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

The NUS & NWPU System for Voice Conversion Challenge 2020

Xiaohai Tian[1], *Zhichao Wang*[2], *Shan Yang*[2], *Xinyong Zhou*[2], *Hongqiang Du*[1,2], *Yi Zhou*[1], *Mingyang Zhang*[1], *Kun Zhou*[1], *Berrak Sisman*[1,3], *Lei Xie*[2] *and Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Audio, Speech and Langauge Processing Group, School of Computer Science,
Northwestern Polytechnical University, China
[3] ISTD Pillar, Singapore University of Technology and Design, Singapore

{eletia, elezmin, haizhou.li}@nus.edu.sg, {zcwang_aslp,lxie}@nwpu.edu.cn,
{syang, xyzhou, hqdu}@nwpu-aslp.org, {yi.zhou, zhoukun, berraksisman}@u.nus.edu

Abstract

This paper presents the NUS & NWPU voice conversion system for Voice Conversion Challenge 2020. Our submission is a Phonetic PosteriorGram (PPG) based voice conversion system, which consists of three modules, including PPG extractor, feature conversion and converted speech signal generation modules. Firstly, a PPG extractor is adopted to extract the speaker independent content features from a speech signal. Then, an encoder-decoder based feature conversion model is used to predict the converted features with the PPG inputs. Finally, a multi-band WaveRNN is utilized to generate the time-domain speech signal from the converted features. The same implementation is used for both intra-lingual and cross-lingual voice conversion tasks. Evaluation results demonstrated the effectiveness of our proposed system.

Index Terms: Voice conversion, intra-lingual, cross-lingual, Phonetic PosteriorGram (PPG), encoder-decoder, multi-band WaveRNN

1. Introduction

Voice conversion (VC) aims to transform the speaker identity from a source speech to a specific target, while maintaining the language content. Typically, a VC system mainly consists of two modules, feature conversion module and vocoder. The feature conversion module is used to transform the speech features of source speaker to that of the target, while a vocoder is used to reconstruct time-domain speech signal with converted features.

With the assumption of the parallel data (the content of source and target speech are the same) is available, various feature conversion techniques are proposed, e.g. Gaussian mixture model (GMM) [1, 2, 3], frequency warping [4, 5, 6], exemplar-based methods [7, 8, 9] and neural network based methods [10, 11]. However, such parallel data is not always available in practice. In order to handle non-parallel data scenarios, non-parallel VC techniques are studied. The INCA algorithm [12] was proposed to align the non-parallel source and target data iteratively. However, the conversion performance was affected by the inaccurate alignment. Recently, generative adversarial network (GAN) based approaches are proposed for non-parallel data VC. For example, cycle-consistent generative adversarial network (CycleGAN) [13] is studied for one-to-one mapping. While StarGAN [14] is able to perform many-to-many conversion. An alternative idea is to disentangle the speaker-dependent component (speaker identity) from the speaker-independent component (speech content). During conversion, only the speaker-dependent component is converted while maintaining the speaker-independent component. An autoencoder network architecture is generally adopted to learn a latent space for speaker-independent representation and a speaker vector is utilized to control the generated speaker identity. To learn the latent representation in a continuous feature space, variational autoencoder (VAE) [15, 16], CycleVAE [17] and autoVC [18] are proposed. While, vector quantised VAE (VQ-VAE) [19] is proposed to learn a discrete latent representation. Among the non-parallel VC approaches, phonetic posteriorgram (PPG) based VC approaches [20, 21, 22] is one of the most popular, where a automatic speech recognition (ASR) system is used to encoded the speech into frame-level speaker independent phonetic representations. A conversion model is trained to learn a mapping between the PPGs to target speech features.

Vocoder plays a crucial role to generate the speech waveform from the converted speech features. It affects the quality of converted speech significantly. Typically, parametric vocoders, e.g. STRAIGHT [23] and WORLD [24], are used in conventional VC system. However, such vocoders are generally designed based on simplified assumptions, which limits the quality of generated speech. To deal with such drawbacks, various neural vocoders have been proposed for high quality speech generation. Neural vocoder is a generative neural network, which is able to directly model the relationships among time-domain speech samples in a data-driven manner. Autoregressive based neural vocoder is one of the most successful approaches, e.g. WaveNet [25], WaveRNN [26, 27] and LPCNet [28]. Recently, some non-autoregressive based neural vocoders have also been proposed, e.g. flow-based vocoders [29], neural source-filter based vocoder [30] and GAN-based vocoders [31, 32, 33].

In this paper, we present our PPG-based average modeling approach for both intra- and cross-lingual VC tasks of VCC 2020 [34]. Specifically, we use an encoder-decoder architecture for feature conversion and WaveRNN-based neural vocoder for converted speech generation. The encoder is based on a CBHG architecture, while the recurrent neural network (RNN) is utilized for decoder. An autoregressive generation is utilized to improve the quality of predicted features. To leverage the public available data and reduce the requirement of target speech, average model based approach is chose in our system. The average model is trained with a multi-speaker database, where a speaker vector is utilized as an auxiliary input to control the speaker identity. Then we adapt a general model towards the target with a small amount of target data. Finally, a multi-band WaveRNN is used for high quality converted speech generation.

10.21437/VCC_BC.2020-26

The organization of this paper is as follows. In Section 2, we present the framework of the submitted system. The system implementation is detailed in Section 3. Section 4 and Section 5 are the evaluation results and conclusion.

2. Framework of proposed system

In this section, we will briefly introduce the framework of our proposed PPG with average modeling VC system. The proposed framework is presented in Fig. 1, which consists of three steps: (a) average conversion model training, (b) target speaker adaptation and (c) run-time conversion. The details will be described as follows.

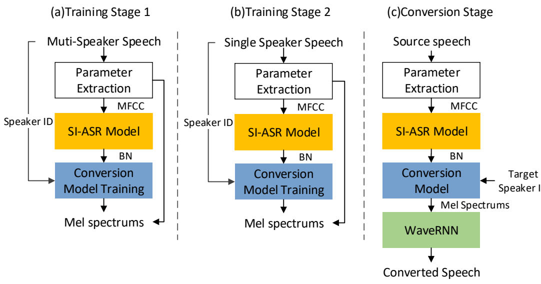

Figure 1: *System diagram of proposed voice conversion system. SI stands for speaker-independent.*

As shown in Fig. 1(a), during training, two types of features, the speaker independent linguistic feature PPG and the speaker dependent acoustic feature Mel-spectrum, are extracted from the multi-speaker corpus. As PPG and Mel-spectrum are extracted from the same utterance, these two feature sequences are initially aligned. Then, the average model is trained to model the relationship between PPG and the corresponding Mel-specturm. An one-hot speaker ID is utilized to control the speaker identity. For a specific target speaker, we adjust the trained SI average model by a small amount of target data, as shown in Fig. 1(b). After adaptation, the conversion model of target speaker is obtained.

At run-time (Fig. 1(c)), given a source speech, we first extract the source PPG and feed it with target speaker ID into the target conversion model to predict the converted Mel spectrogram. Finally, a trained WaveRNN vocoder is used to reconstruct the time-domain signal from the converted Mel spectrogram.

3. System implementation

Our proposed system mainly consists of three modules, PPG extractor, feature conversion module and neural vocoder. In the following subsections, we summarize the each of these modules.

3.1. PPG Extractor

Phonetic PosteriorGram (PPG), an intermediate result of automatic speech recognition (ASR), represents the phonetic content at frame level. As ASR is designed as a speaker invariant system, the extracted PPGs are also considered to be speaker independent. Due to it speaker-independent property, PPGs-based VC has been successfully applied for non-parallel data VC tasks, e.g. intra-lingual [35, 21] and cross-lingual [36, 37] VC.

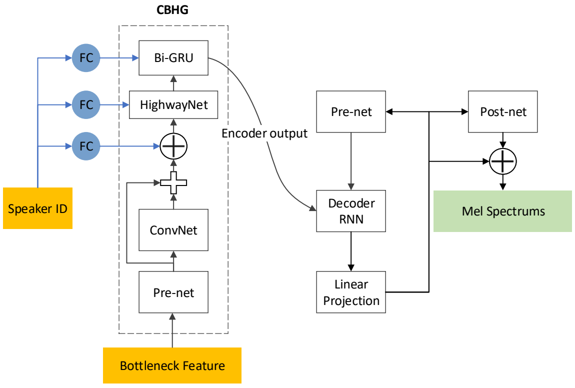

Figure 2: *The network architecture of conversion model.*

Table 1: *Hyper-parameters and network architectures."conv-k-c-ReLU" denotes 1-D convolution with width k and c output channels with ReLU activation. FC stands for fully-connected.*

Speaker Embedding	FC-256-softsign
Encoder Pre-net	FC-256-relu→Dropout(0.3)
Encoder CBHG	Conv1D bank:K=8,conv-k-128-ReLU
	Max pooling:stride=1,width=2
	Conv1D projections:conv-2-256-BactchNorm1D
	Highway net:4 layers of FC-128-ReLU
	Bidirectional GRU:128 cells
Decoder Pre-net	FC-128-relu→Dropout(0.5) →FC-128-relu→Dropout(0.5)
Decoder RNN	1-layer GRU (256 cells)
Linear Projection	FC-80-BN-tanh
Decoder Post-net	Conv1D projections: conv-4-256-BatchNorm1D-tanh

In this work, the ASR model is utilized to extract speaker-independent linguistic content features. The ASR network contained 4 gated recurrent unit (GRU) layers with 512 hidden units in each layer followed by a fully connection bottleneck layer with 256 hidden unites and a soft-max output layer. The 256-dimensional bottleneck feature, represented the content information, is used for conversion model training. As the source speech of both tasks are English, a English PPG extractor is used in our system. Specifically, the clean subsets of Librispeech corpus [38] are used to train the PPG extractor, which consists of 460 hours training data from 1172 speakers. 39-dimensional MFCC are extracted using a 25 ms hamming window with 5 ms shift. The soft-max output layer was 3392, representing 3392 senones. The frame accuracy of the ASR system is 65.26%.

3.2. Conversion model

As shown in Fig.2, an encoder-decoder architecture is employed for conversion model. Following the setting of Tacotron [39], we use CBHG module as the encoder, which consists of a set of 1D convolutional layers, followed by a multi-layer highway network [40] and a bidirectional GRU layer. As the PPGs and the acoustic features are initially aligned, the attention mechanism is not used in our conversion model. In order to control the speaker identity, a one-hot speaker ID is utilized as an auxiliary input of encoder. Specifically, the speaker ID is firstly mapped to high-level representations through three distinct fully-connected (FC) layers. Then the three representations are concatenated with the CNN output, the gate of the highway layer, and the initial state of the GRU, respectively. The decoder consisted of a prenet, a decode recurrent neural network (RNN), a linear projection and a postnet. To further improve the generated speech quality, an autoregressive generation is utilized.

Both VCTK [41] and VCC 2020 corpora are used for model training. 80-dimensional mel-spectrograms with 5 ms frame shift was used as acoustic features. All the audio files are sampled at 24 kHz. The detailed network architecture and hyper-parameters are presented in Table 1.

3.3. Neural Vocoder

To reconstruct waveform from the generated mel-spectrograms, we build a multi-band WaveRNN [26] based vocoder from the ground-truth mel-spectrograms, which can effectively reduce the computational cost compared to the full-band WaveRNN [27]. The architecture of the multi-band WaveRNN is similar to the conventional one [27], where a GRU layer contains 1024 units followed by three feed-forward layers with 512 units are utilized to predict 4 band signals. We firstly adopt all the training data to obtain a speaker-independent WaveRNN model. For each target, we use the target mel-spectrograms generated from training data to fine-tune the basic model.

4. Evaluations and results

4.1. VCC 2020

VCC 2020 consists of two tasks, which are intra-lingual VC and cross-lingual VC. The intra-lingual task requires participants build a VC system with a small semi-parallel database. While, cross-lingual VC is supposed to be a more challenging task, where the source and target speakers utter different languages. The VC system should be capable of handling completely nonparallel training over different languages.

For intra-lingual VC task, it includes 4 English source speakers and 4 English target speakers. Each speaker contained 70 utterances, where 20 sentences are parallel and the content of the other 50 sentences are different. In total, there are $4 \times 4 = 16$ conversion pairs. For cross-lingual VC task, it consists of 4 English source speakers and 6 target speakers with different languages, including 2 Finish speakers, 2 German speakers and 2 Mandarin speakers. Each speaker consists of 70 utterances. There are totally $4 \times 6 = 24$ conversion pairs for this task. All the audio files are sampled as 24 kHz.

4.2. Participants

In total, 33 teams participated the VCC 2020, which are denoted from T01 to T33. Specifically, there are 31 and 28 teams submitted their results for the intra-lingual and cross-lingual conversion tasks, respectively. Among the systems, T11, T16 and T22 are baseline systems provided by organizers. T11 [22] is the top system of VCC 2018, T16 [42] is bulit with CycleVAE and Parallel WaveGAN, while T22 [43] is a cascaded ASR and TTS system. For both tasks, our system is denoted as T25.

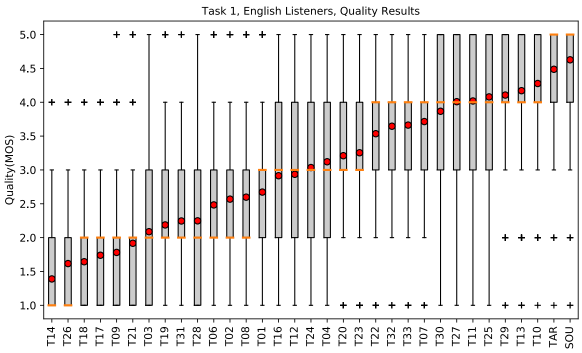

Figure 3: *Average naturalness MOS results of different teams for intra-lingual VC (Task1). Our team is T25.*

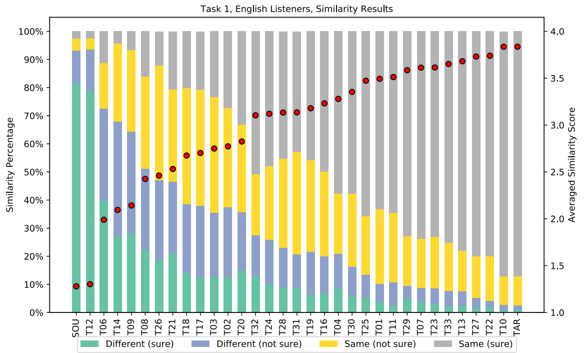

Figure 4: *Average similarity results of different teams for intra-lingual VC (Task1). Our team is T25.*

4.3. Perceptual evaluations

Crowdsource perceptual evaluations were conducted on English and Japanese listeners, respectively. A total of 274 unique listeners are completed the evaluations, including 68 English listeners (32 female, 33 male and 3 for unknown) and 206 Japanese listeners (110 female and 96 male).

4.3.1. Evaluation metrics

Mean opinion score (MOS) is used for speech quality evaluation. During tests, the listeners were asked to rate the naturalness of the converted speech sample on a five-point scale: (1) Bad, (2) Poor, (3) Fair, (4) Good, and (5) Excellent.

To evaluate the speaker similarity, the Same/Different paradigm from the VCC 2018 is used. During tests, the listeners listened to samples from target and converted, then asked to rate speaker similarity of the two samples on a four-point scale: (4) same speaker, absolutely sure, (3) same speaker, not sure, (2) different speaker, not sure, and (1) different speaker, absolutely sure.

4.3.2. Results for intra-lingual VC

Subjective test results on intra-lingual VC task are shown in Figure 3 and Figure 4. the original source and target speakers are denoted as SOU and TAR. Our system is denoted as T25.

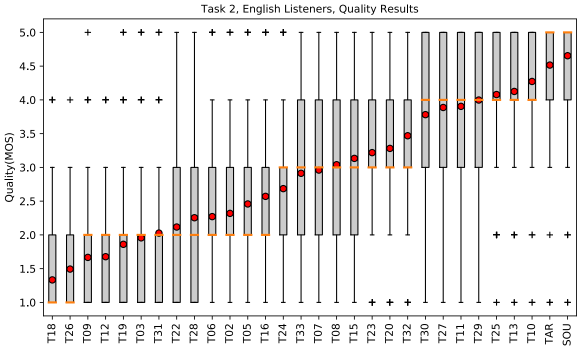

Figure 5: *Average naturalness MOS results of different teams for cross-lingual VC (Task2). Our team is T25.*

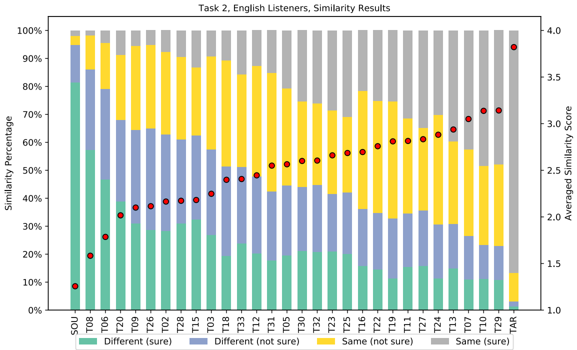

Figure 6: *Average similarity results of different teams for cross-lingual VC (Task2). Our team is T25.*

As shown in Figure 3, our system achieves a naturalness MOS of 4.08 for English listeners. It is observed that T10 performs best among all the systems. While the difference between T13, T29 and our system is not significant. Figure 4 shows that our system achieves a similarity score of 3.47 for English listeners, which is similar to T11 baseline.

4.3.3. Results for cross-lingual VC

Subjective test results on intra-lingual VC task are shown in Figure 3 and Figure 4. Note that the proposed method achieves consistently good results in cross-lingual VC task. It is observed that our system achieves the naturalness and similarity MOS of 4.08 and 2.69, respectively. Overall, our system outperforms all the baselines and rank at 5th place.

5. Conclusions

This paper presented our PPG-based VC system for VCC 2020. The proposed system utilize an autoregressive encoder-decoder model to learn a mapping between speaker-independent PPG to speaker-dependent speech feature. To leverages the public available data for model training, an average model based approach is used. Specifically, an average model is first trained with multi-speaker data. Then a small amount of target speech is used to adapt the average model towards target. Finally, a multi-band WaveRNN is used for converted speech generation. The results of VCC 2020 have demonstrated that the proposed system is able to consistently generate high quality speech for both intra-lingual and cross-lingual VC tasks. Our system achieves average naturalness and similarity MOS of 4.08 and 3.08, respectively.

6. Acknowledgements

This work is supported by Human-Robot Interaction Phase 1 (Grant No. 19225 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore. This research work is partially supported by the National Key Research and Development Program of China (No.2017YFB1002102).

7. References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 1998, pp. 285–288.

[2] 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.

[3] 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.

[4] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[5] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[6] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2014, pp. 211–215.

[7] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Spoken Language Technology workshop (SLT)*, 2012, pp. 313–317.

[9] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1876, 2017.

[10] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (se) minimization training of neural network for voice conversion," in *INTERSPEECH*, 2014.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[12] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2009.

[13] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[14] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[15] 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 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[16] ——, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[17] 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.

[18] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," *arXiv preprint arXiv:1905.05879*, 2019.

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

[20] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data." in *Odyssey*, vol. 2018, 2018, pp. 227–232.

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

[22] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190

[23] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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, 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] A. v. d. 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 preprint arXiv:1609.03499*, 2016.

[26] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[27] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[28] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[29] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[30] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2019.

[31] 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 *Advances in Neural Information Processing Systems*, 2019, pp. 14 910–14 921.

[32] 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.

[33] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multiband melgan: Faster waveform generation for high-quality text-to-speech," *arXiv preprint arXiv:2005.05106*, 2020.

[34] 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, pp. XXX–XXX.

[35] X. Tian, E. S. Chng, and H. Li, "A speaker-dependent wavenet for voice conversion with non-parallel data." in *Interspeech*, 2019, pp. 201–205.

[36] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6790–6794.

[37] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams." in *INTERSPEECH*, 2016, pp. 322–326.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[39] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[41] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[42] P. L. Tobing, Y.-C. Wu, and T. Toda, "The baseline system of voice conversion challenge 2020 with cyclic variational autoencoder and parallel wavegan."

[43] 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, pp. XXX–XXX.