

# CRANK: AN OPEN-SOURCE SOFTWARE FOR NONPARALLEL VOICE CONVERSION BASED ON VECTOR-QUANTIZED VARIATIONAL AUTOENCODER

Kazuhiro Kobayashi<sup>1,2</sup>, Wen-Chin Huang<sup>1</sup>, Yi-Chiao Wu<sup>1</sup>, Patrick Lumban Tobing<sup>1,2</sup>,  
Tomoki Hayashi<sup>1,2</sup>, Tomoki Toda<sup>1</sup>

<sup>1</sup>Nagoya University, Japan,  
<sup>2</sup>TARVO, Inc., Japan,  
kazuhiro.kobayashi@g.sp.m.is.nagoya-u.ac.jp

## ABSTRACT

In this paper, we present an open-source software for developing a nonparallel voice conversion (VC) system named crank. Although we have released an open-source VC software based on the Gaussian mixture model named sprocket in the last VC Challenge, it is not straightforward to apply any speech corpus because it is necessary to prepare parallel utterances of source and target speakers to model a statistical conversion function. To address this issue, in this study, we developed a new open-source VC software that enables users to model the conversion function by using only a nonparallel speech corpus. For implementing the VC software, we used a vector-quantized variational autoencoder (VQVAE). To rapidly examine the effectiveness of recent technologies developed in this research field, crank also supports several representative works for autoencoder-based VC methods such as the use of hierarchical architectures, cyclic architectures, generative adversarial networks, speaker adversarial training, and neural vocoders. Moreover, it is possible to automatically estimate objective measures such as mel-cepstrum distortion and pseudo mean opinion score based on MOSNet. In this paper, we describe representative functions developed in crank and make brief comparisons by objective evaluations.

**Index Terms**— voice conversion, open-source software, vector-quantized variational autoencoder, nonparallel, neural vocoder

## 1. INTRODUCTION

VC is a technique used to convert paralinguistic information such as gender, speaker individuality, and emotions beyond their physical constraints while keeping the linguistic information of a source speech [1]. One of main goals in VC research is to freely control arbitrary factors of a source voice into objective factors depending on the situation in which VC is used. However, control capabilities and the sound quality of the converted voice are usually degraded due to the insufficient modeling accuracy of speech production. If individual

speakers could freely control various factors of the speech, it would open up an entirely new speech communication style.

VC research was initially started to develop a speaker individuality conversion technique enabling a source speaker to change his/her speaker individuality to that of another target speaker while preserving the linguistic content. In this technique, a statistical mapping function that converts acoustic features of the source speech into those of the target speech is preliminarily trained using a parallel dataset consisting of source and target speakers' utterances with the same linguistic contents. To improve the modeling accuracy of the statistical mapping function, several techniques such as the use of the Gaussian mixture model (GMM) [2] and deep neural networks [3] have been proposed.

End-to-end VC [4, 5] is one of the most powerful mapping techniques using a parallel dataset. Unlike conventional statistical mapping techniques, it is not necessary to explicitly calculate alignment functions between source and target utterances. That is, the estimation process of the alignment functions is implicitly included in the model training based on sequence-to-sequence networks and their attention mechanisms. These techniques usually yield considerable improvements of conversion performance compared with the conventional methods using explicit alignment functions. Moreover, it is also possible to convert not only the voice timbre but also prosodic information. On the other hand, it is not straightforward to train the end-to-end mapping function only using a small number of training utterances. Therefore, collecting many parallel utterances usually becomes a burden for users to develop end-to-end VC systems.

To ease the burden of collecting parallel utterances, nonparallel VC has been developed. There are two major nonparallel VC techniques, namely phonetic posteriorgram (PPG)-based VC methods [6, 5] and autoencoder-based VC methods including those using the variational autoencoder (VAE) [7, 8, 9, 10], vector-quantized VAE (VQVAE) [11, 12, 13, 14, 15], and generative adversarial network (GAN) [16, 17, 18, 19]. For the PPG-based VC method, the PPG vector is first estimated using a preliminarily trained

automatic speech recognition (ASR) system. Then, the conversion function is trained using the PPG vector and acoustic features of the target speech. The PPG-based methods achieve relatively higher performance than the autoencoder-based VC methods owing to the speaker-independent linguistic feature of the PPG. However, it is necessary to prepare many training utterances, including contextual information, to build the ASR system for extracting a convincing PPG vector. On the other hand, the autoencoder-based VC methods do not rely on any supervised label such as context labels or parallel utterances, excluding speaker labels. Therefore, the autoencoder-based VC methods are straightforward for building the VC systems compared with the VC methods using parallel utterances and PPG-based VC methods.

In this paper, we introduce an open-source nonparallel VC software based on VQVAE named crank. In addition to the VQVAE-based VC method, we have implemented several components such as WaveNet-like [20] encoder/decoder networks, the hierarchical architecture, the cyclic architecture, and generative adversarial networks. Using crank, One may possible to easily 1) reproduce the VQVAE-based nonparallel VC method using preliminarily stored recipes such as Voice Conversion Challenge (VCC) 2018 and VCC 2020, and 2) develop a nonparallel VC system using one’s own speech corpus. In this paper, we describe 1) technical details and usage, 2) brief comparisons with VCC baseline systems, and 3) experimental results of objective measures calculated using the VCC 2018 dataset.

## 2. NONPARALLEL VOICE CONVERSION BASED ON VQVAE

VQVAE-based voice conversion takes training and conversion phases.

In the training phase, the original feature vector  $\mathbf{x}$  is modeled by the VQVAE consisting of encoder/decoder networks based on the reconstructed loss. The encoder network encodes the original feature vector into the latent vector  $\mathbf{h}$ . The latent vector is quantized into the discrete latent symbol  $\mathbf{q}$ , which minimize the distance between the latent vector  $\mathbf{h}$  and the codebook  $\mathbf{e}$ .

$$\mathbf{q} = \mathbf{e}_k \text{ where } k = \arg \min_j \|\mathbf{h} - \mathbf{e}_j\|_2. \quad (1)$$

Then, the decoder network generates the reconstructed feature vector  $\hat{\mathbf{x}}$  conditioned on the discrete latent symbol  $\mathbf{q}$  and the auxiliary features  $\mathbf{c}_{org}$  such as the speaker code and  $F_0$  of the original speech sample. The objective function of VQVAE is as follows:

$$\mathcal{L}_{obj} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \|\text{sg}[\mathbf{h}] - \mathbf{e}\|_2^2 + \beta \|\mathbf{h} - \text{sg}[\mathbf{e}]\|_2^2, \quad (2)$$

where  $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ ,  $\|\text{sg}[\mathbf{h}] - \mathbf{e}\|_2^2$ , and  $\|\mathbf{h} - \text{sg}[\mathbf{e}]\|_2^2$  are the reconstruction loss, codebook loss, and commitment loss, respectively.  $\beta$  and  $\text{sg}[\cdot]$  indicate the hyperparameter for the

commitment loss and the stop gradient function, respectively. Because there is an  $\arg \min$  function to find the discrete latent symbol, it is not straightforward to optimize this network. To avoid this problem, VQVAE utilizes a straight-through estimator [21] to pass through gradients from the decoder to the encoder via the vector-quantizer.

In the conversion phase, the original feature vector  $\mathbf{x}'$  is first encoded into the latent vector  $\mathbf{h}'$  to find the discrete latent symbol  $\mathbf{q}'$  on the basis of the trained encoder network and codebook. Then, the target auxiliary features  $\mathbf{c}_{tar}$  with the codebook of predicted discrete latent symbols  $\mathbf{q}'$  are fed into the decoder network to generate a converted feature vector  $\hat{\mathbf{y}}'$ .

## 3. CRANK

crank is an open-source software that implements nonparallel VC frameworks. The license of crank is linked to the MIT license. The implementation of crank has been continued on a GitHub repository<sup>1</sup>. In this section, we introduce the basic structure of the crank recipe and representative components. Table 1 shows the fundamental differences in features between crank and successive VCC baseline systems.

### 3.1. Template recipe

In most open-source software for developing a VC system, it is necessary to prepare a speech dataset to run their fundamental functions. To rapidly reproduce and confirm the effectiveness of VQVAE-based VC techniques, we have prepared VCC 2018 [26] and VCC 2020 [27] recipes on the basis of the Kaldi recipe [28]. In these recipes, several steps such as downloading the dataset, feature extraction, training, conversion, and evaluation are automatically processed by simply typing a single command after preparing the execution environment for Python3. Because these recipes are used with a shared template written in shell scripts, it is straightforward to adapt them to one’s dataset<sup>2</sup>.

### 3.2. Feature vector and vocoder

We supported two kinds of the commonly used feature vector in this research field, namely, the mel-cepstrum parameterized from the spectral envelope extracted by CheapTrick [29] and the mel-filterbank. For the VC using mel-cepstrum, it is straightforward to acquire reasonable sound quality using a traditional source-filter vocoder. On the other hand, for the mel-filterbank, it is not straightforward to achieve acceptable sound quality by using only the Griffin–Lim algorithm [30]. To avoid this problem, we have integrated the ParallelWaveGAN [31] vocoder for mel-filterbank decoding<sup>3</sup>.

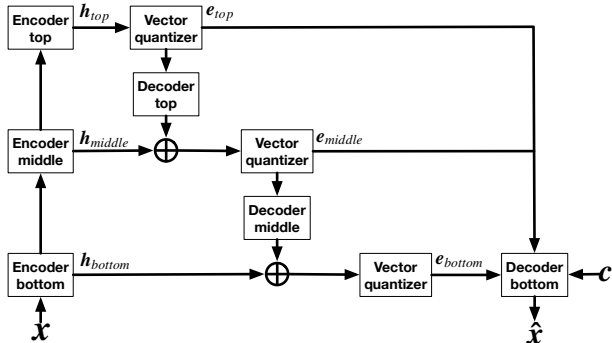
<sup>1</sup><https://github.com/k2kobayashi/crank>

<sup>2</sup>Please see README.md to know how to build an original recipe.

<sup>3</sup>We used an unofficial implementation of ParallelWaveGAN. <https://github.com/kan-bayashi/ParallelWaveGAN>

**Table 1.** Features of crank and successive VCC baseline systems.

| Name              | Year | Method                    | Model       | Requirements       | Open source |
|-------------------|------|---------------------------|-------------|--------------------|-------------|
| Baseline [22]     | 2016 | GMM-based VC              | GMM         | Parallel data      | Yes         |
| sprocket [23]     | 2018 | GMM-based differential VC | GMM         | Parallel data      | Yes         |
| Merlin [24]       | 2018 | DNN-based VC              | DNN         | Parallel data      | Yes         |
| Baseline T11      | 2020 | PPG-based VC              | LSTM        | Speech and context | No          |
| Baseline T16 [25] | 2020 | VAE-based VC              | VAE         | Speech             | Yes         |
| Baseline T22 [10] | 2020 | ASR and TTS               | Transformer | Speech and context | Yes         |
| crank             |      | VQVAE-based VC            | VQVAE       | Speech             | Yes         |



**Fig. 1.** Architecture of hierarchical VQVAE.

### 3.3. Network architecture

For the encoder/decoder architecture of the VQVAE network, we used the WaveNet-like network structure [20] to achieve higher modeling accuracy and higher inference speed than the recurrent neural network-based network. The WaveNet-like network structure has several features such as dilated convolution, gated linear unit, and residual connections. Moreover, as the basis of a work on image generation [32], we used a hierarchical VQVAE structure [14] to achieve higher modeling accuracy.

Figure 1 shows the hierarchical VQVAE structure. The original feature vector passes three encoders to extract latent vectors in each resolution. The top vector-quantizer estimates a discrete latent symbol, and then this symbol is fed into the top decoder to adjust the resolution for adding the middle encoder output. The resulting output is fed into the middle vector-quantizer to estimate the discrete latent symbol, and this latent symbol is also passed through the middle decoder to calculate the discrete latent symbol of the middle stack. Finally, discrete latent symbols calculated in each stack are concatenated and then fed into the bottom decoder to generate a converted feature vector. Note that we did not implement down/up-sample functions among time axes in the hierarchical structure to implement a causal network.

### 3.4. Cyclic architecture

A cyclic architecture for nonparallel VC was initially proposed in a VAE-based VC method [9]. On the basis of this work, we implement the cyclic VQVAE-based VC method. An advantage of the cyclic architecture is that it can include source-to-target conversion flow and target-to-source conversion during its optimization process. The original feature vector is first converted into the converted feature vector conditioned on auxiliary features for target speaker. The converted feature vector is converted into the reconstructed feature vector using auxiliary features consisting of source speaker information to calculate reconstruction loss. By taking into consideration the source-target-source conversion, one can regularize latent features associated with linguistic information among all training speakers. As a result, it is possible to perform stable speaker individuality conversion for any source-target speaker pairs.

### 3.5. Adversarial training

The GAN is one of the most powerful frameworks to generate realistic samples. We implement the least-square GAN framework into the decoder of the VQVAE-based VC. In the VQVAE-based VC method, the discrete latent symbol is estimated from the vector-quantizer through encoder networks. By applying the stop gradient function to the discrete latent symbol, the decoder can be regarded as a generator conditioned on a discrete latent symbol and auxiliary features. For the discriminator, crank also used a WaveNet-like network structure, and auxiliary classifier GAN [33] is also implemented to make the training more stable. Note that it is possible to select either a reconstructed feature vector or a converted feature vector for calculating adversarial loss in our implementation. On the basis of the ParallelWaveGAN [31], we have also implemented a multiresolution short-time Fourier transform (STFT) loss for calculating reconstruction loss.

In VQVAE, a discrete latent symbol is shared over all training speakers as linguistic information. However, as the training process progresses, the VQVAE network easily suffers from overfitting problems. As a result, the predicted discrete latent symbol tends to be speaker-specific linguistic information, degrading conversion quality. To avoid this prob-

lem, inspired by the work on ASR [34], we have implemented an adversarial training procedure for the encoder using a gradient reversal layer. To calculate the adversarial loss of the speaker classifier, we also used the same structure as that of the discriminator.

### 3.6. Objective measures

For the nonparallel VC method, overfitting is one of the biggest problems because it cannot directly optimize source-to-target mapping functions. To avoid this problem, it is reasonable to calculate objective measures that represent conversion performance. To estimate the conversion performance without performing subjective tests, crank automatically calculates mel-cepstrum distortion and the mean opinion score on the basis of the MOSNet [35] using an evaluation set. By calculating mel-cepstrum distortion, one can roughly estimate the conversion performance of speaker individuality. On the basis of MOSNet prediction, it is possible to investigate the sound quality of the converted voice.

## 4. EXPERIMENTS

As brief comparisons between representative functions developed using crank, we evaluated objective measures calculated using VCC 2018 recipes. The sampling rate was set to 22050 Hz. The number of training speakers was 12, and each speaker spoke 80 utterances. We used 75 utterances for training and the remaining five utterances for development. We used the other 35 evaluation utterances in each speaker. The evaluation was performed under the speaker-closed condition (i.e., training speakers were used for the evaluation as well.).

An 80-dimensional mel-filterbank was used as the feature vector. Continuous  $F_0$ , an unvoiced/voice decision symbol, and a speaker code were used as the auxiliary features for the decoder. The ParallelWaveGAN vocoder trained using the same dataset was used as a neural vocoder. We compared mel-cepstrum distortion and predicted naturalness on the basis of MOSNet in this evaluation. The values were averaged among all-speaker pairs. We used a 35-dimensional mel-cepstrum to calculate the distortion. The other settings and resulting voices were described on the website<sup>4</sup>.

The following techniques were compared in this evaluation.

#### Baseline VQVAE

Three-stacked hierarchical VQVAE

#### CycleVQVAE

Baseline VQVAE with cyclic architecture

#### VQVAEGAN

Baseline VQVAE with GAN

**Table 2.** Objective evaluations.

| Method                    | Mel-CD      | MOSNet      |
|---------------------------|-------------|-------------|
| Baseline VQVAE            | 9.89        | 3.53        |
| CycleVQVAE                | 9.66        | 3.54        |
| VQVAEGAN                  | 10.13       | 3.44        |
| CycleVQVAEGAN             | 9.74        | 3.48        |
| CycleVQVAEGAN w/ STFTLoss | <b>9.64</b> | <b>3.59</b> |

#### CycleVQVAEGAN

Baseline VQVAE with cyclic architecture and GAN

#### CycleVQVAEGAN w/ STFTLoss

Baseline VQVAE with cyclic architecture and GAN with STFT loss

The STFT loss means that the network utilizes not only L1 loss but also STFT loss for the reconstruction loss.

Table 2 shows the experimental results of the mel-cepstrum distortion and MOSNet predictions. Compared with the baseline VQVAE, the CycleVQVAE method achieves higher performance in terms of Mel-CD and MOSNet. Moreover, integrating GAN-based training, we can see that the CycleVQVAE w/ STFT method yields the highest performance among methods shown in Table 2. On the other hand, the VQVAEGAN method has a lower performance than the baseline VQVAE method. It is considered that it is not straightforward to optimize the VQVAE decoder network based on the GAN framework, and a cyclic architecture may maintain the stability of the training similarly to CycleGAN-VC [19].

## 5. CONCLUSION

In this paper, we introduced an open-source nonparallel VC software named crank. The main objective of developing crank is to build a nonparallel VC system with limited constraints for collecting the speech corpus. In addition to the vector-quantized variational autoencoder-based VC method, several representative methods such as these using the hierarchical architecture, cyclic architecture, generative adversarial network, and speaker adversarial training have been implemented in crank. Moreover, it also supports the ParallelWaveGAN vocoder to decode a converted mel-filterbank and calculate objective measures such as mel-cepstrum distortion and pseudo mean opinion score on the basis of MOSNet. For our future work, we will continue to develop methods to realize high-quality, easy-to-use nonparallel VC software.

## Acknowledgment

This work was partly supported by JSPS KAKENHI Grant-in-Aid for JSPS Research Fellow Number 19K20295, and JST, CREST Grant Number JPMJCR19A3.

<sup>4</sup><https://k2kobayashi.github.io/crankSamples/>

## 6. REFERENCES

- [1] T. Toda, “Augmented speech production based on real-time statistical voice conversion,” *Proc. GlobalSIP*, pp. 755–759, Dec. 2014.
- [2] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
- [3] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” *Proc. ICASSP*, pp. 4869–4873, Apr. 2015.
- [4] 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 preprint arXiv:1912.06813*, 2019.
- [5] J. Zhang, Z. Ling, and L.-R. Dai, “Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations,” *IEEE/ACM Trans. ASLP*, vol. 28, no. 1, pp. 540–552, 2020.
- [6] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” *Proc. ICME*, pp. 1–6, 2016.
- [7] 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,” *Proc. APSIPA*, pp. 1–6, Dec. 2016.
- [8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder,” *IEEE/ACM Trans. ASLP*, vol. 27, no. 9, pp. 1432–1443, 2019.
- [9] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, “Non-parallel voice conversion with cyclic variational autoencoder,” *Proc. Interspeech*, pp. 674–678, 2019.
- [10] W. Huang, H. Luo, H. Hwang, C. Lo, Y. Peng, Y. Tsao, and H. Wang, “Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion,” *IEEE Trans. TETCI*, vol. 4, no. 4, pp. 468–479, 2020.
- [11] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” *Proc. NIPS*, pp. 6306–6315, 2017.
- [12] D. Shaojin and G.-O. Ricardo, “Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion,” *Proc. Interspeech*, pp. 724–728, 2019.
- [13] 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.
- [14] T. V. Ho and M. Akagi, “Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder,” *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.
- [15] Z. Haitong, “The NeteaseGames system for voice conversion challenge 2020 with vector-quantization variational autoencoder and WaveNet,” *arXiv preprint arXiv:2010.07630*, 2020.
- [16] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, “Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks,” *Proc. Interspeech*, pp. 3364–3368, Apr. 2017.
- [17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks,” *Proc. SLT*, pp. 266–273, 2018.
- [18] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, “High-quality nonparallel voice conversion based on cycle-consistent adversarial network,” *Proc. ICASSP*, pp. 5279–5283, 2018.
- [19] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC2: Improved cyclegan-based non-parallel voice conversion,” *Proc. ICASSP*, pp. 6820–6824, 2019.
- [20] 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,” *arXiv preprint arXiv:1609.03499*, 2016.
- [21] Y. Bengio, N. Léonard, and A. C. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv preprint arXiv:1308.3432*, 2013.
- [22] G. K. Anumanchipalli, K. Prahallad, and A. W. Black, “Festvox: Tools for creation and analyses of large speech corpora,” 2011.
- [23] K. Kobayashi and T. Toda, “sprocket: open-source voice conversion software,” *Proc. Odyssey*, June 2018.
- [24] Z. Wu, O. Watts, and S. King, “Merlin: an open source neural network speech synthesis system,” *Proc. SSW*, pp. 202–207, 2016.
- [25] P. L. Tobing, Y.-C. Wu, and T. Toda, “Baseline system of Voice Conversion Challenge 2020 with cyclic variational autoencoder and Parallel WaveGAN,” *arXiv preprint arXiv:2010.04429*, 2020.
- [26] J. L. Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods,” *Proc. Odyssey*, pp. 195–202, June 2018.
- [27] 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,” *arXiv preprint arXiv:2008.12527*, 2020.
- [28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” *Proc. ASRU*, 2011.
- [29] M. Morise, “Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,” *Speech Commun.*, vol. 67, pp. 1–7, 2015.
- [30] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, 1984.
- [31] R. Yamamoto, E. Song, and J. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” *Proc. ICASSP*, pp. 6199–6203, 2020.
- [32] A. Razavi, A. van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with VQ-VAE-2,” *arXiv preprint arXiv:1906.3432*, 2019.
- [33] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” *Proc. MLR*, vol. 70, pp. 2642–2651, 2017.

- [34] Y. Shinohara, “Adversarial multi-task learning of deep neural networks for robust speech recognition,” *Proc. Interspeech*, pp. 2369–2372, 2016.
- [35] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “MOSNet: Deep learning based objective assessment for voice conversion,” *arXiv preprint arXiv:1904.08352*, 2019.