



VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net architecture

Da-Yi Wu, Yen-Hao Chen, Hung-Yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University

{r07922119, r07921112, , hungyilee} @ntu.edu.tw

Abstract

Voice conversion (VC) is a task that transforms the source speaker’s timbre, accent, and tones in audio into another one’s while preserving the linguistic content. It is still a challenging work, especially in a one-shot setting. Auto-encoder-based VC methods disentangle the speaker and the content in input speech without explicit information about the speaker’s identity, so these methods can further generalize to unseen speakers. The disentangle capability is achieved by vector quantization (VQ), adversarial training, or instance normalization (IN). However, the imperfect disentanglement may harm the quality of output speech. In this work, to further improve audio quality, we use the U-Net architecture within an auto-encoder-based VC system. We find that to leverage the U-Net architecture, a strong information bottleneck is necessary. The VQ-based method, which quantizes the latent vectors, can serve the purpose. The objective and the subjective evaluations show that the proposed method performs well in both audio naturalness and speaker similarity.

Index Terms: voice conversion, vector quantization, skip-connection, disentangled representations

1. Introduction

Voice Conversion (VC) task is to convert the voice of the source speaker to the voice of a target speaker without losing the linguistic information in source speech. To imitate the target speaker, a VC system should modify the tone, accent, and vocalization of the source speaker. In tradition, VC focuses on one-one or many-to-one speaker transformation with parallel data, and it is treated as a statistical problem [1, 2, 3]. However, the above methods need parallel data, which is challenging to collect.

Recently, many-to-many unparallelled VC has been studied. Generative adversarial network (GAN) [4] and its variants, like cycleGAN [5] and starGAN [6], are classic models to tackle unparallelled, many-to-many VC problems. Flow-based models like blow [7] have also been studied, and they transform waveforms directly instead of using acoustic features. These models directly convert the voice without feature disentangling behavior, achieving satisfactory audio quality.

Other works [8, 9, 10, 11, 12, 13] attempt to disentangle the speaker’s unit and content unit in the embedding space. These methods convert the voice by replacing the speaker embedding passed through the decoder. With a pre-trained speaker encoder, AutoVC [14] applies a vanilla auto-encoder and constrains the size of the latent representations to lead the encoder to extract the content information from the audio. Chou et al. [15] bring the idea from the image style-transfer mission, referring the speaker to the style of speech; it maps different speakers to different values of mean and standard deviation, and utilizes AdaIN [16] to transform the speaker in speech. VQVC [17] ap-

plies vector quantization [18] technique to extract the content information, and learn to represent the speaker information by the difference between continuous space and the discrete codes. Due to these models’ ability to disentangle latent space, these approaches can synthesize the voice of the unseen speakers, and can even achieve one-shot VC, that is, the model synthesizes converted audio using only one sentence of each source and target speaker during the inference phase.

To disentangle latent space, the model usually needs a strong bottleneck, which is constrained by GAN [19], layer dimension [14], IN [15], or VQ [17], and due to the constraints applied on the model, the audio quality is sacrificed. To deal with the above issue, we propose VQVC+, a U-Net architecture combining VQ and IN. U-Net connects each encoder layer and its corresponding decoder, and it has recently shown superior performance as a spectrogram generator [20, 21, 22]. However, U-Net [23] is seldom used in voice conversion, and the reason is that its reconstruction capability is too well to make the model lose the ability to disentangle. In this paper, we find out that VQ can form a strong information bottleneck to prevent U-Net from overfitting on the reconstruction task, and with U-Net architecture, VQVC+ can hence synthesize high-quality audio. Also, we compare our method with AutoVC [14] and Chou [15]. The subjective evaluations show that the proposed method reaches a state-of-the-art result in one-shot VC.

2. Methods

2.1. VQVC

VQVC [17] is a one-shot voice conversion system with self-reconstruction loss. The core idea is: *the content information can be represented by discrete codes* [24, 25], *and the speaker information can be viewed as the difference between the continuous representations and the discrete codes*.

As shown in Figure 1, an auto-encoder architecture is used. \mathcal{X} is our whole training set. We denote $\mathbf{X} \in \mathcal{X}$ as an audio segment, represented as a sequence of acoustic features, $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T\}$, where T denotes the audio duration. We denote *enc* as the encoder, *dec* as the decoder, \mathcal{Q} as the quantization codebook, and *Quantize* as the quantization function. Given the audio segment \mathbf{X} , the continuous latent representation $\mathbf{V} \in R^{F \times T}$, the content embedding $\mathbf{C} \in R^{F \times T}$, and the speaker embedding $\mathbf{S} \in R^{F \times T}$ can be derived as

$$\begin{aligned} \mathbf{V} &= \text{enc}(\mathbf{X}), \\ \mathbf{C} &= \text{Quantize}(\mathbf{V}), \\ \mathbf{s} &= \mathbb{E}_t[\mathbf{V} - \mathbf{C}], \quad \mathbf{S} = \underbrace{\{\mathbf{s}, \mathbf{s}, \dots, \mathbf{s}\}}_{T \text{ times}}, \end{aligned} \tag{1}$$

where F is embedding size, and

$$\text{Quantize}(\mathbf{V}) = \{\mathbf{q}_0, \mathbf{q}_1, \dots, \mathbf{q}_T\}, \quad \mathbf{q}_j = \arg \min_{\mathbf{q} \in \mathcal{Q}} (\|\mathbf{v}_j - \mathbf{q}\|_2^2).$$

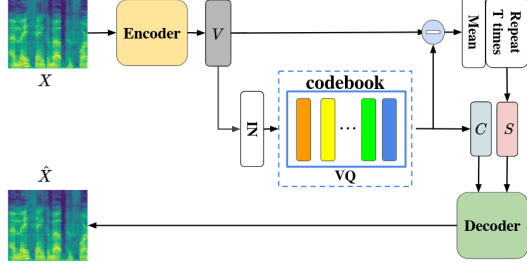


Figure 1: The VQVC architecture. VQ is the vector quantization layer, and IN is the instance normalization layer. VQVC applies IN+VQ layers to separate the content and the speaker information to achieve voice conversion.

Note that the dimension of any vector v_j is equal to the dimension of a code in \mathcal{Q} . The instance normalization (IN) [26] layer is added before the quantization, which is indispensable for good performance. We expect that the speaker information is a kind of global information of the speech. Hence, we derive s by subtracting C from V and then take expectation \mathbb{E}_t on the utterance duration, representing the global information of the audio segment. Afterwards, we get S by repeating s for T times and concatenate them to enforce the dimensions of C and S to be the same. Then, S is added back to C , passed through the decoder, and at last we get the reconstruction

$$\hat{X} = dec(C + S). \quad (2)$$

In the training phase, the reconstruction loss can be written as Equation (3):

$$L_{rec}(\mathcal{Q}, \theta_{enc}, \theta_{dec}) = \mathbb{E}_{X \in \mathcal{X}} [\| \hat{X} - X \|_1]. \quad (3)$$

In addition, the latent loss L_{latent} is added as Equation (4), which minimizes the distance between the discrete codes and the continuous embedding. We denote IN as instance normalization layer.

$$L_{latent}(\theta_{enc}) = \mathbb{E}_t [\| IN(V) - C \|_2^2]. \quad (4)$$

The whole loss can be written as Equation (5):

$$L = L_{rec} + \lambda L_{latent}. \quad (5)$$

During the inference phase, the content embedding C , and the speaker embedding S would be extracted from different speakers.

2.2. VQVC+

Although VQVC can disentangle the linguistic content and the speaker information well, the synthesized audio quality still has room for improvement. VQVC synthesizes the audio that matches the target speaker's characteristic, but the vocalization of the audio is vague. We attribute the problem to the information loss induced by the vector quantization, which makes the decoder unable to reconstruct the content properly. Thus, to improve synthesis quality, we apply the U-Net architecture, which has shown superior performance as a spectrogram generator [23], on VQVC, and we call this new models VQVC+.

Figure 2 illustrates the whole architecture of VQVC+. The encoder is composed of three VQ down-conv modules, which is a variant of VQVC encoder; the decoder is composed of three

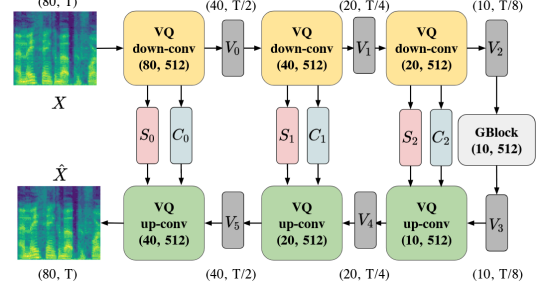


Figure 2: The VQVC+ architecture. VQVC+ applies the U-Net architecture to improve quality, and each sub-module in the encoder is a variant of the VQVC encoder. Quantized output C and the speaker embedding S are skip-connected to the decoder instead of the continuous embedding V .

VQ up-conv modules. To strength the content information received by the decoder layers, the content embedding C and the speaker embedding S are skip-connected into the corresponding decoder layer, which is in a similar light with U-Net.

2.2.1. VQ down-conv module

As shown in Figure 3, VQ down-conv is comprised of two 3×1 kernel 1D-convolution layers, an IN layer, and a vector quantization layer. The Conv1d- c_1 - c_2 - N layer indicates the 1D-convolution layer, whose input channel and the output channel are c_1 and c_2 respectively, and N denotes the stride.

VQ down-conv(c_{in}, c_h) takes a matrix with dimension (c_{in}, T) as input and outputs three components, V , C , and S . V is the embedding of continuous space directly get from the convolution block; C is the quantized matrix of V passing through IN and VQ; S is the speaker embedding as we mentioned in Equation (1). The dimensions of V , C , and S all are ($c_{in}/2, T/2$).

2.2.2. VQ up-conv module

As shown in Figure 4, VQ up-conv takes the output of previous layer, V , and C , S generated from the corresponding encoder layer as inputs. Embeddings are added and upsampled by a factor of 2 within both frequency and time domain. VQ up-conv contains three main components, Group Norm Block (GBlock) [22], TimeUpsampling, and FreqUpsampling.

GBlock consists of two 3×1 kernel 1D-convolution layers and groupnorm following with LeakyReLU, where the size of its input and output are the same.

As shown in Figure 5, the TimeUpsampling module duplicates each vector twice to expand the time dimension; the FreqUpsampling module emphasizes the importance of low-frequency area in mel-spectrogram, so it generates the high-frequency part using its low-frequency part and concatenates them as its output.

C and S are first added and pass through the GBlock, and we add V from the previous layer on it afterward. Then, it passes through the two upsampling modules to get the output.

2.2.3. U-Net

Our architecture can be seen as a variant of U-Net. As shown in Figure 2, each VQ down-conv module generates its own V , C , S , where V is passed through the next VQ down-conv module, and C , S are passed through the corresponding

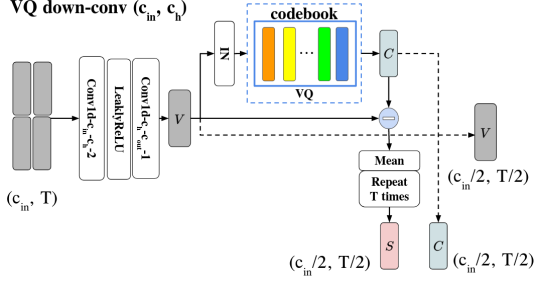


Figure 3: *VQ down-conv module*. The $\text{Conv1d-}c_1\text{-}c_2\text{-}N$ indicates a 1D-convolution layer with input channel c_1 , output channel c_2 , and stride N .

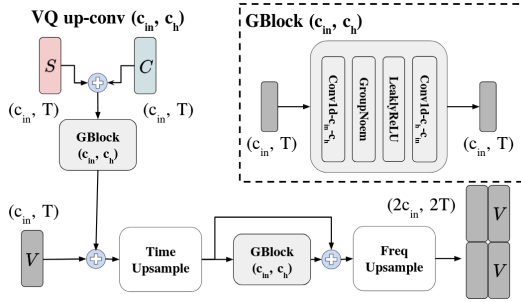


Figure 4: *VQ up-conv module and Group Norm Block (GBlock)*

VQ up-conv module in the decoder. Model is trained with latent loss L_{latent} for each layer and reconstruction loss L_{rec} . We assign equal weight λ in (5) for the latent loss L_{latent} of all layers during training.

3. Experimental Setup

3.1. Datasets

We conduct our experiment on the VCTK dataset[27], which contains about 46 hours of audio from 109 speakers, and there are about 500 sentences for each speaker. We select 20 speakers as our testing set, where we denote them as our unseen speakers. For each audio recording, we remove the silence and randomly choose 3 seconds for training. If the duration of the audio is less than 3 seconds, we repeat it into 3 seconds. Then, we convert the audio from 48000Hz into 22050Hz and perform STFT with 1024 STFT window size and 12.5 milliseconds hop size. Next, we transform the magnitude of the spectrograms to 80-bin mel-scale and take logarithm. To convert the mel-spectrogram back to the waveform, we apply the fast, high-quality pre-trained MelGAN vocoder[28].

3.2. Training details

We train the proposed model using ADAM optimizer with a 0.01 learning rate, and $\beta_1 = 0.9$, $\beta_2 = 0.999$. Our channel size in each *VQ down-conv* and *VQ up-conv* is 512, and our codebook size in each encoder layer is 64. We set the batch size to 32, and latent loss factor in E.q.4, λ , to 0.1. We train our model on 99 speakers for 200k iterations. Further details may be found in our implementation code¹.

¹<https://github.com/ericwudayi/SkipVQVC>

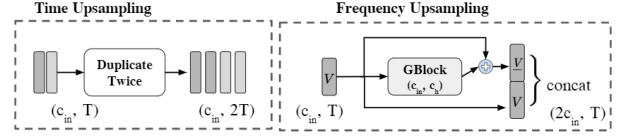


Figure 5: *Upsampling block*

4. Experiments

4.1. Content embedding

We first demonstrate the effects of IN and VQ in Figure 6. We perform t-SNE on V_0 and C_0 of 20 distinct speakers. Different colors represent different speakers. It is shown that for V_0 , the points with the same colors tend to be clustered, whereas there is no obvious group in C_0 .

To further verify that the speaker information is discarded by the IN and VQ layers, we train a speaker classifier based on content embedding. The lower speaker classifier accuracy, the less speaker information contains in content embedding. The classifier is composed of three 1D-convolution layers with 256 hidden nodes followed by a fully connected layer. We compare the result with different quantization settings. Here, Q32(64, 128, 256) represents VQVC+ that the codebook size of each encoder layer is 32(64, 128, 256); IN-only means there is no quantization module after the IN layers; VQVC [17] is the original model without skip-connection, and its codebook size is 128.

We conduct our experiment on the output of each encoder layer, C_0 , C_1 , and C_2 . As shown in Table 1, deeper layers have lower accuracy for every model, and the accuracy of the VQ models are apparently lower than IN-only. IN-only has 71.2% speaker identification rate on C_0 , which indicates that IN-only has no ability to disentangle the content information and the speaker information. We listen to the audio synthesized using IN-only, finding that most of them cannot perform conversion; IN-only just reconstructs the source audio. The codebook size of a model determines how strong the information bottleneck is. Models with smaller codebook size, like Q32, can achieve lower speaker identification rate, but it may lose more content information, which makes higher reconstruction error. On the other hand, models with larger codebook size, like Q256, can reconstruct the audio well, but it may leak some speaker information to its quantized code. We choose Q64 in the following experiments, which achieves an exceptional balance on reconstruction and disentanglement.

Table 1: *Accuracy of identifying speakers on the content embedding and the speaker embedding with different methods. VQVC is the model without skip-connection design. QN means that the size of codebook, Q, in VQVC+ is N. IN-only means no quantization in U-Net. L1Loss is the L1 reconstruction loss.*

Method	$C_0 / C_1 / C_2$ (%)	L1Loss
VQVC	16.0	0.262
Q32	19.5 / 11.8 / 6.8	0.210
Q64	23.2 / 16.6 / 7.0	0.188
Q128	33.3 / 17.0 / 10.3	0.180
Q256	35.8 / 18.1 / 12.5	0.165
IN-only	71.2 / 36.8 / 5	0.145

Table 2: Accuracy of identifying speakers on the speaker embedding S .

Method	$S_0/S_1/S_2$ (%)
VQVC	96.6
Q64	98.3 / 72.2 / 45.4
IN-only	97.4 / 80.1 / 23.1

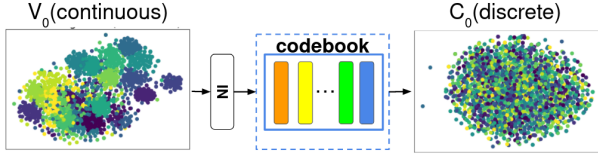


Figure 6: TSNE of embedding space before and after quantization, and different colors represent different speakers.

4.2. Speaker embedding

We do not explicitly add any speaker-related objective or constraint to the encoder, while the speaker embedding is learned properly due to the effective disentanglement of the quantized content embedding. To show that our extraction-based speaker embedding is speaker-correlated, we select 20 unseen speakers to generate their speaker embeddings. For each of S_0 , S_1 , and S_2 , we train a classifier to identify which speaker is it, and the classifier’s architecture is the same as we mentioned in Section 4.1. The higher the classifier accuracy, the better the speaker embedding is. The model used in this experiment is Q64, which is the same one mentioned in Section 4.1. The results are shown in Table 2. It presents that the speaker information in S_0 is learned very well. S_1 and S_2 have 72.2% and 45.4% accuracy, which indicates that the model extracts the speaker embeddings at lower resolution spaces.

4.3. Subjective evaluations

We conduct subjective tests to evaluate the converted audio’s sound quality and its speaker similarity to the target one. We choose AutoVC[14] and Chou[15] as our baseline models, and we train the models using the official code^{2,3} by ourselves. Note that the training data of AutoVC(Chou) has 40(20) speakers in their original paper implementation. While in our experiments, for a fair comparison, all these models are trained with the same 99 speakers, and the output mel-spectrograms are converted to waveform using a pretrained MelGAN vocoder[28].

Figure 7 shows the results of the mean opinion score (MOS) test. The “-S” represents that the speakers of the source and the target audio are in the training set, which means they are seen speakers. Otherwise, the “-U” denotes that the speakers are unseen. We ask the subjects (1) how natural and (2) how similar to the target speaker the converted audio sounds, and they score from 1 (very bad) to 5 (very good) after listening to the converted audio and the target audio. “Oracle” means that the audio is generated from MelGAN vocoder with real mel-spectrograms. Hence, it becomes the upper-bound of these three models. AutoVC uses pre-trained speaker embedding, while Chou and our proposed method do not. The results show that our speaker embedding is more effective in restoring speaker

²<https://github.com/auspicious3000/autovc>

³https://github.com/jjery2243542/adaptive_voice_conversion

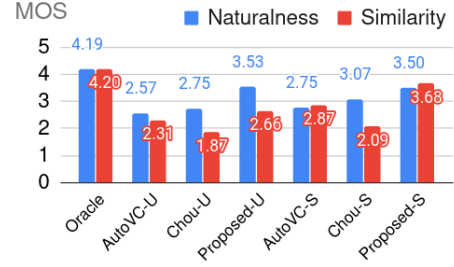


Figure 7: MOS of different methods. “-U” and “-S” are refer to unseen speakers and seen speakers respectively.

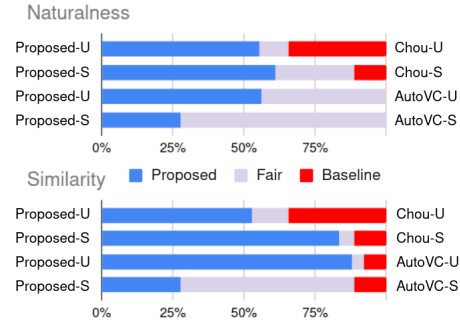


Figure 8: Pairwise comparison. For each row, the blue bar represents the percentage that the proposed method is preferred, and the red bar for the baseline. “-U” and “-S” mean unseen speakers and seen speakers respectively.

information for VC. Moreover, our proposed method also performs the best in the result of naturalness, which indicates that our model generates better mel-spectrograms than other baselines.

Except for the MOS test, we also perform pairwise comparisons between our proposed method and the other two baseline methods. For each question within the questionnaire, we sample audio generated from our model and the baseline model in random order, and ask the subjects which audio sounds more natural and more similar to the target speaker. As shown in Figure 8, our proposed method beat the others in both seen (“-S”) and unseen (“-U”) scenarios. Further, we observe that in seen scenarios (“-S”), AutoVC and our method are comparable. Nevertheless, in unseen (“-U”) scenarios, our method wins for most cases. This implies that our model is more robust to unseen speakers. Meanwhile, it indicates that our method apparently performs better than the baselines for one-shot VC. The generated audio can be found in our demo page.⁴

5. Conclusions

In this paper, we present a new model for one-shot VC. We use the U-Net combined with VQ layers to achieve a high-quality VC. With the well-designed architecture, our proposed model is able to separate the speaker information and the content information effectively in an elegant way with the self-reconstruction loss only. The objective results verify the strong disentanglement of our model, while the subjective results can support our conjecture that the skip-connection design is beneficial for achieving high-quality conversion.

⁴<https://ericwudayi.github.io/VQVC-DEMO>

6. 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] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, “Voice conversion using partial least squares regression,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.
- [4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [5] T. Kaneko and H. Kameoka, “Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks,” in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.
- [6] 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.
- [7] J. Serrà, S. Pascual, and C. S. Perales, “Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion,” in *Advances in Neural Information Processing Systems*, 2019, pp. 6790–6800.
- [8] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, “F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6284–6288.
- [9] H. Lu, Z. Wu, D. Dai, R. Li, S. Kang, J. Jia, and H. Meng, “One-shot voice conversion with global speaker embeddings,” *Proc. Interspeech 2019*, pp. 669–673, 2019.
- [10] A. T. Liu, P. chun Hsu, and H. yi Lee, “Unsupervised end-to-end learning of discrete linguistic units for voice conversion,” *InterSpeech 2019*, 2019.
- [11] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, “Voice conversion across arbitrary speakers based on a single target-speaker utterance,” *InterSpeech 2018*, pp. 496–500, 2018.
- [12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” *ICME 2016*, 2016.
- [13] 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,” *APSIPA 2016*, 2016.
- [14] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” 2019.
- [15] J. chieh Chou, C. chieh Yeh, and H. yi Lee, “One-shot voice conversion by separating speaker and content representations with instance normalization,” *InterSpeech, 2019*, pp. 664–668, 2019.
- [16] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [17] D.-Y. Wu and H. yi Lee, “One-shot voice conversion by vector quantization,” *Icassp, 2020*, pp. 7734–7738, 2020.
- [18] A. van den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.
- [19] J. chieh Chou, C. chieh Yeh, H. yi Lee, and L. shan Lee, “Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations,” *arXiv preprint arXiv:1804.02812*, 2018.
- [20] S. Vasquez and M. Lewis, “Melnet: A generative model for audio in the frequency domain,” *arXiv preprint arXiv:1906.01083*, 2019.
- [21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.
- [22] J.-Y. Liu, Y.-H. Chen, Y.-C. Yeh, and Y.-H. Yang, “Score and lyrics-free singing voice generation,” *arXiv preprint arXiv:1912.11747*, 2019.
- [23] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [24] J. Chorowski, R. J. Weiss, S. Bengio, and A. v. d. Oord, “Unsupervised speech representation learning using wavenet autoencoders,” *arXiv preprint arXiv:1901.08810*, 2019.
- [25] A. H. Liu, T. Tu, H.-y. Lee, and L.-s. Lee, “Towards unsupervised speech recognition and synthesis with quantized speech representation learning,” *arXiv preprint arXiv:1910.12729*, 2019.
- [26] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *arXiv preprint arXiv:1607.08022*, 2016.
- [27] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, “Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2016.
- [28] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Zhen, T. J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” *NeurIPS, 2019*, pp. 14 910–14 921, 2019.