The Sogou speech synthesis system for Blizzard Challenge 2018

Kai Liu, Fanbo Meng, Yang Song, Bo Fan, Wenjun Duan, Wei Chen

Sogou, Beijing, P.R. China

{liukaios3228,songyang206061}@sogou-inc.com

Abstract

This paper presents the Sogou speech synthesis system for Blizzard Challenge 2018. The corpus released to the participants this year is a 6.5-hour children's audio book in British English, which is the same as for the 2017 data release. We build a parametric system for this task. Firstly, a multi-speaker DNN-BLSTM model is applied for mel spectrograms modeling. Then, a modified WaveNet model conditioned on the predicted mel features is used to generate 16-bit speech waveforms at 32 kHz, instead of the conventional vocoder.

This is the first time for Sogou to join the Blizzard Challenge, we have developed speech synthesis for years. The identifier for our system is J, the results show that our submitted system performed good on all the criterion.

Index Terms: Blizzard Challenge 2018, statistical parametric speech synthesis, BLSTM, WaveNet

1. Introduction

The Blizzard Challenge has been held once a year since 2005, in order to better understand and compare research techniques in building corpus-based speech synthesizers on the same data. The basic challenge is to take the released speech database, build a synthetic voice from the data and synthesize a prescribed set of test sentences. The sentences from each synthesizer are the evaluated through listening tests.

The HMM-based statistical parametric speech synthesis (SPSS) method was first proposed and applied successfully in 1999 [1]. In this method, spectrum, pitch and duration are modeled simultaneously in a framework of HMMs. Then, many techniques such as MGE-training [2] and phone duration modeling [3] were proposed to improve the synthesis effect. Due to its maturity and stability , many participating teams prefer HMM as a good baseline system, or as an important part of their systems. And post-filter methods, such as global variance (GV) [4], variance scaling (VS) [5] and modulation spectrum (MS) [6], were also helpful to improve the quality of synthesized speech. In recent years, Deep Neural Networks models have been applied successfully to SPSS [7, 8, 9]. DNN-LSTM models have achieved greater performance in both the front-end text processing [10] and bach-end acoustic modeling [11]. Recently, a postfilter based on a generative adversarial network (GAN) was proposed to compensate for the differences between natural speech and speech synthesized by statistical parametric speech synthesis [12].

In 2007, the unit selection and concatenation system was first used in Blizzard Challenge. The HMM-based unit selection system, which uses maximum likelihood criterion of statistical models to guide the selection of phone-sized candidate units, outperformed all the other systems [13]. Comparing SPSS system, unit selection and concatenation systems has great advantage in similarity , quality and expression. They directly use real speech units from the original corpus for concatenation and speech generation. Since then, many teams have used this method and achieved good results [14, 15, 16]. Moreover, LSTM-RNN based unit selection systems were built to improve the overall performance [17, 18, 19].

Van den Oord et al. proposed WaveNet [21], a fully probabilistic and autoregressive deep neural network, with the predictive distribution for each audio sample conditioned on all previous ones. In Blizzard 2017, the WaveNet system had a good performance [22]. Deep Voice 1 and 3 [23, 24] and the Parallel WaveNet [25] have done more attempts and optimizations. An end-to-end architecture named Tacotron [26, 27], followed by a modified WaveNet model acting as a vocoder, achieve a mean opinion score (MOS) comparable to professionally recorded speech. We had followed this work in the Blizzard 2018. Perhaps due to the limitation or the high expression of the data, we didn't have good results. Inspired by recent literatures, we implemented a DNN-BLSTM based TTS system followed by a modified WaveNet neutral vocoder.

The paper is organized as follows. Section 2 introduces the details of the English task in Blizzard 2018. Section 3 describes each modules of our system building. Section 4 presents the results of the benchmark systems and all the participation. Finally, the convolution is given in Section 5.

2. The task in Blizzard 2018

There is only single task this year.

- **Single task 2018-EH1:** UK English Children's Audiobooks - About 6.5 hours of speech data from professionally-produced childrens audiobooks will be released. This is the same data as used in last years Blizzard Challenge. All data are from a single speaker. The task is to build a voice from this data that is suitable for reading audiobooks to children.

The task is the same as last year's challenge. In the following sections, we will introduce our speech synthesis system in details.

3. Sogou speech synthesis system

As shown in Figure 1, our system consists of two parts, training and synthesis. At training phase, we first use a modified Festival front-end to predict phoneme and other linguistic features, and modify referred to the speech manually. And these annotations are used for ToBI prediction models training. Then we train a HMM model for the force alignment. Thirdly, we train a DNN-LSTM duration model and a DNN-BLSTM acoustic model. Finally, a WaveNet neural vocoder conditioned on the predicted mel spectrograms is trained with the multi-speaker model as an initial model.

At testing phase, expressive linguistic features are predicted by the front-end text analysis. Then feed them into the duration and acoustic models to predict frame-level mel features.

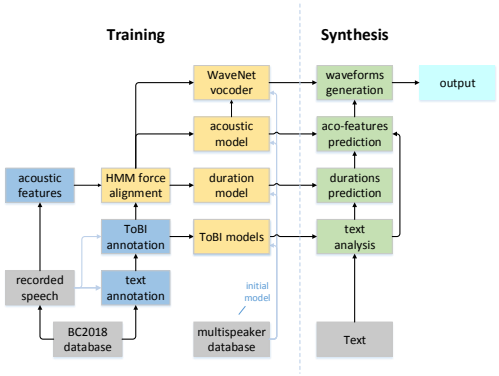

Figure 1: *The flowchart of Sogou TTS system.*

Finally the WaveNet neural model is used to generate speech waveforms sample by sample conditioning on the predicted mel spectrograms.

3.1. Data processing

The audiobook data provided by the organizing committee contains two parts: the transcribed part and the untranscribed part. The phonemes and time boundaries are given in the transcribed part. The untranscribed part corresponds to the pdf files. We manually extract the text from the pdf files. The phoneme sequence is predicted by Festival front-end. Then we use our automatic speech recognition engine to do phoneme level alignment. For the whole corpus, we use Festival front-end to predict the information of stress, accent, part-of-speech(POS), phrase boundary and TOBI boundary tone. Then all the information is manually checked. We also manually divide the emotions of audio into three categories - neutral, happy and sad. The emotion tag is also added to the label. We find that when synthesizing news topics, it is more suitable to adopt neutral emotions. When synthesizing novel topics, it is more suitable to adopt a variety of different emotions. At last, In order to facilitate subsequent training, we divide the long sentence into several clauses.

For the acoustic features, we extracted lf0 and 40-dimensions mcep from waveforms at 16 kHz with 25 ms frame size and 5 ms frame hop. Then we trained a 5-state forced alignment HMM model for the durations. We also extracted 120-dimensions mel spectrograms at 32 kHz using a 50 ms frame length, 12.5 ms frame hop and a Hann window function, which was used to train the acoustic model and as local condition for the WaveNet model.

3.2. Front-end

We use the Festival English front-end to build the initial context label. But the phrase boundary model is configured with a feed-forward layer of 512 nodes and two GRU layers of 128 nodes. A sigmoid output layer is used for phrase boundary prediction. The input feature includes the current and adjacent word embedding, POS and word position. The position of word is normalized by the length of the sentence. In our internal experiment, hundreds of thousands of annotated sentences is used to build the phrase boundary model.

As vowels in some words are pronounced differently in Southern British English, compared with American English. In cases like spot, not, and doctor, the vowel /o/ is pronounced /ɑ/ in American English but /ɒ/ in Southern British English. British and American people also pronounce /r/ differently. In words like car, farm and tour, /r/ sound is almost left out in Southern British English but it is retro-flexes in American English. So we use the Oxford Advanced Learner's Dictionary to replace the CMU Pronouncing Dictionary that comes with Festival English front-end.

3.3. Back-end

For the duration model training, the model is configured with a feed-forward layer of 1024 nodes and two LSTM layers consisting of 256 nodes. The output layer is a feed-forward layer of 6 nodes. Input feature for duration prediction are linguistic features derived from a set of questions about linguistic context. The output feature contains the duration and phoneme and five states. The five state boundaries are obtained by forced alignment of the HMM in advance. In order to obtain better prediction results, we first train the duration model of the hundreds of hours of corpus accumulated in the past, and then adaptively train the duration model of the data of the Blizzard Challenge 2018.

Similarly for the acoustic model training, we used the pre-trained multi-speaker model as an initial model. The inputs of our acoustic model are linguistic features, frame features and the speaker embedding. We trained a DNN-BLSTM model, containing 3 fully connected layers of 1024 hidden ReLU units and a stack of 2 bidirectional LSTM layers with 512 units. The speaker embedding was added to the last fully connected layer before the activation function.

3.4. WaveNet based vocoder

A WaveNet model is trained to replace the conventional vocoder. WaveNet is a fully probabilistic and autoregressive generative model that can generate waveforms directly.

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, x_2, ..., x_{t-1}, \mathbf{h}). \quad (1)$$

where $\mathbf{x} = \{x_1, ..., x_T\}$ is a given waveform, each audio sample x_t is conditioned on the samples at all previous timesteps. $\mathbf{h}$ is conditional inputs, here we use predicted mel spectrograms as local condition and speaker embedding as global condition.

As shown in Figure 2, We use a version of the WaveNet architecture modified from [21] and [23]. As in the original architecture, the model consists of 40 layers, grouped into 4 dilated residual block stacks of 10 layers. In every stack, the dilation rate increases by a factor of 2 in every layer, and no dilation for the first layer. The predicted mel spectrograms are passed through a stack of 2 bidirectional QRNN [28] layers with 256 units. To work with the 12.5 ms frame hop of the spectrogram frames at 32 kHz, the QRNN output is up-sampled 400 times using four transposed convolution layers. We follow Parallel WaveNet [25] and use a 10-component mixture of logistic distributions to generate 16-bit samples at 32 kHz instead of the softmax layer. However, there is no perceptible improvement and the model takes longer to converge.

In order to obtain better prediction results, we first train the duration model of multi-speakers without local condition, then adaptively train the model on the Blizzard 2018 data conditioning on both the speaker embedding and the ground truth mel spectrograms. The synthesized speeches sound much better to use ground truth as local condition then the predicted. This is

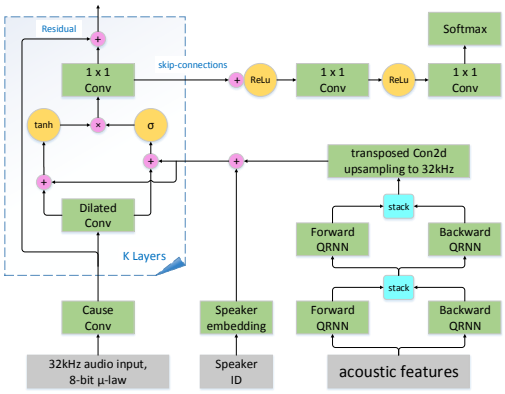

Figure 2: *The modified WaveNet architecture.*

likely because of inherence noise between the predicted features and ground truth. So, we retrained the model conditioned on the predicted mel spectrograms and obtained the best performance. And we get the same observation that adding speakers resulted better compared to training solely on a single speaker, which has been proved in [21].

4. Results

There are 15 systems in total, ten from participating teams, four benchmarks, and one natural speech. System A is natural speech recorded by the original speaker. System B is the Festival unit selection benchmark system by CSTR. System C is the HMM benchmark built using the HTS toolkit. System D and E are both DNN benchmark built using the HTS toolkit, where E employs trajectory training. System F to O are the 10 participating teams, and system J is ours.

Table 1: *Task 2018-EH1*

Sections	Detailed Description
section 1	MOS (variour criteria) - book paragraphs
section 2	MOS (variour criteria) - book paragraphs
section 3	MOS (naturalness) - book sentences
section 4	MOS (naturalness) - book sentences
section 5	Similarity with original speaker
section 6	Semantically unpredictable sentences
section 7	Semantically unpredictable sentences

The evaluation comprised seven sections showed in Table 1. In each section, a set of samples from all the systems were judged by each listener. Finally, our system has shown consistent performance (standing in top tow) in all the criterion for the Challenge. Details are as follows.

4.1. Paragraph test

In paragraph test, seven dimensions of testing were used to evaluate different aspects of synthesized paragraphs, including overall impression, pleasantness, speech pauses, stress, intonation, emotion and listening effort. In each part listeners listened to one whole paragraph from a children's book and chose a score from 1 to 60.

Figure 3 presents the overall impression results of all systems. For the four benchmark systems, the Festival unit selection benchmark system(B) ranks 4th and outperforms the HMM and DNN benchmarks (with trajectory training). Similar to the results of previous years, unit selection system is still competitive and preferred by listeners. Our system outperformed the four benchmark systems and other submitted systems except for systems K and I. And on the other evaluations of paragraphs(pleasantness, speech pauses, etc.), again only system K outperformed ours in some aspects. The mean opinion scores of our system are listed in (Table 2). Over all, our submitted system has very good expression in all paragraph evaluations.

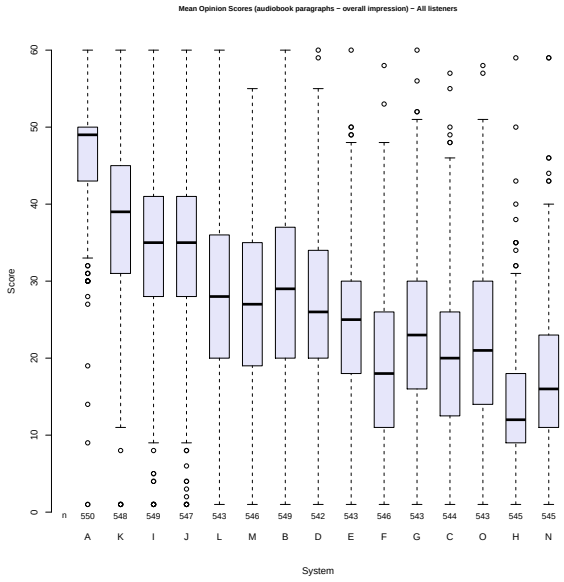

Figure 3: *Overall Impression on book paragraphs of each submitted system.*

Table 2: *Paragraph listening test scores of our system*

Criterion	Mean Opinion Score
overall impression	34
pleasantness	33
speech pauses	36
stress	35
intonation	35
emotion	35
listening effort	34

4.2. Naturalness test

The mean opinion scores for naturalness from all listeners are show in Figure 4. In each part, listeners listened to one sample from children's book sentences, then choose a score which represented how natural or unnatural the sentence sounded. Benchmark system B, D and E perform very close results, they three are obviously better than HMM benchmark system C. Our results are significantly better than benchmark systems and most participates, only system K is better than ours.

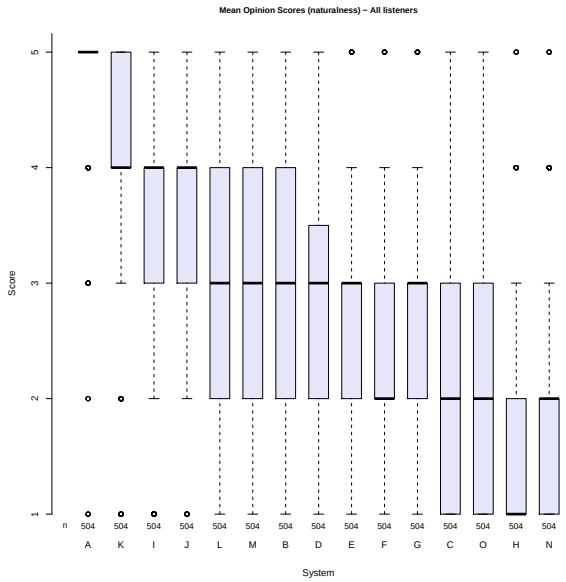

Figure 4: *Mean opinion score for naturalness on book sentences with ratings from all listeners.*

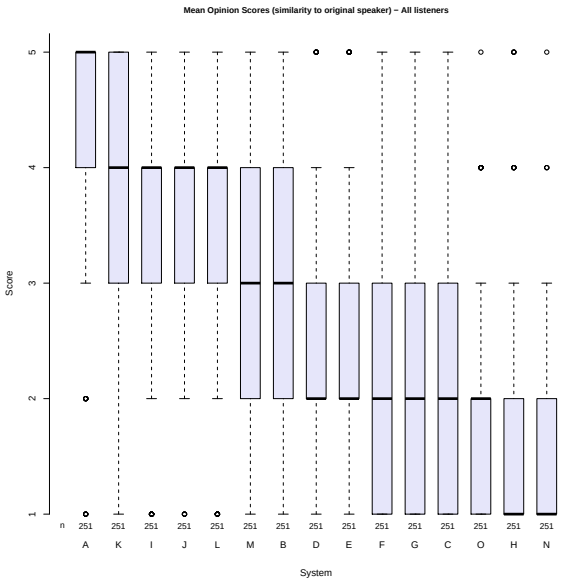

Figure 5: *Mean opinion score for similarity on book sentences with ratings from all listeners.*

4.3. Similarity test

The mean opinion of similarity evaluation results from all listeners is presented in Figure 5, this part is also on children's book sentences. In each part, listeners listened to 4 reference samples of the original speaker and one synthetic sample. They choose a score which represented how similar the synthetic voice sounded to the voice in the reference samples. Unit selection systems always performed outstandingly due to the use of high-quality original speech, while SPSS systems limited by the vocoder and the over-smoothing acoustic model.

As shown in Figure 5, there is no significant difference between the four systems of K, I, J and L. We trained a WaveNet model to generate 16-bit samples at 32 kHz. Because we find that audios with 32 kHz sampling rate sound almost no loss of similarity to the original recordings, and have very high quality.

4.4. Intelligibility test

The word error rates(WER) of all submitted systems are presented in Figure 6. Semantically Unpredictable Sentences (SUS) were designed to test the intelligibility of the synthetic speech. Listeners were allowed to listen to each sentence only once and then typed in what they heard. In the last two terms, DNN and BLSTM based system generally got lower WERs. There are four systems tied for the first place, two DNN benchmark systems, system I and ours. The results show that our system has high intelligibility as well.

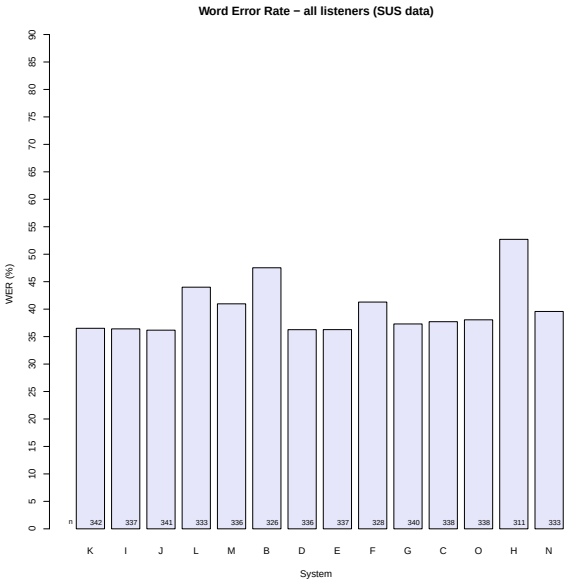

Figure 6: *Word error rate of each submitted system.*

5. Conclusions and future work

This paper presents the details of our submitted system and the results in Blizzard Challenge 2018. We built a DNN-BLSTM based statistical parametric speech synthesis following a WaveNet vocoder at 32 kHz. Our system achieved consistently good performance in all the criterion for the Challenge.

Takuma Okamoto et al. has proposed a subband WaveNet [29] for high-quality synthesis. In future work, we will make more attempts in waveform modeling at high sample rate. And we will continue to investigate multi-speaker speech synthesis.

6. Acknowledgements

We would like to thank Yuchao Zhang from Northwestern Polytechnical University for the work on WaveNet during the internship in Sogou. Thank intern students Jiaqi Wang, Panpan Yang, and Xiaolin Lu et al. for supporting data processing and labeling. And also thank British native speakers Simone, et al. for their expert suggestions and assist in pronunciation.

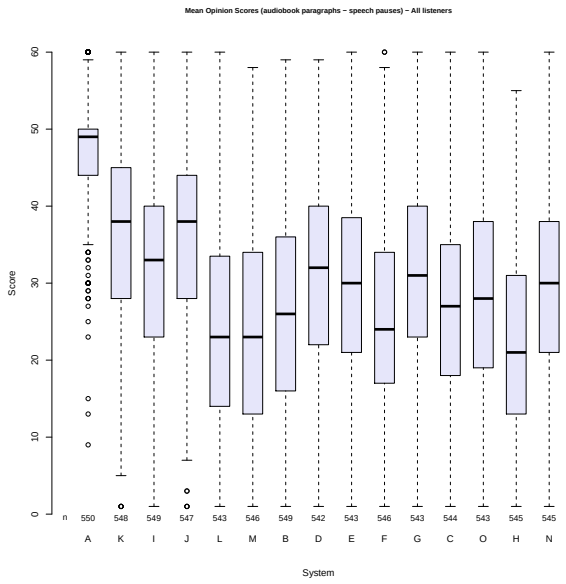

Figure 7: *Speech Pauses on book paragraphs of each submitted system.*

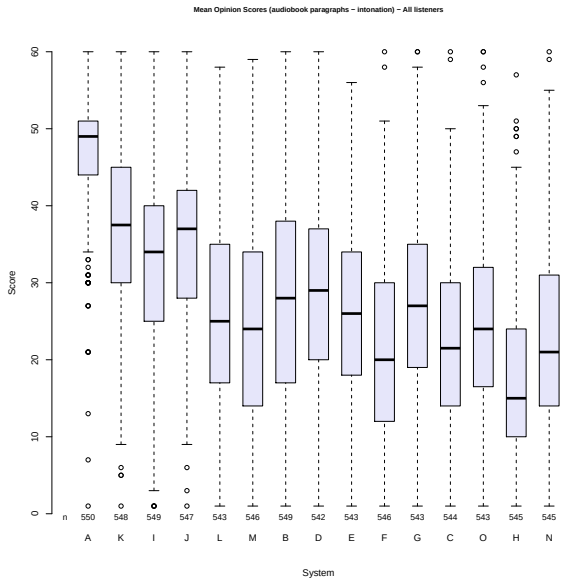

Figure 8: *Intonation on book paragraphs of each submitted system.*

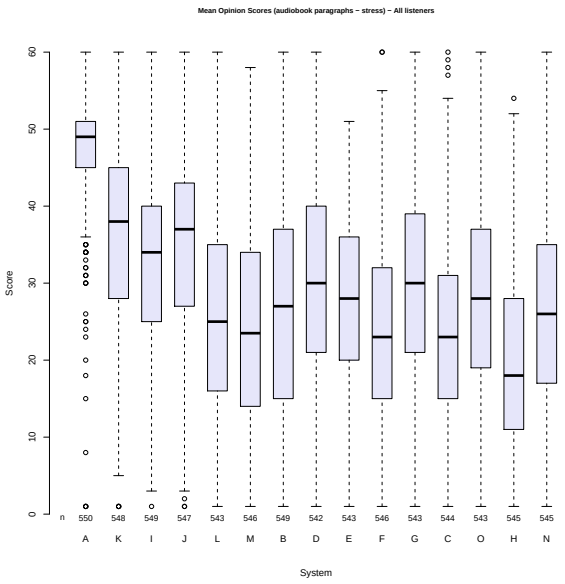

Figure 9: *Stress on book paragraphs of each submitted system.*

7. References

[1] Yoshimura T, Tokuda K, Masuko T, et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Sixth European Conference on Speech Communication and Technology*, 1999.

[2] Wu, Yi-Jian, and Ren-Hua Wang, "Minimum generation error training for HMM-based speech synthesis," *Acoustics, Speech and Signal Processing*, 2006.

[3] Yi-Jian Wu, "Research on HMM-based Speech Synthesis," *Ph.D Thesis, University of Science and Technology of China*, 2006. [in Chinese]

[4] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *IEICE TRANSACTIONS on Information and Systems*, 2007, 90(5): 816-824.

[5] Siln H, Helander E, Nurminen J, et al., "Ways to implement global variance in statistical speech synthesis," *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[6] Takamichi S, Toda T, Neubig G, et al., "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP 2014*, 2014, pp. 290C294.

[7] Ze, Heiga, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," *Acoustics, Speech and Signal Processing (ICASSP)*, 2013 IEEE International Conference on. IEEE, 2013: 7962-7966.

[8] Ling Z H, Kang S Y, Zen H, et al., "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends" *IEEE Signal Processing Magazine*, 2015, 32(3): 35-52.

[9] Wu Z, Valentini-Botinhao C, Watts O, et al., "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis" *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, 2015: 4460-4464.

[10] Ding C, Xie L, Yan J, et al., "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features" in *Automatic Speech Recognition and Understanding*, 2015, pp. 98C102.

[11] Zen, Heiga, and Ha?im Sak., "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis." *Acoustics, Speech and Signal Processing (ICASSP)*, 2015 IEEE International Conference on. IEEE, 2015.

[12] Kaneko T, Kameoka H, Hojo N, et al., "Generative adversarial network-based postfilter for statistical parametric speech synthesis" *Proc. ICASSP*, 2017, 2017: 4910-4914.

[13] Ling Z H, Qin L, Lu H, et al., "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007" *Blizzard Challenge Workshop. 2007*, 2007.

[14] Yu Y, Zhu F, Li X, et al., "Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013" *Blizzard Challenge Workshop. 2013*, 2013.

[15] Chen L H, Ling Z H, Jiang Y, et al., "The USTC System for Blizzard Challenge 2013" *Blizzard Challenge 2013 workshop*, 2013.

[16] Ronanki S, Ribeiro M S, Espic F, et al., "The CSTR entry to the Blizzard Challenge 2017" *The Blizzard Challenge 2017 Workshop, Stockholm.*, 2017.

[17] L.-H. Chen, Y. Jiang, M. Zhou, Z.-H. Ling, and L.-R. Dai, "The ustc system for blizzard challenge 2016" in *The Blizzard Challenge Workshop*, 2016.

[18] Liu L J, Ding C, Jiang Y, et al., "The IFLYTEK system for blizzard challenge 2017" *The Blizzard Challenge 2017 Workshop, Stockholm.*, 2017.

[19] Lu H, Lei M, Meng Z, et al., "The Alibaba-iDST Entry to Blizzard Challenge 2017" *The Blizzard Challenge 2017 Workshop*, 2017.

[20] Lu H, Lei M, Meng Z, et al., "The Alibaba-iDST Entry to Blizzard Challenge 2017" *The Blizzard Challenge 2017 Workshop*, 2017.

[21] Van Den Oord A, Dieleman S, Zen H, et al., "Wavenet: A generative model for raw audio" *CoRR abs/1609.03499*, 2016.

[22] Hu Y J, Ding C, Liu L J, et al., "The USTC system for blizzard challenge 2017" *Proc. Blizzard Challenge Workshop*, 2017.

[23] Arik S O, Chrzanowski M, Coates A, et al., "Deep voice: Real-time neural text-to-speech" *arXiv preprint arXiv:1702.07825*, 2017.

[24] Ping W, Peng K, Gibiansky A, et al., "Deep voice 3: 2000-speaker neural text-to-speech" *arXiv preprint arXiv:1710.07654*, 2017.

[25] Oord A, Li Y, Babuschkin I, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis" *arXiv preprint arXiv:1711.10433*, 2017.

[26] Wang Y, Skerry-Ryan R J, Stanton D, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model" *arXiv preprint*, 2017.

[27] Shen J, Pang R, Weiss R J, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions" *arXiv preprint arXiv:1712.05884*, 2017.

[28] Bradbury J, Merity S, Xiong C, et al., "Quasi-recurrent neural networks" *arXiv preprint arXiv:1611.01576*, 2016.

[29] Okamoto T, Tachibana K, Toda T, et al., "An investigation of subband WaveNet vocoder covering entire audible frequency range with limited acoustic features" in *Proc. ICASSP 2018*, 2018.