The Mobvoi Text-To-Speech System for Blizzard Challenge 2019

Bing Liu[1*], *Yunlin Chen*[2], *Hao Yin*[2], *Yongqiang Li*[2], *Xin Lei*[2], *Lei Xie*[1]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Mobvoi AI Lab, Beijing, China

bingliu@nwpu-aslp.org, {yunlinchen, hao.yin, yongqiangli, mikelei}@mobvoi.com, lxie@nwpu.edu.cn

Abstract

This paper presents the Mobvoi team's text-to-speech system for Blizzard Challenge 2019 (BC2019). The training data provided by this challenge is about 8 hours of speech from one native Mandarin Chinese speaker in talk shows. We built a speech synthesis system based on end-to-end deep learning technology. The system consists of a hybrid front-end that processes both Chinese and English texts, a sequence-to-sequence model that converts the phoneme sequence into a mel spectrogram sequence, and a neural vocoder that generates audio from the mel spectrogram.

Index Terms: text-to-speech, Blizzard Challenge 2019, end-to-end, hybrid front-end, neural vocoder

1. Introduction

To better understand and compare different techniques in building corpus-based speech synthesizers on the same data set, the Blizzard Challenge has been devised. The task is to build a synthetic voice from the released speech data set. A prescribed set of test sentences are synthesized for listening tests [1].

The Blizzard Challenge has been held once a year since 2005 [2]. This year, about 8 hours of speech data from internet talk shows by a well-known Chinese anchor are released. All data are from the same speaker, and the speech is stylistic and expressive.

At present, commonly used speech synthesis technologies can be grouped into the following three categories:

1. Statistical parametric speech synthesis (SPSS).
   This kind of methods characterizes the speech signal using acoustic parameters, and a statistic model is used to build the mapping relationship between the text input and the acoustic output to synthesize arbitrary texts. Depending on the model used, systems in this category can be divided into HMM based [3] and neural networks (DNN [4], RNN or LSTM [5, 6]) based.
2. Unit selection and concatenation.
   For the sentence to be synthesized, such kind of methods first select a set of suitable speech segments from a pre-recorded large speech database and then splice the selected speech segments in the time domain in order to output the synthesized speech. Speech synthesis based on unit selection relies heavily on the size of the speech database and the quality of the unit selection algorithm [7, 8, 9, 10].
3. End-to-end deep learning.
   End-to-end speech synthesis systems mainly includes an attention-based sequence-to-sequence model [11, 12, 13] which maps the text representation to an acoustic representation and a neural vocoder [14, 15, 16] that transforms the acoustic representation into a waveform. End-to-end systems simplify the traditional SPSS model framework, and exceeds the traditional SPSS method and unit selection method in both naturalness and similarity metrics.

Given that the end-to-end approach is the best performing speech synthesis technology, we choose to build our system using Tacotron2 [12] and WaveNet [14]. If Chinese characters are used directly as input, it is difficult to learn the pronunciations of Chinese characters through an end-to-end model due to limited data. Therefore, we use a text analysis module to convert the text into a phoneme sequence, which reduces the difficulty of the model training given the limited data.

As shown in Figure 1, our system consists of three parts. First, we use a hybrid front-end to convert the text into a sequence of phonemes, tones and prosodic boundaries. Second, the sequence is converted to a mel spectrogram sequence via Tacotron2. Finally, high-quality audio is generated through a WaveNet vocoder.

The organization of this paper is as follows. In Section 2, we detail each module in the Mobvoi system. In Section 3, the evaluation results are shown and discussed. Finally, we conclude our work in Section 4.

2. Mobvoi TTS System

2.1. Data processing

The data provided by the organizer contains 480 audio files in MP3 format with a sampling rate of 48 kHz. The audio files are approximately 1 minute on average, with approximately 8 hours in total. We first convert the audio to WAV format and downsample to 16 kHz. In addition, in order to facilitate the training, we cut the long audio recording into short audio segments of no more than 10s per segment. After segmentation, there are 4187 sentences in our training set.

Because the audio is hand-cut, the length of the silence before and after the segmented sentence varies a lot. This problem affects the attention module in Tacotron2 and the WaveNet training. To solve this problem, we use the *sox* tool to split the front and end silence, and re-splice about 200ms at the beginning and end of each audio. We also found that the volume distribution of the audio in the data set is very uneven. So we use an energy-based normalization method to normalize the audio in the data set.

We checked the text and corresponding audio provided by the organizer and found that some of the text and audio were not exactly the same, so we relabeled the text using the data provided by the organizer. This will ensure that the subsequent

* Work done during internship at Mobvoi.

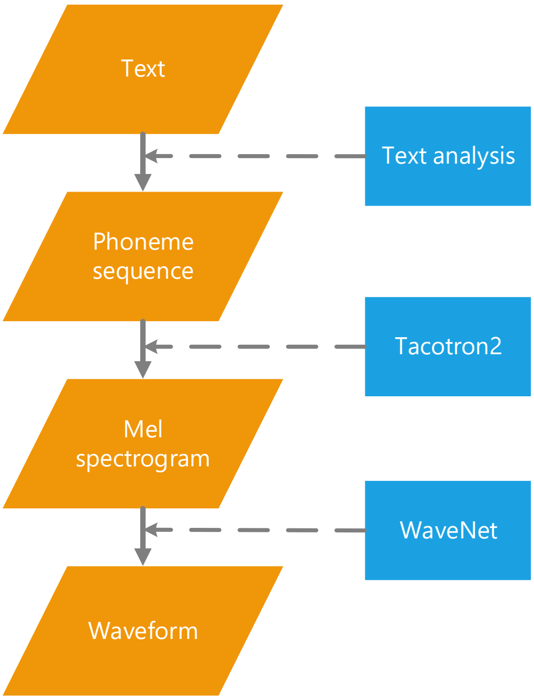

Figure 1: *The architecture of Mobvoi TTS system.*

training can be carried out normally. Otherwise, the attention model in the Tacotron2 can not be trained properly.

2.2. Front-end

The front-end we use is a hybrid front-end that includes both Chinese text analysis and English text analysis. The two front-ends share the segmentation and prosody prediction module. The other modules such as part-of-speech prediction, g2p, etc. are language-specific. The input text first passes through the text normalization module and then passes through a word segmentation module trained by a conditional random fields (CRF) model. Finally, according to the Chinese and English categories of the word after the word segmentation, the front-end of the specific language is used to process and obtain the corresponding phoneme sequence. As suggested in the Tacotron paper [11], the end-to-end model does not require complex linguistic features of the traditional model, so our front-end only needs to predict the phoneme sequence corresponding to the text.

For Chinese, we also need to consider how the tone of the final is combined with the phoneme. There are three ways of combination:

1. Tone and final binding
   Since the finals are closely related to the tone, it is most intuitive to add tonal information by means of tone and finals. It is reasonable theoretically. However, such an approach would make the embedding space of the final and the sound adjustment body too sparse, and the input with the same final and different tone can not share the commonality of pronunciation. Experiments show that the phenomenon of tonal modification is prone to occur when the amount of data is not too large.
2. Tone alone as input
   To reduce the number of categories that embedding representation, we can separate the tone from the final and use it as a way to enter the symbol alone.
3. The phoneme and tone are expressed separately
   In order to distinguish the phoneme and the embedding space of the tone, we can extract the embedding representation of the space and the tone separately, and then combine them by adding or concatenation.

Based on past experimental experience and for the sake of simplicity, we chose the second method. We hope that the back-end model can learn to distinguish between tones and phonemes in the embedding space. Experiments show that the second input mode has a fewer tonal modification, indicating that the back-end model can learn the effect of tones.

Finally, in order to speed up convergence and control the prosody by input, we also place the third-level prosody boundary as input after the corresponding phoneme.

2.3. Back-end

The traditional statistical parametric synthesis method divides the back end into a duration module and an acoustic module. This splitting leads to the cascade transmission error of the model when predicting the actual speech, resulting in over-average of duration and over-smoothing of acoustic features. In order to simplify the overall framework of speech synthesis, the sequence-to-sequence technology based on attention mechanism is gradually beginning to be used in speech synthesis tasks.

The sequence-to-sequence model based on attention mechanism usually consists of three parts. The encoder is mainly used to extract robust sequential representations of text, and the decoder is used to map text representations into acoustic features recursively. The attention mechanism is used as the bridge between the codecs and allows the decoder to selectively focus attention on certain moments of the encoder outputs when generating the acoustic features. On the one hand, the sequence-to-sequence model simplifies the text analysis module(use the original text as input). On the other hand, the attention mechanism is used to learn the alignment mapping between unequal length sequence, which avoids the complexity of manual time-length annotation or forced alignment.

We use the Tacotron2 model as our back-end, accepting the phoneme sequence generated by the front-end to generate the corresponding acoustic features. The acoustic features are a sequence of 80-dim mel-scale filterbank frames, computed from 50ms windows shifted by 12.5ms. The model structure and parameters are consistent with the Tacotron2 paper [12], except that the reduction factor is set to 2 (we tried to train a model of reduction factor = 1, but the alignment result is poor).

We used 20-hour male data to train a basic model (trained to 200k steps) and then fine-tuned the model with the BC2019 data. Compared with directly using the BC2019 data to train the Tacotron2 model, this method can quickly train better alignment, and the quality of audio inferred from this training method is better than direct training.

2.4. Vocoder

Usually, because the speech waveform has a very fast change frequency in a short time(16,000 samples per second or more), researchers rarely model the waveform directly and instead model relatively stable acoustic parameters extracted from the audio. However, the upper bound of the sound quality implied by the traditional vocoder based on the source-filter model severely limits the sound quality of the synthesized speech, making the synthesized speech lack of realism compared with natural speech. The neural-vocoder technique greatly improves the quality of synthesized speech by directly modeling the speech waveform with deep neural network and bypassing the traditional speech vocoder. Google's DeepMind Lab presents a model for directly modeling raw waveforms [14], which uses a deep neural network based on dilated causal convolution to simulate real speech, and the resulting speech sounds better than the baseline(speech generated using traditional vocoder). The optimal speech synthesis system is more natural and almost identical to the human voice.

The network structure we use is basically the same as in WaveNet [14], except for the following two aspects:

1. There are 20 dilated convolution layers, grouped into 2 dilation cycles, i.e., the dilation rate of layer k (k = 0...19) is $2^{k \pmod{10}}$.
2. Instead of predicting discretized buckets with a softmax layer, we follow Parallel WaveNet [15] and use a 10-component mixture of logistic distributions (MoL) to generate 16-bit samples at 16 kHz.

If the WaveNet model is trained using the acoustic parameters of ground truth, when the acoustic parameters predicted by Tacotron2 are used as inputs in the inference, the generated audio has a noticeable noise. In order to eliminate the mismatch of acoustic parameters, we first train the WaveNet model using ground truth features and then fine-tune it on the ground truth-aligned(GTA) predictions of the Tacotron2 model. Experiments show that this method can significantly reduce the noise in inference and further improve the sound quality.

In order to speed up the inference, we adopted the fast generation algorithm of the paper [17]. The algorithm pre-stores some calculated intermediate variables in the form of a cache to provide the sample calculations for a future time. By using this method, the speed of synthesizing speech can be increased by about 60 times(in our experiments). Although Parallel WaveNet [15] can continue to increase the speed of the synthesis by 1000 times, the sound quality has some loss. This challenge only requires offline audio synthesis, so we didn't choose to use Parallel WaveNet.

3. Evaluation Results

There are 26 systems in total, including 24 from participating teams, one benchmark, and one natural speech. System A is a natural speech recorded by the original speaker. System B is the merlin benchmark system. System C to Z are the 24 participating teams, and system E is ours.

Table 1: *Task 2019-EH1*

Sections	Detailed Description
section 1	Naturalness MOS (Mean opinion scores)
section 2	Similarity MOS (Mean opinion scores)
section 3	PER (Pinyin Error Rate)
section 4	PTER (Pinyin+Tone Error Rate)

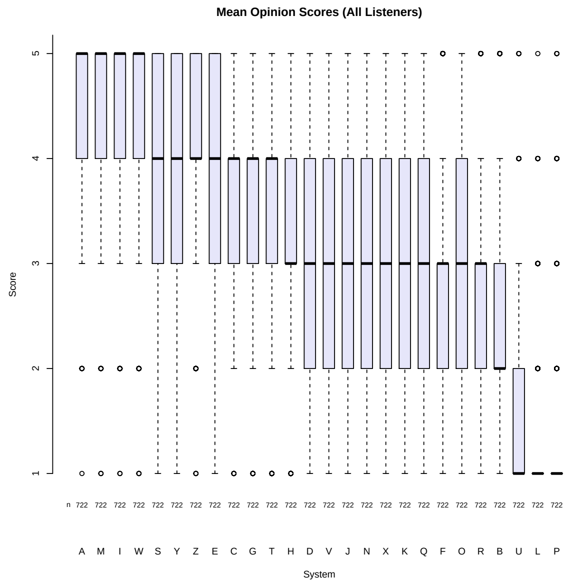

Figure 2: *Mean opinion score, naturalness evaluation, from all listeners.*

The evaluation comprised four sections showed in Table 1. The MOS and Similarity results are based on all the listeners' responses, including paid listeners at Edinburgh, volunteers, and experts. The PER and PTER are mainly based on paid listeners' responses, which produce more reliable results. Finally, our system has achieved good results in all the criteria for the Challenge. Details are as follows.

3.1. Naturalness test

Figure 2 shows the results of the naturalness MOS given by all listeners for all the systems. In this test, listeners were asked to listen to samples and assign scores either on a scale of 1 [Completely Unnatural] to 5 [Completely Natural]. Our system has an average score of 3.9. We believe that if we train a Tacotron2 model with a reduction factor of 1 and increase the audio sample rate from 16 kHz to 48 kHz, our system can achieve a higher naturalness score.

3.2. Similarity test

Figure 3 shows the results of the similarity MOS given by all listeners for all the systems. In this test, each listener was asked to decide how similar the voice in one new sample sounded to the voice in two reference samples either on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person]. Our system has an average score of 3.8 and ranks fourth. This is mainly due to the powerful modeling capabilities of the neural vocoder for the waveform.

3.3. Intelligibility test

Figure 4 and Figure 5 show PER (Pinyin Error Rate) and PTER (Pinyin+Tone Error Rate), respectively. In this test, the listeners were asked to listen to one audio at a time and write down what they heard and to listen to as little audio as possible. Because the tone of Chinese is very important for semantic expression, not only the error rate of pinyin but also the error rate of pinyin

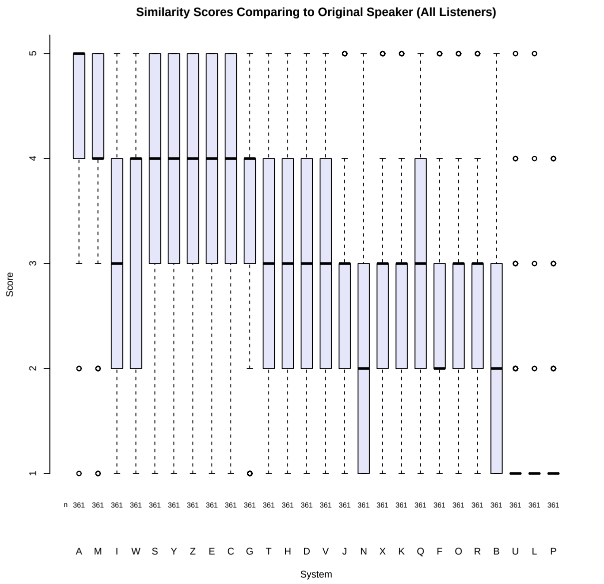

Figure 3: *Mean opinion score, similarity evaluation, from all listeners.*

with tone is counted.

We found that our system does not perform well on long sentence and multiple repeated words. This is mainly due to the fact that the attention module is not robust enough to generate the correct attention to these sentences. We have noticed that there have been many recent papers that improve the robustness of the attention module, such as Monotonic attention [18], Stepwise Monotonic Attention [19], etc. In the future, we will try these methods to improve the intelligibility of our system on Semantically Unpredictable Sentences (SUS).

4. Conclusions

This paper presents the details of our submitted system and summarizes the results in Blizzard Challenge 2019. We built a speech synthesis system based on end-to-end deep learning technology. The system consists of a hybrid front-end that can process both Chinese and English texts, a sequence-to-sequence model that converts the phoneme sequence into a mel spectrogram sequence, and a neural vocoder that generates audio from the mel spectrogram. Our system has achieved good results in all the criteria for the Challenge.

In the future, we will continue to study speaker adaptive techniques based on end-to-end techniques to produce a new voice with a small amount of data. At the same time, we will study different attention mechanisms to improve the robustness of the end-to-end model.

5. Acknowledgements

We would like to thank Mobvoi's TTS labeling team in Wuhan for supporting data processing and labeling. Thanks also to our colleagues Ran Zhang, Zheng Zhang, and Haibin Cao et al. for helpful discussions and advice.

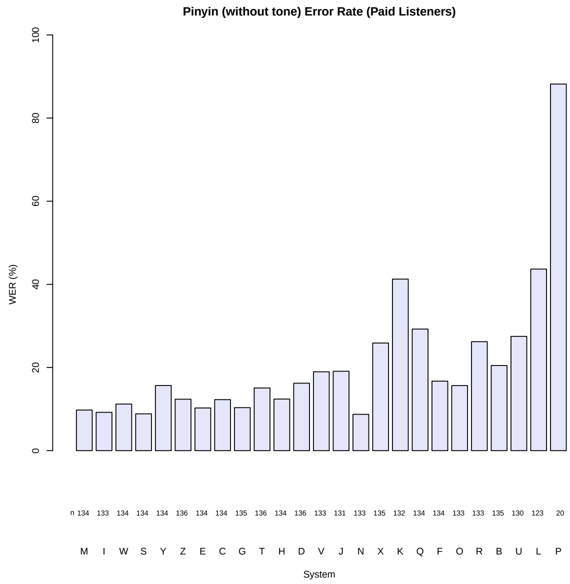

Figure 4: *PER (Pinyin Error Rate) of each submitted system.*

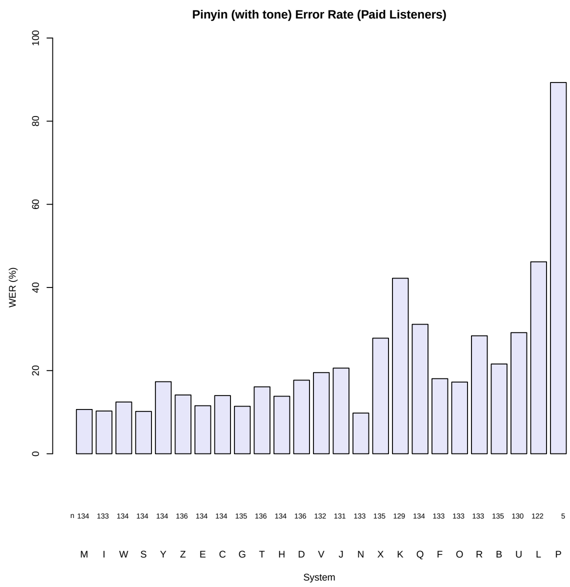

Figure 5: *PTER (Pinyin+Tone Error Rate) of each submitted system.*

6. References

[1] "The blizzard challenge website." [Online]. Available: http://proceedings.mlr.press/v70/raffel17a.html

[2] A. W. Black and K. Tokuda, "The blizzard challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," *Proc Interspeech 2005*, 2005.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009. [Online]. Available: https://doi.org/10.1016/j.specom.2009.04.004

[4] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7962–7966. [Online]. Available: https://doi.org/10.1109/ICASSP.2013.6639215

[5] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1964–1968. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2014/i14_1964.html

[6] 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 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4470–4474. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178816

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990. [Online]. Available: https://doi.org/10.1016/0167-6393(90)90021-Z

[8] A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, and S. Raptis, "A unit selection text-to-speech synthesis system optimized for use with screen readers," *IEEE Trans. Consumer Electronics*, vol. 56, no. 3, pp. 1890–1897, 2010. [Online]. Available: https://doi.org/10.1109/TCE.2010.5606343

[9] Z. Yan, Y. Qian, and F. K. Soong, "Rich-context unit selection (RUS) approach to high quality TTS," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, 2010, pp. 4798–4801. [Online]. Available: https://doi.org/10.1109/ICASSP.2010.5495150

[10] Z. Ling and R. Wang, "Hmm-based hierarchical unit selection combining kullback-leibler divergence with likelihood criterion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, 2007, pp. 1245–1248. [Online]. Available: https://doi.org/10.1109/ICASSP.2007.367302

[11] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 4006–4010. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1452.html

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[13] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[14] 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," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[15] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," *CoRR*, vol. abs/1711.10433, 2017. [Online]. Available: http://arxiv.org/abs/1711.10433

[16] J. Valin and J. Skoglund, "LPCNET: improving neural speech synthesis through linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2019, pp. 5891–5895. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8682804

[17] T. L. Paine, P. Khorrami, S. Chang, Y. Zhang, P. Ramachandran, M. A. Hasegawa-Johnson, and T. S. Huang, "Fast wavenet generation algorithm," *CoRR*, vol. abs/1611.09482, 2016. [Online]. Available: http://arxiv.org/abs/1611.09482

[18] C. Raffel, M. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017, pp. 2837–2846. [Online]. Available: http://proceedings.mlr.press/v70/raffel17a.html

[19] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," *CoRR*, vol. abs/1906.00672, 2019. [Online]. Available: http://arxiv.org/abs/1906.00672