The USTC System for Blizzard Challenge 2019

Yuan Jiang[1,2], *Ya-Jun Hu*[2], *Li-Juan Liu*[2], *Hong-Chuan Wu*[2], *Zhi-Kun Wang*[2], *Yang Ai*[1], *Zhen-Hua Ling*[1], *Li-Rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R. China
[2]iFLYTEK Research, Hefei, P.R. China

yuanjiang@iflytek.com, ay8067@mail.ustc.edu.cn

Abstract

This paper introduces the details of the speech synthesis system developed by the USTC-iFlytek team for Blizzard Challenge 2019. An 8-hour Chinese male talkshow audio corpus was released to the participants this year. A statistical parametric speech system (SPSS) that modeling speech waveforms was built for the task. Firstly, Bidirectional Encoders Representations from Transformers (BERT)-based multi-task models were adopted for the front-end task. LSTM-RNN models were used in duration modeling and acoustic modeling for back-end task. Then, we proposed an autoregressive model to improve the duration modeling, and a generative adversarial network (GAN) to relieve the over-smoothing in acoustic modeling. At last, a WaveNet based neural vocoder was utilized to model speech waveforms from acoustic feature instead of melcepstrum vocoder. The evaluation results show the excellent performance of the submitted system.

Index Terms:Blizzard Challenge 2019, SPSS, BERT, autoregressive, GAN, WaveNet

1. Introduction

The USTC team has been submitting entries to Blizzard Challenge speech synthesis evaluation for twelve years since 2006. In the first participation, we submitted a improved hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) system using line spectral pairs (LSP) [1]. In the next two years, in order to exploit the advantage of the large scale of the released corpus and achieve better performance, an HMM guided unit selection and waveform concatenation system was submitted and achieved promising performance [2][3]. In the challenge of 2009, we adopted the minimum generation error (MGE) criterion in decision tree clustering and used a cross validation method to automatically control the scale of the decision tree [4]. In 2010, as the size of released corpus was growing, a globally covariance tying strategy was utilized to reduce the footprint of the model, as well as improve the model training efficiency [5]. In addition, a syllable-level F0 model was further introduced to consider the long term prosody correlations between unit candidates to be concatenated. In the Blizzard Challenge 2011, we proposed an improved unit selection criterion, maximum log likelihood ration (LLR) criterion, to improve the performance of unit selection [6]. In 2012, a set of audiobook corpus with different recording channels were released. We utilized a channel equalization method to compensate these channel differences [7]. A large corpus with hundreds of hours of unaligned audiobooks was released in Blizzard Challenge 2013. The scale of the corpus was a challenge to both the computation efficiency and robustness of the submitted system. A phone dependent model clustering method was utilized to enable parallel training of HMMs on such a large corpus. We also proposed a weight optimization method to automatically tune the weights of each component in the costs of our unit selection criterion [8]. In Blizzard Challenge 2013, 2014 and 2015, corpus of many Indian languages were released to non-native participants. We adopted a letter-to-sound (L2S) [9] method to build frontend text processing for Hindi, and used a simple character based front-end for other Indian languages [8]. We also adopted deep neural network DNN-based data driven spectral post-filtering techniques [10] and modulation spectrum [11] based ones to improve the quality of synthetic speech [12]. A non-uniform units were used for unit selection and concatenation in our system to improve the stability of our system for Blizzard Challenge 2015 [13]. Last three year, a highly expressive children's audiobook corpus was released for system construction. In our submitted system, an long short term memory LSTM-based recurrent neural networks were adopted for tone and breaking indices (ToBI) prediction to achieve high expressiveness. Another DNN-based unit embedding model was built and the unit vector was adopted as phone unit feature, which was used to evaluate contextual similarities between candidates and target units during the unit selection time [14] [15] [16].

Unit selection systems always achieve excellent performance at the Blizzard Challenge in recent years. Due to the over-smoothing problem in acoustic modeling and the restriction of vocoder, SPSS system performs not good enough in voice quality and similarity [17]. However, SPSS is still a hot research topic in academia and widely used in industry because of its flexibility and small footprint. As reported in recent literature, deep learning techniques have been applied successfully to SPSS [18]. LSTM-RNN has achieved great performance in both the front-end text processing [19] and back-end acoustic modeling [20]. Moreover, a generative adversarial network (GAN) based post-filtering was proposed to compensate for the differences between natural speech and synthetic speech in SPSS [21]. The performance of these methods is still constrained by the framework of two step (feature extraction and acoustic modeling) optimization and phase information is lost by a mel-cepstrum vocoder. Therefore, some researchers tried to model speech waveforms using neural networks. Oord et al. [22] proposed WaveNet, a deep convolutional neural network for generating raw audio waveforms. The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones. Kalchbrenner [23] proposed WaveRNN to increase the efficiency of audio sampling from sequential models with Recurrent Neural Networks. WaveRNN and WaveNet are also adopted as neural vocoder that generates raw waveform samples from intermediate representations [24] [25]. There are also some research results in end-to-end speech synthesis, including Char2Wav [24],

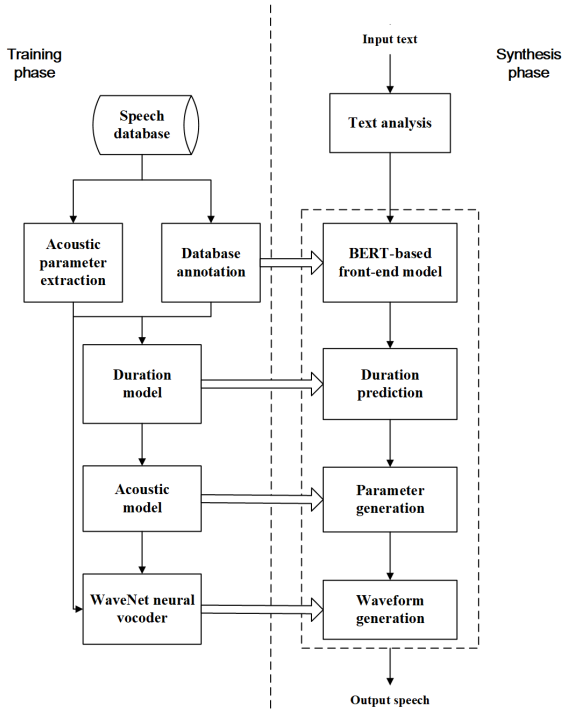

Figure 1: *Flowchart of the Our parametric system.*

Tacotron [26] and Tacotron 2 [27], which synthesis speech directly from characters based on sequenceto- sequence [28] with attention paradigm [29]. In order to further advance the state of SPSS, we built a parametric system from the following 4 points: (1) BERT-based front-end prediction, (2) duration modeling with an autoregressive model structure, (3) GAN-based multi-task acoustic modeling, (4) WaveNet-based neural vocoder to generate raw waveform samples from intermediate acoustic features. Finally, evaluation results showed the effectiveness of the proposed system.

The rest of this paper is organized as follows: Section 2 presents the methods used in our system. Section 3 describes system building. Section 4 shows the evaluation results. Conclusion is given in the end.

2. Framework

In this section, we will briefly introduce the framework of our proposed parametric system. As indicated in Figure 1, our SPSS system consists of two parts, the training phase and the synthesis phase.

We followed this flowchart and constructed our submitted system this year. A detailed description of the training and synthesis procedure will be presented as follows.

2.1. Training phase

At the training phase, manual annotations were performed at first, including Pinyin(with tone), prosodic word boundary, prosodic phrase boundary and focus position will be checked manually. in advance. These annotations were used for BERT-based prediction models training at front-end module in synthesis phase. Frame-level acoustic features were extracted, including mel-cepstrum, F0s and voice/unvoiced (U/V) information. An HMM alignment was conducted to obtain phoneme boundaries. Then, we applied LSTM-RNN models to duration and acoustic modeling.

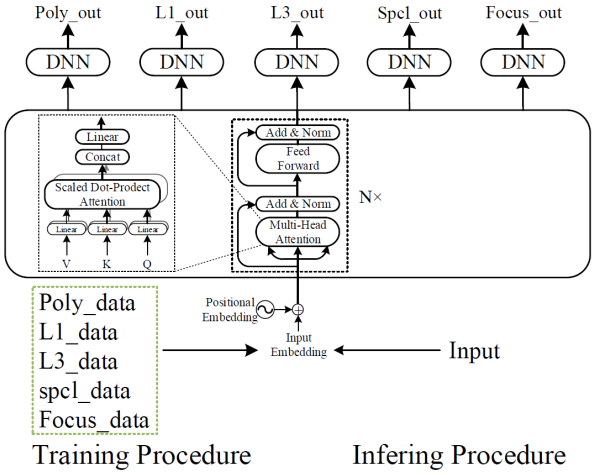

Figure 2: *BERT-based Front-end model.*

2.2. Synthesis phase

There are three major steps in synthesis stage. In the baseline system, expressive linguistic features were extracted from input text via text analysis by BERT-based front-end module, then fed into duration prediction and parameter generation module. WaveNet based neural vocoder took the post-processed acoustic feature as condition and generated speech waveforms sample by sample.

3. System Building

3.1. BERT-based front-end model

In the TTS synthesis system, the main front-end procedure includes text processing, grapheme-to-phoneme (G2P) conversion and prosody prediction from text. The target of the text processing is usually special marks conversion, such as converting Arabic numerals to values or strings. The target of the G2P conversation in Chinese TTS system is to convert Chinese character to pinyin. Most of the time, the character and pinyin are one-to-one mapping. Yet there are more than 900 polyphones in Chinese characters, which means one single character could have several different pronunciations according to usages and meanings in sentences. The target of prosody prediction is to break a sentence into phrases and focus on words. The strength of the break is usually classified as Prosodic Word (L1) or Prosodic Phrase (L3), which means a minor break with 1-4 characters or a major break with 5-7 characters respectively. The focus words are usually emphasized words. This occurs when a speaker wants to draw attention to particular words. In brief, the front-end tasks of the Chinese TTS synthesis system are special marks procession, polyphones classification, breaks prediction and focuses prediction. We use multi-task model based on BERT (Bidirectional Encoders Representations from Transformers)[30] as the front-end model in TTS synthesis system. Figure 2 illustrates the framework of BERT-based multi-task model.

In the pre-training procedure, firstly, we used Chinese unsupervised corpus in novel and news fields to pre-train BERT-base model. The model size of BERT-base is 12 transformer

encoder blocks with hidden size as 512 and self-attention heads as 8. We adopt masked language model (MLM) as the pre-training task to train a deep bidirectional representation. We randomly mask 15% of the input characters and then predict the masked characters. In the fine-tuning procedure, we used annotation data to model the front-end tasks mentioned above, such as Polyphones classification, L1 prediction, L3 prediction, special marks procession and focuses prediction. We use multi-task learning model for the front-end tasks. More specifically, the five tasks share the BERT language model (LM) as an encoder, and use DNN model as a decoder respectively. The decoder maps the character representations of BERT LM output to target spaces. We use mini-batch based stochastic gradient descent (SGD) and cross entropy loss to learn the parameters. In each epoch, the mini-batch is selected from the mixed annotation data from different tasks, with an index to distinguish data sources. The BERT LM encoder is updated for all the tasks, and the DNN decoder are updated according to the task-specific objective for each task. In the inferring procedure, we use sentences on L4 level as input data. For every character of different tokens such as Chinese, English and symbol, there will be five kinds of output, Poly-out, L1-out, L3-out, Spcl-out and Focus-out. With a softmax classifier, we get the labels of max probability. According to the Poly-out labels, the pinyin of the Chinese polyphones is predicted. According to the L1-out and L3-out labels, the system decides whether there is a L1 break or L3 break. According to the spcl-out label, the system converts the special marks such as Arabic numerals to values or strings. According to the Focus-out label, the system decides whether there is an emphasized character.

3.2. Duration modeling

Speech duration modeling is critical for the expressiveness of SPSS. Speech duration is an important part of speech prosody, and it will be used as input for the following acoustic modeling, further affecting the intonation of speech. An HMM-based alignment was performed to get the phoneme durations for duration modeling. To better fit the expressive corpus, we built an autoregressive model for quantified speech durations. The framework of the model is shown in figure 3. The duration model followed an autoregressive fashion. Since there is little correlation in speech duration values, the context and duration values were concatenated as the auto-regressive input. The autoregressive input was fed to a causal encoder and then the output was shifted as the previous embedding input for the decoder. On the other side, the context was fed to a context encoder to get the context embedding. The previous embedding and context embedding are sent to the decoder to predict the duration of current phoneme. The context encoder was composed of convolution banks and bi-directional LSTM and the causal encoder consisted of convolution banks with causal convolutions and LSTM. The decoder was composed of two LSTMs.

The context input included phoneme identifications, tones of the phonemes, and other expressive linguistic features, such as ToBI annotations and stress flags. We quantified the duration in the logarithm domain and predicted it with the softmax layer in the model. The model parameters were estimated using cross entropy criteria. The model was first pre-trained on a large multi-speaker corpus and then fine-tuned on the target dataset.

3.3. Acoustic modeling

Fundamental frequency (F0), 41 dimensional mel-cepstra (MCEP), band aperiodicity (BAP) were adopted as the acoustic

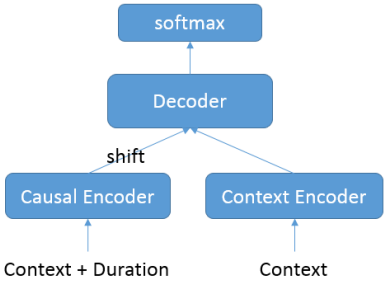

Figure 3: *Framework of the autoregressive duration model.*

features. An LSTM-RNN model was used to model the acoustic features. Conventional LSTM-RNN based acoustic model suffers from the over-smoothing problem due to the MSE training criteria, leading to degraded speech quality. To alleviate this problem, we adopted generative adversarial network (GAN) as a regularizer for the LSTM-RNN acoustic model. The framework of the acoustic model is shown in Figure 4.

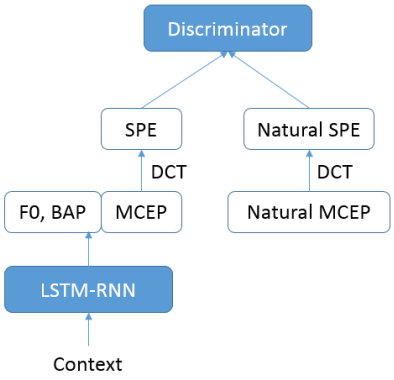

Figure 4: *Framework of the acoustic model.*

The input context feature was fed to the LSTM-RNN (the generator) acoustic model to generate F0, BAP, and MCEP. The MCEP was then transformed to the Mel spectral envelope (SPE) space by discrete cosine transform (DCT). The natural MCEP was also transformed to SPE by DCT. A discriminator was used to discriminate the generated SPE and natural SPE. The discriminator was trained to best discriminate the two inputs, i.e. the discriminator loss of GAN. The generator was trained in a multi-task fashion, composed of the traditional MSE loss and the generator loss of GAN. Least square GAN (LSGAN) was chosen as the criteria for GAN training.

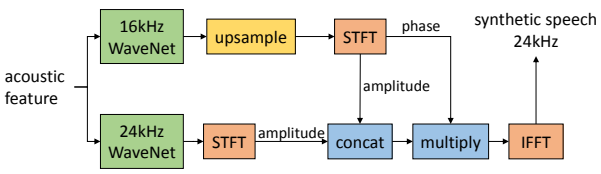

Figure 5: *The flowchart of waveform generation by WaveNets.*

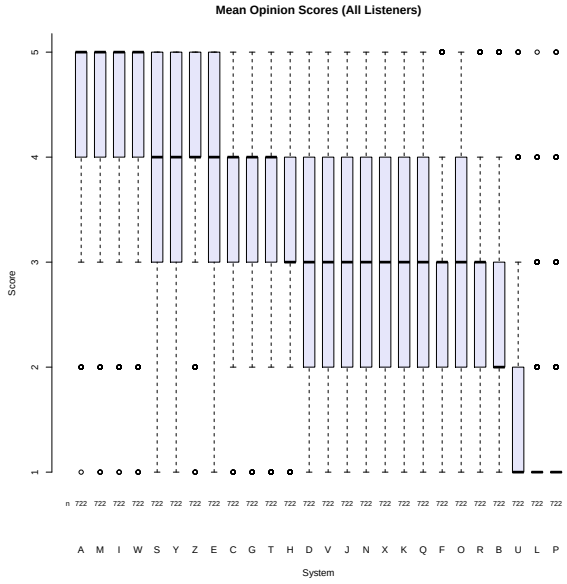

Figure 6: *Boxplot of naturalness scores of each submitted system for all listeners.*

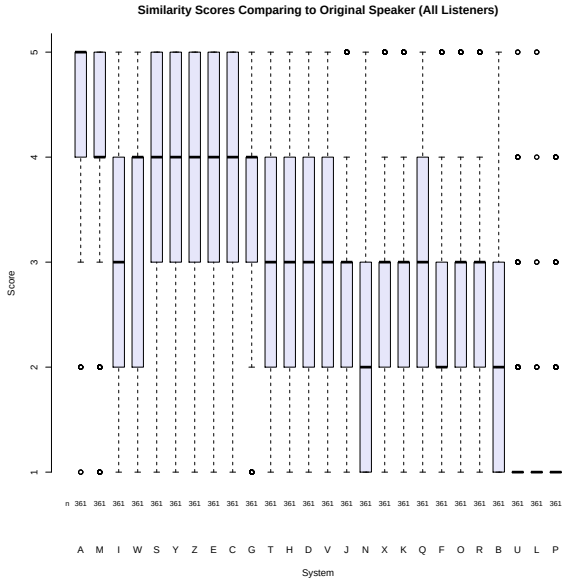

Figure 7: *Boxplot of similarity scores of each submitted system for all listeners.*

3.4. WaveNet-based neural vocoder

In the generation phase of the conventional vocoder-based speech synthesis system, the quality of synthesized speeches is degraded due to two major factors. They are the lack of phase prediction and the artifacts caused by vocoder synthesizer respectively. In order to address these two problems, WaveNet based neural vocoder is proposed for waveform generation instead, which greatly improves the quality of synthetic speech. WaveNet is a neural autoregressive generative model that models waveform directly. The WaveNet based vocoder is realized by the use of conditional WaveNet, in which acoustic feature is set as the conditional input to guide waveform generation. Given a sequence of waveform $\boldsymbol{X} = \{x_1, x_2, ..., x_t\}$, the joint probability of all these samples is represented as follows:

$$p(\boldsymbol{x}|\boldsymbol{h};\boldsymbol{\theta}) = \prod_{t=1}^{T} p(x_t|x_1, x_2, ..., x_{t-1}, \boldsymbol{h};\boldsymbol{\theta}) \tag{1}$$

Where $\boldsymbol{h}$ is the acoustic feature vector, $\boldsymbol{\theta}$ is the parameter set of this model. $p(x_t|x_1, x_2, ..., x_{t-1}, \boldsymbol{h};\boldsymbol{\theta})$ denotes the long range relationship among waveform samples. In WaveNet, it is modelled with the use of a stack of dilated causal convolutional layers. In our system, we adopted this WaveNet based neural vocoder for waveform generation. The acoustic feature used was the joint feature vector of Mel-cepstrum, F0 and the u/v decision. Besides, we made three improvements to the usage of the basic WaveNet model in order to enhance synthetic speech quality. Firstly, we modelled the samples with a single variance-bounded Gaussian distribution introduced in ClariNet[31], which could relieve the quantization noise in synthetic speeches brought by previous categorical distribution. The mean μ_t and variance σ_t of each audio sample distribution are predicted by model conditioned on the samples at all previous time-steps and the current acoustic feature:

$$\begin{aligned} \mu_t, \sigma_t &= WaveNet(x_t|x_1, x_2, ..., x_{t-1}; \boldsymbol{h}) \\ p(x_t|x_1, x_2, ..., x_{t-1}; \boldsymbol{h}) &= N(\mu_t, \sigma_t) \end{aligned} \tag{2}$$

Secondly, the model was trained with an initialization of a pre-trained multi-speaker model in order to improve training stability. The training process of this multi-speaker model and the model adaptation process were performed similar to those in [32]. Finally, as we found that 16kHz WaveNet could generate better low frequency harmonic than 24kHz WaveNet, we trained two WaveNets with different sample rates(16kHz and 24kHz) and generated the final 24kHz waveforms by combining the generated waveforms of these two. The generation procedure was shown in Figure 5. Specifically, a piecewise linear combination was used to obtain the concatenated amplitude spectra.

As the WaveNet architecture in [33], we adopted 24 dilated convolution layers, grouped into 4 dilation cycles, i.e., the dilation rate of layer $k(k = 0, 1, ..., 23)$ is $2^{k(mod6)}$. The filter width is 2 for 16kHz WaveNet, and 3 for 24kHz WaveNet. The lower bound variance is -10(in log scale). The model is optimized with Adam algorithm.

4. Evaluations

In this section, we will present the official evaluation results of our system. Our system identifier is M. There are 25 systems, including 1 benchmarks and 24 submitted systems, plus the natural speech were evaluated. System Z is an unit selection system [16] for comparison. The identifiers for the benchmark systems and our system are:

- A: Natural speech
- B: Benchmark merlin
- M: Our system
- Z: IIM-USTC system

4.1. Naturalness test

Figure 6 shows the boxplot of mean opinion scores(MOS) of each system on naturalness. Our system achieved MOS of 4.5,

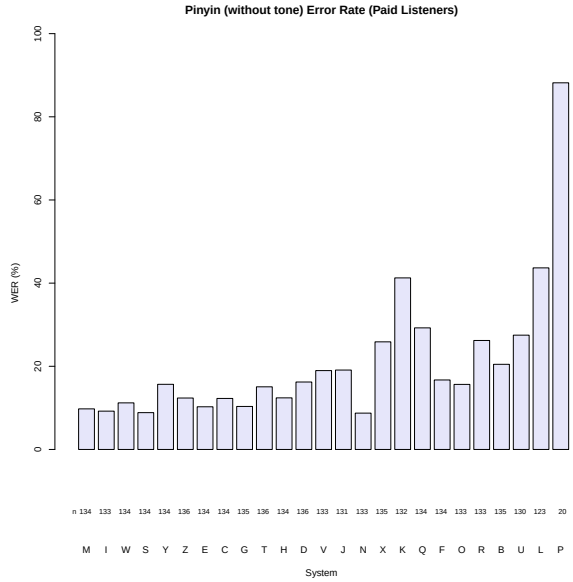

Figure 8: *Pinyin (without tone) Error Rate (Paid Listeners).*

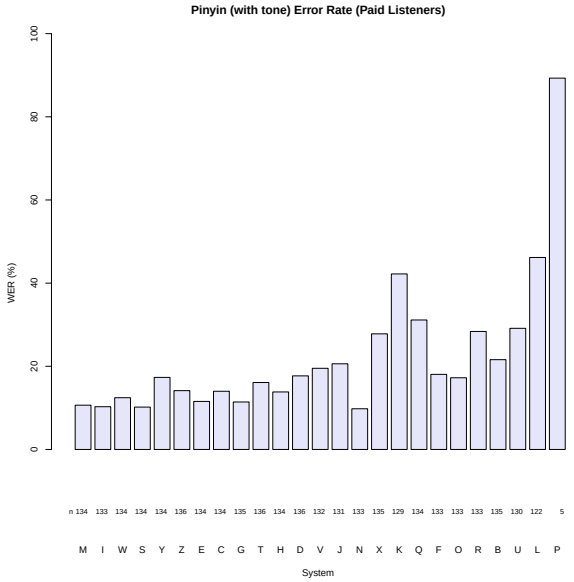

Figure 9: *Pinyin (with tone) Error Rate (Paid Listeners).*

The results indicate that our system outperforms all the other participants on naturalness. Besides, Wilcoxon signed rank tests show that the difference between our system and any other participant system on naturalness is significant. The score of 4.5 is very high in history of Blizzard Challenge naturalness test.

4.2. Similarity test

Figure 7 presents the boxplot MOS of each submitted system on similarity. Our system M achieved a mean opinion similarity score of 4.1. The socre is highest in all submitted systems, but the difference is not significant between our system and system S and Z. As we know, system Z is the same unit selection system as last year's best system on similarity. This shows that our proposed parametric waveform modeling system performs as well as unit selection system in similarity comparing. This may be related to the poor quality of the audio copus.

4.3. Intelligibility test

As shown in Figure 8 and 9, the Pinyin error rate (PER) of our system is 9.8%and the Pinyin with tone error rate (PTER) is 10.7%. The lowest PER score is 8.7% of system N. but Wilcoxon signed rank tests indicate that the difference is not significant compared our system to system N.

5. Conclusions

This paper presented the details of building the USTC system for the evaluation of Blizzard Challenge 2019. We built a parametric system that modeling speech waveforms. a BERT based models were used in our system for front-end text processing. An autoregressive LSTM model was used for duration modeling, and an LSTM-RNN model was used to model the acoustic features. Then, we adopted a generative adversarial network (GAN) as a regularizer for the LSTM-RNN acoustic model, to relieve the over-smoothing in acoustic modeling. In order to break the constraint of traditional mel-cepstrum vocoder, a WaveNet based neural vocoder was utilized to model speech waveforms from acoustic feature. The effectiveness of our system is verified by official evaluation results. Our system achieved an extremely well performance and surpassed system Z which used unit selection method.

6. References

[1] Y.-P. W. L. Q. R.-H. W. Zhen-Hua Ling, Yi-Jian Wu, "USTC system for blizzard challenge 2006 an improved hmm-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[2] 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 *et al.*, "The USTC and I-FLYTEK speech synthesis systems for Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007.

[3] Z.-H. Ling, H. Lu, G.-P. Hu, L.-R. Dai, and R.-H. Wang, "The USTC systems for Blizzard Challenge 2008," in *Blizzard Challenge Workshop*, 2008.

[4] H. Lu, Z.-H. Ling, M. Lei, C.-C. Wang, H.-H. Zhao, L.-H. Chen, Y. Hu, L.-R. Dai, and R.-H. Wang, "The USTC systems for Blizzard Challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[5] J. Yuan, Z.-H. Ling, M. Lei, C.-C. Wang, h. lu, Y. Hu, L.-R. Dai, and R.-H. Wang, "The USTC systems for Blizzard Challenge 2010," in *Blizzard Challenge Workshop*, 2010.

[6] L.-H. Chen, C.-y. Yang, Z.-H. Ling, Y. Jiang, L.-R. Dai, Y. Hu, and R.-H. Wang, "The USTC systems for Blizzard Challenge 2011," in *Blizzard Challenge Workshop*, 2011.

[7] Z.-H. Ling, X.-j. Xia, Y. Song, C.-y. Yang, L.-H. Chen, and L.-R. Dai, "The USTC systems for Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, 2012.

[8] L.-H. Chen, Z.-H. Ling, J. Yuan, Y. Song, X.-j. Xia, Y.-q. Zu, R.-q. Yan, and L.-R. Dai, "The USTC systems for Blizzard Challenge 2013," in *Blizzard Challenge Workshop*, 2013.

[9] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *International Speech Communication Association*, 1998, pp. 77–80.

[10] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, "Dnn-based stochastic postfilter for hmm-based speech synthesis," in *Proc. INTERSPEECH*, 2014, pp. 1954–1958.

[11] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *Proc. ICASSP*, 2014, pp. 290–294.

[12] L.-H. Chen, Z.-H. Ling, Y.-q. Zu, R.-q. Yan, Y. Jiang, X.-j. Xia, and Y. Wang, "The USTC systems for Blizzard Challenge 2014," in *Blizzard Challenge Workshop*, 2014.

[13] L.-H. Chen, Z.-H. Ling, X.-j. Xia, Y. Jiang, Y.-q. Zu, and R.-q. Yan, "The USTC systems for Blizzard Challenge 2015," in *Blizzard Challenge Workshop*, 2015.

[14] L.-H. Chen, Y. Jiang, M. Zhou, Z.-H. Ling, and L.-R. Dai, "The USTC systems for Blizzard Challenge 2016," in *Blizzard Challenge Workshop*, 2016.

[15] L.-J. Liu, C. Ding, Y. Jiang, M. Zhou, and S. Wei, "The IFLYTEK system for Blizzard Challenge 2017," in *Blizzard Challenge Workshop*, 2017.

[16] Y. Jiang, X. Zhou, C. Ding, Y.-j. Hu, Z.-H. Ling, and L.-R. Dai, "The USTC system for Blizzard Challenge 2018," in *Blizzard Challenge Workshop*, 2018.

[17] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[18] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[19] 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 *Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 98–102.

[20] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for lowlatency speech synthesis," in *Proc. ICASSP*, 2014, pp. 290–294.

[21] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4910–4914.

[22] A. van den 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," *CoRR abs/1609.03499*, 2016.

[23] 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.

[24] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[25] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *Interspeech*, 2017, pp. 1118–1122.

[26] 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: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[27] 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.

[28] 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.

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

[30] D. J, C. M. W, and e. a. Lee K, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[32] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "Wavenet vocoder with limited training data for voice conversion." in *Interspeech*, 2018, pp. 1983–1987.

[33] 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.