The RoyalFlush Synthesis System for Blizzard Challenge 2020

Jian Lu, Zeru Lu, Ting He, Peng Zhang, Xinhui Hu, Xinkang Xu*

Zhejiang Hithink RoyalFlush AI Research Institute, HangZhou, P.R. China

lujian@myhexin.com

Abstract

The paper presents the RoyalFlush synthesis system for Blizzard Challenge 2020. Two required voices are built from the released Mandarin and Shanghainese data. Based on end-to-end speech synthesis technology, some improvements are introduced to the system compared with our system of last year. Firstly, a Mandarin front-end transforming input text into phoneme sequence along with prosody labels is employed. Then, to improve speech stability, a modified Tacotron acoustic model is proposed. Moreover, we apply GMM-based attention mechanism for robust long-form speech synthesis. Finally, a lightweight LPCNet-based neural vocoder is adopted to achieve a nice traceoff between effectiveness and efficiency.

Among all the participating teams of the Challenge, the identifier for our system is N. Evaluation results demonstrates that our system performs relatively well in intelligibility. But it still needs to be improved in terms of naturalness and similarity.
Index Terms: Blizzard Challenge 2020, speech synthesis, end-to-end, attention, LPCNet

1. Introduction

The purpose of the Blizzard Challenge, which has been held annually since 2005, is to better understand and compare speech synthesis techniques proposed by different participants on the same corpus.

Speech synthesis or text-to-speech(TTS) is a technique that converts the normal text into human-like speech. Intelligibility and naturalness are two key points of a speech synthesis system. Until now, there are mainly three types of popular speech synthesis techniques described as below.

- **Concatenation synthesis:** Concatenation synthesis is based on the concatenation of recorded speech units. Ling et al. [1] presented HMM-based unit selection method to determine the selected speech units for generating speeches in Blizzard Challenge 2007. Concatenation TTS directly selects natural speech units from a recorded speech database, which enables the system to generate speech with natural quality. However, as the footprint of the stored data is reduced, desired units may be unavailable in the database, and audible discontinuities may result.
- **Statistical parametric speech synthesis:** Statistical parametric speech synthesis(SPSS) can be divided into HMM-based synthesis and NN-based synthesis. HMM-based synthesis is a method to synthesize speech at the foundation of hidden Markov models. In this approach, the frequency spectrum (vocal tract), fundamental frequency (voice source) and duration (prosody) of speech are simultaneously modeled by using HMMs [2]. On the other hand, NN-based method is based on deep nueural network(DNN) [3] or long short term memory(LSTM) [4]. Compared with concatenation synthesis, speech synthesized by SPSS uses a relatively small corpus, but its quality is relatively poor. However, because of its stability, the SPSS method has been utilized in practical applications before the emergence of end-to-end approach.
- **End-to-end synthesis:** In recent years, end-to-end speech synthesis technologies have made rapid progress and achieved remarkable performances. In 2017, google proposed Tacotron 2 [5] which predicts mel spectrograms, following with the Wavenet vocoder [6], and achieved synthesized speech with high quality close to human beings. FastSpeech [7] and FastSpeech2 [8] were proposed by Microsoft to solve the problem of slow inference. WaveRNN [9], a single-layer RNN vocoder that matches the quality of WaveNet, was proposed in 2018. LPCNet [10], a WaveRNN variant can be deployed on mobile phones. Compared with above two approaches, end-to-end synthesis simplies traditional pipeline and is capable of generating better speech.

Following the recent progress of speech synthesis, we adopt an end-to-end architecture for tasks in Blizzard Challenge 2020. A modified Tacotron model is proposed to better predict acoustic features. Location-relative GMM attention [11, 12] is applied as a replacement for Location-sensetive attention [13, 5] in Tacotron 2. LPCNet vocoder is used to generate waveforms from the predicted acoustic features. Moreover, we generate prosody boundary labels for the released Mandarin data with help of an annotation tool, and use phoneme sequences and prosody labels as inputs of the modified Tacotron model to alleviate the controllability problem.

The remainder of this paper is organized as follows. In Section 2, we introduce the tasks in Blizzard Challenge 2020. In Section 3, our system is described in detail. The evaluation results are given in Section 4. Finally, Section 5 presents some concluding remarks to end the paper.

2. The tasks in Blizzard Challenge 2020

There are two tasks in Blizzard Challenge 2020 as follows:

- **Hub task 2020-MH1: Mandarin Chinese Found Data** - About 9.5 hours of speech data from one native Mandarin speaker is provided. The hub task is to build a voice from this data.
- **Spoke task 2020-SS1: Shanghainese Found Data** - About 3 hours of speech data from one native Shanghainese speaker is provided. The spoke task is to build a voice from this data.

While hub task is very similar to the task of last year, spoke task introduces Shanghainese to Blizzard Challenge for the first time. We will describe our systems for these 2 tasks in the following section.

*Work done during internship at RoyalFlush

10.21437/VCC_BC.2020-9

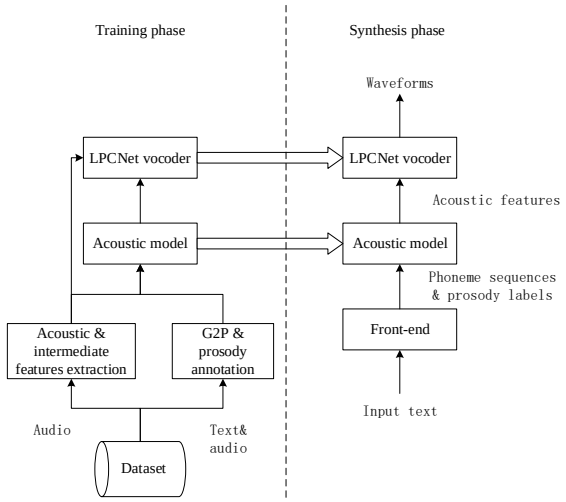


Figure 1: *The overall architecture of RoyalFlush TTS system.*

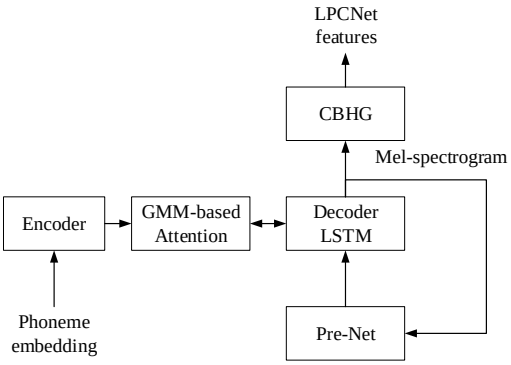


Figure 2: *The architecture of acoustic model.*

3. System description

The overall architecture of the RoyalFlush speech synthesis system is shown in Figure 1. It consists of two parts, the training phase and the synthesis phase. We construct our systems for the hub task and the spoke task with the same architecture. The two systems have only slight differences in linguistic features, which will be described as follows.

In training phase, the <text, audio> pairs from the released datasets are utilized. For the released Mandarin text in hub task, an internal grapheme to phoneme(G2P) tool and a prosody annotation tool are used to generate linguistic features, including phoneme sequence and prosody labels. Here, the prosody labels are in the form of typical three-layer structure[14]. Meanwhile, for the spoke task, the phoneme sequences are provided directly from the Challenge committee, they are directly used as the linguistic features. In this work, we define the LPCNet features as consisting of 18 Bark-scale [15] cepstral coefficients(BFCC) and 2 pitch parameters. The 20-dimensional LPCNet features are extracted from audio as acoustic features, while 80-dimensional mel-spectrgrams are also extracted as intermediate features of acoustic model. The linguistic features, acoustic features, as well as intermediate features are used for training the modified Tacotron acoustic model, and LPCNet features are used independently for training LPCNet vocoder.

In synthesis phase, we adopt front-end to convert Mandarin text in evaluation set of the hub task to linguistic features. However, we directly use phoneme sequences, which are provided by the challenge committee in evaluation set of the spoke task as linguistic features. It is worth noting that we do not split long paragraphs into short sentences, because our acoustic model is robust to synthesize long sentences. The linguistic features are fed into the system and speech waveforms are then generated.

3.1. Data processing

All training speech data are provided by Challenge committee, no external data are used in our systems.

Mandarin data for the hub task contains 4365 audio files, with a sampling rate of 48kHz. The files are about 8 seconds on average, and 9.5 hours in total. In the data processing stage, both the audio and text are processed. Firstly, the audio are down-sampled to 16 kHz. Then we use a prosody annotation tool to automatically split long speech into short ones and add prosody labels to corresponding text according to prosody boundary detection. Then we use a G2P tool to convert the text into phoneme sequence. Linguistic features, including phoneme sequence and prosody labels, are taken as input of modified Tacotron acoustic model.

Shanghainese data for the spoke task contains 1900 audio files, with a sampling rate of 16kHz. The files are about 5.6 seconds on average, and 3 hours in total. As the dataset has been provided with phoneme transcriptions, we directly take these Shanghainese phoneme sequence as input of acoustic model without use of prosody annotation or G2P module.

Both mel-spectrgrams and LPCNet features are extracted for acoustic model training. Here, mel-spectrgrams are extracted with 10ms hop size to match the size of LPCNet features, and LPCNet features are extracted using mozilla's official tool [16]. LPCNet features are used as target output of acoustic model, while mel-spectrgrams are introduced as intermediate representation of acoustic model as mentioned above.

3.2. Front-end

Front-end is used to transform input Mandarin text into linguistic features in synthesis phase. It works in following process. Mandarin text first passes through a rule-based text normalization module, then a Conditional Random Field(CRF) [17] based prosody prediction module is utilized to predict the three-layer prosody boundary labels for the normalized Mandarin text. Finally, a G2P module, with polyphone disambiguation [18] as its core component, is used to acquire linguistic features in the phoneme level.

3.3. Acoustic model

Traditional SPSS systems are complex, which generally consist of a duration model and an acoustic model. These modules need laborious feature engineering, and errors from independently trained modules will be accumulated. To avoid such problems, we adopt end-to-end architecture with attention mechanism for our acoustic model, for which no extra module is needed, including duration model.

The implementation of acoustic model in our system mainly refers to Tacotron 2, however we modify the architecture to better predict acoustic features, as shown in Figure 2. Different from original architecture, the acoustic model takes lin-

guistic features generated by front-end as input, and the 20-dimensional acoustic features for LPCNet vocoder as output. The detailed descriptions of each component will be presented as follows.

3.3.1. Encoder

The encoder is used to covert linguistic features into text hidden representations. It consists of 3 convolutional layers and 1 bi-directional LSTM [19]. The configurations of these layers are the same as in Tacotron 2.

3.3.2. Attention

Attention mechanism [20] is widely used in seq-to-seq models [21] to align input and output sequences. Tacotron [22] adopts content-based [23] attention while Tacotron 2 adopts location-sensitive attention. However, these systems sometimes suffer from alignment failures, which may lead to missing characters or incomplete synthesis. Most importantly, they lack ability to process long text paragraphs like the cases in evaluation set of the hub task.

With prior knowledge that text position progresses nearly linearly to time in TTS domain, location-relative GMM-based attention was firstly introduced in [11]. Google further modified the GMM-based attention [12] in improving alignment speed and consistency during training.

This attention mechanism uses mixture of K Gaussions to produce attention weight $\phi_{i,j}$, which indicates the alignment weight at decoder time step i attending to text position j. The implementing form of GMM-based attention is shown in Equation (1), where κ_i, the mean of each Gaussian component, indicates the attending central location, β_i indicates the attending boundary, and α_i indicates the importance of each component within the mixture.

$$\phi_{i,j} = \sum_{k=1}^{K} \alpha_i^k \exp\left(-\beta_i^k \left(\kappa_i^k - j\right)^2\right) \tag{1}$$

To compute the parameters of the attention weight, intermediate parameters $\left(\hat{\alpha}_i, \hat{\beta}_i, \hat{\kappa}_i\right)$ are firstly computed by applying a densely-connected layer in Equation (2), where s_i is the decoder RNN hidden state. Then, the final parameters $\alpha_i, \beta_i, \kappa_i$ are computed by Equations (3)-(5).

$$\left(\hat{\alpha}_i, \hat{\beta}_i, \hat{\kappa}_i\right) = W s_i + b \tag{2}$$

$$\alpha_i = \exp\left(\hat{\alpha}_i\right) \tag{3}$$

$$\beta_i = \exp\left(\hat{\beta}_i\right) \tag{4}$$

$$\kappa_i = \kappa_{i-1} + \exp\left(\hat{\kappa}_i\right) \tag{5}$$

As we see in Equation (5), $\exp\left(\hat{\kappa}_i\right)$ is always positive, which indicates that the GMM-based attention is monotonic and location-relative.

With GMM-based attention applied in our acoustic model, text as long as hundreds of Mandarin characters can be synthesized, while the speech naturalness is preserved.

3.3.3. Decoder

The decoder is used to predict mel-spectrograms from the text hidden representation by using an autoregressive recurrent neural network. The components of the decoder are the same as the original Tacotron 2. The mel-spectrograms predicted at previous time step are passed through a 2 densely-connected layers known as Pre-Net. Output of decoder LSTM which consists of 2 LSTM layers with 1024 units then passes through two linear transforms separately to predict mel-spectrograms and stop tokens. 5-layer convolutional network is applied to improve mel-spectrograms reconstruction.

3.3.4. Postnet - CBHG module

We don't adopt Tacotron 2 architecture to predict LPCNet features directly, due to the instability of synthesized speech. Instead, mel-spectrograms are treated as intermediate features, and CBHG module proposed in Tacotron is used as postnet to transform mel-spectrograms into LPCNet features. Consisting of a bank of 1-D convolutions filters, highway network [24] and bidirectional GRU [25], CBHG is a powerful module for extracting acoustic representations. With this approach, we aim to improve the stability of synthesized speech.

3.4. Vocoder

LPCNet vocoder is a WaveRNN variant which produces speech waveforms from 20 features consisting of 18 cepstral coefficients and 2 pitch parameters. Combining linear predictor [26] which represents vocal tract response, with neural network which predicts LPCNet residual(vocal source signal), the LPCNet vocoder achieves extremely high efficiency while high speech quality is remained.

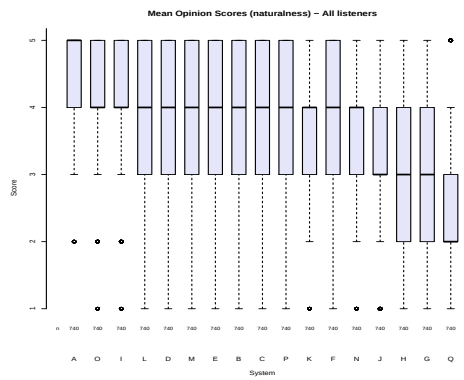
We combine the LPCNet vocoder with the modified Tacotron acoustic model. For hub task, we trained the network for 40 epochs with a batch size of 64. While for spoke task, we trained for 90 epochs. The AMSGrad [27] optimizer is used and models are trained from scratch. Only datasets provided by Challenge committee are used without any external data. Running 3 times faster than real time on a single 2.4GHz Intel Xeon E5 cpu, we have comfirmed that LPCNet achieves higher speech quality than Griffin-lim [28].

4. Evaluation results

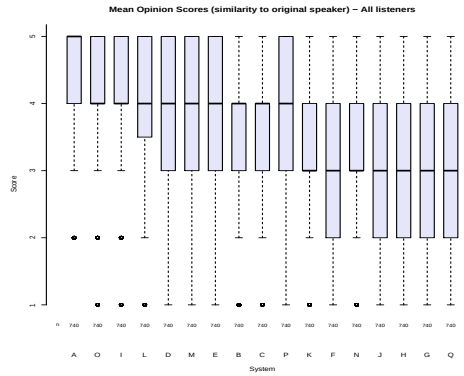
In Blizzard Challenge 2020, 17 systems were evaluated for the hub task, and 9 systems were evaluated for the spoke task. For each task, a natural speech set A was also provided for reference. Among all the participating teams, our system is identified as N in both tasks.

The evaluation results of hub task for all participating systems, including naturalness test, similarity test, intelligibility test and paragraph test, are shown in Figure 3. Meanwhile, the evaluation results of spoke task are shown in Figure 4 with only naturalness test, similarity test and intelligibility test. Here, mean opinion score (MOS) is used to represent the naturalness of the system. Pinyin with tone error rate (PTER), as well as intelligibility score, is used to indicate the intelligibility of the system. Similarity represents how similar the synthetic voice sounds like the original speaker. Paragraph test is evaluated with various criteria, such as pleasantness, emotion and stress.

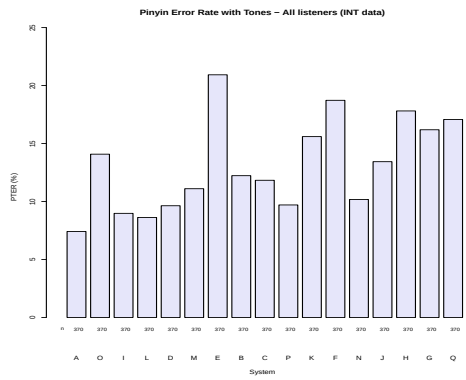
Our system performs relatively well in intelligibility test of hub task, and achieves improvements in all aspects compared with our system of last year. But there are still much space to be improved in naturalness and similarity. One reason why our system is not ideal is that we use the LPCNet vocoder, which keeps a tradeoff between effectiveness and efficiency. Due to resource limitations in both human and computations, we didn't explore more effective and complicated vocoders.



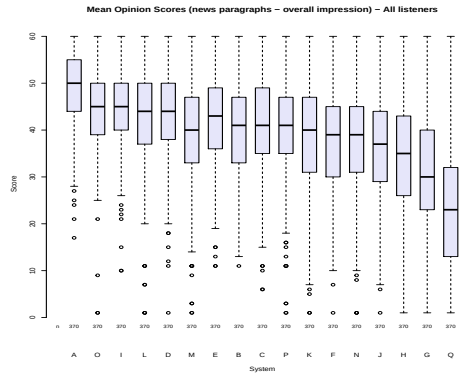
(a) Naturalness Test - MOS



(b) Similarity Test - Similarity

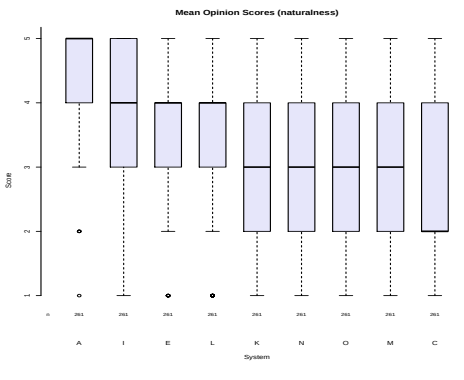


(c) Intelligibility Test - PTER

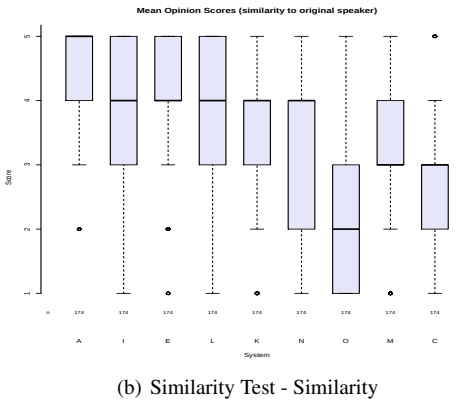


(d) Paragraph test - various criteria

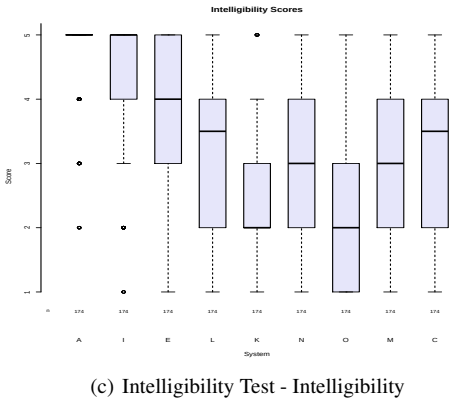
Figure 3: *Evaluation results of hub task.*



(a) Naturalness Test - MOS



(b) Similarity Test - Similarity



(c) Intelligibility Test - Intelligibility

Figure 4: *Evaluation results of spoke task.*

5. Conclusions

This paper gives the description of our submitted system and the evaluation results in Blizzard Challenge 2020. We built an end-to-end architecture based on a modified Tacotron acoustic model, followed by a LPCNet vocoder. Prosody labels are introduced for fine-grained control of synthesized speech. GMM-based attention is applied for robust long-form speech synthesis. Our system achieved relatively good performance in several evaluation aspects for the challenge. However, there are still much work to do in terms of naturalness and similarity. In future work, more effective front-end module and high-quality neural vocoders such as WaveGlow are the main directions of our system.

6. References

[1] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen, and G.-P. Hu, "The ustc and iflytek speech synthesis systems for blizzard challenge 2007."

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*, 1999.

[3] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[4] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[6] 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.

[7] 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 *Advances in Neural Information Processing Systems*, 2019, pp. 3171–3180.

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

[9] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[10] 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.

[11] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[12] 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," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6194–6198.

[13] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[14] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for chinese speech synthesis using blstm-rnn and embedding features," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 98–102.

[15] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[16] J.-M. Valin and J. Skoglund, "Mozilla lpcnet toolbox," https://github.com/mozilla/LPCNet.

[17] G.-A. Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[18] C. Shan, L. Xie, and K. Yao, "A bi-directional lstm approach for polyphone disambiguation in mandarin chinese," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[22] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[23] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[24] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[26] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[27] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.

[28] G. D. W. and L. Jae, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics Speech & Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.