The NLPR Speech Synthesis entry for Blizzard Challenge 2019

Jianhua Tao[1,2,3], Ruibo Fu[1,2], Zhengqi Wen[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{ jhtao, ruibo.fu, zqwen}@nlpr.ia.ac.cn

Abstract

The paper describes the CASIA speech synthesis system entry for Blizzard Challenge 2019. About 8 hours of speech data from online talkshow is adopted as the training data for the construction this year. Our synthesis system is built based on the multi-speaker end-to-end speech synthesis system. And LPCNet based neural vovoder is adpted to improve the quality . Different from our previous system, some improvements about data pruing and speaker adaptation strategies were made to improve the robustness of our system. In this paper, the whole system structure, data pruing method, and duration control will be introduced and discussed. Finally, the results of the listening test will be presented.

Index Terms: speech synthesis, data pruing, LPCNet, end-to-end, Blizzard Challenge 2019

1. Introduction

This paper describes details about our sixth entry speech synthesis for Blizzard Challenge. The task of this year is to build a voice from the provided data, suitable for expressive text-to-speech (TTS) from plain text input. Mandarin Chinese speaker collected from talk shows is full of expressiveness and the quality of the recording speech is low with missing the high frequency information. Besides, the speech of talk shows speaker is very high which decrease the accuracy of the force aglignment by Automatic Speech Recognition (ASR) technologies.

End-to-end speech synthesis have made rapid progress in recent years, and achieved state-of-art performance [1–3]. The single speaker speech synthesis system, usually neutral speaking style, is approaching the extreme quality close to human expert recording. And the interests in style control and speaker adaptation speech synthesis with limited corpus also keep rising. Recently, there also published many promising works in this topic, such as transferring prosody and speaking style within or cross speakers based on end-to-end TTS model [4–6].

Generally, speaker adaptive training methods mainly focus on two aspects. One aspect is the speaker style representations, which is the one of inputs in the system. The methods mainly used fixed global speaker style representations for speaker recognition, such as i-vectors [7], d-vector [8]. It is not optimal for the multi-speaker speech synthesis and adaptation task. Therefore, methods [9, 10] that extracted trainable speaker representations from waveform were proposed in the statistic parametric speech synthesis framework. In the end-to-end speech synthesis framework, [2,11] use trainable speaker embeddings for multi-speaker speech synthesis. Besides, an extension to the Tacotron speech synthesis architecture that learns a latent embedding space of prosody, derived from a reference acoustic representation containing the desired prosody was proposed [4]. Another aspect is vocoder, which is the output of the system. Speaker dependent layers with vocoder STRAIGHT [12], WORLD [13] is applied in the conventional statistical parametric speech synthesis method. And the speaker embedding information was also applied in the neural network based vocoder like WaveNet [14], SampleRNN [15] in the voice conversion tasks [16-19].

To select a vocoder which is apt for adaptative training and online system. Recent research work [20] reported that LPCNet could achieve significantly higher quality than WaveRNN [21] for the same network size and that high quality LPCNet speech synthesis is achievable with a complexity under 3 GFLOPS. Therefore, we applied LPCNet based spectrogram-to-audio neural vocoder use it with Tacotron as a replacement for Griffin-Lim audio generation. And we used the trainable speaker embedding from Tacotron to do an adaptative training on LPCNET.

Considering the low quality of provided recording, we use the following stragies to improve the performance of our system. Firstly, we use a data puring model to automatic select the unmatching data pairs, which improves the efficiency of manual checking process and shorts the time on the database building. Combining with manual checking, the performance of our submitted final system is improved significantly. Secondly, we add phone during predicting model to control the speech generation by Tacotron, which prevents the speed the synthestic speech is too high. It improves the intelligibility and rhythm of the synthestic speech. Thirdly, we using bandwith expansion technologies by adding different style embeddings to further improve the quality of synthestic speech based on our own external data. Fourthly, the adaptative LPCNet is adopted to further improve the quality of synthestic speech.

The rest of the paper is organized as follows. Section 2 gives an overview of our methods used for system construction. Section 3 gives an detailed introduce of the database processing. Section 4 introduces the phone duration model. In section 5, the evaluation results of our system in Blizzard Challenge 2019 are shown and discussed. The conclusions are presented in section 6.

2. System Overview

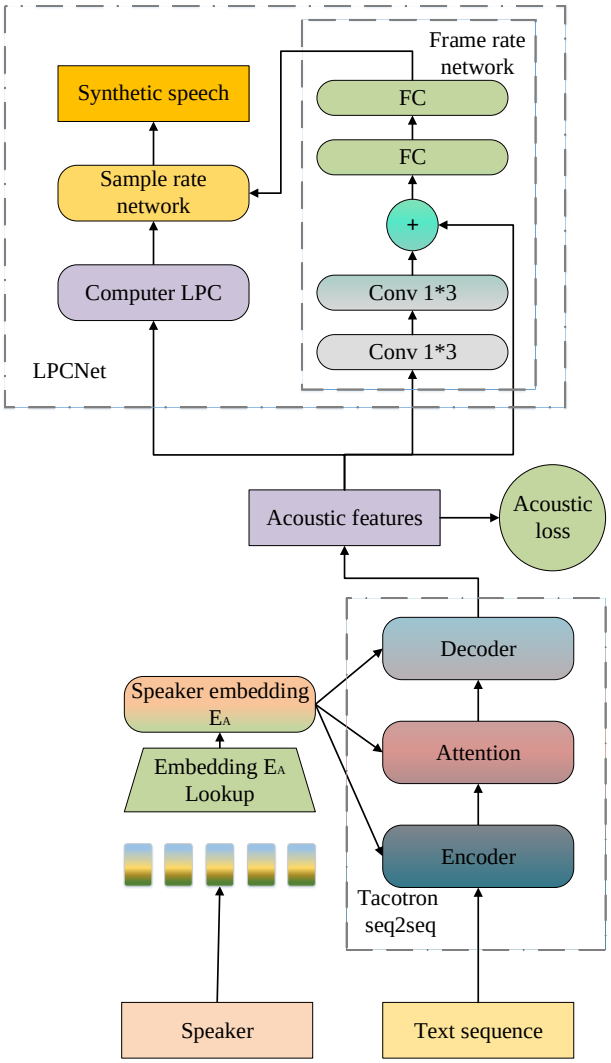

Figure 1: An overview of our system.

The whole network consists of two components, as shown in the Figure 1. Tacotron as proposed in [21] does not include explicit modeling of speaker identity; however, due to the flexibility of all neural sequence-to-sequence models, learning multi-speaker models via conditioning on speaker identity is straightforward. In the acoustic model part, we follow a modified scheme based on [11] to model multiple speakers. For each speaker in the dataset, a R^{d_s} embedding vector is initialized with Glorot [22] initialization. And we form the separate speaker embeddings $\hat{E}_A$. For each step of the training process, the speaker embeddings would be updated. For each example, the d_s -dimensional speaker embedding corresponding to the true speaker of the example is concatenated to encoder, decoder and attention of the Tacotron. In the neural vocoder part, we deploy the LPCNet, which significantly improve the efficiency of speech synthesis and remain high quality. In the frame rate network of LPCNet, we combine the trainable speaker embeddings from Tacotron with the output of convolution layer and the acoustic features that Tacotron predict.

3. Database Processing

The database of this year task is very chanllenge. Filrstly, the speed of the talking speaker is very high, which makes it hard for manual segmentation. We use an ASR model and silence detection model to automatic segmentation of the provided data. But there are still a lot of mistakes in the corpus. We use the forced aglignment by the ASR technologies to further check the matching between the audio and the text. Furthermore, we also use a trained model to eliminate the unmatching audio-text data pairs. The model can the find the mistakes of the text more thouroughly than the forced aglignment by ASR.

Due to the low quality of the provided corpus, we also use external data to improve the performance of our submitted system. The external data could improve the stability and quality of our system. All the data are processed by the same technologies and checked by annotator.

4. Phone Duration Modeling

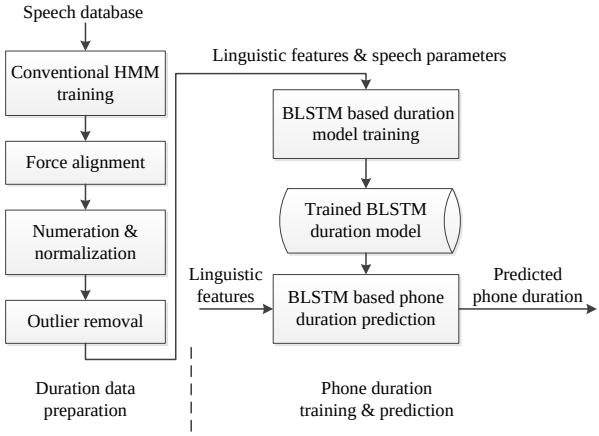

Figure 2: *BLSTM based phone predition in our system.*

The prosody in the provided database is not very well, thus we use external data to train a duration model separately. The duration model is added to the input of end-to-end system by combining the word embedding and duration embedding together to control the prosody output to adapt different styles. In these tasks, its powerful sequence modeling has been proved. Here, we consider duration modeling at phone level, for the audiobook data, using BLTSM. BLSTM based duration method with outlier removal is shown in Figure 2. This framework is general, thus it is easy to replace BLSTM with other machine learning methods for duration prediction purpose. Some important steps involved are briefly discussed below.

In duration data preparation part, force alignment is carried out at phone level after conventional HMM training. This step is to segment each utterance into a sequence of shorter and simple speech units, thus each unit can be modelled independently in subsequent steps. Unlike decision tree, BLSTM can only handle numeric features, thus it is necessary to encode all nominal features to be numeric values. Normalization is immediately carried out to transfer feature values into a limited interval.

In phone duration training and prediction part, BLSTM is trained with “cleaned” training samples and stored. After that, phone level duration are predicted by BLSTM for any given linguistic features of full context label.

5. System Building

5.1. Speech database

The speech database is An estimated 8 hours of speech from one native Mandarin Chinese speaker collected from talk shows for the Blizzard Challenge 2019, which is recorded by Zhenyu Luo. It contains about different speaking styles, which is recorded by different time and environment. Each provided utterance is 60 seconds, which is long for the training.

We also build a multi-speaker speech synthesis system by adding external The composition of the external Mandarin database is shown in the Table 1. All the wav files are sampled at 24kHz. And apply a first-order pre-emphasis filter $E(z) = 1 - \alpha z^{-1}$ to the training data, with $\alpha = 0.97$. In this work, we limit the input of the synthesis to just 22 features: The 20-dim Bark-scale [23] cepstral coefficients, and 2 pitch parameters (period, correlation) are extracted directly from recorded speech samples. The cepstrum uses the same band layout as [24] and the pitch estimator is based on an open-loop cross-correlation search. The input text is processed by our G2P frontend to transform to the phone sequences with tone in vowel.

Table 1: Composition of external Mandarin database

Sentence number		Training	Validation	Test	Speaker
Training	Large set	9,000×8	500×8	500×8	2 M 6 F
	Small set	900×4	50×4	50×4	1 M 3 F
	Total	75,600	4200	4200	3 M 9 F
(M=Male; F= Female)					

The task (**Single task 2019-EH1**) is to Build a voice from the provided data, suitable for expressive text-to-speech (TTS) from plain text input.

5.2 Building system

For the Tacotron training, we set output layer reduction factor $r = 2$. We use the Adam optimizer [25] with learning rate decay, which starts from 0.001 and is reduced to 0.0005, 0.0003, and 0.0001 after 500K, 1M and 2M global steps, respectively. The post-processing net is discarded. We use a simple loss for seq2seq decoder, which is the acoustic loss. Besides, the "stop token" prediction [1] is used during inference to allow the model to dynamically determine when to terminate generation instead of always generating for a fixed duration, which is the stop token loss. The combined cost is the sum of PIP loss, acoustic loss and stop token loss with equal weights. We train using a batch size of 32, where all sequences are padded to a max length. For the Tacotron adaptation, the learning rate is set to 0.0001. After about 600K global steps, there are about 2-3K global steps for adaptative training.

For the LPCNet training, the network was trained for 120 epochs, with a batch size of 64, each sequence consisting of 15 10-ms frames. We use the AMSGrad [26] optimization method (Adam variant) with a step size $\alpha = \frac{\alpha_0}{1+\delta \cdot b}$ where $\alpha_0 = 0.001$, $\delta = 5 \times 10^{-5}$, and b is the batch number. For the LPCNet adaptation, there are about 10 epochs for adaptative training.

Our proposed system consists two part: Acoustic model based on Tacotron, neural vocoder based on LPCNet. Therefore, we have following groups of baseline systems.

Taco Baseline1 (Mono-Taco)： The basic structure is similar as Tacotron 2 [31], only provided data is used to bulid a mono speaker speech synthesis system.

Taco Baseline2 (Multi-Taco-1)： The basic structure is similar as Tacotron 1, which has a post-processing net after decoder layer. We replace the spectral magnitude to the vocoder parameters for WORLD[13] or LPCNet. And we add stop prediction and speaker embeddings.

Taco Baseline3 (Multi-Taco-2)： Compared with Taco Baseline 1, The post-processing net is discarded just like our proposed method. Only one speaker embedding is deployed to train together with Tacotron.

To compare the performance of vocoders, we set WORLD and LPCNet as two baselines.The following Table 2 is the abbreviations for several acoustic model + vocoder combinations.

Table 2 Abbreviations for Acoustic model + Vocoder

	WORLD	LPCNet
Mono-Taco	M-T-W	M-T-L
Multi-Taco-1	T-1-W	T-1-L
Multi-Taco-2	T-2-W	T-2-L
Multi-Taco-Cleaned	T-C-W	T-C-L
Multi-Taco-Duration	T-D-W	T-D-L

5.3 Internal evaluation

We conduct an internal evaluation to validate the effectiveness of the our submitted system.

Quality

We first compare different vocoders and different Tacotron. By observing the preference scores of subjective evaluations in the Table 3 it is worth noticing that the LPCNet can significantly improve the quality of synthetic speech. Besides, the Tacotron with post-processing net can not improve the quality of synthetic speech significantly. The dimensionality of acoustic features for LPCNet is 22, while the dimensionality of acoustic features for WORLD is 187. The discard of post-processing net and decreasing the dimensionality of predicting acoustic features can reduce the complexity of Tacotron model. Therefore, in the following sections we mainly concentrate on Tacotron 2 structure.

Table 3: Preference scores on quality of synthetic speech

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-1-W	14.67	16.98	68.35	T-2-L
T-2-W	11.67	13.73	74.60	T-1-L
T-1-W	40.45	25.82	33.73	T-2-W
T-1-L	36.42	29.94	33.64	T-2-L

Naturalness

By observing the preference scores of subjective evaluations in the Table 4 it is worth noticing that by using model to clean the database and adding the duration control can improve the performance of the submitted system.

Table 4: Preference scores on naturalness of synthetic speech

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-C-L	48.93	13.26	37.81	T-2-L
T-D-L	62.48	17.14	20.38	T-2-L
T-D-L	52.69	8.94	38.37	T-C-L

Similarity

By observing the preference scores of subjective evaluations in the Table 5 it is worth noticing that the deployment of LPCNet can improve the similarity of synthestic speech. By using adaptative training on the LPCNet, the neural vocoder can be more suitable for the target speaker.

Table 5: Preference scores on similarity of synthetic speech

System A	Scores A(%)	Scores Neutral(%)	Scores B(%)	System B
T-C-L	53.62	6.94	39.44	T-C-W
T-D-L	58.73	12.75	28.52	T-D-W

Evaluation results

24 participants attend the evaluation for **Single task 2019-EH1.** The Mean Opinion Scores, similarity (MOS) and intelligibility (word error rate (WER)) were calculated. The results are shown in Figure 3- Figure 6, where system A identifies natural speech and indentidy of our system is G. For all these three evaluation (naturalness, similarity and intelligibility) results, our system only ranks average level.

Discussion of the results

From the evaluation result, there is still a great gap between our system to the top one. There are many reasons leading this results. And the mainly one is that the LPCNet neural vocoder we used had a large gap with WaveNet on the quality. The big background noise produced by the LPCNet lead to a significant influence on the perception, which leads to a low scores on the MOS results. And the Pinyin error rate (with/without tone) achieved fair results, which demonstrated the effiectiveness of our propsed methods. These results reminder us there is still many works need to be done, especially on improving the quality of the synthestic speech.

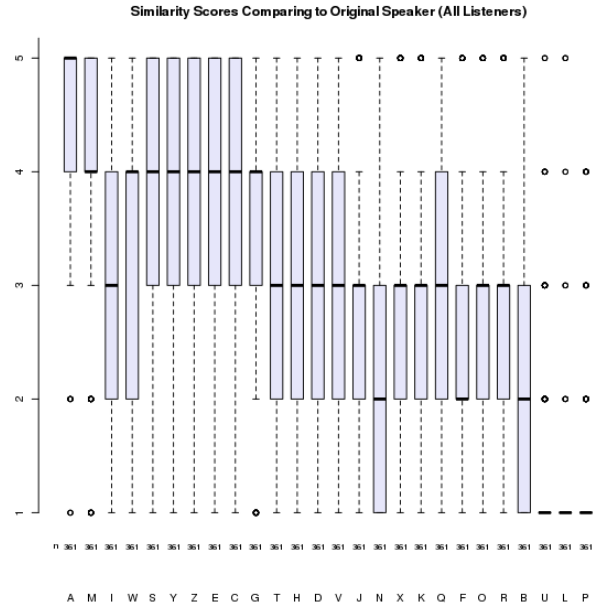

Figure 3: Boxplot of Similarity Scores.

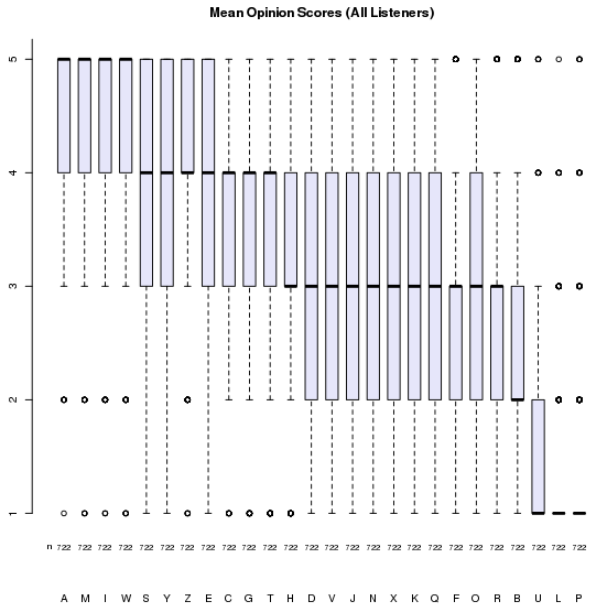

Figure 4: Boxplot of Mean Opinion Scores.

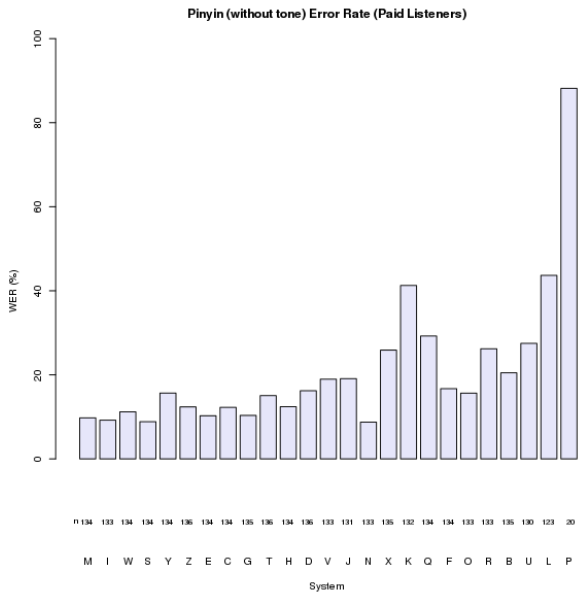

Figure 5: Pinyin (without tone) Error Rate

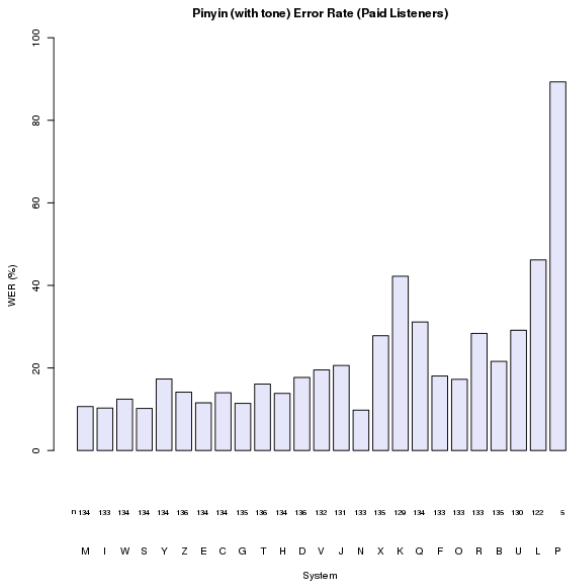

Figure 6: Pinyin (with tone) Error Rate

6. Conclusion

In this paper, the multi-speaker end-to-end speech synthesis system built for Blizzard Challenge 2019 by CASIA is introduced. There are several improvements from our previous Challenge system. The first one is the use of end-to-end system. The second one is the use of data pruing and speaker adaptation strategies. The final one is the duration control approaches. The internal evaluation results show that the effectiveness of these three techniques. Also, the evaluation results from the Blizzard Challenge committee shows that, the naturalness, similarity and intelligibility of our system are of average level. Many works need to be done, especially on improving the equality of the synthestic speech.

7. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), and Inria-CAS Joint Research Project (No.173211KYSB20170061).

8. References

[1] J. Shen, R. Pang, R. J. Weiss, et al, “Natural TTS synthesis by conditioningWaveNet on mel spectrogram predictions,” in Proceedings ICASSP-2018-IEEE International Conference on Acoustics, Speech, and Signal Processing,2018, pp. 4779-4783.

[2] W. Ping, K. Peng, A. Gibiansky, et al, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning”, in Proceedings ICLR 2018-International Conference on Learning Representations, 2018.

[3] N.Li, S. Liu, Y. Liu, et al, “Close to human quality TTS with transformer,” arXiv preprint arXiv:1809.08895, 2018.

[4] RJ Skerry-Ryan, E. Battenberg, Y. Xiao, et al, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,”in Proceedings ICML 2018-Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 4700–4709.

[5] Y. Wang, D. Stanton, Y. Zhang, et al, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in Proceedings ICML 2018-Proceedings of the 35th International Conference on Machine Learning, 2018.

[6] D. Stanton, Y. Wang, and RJ Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” arXiv preprint arXiv:1808.01410, 2018.

[7] Z. Wu, P. Swiqetojanski, C. Veaux, et al. “A study of speaker adaptation for DNN-based speech synthesis,” in Proceedings INTERSPEECH 2015 –Annual Conference of the International Speech Communication Association,2015.

[8] Y. Zhao, D. Saito, N. Minematsu, “Speaker Representations for Speaker Adaptation in Multiple Speakers’ BLSTM-RNN-Based Speech Synthesi,” in Proceedings INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association,2017. 2268-2272.

[9] M. Wan, G. Degottex, M. Gales, “Waveform-Based Speaker Representations for Speech Synthesis,” in Proceedings INTERSPEECH 2018 –Annual Conference of the International Speech Communication Association,2018.

[10] R. Fu, J. Tao, Z. Wen, et al, “Phoneme dependent speaker embedding and model factorization for multi-speaker speech synthesis and adaptation,” in Proceedings ICASSP-2019-IEEE International Conference on Acoustics, Speech, and Signal Processing,2019.

[11] S. Arik, G. Diamos, A. Gibiansky, et al. “Deep Voice 2: Multi-Speaker Neural Text-to-Speech,” in NIPS- Annual Conference on Neural Information Processing Systems, 2017

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveign´e, “Restructur-ing speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction,” Speech Communication, vol.27, no.3-4, pp.187–207, 1999.

[13] M. Morise, F. Yokomori, K. Ozawa, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” Ieice Transactions on Information & Systems, 2016, 99(7):1877-1884.

[14] A. V. D. Oord, S. Dieleman, H. Zen, et al. “WaveNet: A Generative Model for Raw Audio,” in Proceedings INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association,2017.

[15] S. Mehri, K. Kumar, I. Gulrajani, et al. “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model”. in Proceedings ICLR 2017-International Conference on Learning Representations, 2017.

[16] K. Chen, B. Chen, J. Lai, et al. “High-quality Voice Conversion Using Spectrogram-Based WaveNet Vocoder” in Proceedings INTERSPEECH 2018–Annual Conference

of the International Speech Communication Association,2018.2018-1528.
[17] L. Liu, Z. Ling, Y. Jiang, et al. “WaveNet Vocoder with Limited Training Data for Voice Conversion,” in Proceedings INTERSPEECH 2018–Annual Conference of the International Speech Communication Association,2018. 2018-1190.
[18] B. Sisman, M. Zhang, H. Li, “A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder,” in Proceedings INTERSPEECH 2018–Annual Conference of the International Speech Communication Association,2018.1978-1982, 2018-1131.
[19] C. Zhou, M. Horgan, V. Kumar, C. Vasco, et al, “Voice Conversion with Conditional SampleRNN,” in Proceedings INTERSPEECH 2018–Annual Conference of the International Speech Communication Association,2018.1973-1977, 2018-1121.
[20] J. M. Valin , J. Skoglund, “LPCNet: Improving Neural Speech Synthesis Through Linear Prediction”. in Proceedings ICASSP 2019- IEEE International Conference on Acoustics, Speech, and Signal Processing,2019.
[21] Y. Wang, R. Skerry-Ryan, D. Stanton, et al, “Tacotron: Towards End-to-End Speech Synthesis,” in Proceedings INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association,2017,4006-4010.
[22] X. Glorot, Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256, 2010.
[23] B.C.J. Moore, An introduction to the psychology of hearing, Brill, fifth edition, 2012.
[24] J. M. Valin, “A hybrid DSP/deep learning approach to realtime full-band speech enhancement,” in Proceedings MMSP 2018-Multimedia Signal Processing Workshop, 2018.
[25] D. Kingma, J. Ba , “Adam: A method for stochastic optimization,” in Proceedings ICLR 2015-International Conference on Learning Representations, 2015.
[26] S. J. Reddi, S. Kale, and S. Kumar, “On the convergence of adam and beyon, ” in Proceedings ICLR 2018-International Conference on Learning Representations, 2018.