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

The HITSZ TTS system for Blizzard challenge 2020

Huhao Fu, Yiben Zhang, Kailong Liu, Chao Liu

Harbin Institute of Technology, Shenzhen

huhaofu@gmail.com

Abstract

In this paper, we present the techniques that were used in HITSZ-TTS[1] entry in Blizzard Challenge 2020. The corpus released to the participants this year is about 10-hours speech recordings from a Chinese male speaker with mixed Mandarin and English speech. Based on the above situation, we build an end to end speech synthesis system for this task. It is divided into the following parts: (1) the front-end module to analyze the pronunciation and prosody of text; (2) The phoneme-converted tool; (3) The forward-attention based sequence-to-sequence acoustic model with jointly learning with prosody labels to predict 80-dimensional Mel-spectrogram; (4) The Parallel WaveGAN based neural vocoder to reconstruct waveforms.

This is the first time for us to join the Blizzard Challenge, and the identifier for our system is G. The evaluation results of subjective listening tests show that the proposed system achieves unsatisfactory performance. The problems in the system are also discussed in this paper.

Index Terms: speech synthesis, end-to-end, Evotron, Parallel WaveGAN

1. Introduction

To promote speech synthesis technology, Blizzard Challenge (BC) has been held every year to evaluate Text-to-Speech (TTS) systems since 2005. In Blizzard Challenge 2020, The MH1 task is to build a Chinese TTS system based on a 10-hours speech dataset from a single speaker. A testing set of sentences is also provided to evaluate the model's performance. There are four evaluation items for the task, namely; similarity, naturalness, error rate, and overall feeling of the paragraph; the overall feeling of the paragraph is divided into 6 sub-items: pleasure, pause rhythm, accent, tone, emotion, and hearing resistance.

Recent work on neural text-to-speech (TTS) can be divided into two parts. In the first part, statistical parameter speech synthesis methods [1,2,3] with deep neural network architectures. Based on the DNN framework, more novel architectures or variants have been proposed to improve the performance of synthesized speech. Note that the traditional frameworks require an additional module to align linguistic and acoustic representations, and the wrong align errors may propagate to the latter synthesis model. In the other part, the attention-based sequence-to-sequence [4] e2e TTS model (like Tacotron) [5-8] that they can be trained on <text, audio> pairs have been proposed to eliminate the need for complex sub-systems that need to be developed and trained separately and thus simplify the pipeline of traditional systems. It predicts the spectrum directly from phonemes and combining with the neural vocoder [9,10,11] it achieves superior performance over the conventional structures.

Besides audio generations, especially for Mandarin TTS, text analysis also plays an important role in the text to speech system. The G2P and prosody break boundary directly influences the intelligibility and naturalness of the synthesized audio [12]. To improve the performance of our system, we use a text analyzer to extract the linguistic feature and prosody information. Besides, we adopt a forward location sensitive attention mechanism [13] to stabilize the generating process for long sentences and decrease the synthesis wrong such as repeat, skip.

This paper is structured as follows: section2 describes the data process and text analyzer, Section 3 proposed HITSZ TTS system. Section 4 describes the evaluation results and system performance. The conclusion is given in Section 5.

2. The task in Blizzard 2020

The BC 2020-MH1 is Mandarin synthesis. The data corpus provided by the organizer consists of Mandarin male news, which is 5566 audio files at a 48kHz sampling rate with a total duration of 10 hours and the corresponding texts. All files in this database are in ".wav" format. In this challenge, we have two aims, the first aim of our proposed system is to generate a new voice as similar as possible to the voice of the target speaker and the second aim is to ensure the synthesized audio more natural. More details of our system will be introduced in the following sections.

3. System building

As shown in Figure 1, our speech synthesis system takes an end to end architecture. At the training phase, we use an open-source front-end tool pypinyin[2] to predict phoneme, tone and other linguistic features. Different from the traditional TTS methods, our systems replace the duration model with the attention mechanism in the acoustic model. And then, syntactic features are generated by the Stanford parser tool[14]. Meanwhile, we use the variant transformer block structure as an encoder in the seq2seq model with some useful training tricks. Furthermore, to reduce the unnatural breaking times in the synthesized audio, we introduced the break prosody embedding as additional inputs here [15]. Finally, To improve the inference speed while ensuring the synthesized audio quality, a GAN-based neural vocoder [11] which was conditioned on the Mel-spectrograms was trained on the target male dataset.

At the synthesis phase, the phoneme inputs and linguistic features are predicted by the open-source tools and the

[1] HITSZ-TTS: The text-to-speech system of Harbin Institute of Technology, Shenzhen.

[2] https://github.com/mozillazg/python-pinyin

10.21437/VCC_BC.2020-11

prosodic break label is predicted by the prosodic break predictor. Then we feed those features into the acoustic model to predict the Mel-spectrogram. Finally, the Parallel WaveGAN is used to generate waveform samples conditioning on the predicted Mel-spectrograms.

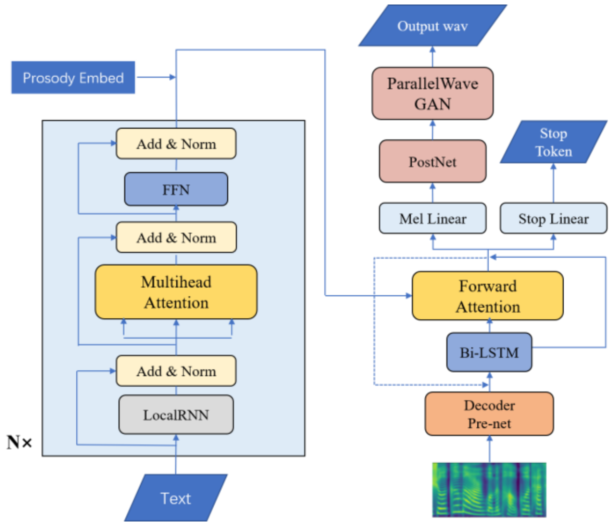

Figure 1: The architecture of our system.

3.1. Data Preparation

3.1.1. *Linguistic features*

To improve the model performance of language representations, we extract several linguistic features as an additional input of the model. Firstly, we adopted a Jieba text-analyzer tool to get the linguistic feature, such as segment token (Seg), part-of-speech (POS). we also extract the syntactic feature (lowest common ancestor, syntactic distance) [16] by Stanford parser tool. The above features are taken as part of inputs to the acoustic model.

3.1.2. *Acoustic features*

All audios used in the training phase were firstly down-sampled to 24 kHz, The leading and trailing silence are trimmed to a fixed length. We use the Mel-frequency spectrogram as the acoustic representation of audio signals. It's suitable for the acoustic model and neural vocoder to generate high-quality speech. The targets Mel-spectrograms features are computed through a short-time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop size, and a Hann window function, passed through 80-channel Mel-scale filterbanks spanning 0 Hz to 12 kHz followed by log dynamic range compression.

3.2. Phrase break prediction model

To get more natural speech, we design a phrase break prediction model to predict the phrase break label. In our system, the typical three-layer prosody is employed as the label which consists of prosody word (PW), prosodic phrase (PPH), and intonational phrase (IPH) [17]. The Bert-BiLSTM-CRF is adopted to predict the phrase break label. In this model, each word in an input sentence is mapped to a sequence of word embeddings by a trained BERT model[18], which is passed through BiLSTM. Finally, we use a CRF layer[19][20] as the decoder layer to produce a prosodic break label. Then the CRF layer can learn the decoder process by maximum likelihood estimation. It is shown in Figure 2.

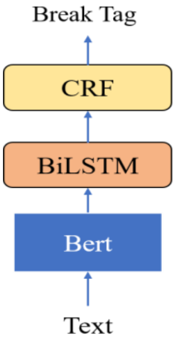

Figure 2: The architecture of our Phase predictor.

3.3. Seq2seq acoustic model

For achieving expressive speech synthesis, generating natural prosody is meaningful which is much hard for SPSS. In our system, we adopt the attention-based seq2seq framework as an acoustic model to predict Mel-spectrogram from the text representation. Similar to common seq2seq TTS models, such as Tacotron, our system contains a text encoder, a forward location sensitive attention[13] module, and an auto-regressive decoder. Figure 1 shows the building blocks of our system.

For the encoder part, we adopt the transformer network as described in Transformer TTS [7], which is also proved to improve the performance of prosodic phrasing. This is inspired by Transformer network [21], where self-attention plays a vital role in modeling global dependency. And we also use the Local-RNN [22] module to enhance the local relations of Transformer based TTS model. Considering the attention module is the key factor that directly affects the stability of the end-to-end systems, so we choose the forward attention in our system as it shows robustness in generating long sentences than location-sensitive attention.

As for the decoder part, we followed the architecture of Tacotron2 to predict Mel spectrogram. The decoder is an autoregressive recurrent neural network that contains 2 dense layers of 256 hidden ReLU units as pre-net, which is essential for learning attention. And the sub-network of 2 uni-directional GRU layers with residual connections is adopted to produce the attention query as each decoder step. Then the GRU output is concatenated with the attention context to predict Mel spectrogram. At last, a CNN based PostNet is adopted to improve the quality of the generated Mel-spectrogram.

3.3.1. *Modeling local structures with Local-RNN*

As proposed in [22], We replace the positional encoder embedding with the Local-RNN to improve the ability in local dependency modeling. Like a sliding window, the Local-RNN is reorganizing the original long sequence into many short sequences that only contain local information and are processed by a shared RNN independently and identically. For example, we define the window size is L, give a short sentence $(X_{t-L-1},\cdots,X_t)$ as the input of Local-RNN, and then the last hidden state is used as the latent representation for the local short sequence:

$$h_t = LocalRNN(X_{t-L-1}, \cdots, X_t) \quad (1)$$

Finally, we can calculate all corresponding local latent representations for all positions in the input sequence by Local-RNN.

3.3.2. *The loss for acoustic modeling*

In addition to base Tacotron loss, we use guided attention loss for faster attention convergence. we also use the Structural Similarity Index (SSIM) loss [23] to increase the stability of the training and make Mel-spectrograms less blurry. High-quality vocoder can make the audio quality difference caused by spectral blurring more obvious. To make the Mel-spectrograms texture more clear and more coherent, we introduce the first-order difference loss between the true Mel-spectrogram and predict Mel-spectrogram. We define the true el-spectrogram is g, and the predict Mel-spectrogram is p, we define g_{diff} is the difference between t time step and t-1 time step on g, and p_{diff} is also calculated in the same way. And then the formula of first-order difference loss is shown as follows:

$$L_{diff} = MSE(g_{diff} - p_{diff}) \quad (2)$$

The above losses are very useful for our system in this challenge.

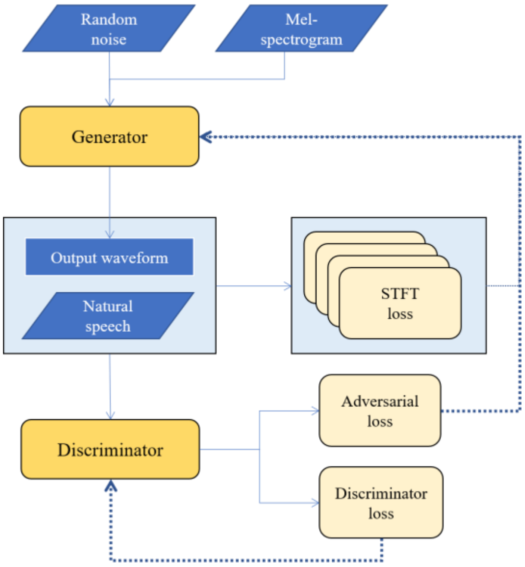

Figure 3: The architecture of Parallel WaveGAN.

3.4. Parallel WaveGAN Vocoder

In consideration of the real-time ratio, we choose Parallel WaveGAN [11] as the vocoder model. As shown in Figure 3. a WaveNet-based model with non-causal convolutions is used as a generator, which is a non-autoregressive model that is conditioned on Mel-spectrogram to generate a waveform in parallel. The generator can generate a 24kHz speech waveform faster than in realtime. The multi-resolution short-time Fourier transform (STFT) auxiliary loss (L_{aux}) is represented as follows:

$$L_{aux}(G) = \frac{1}{M}\sum_{m=1}^{M} L_s^{(m)}(G) \quad (3)$$

M denotes the number of STFT losses. The final loss function for the generator is defined as follows:

$$L_G(G, D) = L_{aux}(G) + L_{adv}\, \lambda_{adv}(G, D) \quad (4)$$

Where L_{adv} denotes the adversarial loss and λ_{adv} denotes the balancing factor of the two-loss terms. The use of STFT losses improved the stability and efficiency in the training process so that the entire model can be easily trained. In addition, with the training method of the generative adversarial network (GAN), the generator can effectively capture the time-frequency distribution of natural speech waveform.

4. Evaluation results

A total of 17 systems were evaluated, at last, 16 from participating teams. In this section, we will discuss the evaluation results in detail. System A is natural speech. Our designated system identification letter is “G”. Audio samples are available at https://whitefu.github.io/blizarrd_2020/.

Table 1: Task 2020-MH1

Sections	Details Description
Section 1	MOS
Section 2	Similarity
Section 3	Paragraph Test
Section 4	Pinyin (with tone) Error rate

The evaluation criteria include four sections as shown in Tabel 1. The synthesized and natural audios were carefully scored in every section by paid listeners, volunteers, and speech experts. Our system achieves unsatisfactory performance in the criteria for the Challenge. Below in each subsection, we will report the detailed results and analyze the reason for the bad performance. We hope it can help the other participant to avoid some mistakes.

4.1. Naturalness evaluation

The boxplot evaluation results of all systems on naturalness are shown in Figure 4. Our system has an average score of 3. Analyzing the results, we guess that the main influence factors are the performance of the vocoder and the prosody of generated speech. For improving the real-time ratio, we choose Parallel WaveGAN as our final vocoder and reduce some parameters by reducing some layers. The performance of this model is superior to the Griffin-Lin algorithm. We also considered using WaveNet as the vocoder part of the system, but without special processing, but the synthesis speed of WaveNet is too slow. And we listen to some unnatural pauses in the synthesized audio. This is because the number of pause corpus is very small (less than 10,000 pieces), which leads to the poor performance of the prosodic prediction model, and the f1 value of the PPH label is less than 40%. In the future, we will deal with the problem by augmenting the pause corpus or introduce other new methods, such as adding a learnable gaussian bias.

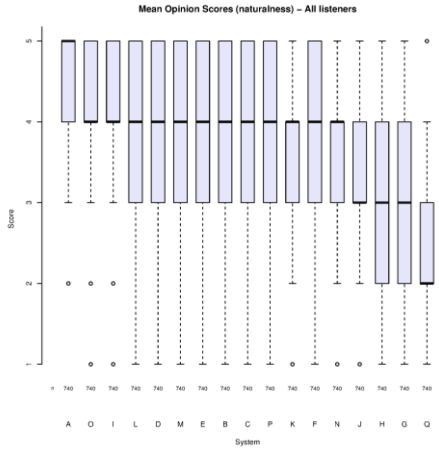

Figure 4: Boxplot of naturalness scores of each submitted system for all listeners.

4.2. Similarity evaluation

Figure 5 presents the mean opinion of similarity evaluations for all systems. In this test, the listeners are asked to judge whether the generated speech is similar to the target speaker. Our system has an average score of 3.0, one point below the first place system. Analyzing the results, we believe that the vocoder performance directly affects the speaker similarity, Meanwhile, the generated prosody from the acoustic model may also have a clear influence on the perceived speaker similarity by the listeners. We have noticed there have been many recent papers about high-quality and real-time vocoder, such as multi-band MelGAN [24] and multi-band LPCNet [25], etc. In the future, we will try to use the new vocoder methods and design Frequency band extension to increase the audio sample rate from 24 kHz to 48 kHz to achieve a higher similarity score.

4.3. Paragraph evaluation

Figure 6 shows that the MOS results of the new paragraph's overall impression given by all listeners for all the systems. In this test, as expected, the original natural speech achieves the highest score of 50. System G achieves a score of 30. We find some skip and repeated errors and pronunciation mistakes in the synthesized speech. The main reason is that forward attention is not robust enough to generate super-long sentences. We prepare two types of paragraph synthesized speech before submit, the one method is synthesized of the whole sentence and the other one method is segmented text line from paragraph level to shorter sentences, synthesized each shorter sentence independently, and concatenated synthesized speech to compose longer speech wave restored at the paragraph level. And the speech synthesized by the second method is better than the first method. But we want to test the long sentence generalization of our system, So we choose the speech synthesized by the first method as our final submit.

4.4. Pinyin and tone error

As we know, the pronunciation of Chinese character is also affected by the tones, so the Pinyin with tone error rate (PTER) are usually used to evaluate the intelligibility of each system. The results are shown in Figure 7. The PTER of system L is 8.6% performs much better than other participants. And the PTER of our system is 16.2% on the intelligibility test. We find that tone error often occurs in long sentences, multiple repeated words, and polyphone words. This is mainly due to the fact the attention mechanism module is not robust enough to generate the correct mapping relation between text and Mel-spectrogram. And we only use the open-source tool, pypinyin, to convert Chinese characters to pinyin (with tone) sequence, it's not good enough to deal with the polyphonic problem. We have noticed that there have many great attention modules that are suitable to generate long-form sentences, such as Step-wise Monotonic Attention [26], Dynamic Convolution attention [27], etc. In the future, we will try these attention methods and design some polyphone disambiguation methods to improve the intelligibility of our system.

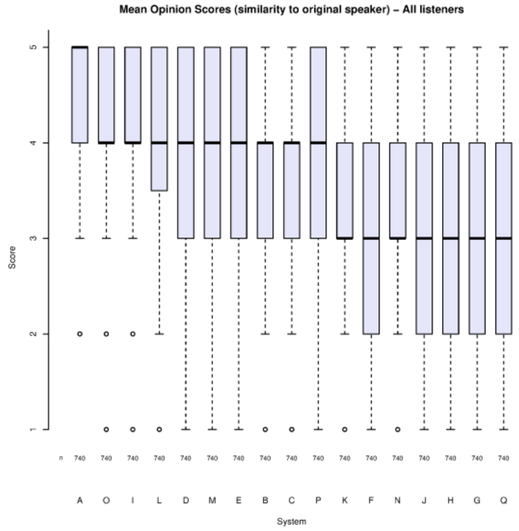

Figure 5: Boxplot of similarity scores of each submitted system for all listeners

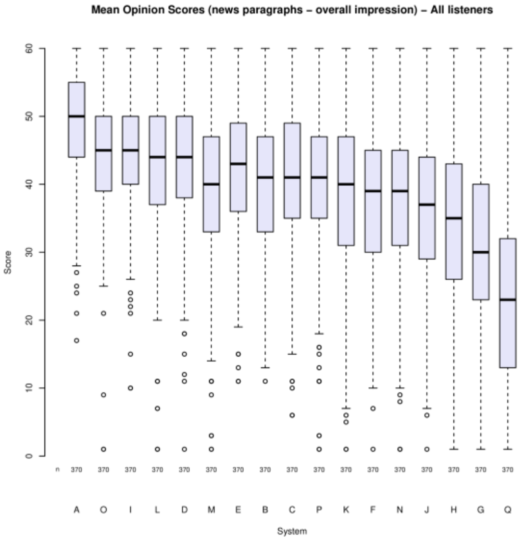

Figure 6: Boxplot of paragraphs MOS scores of each submitted system for all listeners

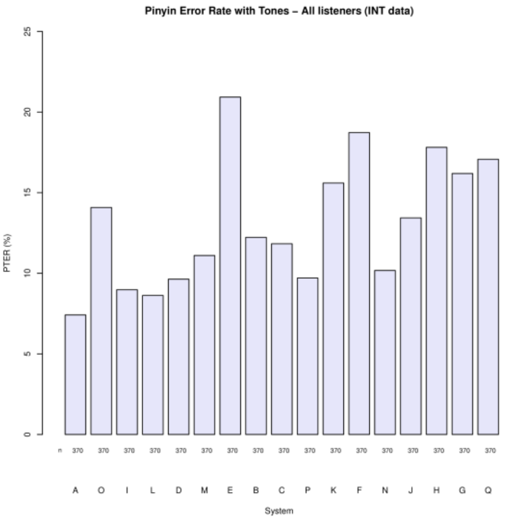

Figure 7: Pinyin error rate (with tone) of each submitted system.

5. Conclusions

This paper presents the details of our submitted system and the result in the Blizzard Challenge 2020. We built a forward attention based end to end speech synthesis system followed by a Parallel WaveGAN vocoder at 24 kHz sampling rate. The results of the listening test for our system are not good. According to the subjective evaluation results, we found many problems in our TTS system, and there is still much space for improvement in our method.

In the future, we will introduce more robust attention mechanisms, such as stepwise attention [26] and dynamic convolutions attention [27], to solve the long-form sentence synthesis and alleviate the times of synthesis errors, like skipping and repeat. At the same time, we will study a real-time vocoder which can generate high-fidelity speech and try to achieve good performance in all criteria for further speech synthesis challenge.

6. References

[1] A. W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 4, 2007.

[2] H. Ze, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7962–7966, 2013.

[3] N. Adiga and S. R. M. Prasanna, “Acoustic Features Modelling for Statistical Parametric Speech Synthesis : A Review Acoustic Features Modelling for Statistical Parametric Speech Synthesis : A Review,” vol. 4602, 2018.

[4] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, “Speech Driven Backchannel Generation using Deep Q-Network for Enhancing Engagement in Human-Robot Interaction,” pp. 4445–4449, 2019.

[5] Y. Wang *et al.*, “Tacotron: Towards end-To-end speech synthesis,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 4006–4010, 2017.

[6] J. Shen *et al.*, “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 4779–4783, 2018,.

[7] Li, Naihan, et al. "Neural speech synthesis with transformer network." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

[8] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech: Fast, robust and controllable text to speech,” in NeurIPS 2019, November 2019.

[9] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” arXiv preprint arXiv:1609.03499, 2016.

[10] N. Kalchbrenner *et al.*, “Efficient neural audio synthesis,” *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 6, pp. 3775–3784, 2018.

[11] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” arXiv preprint arXiv:1910.11480, 2019.

[12] C. Lu, P. Zhang and Y. Yan, "Self-attention Based Prosodic Boundary Prediction for Chinese Speech Synthesis," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7035-7039.

[13] J. X. Zhang, Z. H. Ling, and L. R. Dai, “Forward Attention in Sequence- To-Sequence Acoustic Modeling for Speech Synthesis,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 4789–4793, 2018, doi: 10.1109/ICASSP.2018.8462020.

[14] Chen, Danqi, and Christopher D. Manning. "A fast and accurate dependency parser using neural networks." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[15] Wu, Pengfei, et al. "End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training." 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2019.

[16] H. Guo, F. K. Soong, L. He, and L. Xie, “Exploiting Syntactic Features in a Parsed Tree to Improve End-to-End TTS,” pp. 4460–4464, 2019, [Online]. Available: http://arxiv.org/abs/1904.04764.

[17] Y. Qian, Z. Wu, X. Ma, and F. Soong, “Automatic Prosody Prediction and Detection with Conditional Random Field (CRF) Models,” pp. 135–138, 2010.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[19] V. Klimkov *et al.*, “Phrase break prediction for long-form reading TTS: Exploiting text structure information,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2017-Augus, pp. 1064–1068, 2017, doi: 10.21437/Interspeech.2017-419.

[20] Y. Zheng, J. Tao, Z. Wen, and Y. Li, “BLSTM-CRF based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 47–51, 2018, doi: 10.21437/Interspeech.2018-1472.

[21] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[22] Y. Zheng, X. Li, F. Xie and L. Lu, "Improving End-to-End Speech Synthesis with Local Recurrent Neural Network Enhanced Transformer," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6734-6738.

[23] L. Sheng, D.-Y. Huang, and E. N. Pavlovskiy, “High-quality Speech Synthesis Using Super-resolution Mel-Spectrogram,” 2019, [Online]. Available: http://arxiv.org/abs/1912.01167.

[24] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, “Multi-band MelGAN: Faster Waveform Generation for

High-Quality Text-to-Speech," pp. 1–5, 2020, [Online]. Available: http://arxiv.org/abs/2005.05106.

[25] J.-M. Valin and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5891–5895.

[26] He, Mutian, Yan Deng, and Lei He. "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS." arXiv preprint arXiv:1906.00672 (2019).

[27] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," arXiv preprint arXiv:1910.10288, 2019.