The USTC System for Blizzard Challenge 2016

Ling-Hui Chen[†‡], *Yuan Jiang*[‡], *Ming Zhou*[‡], *Zhen-Hua Ling*[†], *Li-Rong Dai*[†],

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

chenlh@mail.ustc.edu.cn

Abstract

This paper introduces the details of the speech synthesis entry developed by the USTC team for Blizzard Challenge 2016. A 5-hour corpus of highly expressive children's audiobook was released this year to the participants. An hidden Markov model (HMM)-based unit selection system was built for the task. In addition, we utilized deep neural networks to improve the performance of our system, in both the front-end text processing and back-end acoustic modeling for unit selection. Firstly, an long short term memory (LSTM)-based recurrent neural networks (RNN) were adopted for tone and breaking indices (ToBI) prediction. Secondly, another LSTM-RNN was adopted to extract distributional representation of contextual features. The context embeddings can be used for evaluating contextual similarities between candidate and target units at the unit selection time. The evaluation results show the effectiveness of the submitted system. Our system achieved the highest scores in all metrics.

Index Terms: Unit selection, hidden Markov models, long short term memory, recurrent neural network

1. Introduction

The USTC team have been submitting entries to the Blizzard Challenge speech synthesis evaluation for eleven years since 2006. In 2006, our first HMM-based statistical parametric speech synthesis system using line spectral pairs (LSP) was submitted [1]. In the coming two years, in order to achieve better performance of our system, an HMM guided unit selection and waveform concatenation method was adopted to exploit the advantage of the large scale of the released corpus [2] [3]. The submitted hybrid system achieved promising performance. Next year 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]. A globally covariance tying strategy was utilized in order to reduce the footprint of the model, as well as improving the modeling training efficiency in 2010 [5], as the size of released corpus is growing. 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, an improved unit selection criterion, maximum log likelihood ration (LLR) criterion, was proposed [6] to improve the performance of uint selection. The evaluation tasks were becoming more and more difficult since 2012. Expressive corpus, such as audiobooks, and many Indian language were released for system building. In the Blizzard Challenge 2012, a set of audiobook corpus with different recording channels were release. A channel equalization method were utilized to compensate these channel differences [7]. A large corpus with hundreds of hours of unaligned audiobooks were release in Blizzard Challenge 2013. The scale of the corpus was a challenge to both the computational efficiency and robustness of the participants' system. We utilized the phone dependent model clustering method 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]. Besides, corpus of many Indian languages were released to non-Indian participants in Blizzard Challenge 2013, 2014 and 2015. We adopted letter-to-sound (L2S) [9] methods to build front-end text processing for Hindi, and used simple charactor 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].

This year, the challenge went back to English tasks. A highly expressive children's audiobook corpus were released to participants for system construction. The 5-hour speech corpus was relatively small for building robust unit selection system. In our system this year, we proposed 3 points to achieve this: 1) an LSTM-RNN based front-end was adopted to enable more expressive ToBI prediction, 2) expressive labels, such as dialogue tags and sentence types (obtained from punctuations) are used in our contextual information as expressive feature, 3) another LSTM-RNN based distributed contextual representation was adopted to provide a better metric for evaluating contextual differences between unit candidates. Internal experiments and evaluation results showed the effectiveness of the proposed system.

This paper is organized as follows: Section 2 reviews the baseline system of the USTC unit selection system. Section 3 presents the details of system construction in Blizzard Challenge 2016 will elaborated as well as some internal experimental results. Section 4 shows the evaluation results of our proposed system together with according further analysis. Lastly in section 5, some conclusions and potential future research are given.

2. Baseline Systems

2.1. HMM-based parametric method vs. unit selection

There are typically two kinds of approaches to build the baseline system: the HMM-based parametric speech synthesis system and the HMM-guided unit selection and waveform concate-

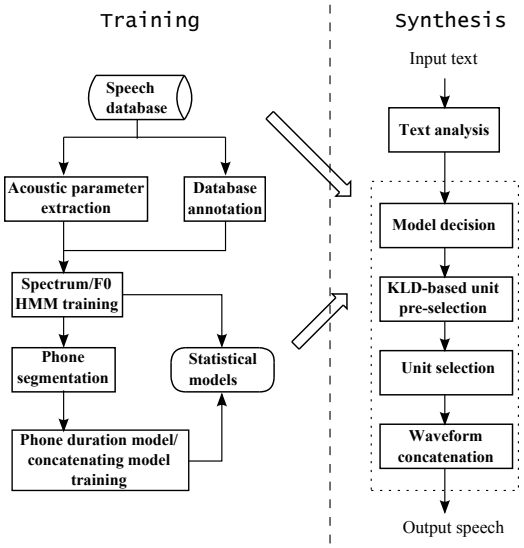

Figure 1: The flowchart of USTC unit selection system.

nation methods. The unit selection methods may achieve high quality of synthetic speech. However, their robustness of the method on a small size of highly expressive corpus is a challenge. The parametric methods, on the other hand, can produce stable speech. However, the speech quality and expressiveness of the synthetic speech was much poorer than unit selection ones. Therefore in the beginning, we conducted a listening test to compare these two different methods. A preference test between these two methods on the overall impression of synthetic speech was conducted on the Amazon Mechanical Turk (ATM) crowd sourcing platform [1]. 40 sentences were used for each system in the test. 10 subjects participated in the test. The results shown in Tabel 1 indicates that although there are several instability in synthetic speech of the unit selection method, it is significantly better than the parametric based one.

Table 1: Result of preference test comparing HMM-based parametric speech synthesis and unit selection method.

HMM	Unit Selection	N/P	p-value
6.94	88.34	4.72	< 0.001

2.2. The USTC unit selection system

In this section, we will briefly introduce the baseline system of our HMM-based unit selection system. As indicated in Figure 1, our HMM-based unit selection system consist of two parts, the training phase and the synthesis phase.

2.2.1. Training phase

At the training stage, several HMM [14] based acoustic models are trained in advance. These models are used to guide the unit selection at synthesis time. There are six sets of HMM based acoustic models in total, including a spectral model, an F0 model, a phone duration model, a concatenating spectral model, a concatenating F0 model, as well as a syllable-level F0 model.

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

The spectral model, F0 model and phone duration model are trained using the same methods as a conventiona HMM based parametric speech system [14]. Frame-level acoustic features are used for model training, including mel-cepstral coefficients and F0s. The duration of each phone unit is segmented by spectral model and F0 model using a viterbi based force align method.

Concatenation models are trained to model the distributions of acoustic difference in spectrum and F0 at the phone boundaries. The features for model training are the delta and delta-delta of spectrum and F0. In addition, a syllable-leval F0 model, which is trained with F0 features extracted from the vowels of two adjacent syllables, are used to capture the long term prosody dependence in F0.

The multi-space distribution HMMs (MSD-HMMs) [15] are adopted to model the continuous probability HMMs with F0 feature. The decision tree based context clustering is adopted to deal with the data sparsity problems. The minimum description length (MDL) [16] based model clustering is utilized to control the size of the decision trees. The phone durations, concatenating spectral features, concatenating F0 features and syllable-level F0 features are extracted using state-frame alignment information.

2.2.2. Synthesis phase

There are two steps in the synthesis phase: unit selection and waveform concatenation. A sequence of phone units are selected under the Maximum Likelihood criterion for the input sentence to be synthesized. Let N be the number of phonemes in the utterance to be synthesized with context feature sequence C. At the unit selection stage, a sequence of phone unit candidates $\boldsymbol{U} = \{u_1, u_2, \cdots, u_N\}$ are search out from the database under the following statistical criterion

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{6} w_m [\log P(\boldsymbol{X}(\boldsymbol{U}, m)|C, \lambda_m) - w_{KLD} D_m(C(\boldsymbol{U}), C)], \quad (1)$$

where λ_m indicates the acoustic models described in the previous section, and w_m corresponds to their weights. The weights were manually tuned on a development set. $\boldsymbol{X}(\boldsymbol{U}, m)$ and $C(\boldsymbol{U})$ extract corresponding acoustic features and context features from a phone unit, $D_m()$ denotes the Kullback-Leibler divergence (KLD) [17] of corresponding acoustic model. A dynamic programming (DP) search algorithm is applied to find the optimal unit sequence, and a KLD-based unit pre-selection method is adopted to reduce the computational complexity in the DP based search.

Finally, in the concatenation step, the waveforms of every two consecutive candidate units in the optimal unit sequence are concatenated to produce the synthetic speech. The cross-fade technique [18] is used here to smooth the phase discontinuity at the concatenation points of unit boundaries.

3. System Building

3.1. LSTM-based recurrent neural network

In recent years, the long short term memory based recurrent neural network has been successfully applied to many tasks, such as acoustic modeling in speech recognition [19] and speech synthesis [20], natural language understanding [21], because of its powerful ability in modeling sequential features.

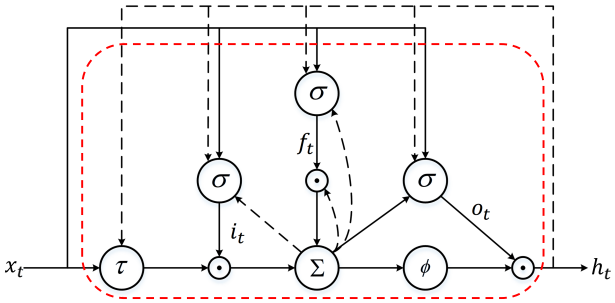

Figure 2: The structure of a long short term memory cell.

Fig. 2 shows the structure of a LSTM cell, which is included in the red block. x_t is the input of the cell while h_t is the output of the cell. Dashed lines represent the recurrent connection from the output of the cell at the previous time step. There are three gates in the structure: the input gate i_t, forget gate f_t and output gate o_t. The input gate enables the memory cell to memorize or ignore the information in the input of current time step. The forget gate, on the other hand, enables the cell to memorize or clear the information in the previous inputs of the sequence. The output gate controls the information output flow of current cell. With these gates, an LSTM cell has the ability to model the sequential characteristics of the input frames from the beginning to current time step.

The formulas of the LSTM cell is given by:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \quad (2)$$
$$f_t = \sigma(W_{fi}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \quad (3)$$
$$a_t = \tau(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (4)$$
$$c_t = f_t c_{t-1} + i_t a_t, \quad (5)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \quad (6)$$
$$h_t = o_t \phi(c_t), \quad (7)$$

where W_* and b_* are the parameters of the cell, $\sigma(\cdot)$, $\tau(\cdot)$ and $\phi(\cdot)$ are activation functions.

3.2. ToBI prediction

Firstly, we used LSTM-RNNs for ToBI prediction. ToBI tags are important for prosody modeling of standard English [22]. Three LSTM-RNNs were used to predict the accent, phrase boundary and boundary tone separately from the text to be synthesized. ToBI prediction is a sequence labelling, to which LSTM-RNN has been successfully applied [23] [24]. The key point is to define the input and output of the model.

3.2.1. Accent prediction

The input feature for accent prediction includes 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. An one-hot vector was used as the word feature. Another one-hot vector was used as POS feature. 35 POS tags were used in our system. The position of word was normalized by the length of the sentence. Number of phonemes and number of stresses were given by the Lexicon. The word frequencies were obtained from an additional large scale corpus. Words with lower frequency tend to be accented more frequently. On the other hand, all words in our dictionary are in lower case. However, the case sensitivity of words may be important for ToBI predicting. Therefore, we additionally used four one-hot features for different word case style: in lower case, in upper case, first character in upper case and others. The output layer of the LSTM-RNN is a binary classification layer, predicting the probability of the current word being accented.

3.2.2. Phrase boundary prediction

The input feature for phrase boundary prediction includes the word, POS tag, word position and word case style. These features are the same as the ones described in section 3.2.1 except the word position feature. The absolute position of current word from the beginning and end of the sentence were used as the word position feature.

A softmax output layer was used for phrase boundary prediction. The probability of three classes were predicted: beginning, intermediate and end of a phrase boundary.

3.2.3. Boundary tone prediction

All the input features for accent predicting were used in the input feature for boundary tone prediction. In addition, the boundary tone should be predicted only for the word in the end of a phrase or sentence. Therefore, an additional feature indicating whether the current word is the end of a phrase or sentence. This feature is given by the phrase boundary prediction model.

A 6-class softmax output layer was used for phrase boundary tone predicting, including the beginning, intermediate, end of a phrase with L-L tone and beginning, intermediate, end of a phrase with L-H tone.

3.2.4. Internal experiment

In our internal experiment, a corpus of 49,700 sentences was used to build these three models. 44,730 sentences were used to train the models and the remaining 4,970 sentences were used to tune the parameters. The released training data of Blizzard Challenge 2016 were manually annotated for evaluation. There are 3753 sentences in the test set.

We compared the proposed LSTM-RNN based ToBI prediction method with a conventional decision tree (DT) based one [25]. Bi-directional LSTM-RNNs were used in this experiment. The results (F-scores) are shown in Table 2. It can be seen that the LSTM-RNNs significantly outperform the DT in all tasks.

Table 2: F-score of decision tree and LSTM-RNN on ToBI predicting.

	DT	LSTM-RNN
Accent	0.361	0.409
Phrase Boundary	0.486	0.504
Boundary Tone	0.302	0.402

3.3. Acoustic modeling

3.3.1. LSTM-RNN based context embedding

In the conventional HMM-based speech synthesis system, the context features are manually designed and they are mostly discrete [14]. This may cause several problems:

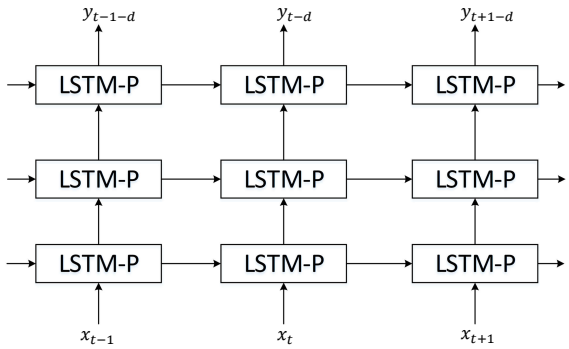

Figure 3: The structure of the LSTM-RNN for acoustic modeling.

1) It is difficult to cluster these features. Although decision tree based context clustering [16] can be applied, there are still many problems [26].
2) Manually designed context feature cannot include sufficient context information in the whole sentence.
3) Most importantly, it is difficult to directly evaluate the distance between two different context features.

A DNN based context embedding method was proposed in [27]. The DNN was used to transform input discrete context features into fixed dimensional continuous features. Therefore, context features can be evaluated using Euclid distance or KLD in the embedding space. This method can address the problem 1) and 3) mentioned above and it was applied to unit selection based speech synthesis [28]. However, the DNN is a frame-wise model, it can not encode sufficient information about the whole input contextual sequence.

In this paper, we propose to use an LSTM-RNN for context embedding instead of a DNN. An LSTM-RNN is a sequential model with powerful ability in memorized the while sequence, it has the potential ability to address these three problems simultaneously. The structure of the LSTM-RNN is shown in Fig. 3. There are three hidden layers in this architecture. Each hidden layer has a projection output [29] in order to compress the dimensionality of layer output. Therefore the number of parameters and computational cost are significantly reduced. Note that in our structure, a delayed output is adopted, which means that $t-d$-th frame of acoustic feature is output at time step t. With this structure, the model can see a few future information with a uni-directional recurrent structure. In this paper, time delay d is set to 10 frame.

At unit selection time, the output of the first hidden layer is used as the context embedding vector. Therefore, the statistical criterion for unit selection becomes

$$\begin{aligned}\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{6} w_m [&\log P(\boldsymbol{X}(\boldsymbol{U}, m)|C, \lambda_m) \\ &- w_{KLD} D_m(C(\boldsymbol{U}), C)] \\ &- w_{ce} D_e(H(C(\boldsymbol{U})), H(C)), \end{aligned} \quad (8)$$

in which the first two terms of the right hand side of equation is exactly the same as those in the baseline system. $H(C)$ is the context embedding sequence of context feature sequence C. D_e denotes the Euclid distance between two context embedding sequences. Phone-level linear interpolation was utilized to normalize the lengths of phone units with duration. w_{ce} is the weight of this term that need to be tuned.

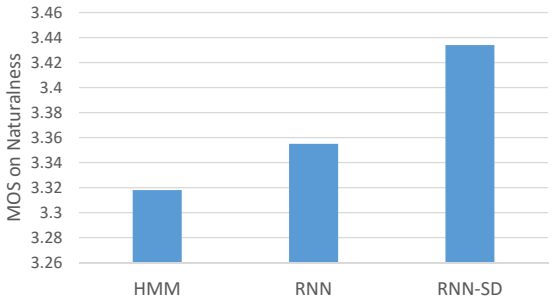

Figure 4: Mean opinion score on naturalness of three compared systems.

3.3.2. *Dialogue and sentence type embedding*

Since the corpus for the challenge this year is highly expressive, our conventional context feature is insufficient for prosody modeling. To enrich the context feature, we added the dialogue embedding and sentence type embedding into the input of the LSTM-RNN. Dialogue embedding indicates whether the current phoneme is a dialogue in the story. Sentence type was obtained from the raw text according to its punctuation.

3.3.3. *Internal experiment*

Table 3: Systems compared in the subjective text.

system	description
HMM	The HMM-based baseline unit selection system.
RNN	HMM + context embedding.
RNN-SD	RNN + sentence type embedding and dialogue embedding.

We conducted a listening test on the AMT to verify the performance of the proposed method. Table 3 presents the three systems that were compared in the test. Results in Fig. 4 proved the effectiveness of the proposed method. The RNN-SD 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 L.

Fig. 5 presents the boxplot of mean opinion scores (MOS) of each submitted system on similarity. Since we built our unit selection and waveform concatenation system directly using speech recordings with high sampling rate of 44.1 kHz, our system L achieved a high mean opinion similarity score of 4.2, which is higher than all other submitted systems. Our system is significantly better than other systems except system M, whose MOS is 3.9.

Fig. 6 shows the boxplot of MOS of each system on naturalness. Our system also achieve MOS of 4.2, higher than all other systems. And our system is the only system that is higher than 4.0. The difference between our system and other submitted system is significant except M, whose MOS is 3.9.

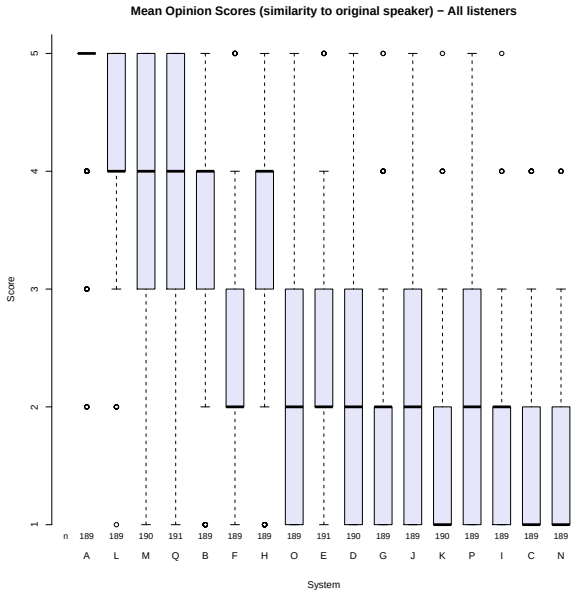

Figure 5: Boxplot of similarity scores of each submitted system.

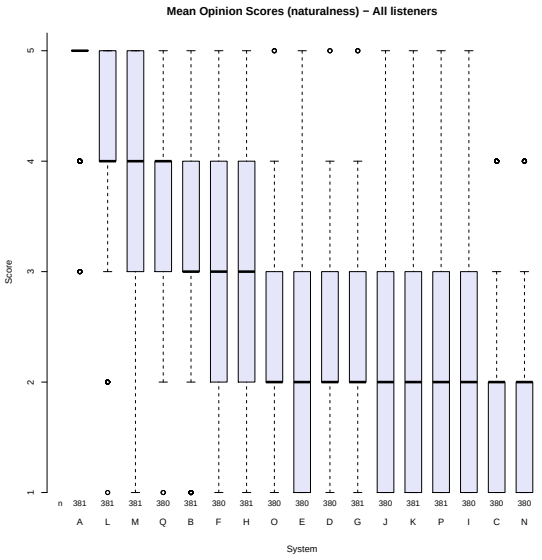

Figure 6: Boxplot of naturalness scores of each submitted system.

Our system also performed the best on the intelligibility test. As shown in Fig. 7, the word error rate (WER) of our system is 26%. Better units can be selected using our new criterion with the LSTM-RNN based context embedding for unit selection. This leads to a better WER result in the evaluation, even better than the HMM base parametric speech synthesis approaches.

The scores of our system in the paragraph test are presented in Table 4. An additional comparison between our system and the best system other than our system, which is system M, is shown in Fig 8. It can be seen that our system outperform other systems in all metrics.

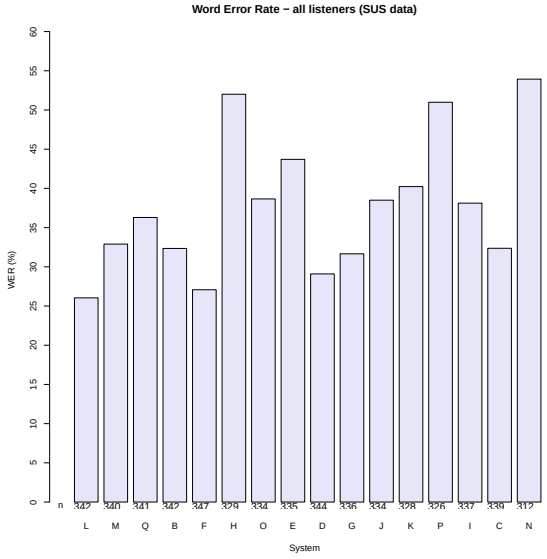

Figure 7: WER of each submitted system.

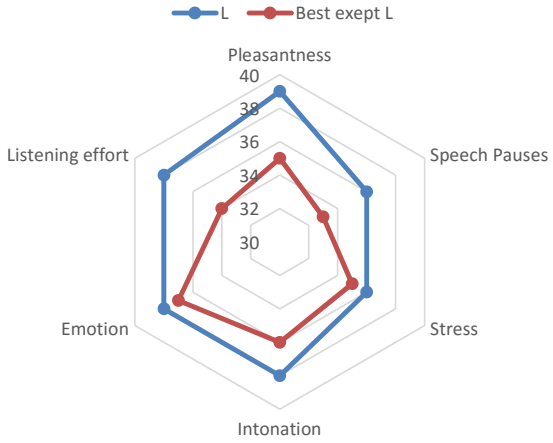

Figure 8: The structure of LSTM-RNN for acoustic modeling.

Table 4: Mean opinion scores of paragraph test.

	MOS
Pleasantness	39
Speech Pauses	36
Stress	36
Intonation	38
Emotion	38
Listening effort	38

5. Conclusions

This paper presented the details of building the USTC system for the evaluation of Blizzard Challenge 2016. The LSTM-RNN based models were used in our system in both front-end text processing and back-end acoustic modeling. We adopted them for ToBI prediction, such as accent, phrase boundary and boundary tone prediction. Context embeddings were also extracted by an LSTM-RNN to help the acoustic modeling and

unit selection. Dialogue embeddings and sentence type embeddings were included for better prosody modeling to enrich the expressiveness of synthetic speech. The effectiveness of our system is verified by both our internal experiments and official evaluation results. Our system outperformed all other submitted systems. The future work will be further investigating the LSTM-RNN based context embedding in unit selection based speech synthesis.

6. References

[1] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for blizzard challenge 2006: an improved HMM-based speech synthesis method," in *Blizzard Challenge Workshop*, 2006.

[2] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, "The USTC and iflytek speech synthesis systems for blizzard challenge 2007," in *Blizzard Challenge Workshop*, 2007.

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

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

[5] Y. Jiang, Z. Ling, M. Lei, C. Wang, H. Lu, Y. Hu, L. Dai, and R. 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 L.-R. Dai, "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." 3rd ESCA Workshop on Speech Synthesis, 1998, pp. 77–80.

[10] L.-H. Chen, T. Ratio, C. Valentini-Botinhao, Y. Junichi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," in *Proc. Interspeech*, 2014.

[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.

[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, Jiang, Y, and Y.-Q. Zu, Yi-Qingan, "The USTC system for blizzard challenge 2015," in *Blizzard Challenge Workshop*, 2015.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech.*, vol. 5, 1999, pp. 2347–2350.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, 1999, pp. 229–232.

[16] T. W. K. Shinoda, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust Soc. Japan (E)*, vol. 21, no. 2, 2000.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist*, vol. 22, no. 1, 1951.

[18] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *5th ISCA Speech Synthesis Workshop*, 2004.

[19] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[20] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[21] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 189–194.

[22] K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg, "Tobi: A standard scheme for labeling prosody," in *Proceedings of the Second International Conference on Spoken Language Processing*, 1992, pp. 867–879.

[23] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, ser. Studies in Computational Intelligence. Springer, 2012, vol. 385.

[24] 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.

[25] A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech communication*, vol. 33, no. 1, pp. 135–151, 2001.

[26] 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*, 2013, pp. 7962–7966.

[27] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in *Proc. Interspeech*, 2015.

[28] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5145–5149.

[29] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014, pp. 338–342.