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

# The Tencent speech synthesis system for Blizzard Challenge 2020

*Qiao Tian, Zewang Zhang, Ling-Hui Chen, Heng Lu, Chengzhu Yu, Chao Weng, Dong Yu*

Tencent, China

{briantian, zewangzhang, nedchen, bearlu, czyu, cweng, dyu}@tencent.com

## Abstract

This paper presents the Tencent speech synthesis system for Blizzard Challenge 2020. The corpus released to the participants this year included a TV's news broadcasting corpus with a length around 8 hours by a Chinese male host (2020-MH1 task), and a Shanghaiese speech corpus with a length around 6 hours (2020-SS1 task). We built a DurIAN-based speech synthesis system for 2020-MH1 task and Tacotron-based system for 2020-SS1 task. For 2020-MH1 task, firstly, a multi-speaker DurIAN-based acoustic model was trained based on linguistic feature to predict mel spectrograms. Then the model was fine-tuned on only the corpus provided. For 2020-SS1 task, instead of training based on hard-aligned phone boundaries, a Tacotron-like end-to-end system is applied to learn the mappings between phonemes and mel spectrograms. Finally, a modified version of WaveRNN model conditioning on the predicted mel spectrograms is trained to generate speech waveform. Our team is identified as *L* and the evaluation results shows our systems perform very well in various tests. Especially, we took the first place in the overall speech intelligibility test.

**Index Terms**: Blizzard Challenge, DurIAN, Tacotron, WaveRNN , Speech Synthesis, attention networks

## 1. Introduction

The Blizzard Challenge has been organized annually since 2005 with the purpose to evaluate different speech synthesis techniques based on same provided training database. It has made great contribution to the speech synthesis society. Conventional speech synthesis methods, like waveform concatenation [1] and statistical parameter speech synthesis (SPSS) [2, 3, 4] or the hybrid of the above mentioned two [5, 6], has dominated the Challenge for a long time. Unit selection and concatenation system [1] outperforms conventional SPSS system in speech quality because it directly selects and concatenates natural speech recordings. However, a large and carefully labelled speech corpus is a must for Unit selection and concatenation system. And expert knowledge is needed to tune various weights in the system to balance amongst different features. On the contrary, the advantages of SPSS systems lies in its flexibility in modification and small in footprint size. However, the speech quality of the generated speech is harmed by the parameter over-smoothing problem in both acoustic modeling step and vocoder speech rendering step.

In the recent 5 years, speech synthesis systems based on deeper and larger neural networks models have achieved near to human speech synthesis naturalness. In particular, large auto-regressive networks based models like Tacotron and WaveNet [7] take advantage of the causal properties of speech, and demonstrate their strong performance over traditional speech synthesis methods. However, conventional attention based end-to-end methods are lack of robustness when generating speech as they produce unpredictable artifacts where random words in the source text are repeated or skipped in generated speech [8, 9] esp. In [10], we have proposed the Duration Informed Attention Network for multimodal speech synthesis, in which a separate phoneme duration model is trained along with attention based acoustic neural network models. During generation process, phone duration sequence is first generated by phoneme duration model, then the encoded text embedding is expended by repeating according to the generated phone duration. Finally, expended text embedding is input into local attention networks to render output states. Experiments indicate that the proposed DurIAN system could synthesize speech with the naturalness and quality on par with the current state of the art end-to-end system Tacotron 2, at the same time effectively avoid the word skipping and repeating errors in generated speech [10].

We followed our recent works to build TTS systems for the two task in Blizzard Challenge 2020. In 2020-MH1 task, a robust neural speech synthesis system with AdaDurIAN and WaveRNN vocoder was built. In order to generate high fidelity, the modified version of WaveRNN vocoder was modeled to generate 48 kHZ waveform directly. In task 2020-SS1, an end-to-end speech synthesis system was trained with Tacotron acoustci model and WaveRNN vocoder, without extra alignment modules. Moreover, we trained our SS1 system on the phoneme sequence offered by blizzard challenge host without using any extra front-end and data set. Multiple subjective and objective tests conducted by the Challenge demonstrate the robustnestt and naturalness of our systems.

This paper is presented as follows: Section 2 describes the details tasks in Blizzard 2020. Section 3 introduces our implemented systems for the two tasks in detail. Section 4 introduces the subjective evaluation results. Finally, Conclusion and future work are given in Section 5.

## 2. The task in Blizzard 2020

There are two tasks this year. One task 2020-MH1: Mandarin Chinese Found Data - About 8 hours of speech data from an TV news by a Chinese male host. All data are from a single speaker. The task is to build a voice from this data that is suitable for expressive TTS. The other task 2020-SS1: Shanghaiese Found Data. About 6 hours of speech data. For 2020-MH1, we adopt the AdaDurIAN model, and for 2020-SS1, we adopt a GMM-based tacotron-like end to end acoustic model from scratch without extra hard-align information and front end processing. The overall architecture is shown in Figure 1. In the following sections, we will introduce our speech synthesis system in details.

## 3. Tencent speech synthesis system

As showed in Figure 2, 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 prosody



10.21437/VCC_BC.2020-4

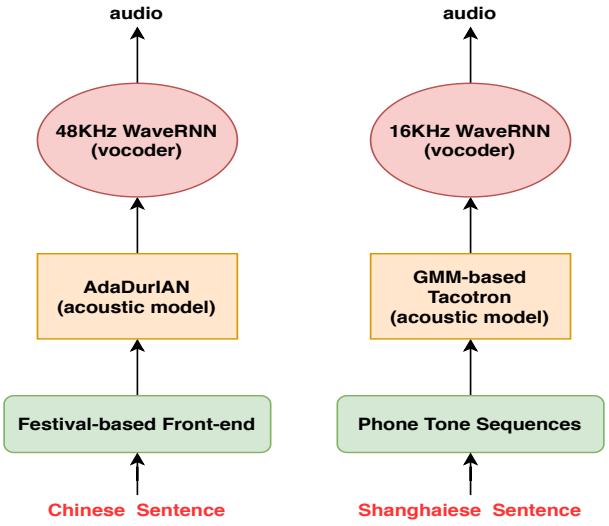


Figure 1: *Overall architecture of our systems for Blizzard 2020.*

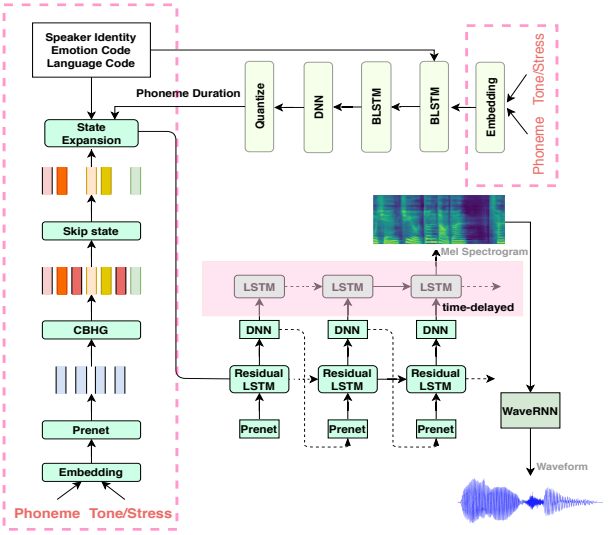


Figure 2: *2020-MH1 System architecture.*

boundaries firstly. Although we can use knowledge distillation to get duration from a Tacotron-like model, we found that using HMM-based force-align model makes the duration more accurate. And then, representations of phonemes, tones and prosody boundaries are generated from a CBHG-like content encoder, which is speaker independent. Thirdly, we trained a multi-speaker acoustic model based on AdaDurIAN [11], which is a variant of DurIAN [10]. Code-switch language embedding is adopted here. Fourthly, we re-trained the model alone on the corpus offered by Blizzard 2020. Finally, a WaveRNN neural vocoder conditioned on the mel-spectrograms is trained with the multi-speaker which consists of a lot of internal male Mandarin Chinese data and the offered data.

At synthesis phase, phones, tones and prosody boundaries are predicted by the Festival based front-end tool. Then, we fed those sequences into the acoustic model to predict the mel-spectrograms. Finally, the WaveRNN 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 AdaDurIAN model. Firstly, we adopted a modified festival frontend toolkit [12] to convert raw text to the phoneme level full-context linguistic features. Then, with a modified mixed-lingual question set, the HTS format [13, 14] 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 for extracting acoustic features. The beginning silence and ending silence are trimmed to a fixed length. Then 80-dimensional mel-spectrograms with 240 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 WaveRNN. 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.

On one hand, it's pretty hard for SPSS to generate expressive speech, on the other hand, Tacotron-like acoustic models often meet word skipping, missing or repeating errors. Therefore, we adopted a robust and controllable acoustic model AdaDurIAN for 2020-MH1, which is built by a sequence to sequence neural network with windowed content-based attention to convert linguistic features to duration of each phoneme and then predict mel-spectrograms. As showed in Figure 2, similar to Tacotron, convolution-bank-highway-GRU(CBHG) module was used in encoder. The encoder is speaker-independent for improved generalization when faced with low-resource settings. The output states of encoder is expanded according to the duration predicted by a bi-directional LSTM model. As described in [11], we add speaker identity embedding, language embedding and emotion embedding to the expanded encoder states. The input of decoder is generated by a windowed content-based attention. After attentive rnn, two decoder LSTM layers are followed. At last, a time-delayed LSTM layer is used to enhance generated mel-spectrograms.

#### *3.1.3. Pretrained acoustic modeling*

As discussed above, the task of Blizzard 2020-MH1 is mainly to generate Mandarin Chinese. To this end, it is worth nothing to train a mixed-lingual system. However, the corpus offered by Blizzard 2020 has a few english words which indicates it's necessary to use external data. We followed the work of multi-speaker Tacotron in [15] 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 2020-MH1 a bit of time for better speech quality and higher speaker similarity.

### 3.2. WaveRNN based neural vocoder

At last year, a powerful enhanced WaveNet neural vocoder [16] was adopted by us to achieve high quality speech. But WaveNet is too heavy for speech generation. In this year, we adopt a WaveRNN based neural vocoder for fast batch generation on GPU, which has very little parameters compared to WaveNet. As claimed in [17], WaveRNN-2048 can get no significant differences with WaveNet in MOS test. So a big WaveRNN with GRU hidden size 2048 was used by us to obtain high quality generated speech. The WaveRNN model the waveform at 48 kHZ sampling rate of as the offered in task 2020-MH1 for high fidelity speech synthesis, which conditioned on conventional 24 kHZ mel spectrograms produced by acoustic model. For 2020-SS1, a 16 kHZ WaveRNN was adopted as the corpus is sampling at 16 kHZ. All the neural vocoder adopted are the variant of WaveRNN, which was proposed in FeatherWave [18]. We used 12bit $\mu$-law and 4 band to efficiently model discretized waveform and fast generation. Condition network consists of five $1x3$ convolution layers with channel size 512. Before passed into sample rate network, extracted local features are repeated to match the sampling rate of target waveform.

Specially, as acoustic features are extracted with hop size 240 from 24 kHZ audio in task 2020-MH1, local features are repeated 480 times to generate 48 kHz waveform.

## 4. Subjective results

### 4.1. 2020-MH1

A total of 17 systems were evaluated at last, 16 from participating teams and one natural speech. Our system is identified as $L$. System $A$ is natural speech recorded by the original speaker. System $B$ to $Q$ are the 24 participating teams.

Table 1: *Task 2020-MH1*

| Sections | Detailed Description |
|---|---|
| section 1 | SIM (news) |
| section 2 | SIM (PSC) |
| section 3 | MOS (news) |
| section 4 | MOS (PSC) |
| section 5 | MOS (news) |
| section 6 | INT |

The evaluation criteria includes six 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 INT test.

#### *4.1.1. Naturalness test*

The boxplot evaluation results of all systems on naturalness is showed in Figure 3. System $O$, $I$, $L$ perform better than other systems. It's proved that our DurIAN-based speech synthesis system has shown superiority over most other systems.

#### *4.1.2. Similarity test*

Figure 4 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

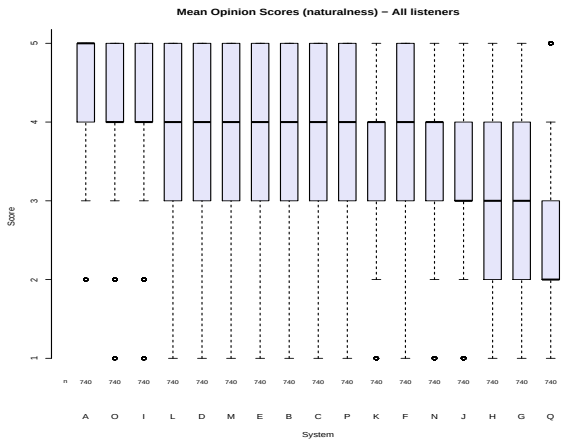


Figure 3: *Boxplot of naturalness scores of each submitted system for all listeners*

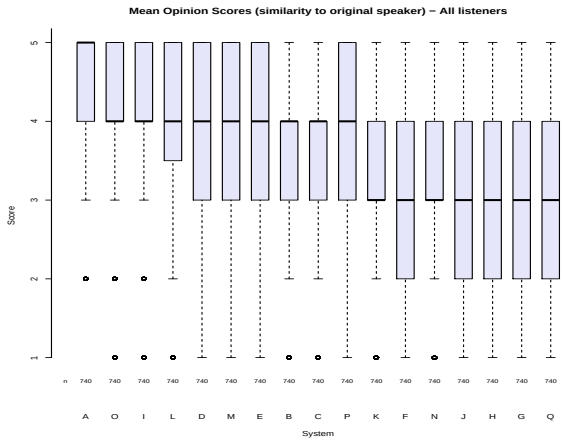


Figure 4: *Speaker Similarity Scores of each submitted system for all listeners*

system achieves the third highest score and shows significant advantages than many other participants.

#### *4.1.3. Pinyin Error Rate with tone test*

The Pinyin with tone error rate(PTER) of all participant systems are presented separately in Figure 5 which are only scored by paid listeners. In this test, system $O$ and system $I$ perform better than other participants. Our system $L$ still has the third highest score of speaker similarity.

### 4.2. 2020-SS1

A total of 9 systems were evaluated at last, 8 from participating teams and one natural speech. Our system is identified as $L$. System $A$ is natural speech recorded by the original speaker.

#### *4.2.1. Naturalness test*

The boxplot evaluation results of all systems on naturalness is showed in Figure 6. The top three systems $I$, $E$, $L$ perform similarly. It's proved that our Tacotron-based speech synthesis system has shown superiority over most other systems.

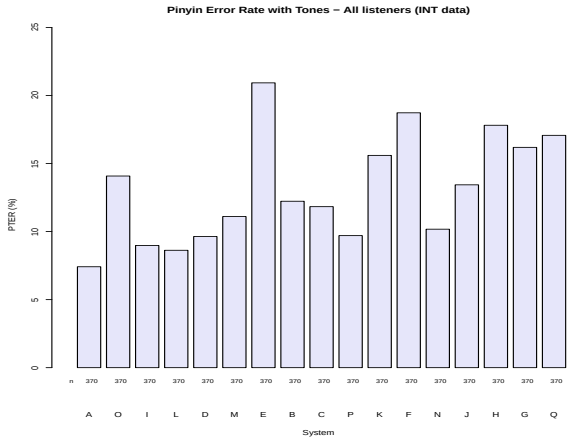


Figure 5: *Pinyin Error Rate with Tones of each submitted system for all listeners*

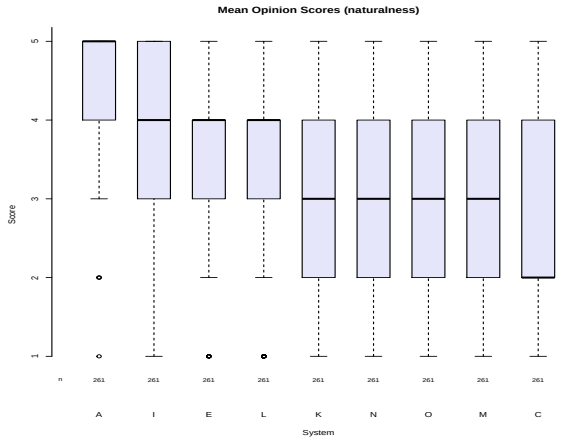


Figure 6: *Boxplot of naturalness scores of each submitted 2020-SS1 system for all listeners*

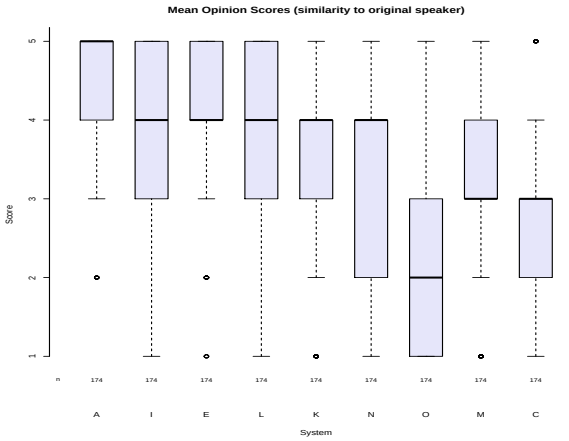


Figure 7: *Speaker Similarity Scores of each submitted 2020-SS1 system for all listeners*

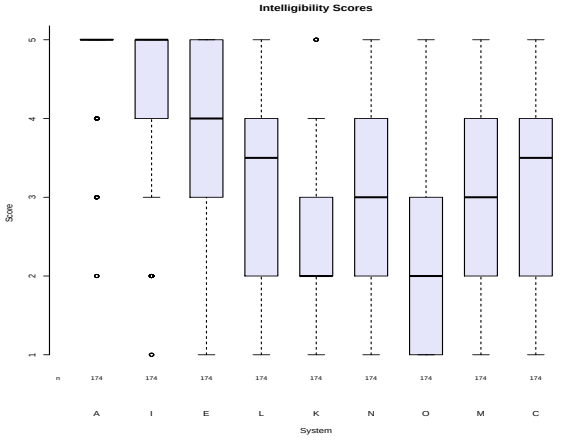


Figure 8: *Intelligibility scores of each submitted 2020-SS1 system for all listeners*

#### *4.2.2. Similarity test*

Figure 7 presents the mean opinion of similarity evaluation results for all systems which are scored by all listeners. It's obvious that system $E$ performs best. Our system $L$ achieves the second highest score and still performs better that many other systems.

#### *4.2.3. Intelligibility Test*

The intelligibility scores of all participant systems are presented in Figure 8. In this test, system $I$ performs much better than other participants. Our system $L$ still has the third highest score of intelligibility.

## 5. Conclusions and future work

This paper presents the details of our submitted system and the results in Blizzard Challenge 2020. We built a duration-informed AdaDurIAN system for 2020-MH1 task and GMM attention based end to end speech synthesis for 2020-SS1 task, both systems are followed by a high quality WaveRNN vocoder. Especially, the 48 kHZ sampling rate waveform as the original audio in 2020-MH1 were reconstructed by WaveRNN with conventional 24 kHZ melspec features. Our system achieved good performance in most sections of test for this year's challenge, the highest score was obtained by us in intelligibility test.

Recently, transformer-based parallel TTS system FastSpeech2 [19] with pitch and energy variances as input has been proposed, and GAN based parallel vocoder such as GAN-TTS [20], HooliGAN [21], VocGAN [22] get compared MOS scores with the powerful auto-regressive neural vocoder, such as WaveRNN. We will also make attempts to build a more robust, high quality and controllable parallel TTS 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] Z.-H. Ling, H. Lu, G.-P. Hu, L.-R. Dai, and R.-H. a. Wang, "The ustc system for blizzard challenge 2008," in *Blizzard Challenge Workshop*, 2008.

[6] H. Lu, Z.-H. Ling, M. Lei, *et al.*, "The ustc system for blizzard challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, "Adadurian: Few-shot adaptation for neural text-to-speech with durian," *arXiv preprint arXiv:2005.05642*, 2020.

[12] 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.

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

[14] 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.

[15] 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.

[16] Q. Tian, J. Chen, and S. Liu, "The tencent speech synthesis system for blizzard challenge 2019," in *Proc. Blizzard Challenge Workshop*, 2019.

[17] 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.

[18] Q. Tian, Z. Zhang, H. Lu, L.-H. Chen, and S. Liu, "Featherwave: An efficient high-fidelity neural vocoder with multi-band linear prediction," *arXiv preprint arXiv:2005.05551*, 2020.

[19] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[20] 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," *arXiv preprint arXiv:1909.11646*, 2019.

[21] O. McCarthy and Z. Ahmed, "Hooligan: Robust, high quality neural vocoding," *arXiv preprint arXiv:2008.02493*, 2020.

[22] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.