Interspeech 2018
2-6 September 2018, Hyderabad

WaveNet Vocoder with Limited Training Data for Voice Conversion

Li-Juan Liu[1], *Zhen-Hua Ling*[2], *Yuan-Jiang*[1], *Ming-Zhou*[1], *Li-Rong Dai*[2]

[1]iFLYTEK Research, iFLYTEK Co., Ltd.
[2]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China

{ljliu, yuanjiang, mingzhou}@iflytek.com,{zhling, lrdai}@ustc.edu.cn

Abstract

This paper investigates the approaches of building WaveNet vocoders with limited training data for voice conversion (VC). Current VC systems using statistical acoustic models always suffer from the quality degradation of converted speech. One of the major causes is the use of hand-crafted vocoders for waveform generation. Recently, with the emergence of WaveNet for waveform modeling, speaker-dependent WaveNet vocoders have been proposed and they can reconstruct speech with better quality than conventional vocoders, such as STRAIGHT. Because training a WaveNet vocoder in the speaker-dependent way requires a relatively large training dataset, it remains a challenge to build a high-quality WaveNet vocoder for VC tasks when the training data of target speakers is limited. In this paper, we propose to build WaveNet vocoders by combining the initialization using a multi-speaker corpus and the adaptation using a small amount of target data, and evaluate this proposed method on the Voice Conversion Challenge (VCC) 2018 dataset which contains approximately 5 minute recordings for each target speaker. Experimental results show that the WaveNet vocoders built using our proposed method outperform conventional STRAIGHT vocoder. Furthermore, our system achieves an average naturalness MOS of 4.13 in VCC 2018, which is the highest among all submitted systems.

Index Terms: voice conversion, WaveNet, vocoder, adaptation

1. Introduction

Voice conversion (VC) aims to process the speech from one speaker (source speaker) in order to make it sound like being uttered by another speaker (target speaker) while keeping linguistic contents unchanged. Various approaches have been proposed to achieve this task. Among them, the statistical parametric voice conversion (SPVC) approach has attracted most attention in recent years. In this approach, the acoustic features extracted from the source speech are first mapped toward the target speaker using a conversion model, which could be a Gaussian mixture model (GMM) [1, 2, 3], an artificial neural network (ANN) [4, 5] and so on. Then, the converted acoustic features are sent into a vocoder to reconstruct the speech waveforms of the target speaker. Although this approach owns the advantages of building systems automatically and producing stable conversion output, the converted speech usually suffers from the degradation of speech quality and the lack of similarity to the target speaker.

One reason is the inadequacy of conversion models to capture the complex mapping relationship between source and target acoustic features, resulting in the over-smoothness of generated spectra. Many methods have been proposed to improve the performance of VC by alleviating the over-smoothing problem. Various deep neural networks (DNN) were designed to boost the ability of conventional GMMs [6, 7, 8, 9, 10]. Some methods tried to develop novel training criteria, such as minimizing sequence errors (SE) [11] and generative adversarial networks (GAN) [12]. Moreover, to compensate the smoothed spectral details, additional features, such as global variance (GV) [3] or modulation spectrum (MS) [13], were utilized.

The other reason is that artifacts are introduced by using vocoders for waveform reconstruction. Conventional source-filter vocoders are designed under assumptions about the speech production mechanism. Some waveform details, such as the phase information, are usually discarded during parameterization. To avoid this problem, a method of conducting VC at waveform level directly was proposed [14]. A differential GMM (DIFFGMM) was estimated for waveform modulation. This method can obtain high quality of converted speech for intra-gender conversion pairs. However, the quality still degraded due to the F_0 conversion in inter-gender conversions.

Recently, the naturalness of statistical parametric speech synthesis (SPSS) has been significantly improved benefitting from the emergence of WaveNet for direct waveform modeling and generation [15]. Speaker-dependent WaveNet vocoders that can reconstruct waveforms from intermediate acoustic representations (e.g., acoustic features extracted by conventional vocoders, mel-spectrograms) were proposed [16, 17, 18, 19, 20]. The effectiveness of incorporating WaveNet vocoder into SPVC for waveform generation has also been investigated [21]. This work adopted a 1-hour dataset for experiments because the speaker-dependent training of WaveNet vocoders always requires a relatively large training set of a specific speaker. This makes it infeasible to apply WaveNet vocoders to general VC task with small training sets. The method of building speaker-independent WaveNet vocoders was then studied [22]. Although this method required no training data of unknown speakers, it only achieved comparable performance to STRAIGHT.

In this paper, we explore the approaches of building WaveNet vocoders with limited training data for VC. Speaker adaptation techniques have been developed for training acoustic models with a few samples in speech synthesis and voice conversion [23, 24, 25]. This paper investigates the performance of applying similar speaker adaptation techniques to build WaveNet vocoders with limited training data. First, an initialization model is trained with a multi-speaker corpus. Then, it is fine-tuned with the small amount of adaptation data from the target speaker. This learnt WaveNet vocoder is used for reconstructing waveforms from the converted acoustic features during the VC process. We evaluate the performance of the proposed method in Voice Conversion Challenge (VCC) 2018, which provided approximately 5 minutes of recorded speech for each target speaker. The experimental results show that our proposed method can obtain better speech quality

10.21437/Interspeech.2018-1190

than STRAIGHT and help to improve the VC quality on both naturalness and similarity.

This paper is organized as follows: a brief overview of WaveNet vocoder is described in Section 2. The details of our proposed methods are introduced in Section 3. Section 4 shows experimental setups and results. The conclusion is given in the end.

2. WaveNet Vocoder

WaveNet [15] is an autoregressive generative neural network and has been recently proposed to model raw audio waveforms directly. For a waveform sequence $\boldsymbol{x} = [x_0, x_1, ..., x_{T-1}]$, WaveNet models the probability of generating $\boldsymbol{x}$ as the product of conditional distributions, i.e.,

$$p(\boldsymbol{x}; \boldsymbol{\lambda}) = \prod_{t=0}^{T-1} p(x_t | x_0, x_1, ..., x_{t-1}; \boldsymbol{\lambda}). \quad (1)$$

To learn the long-term dependencies among temporal waveform samples efficiently, WaveNet utilizes stacks of causal dilated convolution layers and gated activation units. The calculation at each layer is

$$\boldsymbol{z} = tanh(\boldsymbol{W}_{f,k} * \boldsymbol{y}) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{y}), \quad (2)$$

where $\boldsymbol{y}$, $\boldsymbol{z}$ are the input and output vectors, k denotes the layer index, f and g represent the filter and gate respectively, $\boldsymbol{W}_{f,k}$ and $\boldsymbol{W}_{g,k}$ are trainable weight matrices, $*$ denotes a convolution operator, $\odot$ is an element-wise multiplication operator, and $\sigma(\cdot)$ denotes a sigmoid function.

WaveNet vocoders aim to recover time-domain waveform samples from intermediate representations of speech signals. They are developed based on the conditional version of WaveNet, which is realized by adding an extra input to each layer to control waveform generation. Then, the calculation at each layer output becomes

$$\boldsymbol{z} = tanh(\boldsymbol{W}_{f,k} * \boldsymbol{y} + \boldsymbol{V}_{f,k} * \boldsymbol{h}) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{y} + \boldsymbol{V}_{g,k} * \boldsymbol{h}), \quad (3)$$

where $\boldsymbol{h}$ denotes the condition feature vector, $\boldsymbol{V}_{f,k}$ and $\boldsymbol{V}_{g,k}$ are trainable convolution filters.

In WaveNet vocoders, the intermediate representations are acoustic descriptions extracted from waveforms, such as the acoustic features extracted by conventional vocoders (e.g., mel-cepstra and F_0 extracted by STRAIGHT analysis) [16, 17] or the raw mel-spectrograms given by STFT [18, 20]. These intermediate representations are used as the condition input $\boldsymbol{h}$ in Eq.(3). Then the mapping from the intermediate representations to the time-domain waveforms is learnt in a data-driven way. Therefore, WaveNet vocoders are capable to recover some waveform details that are not contained in the intermediate acoustic representation, such as the phase information. Furthermore, WaveNet vocoders get rid of the constraint of linear filtering since neural networks provide the flexibility of mapping input toward output in a non-linear way.

On the other hand, in contrast to conventional vocoders, WaveNet vocoders need to be trained beforehand. The speech data of a few hours from the reference speaker is usually adopted in order to train a high-quality WaveNet vocoder in a speaker-dependent way.

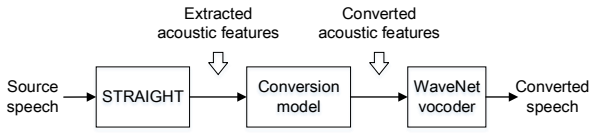

Figure 1: *The process of voice conversion using a WaveNet vocoder.*

3. Proposed Methods

3.1. Building WaveNet vocoders with limited training data

Speaker adaptation methods have been proposed to facilitate model training with small datasets, such as personalized speech synthesis with a few samples [23]. In these methods, a speaker-dependent model is usually trained by adapting a pre-trained multi-speaker model for the reference speaker. Considering common characteristics shared among different speakers, such as speech pronunciation, it is supposed that speaker adaptation based on the learnt multi-speaker model would facilitate the model training for the reference speaker, thus decrease the amount of required data. We investigate the effectiveness of applying this technique to building WaveNet vocoders with limited training data of the reference speaker. The proposed method includes two steps: the training of an initialization model with a multi-speaker dataset and the adaptation with the limited training data of the reference speaker.

The acoustic features (mel-cepstra and F_0) extracted by STRAIGHT vocoder are used as the intermediate representation features in this study. In order to get the initialization model, a unified WaveNet vocoder model is first trained with a multi-speaker dataset. The acoustic features augmented with a speaker embedding vector are used as the condition input. The network parameters and the speaker embedding vectors are learnt simultaneously following the WaveNet framework. It is expected that the speaker embedding vectors can capture speaker-related information [18, 22]. Then, these learnt speaker embedding vectors are discarded and only the model parameters dealing with the acoustic features are used as the initial model parameters for adaptation. In the adaption step, the speaker-dependent WaveNet vocoder is trained by updating all initial model parameters using the training data from the reference speaker.

3.2. Voice conversion using a WaveNet vocoder

Fig. 1 presents the process of voice conversion using a WaveNet vocoder [21], in which the WaveNet vocoder employs the acoustic features extracted by STRAIGHT as input [17]. In this process, the acoustic features of the source speaker are first converted toward the target speaker using a conversion model. Then, the waveform samples of the converted speech are synthesized by sending the converted acoustic features into the WaveNet vocoder built for the target speaker.

In this paper, the conversion model is built based on a framework similar to the context posterior probabilities (CPPs) based VC method [26]. A speaker-independent content feature extractor is built first, which maps acoustic features toward linguistic labels for each frame. This model is used to extract content features from source speech at the conversion stage. Then a speaker-dependent acoustic feature predictor is trained using the training data of the target speaker, which converts the content features toward the acoustic features of the target

speaker. Because the content feature extractor and the acoustic feature predictor are trained separately, this conversion model can deal with parallel and non-parallel VC tasks in the same way. Finally, the WaveNet vocoder of target speaker, which is built following the method introduced in Section 3.1, is used to reconstruct waveforms from the converted acoustic features.

4. Experiments

4.1. Experimental setups

We evaluated the proposed method on the VCC 2018 HUB task [27], which was a parallel training task. The database included 4 source speakers (2 female and 2 male) and 4 target speakers (2 female and 2 male), respectively. They were all professional US English speakers. 81 sentences of each speaker were released for training. The total duration of recordings for each speaker was about 5 minutes. The number of sentences for testing was 35. The recordings were sampled at 22.05kHz with 16bit resolution. We downsampled them to 16kHz/16bit for experiments. For the multi-speaker WaveNet vocoder training, an internal dataset of iFlytek was employed. It consisted of recordings from 20 speakers (10 male and 10 female), with a total duration of approximately 80 hours.

In our experiments, we randomly chose 76 sentences from each speaker for training and the remaining 5 sentences for validation. 41-dimensional mel-ceptral coefficients, 3-dimensional logarithmic fundamental frequency (static, delta and delta-delta) and a unvoiced/voiced (U/V) flag were used to compose the condition vector of acoustic features at each frame in the WaveNet vocoder. STRAIGHT vocoder was employed to extract those features as well as 5-band aperiodicities at 5 ms frame shift. The F_0 parts in unvoiced regions were interpolated and the extracted F_0 values were manually revised in order to prevent the extraction errors from affecting the training of WaveNet vocoder.

We trained a WaveNet vocoder for each target speaker. The architecture of WaveNet vocoders included 4 blocks, with 10 dilation layers in each. The dilation in each block started from 1 and was doubled for every layer until it was up to 512. The filter size of causal dilated convolution was 2. For the skip-connection and the 1×1 convolution layer before the softmax output layer, the number of channels was set to 256. The number of channels in all the other convolution layers, such as the residual connections, the filter and gating convolution layers, was set to 100. Besides, to alleviate the quantization noise in synthetic speeches, we modelled the 10-bit (μ-law) waveforms with a 1024-way categorical distribution [28]. The conditional network consisted of 5 convolution layers and the network output was repeated 80 times directly in order to match the temporal resolution of waveforms. The dimensionality of the speaker embedding vector was 50. All WaveNets were optimized with the Adam optimization method [29] with a constant learning rate 1×10^{-5}. The multi-speaker WaveNet was trained for 440,000 steps. During adaptation, the WaveNet models were further updated for 10,000 steps.

When building the conversion model introduced in Section 3.2, the speaker-independent content feature extractor was estimated using hundreds of hours of recordings with aligned phonetic transcriptions. The acoustic feature predictor for each target speaker consisted of one feedforward layer, two recurrent layers of long-short term memory with projection (LSTMP) and one linear output layer. The number of units in each hidden layer was 512 while the number of the projection units in LSTMP layers was 256. As BAPs can help to improve the quality of reconstructed speech [30], the 5-band BAPs along with the full 45-dimensional acoustic features used in WaveNet vocoder were predicted together in this network. Stochastic gradient descent (SGD) algorithm was used to train this model with a learning rate of 1×10^{-3} and a momentum of 0.9. For each target speaker, the model parameters were not learnt from scratch. They were initialized by a pre-trained multi-speaker model to deal with issue of limited training data.

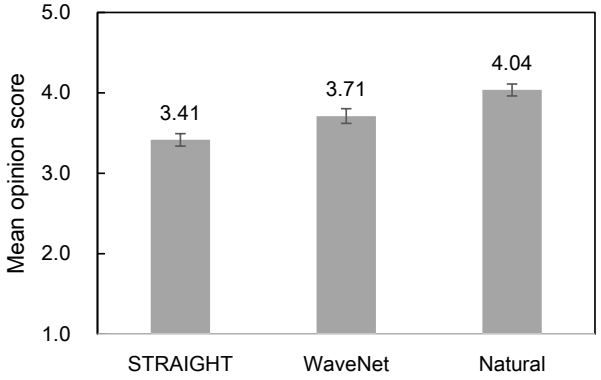

Figure 2: *Mean opinion scores on speech quality of reconstructing waveforms using STRAIGHT vocoder and WaveNet vocoder respectively. Error bars show* 95% *confidence interval.*

4.2. Performance of WaveNet vocoders

To evaluate the performance of WaveNet vocoders trained using our proposed methods, we conducted a mean opinion score test to compare the waveforms synthesized by the learnt WaveNet vocoder with those synthesized by STRAIGHT vocoder together with natural waveforms. 7 non-native English speakers took part in this test. The results shown in Fig. 2 demonstrate that the WaveNet vocoder outperformed STRAIGHT on speech quality significantly. It indicates the effectiveness of our proposed methods in building WaveNet vocoders with only 5 minutes of training data.

4.3. Performance of VC with WaveNet vocoders

Furthermore, we conducted subjective evaluations to assess the naturalness and similarity of voice conversion using WaveNet vocoders. Two systems were compared in our experiments, which were

- VC-STRAIGHT: Conventional VC with STRAIGHT vocoder for waveform generation.
- VC-WaveNet: VC with WaveNet vocoder for waveform generation. The WaveNet vocoder of each target speaker was trained following the proposed adaptation method.

A mean opinion score test was adopted to evaluate the naturalness of converted speech while the speaker similarity was evaluated by a preference test. 40 sentences, randomly selected from 8 conversion pairs (including all conversion types) were rated by 7 non-native English speakers.

4.3.1. Naturalness

Fig. 3 shows the results of the MOS test. The error bar presents 95% confidence interval. We can see that WaveNet vocoders

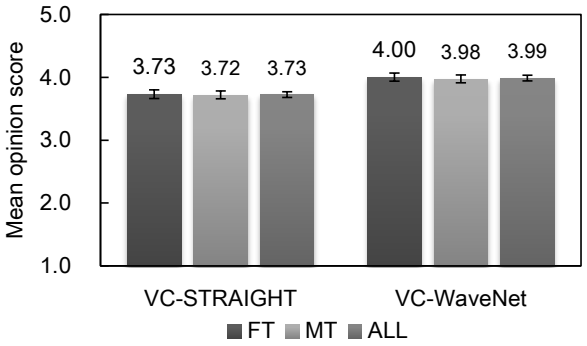

Figure 3: *Mean opinion scores on naturalness of VC systems using STRAIGHT vocoder and WaveNet vocoders respectively. Error bars show 95% confidence interval. "FT", "MT", and "All" denote the conversion pairs of female target speakers, male target speakers and all target speakers.*

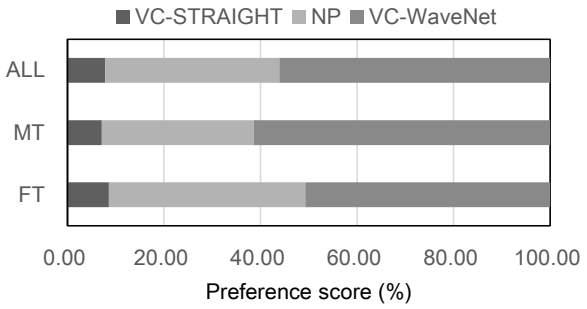

Figure 4: *Preference scores on speaker similarity of VC systems using STRAIGHT vocoder and WaveNet vocoders. "NP" stands for no preference. The p-values of the t-tests for MT, FT and ALL are* 1.1×10^{-18}, 2.2×10^{-12} *and* 3.3×10^{-29} *respectively.*

significantly improved the naturalness of the converted speech comparing with STRAIGHT. This conclusion is consistent among conversion pairs of female target speakers (FT) and male target speakers (MT).

4.3.2. Speaker similarity

The results of the preference tests on speaker similarity are summarized in Fig. 4. Similar to the naturalness results presented in Fig. 3, VC-WaveNet outperformed VC-STRAIGHT on speaker similarity as well. More than half of the test utterances generated by VC-WaveNet were considered to be more similar to the target speakers than VC-STRAIGHT.

4.4. Evaluation results of VCC 2018

Fig. 5 shows the evaluation results of all systems in VCC 2018. Our system performed best among all participants on both naturalness and speaker similarity. Benefitting from effectively training WaveNet vocoders with limited target data, our system outperformed all the other systems on naturalness significantly. Our system achieved a MOS score of 4.13, while the MOS score of natural speech (target) was 4.64.

The baseline system B01 was a vocoder-free system based on the DIFFGMM [14]. We can see that B01 performed better than all other systems except our system on naturalness.

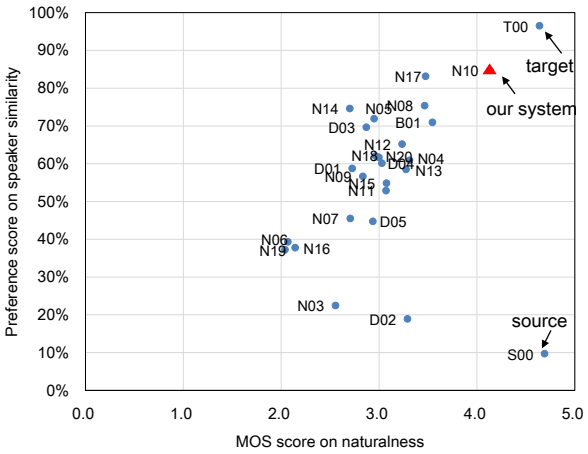

Figure 5: *Scatter plot of the overall naturalness and speaker similarity scores of all systems in VCC 2018.*

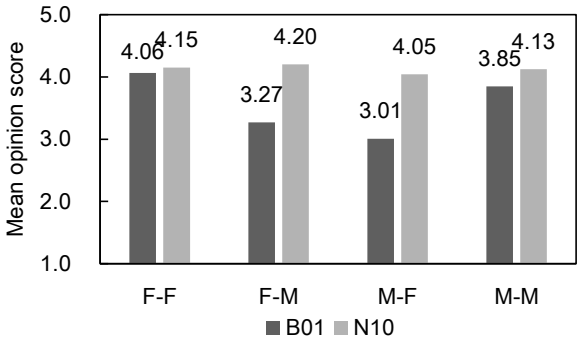

Figure 6: *Mean opinion scores of different conversion pairs of B01 and N10 (our system) in VCC 2018.*

Fig. 6 presents the MOS scores of our system (N10) and B01 on different types of conversion pairs. We can see that B01 performed well on intra-gender conversion pairs. However, the performance greatly degraded on inter-gender conversions. Compared with B01, our system was capable to achieve stable performance on all conversion pairs.

5. Conclusion

In this paper, we have proposed an approach to build WaveNet vocoders with limited training data for VC. The speaker-dependent WaveNet vocoders are estimated by adapting an initialization model, which is learnt using a multi-speaker dataset. Benefitting from the use of this speaker adaptation technique, a stable speaker-dependent WaveNet vocoder can be obtained with only 5 minute training data of target speaker. Both the results of our internal experiments and VCC 2018 evaluations demonstrate the effectiveness of this method. The proposed method can also be applied to improve the quality of personalized speech synthesis systems. We will investigate this topic in the future. Applying speaker adaptation techniques to other neural vocoders, such as SampleRNN-based ones [31], will also be a task of our future work

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] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, 1998, pp. 285–288.

[3] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, vol. 1, 2005, pp. I–9.

[4] B. Makki, S. Seyedsalehi, N. Sadati, and M. N. Hosseini, "Voice conversion using nonlinear principal component analysis," in *Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007. IEEE Symposium on*. IEEE, 2007, pp. 336–339.

[5] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, July 2010.

[6] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *Proc. Interspeech*, 2013, pp. 3052–3056.

[7] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[8] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *Spoken Language Technology Workshop (SLT)*, 2014, pp. 19–23.

[9] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869–4873.

[10] Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using input-to-output highway networks," *IEICE Transactions on Information and Systems*, vol. 100, no. 8, pp. 1925–1928, 2017.

[11] 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 *Proc. Interspeech*, 2014, pp. 2283–2287.

[12] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[13] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," in *Proc. ICASSP*, 2015, pp. 4859–4863.

[14] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016." in *Proc. Interspeech*, 2016, pp. 1667–1671.

[15] A. Van Den 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.

[16] Y.-J. Hu, C. Ding, L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "The USTC system for Blizzard Challenge 2017."

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

[18] S. O. Arık, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.

[19] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[21] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," *Proc. Interspeech*, pp. 1138–1142, 2017.

[22] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 712–718.

[23] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[24] S. Takaki, S. Kim, and J. Yamagishi, "Speaker adaptation of various components in deep neural network based speech synthesis," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 153–159.

[25] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, vol. 4, 2007, pp. IV–1249.

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

[27] J. Lorenzo-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," in *Submitted to Odyssey 2018*.

[28] Y.-J. Hu, L.-J. Liu, C. Ding, Z.-H. Ling, and L.-R. Dai, "The USTC system for Blizzard Machine Learning Challenge 2017-ES2," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 650–656.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," pp. 2266–2269, 2006.

[31] Y. Ai, H.-C. Wu, and Z.-H. Ling, "SampleRNN-based neural vocoder for statistical parametric speech synthesis," in *Proc. ICASSP*, 2018 (accepted).