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

The NU Voice Conversion System for the Voice Conversion Challenge 2020: On the Effectiveness of Sequence-to-sequence Models and Autoregressive Neural Vocoders

Wen-Chin Huang⋆, Patrick Lumban Tobing⋆, Yi-Chiao Wu⋆, Kazuhiro Kobayashi⋆, Tomoki Toda

Nagoya University, Japan

{wen.chinhuang, patrick.lumbantobing, yichiao.wu}@g.sp.m.is.nagoya-u.ac.jp,
kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

In this paper, we present the voice conversion (VC) systems developed at Nagoya University (NU) for the Voice Conversion Challenge 2020 (VCC2020). We aim to determine the effectiveness of two recent significant technologies in VC: sequence-to-sequence (seq2seq) models and autoregressive (AR) neural vocoders. Two respective systems were developed for the two tasks in the challenge: for task 1, we adopted the Voice Transformer Network, a Transformer-based seq2seq VC model, and extended it with synthetic parallel data to tackle nonparallel data; for task 2, we used the frame-based cyclic variational autoencoder (CycleVAE) to model the spectral features of a speech waveform and the AR WaveNet vocoder with additional fine-tuning. By comparing with the baseline systems, we confirmed that the seq2seq modeling can improve the conversion similarity and that the use of AR vocoders can improve the naturalness of the converted speech.

Index Terms: voice conversion, voice conversion challenge, sequence-to-sequence, neural vocoder, autoregressive modeling

1. Introduction

Voice conversion (VC) aims to convert in speech from a source to that of a target without changing the linguistic content [1, 2]. The voice conversion challenge[1] (VCC) aims to better understand advances in VC techniques. This year [3], two tasks were considered: the first task was *intra-lingual semiparallel* VC, where only a small subset of the training set was parallel, with the rest being nonparallel; the second task was *cross-lingual* VC, where the training set of the source speaker is different from that uttered by the target speaker in language and content, thus nonparallel in nature. In conversion, the source speaker's voice in the source language is converted as if it was the target speaker's voice.

In recent years, deep learning has been a game changer in many research fields, and VC is no exception. The theoretically unlimited expressive power of deep neural networks (DNNs) makes it possible to model various complex characteristics that are essential to the performance of VC systems, such as the high resolution nature of speech signals and the conversion of prosody and speaking rates. These characteristics have been addressed by two epoch-making technologies: sequence-to-sequence (seq2seq) models and neural vocoders.

First, as most VC studies have focused on frame-by-frame conversion models, i.e., the converted speech and the source speech are always of the same length, the modeling of speaking rate is largely restricted, and so are other prosody-related factors such as the F0 contour. Seq2seq models [4], which are often equipped with an attention mechanism [5, 6] to implicitly learn the alignment between the source and output sequences and generate outputs of various lengths and capture long-term dependences, are therefore suitable for converting prosody in VC. Since the suprasegmental characteristics of F0 and duration patterns well handled in seq2seq VC models are closely correlated with the speaker identity, it has been shown that seq2seq VC models can outperform conventional frame-wise VC systems, especially in terms of conversion similarity [7, 8].

On the other hand, neural-based speech generation models [9–22] have been proposed to directly model speech waveforms, to overcome the naturalness degradation caused by the loss of phase and temporal details from the oversimplified assumptions in conventional parametric-based vocoders [23, 24]. Although autoregressive (AR) models [9–14] achieve marked high-fidelity speech generation, the AR mechanism greatly limits the generative speed or model complexity. That is, to achieve real-time generation, a compact AR model with specific prior knowledge [13, 14] is required. On the other hand, on the basis of the parallel computing advantage of convolution neural networks (CNNs), many non-AR models such as flow-based [15–17], source-filter-based [18], and GAN-based [19–22] models have been proposed. However, the naturalness of VC speech generated by these non-AR models has not been comprehensively evaluated and compared with AR models. Therefore, we provide more insight via the NU VC system in this paper.

Here, we describe our submission to the VCC2020. For task 1, we extended our previously proposed Transformer-based seq2seq VC framework, the Voice Transformer Network (VTN) [25, 26] with synthetic parallel data, where we used text-to-speech (TTS) systems to generate synthetic data for parallel learning. For task 2, we used the frame-based cyclic variational autoencoder (CycleVAE) [27] framework to model the spectral features of the speech waveform and the AR WaveNet vocoder [9–11] as the waveform generator with additional fine-tuning. We aim to answer the following research questions:

- In what aspects can seq2seq VC models outperform frame-based VC?
- How can VC systems benefit from AR neural vocoders over non-AR neural vocoders?

2. Task 1: VTN with synthetic parallel data

2.1. Motivation

Most seq2seq VC models are data-hungry, requiring over 1 h of data to generalize well. Although we previously proposed a TTS-based pretraining technique that can generalize seq2seq

⋆ Equal contribution.

[1] http://www.vc-challenge.org/

10.21437/VCC_BC.2020-25

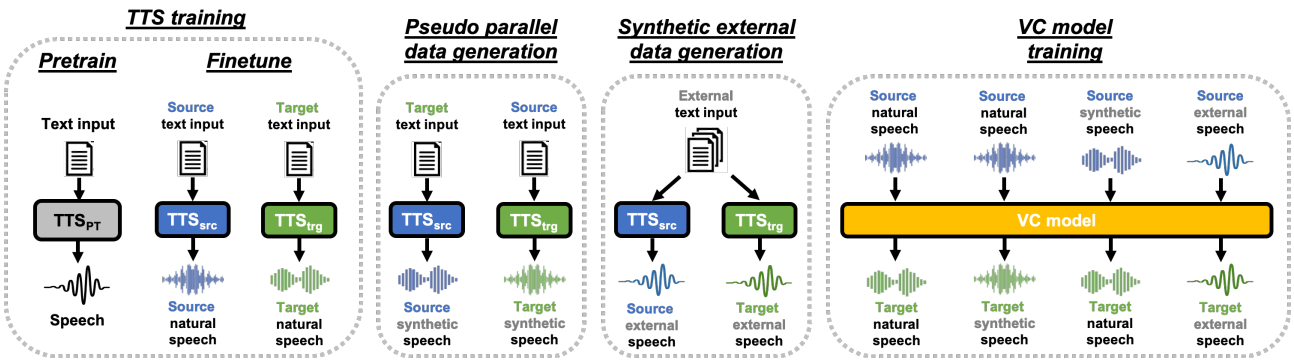


Figure 1: *Overview of the training procedure of task 1 system. The speech utterances generated by a TTS system are shown in grey.*

VC models to only 5 min of training data [25, 26], a major limitation of most existing seq2seq VC models, including ours, is that they can only achieve *parallel* VC, i.e., a requirement of a parallel corpus consisting of pairs of speech samples with identical linguistic contents uttered by both source and target speakers. The application of current seq2seq VC models to a semiparallel setting such as task 1 in VCC2020 is therefore not straightforward, since among the 70 utterances in total, only 20 are parallel and the remaining 50 are not.

Inspired by [28], we utilized TTS models to generate synthetic parallel data for VTN training. Considering the corpus in task 1 being semiparallel, there are three types of speech to use:

1. **Natural speech**: the utterances in the corpus.
2. **Synthetic pseudo-parallel speech**: The synthetic target/source counterpart of the natural speech is generated by inputting the transcription of a source/target natural speech to a target/source TTS system.
3. **Synthetic external speech**: A synthetic pair of utterances with identical content is generated by inputting a set of external text to the source and target TTS systems.

As a result, there are four types of parallel data pairs available for training the VC model in total:

1. <source natural, target natural>
2. <source synthetic pseudo-parallel, target natural>
3. <source natural, target synthetic pseudo-parallel>
4. <source synthetic external, target synthetic external>

2.2. System Overview

In this section, we describe the training and conversion procedure in detail. Figure 1 illustrates the training procedure of our task 1 system.

2.2.1. TTS training

To generate a synthetic speech, TTS systems for the source and target speakers need to be trained first. We adopted a seq2seq-based TTS model considering the success of seq2seq models in the field of TTS. However, the size of the training set of each target speaker in VCC2020 is too limited to train seq2seq TTS models from scratch. In light of this, we employed a pretraining–finetuning scheme that first pretrains on a large TTS dataset followed by fine-tuning on the limited source or target speaker dataset. This allowed us to successfully train on even 5 min of data.

2.2.2. Pseudo-parallel data generation

For the source/target natural utterances in the corpus, we can generate the target/source counterpart to form a <natural, synthetic> data pair. For instance, by simply passing the transcription of a source utterance to the target TTS system mentioned in Section 2.2.1, the output will be of the identity of the target speaker but with the same content as that of the source utterance. This utterance can therefore be used together with the source natural utterance to form a parallel data pair.

2.2.3. Synthetic external data generation

Even with pseudo-parallel data, the total number of valid parallel data pairs is still only 120. For data augmentation, we used an external set of text data as input and generated synthetic voices using the source and target TTS systems. Since we pass the same contents, this set of synthetic external data is parallel in nature and can be used to train the seq2seq VC model. In our initial experiments, we found that including such data can improve the intelligibility of the converted speech.

2.2.4. VC model training and conversion process

Using the four types of parallel data pairs described in Section 2.1, we can extract acoustic features from them as inputs and outputs to train the seq2seq VC model in a parallel manner as usual. In the conversion phase, given a source speech utterance, we first extract the acoustic feature and pass it to the seq2seq VC model to obtain the converted acoustic features. A neural vocoder is then used to synthesize the final converted waveform.

2.3. Implementation

The seq2seq VC model was of the same architecture as described in [26], and we used the official implementation[2]. For TTS model training, we directly used the recipes provided in the open-source implementation of the VCC2020 seq2seq baseline[3]. The external text that we used was from the CMU ARCTIC dataset [29] which contained 1132 utterances. For the neural vocoder, we used the non-AR, faster than real-time Parallel WaveGAN (PWG) [19], and we directly used the pretrained models from the VCC2020 seq2seq baseline.

[2] https://github.com/espnet/espnet/tree/master/egs/arctic/vc1

[3] https://github.com/espnet/espnet/tree/master/egs/vcc20

3. Task 2: CycleVAE and WaveNet vocoder

In this section, we describe the NU system for Task 2 using the CycleVAE-based [27] spectral model and WaveNet-based [11] vocoder. Compared with that in the official baseline systems [30], which also utilizes the CycleVAE-based spectral model, we used the AR neural vocoder, i.e., WaveNet, instead of a non-AR neural vocoder, i.e., PWG. Furthermore, we also used more variations of speech data to support the development of the WaveNet vocoder for producing high-quality synthetic speech. Therefore, compared with the CycleVAE baseline system T16 [30], the differences are as follows: (1) the use of more speech data for training CycleVAE and neural vocoder models, and (2) the use of an AR-based neural vocoder instead of a non-AR one.

3.1. CycleVAE-based spectral model

CycleVAE [27] is a frame-based nonparallel spectral modeling framework based on the variational autoencoder (VAE) [31] and a cycle-consistent approach that recycles converted spectra for generating cyclic-reconstructed spectra that can be utilized in the optimization. This technique has been found to be very effective, compared with only using reconstructed spectra in the optimization [32], to improve both the latent space condition, i.e., to be more speaker-independent, and the accuracy of the converted spectra. In this work, we follow the CycleVAE architecture described in [27], where the differences are the use of a standard Laplacian prior, the use of only two cycles, and the use of a unified encoder–decoder for many-to-many VC.

3.2. WaveNet vocoder

The WaveNet vocoder [9, 10] is an AR neural vocoder that can produce synthetic speech waveforms with natural quality. It consists of a stack of dilated convolutional blocks to effectively capture the receptive field of waveform samples. In this work, we follow the architecture of the shallow WaveNet vocoder using softmax output as in [11], with an extension of the fine-tuning procedure to reduce the mismatches [33] between naturally extracted spectra and converted spectra generated from a VC model, such as from CycleVAE.

3.3. Training and conversion process

The development procedure for the CycleVAE and WaveNet vocoder in task 2 is shown in Fig. 2. We used additional speech data of 24 speakers from the VCTK [34] corpus, 12 males and 12 females, each with 315 utterances, to accompany the VCC 2020 dataset. To train the CycleVAE-based spectral model, we also performed data augmentation using waveform similarity and overlap add (WSOLA)-based F0 transformation [35] to produce modified speech waveforms from the speech waveforms available from the VCC 2020 dataset. Specifically, for each of the 10 target speakers, F0 ratios with respect to each of the four source speakers were used to generate modified speech waveforms producing 40 additional speakers.

On the other hand, in the training of the WaveNet vocoder, we carried out a four steps training/fine-tuning procedure to reduce the mismatches between natural and converted spectra. In the first step, we used natural speech features of the 38 speakers from VCTK and VCC 2020 datasets to train a natural multi-speaker model. In the second step, we retrained the first model using speech features containing reconstructed spectra obtained from the CycleVAE model, i.e., from the 38 speakers. In the third step, we fine-tuned the second model using the generated speech features of only 14 speakers from the VCC 2020 dataset.

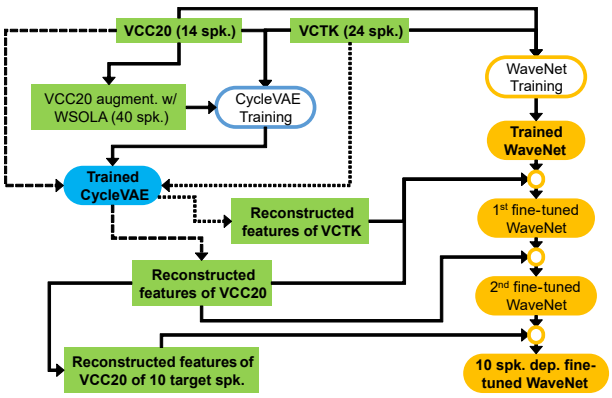


Figure 2: *Development flow of CycleVAE-based spectral model and WaveNet vocoder for task 2 of NU system.*

Finally, we carried out the last fine-tuning using the generated features of only the target speakers to produce speaker-dependent WaveNet vocoders.

The above second training phase was performed in $\sim$5.5 days with NVIDIA Titan V with the number of optimization steps similar to that in the first training. The third and fourth phases were performed until the overfit condition was reached for the development set, which contained 10 utterances of the total 70 utterances for each speaker in the VCC 2020 dataset.

As the speech features, instead of the mel-spectrogram, we used WORLD-based [24] features for both CycleVAE and WaveNet containing unvoiced/voiced (U/V) decisions, continuous log fundamental frequency (log-F0) values, aperiodicity parameters, and 49-dimensional mel-cepstrum [36] parameters including 0-th power extracted from the WORLD-based spectral envelope. CycleVAE models only the estimation of mel-cepstrum parameters. In the conversion phase, we simply generated the converted spectra using the speaker code (one-hot vector) of the desired target speaker and the latent features of the source speaker as in [27]. Then, linear transformation of log-F0 values [2] of the source speaker was performed to convert the pitch of the source to that of the target. Finally, to generate the converted speech waveform, we fed the converted speech features containing input U/V decisions, converted F0, input aperiodicity, and converted spectra to the WaveNet vocoder.

4. Evaluation Results

The VCC2020 organizing committee conducted a large-scale subjective test on all submitted systems for both tasks 1 and 2. The evaluations included naturalness and similarity tests. We first describe the evaluation protocols. Then, we compare the performance of our entry with a subset of the participating systems, as briefly described in Table 1. The results are shown in Figure 3.

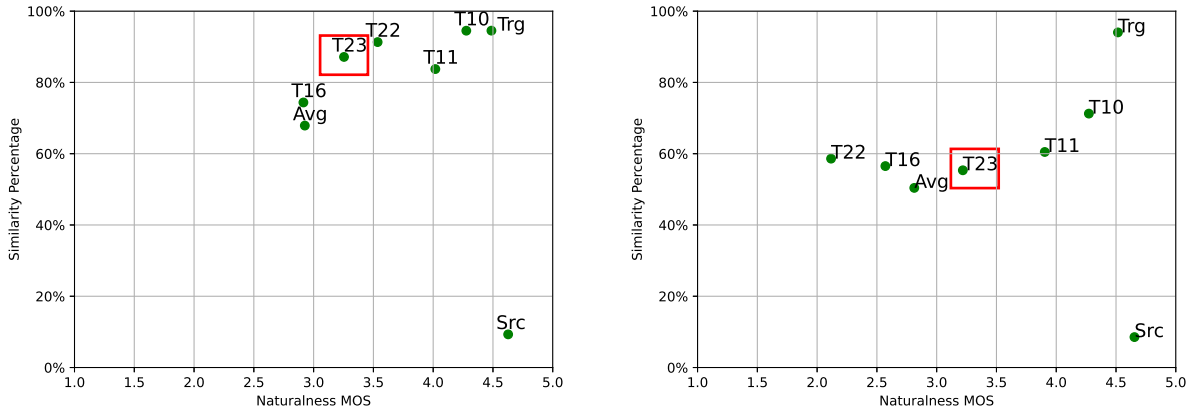
4.1. Evaluation protocol

In the naturalness test, a five-point mean opinion score (MOS) test was adopted, where listeners were instructed to rate the naturalness of each speech clip from 1 to 5. In the similarity test, listeners were presented a converted utterance and a ground truth target utterance, and they were asked to determine whether the two utterances were spoken by the same person on a four-point scale. Figure 3 shows the overall results[4].

[4]Although the official listening report contained results from Japanese and English listeners, we only report results of English lis-

Table 1: *Subset of systems that participated in VCC2020.*

Team ID	Description	Conversion Model	Vocoder	AR
T10	VCC2020 top system	PPG → **seq2seq** → mel filterbanks	WaveNet	✓
T11 [37]	Baseline	PPG → **LSTM** → STRAIGHT features	WaveNet	✓
T16 [27, 30]	Baseline	WORLD features → **CycleVAE** → WORLD features	PWG	✗
T22 [38]	Baseline	Mel filterbanks → **seq2seq ASR** → text → **seq2seq TTS** → mel filterbanks	PWG	✗
T23 task 1	NU system	mel filterbanks → **seq2seq** → mel filterbanks	PWG	✗
T23 task 2	NU system	WORLD features → **CycleVAE** → WORLD features	WaveNet	✓



(a) *Task 1* (b) *Task 2*

Figure 3: *Scatter plot of naturalness and similarity scores. Our system, T23, is indicated by the red boxes.*

4.2. Task 1 results

Figure 3(a) shows the scatter plot of the naturalness and similarity results of task 1. First, our system showed better results than the average of all submitted systems in terms of both naturalness and similarity. Compared with the T16 baseline, which also used the non-AR PWG, our system was superior in terms of both naturalness and similarity. As T16 used a frame-based method (Cycle-VAE), this result demonstrates the effectiveness of seq2seq modeling. Next, without using the AR WaveNet vocoder as T16 did, our system still performed better in terms of similarity, again showing the power of seq2seq models when it comes to modeling speaker identity. Finally, our system could not outperform the last baseline, T22, which also used PWG and seq2seq conversion models. This suggests that parallel acoustic feature mapping may not be the best approach when it comes to seq2seq VC modeling, although an integrated model should theoretically outperform a cascade model such as T22. Analysis on this result will be an important future work.

4.3. Task 2 results

The result of task 2 is shown in Figure 3(b). It can be observed that our system could yield significant naturalness improvements compared with the T16 baseline, which used CycleVAE-based spectral modeling but with a non-AR neural vocoder, i.e, Parallel WaveGAN. Even though the condition of CycleVAE training was not exactly the same, where our system used more data than the T16 baseline, the quality of speech waveforms is heavily affected by the use of the neural vocoder. This was also observed in our internal assessment, where we found that the quality of our system and that of the T16 baseline when using a conventional vocoder, such as WORLD, was similar. Thus, further improvements of feature mapping accuracy, to improve not only naturalness but also speaker similarity, can be achieved by using more constraints for speaker-independent space, such as PPG-based VC, where all of the best systems in VCC 2020 used this approach, as well as baseline T11 and the winner T10.

teners since the two listener groups shared a similar tendency of their results.

5. Conclusions

In this paper, we described the NU VC systems for the VCC2020. The task 1 system was based on VTN, a seq2seq VC model, with the help of synthetic parallel data generated with a TTS system. The task 2 system combined CycleVAE, a frame-based spectral mapping model, with WaveNet, an AR neural vocoder. By comparing with the baseline systems, we confirmed that seq2seq modeling and AR neural vocoders can improve conversion similarity and naturalness of the converted speech, respectively. For future work, we will investigate various aspects of the synthetic parallel data method, such as how the choice, size, and quality of the synthetic data affect the conversion performance. Also, the combination of seq2seq VC models and AR vocoders would be another promising direction.

6. Acknowledgements

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

7. References

[1] Y. Stylianou, O. Cappe, 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 Trajec-

tory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

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

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Sep. 2015, pp. 1412–1421.

[7] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6805–6809.

[8] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-Sequence Acoustic Modeling for Voice Conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR arXiv preprint arXiv:1609.03499*, 2016.

[10] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, 2017, pp. 1118–1122.

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

[12] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, Apr. 2017.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, July 2018, pp. 2415–2424.

[14] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5891–5895.

[15] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, May 2019.

[16] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, May 2019, pp. 3617–3621.

[17] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A generative flow for raw audio," in *Proc. ICML*, June 2019, pp. 3370–3378.

[18] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, May 2019, pp. 5916–5920.

[19] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14 910–14 921.

[21] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, Apr. 2020.

[22] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-periodic parallel wavegan vocoder: A non-autoregressive pitch-dependent dilated convolution model for parametric speech generation," in *Proc. Interspeech*, Oct. 2020.

[23] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[24] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[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 preprint arXiv:1912.06813*, 2019, to appear in Interspeech 2020.

[26] ——, "Pretraining techniques for sequence-to-sequence voice conversion," *arXiv preprint arXiv:2008.03088*, 2020.

[27] 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, pp. 674–678.

[28] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[29] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[30] P. L. Tobing, Y.-C. Wu, and T. Toda, "The baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel wavegan," in *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.

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

[32] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, 2016, pp. 1–6.

[33] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with CycleRNN-based spectral mapping and finely tuned WaveNet vocoder," *IEEE Access*, vol. 7, pp. 171 114–171 125, 2019.

[34] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[35] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *Proc. SLT*, 2016, pp. 693–700.

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

[37] 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, pp. 1983–1987.

[38] 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*, 2020.