INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Cyclic Spectral Modeling for Unsupervised Unit Discovery into Voice Conversion with Excitation and Waveform Modeling

Patrick Lumban Tobing[1], *Tomoki Hayashi*[1], *Yi-Chiao Wu*[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, hayashi.tomoki, yichiao.wu}@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

We present a novel approach of cyclic spectral modeling for unsupervised discovery of speech units into voice conversion with excitation network and waveform modeling. Specifically, we propose two spectral modeling techniques: 1) cyclic vector-quantized autoencoder (CycleVQVAE), and 2) cyclic variational autoencoder (CycleVAE). In CycleVQVAE, a discrete latent space is used for the speech units, whereas, in CycleVAE, a continuous latent space is used. The cyclic structure is developed using the reconstruction flow and the cyclic reconstruction flow of spectral features, where the latter is obtained by recycling the converted spectral features. This method is used to obtain a possible speaker-independent latent space because of marginalization on all possible speaker conversion pairs during training. On the other hand, speaker-dependent space is conditioned with a one-hot speaker-code. Excitation modeling is developed in a separate manner for CycleVQVAE, while it is in a joint manner for CycleVAE. To generate speech waveform, WaveNet-based waveform modeling is used. The proposed framework is entried for the ZeroSpeech Challenge 2020, and is capable of reaching a character error rate of 0.21, a speaker similarity score of 3.91, a mean opinion score of 3.84 for the naturalness of the converted speech in the 2019 voice conversion task.

Index Terms: unsupervised speech unit discovery, cyclic modeling, spectral, voice conversion, excitation, WaveNet vocoder

1. Introduction

Recently, a lot of works have been carried out on the so-called zero resource settings of speech processing [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. In a zero resource setting, the textual information of the speech data is not provided, e.g., segmented annotations, which is typical in a low resource language and, to a certain degree, in speech acquisition for infants. Notably, in 2015 [4, 5], 2017 [7], and 2019 [10], the Zero Resource Speech Challenge (ZSC) has been organized, where the objective is to perform the discovery of subword/wordlike speech units (speech sounds representations) without any textual annotations, i.e., in an unsupervised manner. Further, in ZSC 2019 [10], it is extended into a speech synthesis task to generate the same linguistic contents as the input speech, but with a different voice characteristic, i.e., voice conversion (VC) [14]. In this paper, we present our system for the ZSC 2020 on the unsupervised discovery of subword units into the VC pipeline of 2019 task.

A lot of works have been proposed for the unsupervised discovery of subword units [3, 6, 8, 11, 12, 13]. These methods are mainly evaluated in terms of ABX discriminability for triphones with different centering phonemes [15, 16]. It is logical that the ideal subword units have to be free as much as possible from speaker-dependent characteristics (paralinguistic), e.g., prosody or voice timbre. Further, in order to acquire reliable final product of a speech synthesis pipeline, it might be useful to develop these kind of frameworks in an end-to-end manner.

To obtain such capabilities, in this work, we propose to use a system that is based on a spectral modeling method for VC. Specifically, the spectral modeling is based on variational autoencoder (VAE) [17] framework, where a prospective speaker-independent latent space, used for the subword units, is regularized with a chosen prior distribution. The speaker-dependent space, however, is defined by the use of a one-hot speaker code. The objective includes the spectral reconstruction term and the regularization term. In a conventional VAE-based VC [18], though, the performance is still limited due to the minimal treatment for the latent space to be disentangled from the speaker-dependent traits.

In this paper, to possibly extract a cleaner latent space in an unsupervised manner, we propose to use a cyclic framework of VAE (CycleVAE) [19] to disentangle the speaker-dependent traits by means of marginalization of all possible speaker conversion pairs. This can be achieved by sampling a target speaker from the dataset at each cycle, generate the corresponding converted spectra, and recycle the converted spectra to obtain the cyclic reconstructed spectra of the input speaker. In addition, inspired by [20], we also extract the speaker-posterior, together with the latent-posterior, from the encoder. Subsequently, to obtain a lower bitrate value for subword units, as one of the objective in 2019 task of ZSC 2020, we also propose to use a discrete latent space, which is an extension of the conventional vector-quantized VAE (VQVAE) [21, 11, 12, 13] into a cyclic framework of VQVAE (CycleVQVAE).

Finally, to provide the capability of speech waveform generation using the discovered subword units, we utilize additional excitation and waveform modelings. For the CycleVQVAE-based spectral model, the excitation module is developed in a separate manner, whereas for the CycleVAE-based model, it is developed in a joint manner through an ad-hoc modification of the variational lower bound. On the other hand, WaveNet vocoder [22, 23] is used as the framework for the waveform modeling. The results of ZSC 2020 demonstrate that the proposed framework is capable of reaching highly competitive scores in the VC of 2019 task.

2. Spectral Modeling based on Variational Autoencoder

Let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_T^\top]^\top$ be the sequence of speech feature vectors, where $\boldsymbol{x}_t = [\boldsymbol{e}_t^{(x)^\top}, \boldsymbol{s}_t^{(x)^\top}]^\top$. The D_e-th dimensional excitation and the D_s-th dimen-

http://dx.doi.org/10.21437/Interspeech.2020-2559

sional spectral feature vectors are respectively denoted as $\boldsymbol{e}_t^{(x)\top} = [e_t^{(x)}(1), \ldots, e_t^{(x)}(D_e)]^\top$ and $\boldsymbol{s}_t^{(x)\top} = [s_t^{(x)}(1), \ldots, s_t^{(x)}(D_s)]^\top$ at time t.

In this work, we assume that the probability density function (pdf) of the speech feature vector $\boldsymbol{x}_t$ at time t is composed as $p(\boldsymbol{x}_t) \propto p(\boldsymbol{s}_t^{(x)})p(\boldsymbol{e}_t^{(x)})$. Hence, the term in objective function $p(\boldsymbol{x}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t)$ can be written as follows:

$$p(\boldsymbol{x}_t) = \int \sum_{\boldsymbol{c}\in\boldsymbol{C}} p(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c})p(\boldsymbol{e}_t^{(x)})p(\boldsymbol{c}|\boldsymbol{x}_t)p(\boldsymbol{z}_t)\mathrm{d}\boldsymbol{z}_t \simeq \int p(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)})p(\boldsymbol{c}^{(x)}|\boldsymbol{x}_t)p(\boldsymbol{z}_t)\mathrm{d}\boldsymbol{z}_t, \quad (1)$$

where $\boldsymbol{z}_t$ denote the latent feature vector at time t. $\boldsymbol{c}$ denotes a time-invariant one-hot speaker-code, and $\boldsymbol{c}^{(x)}$ is the speaker-code for the input speech features $\boldsymbol{x}_t$. $\boldsymbol{C}$ is the set of speaker-codes for all available speakers. The excitation term $p(\boldsymbol{e}_t^{(x)})$ is assumed to be constant due to only spectral modeling,

In a variational autoencoder (VAE) [17], an inference network is used to model the approximate of true posterior of the latent features $p(\boldsymbol{z}_t|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t,\boldsymbol{z}_t)}{p(\boldsymbol{x}_t)}$ as follows:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t) \simeq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_t) + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{x}_t)||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{x}_t)), \quad (2)$$

where the variational lower bound is given by

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_t) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{x}_t)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_t^{(x)}|\boldsymbol{z}_t, \boldsymbol{c}^{(x)}) + \log p_{\boldsymbol{\phi}}(\boldsymbol{c}^{(x)}|\boldsymbol{x}_t) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{x}_t)||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)). \quad (3)$$

The sets of inference parameters (encoder) and generative parameters (decoder) are denoted as $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively.

3. Cyclic Spectral Modeling for Unsupervised Unit Discovery

3.1. CycleVAE-based spectral modeling

In CycleVAE-based spectral modeling, the variational lower bound is defined as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_t) = &\sum_{n=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{x}_{n,t})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})] \\ &+ \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{y}_{n,t})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|y)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}, \boldsymbol{c}^{(x)})] \\ &- D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{x}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{y}_{n,t})||p_{\boldsymbol{\theta}}(\boldsymbol{z}_t)) \\ &+ \log p_{\boldsymbol{\phi}}(\boldsymbol{c}_{n,t}^{(x)} = \boldsymbol{c}^{(x)}|\boldsymbol{x}_{n,t}) + \log p_{\boldsymbol{\phi}}(\boldsymbol{c}_{n,t}^{(y)} = \boldsymbol{c}^{(y)}|\boldsymbol{y}_{n,t}), \quad (4)\end{aligned}$$

where

$$\boldsymbol{s}_{n,t}^{(x)} = g_{\boldsymbol{\theta}}(\boldsymbol{z}_{n,t}^{(x)}, \boldsymbol{c}^{(x)});\ \boldsymbol{s}_{n,t}^{(x|y)} = g_{\boldsymbol{\theta}}(\boldsymbol{z}_{n,t}^{(y)}, \boldsymbol{c}^{(x)}), \quad (5)$$

$$\boldsymbol{z}_{n,t}^{(x)} = f_{\boldsymbol{\phi}}(\boldsymbol{x}_{n,t})^{(\mu)} - f_{\boldsymbol{\phi}}(\boldsymbol{x}_{n,t})^{(\sigma)} \odot \boldsymbol{\epsilon}, \quad (6)$$

$$\boldsymbol{z}_{n,t}^{(y)} = f_{\boldsymbol{\phi}}(\boldsymbol{y}_{n,t})^{(\mu)} - f_{\boldsymbol{\phi}}(\boldsymbol{y}_{n,t})^{(\sigma)} \odot \boldsymbol{\epsilon}, \quad (7)$$

$$\boldsymbol{\epsilon} = \mathrm{sign}(\boldsymbol{U}) \ln(1 - 2|\boldsymbol{U}|), \text{ s. t. } \boldsymbol{U} \sim (-1/2, 1/2], \quad (8)$$

$$\boldsymbol{x}_{n,t} = [\boldsymbol{e}_t^{(x)\top}, \boldsymbol{s}_{n-1,t}^{(x|y)\top}]^\top;\ \boldsymbol{s}_{0,t}^{(x|y)} = \boldsymbol{s}_t^{(x)}, \quad (9)$$

$$\boldsymbol{y}_{n,t} = [\boldsymbol{e}_t^{(y|x)\top}, \boldsymbol{s}_{n,t}^{(y|x)\top}]^\top;\ \boldsymbol{s}_{n,t}^{(y|x)} = g_{\boldsymbol{\theta}}(\boldsymbol{z}_{n,t}^{(x)}, \boldsymbol{c}^{(y)}), \quad (10)$$

$$\boldsymbol{c}_{n,t}^{(x)} = f_{\boldsymbol{\phi}}(\boldsymbol{x}_{n,t})^{(c)};\ \boldsymbol{c}_{n,t}^{(y)} = f_{\boldsymbol{\phi}}(\boldsymbol{y}_{n,t})^{(c)}, \quad (11)$$

$$\boldsymbol{c}^{(y)} \in \boldsymbol{C} \setminus \boldsymbol{c}^{(x)}, \quad (12)$$

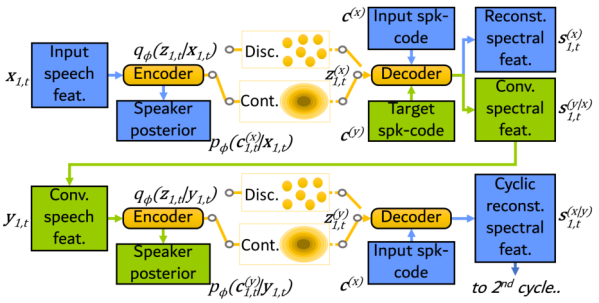

Figure 1: *Cyclic spectral modeling based on variational autoencoder (VAE) for unsupervised unit discovery and spectral features (feat.) conversion (conv.) with discrete (disc.), i.e., CycleVQVAE, or continous (cont.), i.e., CycleVAE, latent space. Target speaker (spk) is sampled from the dataset at each cycle.*

$\boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{z}_t) = \mathcal{L}(\boldsymbol{0}, \boldsymbol{1})$, and $\boldsymbol{y}_{n,t}$ denotes the converted speech features at the n-th cycle of time t. $\boldsymbol{e}_t^{(y|x)}$ denotes the excitation feature vector of the converted speaker $\boldsymbol{c}^{(y)}$ at time t, such as linear-transformed fundamental frequency (F0) from that of the input speaker $\boldsymbol{c}^{(x)}$. The number of cycles is denoted as N. A signed function is denoted as $\mathrm{sign}(\cdot)$. $\mathcal{L}(\boldsymbol{0}, \boldsymbol{1})$ denotes the standard Laplacian distribution. The encoder and decoder network functions are denoted as $f_{\boldsymbol{\phi}}(\cdot)$ and $g_{\boldsymbol{\theta}}(\cdot)$, respectively. The output of the encoder network is consisted of the location μ and the scale σ parameters of the approximate latent-posterior $q_{\boldsymbol{\phi}}(\cdot)$ and the logits c of the speaker-posterior $p_{\boldsymbol{\phi}}(\cdot)$.

At each n-th cycle, the converted speaker $\boldsymbol{c}^{(y)}$ is randomly sampled from the set of speakers $\boldsymbol{C}$ excluding the input speaker $\boldsymbol{c}^{(x)}$. As the number of optimization steps increases, it converges towards marginalization of all possible speaker conversion pairs. Given input speech feature vector $\boldsymbol{x}_{1,t}$, the subword units are assumed to be represented by the latent feature vector $\boldsymbol{z}_{1,t}^{(x)}$ at time t. The flow of CycleVAE is illustrated in Fig. 1 using the continuous part (Cont.) of latent space configuration.

3.2. CycleVQVAE-based spectral modeling

In this paper, we introduce cyclic vector-quantized variational autoencoder (CycleVQVAE)-based spectral modeling, which is depicted using the discrete part (Disc.) of latent space configuration in Fig. 1. Following Eq. (4), the variational lower bound is defined as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\varphi}; \boldsymbol{x}_t) = &\sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}^{(x)}, \boldsymbol{c}^{(x)}) \\ &+ \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_{n,t}^{(x|y)} = \boldsymbol{s}_t^{(x)}|\boldsymbol{z}_{n,t}^{(y)}, \boldsymbol{c}^{(x)}) \\ &+ \log p_{\boldsymbol{\phi}}(\boldsymbol{c}_{n,t}^{(x)} = \boldsymbol{c}^{(x)}|\boldsymbol{x}_{n,t}) + \log p_{\boldsymbol{\phi}}(\boldsymbol{c}_{n,t}^{(y)} = \boldsymbol{c}^{(y)}|\boldsymbol{y}_{n,t}), \\ &- D(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{x}_{n,t}), \boldsymbol{z}_{n,t}^{(x)}) - D(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{y}_{n,t}), \boldsymbol{z}_{n,t}^{(y)}), \quad (13)\end{aligned}$$

where

$$\boldsymbol{z}_{n,t}^{(x)} = \min_{\boldsymbol{v}_{\boldsymbol{\varphi}} \in \boldsymbol{V}_{\boldsymbol{\varphi}}} ||q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{x}_{n,t}) - \boldsymbol{v}_{\boldsymbol{\varphi}}||, \quad (14)$$

$$\boldsymbol{z}_{n,t}^{(y)} = \min_{\boldsymbol{v}_{\boldsymbol{\varphi}} \in \boldsymbol{V}_{\boldsymbol{\varphi}}} ||q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{y}_{n,t}) - \boldsymbol{v}_{\boldsymbol{\varphi}}||, \quad (15)$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{x}_{n,t}) = f_{\boldsymbol{\phi}}(\boldsymbol{x}_{n,t})^{(z)}, \quad (16)$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}_{n,t}|\boldsymbol{y}_{n,t}) = f_{\boldsymbol{\phi}}(\boldsymbol{y}_{n,t})^{(z)}, \quad (17)$$

$||\cdot||$ denotes a vector norm, $D(\cdot, \cdot)$ denotes a specified distance function, and $\boldsymbol{V}_{\boldsymbol{\varphi}}$ denotes the set of VQ codebook parameters.

In order to bypass the min function, straight-through estimator [21] is used at each n-th cycle to allowing backpropagation path from the decoder $g_{\boldsymbol{\theta}}(\cdot)$ into the encoder $f_{\boldsymbol{\phi}}(\cdot)$. On the other hand, the VQ codebook $V_{\boldsymbol{\varphi}}$ is optimized only with the distance term $D(q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1,t}|\boldsymbol{x}_{1,t}), \boldsymbol{z}_{1,t}^{(x)})$ and the conditional pdf $p(\boldsymbol{s}_{1,t}^{(x)}|\boldsymbol{z}_{1,t}^{(x)}, \boldsymbol{c}^{(x)})$ of the 1st cycle.

4. Excitation and Waveform Modeling for Speech Generation

To accomodate the voice conversion task in the ZSC 2020, we present a novel approach to allow excitation and waveform modeling, where the pipeline is depicted in Fig. 2.

4.1. Excitation modeling for CycleVQVAE and CycleVAE

For the CycleVQVAE-based spectral modeling, we use a separate excitation network to estimate the excitation feature vector $\boldsymbol{e}_t^{(x)}$ of speaker $\boldsymbol{c}^{(x)}$ given the spectral feature vector $\boldsymbol{s}_t^{(x)}$ at time t. Hence, the term in objective function $p(\boldsymbol{e}|\boldsymbol{s}) = \prod_{t=1}^{T} p(\boldsymbol{e}_t|\boldsymbol{s}_t)$ can be denoted as follows:

$$p(\boldsymbol{e}_t|\boldsymbol{s}_t) = \sum_{\boldsymbol{c}^{(x)} \in \boldsymbol{C}} p_{\boldsymbol{\vartheta}}(\boldsymbol{e}_t^{(x)}|\boldsymbol{s}_t^{(x)}, \boldsymbol{c}^{(x)}) p(\boldsymbol{c}^{(x)}), \tag{18}$$

where $\boldsymbol{\vartheta}$ denotes the set of excitation network parameters. $p(\boldsymbol{c}^{(x)}) = 1$ for $\boldsymbol{s}_t^{(x)}$ and 0 otherwise.

On the other hand, in the CycleVAE-based spectral modeling, we use an ad-hoc modification of variational lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_t)$ in Eq.(4) to perform a joint optimization as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\vartheta}; \boldsymbol{x}_t) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_t) + \log p_{\boldsymbol{\vartheta}}(\boldsymbol{e}_{1,t}^{(x)} = \boldsymbol{e}_t^{(x)}|\boldsymbol{s}_{1,t}^{(x)}, \boldsymbol{c}^{(x)}) \\ + \log p_{\boldsymbol{\vartheta}}(\boldsymbol{e}_{1,t}^{(y|x)} = \boldsymbol{e}_t^{(y|x)}|\boldsymbol{s}_{1,t}^{(y|x)}, \boldsymbol{c}^{(y)}).\end{aligned} \tag{19}$$

4.2. Waveform modeling with WaveNet vocoder

In order to generate speech waveform, we propose to use WaveNet-based waveform modeling [22, 23], where the pdf of waveform samples $\boldsymbol{w} = [w_1, \ldots, w_T]^\top$ is given by

$$p(\boldsymbol{w}) = \prod_{t=1}^{T} p_{\boldsymbol{\psi}}(w_t|\boldsymbol{w}_{t-p}, \boldsymbol{x}_t). \tag{20}$$

$\boldsymbol{w}_{t-p}$ denotes the p past waveform samples, and $\boldsymbol{x}_t$ denotes the conditioning speech features, e.g., spectral and excitation feature vectors, at time t.

Furthermore, to reduce the mismatches between natural speech features $\boldsymbol{x}_t$ used in the training, and estimated speech features, e.g., converted spectral features $\boldsymbol{s}_{1,t}^{(y|x)}$, a WaveNet fine-tuning procedure is introduced [24, 25]. Following Eq.(20), the objective function in the fine-tuning is as follows:

$$p(\boldsymbol{w}|\hat{\boldsymbol{\psi}}) = \prod_{t=1}^{T} p_{(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}})}(w_t|\boldsymbol{w}_{t-p}, \hat{\boldsymbol{x}}_t), \tag{21}$$

where $\hat{\boldsymbol{x}}_t = [\boldsymbol{s}_{1,t}^{(x)\top}, \boldsymbol{e}_t^{(x)\top}]^\top$, and $\hat{\boldsymbol{\psi}}$ denotes the parameters set of a pretrained WaveNet.

5. Experimental Evaluation

5.1. Experimental conditions

We used WORLD [26, 27] package to parameterize the speech signal into spectral and excitation features. Specifically, 49-th dimensional mel-cepstrum parameters [28] including 0-th power were used as the spectral envelope features. On the other hand, log-continuous F0 values, including unvoiced/voiced (U/V) binary decisions, and 1-dimensional code-aperiodicity, were used as the excitation features. The sampling rate of the speech signal was 16000 Hz. The frame shift was set to 10 ms. The number of FFT points for analysis was 1024. The frequency warping coefficient was set to 0.455.

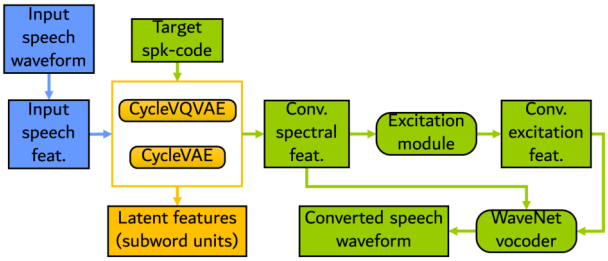

Figure 2: *Flow of the proposed frameworks to perform discovery of subword units using the latent space representations of CycleVAE/CycleVQVAE spectral model, and into voice conversion with additional excitation module and WaveNet vocoder.*

The dataset of the 2019-task [10] in ZSC 2020 [1]consisted of English and Surprise [29, 30] languages, where each has a subset of voice dataset and unit dataset. The voice dataset of English language consisted of 1 male and 1 female of about 4.6 hours (hr) in total, whereas that of the Surprise language consisted of 1 female of about 1.5 hr. The unit dataset of English language consisted of another 100 speakers of about 15.6 hr in total, whereas that of the Surprise language consisted of another 112 speakers of about 15 hr in total. The test dataset of each language consisted of speech data from speakers that are outisde of the speakers set $\boldsymbol{C}$ of the training data.

Each of the CycleVQVAE- and CycleVAE-based spectral models [2], described in Section 3, were developed using the speech data of all 215 speakers in the 2019-task. CycleVAE adopted the joint excitation modeling with objective function given in Eq. (19). CycleVQVAE adopted the separate excitation modeling with objective in Eq. (18), where the excitation module was trained using only the speech data of the 3 target speakers in the 2019-task. Note that we also included reconstructed $\boldsymbol{s}_{1,t}^{(x)}$ and cyclic reconstructed $\boldsymbol{s}_{1,t}^{(x|y)}$ spectral features to develop the excitation module of CycleVQVAE.

The network architectures for CycleVQVAE and CycleVAE models were based on [19]. The number of cycles was set to 2 and 3 for CycleVQVAE and CycleVAE, respectively. The number of latent dimensions was set to 50 and 32 for CycleVQVAE and CycleVAE, respectively. The number of centroids was set to 50 for CycleVQVAE. Adam [31] was used to optimize CycleVQVAE, while rectified Adam [32] was used to optimize CycleVAE. The learning rate was set to 0.0001. Weight normalization [33] was used for convolution layers in CycleVAE. As input excitation features for CycleVAE, we used additional U/V decisions for code-aperiodicity, and transformed the code-aperiodicity values into their continuous negative-log.

For the waveform model, we used a shallow WaveNet vocoder using discrete output (softmax), which was exactly the same as in [34]. A multispeaker WaveNet vocoder [35] was trained using the 3 target speakers. Fine-tuning using reconstructed spectral features $\boldsymbol{s}_{1,t}^{(x)}$ [24, 25], as given in Eq.(21), was

[1] https://zerospeech.com/2020/results

[2] An implementation package has been made available at: https://github.com/patrickltobing/cyclevae-vc-neuralvoco

Table 1: *Results on mel-cepstral distortion (MCD) between converted and target spectral features, and on root-mean-square error (RMSE) and cosine similarity (Cos-Sim) between latent features of source and target utterances for CycleVQVAE spectral modeling on English test dataset of 2019-task in ZSC 2020. Dynamic time warping was performed on only speech frames. The number of latent-dimensions (Lat) and centroids (Ctr) were set the same, i.e.,* 32 *or* 50*. The number of cycles (Cyc) were set to* 0*,* 1*, or* 2*. The target speaker was male (M) or female (F).*

Lat-Ctr-Cyc	MCD [dB]		RMSE		Cos-Sim	
	M	F	M	F	M	F
32-32-0	6.46	6.15	0.29	0.26	0.68	0.73
32-32-1	6.19	5.98	0.20	0.18	0.82	0.85
32-32-2	**6.01**	**5.87**	**0.18**	**0.16**	**0.89**	**0.91**
50-50-2	**5.99**	**5.89**	**0.08**	**0.07**	**0.92**	**0.93**

performed on only CycleVQVAE for each target speaker.

5.2. Experimental results

5.2.1. Internal evaluation results

In our internal evaluations, we performed objective measurements on the English test dataset to confirm the CycleVQVAE framework. The English test dataset contained 193 utterances to be converted into either of the 2 target English speakers. Given that the parallel target utterances were also provided, we could compute the feature distance using dynamic-time-warping (DTW) between the source/converted and the target utterances. For CycleVAE, we based its hyperparameters on the previous work [19].

In the first internal evaluation, we measured mel-cepstral distortion (MCD) between the converted spectra and the target spectra in the English test. Secondly, we measured the root-mean-square error (RMSE) and cosine-similarity [36] between the VQ latent features estimated from the source utterances and that from the target utterances. The results of spectral distortion and latent features measurement, using 0, 1, or 2 cycles, with 32 or 50 latent-dimensions/centroids, are shown in Table 1. The results show that the use of 2 cycles could provide better accuracy of spectral conversion. Moreover, it can also be observed that marginalization over possible conversion pairs in training, i.e., by the use of cycles, could provide more similar latent spaces for two same utterances between different speakers.

5.2.2. Official evaluation results

The official evaluation results from ZSC 2020 include objective and subjective evaluations. The objective evaluation consisted of ABX discriminability test, of discovered units, between triphones with different centering phoneme, and of character error rate (CER) for the synthesized speech. The subjective evaluation consisted of mean opinion score (MOS) test for naturalness of synthesized speech and a test of its similarity (Sim) to the reference target speaker. Additionaly, the bitrate values were also computed with respect to the discovered subword units. Note that for ABX and bitrate tests, the test datasets consisted of additional audios, each having short durations, totaling in about 1.65 hr (13529 utterances) and 1.37 hr (10189 utterances) for English and Surprise test sets, respectively. The number of synthesized utterances for Surprise test set was 150.

The official evaluation results are shown in Table 2. It can be observed that for Surprise set, CycleVAE yields better score with 3.84 MOS, 0.21 CER, and 3.91 Sim. As for English set, CycleVQVAE yields better score in MOS and Sim with values of 3.4 and 3.79, respectively, while CycleVAE yields better CER with a value of 0.31. In terms of ABX score and bitrate, the Topline system yields the best values.

Table 2: *Official evaluation results on mean opinion score (MOS) test of naturalness, character error rate (CER), and speaker similarity (Sim) of the synthesized (converted) speech for English and Surprise test sets. Additional sets of audios were used for ABX discriminability test between triphones with different centering phoneme, and for bitrate calculation of the discovered units. Systems include the topline and the baseline of ZeroSpeech 2020, and our CycleVQVAE- and CycleVAE-based spectral modeling for unit discovery, with their excitation and waveform models for voice conversion.*

Surprise	MOS	CER	Sim	ABX	Bitrate
Topline	3.49	0.33	3.77	**16.09**	**35.2**
Baseline	2.23	0.67	3.26	27.46	74.55
CycleVQVAE	3.28	0.33	3.64	18.13	463.75
CycleVAE	**3.84**	**0.21**	**3.91**	24.42	1745.79
English	**MOS**	**CER**	**Sim**	**ABX**	**Bitrate**
Topline	2.52	0.43	3.1	**29.85**	**37.73**
Baseline	2.14	0.77	2.98	35.63	71.98
CycleVQVAE	**3.4**	0.46	**3.79**	30.54	468.23
CycleVAE	3.31	**0.31**	3.16	36.29	1739.85

From our investigations, we have found that the use of separate excitation module as in CycleVQVAE significantly degrades the phonetic accuracy in converted speech, due to the high-error in excitation estimation using generated spectral features. Hence, its worst CER values compared to CycleVAE with joint excitation modeling. However, for higher F0 into lower F0 conversion, as in English sets, the ad-hoc joint excitation module seems to underestimate the converted prosody, hence, the CycleVAE has lower Sim value in English test. Further, the CycleVAE also uses refined input excitation features, where it could produce less oversmoothed spectral trajectory compared to CycleVQVAE, hence, its acceptably higher MOS scores even though it does not use fine-tuned WaveNet vocoder. To improve, we will investigate the use two separate latent spaces for a cyclic spectral and excitation modeling.

6. Conclusions

We have presented a novel framework for unsupervised subword units discovery based on cyclic spectral modeling with variational autoencoder (VAE). Continuous latent space is proposed through cyclic VAE (CycleVAE), while discrete latent space is proposed through cyclic vector-quantized VAE (CycleVQVAE). Further, speech generation, e.g., for voice conversion, is made possible by additional excitation module, in a separate or in a joint modeling manner to spectral module, and by WaveNet vocoder. The experimental results demonstrate that the proposed framework is capable of achieving a mean opinion score of 3.84, a character error rate of 0.21, and a speaker similarity score of 3.91 in the ZeroSpeech 2020 challenge results of 2019-task. Future work includes joint cyclic spectral and excitation modeling with separate latent spaces, and an improvement of neural vocoder fine-tuning.

7. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 17H06101 and JST, CREST Grant Number JPMJCR19A3.

8. References

[1] J. Glass, "Towards unsupervised speech processing," in *Proc. ISSPA*, Montreal, Canada, Jul. 2012, pp. 1–4.

[2] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose *et al.*, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 8111–8115.

[3] P. K. Muthukumar and A. W. Black, "Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 2594–2598.

[4] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3169–3173.

[5] M. Versteegh, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015: Proposed approaches and results," in *Proc. SLTU*, Yogyakarta, Indonesia, May 2016, pp. 67–72.

[6] L. Ondel, L. Burget, and J. Černockỳ, "Variational inference for acoustic unit discovery," *Procedia Comput. Sci.*, vol. 81, pp. 80–86, 2016.

[7] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 323–330.

[8] M. Heck, S. Sakti, and S. Nakamura, "Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 740–746.

[9] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel *et al.*, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking rosetta" JSALT 2017 workshop," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 4979–4983.

[10] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black *et al.*, "The zero resource speech challenge 2019: TTS without T," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1088–1092.

[11] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2041–2053, 2019.

[12] R. Eloff, A. Nortje, B. van Niekerk, A. Govender, L. Nortje, A. Pretorius, E. van Biljon, E. van der Westhuizen, L. van Staden, and H. Kamper, "Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1103–1107.

[13] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1118–1122.

[14] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. of the Acoust. Soc. of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[15] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 1781–1785.

[16] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task (ii): Resistance to noise," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 915–919.

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

[18] 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.

[19] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 674–678.

[20] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 9, pp. 1432–1443, 2019.

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

[22] 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.

[23] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1118–1122.

[24] 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 *Proc. EUSIPCO*, A Coruña, Spain, Sep. 2019, pp. 1–5.

[25] 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.

[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] C.-C. Hsu, "Pyworldvocoder - a python wrapper for world vocoder." [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[28] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 1043–1046.

[29] S. Sakti, R. Maia, S. Sakai, T. Shimizu, and S. Nakamura, "Development of HMM-based Indonesian speech synthesis," in *Proc. O-COCOSDA*, Kyoto, Japan, Nov. 2008, pp. 215–220.

[30] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project," in *Proc. TCAST*, Hyderabad, India, Jan. 2008, pp. 19–24.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[32] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[33] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Adv. NIPS*, Barcelona, Spain, Dec. 2016, pp. 901–909.

[34] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Efficient shallow WaveNet vocoder using multiple samples output based on Laplacian distribution and linear prediction," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 7204–7208.

[35] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 712–718.

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