frame t, and the speaker-code, respectively. Conditioned on a time-invariant speaker-code feature vector sequence $\boldsymbol{c}^{(x)}$, the marginal likelihood of an input feature vector sequence $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ is then given by

$$p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{c}^{(x)}) = \prod_{t=1}^{T} \int p_{\boldsymbol{\theta}}(\boldsymbol{X}_t|\boldsymbol{z}_t, \boldsymbol{c}^{(x)}) p_{\boldsymbol{\theta}}(\boldsymbol{z}_t) \mathrm{d}\boldsymbol{z}_t, \quad (1)$$

where $\boldsymbol{z}_t$ denotes the D_z-dimensional latent feature vector at time t. In a VAE-based framework [25], the above intractable marginal likelihood, and the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{X}_t)$ of the latent variable $\boldsymbol{z}_t$, are solved by the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{X}_t)$, approximating the true posterior.

Specifically, a VAE-based VC [19] model is optimized by maximizing the following variational lower bound

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

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler (KL)-divergence between two probability distributions, and the prior distribution of the latent space is denoted with $p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)$. The parameters set of the encoder and the decoder networks are denoted with $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Given a sampled latent feature vector $\hat{\boldsymbol{z}}_t^{(x)}$ as follows

$$\hat{\boldsymbol{z}}_t^{(x)} = f_{\boldsymbol{\phi}}^{(\mu)}(\boldsymbol{X}_t) + f_{\boldsymbol{\phi}}^{(\sigma)}(\boldsymbol{X}_t) \odot \boldsymbol{\epsilon} \text{ s.t. } \boldsymbol{\epsilon} \sim \mathcal{L}(\boldsymbol{0}, \boldsymbol{I}), \quad (3)$$

where $f_{\boldsymbol{\phi}}(\cdot)$ denotes the encoder network and $\mathcal{L}(\boldsymbol{0}, \boldsymbol{I})$ denotes a standard Laplacian distribution, the reconstructed spectral feature vectors $\hat{\boldsymbol{s}}_t^{(x)}$ and the converted spectral feature vectors $\hat{\boldsymbol{s}}_t^{(y|x)}$ are respectively given by

$$\hat{\boldsymbol{s}}_t^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(x)}), \quad (4)$$

$$\hat{\boldsymbol{s}}_t^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_t^{(x)}, \boldsymbol{c}^{(y)}), \quad (5)$$

where $g_{\boldsymbol{\theta}}(\cdot)$ denotes the decoder network and $\boldsymbol{c}^{(y)}$ denotes the D_c-dimensional feature vector of the target speaker-code. However, the converted spectra is not optimized in this optimization process, hence, the conversion performance of VAE-based VC is limited.

During the training process of the CycleVAE, given 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 $\boldsymbol{c}^{(x)}$ and $\boldsymbol{c}^{(y)}$, a set of network parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ is updated using the following variational lower bound [25]:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{X}_t, \boldsymbol{c}^{(x)}, \boldsymbol{c}^{(y)}) = &\sum_{n=1}^{N} -D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{X}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_{n,t})) \\ &- D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_{n,t})) \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{X}_{n,t})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})] \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\hat{\boldsymbol{Y}}_{n,t})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})], \quad (6)\end{aligned}$$

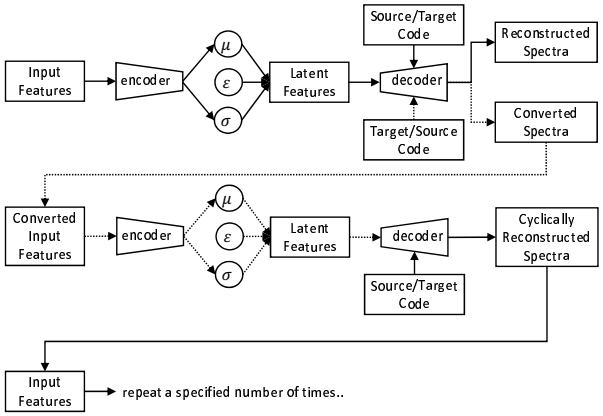

Fig. 1. Training process for the CycleVAE. The process is repeated a specified number of times.

Here, $\boldsymbol{z}_t$ is a D_z-dimensional latent feature vector at time t, $\boldsymbol{s}_{n,t}^{(x)}$ and $\boldsymbol{s}_{n,t}^{(x|x)}$ are random variables, and $\boldsymbol{s}_t^{(x)}$ is an observed value. Also,

$$\hat{\boldsymbol{Y}}_{n,t} = [\hat{\boldsymbol{e}}_t^{(y|x)\top}, \hat{\boldsymbol{s}}_{n,t}^{(y|x)\top}]^\top, \tag{7}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(y|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(y)}), \tag{8}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(x)}, \boldsymbol{c}^{(x)}), \tag{9}$$

$$\boldsymbol{X}_{n,t} = [\boldsymbol{e}_t^{(x)\top}, \hat{\boldsymbol{s}}_{n-1,t}^{(x|x)\top}]^\top, \tag{10}$$

$$\hat{\boldsymbol{s}}_{n,t}^{(x|x)} = g_{\boldsymbol{\theta}}(\hat{\boldsymbol{z}}_{n,t}^{(y|x)}, \boldsymbol{c}^{(x)}), \tag{11}$$

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

where $\hat{\boldsymbol{Y}}_{n,t}$, $\hat{\boldsymbol{e}}_t^{(y|x)}$ and $\hat{\boldsymbol{s}}_{n,t}^{(y|x)}$ are the converted feature vectors of the input, excitation, and spectra, respectively. $\hat{\boldsymbol{s}}_{n,t}^{(x|x)}$ denotes the cyclic reconstructed spectra at the n-th cycle, while $g_{\boldsymbol{\theta}}(\cdot)$ and $f_{\boldsymbol{\phi}}(\cdot)$ denote the encoder and decoder networks, respectively. The index of the n-the cycle is denoted as n, and the total number of cycles is N. For example, 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 $\boldsymbol{X}_{1,t} = \boldsymbol{X}_t$. Fig. 1 shows the training process of the CycleVAE-VC. We then employ the CycleVAE-based VC framework for the development of cross-lingual VC.

B. WaveNet vocoder

The WaveNet vocoder is a deep, auto-regressive, neural vocoder which directly generates speech waveforms from the given auxiliary features [24]. Given a sequence of auxiliary features $\boldsymbol{h} = \left[\boldsymbol{h}_1^\top, \ldots, \boldsymbol{h}_T^\top\right]^\top$, the conditional probability distribution function of the sequence of waveform samples $\boldsymbol{x} = [x_1, \ldots, x_N]^\top$ is given by:

$$P(\boldsymbol{x}|\boldsymbol{h}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P\left(x_t|\boldsymbol{x}_{<t}, \boldsymbol{h}_t, \boldsymbol{\lambda}\right) \tag{13}$$

where $\boldsymbol{x}_{<t}$ denotes the previous waveform samples, with respect to waveform sample x_t at time t, that are proportional to a number of receptive fields of the WaveNet vocoder model, and where $\boldsymbol{\lambda}$ denotes a set of WaveNet vocoder model parameters. In short, given a data sequence pair ($\boldsymbol{h}$, $\boldsymbol{x}$), the WaveNet vocoder learns to map acoustic features $\boldsymbol{h}$ to time-domain signals $\boldsymbol{x}$. The WaveNet vocoder can generate natural-sounding speech almost identical to human speech when features from natural speech are provided as the auxiliary features.

C. Fine-tuning approaches for WaveNet vocoder and their drawbacks

The WaveNet vocoder has been used in many VC frameworks in recent years to improve conversion quality. However, despite its high potential, developing a WaveNet vocoder usually requires quite a large amount of data in order to obtain reasonable quality speech samples. In general, developing a speaker-dependent WaveNet vocoder [24] requires about 1 hour of speech data from the target speaker. Since the amount of target speaker speech data that is available is usually limited, this approach is not flexible enough for practical usage. Therefore, various fine-tuning techniques [26], [27], [28] have been proposed which use multi-speaker models, including [29], in which the vocoder is actually trained using multi-speaker speech data. Although these fine-tuning techniques are effective when sufficient target speech data is unavailable, they still suffers from a quality degradation problem since the characteristics of the natural features used during the training phase and the converted features used in the conversion phase differ.

D. Fine-tuning the WaveNet vocoder using the self-reconstructed features of the CycleVAE

One method that has been used to address the quality degradation issue discussed above is to utilize the self-reconstructed features of the target speaker generated by the VC model for fine-tuning [30], [31]. In VAE-VC, for example, self-reconstructed features have an identical temporal structure to the original target waveforms, so they can be directly used for WaveNet vocoder training without time-alignment. In [30], the self-reconstructed features generated by a VAE-VC model were shown to be similar to the converted features, and effective for alleviating the quality mismatch issue. In [31], a cyclic spectral conversion network is trained, and the self-predicted features from the network are used for fine-tuning. Since CycleVAE is also capable of generating these kinds of features, we hypothesize that the self-reconstructed features generated by the CycleVAE, as shown in Eq. (9) and Eq. (11), are also similar to the converted features, and thus are also useful for alleviating the quality mismatch issue. Specifically, given a set of self-reconstructed features, generated by the CycleVAE using either Eq. (9) or Eq. (11), the WaveNet vocoder learns how to map the features to time-domain signals during the fine-tuning process using the following equation:

$$P(\boldsymbol{x}|\hat{\boldsymbol{h}}, \hat{\boldsymbol{\lambda}}) = \prod_{t=1}^{T} P(x_t|\boldsymbol{x}_{<t}, \hat{\boldsymbol{h}}_t, \hat{\boldsymbol{\lambda}}), \tag{14}$$

where $\hat{\boldsymbol{h}}_t = [\boldsymbol{e}_t^{(x)^\top}, \hat{\boldsymbol{s}}_{1,t}^{(x)^\top}]^\top$ or $\hat{\boldsymbol{h}}_t = [\boldsymbol{e}_t^{(x)^\top}, \hat{\boldsymbol{s}}_{n-1,t}^{(x|x)^\top}]^\top$, and $\hat{\boldsymbol{\lambda}}$ is a set of pre-trained parameters. Thus, the quality mismatch issue should be solved since the characteristics of the features used in training are expected to be more similar to the converted features than natural features $\boldsymbol{h}$.

III. Proposed methods

A. Cross-Lingual VC using a CycleVAE

In this paper, we work on a CycleVAE-based cross-lingual VC framework for the conversion of speech between an English speaker and a Japanese speaker. As we mentioned earlier, we specifically focus on the task of using the CycleVAE to maintain the pronunciation of the native English speaker, while converting only the speaker characteristics into those of the target Japanese speaker.

B. Cross-Lingual VC using a CycleVAE and WaveNet vocoder

We also introduce a WaveNet vocoder in the speech waveform generation phase using the converted spectra from the CycleVAE, in order to further improve conversion quality. To fine-tune the WaveNet vocoder, we develop a CycleVAE model that includes additional English speakers besides the source and target speakers for training, in order to generate additional self-reconstructed features. These features are expected to act as regularizing data to prevent the model from over-fitting during the fine-tuning process, as opposed to using only the self-reconstructed features of the target speaker. Furthermore, to maximize the number of self-reconstructed features of the target speaker that are available, we use various cyclically reconstructed spectra Eq. (11), which were conditioned using not only the source speaker's codes, but also the other speakers' codes. This allows us to increase the amount of self-reconstructed target speaker features that are available. This process can be summarized as follows:

Step 1: Develop a multi-speaker WaveNet vocoder model (which includes the target speaker) and a CycleVAE-VC model (which includes additional speakers).

Step 2: Fine-tune the pre-trained WaveNet vocoder model with the self-reconstructed features of the target speaker and the additional speakers.

Step 3: Conduct further fine-tuning of the WaveNet vocoder model using only the self-reconstructed features of the target speaker.

Fig. 2 shows the method to generate the cyclically reconstructed spectra of each target, and Fig. 3 shows the overall conversion flow between the CycleVAE and WaveNet vocoder.

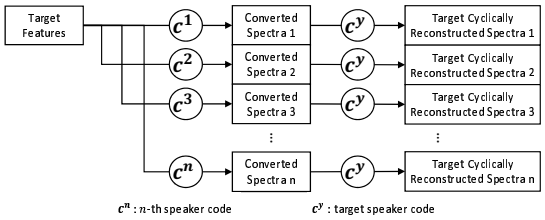

Fig. 2. Generation process for each cyclically reconstructed target spectra. Each spectra is conditioned using the n-th speaker's code, except for the target speaker's spectra.

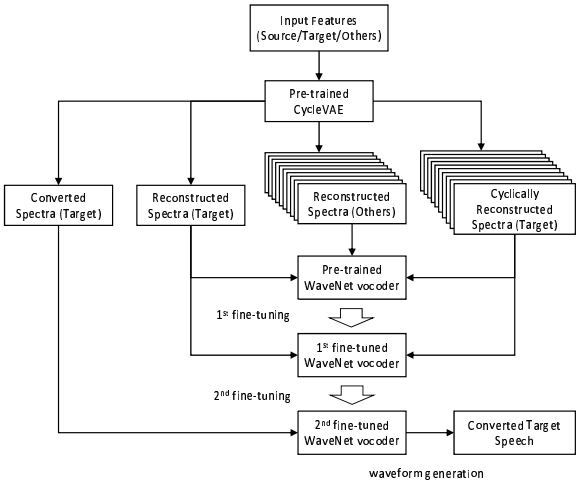

Fig. 3. Overview of proposed VC method using a CycleVAE and WaveNet vocoder.

IV. Experiments

We used one male English speaker from the VCC2018 database [32] as our source speaker, one male Japanese speaker (whose speech samples were recorded separately) as our target speaker, and 24 other English speakers (12 male and 12 female) from the VCTK corpus [33] as additional speakers. The speech data for the source speaker included 81 utterances in English, while the speech data for the target speaker included 81 utterances in Japanese, 31 American English utterances and 31 Japanese-accented English utterances. The speech data from the each additional English speaker included 339 English utterances. The sampling rate for all of the speech data was 24000 Hz.

We used WORLD [34] to extract f_0, aperiodicity, and spectral envelope as speech parameters. We used 49-dimensional Mel-cepstrum coefficients as our spectral envelope parameters. As excitation features, we used the log-scale of continuous f_0, which included an unvoiced/voiced binary decision feature, and two-dimensional aperiodicity coding coefficients. The number of FFT points was 1,024, and the frameshift length was set to 5 ms.

We used a recurrent neural network (RNN) -based model for the encoder and decoder of our CycleVAE model. The details of the model's architecture are as follows: 2 dilated convolutional layers with a kernel size of 3, with 1 and 3 dilations, respectively, a gated recurrent unit (GRU) [35] with 1,024 hidden units and 1 hidden layer, and a linear output layer. The output frame was also fed back into the GRU. We used fixed normalization and de-normalization layers, which were determined based on the statistics of the training data, located before the convolutional layers and after the output layers, respectively. Dropout [36] layers were used with 0.5 probability after the convolutional and GRU layers. We

initialized the network parameters using the Glorot method [37], and optimized them using Adam [38] with a learning rate of 0.0001. Batch frame size and the total number of cycles were set to 24 and 2, respectively. The dimensions of latent features were set to 32.

We adopted the Shallow WaveNet vocoder [39] as our WaveNet vocoder model since the amount of target speech data that was available was limited. The multi-speaker vocoder model was trained in advance using the speech of the additional speakers and the target speaker, then we applied the fine-tuning method explained in the previous section. The details of the model architecture are as follows: The model for the softmax output was trained. The dilation depth was set to 3, and the number of dilation sequence repetitions was 2. The numbers of channels for the residual blocks and skip connections were 192 and 256, respectively. Two convolution layers with a kernel size of 3 and a dilation size of 2 were used to capture the context of auxiliary speech parameters. A noise shaping technique [40] was used to reduce errors in the higher-frequency region. The dropout rate, initialization, and optimization methods of the network parameters were the same as those used during CycleVAE training.

We trained two English-to-Japanese CycleVAE conversion models, which we called Proposed(CycleVAE) and Proposed(CycleVAE+WNV), as well as a baseline parallel learning VC model, to evaluate the conversion accuracy of the proposed method. Proposed(CycleVAE) was trained using speech from the target and source speakers, as well as speech from the additional 24 English speakers. The training data from the target speaker included 50 Japanese utterances. Proposed(CycleVAE+WNV) was a combination of the Proposed(CycleVAE) and the WaveNet vocoder, which was trained separately and fine-tuned using the method described above. Fine-tuning was done using all of the training and test data from the target and additional speakers that was used when developing the CycleVAE. While performing fine-tuning, we used 1 sample of self-reconstructed features from the CycleVAE. For our baseline parallel learning method, we used a spectral mapping model based on an RNN [31]. In [31], the WaveNet vocoder is used as the vocoder, but we used WORLD instead so that we could first simply compare the performance of the RNN-VC and the CycleVAE-VC. The baseline model was trained using 50 accented English utterances from the target speaker.

A. Objective evaluation

For objective evaluation, we used Mel-cepstral distortion (MCD) [41] to measure the quality of the converted speech. MCD values were calculated by comparing spectra of the target speaker's English utterances with the converted source-to-target spectra, using DTW alignment [9]. Table I shows the MCD values of the converted speech for each model. We can see that our proposed methods outperform the baseline method, demonstrating that our proposed methods can generate more accurately converted speech than the baseline method. However, the distance between our two proposed methods was trivial. This is mainly because generated speech from WaveNet vocoder tends to slightly fluctuate, while a generated speech from CycleVAE relatively retains the original shape of the input speech waveform.

TABLE I
OBJECTIVE VOICE CONVERSION ACCURACY FOR BASELINE AND PROPOSED METHODS, USING MEL-CEPSTRAL DISTORTION FOR COMPARISON.

	MCD [dB]
Baseline(RNN Parallel)	8.47
Proposed(CycleVAE)	7.67
Proposed(CycleVAE+WNV)	7.70

B. Subjective evaluation

We conducted a Mean Opinion Score (MOS) test and a speaker identification test to perceptually quantify the conversion accuracy of each model. In the MOS test, listeners rated the naturalness of the converted speech produced by each model on a scale of 1 (very bad) to 5 (very good). In the speaker identification test, participants listened to a sample of the Japanese speaker's Japanese speech and a sample of reference speech, and were asked to judge whether the two speech samples could have been produced by the same speaker by choosing from among four possible responses; 1 = same speaker (sure), 2 = same speaker (not sure), 3 = different speaker (not sure), or 4 = different speaker (sure). The analysis-synthesis of the Japanese speaker's English speech, the Japanese speaker's accented English speech, the source English speaker's English speech and the converted speech using each VC method were used in both experiments. 31 utterances of each type of speech were used in the MOS test, and 10 utterances of each type of speech, randomly selected for each subject, were used in the speaker identification test. The number of subjects who participated in each experiment was 10.

The results of these subjective evaluations are shown in Table II. We conducted a t-test between the baseline method and each of our proposed methods on the results of the MOS test, and a significant difference was found in each comparison at p-value$<$0.01. This result shows that our proposed methods significantly improved the naturalness of the converted speech when compared to the baseline method. Moreover, our two proposed methods were also found to be significantly different from one another at p-value$<$0.01. This means that the speech converted using the CycleVAE-VC becomes even more natural when it is combined with a fine-tuned WaveNet vocoder. Having said that, there is still room for improvement in both cases because neither method achieved the level of naturalness of the analysis-synthesis speech of the target speaker.

We can confirm that both of our proposed methods achieved higher speaker similarity to the target speaker than the baseline method, based on the results of the speaker identification test. We can also see that the Proposed(CycleVAE+WNV) method achieved a higher similarity score than Proposed(CycleVAE). This result means that the speaker similarity of the converted

TABLE II
SUBJECTIVE VOICE CONVERSION ACCURACY, IN TERMS OF NATURALNESS AND SPEAKER IDENTIFICATION. MEAN OPINION SCORE VALUES ARE SHOWN WITH 95% CONFIDENCE INTERVALS. SPEAKER SIMILARITY SCORES WERE COMPUTED BY COMBINING THE "SAME SPEAKER (SURE)" AND "SAME SPEAKER (NOT SURE)" RESPONSES.

	MOS	Correct Rate [%]
Source	-	0.00
Baseline(RNN Parallel)	1.44±0.06	44.44
Proposed(CycleVAE)	2.76±0.08	63.33
Proposed(CycleVAE+WNV)	3.21±0.09	67.77
Target(English)	4.33±0.08	70.00
Target(Accented English)	4.45±0.07	91.11

speech from CycleVAE can be improved by using a WaveNet vocoder. Regarding the target speaker's analysis-synthesis accented English speech, they were correctly identified by 20% of the participants when compared to the English speech without an accent. This is probably because the target speaker's English speech had quite a heavy Japanese accent, so that is almost sounded like Japanese speech. All of the speech samples are available at: "https://bit.ly/2KSbfIB"

V. CONCLUSIONS

In this paper, we proposed a cross-lingual VC method using CycleVAE. Our experimental results showed that the proposed method outperformed a conventional, parallel learning method using an RNN, in terms of the quality, naturalness, and speaker similarity of the converted speech. We achieved further improvement in performance by using a WaveNet vocoder in the waveform generation process, which was also objectively and subjectively confirmed. As future work, we plan to explore methods of implementing cross-lingual VC in real-life applications.

VI. ACKNOWLEDGMENTS

This work was partially supported by JST, CREST and JPMJCR19A3.

REFERENCES

[1] Y. Stylianou, O. Cappé, 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] 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.

[3] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[4] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "A probabilistic interpretation for artificial neural network-based voice conversion," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 552–558.

[5] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2011.

[6] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[7] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1876, 2017.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 313–317.

[9] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[10] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *The Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 76–82, 1991.

[11] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," pp. 2177–2185, 2002.

[12] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," *Proc. ICSLP, Pittsburgh, PA*, pp. 2446–2449, 2006.

[13] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," 2009.

[14] 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)*. IEEE, 2016, pp. 1–6.

[15] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[16] 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," *arXiv preprint arXiv:1704.00849*, 2017.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[18] T. Nakashika, T. Takiguchi, Y. Minami, T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2032–2045, 2016.

[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 *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[20] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5274–5278.

[21] 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," *arXiv preprint arXiv:1704.00849*, 2017.

[22] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," *arXiv preprint arXiv:1907.10185*, 2019.

[23] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.

[24] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder." in *Interspeech*, vol. 2017, 2017, pp. 1118–1122.

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

[26] Y. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018," EasyChair Preprint no. 65, EasyChair, 2018.

[27] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted WaveNet vocoder." in *Interspeech*, 2018, pp. 1978–1982.

[28] W.-C. Huang, C.-C. Lo, H.-T. Hwang, Y. Tsao, and H.-M. Wang, "WaveNet vocoder and its applications in voice conversion," in *Proc. The 30th ROCLING Conference on Computational Linguistics and Speech Processing (ROCLING)*, 2018.

[29] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.

[30] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Refined WaveNet vocoder for variational autoencoder based voice conversion," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[31] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cycleRNN-based spectral mapping and finely tuned WaveNet vocoder," *IEEE Access*, vol. 7, pp. 171 114–171 125, 2019.

[32] 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," *arXiv preprint arXiv:1804.04262*, 2018.

[33] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSRT vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[34] M. Morise, F. Yokomori, and K. 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.

[35] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

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

[39] P. L. Tobing, Y. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Efficient shallow WaveNet vocoder using multiple samples output based on laplacian distribution and linear prediction," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7204–7208.

[40] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5664–5668.

[41] 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.