Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder

Tuan Vu Ho¹ and Masato Akagi¹

¹Japan Advanced Institute of Science and Technology
tuanvu.ho@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

This paper proposes a hierarchical latent embedding structure for Vector Quantized Variational Autoencoder (VQVAE) to improve the performance of the non-parallel voice conversion (NPVC) model. Previous studies on NPVC based on vanilla VQVAE use a single codebook to encode the linguistic information at a fixed temporal scale. However, the linguistic structure contains different semantic levels (e.g., phoneme, syllable, word) that span at various temporal scales. Therefore, the converted speech may contain unnatural pronunciations which can degrade the naturalness of speech. To tackle this problem, we propose to use the hierarchical latent embedding structure which comprises several vector quantization blocks operating at different temporal scales. When trained with a multi-speaker database, our proposed model can encode the voice characteristics into the speaker embedding vector, which can be used in one-shot learning settings. Results from objective and subjective tests indicate that our proposed model outperforms the conventional VQVAE based model in both intra-lingual and cross-lingual conversion tasks. The official results from Voice Conversion Challenge 2020 reveal that our proposed model achieved the highest naturalness performance among autoencoder based models in both tasks. Our implementation is being made available at ¹.

Index Terms: Voice Conversion Challenge 2020, cross-lingual, variational autoencoder, hierarchical structure.

1. Introduction

Voice conversion (VC) is a subset of voice transformation method for altering speaker characteristics while preserving the linguistic information [1]. Conventionally, VC can be seen as a mapping problem between source waveform and target waveform [2]. This perspective requires learning a mapping function using parallel training data, in which the source and target waveform shares the same linguistic information. However, parallel training data cannot be collected in some situations such as in cross-lingual VC. Therefore, VC methods for non-parallel data are increasingly gaining more attention in recent years.

One of the straightforward methods for non-parallel VC (NPVC) is to concatenate speech recognition (ASR) with text-to-speech (TTS) model [3, 4, 5]. These methods often achieve the highest performance with highly natural converted speech [6]. However, both the ASR and TTS models must be trained on an enormous amount of transcribed speech data, which is often very expensive to construct. This constraint limits the applicability of the ASR-TTS approach in a practical situation.

In contrast, NPVC based on deep generative model such as Generative Adversarial Network (GAN) and Variational Autoencoder (VAE) can be trained without transcribed data.

Therefore, this type of NPVC model can be easily constructed from scratch using vastly available of untranscribed speech, thus reducing the development cost. With the recent advances of deep generative model, state-of-art GAN based VC [7, 8, 9] and VAE based VC [10, 11] have narrowed down the performance gap with ASR-TTS approaches. Although GANs come with a nice theoretical justification that the generated data should match the distribution of true data, it is widely known that the adversarial training is fragile and unstable. Moreover, while there are many studies on GAN-based VC, neither of them give strong evidence that the data distribution learned by Discriminator corresponds to human speech perception. In contrast, VAE can be easily trained. However, the VAE often suffers from the posterior collapse problem caused by Kullback-Leibler divergence (KLD) [12], which reduces the useful information received by the decoder for speech reconstruction.

A recently proposed Vector Quantized VAE (VQVAE) [13] model with discrete latent space avoids the posterior collapse problem by not optimizing the KLD but learning the categorical prior instead. Since linguistic information can be regarded as categorical data, discrete latent space is suitable to represent linguistic information. The VQVAE has been successfully applied in various speech processing tasks [14, 15, 16]. However, the linguistic information conveys different levels of semantic structure that spans at different temporal scales (e.g phonemes, syllables). Therefore, a single vector quantizer operating at a fixed temporal scale is inefficient to capture various levels of semantic structure, hence reducing the naturalness of converted speech. To tackle this problem, we propose the hierarchical latent embedding VQVAE (HLE-VQVAE) to capture the linguistic information at various temporal scales. As shown in the next sections, the proposed scheme can improve the performance of VC system and provide highly natural converted speech for both intra-lingual and cross-lingual tasks.

2. Baseline method

2.1. Vector Quantized Variational Autoencoder based Voice Conversion

The VQVAE can be regarded as a communication systems, in which the input feature vector x is compacted into latent vector...
z by a non-linear transformation (encoder). The latent vector z is then quantized to discrete variable q based on its distance to pseudo-vectors in the codebook e_k, k ∈ 1…K

\[ q = e_k \text{ where } k = \arg \min_k |z - e_k| \] (1)

Finally, the decoder reconstructs the input vector from the discrete latent vector q and one-hot speaker embedding s_m of the source speaker. The latent codebook is updated simultaneously with other parameters of the model during training process. Due to the use of argmin function in quantization process, the computation graph is disconnected and the model cannot be trained with back-propagation. Therefore, straight-through reparameterization trick [13] is used to avoid this problem:

\[ z = \text{Enc}(x) \]
\[ q = \text{Quantize}(z) \]
\[ q_{st} = z + sg(q - z) \]
\[ x_{dec} = \text{Dec}(q_{st}, s_m) \] (2)

where x_{dec} is the reconstructed feature vector, q_{st} is straight-through variable from which gradient is copied to z, Enc(·) is the encoder function, Dec(·) is the decoder function, Quantize(·) is quantization function, and sg(·) is the stop-gradient operator. The model parameters are obtained by minimizing the following objective function:

\[ \mathcal{L}_\text{QVAE} = \|x - x_{dec}\|^2_2 + \|z - sg(q)\|^2_2 + \beta \|sg(z) - q\|^2_2 \] (3)

where \|x - x_{dec}\|^2_2 is the reconstruction loss, \|z - sg(q)\|^2_2 is the quantization loss, \|sg(z) - q\|^2_2 is the codebook loss, and \( \beta \) is a hyper-parameter to control the reluctance to change of the codebook loss.

At the inference step, providing the source mel-cepstrum and the speaker embedding of target speaker, the model outputs the converted mel-cepstrum containing the target voice characteristics. The overview of conventional QVAE based VC is shown in Fig. 1.

3. Proposed method

In this section, the QVAE model with hierarchical latent embedding structure (HLE-QVAE) is proposed. Following this, we also describe our method to adapt the intra-lingual VC model for cross-lingual VC task.

3.1. Hierarchical Latent Embedding QVAE

In conventional QVAE, input data are encoded to latent embedding variable at a fixed temporal scale. However, the semantic structure of speech contains different levels that span across different temporal scales. Inspired by the work of [17] on image generation, a hierarchical structure is used to better capture different information at different temporal scales.

The overview of our proposed model with 3 stages of hierarchical structure is shown in Fig. 2. Each stage consists of an encoder network, a quantizer and a decoder network. At stage, the encoder downsamples its input and the decoder upsamples its input by the same factor. Except for the top encoder, each encoder output is split along channel dimension into 2 parts: the latent variable z_n and hidden variable u_n. The latent variable z_n is then discretized to q_n, while hidden variable u_n is passed to the next encoder. On the decoder side, the discrete latent variable of the current stage is concatenated with the decoded hidden variable v_n from previous stage before passing through the decoder network. Similar to vanilla QVAE based VC, each decoder in the proposed model is conditioned by the same speaker embedding s_m.

At the training phase, providing the mel-cepstral sequence as input, the model is trained to minimize the following objective functions:

\[ \mathcal{L}_{\text{HLE-QVAE}} = \|x - x_{dec}\|^2_2 + \sum_{n=1}^{N} \left( \|z_n - sg(q_n)\|^2_2 + \beta \|sg(z_n) - q_n\|^2_2 \right) \] (4)

where \( N \) is the number of hierarchical stage, \( \beta \) is set to 0.25 in our study.

3.2. Learnable speaker embedding

One-hot speaker embedding has the drawback that the number of speaker embedding is fixed by the dimension of the one-hot vector. Follow the study [18], our proposed model uses learnable speaker embeddings which are jointly optimized with other models parameters during the training phase by using back-propagation. The speaker index is used to select the corresponding speaker embedding in speaker codebook.

3.3. Cross-lingual adaptation

The advantage of our VC scheme is that only the target speaker embedding is needed to mimic the voice characteristics of the target. To obtain the target speaker embedding of foreign language, the latent codebook from the pretrained intra-lingual
model and the random-initialized speaker embedding are fine-tuned on the target data. After the target speaker embedding is obtained, the model generates converted mel-cepstrum using the similar inference step described in Section 2.1.

4. Experiments

In this section, we describe the results of the objective and subjective measurements to explain our model selection for Voice Conversion Challenge 2020 (VCC2020). Then, we show the official results of the VCC2020 to demonstrate the performance of our submitted system. To conveniently compared the models that we tested, we name the models as follows:

- **VQVAE**: vanilla VQVAE model with 1 stage of quantization.
- **HLE-VQVAE-2**: the proposed HLE-VQVAE model with 2 stage of quantization.
- **HLE-VQVAE-3**: the proposed HLE-VQVAE model with 3 stage of quantization.

4.1. Dataset

The VCC2020 training set consists of 4 source English speakers, 4 target English speakers, and 2 target speakers of each foreign language (Finnish, German, and Mandarin). Each speaker in the VCC2020 training set utters a sentence set consisting of 70 sentences. Besides, a subset of the CSTR VCTK dataset [19] containing all utterances from the first 100 speakers was used in combination with the VCC2020 training set to train the models. We directly used VCC2020 evaluation data for testing. In the pre-processing step, the audio file is down-sampled to 24 kHz and normalized to $[-1.0, 1.0]$ range. Then, an 80-dimension mel-spectrogram is extracted using the Short-time Fourier Transform (STFT) and mel-filterbank. The window length of STFT is set to 2048 and the hop-length is 500. The mel-spectrum is transformed into mel-cepstrum by applying Inverse Discrete Fourier Transform on the log-magnitude mel-spectrum. To reconstruct the waveform, we used the Parallel WaveGAN vocoder [20] which has been trained on the VCTK dataset for 1000k iterations.

4.2. Implementation details

For the proposed model, the downsampling and upsampling factors for each encoder and decoder are set to 2. The codebook at each stage contains 128 atoms of 64 dimensions. The encoder and decoder are implemented by stacking multiple non-causal dilated WaveNet-like structures [21] as shown in Fig. 3.

For the baseline model, we implemented a vanilla VQVAE model with a similar encoder and decoder structure as the proposed model. As the baseline model has 1 stage, the feature vector is downsampled by the factor of 2 before quantized using a codebook containing 256 atoms of 64 dimensions.

The dimension of speaker embedding in all models is 16. The model parameters were optimized using Adam [22] with learning rate of $0.0005$ and gradually reduced to $0.0002$ after 10 epochs. For intra-lingual task, all models were trained with 200 epochs with batch size 32. For cross-lingual adaptation, all models are fine-tuned with 1000 epochs for each target speaker.

4.3. Visualization of Speaker Embedding

Principle component analysis (PCA) is used to visualize the learned speaker embedding. As shown in Fig. 4, it can be seen that the speakers are well clustered by genders. This indicates that the speaker embedding can encode meaningful voice characteristics of the speakers without any additional speaker information.

4.4. Objective test

The modulation spectrum (MS) of the parameter temporal trajectory is one of the well-known metrics to measure the quality of synthetic speech [23]. The MS of converted mel-cepstrum is measured by taking Discrete Fourier transformation on each cepstral sequence. Then, root-mean-squared errors (RMSEs) between the logarithmic MS of natural speech and converted speech from different models are calculated. It should be expected that the lower the RMSEs, the better quality of converted speech. We measure the RMSEs on all the converted utterances and average across all mel channels and modulation frequencies. The results shown in Table 1 indicate that the mel-spectral sequences obtained from our proposed models are closest to the target speaker in terms of MS. In particular, the HLE-VQVAE-3 outperformed the HLE-VQVAE-2 in most cases except for cross-lingual and cross-gender VC.

4.5. Subjective test

We conducted the AB naturalness test and ABX similarity test to compare the performance of 3 models. Due to time constraint, we only tested the converted speech between 2 source speakers (SEF1 and SEM1) and 4 target speakers (English speakers: TEF1 and TEM1, German speakers: TGF1 and TGM1). Two sentences (E30001 and E30002) were selected from each source-target pairs to form the listening test set. Therefore, the listening test set consisted of 48 converted utterance pairs (2 sentences × 8 source-target speaker pairs × 3 model pairs). As for reference stimuli in the ABX similarity test, we randomly selected the original utterances of the target speakers from the VCC2020 training set. There were 12 partic-

<table>
<thead>
<tr>
<th>Method</th>
<th>VQVAE</th>
<th>HLE-VQVAE-2</th>
<th>HLE-VQVAE-3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intra-lingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same-gender</td>
<td>0.267</td>
<td>0.258</td>
<td>0.238</td>
</tr>
<tr>
<td>Cross-gender</td>
<td>0.313</td>
<td>0.302</td>
<td>0.280</td>
</tr>
<tr>
<td><strong>Cross-lingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same-gender</td>
<td>0.431</td>
<td>0.477</td>
<td>0.472</td>
</tr>
<tr>
<td>Cross-gender</td>
<td>0.454</td>
<td>0.414</td>
<td>0.430</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.375</td>
<td>0.364</td>
<td>0.359</td>
</tr>
</tbody>
</table>

Table 1: Comparison of RMSE between target and converted logarithmic MS averaged over all mel channels and modulation frequencies. Smallest RMSE value is highlighted in bold.

![2D PCA visualization of learned speaker embedding](image.png)

Figure 4: 2D PCA visualization of learned speaker embedding by HLE-VQVAE-3 model from VCC2020 dataset (VCC2020) and VCTK dataset (VCTK male and VCTK female). The horizontal and vertical axes are the first and second principal components, respectively.
This study is supported by NII-CRIS and Grant-in-Aid for Scientific Research (Grant number: 20H04207).

5. Conclusion

This paper has proposed a VC model based on VQVAE with a hierarchical latent structure to improve the quality of converted speech. We have shown that our proposed model outperformed the vanilla VQVAE based VC model in both objective and subjective evaluation. Results from the official listening test in VCC2020 shown that our submitted HLE-VQVAE-3 model was comparable with the average performance of PPG/ASR-TTS models and superior to other autoencoder VC models in term of naturalness. However, there are still rooms to improve the similarity performance of the proposed model. Since our proposed model works purely in the acoustic domain, it can be easily adapt to other VC tasks such as speech enhancement, one-shot VC, etc.

6. Acknowledgments

This study is supported by NII-CRIS and Grant-in-Aid for Scientific Research (Grant number: 20H04207).
7. References


