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

The Blizzard Challenge aims to better understand and compare different techniques in building corpus-based speech synthesizers on the same data set. Specifically, the task is to build a synthetic voice from the released speech data set and a prescribed set of test sentences is 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 very 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 statistical 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 network (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 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 involve 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 exceed the SPSS pipeline and unit selection method in both naturalness and quality.

Given that the end-to-end approach achieves superior performance, we choose to build our system using Tacotron2 [12] and WaveNet [14], the most popular end-to-end framework. 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 input 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 input 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.

2. Mobvoi TTS System

2.1. Data processing

The data set 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 model training, we cut the long audio recording into short audio segments with 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 as well. 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 find 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 check the text and corresponding audio provided by the organizer and find that some of the text and audio are not exactly the same. Hence we relabel the text using the data provided by the organizer. This can ensure that the subsequent model training can be carried out normally. Otherwise, the attention model in the Tacotron2 can not be trained properly, according to our experiments.

[*] Work done during internship at Mobvoi.

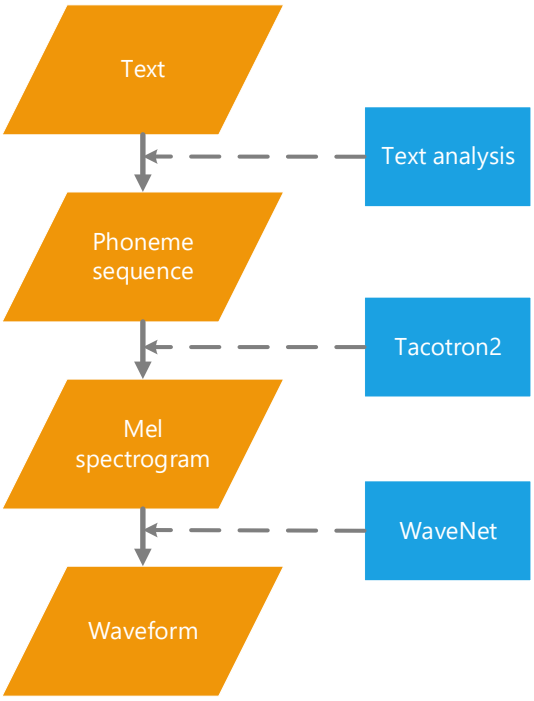

Figure 1: *The architecture of the Mobvoi TTS system.*

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 a text normalization module and then passes through a word segmentation module and prosody prediction 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 [11], the end-to-end Tacotron model does not require complex linguistic features as in 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 finals is combined with the phoneme. There are three ways of combination that we can consider.

1. Tone and finals binding
   Since the finals are closely related to the tone, it is intuitive to add tonal information by means of the finals with tone (e.g. "a1", "a2" and so on). The number of expanded phonemes is the number of initials plus the product of the number of finals and the number of tones. Hence such an approach would make the embedding space of phonemes with tone too sparse, and the input with the same finals and different tones can not share the similarity of pronunciation. Experiments show that there are many inaccurate tonal pronunciations when the amount of data is not large enough.
2. Tone alone as input
   To reduce the number of categories for embedding representation, we can separate the tone from the finals with tone and use it as a phoneme (e.g. "a 1", "a 2" and so on). The number of expanded phonemes is the sum of the number of initials, the number of finals and the number of tones. This method maps phonemes and tones into the same space and does not require modification of the embedding network in the back-end.
3. The finals and tone are expressed separately
   In order to distinguish the embedding space of phonemes and tones, we can extract the embedding representation of phonemes and tones separately, and then combine them by adding or concatenation.

Based on our past experimental experience and for the sake of simplicity, we finally choose 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 fewer inaccurate tonal pronunciation, indicating that the back-end model can learn the tones effectively.

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

2.3. Back-end

The traditional SPSS 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 an 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 become mainstream in speech synthesis.

The sequence-to-sequence model based on attention mechanism usually consists of three parts. The encoder is mainly used to extract sequential representations of text, and the decoder is used to map the text representations into acoustic features recursively. The attention mechanism is used as a bridge that allows the decoder to selectively focus 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 learns the alignment mapping between two unequal length sequences, which avoids manual duration annotation or forced alignment in traditional methods.

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

We use an extra set of 20-hour data from another Chinese male speaker to first train a basic model and then fine-tuned the model with the BC2019 data. Compared with directly training using the BC2019 data, this method can quickly achieve better attention alignment, and the quality of audio inferred from this training method is better than direct training.

2.4. Vocoder

Usually, because speech waveform has a very fast change in frequency in a short time (16,000 samples per second or more), researchers rarely model the waveform directly. Instead, they model relatively stable acoustic parameters extracted from the audio, i.e, the traditional vocoding method. However, the upper bound in the time-frequency domain 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 human speech. The recent neural-vocoder technique greatly improves the quality of synthesized speech by directly modeling the speech waveform with a generative deep neural network and bypassing the traditional speech vocoder. Google DeepMind has recently proposed WaveNet, a generative model that directly model raw waveforms [14]. WaveNet uses a deep neural network based on dilated causal convolution to simulate real speech, resulting in high fidelity speech that reaches human parity.

The network structure we use is basically the same as in the original 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 $2^{k(\text{mod } 10)}$.
2. Instead of predicting discretized buckets with a softmax layer, we follow the Parallel WaveNet [15] and use a 10-component mixture of logistic distributions (MoL) to generate 16-bit samples at 16 kHz.

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

In order to speed up the inference, we adopt the fast generation algorithm in [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 (according to our experiments). Although Parallel WaveNet [15] can continue to increase the inference speed by 1000 times, the sound quality has a noticeable loss. As this challenge only requires offline audio synthesized, we do not choose to use Parallel WaveNet this time.

3. Evaluation Results

There are 26 systems in the official evaluation 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 our system.

The evaluation includes four sections, as shown 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. Details are as follows.

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)

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 a higher level, e.g. 48 kHz, our system may achieve a better naturalness score.

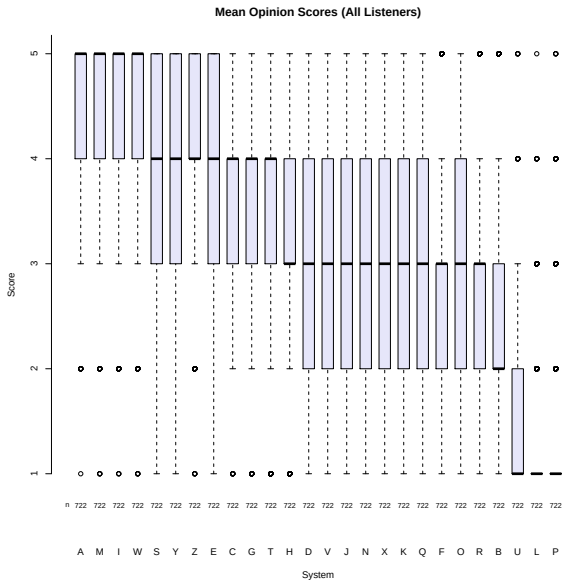

Figure 2: *Naturalness evaluation in terms of mean opinion score*

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 is 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 with high speaker similarity. We believe this level of speaker similarity is mainly due to the powerful modeling capabilities of the WaveNet neural vocoder.

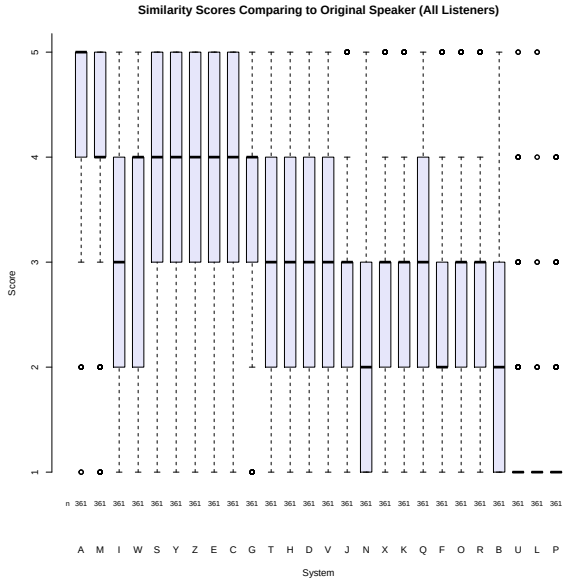

Figure 3: *Similarity evaluation in terms of mean opinion score*

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 are 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 the Chinese syllable is very important for semantic expression, both the error rate of Pinyin and the error rate of Pinyin with tone are counted.

We find that our system does not perform well in long sentences 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.

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

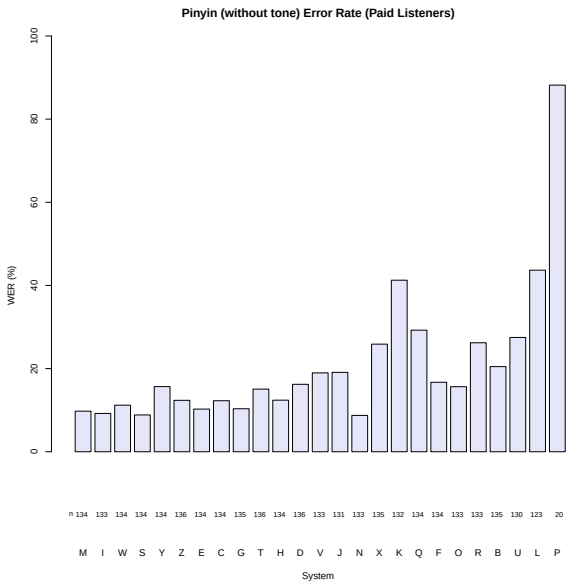

Figure 4: *PER (Pinyin Error Rate) of submitted systems.*

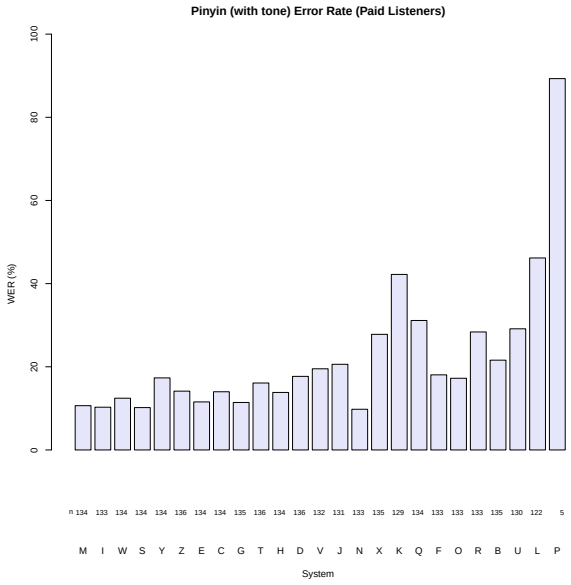

Figure 5: *PTER (Pinyin+Tone Error Rate) of submitted systems.*

colleagues Ran Zhang, Zheng Zhang, and Haibin Cao for helpful discussions and advice.

6. References

[1] "The blizzard challenge website." [Online]. Available: https://www.synsig.org/index.php/Blizzard$_Challenge$

[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