Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020
30 October 2020, Shanghai, China

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

*Tuan Vu Ho*[1] *and Masato Akagi*[1]

[1]Japan 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 auoencoder, 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.

[1]Our implementation: https://github.com/tuanvu92/VCC2020

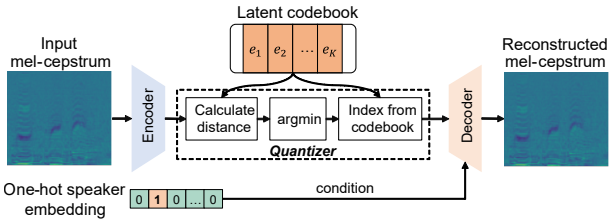


Figure 1: *Conventional VQVAE-based VC.*

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 $\mathbf{x}$ is compacted into latent vector



10.21437/VCC_BC.2020-20

$\mathbf{z}$ by a non-linear transformation (encoder). The latent vector $\mathbf{z}$ is then quantized to discrete variable $\mathbf{q}$ based on its distance to pseudo-vectors in the codebook $\mathbf{e}_k, k \in 1...K$.

$$\mathbf{q} = \mathbf{e}_k \text{ where } k = \underset{k}{argmin} \, \|\mathbf{z} - \mathbf{e}_j\| \tag{1}$$

Finally, the decoder reconstructs the input vector from the discrete latent vector $\mathbf{q}$ and one-hot speaker embedding $\mathbf{s}_m$ 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:

$$\begin{aligned} \mathbf{z} &= Enc(\mathbf{x}) \\ \mathbf{q} &= Quantize(\mathbf{z}) \\ \mathbf{q}_{st} &= \mathbf{z} + sg(\mathbf{q} - \mathbf{z}) \\ \mathbf{x}_{dec} &= Dec(\mathbf{q}_{st}, \mathbf{s}_m) \end{aligned} \tag{2}$$

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

$$\begin{aligned} \mathcal{L}_{\text{VQVAE}} = \|\mathbf{x} - \mathbf{x}_{dec})\|_2^2 + \|\mathbf{z} - sg(\mathbf{q})\|_2^2 + \\ \beta \, \|sg(\mathbf{z}) - \mathbf{q}\|_2^2 \end{aligned} \tag{3}$$

where $\|\mathbf{x} - \mathbf{x}_{dec})\|_2^2$ is the reconstruction loss, $\|\mathbf{z} - sg(\mathbf{q})\|_2^2$ is the quantization loss, $\|sg(\mathbf{z}) - \mathbf{q}\|_2^2$ 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 VQVAE based VC is shown in Fig. 1.

## 3. Proposed method

In this section, the VQVAE model with hierarchical latent embedding structure (HLE-VQVAE) 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 VQVAE

In conventional VQVAE, 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 scale. 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 $n$, 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 $\mathbf{z}_n$ and hidden variable $\mathbf{u}_n$. The latent variable $\mathbf{z}_n$ is then discretized to $\mathbf{q}_n$, while hidden variable

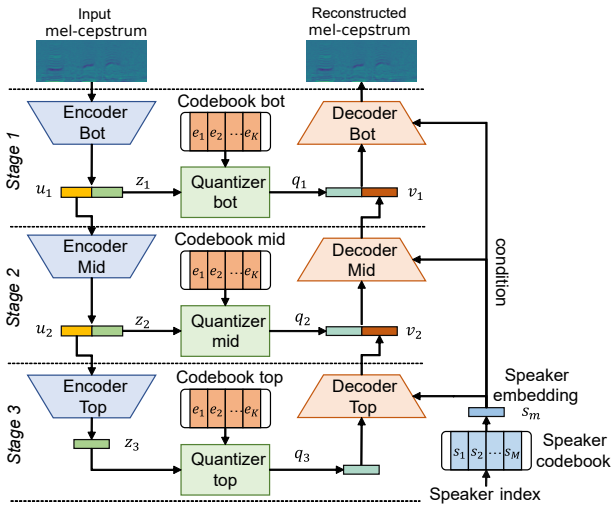


Figure 2: *Diagram of our submitted 3-stage HLE-VQVAE.*

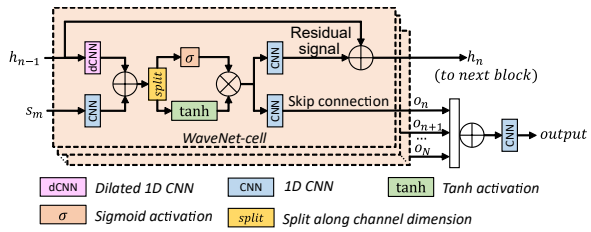


Figure 3: *Stacks of non-causal dilated Wavenet-like structure in encoder and decoder.*

$\mathbf{u}_n$ 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 $\mathbf{v}_n$ from previous stage before passing through the decoder network. Similar to vanilla VQVAE based VC, each decoder in the proposed model is conditioned by the same speaker embedding $\mathbf{s}_m$.

At the training phase, providing the mel-cepstral sequence with speaker embedding of source speaker, the model is trained to minimize the following objective functions:

$$\begin{aligned} \mathcal{L}_{\text{HLE-VQVAE}} = \|\mathbf{x} - \mathbf{x}_{dec})\|_2^2 + \\ \sum_{n=1}^{N} \left( \|\mathbf{z}_n - sg(\mathbf{q}_n)\|_2^2 + \beta \, \|sg(\mathbf{z}_n) - \mathbf{q}_n\|_2^2 \right) \end{aligned} \tag{4}$$

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 300. 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 32 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 clusterized by genders. This indicates that the speaker embedding can encode meaningful voice characteristics of the speakers without any additional speaker information.

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.*

| Method | | VQVAE | HLE-VQVAE-2 | HLE-VQVAE-3 |
|---|---|---|---|---|
| Intra-lingual | Same-gender | 0.267 | 0.258 | **0.238** |
| | Cross-gender | 0.313 | 0.302 | **0.280** |
| Cross-lingual | Same-gender | 0.431 | 0.477 | **0.472** |
| | Cross-gender | 0.434 | **0.414** | 0.430 |
| Average | | 0.375 | 0.364 | **0.359** |

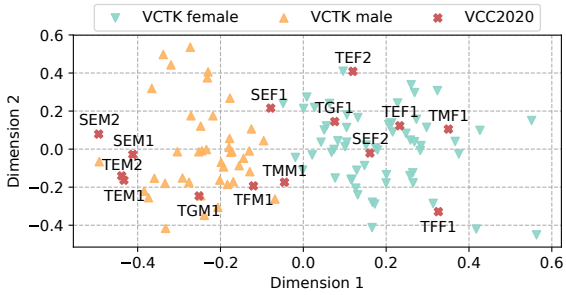


Figure 4: *2D PCA visualization of learned speaker embedding by HLE-VQVAE-3 model from VCC2020 dataset (**VCC2020**) and VCTK dataset (**VCTK male** and **VCTK female**). The horizontal and vertical axes are the first and second principal components, respectively.*

## 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 target 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 $\times$ 8 source-target speaker pairs $\times$ 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-

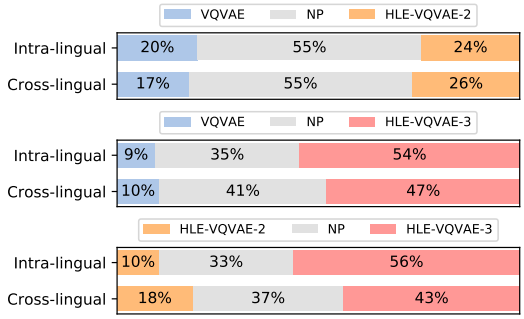


Figure 5: *Preference score of AB naturalness test.* ***NP*** *means no preference.*

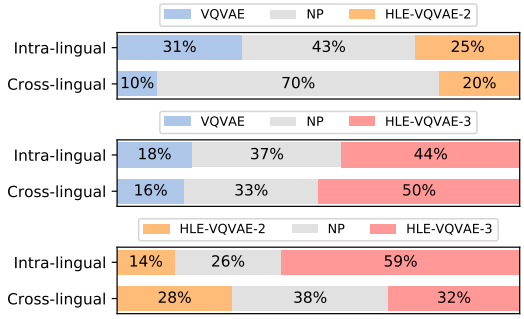


Figure 6: *Preference score of ABX speaker similarity test.* ***NP*** *means no preference.*

ipants with good English proficiency joined both listening tests. Each participant rated 24 random pairs of converted utterances for each test.

The results of the AB naturalness test are shown in Fig. 5. It can be seen that the HLE-VQVAE-3 model outperformed the VQVAE and HLE-VQVAE-2 in terms of naturalness performance for both intra-lingual and cross-lingual conversion. The result of the ABX similarity test shown in Fig. 6 indicates that the HLE-VQVAE-3 model was slightly better than the HLE-VQVAE-2 model in cross-lingual VC. In other cases, the HLE-VQVAE-3 significantly outperformed the HLE-VQVAE-2 and VQVAE model. These results were also aligned with the objective measurement shown in Section 4.4.

### 4.6. Voice Conversion Challenge 2020 results

The VCC2020 organizers conducted 2 large-scale listening tests to evaluate the speech naturalness and speaker similarity of converted speech [24]. In the naturalness test, listeners were asked to evaluate voice quality on a scale from 1 (Bad) to 5 (Excellent). In the speaker similarity test, listeners were asked to judge whether or not the converted and target utterances were spoken by the same person, and then evaluate using a 4-point scale that varies from "Different (sure)" to "Same (sure)".

To conveniently evaluate the performance of our submitted systems, we compare the score of our submitted system with different types of VC models, which is named as follows:

- **PPG/ASR-TTS**: text-dependent models including Phonetic Posteriorgram VC [3], concatenation of speech recognition (ASR) and text-to-speech (TTS) system , and leveraging TTS for VC methods [25]. Speech transcription is required to train these types of model.
- **AE**: Autoencoder based models including VQVAE, CycleVAE [26], AutoVC [27], and one-shot VC [28]. These types of VC models share the same paradigm as our submitted model.

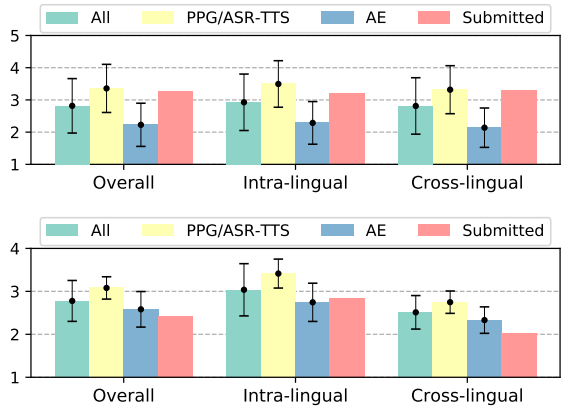


Figure 7: *Average MOS score (top) and similarity score (bottom) with standard deviation of English listeners from all models (**All**), text-dependent models (**PPG/ASR-TTS**), autoencoder based models (**AE**), and our submitted model (**Submitted**).*

The results of naturalness MOS score and similarity score are summarized in Fig. 7. In both intra-lingual VC and cross-lingual VC tasks, the naturalness performance of our submitted model is significantly higher than the average of autoencoder based models and is comparable with the average of PPG/ASR-TTS based models. In terms of similarity performance, our submitted model still achieves a higher score than the average of autoencoder based VC in intra-lingual VC task. However, there is a decline in similarity score of our submitted model in cross-lingual VC task. We speculate that this might be due to the lack of an explicit input $F_0$ information in our VC model. Since the mean and the variance of $F_0$ is one of the important cues for speaker individuality [29], the speaker embedding may encode the $F_0$ statistics embedded in the mel-cepstrum. However, since different languages may have distinctive shape of $F_0$ contour which is reflected in $F_0$ statistics, the estimation of speaker embedding of foreign speaker will be biased. By providing the decoder with an explicit $F_0$ information, the speaker embedding will be freed from capturing $F_0$ mean and variance, hence increasing the accuracy of modeling speaker characteristics.

## 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


[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, 2017.

[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, pp. 2222–2235, 2007.

[3] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. Interspeech 2018*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190

[4] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020.

[5] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016.

[6] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z.-H. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, 2018.

[7] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *INTERSPEECH*, 2017.

[8] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[9] 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.

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

[11] T. Dinh, A. Kain, and K. Tjaden, "Using a manifold vocoder for spectral voice and style conversion." in *INTERSPEECH*, 2019, pp. 1388–1392.

[12] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue pca directions (by accident)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 406–12 415.

[13] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[14] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (vq-vae) autoregressive neural $f_0$ model for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 157–170, 2020.

[15] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge," *arXiv preprint arXiv:2005.09409*, 2020.

[16] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion." in *INTERSPEECH*, 2019, pp. 724–728.

[17] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 866–14 876.

[18] T. V. Ho and M. Akagi, "Non-parallel voice conversion with controllable speaker individuality using variational autoencoder," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 106–111, 2019.

[19] C. Veaux, J. Yamagishi, and K. Macdonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[20] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[21] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *ArXiv*, vol. abs/1609.03499, 2016.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[23] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[24] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020 — intra-lingual semi-parallel and cross-lingual voice conversion —," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*. ISCA, 2020.

[25] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *ArXiv*, vol. abs/1912.06813, 2019.

[26] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-Parallel Voice Conversion with Cyclic Variational Autoencoder," in *Proc. Interspeech 2019*, 2019, pp. 674–678.

[27] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Zero-shot voice style transfer with only autoencoder loss," *ArXiv*, vol. abs/1905.05879, 2019.

[28] J.-C. Chou, C. chieh Yeh, and H. yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *ArXiv*, vol. abs/1904.05742, 2019.

[29] M. Akagi and T. Ienaga, "Speaker individuality in fundamental frequency contours and its control," *J. Acoust. Soc. Jpn. (E)*, vol. 18, no. 4, pp. 73–80, 1997.