The Tencent speech synthesis system for Blizzard Challenge 2019

Qiao Tian, Jing Chen, Shan Liu

Tencent Technology Co., Ltd

briantian@tencent.com

Abstract

This paper presents the Tencent speech synthesis system for Blizzard Challenge 2019. The corpus released to the participants this year is a about 8 hours of speech data from an internet talk show by a well-known Chinese character. We built a end to end speech synthesis system for this task. Firstly, a multi-speaker Tacotron-like acoustic model fed on nonalignment linguistic feature and sentence embedding by Bert were employed for mel spectrograms modeling. Then the model was re-trained only on the corpus offered. At last, a modified multi-speaker WaveNet model conditioned on the predicted mel features was trained to generate 16-bit speech waveforms at 24 kHz, instead of the conventional vocoder. For achieving higher quality, channel embedding was incorporated in WaveNet. The evaluation results shows that the system we submitted performs good in various criteria which indicated the superiority of our system.

Index Terms: Blizzard Challenge 2019, end to end speech synthesis, Tacotron, Bert, WaveNet, channel embedding

1. Introduction

In order to promote speech synthesis technology, the Blizzard Challenge has been organized annually since 2005 in building corpus-based speech synthesis systems on the same data. Unit selection based waveform concatenation methods [1] and statistical parameter speech synthesis (SPSS) approach [2, 3, 4] have been the most popular methods in the past years. As the direct using of raw segments of natural speech corpus, high quality speech could be generated by unit selection based waveform concatenation systems which is much better statistical parameter speech synthesis (SPSS) systems. However, large speech corpus and expert fine-tuning are required for building a conventional waveform concatenation system. Much differently, a SPSS system is easy to set up which parametrizes waveforms and builds acoustic models to predict the acoustic features. In a SPSS system, a vocoder, such as STRAIGHT [5], is used to generate waveforms from the predicted features by acoustic model. And the prosody is mainly reconstructed by the duration model. The advantages in flexibility and light footprint of SPSS systems is much charming. However, as the lossy signal processing based vocoder and the over-smoothing problem of acoustic model, the fidelity of generated speech is limited. Meanwhile, the prosody of synthesized speech suffers from less expressive problem.

Recently, benefiting from the in-depth research of neural text to speech(TTS), since 2017, some systems in Blizzard have replaced conventional world vocoder with powerful neural vocoder such as WaveNet [6] to generate speech waveforms directly, which is a very deep auto-regressive model conditioned on previous sample points and could outperform than unit selection on speech quality. In last year's Blizzard Challenge, WaveNet showed good results. More deep work has been done such as in Deep Voice 1, 3 [7, 8] and Parallel WaveNet [9] for better performance. What's more, Generative adversarial Network(GAN) has been introduced to combine [10, 11] with Parallel WaveNet. To achieve more natural and expressive prosody, an end to end speech synthesis architecture named Tacotron [12] has been proposed by Yuxuan Wang et al.. Following with a neural vocoder such as WaveNet, neural TTS system was introduced by [13].

We followed the recent work of neural TTS to complete the task in Blizzard 2019 by implementing an end to end speech synthesis system with WaveNet vocoder. Firstly, for achieving high expressive speech, a modified multi-speaker Tacotron-like acoustic model is implemented to predict mel-spectrograms. Moreover, we proposed to use nonalignment full-context label directly which is widely used for statistical parametric speech synthesis duration modeling as the inputs of Tacotron-like model. Sentence embedding by pre-trained Bert [14] model is also incorporated to the acoustic model. A modified multi-speaker WaveNet is implemented to reconstruct waveforms from predicted mel-spectrograms. To enhance speech quality, we incorporated channel embedding in WaveNet. The results of subjective evaluation and our experiment showed the superiority of our submitted system.

This paper is organized as follows: Section 2 describes the details task in Blizzard 2019. Section 3 presents the implemented method adopted in our system. Section 4 introduces the subjective evaluation results. Finally, Conclusion and future work are presented in Section 5 in the end.

2. The task in Blizzard 2019

There is only single task this year. Single task 2019-MH1: Mandarin Chinese Found Data - About 8 hours of speech data from an internet talk show by a well-known Chinese character will be released. All data are from a single speaker. The task is to build a voice from this data that is suitable for expressive TTS. The task is a new challenge which is different from last year. In the following sections, we will introduce our speech synthesis system in details.

3. Tencent speech synthesis system

As showed in Figure 1, our speech synthesis system performs in an end to end manner. At training phase, we use a modified Festival front-end to predict phoneme, tone and other linguistic features firstly. Differently from SPSS, we do not use a HMM model for the force alignment. Instead, alignments are learned with attention mechanism in acoustic model. And then, sentence embeddings are generated by a pre-trained Bert model. Thirdly, we trained an attention based multi-speaker acoustic model which is a variant version of Tacotron 2 [13], nonalignment full-context label with sentence embedding are used as inputs here. Fourthly, we re-trained the model alone on the corpus offered by Blizzard 2019. Finally, a WaveNet neural vocoder conditioned on the mel-spectrograms is trained with

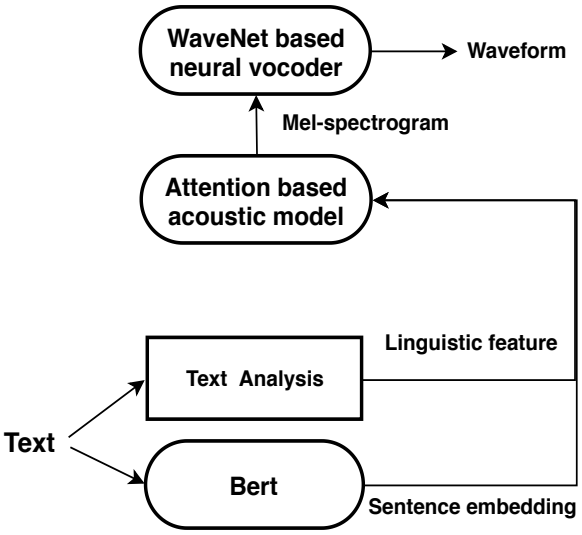

Figure 1: *System architecture.*

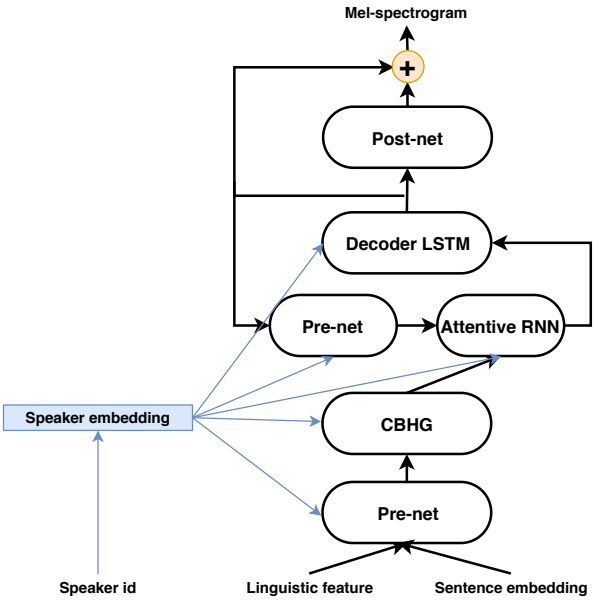

Figure 2: *The acoustic model architecture.*

the multi-speaker which consists of a lot of internal male Mandarin Chinese data and the offered data. Channel embedding with WaveNet is proposed by us to get better quality of generated waveforms.

At synthesis phase, full-context label is predicted by the Festival based front-end, the sentences embedding are generated by a Bert model. Then, we fed those into the acoustic model to predict the mel-spectrograms. Finally, the WaveNet with channel embedding is used to generate waveforms point by point conditioning on the predicted mel spectrograms.

3.1. Data Preparation

3.1.1. Linguistic features

Full-context label which is widely used in the parametric system as the inputs of duration model is incorporated by us to Tacotron-like model. Firstly, we adopted a modified festival frontend toolkit [15] to convert raw text to the phoneme level full-context linguistic features. Then, with a modified mixed-lingual question set, the HTS format [16, 17] full-context labels are generated, which consists of 559 questions, to binary features. Then the normalized binary features are taken as part of inputs to the acoustic model.

3.1.2. Acoustic features

All audios used in our system were firstly re-sampled to 24 kHz. The beginning silence and ending silence are trimmed to a fixed length. Then 80-dimensional mel-spectrograms with 256 hop-size were extracted from audios as the acoustic model target output. Our mel-spectrograms extracting methods is mostly as the same as Tacotron 2. The Pairs of mel-spectrograms and 16-bit audios were used in the training of WaveNet. We used 80 channel mel filter-bank spanning 50 Hz to 12 kHz after transforming the short-time Fourier transform(STFT) linear magnitude to the mel scale, followed by log dynamic range compression.

3.2. Attention based acoustic model

For achieving expressive speech synthesis, generating natural prosody is meaningful which is much hard for SPSS. We trained an end to end Tacotron-like acoustic model which is a sequence to sequence neural network with attention to convert linguistic features to mel-spectrograms instead of conventional SPSS model(duration and acoustic model). As showed in Figure 2, similar to Tacotron, convolution-bank-highway-GRU(CBHG) module was used in encoder. As in [18], we also employed the Graves Gaussian Mixture Model (GMM)-based monotonic attention mechanism [19]. After attentive rnn, two decoder LSTM layers are followed. At last, a CNN based post-processing network is used to enhance generated mel-spectrograms. Stop token is predicted as the same as in Tacotron 2, for simplicity, we omitted it in Figure 2.

3.2.1. Full-context label and sentence embedding

The conventional inputs for end to end speech synthesis is phoneme sequences or character sequences [13, 20]. As Mandarin Chinese is a tone based language, tone sequences could be added for better performance. Instead, we employ conventional input feature of duration model in SPSS. The results indicates full-context label is more efficient for training and controllable for synthesis. What's more, with full-context label, it's easy to train a mixed lingual neural TTS system as the same as SPSS which is part of Blizzard 2019's task. Recently, Bert has showed great success in many NLP tasks, we incorporated pre-trained Bert model as our sentence encoder to generate sentence embeddings. The full-context label and Bert based sentence embedding are combined as our end to end acoustic model's input. With sentence embedding, better prosody are predicted by acoustic model, especially in long sentences.

3.2.2. Multi-speaker acoustic modeling

As discussed above, the task of Blizzard 2019 is mainly to generate Mandarin Chinese and a little bit of English speech. To this end, it is worth nothing to train a mixed-lingual system. However, the corpus offered by Blizzard 2019 has a few english words which indicates it's necessary to use external data. We followed the work of multi-speaker Tacotron in [21] to build

a multi-speaker acoustic model incorporating the corpus offered with our huge internal male Mandarin Chinese data which includes many mixed-lingual speakers. As showed in Figure 2, speaker embeddings are incorporated into CBHG encoder, attention rnn and decoder LSTM to make those part speaker dependent except post-processing network. After multi-speaker acoustic model converged, we continued finetuning the model with the corpus of Blizzard 2019 a bit of time for better speech quality.

3.2.3. Loss for acoustic modeling

As proposed in [22], L_2 loss with an additional L_1 loss could be more robust on noisy training data. We tried it on our experiments as the corpus of Blizzard 2019 is a bit noisy, it's practical as it claimed. But tone error problem has arisen when using the combining loss. Obviously, Mandarin Chinese is a tone based language which is much different from English. To fix this, we modified the combining strategy. Firstly, only L_2 loss was used as the same as in [13] for robust tone learning, then only L_1 loss was used for generating more clean mel-spectrograms.

As claimed in LibriTTS [23], model configuration could influence the generated speech quality for male speakers, such as filter-bank configuration and modeling dependency of time-domain signals by neural vocoder. Meanwhile, we observed that the energy of the corpus used in Blizzard 2019 mostly centralizes in very low frequency. With this prior, a low frequency priority loss was used in acoustic modeling.

3.3. WaveNet based neural vocoder

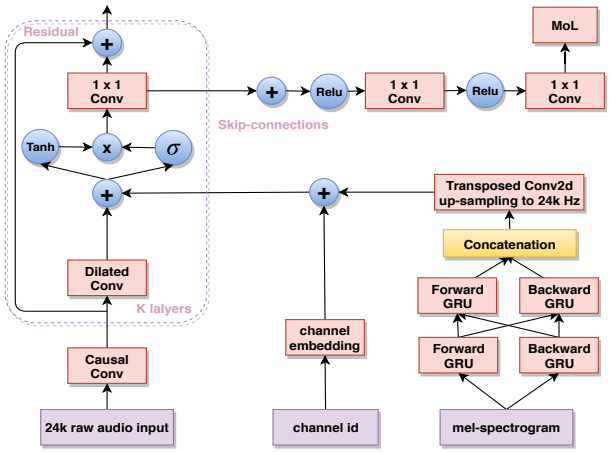

Figure 3: *The modified WaveNet architecture.*

In order to improve the fidelity of speech, we chose WaveNet as our neural vocoder which is a powerful probabilistic and auto-regressive generative model that models waveforms directly in time domain instead of conventional signal processing based non-trainable vocoder. The joint probability of a given waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ is conditioned over all previous time steps:

$$p(\mathbf{x}, \mathbf{c}, \mathbf{h}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}, \mathbf{h}, \mathbf{c}). \qquad (1)$$

Here, c is local conditional inputs, whereas h is the model parameters. For speech synthesis, the predicted mel-spectrograms are used as local condition.

As Blizzard 2019 offered data set is a bit small, we trained a multi-speaker WaveNet incorporating offered data and huge internal male Mandarin Chinese data without speaker embedding as global condition. As shown in Figure 3, We adopted a variant version of the WaveNet architecture modified from [9] and [7]. The same as in the architecture in [13], our model consists of 24 dilated CNN layers, grouped into 4 residual block stacking with 6 layers. For every stack, the dilation rate increases by a factor of 2 in every layer, and no dilation for the first layer.

Parallel WaveNet uses a 10-component mixture of logistic(MoL) distributions to generate 16-bit samples at 24 kHz instead of the 8 or 10 bit softmax layer to reserve high fidelity in generated speech. However, the model needs longer to converge with MoL, to fix this, we used a pre-trained model as the initial model for the multi-speaker model which converged fast as speaker embedding are not used. After training on ground-truth mel-spectrograms, model is also trained on the predicted. Benefiting from the powerful model, the generated waveforms on predicted mel-spectrograms sounds almost as clean as the ground-truth.

As claimed in [7], an inappropriate condition network could suffer from pronunciation problem on the predicted mel-spectrograms which could be much obvious in tone based language as WaveNet converged like the right one. We used the same architecture of in [7], but replaced QRNN with GRU for simplification as in [24]. So, the mel-spectrograms are firstly fed into a stack of two bidirectional GRU layers with 256 unit. Then four transposed conventional layers are used to up-sampled with 256 times to 24 kHz as hop size is 256. Finally, the condition information is incorporated to every dilated layer. With the modified condition network, the generated speech sounds highly intelligibility.

3.3.1. Channel embedding

Compared to our internal data, the offered data by Blizzard 2019 is more noisy with about 21 kHz sampling rate and is probably processed by MP3 format in a low bit rate. We tried to generate high quality waveforms as our internal data at 24 kHz sampling rate by multi-speaker WaveNet. To this end, we proposed channel embedding in multi-speaker WaveNet. The learnable embeddings were designed to distinguish the quality of waveforms into two parts in a speaker independent manner. As shown in Figure 3, channel embedding are injected to every dilated CNN layers as global condition. So, Equation 1 needs to be re-defined as:

$$p(\mathbf{x}, \mathbf{c}, \mathbf{g}, \mathbf{h}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}, \mathbf{h}, \mathbf{c}, \mathbf{g}). \qquad (2)$$

Here, g is channel embedding.

As we have huge internal male Mandarin Chinese data, data augments is a obvious method to make the embedding learning well. Firstly, we reserved the original internal data at 24 kHz as high quality data and down-sampled the copy of internal data to 21 kHz encoded with MP3 format with 64 bit rate. Then, the low quality waveforms were up-sampled back to 24 kHz encoded with PCM format. At last, We got two part internal data at 24 kHz sampling rate for generating input-output data pairs, one is of low quality as the offered data, another is of high quality. In training phase, low quality processed internal data and offered data in Blizzard 2019 is with channel id 0, another part is 1. In inference phase, only channel id 1 is used for high quality. We also tried to use one hot instead of learnable

embedding, our results indicated a small embedding is much better for this task.

4. Subjective results

A total of 26 systems were evaluated at last, 24 from participating teams, one benchmark Merlin [25] and one natural speech. Our system is identified as W. System A is natural speech recorded by the original speaker, whereas system B is the Merlin speech synthesis system. System C to Z are the 24 participating teams.

Table 1: *Task 2019-MH1*

Sections	Detailed Description
section 1	MOS
section 2	Similarity
section 3	Pinyin (without tone) Error Rate
section 4	Pinyin (with tone) Error Rate

The evaluation criteria includes four sections as shown in Table 1. The synthesized and natural audios were carefully scored in every section by three types of listeners who is involved by paid listeners, online volunteers and speech experts. Overall, our system shows good performance in most of the challenge criteria, especially in naturalness test.

4.1. Naturalness test

The boxplot evaluation results of all systems on naturalness is showed in Figure 4. System M, W, I perform better than other systems, whereas most system are obviously better than Merlin benchmark system B. In synthesized systems, the result of our system W shows no significant gap to system I and better than most of the participation, only system M is obviously better than ours. It's proved that our end to end speech synthesis system has shown superiority over other systems as using the channel embedding enhanced multi-speaker WaveNet.

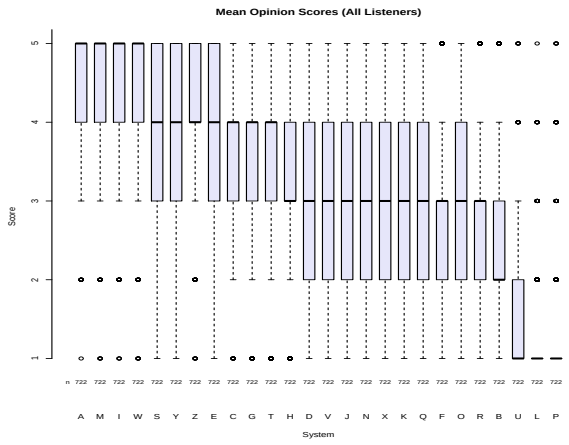

Figure 4: *Boxplot of naturalness scores of each submitted system for all listeners*

4.2. Similarity test

Figure 5 presents the mean opinion of similarity evaluation results for all systems which are scored by all listeners. In this section, each listener should score the synthesized audio in two fixed reference samples of natural speech for all systems. Our system achieves better result than system I and shows significant advantage than benchmark system and many participants. As channel embedding was adopted in multi-WaveNet for high quality generated waveforms which has more energy in medium frequency, the similarity to the corpus offered in Blizzard 2019 which almost only has energy in very low frequency drops a little compared to the result in naturalness test.

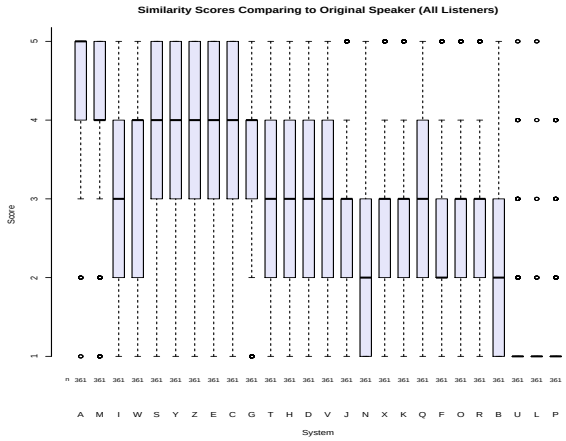

Figure 5: *Boxplot of similarity scores of each submitted system for all listeners*

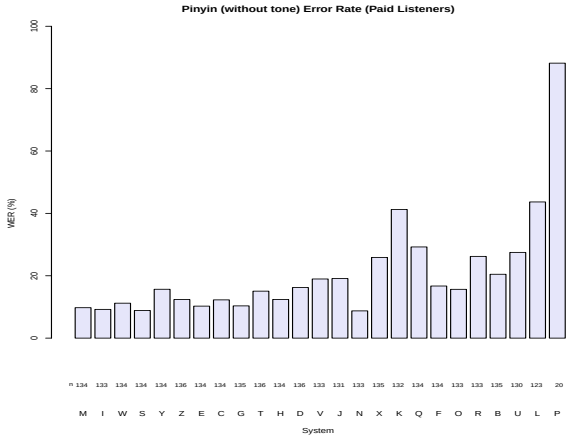

Figure 6: *Pinyin without tone error rate scores of each submitted system for paid listeners*

4.3. Pinyin and tone error rate test

The Pinyin without tone error rate(PER) and Pinyin with tone error rate(PTER) of all participant systems are presented separately in Figure 6 and Figure 7 which are only scored by paid listeners. PER and PTER could show the intelligibility of each system. In this two sections, system N performs much better than other participants, whereas the result of system N in naturalness and similarity test are not so good. The scores of System N, M, I, S, etc. are higher than us which indicates that our end to end acoustic model has more work to do in optimization for intelligibility.

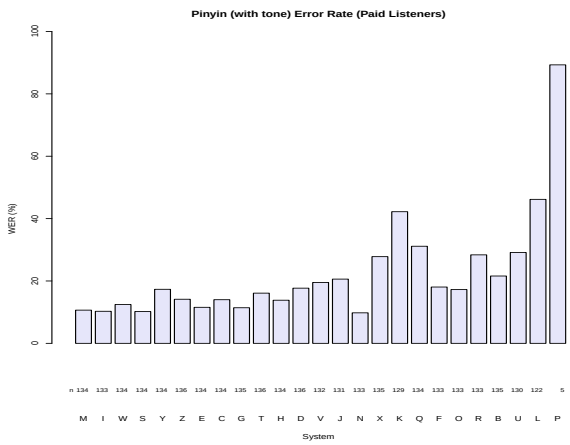

Figure 7: *Pinyin with tone error rate scores of each submitted system for paid listeners*

5. Conclusions and future work

This paper presents the details of our submitted system and the results in Blizzard Challenge 2019. We built a GMM attention based end to end speech synthesis followed by a WaveNet vocoder at 24 kHz sampling rate. Our system achieved good performance in the most criterion for this challenge, especially in naturalness test. In attention based acoustic model, we proposed a new features representation as the input which combines nonalignment full-context label with Bert sentence embedding for efficiency of training and controllability in inference phase. For more clean waveforms and better tone learning, we adopted a new strategy L_1 and L_2 Loss. In WaveNet based neural vocoder, we proposed a channel embedding to enhance the generated speech fidelity.

A stepwise monotonic attention [26] is proposed for robust end to end TTS. We will also make attempts to build a more robust end to end acoustic model in our future work.

6. References

[1] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen *et al.*, "The ustc and iflytek speech synthesis systems for blizzard challenge 2007," in *Blizzard Challenge Workshop*, 2007.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[4] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[5] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[6] 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." in *SSW*, 2016, p. 125.

[7] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 195–204.

[8] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[9] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[10] Q. Tian, X. Wan, and S. Liu, "Generative adversarial network based speaker adaptation for high fidelity wavenet vocoder," *arXiv preprint arXiv:1812.02339*, 2018.

[11] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," *arXiv preprint arXiv:1904.04472*, 2019.

[12] 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: A fully end-to-end text-to-speech synthesis model," *arXiv preprint*, 2017.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen, "The festival speech synthesis system, version 1.4. 2," *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*, vol. 6, pp. 365–377, 2001.

[16] H. Zen, "An example of context-dependent label format for hmm-based speech synthesis in english," *The HTS CMUARCTIC demo*, vol. 133, 2006.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[18] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[19] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[20] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," *arXiv preprint arXiv:1904.06063*, 2019.

[21] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[22] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[24] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding."

[25] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[26] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," *arXiv preprint arXiv:1906.00672*, 2019.