The USTC System for Blizzard Challenge 2017

Ya-Jun Hu[1], *Chuang Ding*[2], *Li-Juan Liu*[2], *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

hyjl5475@mail.ustc.edu.cn

Abstract

This paper introduces the details of the speech synthesis system developed by the USTC team for Blizzard Challenge 2017. A 6.5-hour corpus of highly expressive children's audiobook was released to the participants this year. A parametric system that modeling speech waveforms was built for the task. Firstly, long short term memory (LSTM)-based recurrent neural networks (RNN) were adopted for the baseline system, including tone and breaking indices (ToBI) prediction, duration modeling and acoustic modeling. Then, we proposed a generative adversarial network (GAN) based post-filtering to relieve the over-smoothing in acoustic modeling and compensate for the differences between natural and synthetic spectrum in the baseline system. At last, a WaveNet based neural vocoder was utilized to model speech waveforms from acoustic feature instead of mel-cepstrum vocoder. The evaluation results show the effectiveness of the submitted system.

Index Terms: statistical parametric speech synthesis, LSTM-RNN, GAN, WaveNet

1. Introduction

The USTC team have been submitting entries to Blizzard Challenge speech synthesis evaluation for twelve years since 2006. In the first participation, we submitted an 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 is growing, a globally covariance tying strategy was utilized to reduce the footprint of the model, as well as improving the modeling 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 were released in Blizzard Challenge 2013. The scale of the corpus was a challenge to both the computational 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 an 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 letter-to-sound (L2S) [9] method to build frontend text processing for Hindi, and used 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 year, a 5-hour 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 (RNN) were adopted for tone and breaking indices (ToBI) prediction to achieve high expressiveness. And another LSTM-RNN was adopted to extract distributional representation of contextual features, which is used to evaluate contextual similarities between candidate and target units at the unit selection time [14]. This year, about 6.5 hours of British English speech data from a single female talker was released, which comprises speech data already released for the 2016 challenge.

Unit selection systems always achieve excellent performance in the Blizzard Challenge every year. Due to the over-smoothing in acoustic modeling and the restriction of vocoder, SPSS system performs not good enough in voice quality and similarity [15]. 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 [16]. LSTM-RNN has achieved great performance in both the front-end text processing [17] and back-end acoustic modeling [18]. Moreover, a generative adversarial network (GAN) based post-filtering was proposed to compensate for the differences between natural speech and synthetic speech in SPSS [19]. 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. Tokuda et al. [20] [21] attempted to model raw speech waveforms using the neural network-based SPSS framework with a specially designed output layer. 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. Mehri et al. [23] proposed SampleRNN for unconditional audio

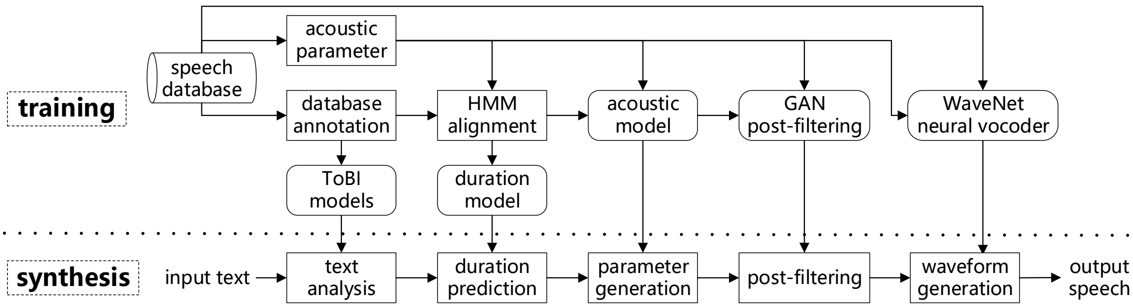

Figure 1: *The flowchart of USTC parametric system.*

generation based on generating one audio sample at a time. The model combines autoregressive multilayer perceptrons and stateful RNN in a hierarchical structure and models extremely long-term dependencies in audio signals. SampleRNN 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] and Tacotron [26], which synthesis speech directly from characters based on sequence-to-sequence [27] with attention paradigm [28].

In order to further advance the state of SPSS, we built a parametric system from the following 3 points: (1) LSTM-RNN based expressive ToBI prediction, duration and acoustic model for baseline system, (2) GAN based post-filtering to relieve over-smoothing in acoustic modeling, (3) WaveNet based neural vocoder to generate raw waveform samples from intermediate acoustic features. Internal experiments and evaluation results showed the effectiveness of the proposed system.

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.

2.1. Training phase

At the training phase, ToBI annotations were performed manually in advance. These annotations were used for LSTM-RNN based ToBI prediction models training to enable personalized ToBI tags. Expressive labels, such as dialogue tags and sentence types were used in our contextual information. Frame-level acoustic features were extracted, including mel-cepstrum, F0s and voice/unvoiced (U/V) information. An HMM alignment was conducted to obtain 5 states and phoneme boundaries. Then, we applied LSTM-RNN models to duration and acoustic modeling.

To relieve over-smoothing in acoustic modeling, we utilized a GAN based post-filtering composed of a generator and a discriminator. The learned post-filtering was trained with synthesized mel-cepstrum as condition and enabled the generator to generate natural spectral texture. Then a WaveNet based neural vocoder was trained to learn the predictive distribution for each audio sample conditioned on all previous ones.

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, then fed into duration prediction and parameter generation module. The synthesized acoustic feature worked as conditional parameter in GAN based post-filtering. WaveNet based neural vocoder took the post-processed acoustic feature as condition and generated speech waveforms sample by sample.

3. System Building

3.1. ToBI prediction

ToBI tags are important for prosody modeling of standard English, especially for a expressive speech corpus. After all ToBI information were annotated, three LSTM-RNNs were trained separately for accent prediction, phrase boundary prediction and boundary tone prediction. We have built these ToBI models using a large corpus last year, so the new personalized ToBI models were trained using last year's model as initial model.

A set of linguistic features were extracted for prediction. The input feature for accent prediction included word feature, part-of-speech (POS) tag, position of current word in the sentence, number of phonemes in current word, number of stresses in current word, word frequency and word case style. We adopted a binary classification layer with cross-entropy criteria to predict the probability of the current word being accented. The input features for phrase boundary are the same as the ones for accent prediction except the absolute position of current word in sentence. The probability of three classes were predicted: beginning, intermediate and end of a phrase boundary. All the input features for accent predicting are used in the input feature for boundary tone prediction, except an additional feature that indicates whether the current word is the end of a phrase or sentence. All words are categorized with 6-class, including the beginning, intermediate, end of a phrase with L-L tone and beginning, intermediate, end of a phrase with L-H tone.

3.2. Duration modeling

Speech sound duration is an important component in the prosody of synthetic speech, especially in generating expressive and conversational speech for audiobooks. Expressive linguistic features including abundant ToBI annotations and dialogue and sentence types were used to improve duration prediction.

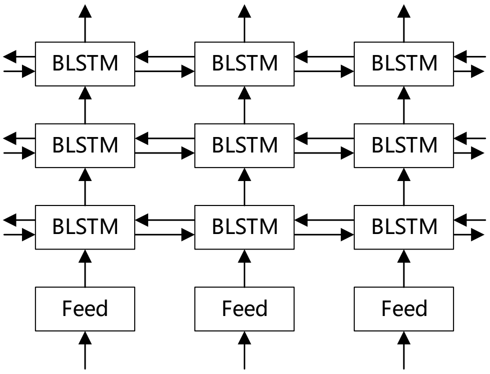

Figure 2: *The structure of the LSTM-RNN for acoustic modeling.*

40 dimensions mel-cepstrum, F0 and respective delta features were extracted from 16 kHz speech data. An HMM based alignment was performed to generate 5 states and phoneme boundaries. In duration modeling, we applied bidirectional LSTM-RNN with a 6-dimensions linear output layer to predict 5 states and phoneme durations. The input and output features were normalized separately and the model was trained with mean square error (MSE) criteria.

3.3. Acoustic modeling

The major factor that degrades the naturalness of the synthetic speech from SPSS is the accuracy of acoustic modeling. In the conventional HMM-based acoustic modeling, the Gaussian distributions are estimated by averaging all observations associated with a given decision tree leaf node. Although this averaging process improves the robustness of parameter estimation and generation, the detailed characteristics of the speech parameters are often lost. Therefore, the reconstructed spectral envelopes are typically over-smoothed, which leads to the muffled voice quality of the synthetic speech. LSTM-RNN, which is designed to model temporal sequences and their long-term dependencies, has been successfully applied to acoustic modeling for SPSS and has shown the potential to produce more natural synthetic speech. As our baseline system, we adopted a bidirectional LSTM-RNN, which can access input features at both past and future frames.

Speech with higher frequency always sounds more pleasant. However, in the WaveNet based neural vocoder, it usually takes a week to train and higher frequency needs longer training time. Moreover, a robust and effective neural vocoder requires more training data. As the amount of training data is limited, to balance the quality of synthetic speech and training time, we adopted 128 dimensions mel-cepstrum extracted from 22k speech wav using STRAIGHT, with 1-dim energy, 1-dim F0, 1-dim U/V decision and 5-dim aperiodicity ratio. Since the corpus for the challenge this year is highly expressive, the conventional context feature including quinphone and ToBI information, is insufficient for prosody modeling. To enrich the context feature, we added the dialogue embedding and sentence type into the input of the LSTM-RNN. Dialogue embedding indicates whether the current phoneme is in a dialogue in the story. Sentence type was obtained from the raw text according to its punctuation.

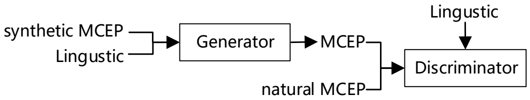

Figure 3: *The framework of GAN based post-filtering.*

The structure of the LSTM-RNN is shown in Figure 2. There are four hidden layers in this architecture stacked by a feedforward layer and three bidirectional LSTM-RNN layers with 1024 nodes in each layer. The network was trained using Stochastic Gradient Descent (SGD) algorithm. The training stopped if no new best error on the validation set could be achieved within the last 20 epochs.

3.4. GAN-based post-filtering

Conventional deep learning based SPSS methods generally adopt minimum MSE as the criteria for acoustic modeling, which could guarantee the robustness of generated acoustic features. However the generated acoustic features are over-smoothed and the generated speech sounds muffled due to the MMSE criteria. Recently, GAN [29] has been proposed as described in previous sections. As an alternative training method for generative model, GAN is a promising method to relieve the over-smoothing effect of acoustic models trained by MMSE criteria and improve the segmental quality of synthesized speech.

A straightforward application to utilize GANs in SPSS systems is to predict acoustic features given linguistic features directly by a GAN. However, preliminary experiments showed that there were some artifacts in the generated spectrum, which makes the synthetic speech sound uncomfortable.

In this section, a method to utilize GANs in SPSS system and walk around the problem of robustness was introduced. We proposed a GAN-based post-filter method, in which linguistic feature and synthetic mel-cepstrum were sent to the generator (G) of GAN as condition to predict the natural mel-cepstrum. The discriminator (D) in GAN tried to discriminate the natural mel-cepstrum and the generated ones by G conditioned on linguistic features. The framework of the model is shown in figure 3.

During test, synthetic mel-cepstrum predicted by baseline system was sent to the GAN along with linguistic feature as conditions for GAN prediction, which makes this method a GAN-based post-filter. Several kinds of GANs have been tried in preliminary experiments, including conventional GAN [29], wasserstein GAN (wGAN) [30] and least square GAN (LSGAN) [31], and LSGAN is adopted in our system for its ability of fast convergence.

3.5. WaveNet-based neural vocoder

In the generation phase of the present vocoder-based speech synthesis system, the quality of synthesized speeches are 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, we proposed a WaveNet based neural vocoder for waveform generation

Table 1: *Systems compared in the subjective text.*

system	description
LSTM-RNN	The LSTM-RNN based baseline system
GAN	GAN based postfilter on baseline system
WaveNet	WaveNet neural vocoder on GAN generation

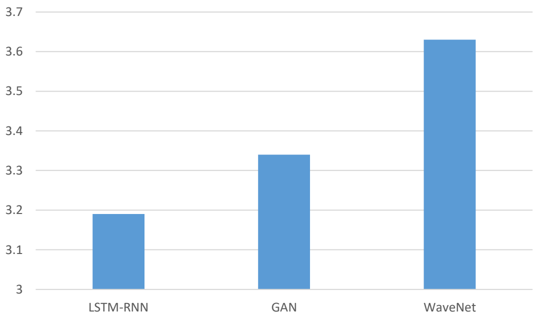

Figure 4: *Mean opinion score on naturalness of three compared systems.*

instead.

WaveNet is a neural autoregressive generative model that can generate waveforms directly. Given a waveform $\boldsymbol{x} = \{x_0, x_1, ..., x_{T-1}\}$, the joint probability of all these samples is represented as follows:

$$p(\boldsymbol{x}; \boldsymbol{\lambda}) = \prod_{t=0}^{T-1} p(x_t | x_0, x_1, ..., x_{t-1}). \tag{1}$$

As equation(1) indicates, each waveform sample is only related to the samples at all previous timesteps. So that WaveNet can generate waveforms sample by sample with no demands for speech processing related assumptions or manipulations. Thus no artifacts will be brought in.

In the framework of WaveNet, the conditional probability distribution $p(x_t|x_0, x_1, ..., x_{t-1})$ in equation(1) is modeled by a stack of convolutional layers. The output of each layer is:

$$\boldsymbol{z} = tanh(\boldsymbol{W}_{f,k} * \boldsymbol{y}) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{y}), \tag{2}$$

where $\boldsymbol{y}$, $\boldsymbol{z}$ are the input and output vectors, k denotes the layer index, f and g represent the filter and gate, respectively, $\boldsymbol{W}_{f,k}$ and $\boldsymbol{W}_{g,k}$ are trainable convolution filters, $*$ denotes a convolution operator, $\odot$ is an element-wise multiplication operator, $\sigma(\cdot)$ denotes a sigmoid function.

Additional input can be added to WaveNet to guide the waveform generation. Now given the additional input $\boldsymbol{h}$, the waveform is modeled by a conditional probability distribution$P(\boldsymbol{x}|\boldsymbol{h})$. The activation function from equation(2) becomes:

$$\boldsymbol{z} = tanh(\boldsymbol{W}_{f,k} * \boldsymbol{y} + \boldsymbol{V}_{f,k} * \boldsymbol{h}) \odot \sigma(\boldsymbol{W}_{g,k} * \boldsymbol{y} + \boldsymbol{V}_{g,k} * \boldsymbol{h}), \tag{3}$$

where $\boldsymbol{V}_{f,k}$ and $\boldsymbol{V}_{g,k}$ are learnable convolution filters. By conditioning WaveNet on linguistic features, the text-to-speech(TTS) system achieved state-of-the-art performance. It outperformed both LSTM-RNN based SPSS and the HMM-based unit selection.

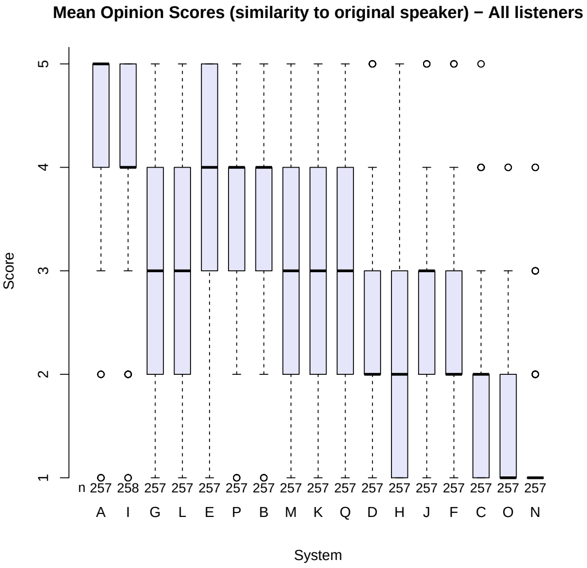

Figure 5: *Boxplot of similarity scores of each submitted system.*

Considering of the successful application of WaveNet in TTS, we try an alternative use of the conditional WaveNet in our system. By conditioning WaveNet on acoustic features, a WaveNet based neural vocoder is implemented. It can learn the relationship between acoustic features and waveform samples automatically. This neural vocoder is used to replace the conventional vocoder so that waveforms can be generated directly. The quality of synthetic speeches is supposed to be further improved. The acoustic features include mel-cepstrum, $F0$ and the U/V decision in our system. During the training stage, natural acoustic features extracted by STRAIGHT vocoder are used. While at synthesis stage, the predicted acoustic feature is fed to the input of WaveNet and waveforms are generated through this neural vocoder.

As the original 8bit quantization introduced quantization noise in synthetic speeches, we proposed to use a 10bit quantization scheme instead, in order to alleviate this problem. WaveNet with 3 blocks, which was 30 layers in total, was used in our system. The model was optimized with Adam algorithm.

3.6. Internal experiment

We conducted a listening test on the Amazon Mechanical Turk (ATM) crowd sourcing platform[1], to verify the performance of the proposed method. Table 1 presents the three systems that were compared in the test. Results in figure 4 proves the effectiveness of the proposed method. The WaveNet neural vocoder system was used to build our final submitted voices.

4. Evaluation

In this section, we will present the official evaluation results of our system. Our system identifier is G.

Figure 5 presents the boxplot of mean opinion scores (MOS) of each submitted system on similarity. Our system G achieved a similarity score of 3.1, which shows no statistical difference with system K, L, M, Q and ranks 5th in all

[1] https://www.mturk.com

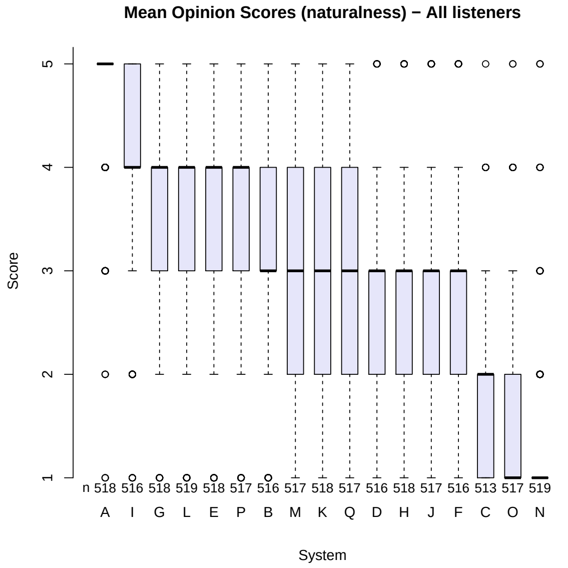

Figure 6: *Boxplot of naturalness scores of each submitted system.*

Table 2: *Mean opinion scores of paragraph test.*

	I	G	Best except G and I
Overall	38.1	32.8	31.6
Pleasantness	38.1	31.8	32.4
Speech Pauses	36	34	31
Stress	36	33	32
Intonation	37	33	32
Emotion	38	34	34
Listening effort	36.6	32.8	31.1

submitted systems. As we know, system B is the Festival unit selection benchmark system and system I is HMM-based unit selection system from IFLYTEK research. This shows that our proposed parametric waveform modeling system performs not good enough in similarity comparing with unit selection systems. The accuracy of modeling needs to be further improved.

Figure 6 shows the boxplot of MOS of each system on naturalness. Our system achieved MOS of 3.6, which shows no statistical difference with system E, L, P and ranks 2nd place in all submitted system. It's a considerable performance comparing with unit selection systems.

As shown in Figure 7, the word error rate (WER) of our system is 35% on the intelligibility test, ranking 5th in all submitted systems. We found that there were some U/V error in the synthesised speech of our system, which was probably one of factors that degrading the performance in intelligibility.

The scores of our system in the paragraph test are presented in Table 2. Our system ranks 2nd in overall impression and all other subjective metrics except pleasantness. Therefore, we compared our system G, the best system I and the highest score except system G and I. Figure 8 showed the comparison. The results indicated that the generation of wavenet based neural

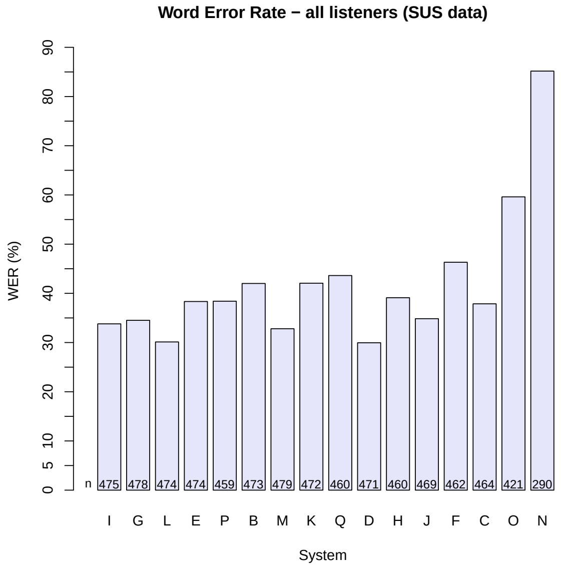

Figure 7: *Boxplot of WER of each submitted system.*

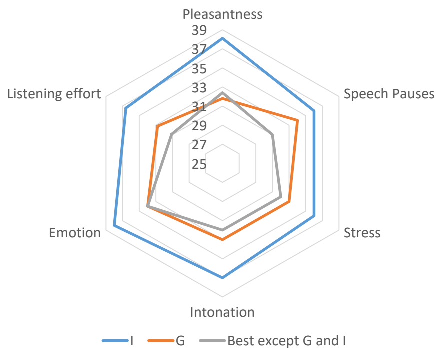

Figure 8: *Comparison between system I, G and best except G and I.*

vocoder performs slight weak in pleasantness.

5. Conclusions

This paper presented the details of building the USTC system for the evaluation of Blizzard Challenge 2017. We built a parametric system that modeling speech waveforms. The LSTM-RNN based models were used in our baseline system for front-end text processing, back-end duration modeling and acoustic modeling. Then, we adopted a GAN based post-filtering 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 both our internal experiments and official evaluation results. Our system achieved a considerable performance with unit selection system. The future work will be further investigating the WaveNet based neural vocoder to achieve a more stable and robust SPSS system.

6. References

[1] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "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 iflytek 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 system 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 system for blizzard challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[5] Y. Jiang, Z.-H. Ling, M. Lei, C.-C. Wang, L. Heng, Y. Hu, L.-R. Dai, and R.-H. Wang, "The ustc system 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 system 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 system for blizzard challenge 2012," in *Blizzard Challenge Workshop*, 2012.

[8] L.-H. Chen, Z.-H. Ling, Y. Jiang, Y. Song, X.-J. Xia, Y.-Q. Zu, R.-Q. Yan, and D. Li-Rong, "The ustc system 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," *International Speech Communication Association*, pp. 77–80, 1998.

[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 system 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 system 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 system for blizzard challenge 2016," in *Blizzard Challenge Workshop*, 2016.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4215–4219.

[21] ——, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Proc. ICASSP*, 2016, pp. 5640–5644.

[22] 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.

[23] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[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] T. Akira, H. Tomoki, K. Kazuhiro, T. Kazuya, and T. Tomoki, "Speaker-dependent wavenet vocoder." in *Proc. INTERSPEECH*, 2017.

[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] 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.

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

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *arXiv preprint ArXiv:1611.04076*, 2016.