# The USTC System for Blizzard Challenge 2018

*Yuan Jiang*[1 2], *Xiao Zhou*[1], *Chuang Ding*[2], *Ya-jun Hu*[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

## Abstract

This paper introduces the USTC speech synthesis system for Blizzard Challenge 2018. The task is to build a speech synthesis system on a 6.5-hour children's audio book corpus. The submitted system followed our previous one proposed in Blizzard Challenge 2017. A hidden Markov model (HMM)-based unit selection system was built with improvements in both the front-end text processing and back-end acoustic modeling. In the front-end, long short term memory(LSTM)-based recurrent neural networks(RNN) were adopted for tone and breaking indices (ToBI) prediction. In the back-end, two models were built for unit selection, a LSTM-RNN based acoustic model was built and the hidden layer was adopted as context embedding feature, a DNN based unit embedding model was built and the unit vector was adopted as phone unit feature. Evaluation results demonstrated that our system performed good on all aspects of paragraph test, which proved the effectiveness of our proposed system.

**Index Terms**: Blizzard Challenge 2018, speech synthesis, unit selection, HMM, LSTM, unit embedding

## 1. Introduction

Blizzard Challenge was organized annually since 2005 to better understand and compare research techniques in building corpus-based speech synthesis systems on the same data. During the past thirteen years, unit selection based waveform concatenation approaches and statistical parametric speech synthesis (SPSS) approaches have been the most popular methods.

Benefiting from the direct use of natural speech segments, unit selection based waveform concatenation systems could generate speech segments resembling natural speech [1, 2]. The speech quality of unit selection based systems surpass that of SPSS systems by a large margin. The main deficiencies are the demands for large speech corpus and expert fine-tuning. On the other hand, statistical parametric speech synthesis (SPSS) methods try to parameterize waveforms and build acoustic models to predict the acoustic features [3, 4, 5]. Then, a vocoder, such as STRAIGHT [6], is used to reconstruct the speech waveform given the acoustic features. SPSS systems have advantages on flexibility and small footprint, but the speech quality of SPSS systems is limited by the vocoder. Some approaches have been proposed to integrate feature extraction with acoustic model training in the hidden Markov model(HMM)[7] based or deep neural network(DNN) based SPSS systems[8, 9]. Although these methods could generate waveforms directly, they suffer from similar problems as in vocoder approaches due to the use of hidden feature extraction process. Recently, a neural network based autoregressive model name WaveNet was proposed [10], and could generate speech waveforms directly. WaveNet outperformed the baseline HMM-based unit selection system[11] on speech naturalness. However, it demands larger corpus to train and suffers from low efficiency problem in the point by point autoregressive generation process.

This year, participants in Blizzard Challenge were asked to construct speech synthesis systems based on a 6.5-hour children's audio book corpus which was the same as the corpus used last year. Although it is demonstrated in [10] that WaveNet-based system outperformed both the statistical parametric system[12] and HMM-based unit selection system on a 24.6-hour dataset, it is in doubt whether it can achieve similar performance on this comparatively small dataset. We implemented a WaveNet-based TTS system last year[13], and the results showed problems in speech naturalness and intelligibility. Therefore, we chose the unit selection based method for the task this year. Several techniques were proposed to improve the performance of the baseline HMM-based unit selection system. A long short term memory(LSTM)-based recurrent neural networks(RNN) was adopted for tone and breaking indices(ToBI) prediction in the front-end text analysis. Another LSTM-RNN based acoustic model was built and the hidden layer was adopted as context embedding features for unit selection in the back-end, an DNN-based unit embeddings model was used to represent the acoustic characteristics of phone-sized candidate units with fixed-length vectors. The system was constructed following the framework of our system for Blizzard Challenge 2017[2]. Evaluation results demonstrated the superiority of our submitted system.

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

## 2. Method

In our conventional HMM-based unit selection system, the statistical modeling techniques used in the HMM-based parametric synthesis were introduced in order to incorporate the advantages of SPSS systems. Although the system robustness can be improved, there are still some problems. First, decision tree is a simple model that it cannot express complex context dependencies, as well as the long term relationship between consecutive linguistic frames. This limitation will degrade the unit selection results. Suboptimal units may be selected for the target context. Second, as only concatenation costs at the boundaries of candidate units are evaluated, the expressiveness consistency of the selected unit sequence cannot be guaranteed. Systems built on high expressiveness corpus, such as the provided audio book corpus with various style variation, will suffer from instability problems. In order to alleviate those problems, we proposed to add context embedding features to guide the unit selection process last year, which was inspired by [14]. A LSTM-RNN

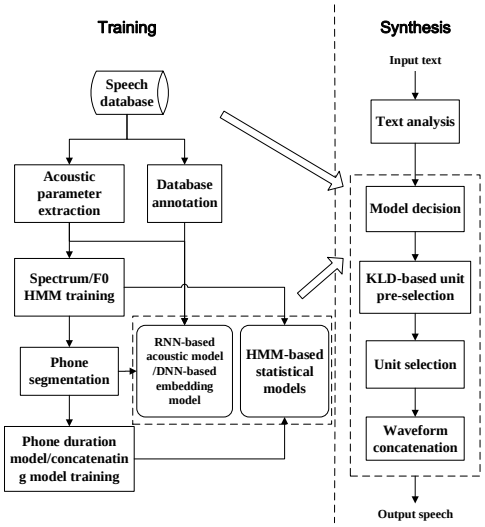


Figure 1: *Flowchart of the USTC unit selection system.*

based acoustic model was used to extract these features in our system. Deep neural networks have shown superiority over decision trees in dealing with context features[4]. Context feature dependencies within one frame or between consecutive frames can be compactly represented by deep neural networks. So the first problem can be addressed. As deep neural network based acoustic models can generate stable feature sequence, expressiveness consistency of the selected units is supposed to be improved when the context embedding feature sequence of the target context are used to guide unit selection procedure[2]. In this work, one more phone-level DNN acoustic models was built to predict the embedding vectors and are integrated into unit selection criterion[15]. Phone-level models can better capture the dependencies among consecutive candidate units and are expected to be more appropriate for the unit selection. Therefore, a conventional statistical models, a LSTM-RNN based acoustic model and a DNN based unit embedding model were trained in the training phase of our system. In synthesis phase, the distance cost of frame and phone-level embedding features was incorporated to the selection criterion and used to guide unit selection process. Figure 1 illustrates this upgraded framework of HMM-based unit selection system.

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

#### *2.1.1. HMM based statistical models training*

We used phone as the basic segment for unit selection. 6 context dependent statistical models considering different features were estimated. These models included a spectral model, a F0 model, a phone duration model, a concatenating spectral model, a concatenating F0 model, and a syllable-level F0 model.

The spectral model and F0 model were used to describe frame-level spectral and F0 distribution respectively. Phone duration model was used to present the distribution of frame numbers within a phone. All these three models were estimated under the training framework of the traditional HMM-based statistical parametric speech synthesis system[3], where spectra were modeled by a continuous probability distribution and F0s were

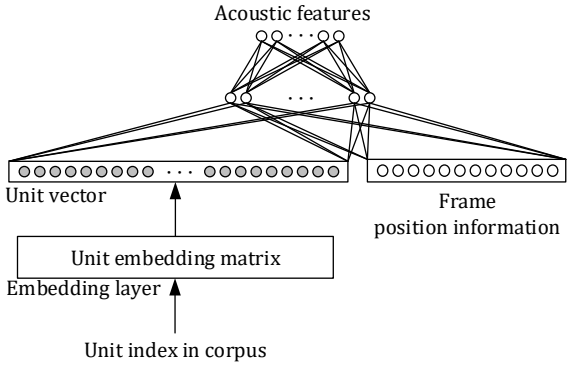


Figure 2: *Flowchart of the Unit2Vec model for learning unit embeddings.*

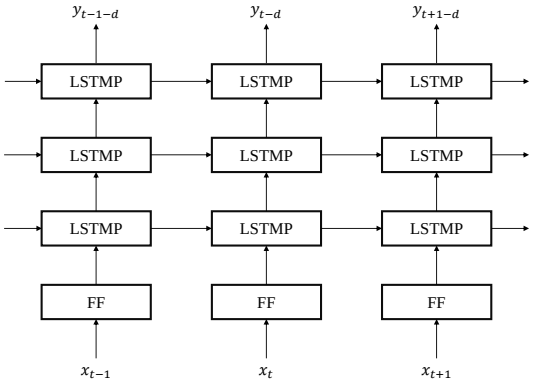


Figure 3: *The structure of our LSTMP-RNN based acoustic model with d frames delay.*

modeled by a multi-space probability distribution(MSD)[16]. After the spectral and F0 models were obtained, the state and phone boundary information were generated by using a viterbi based force alignment algorithm. Then phone duration model was trained using those information. All these three models were optimized based on ML criterions.

Concatenation models were used to model the distributions of difference features of spectrum and F0 at the phone boundaries separately. Delta and delta-delta feature vectors of spectra and F0s were used for model training. In addition, a syllable-level F0 model, which were trained using F0 features extracted from the vowels of two adjacent syllables, was used to capture the long term prosody dependency in F0s. The alignment information generated above was used to extract those features for concatenation models training.

During the training process, decision tree based model clustering technique was applied to cope with the data-sparsity problems. And the minimum description length(MDL) based model clustering was utilized to control the size of decision trees[17].

#### *2.1.2. LSTM-RNN based acoustic model training*

In [14], a DNN based acoustic model was used for context embedding feature extraction. As DNNs deal with each frame independently, correlations between consecutive linguistic frames

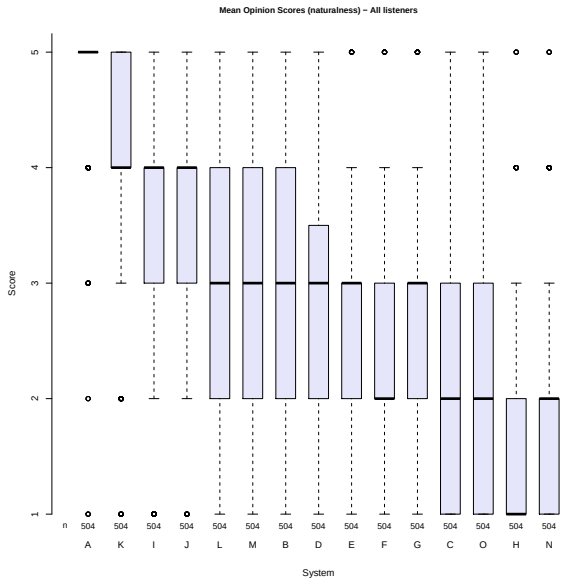


Figure 4: *Boxplot of naturalness scores of each submitted system for all listeners.*

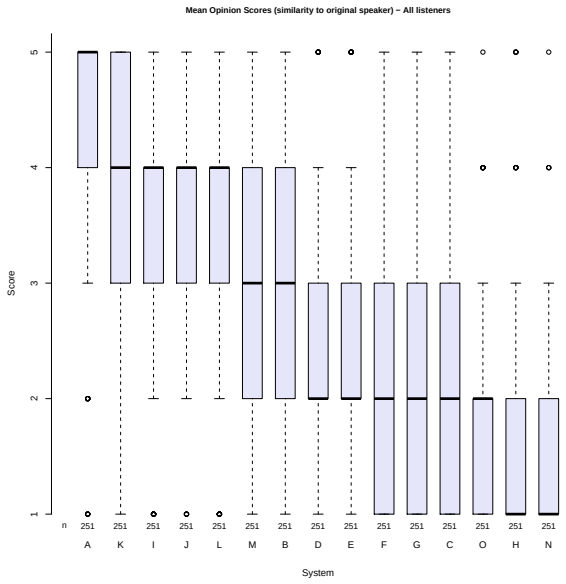


Figure 5: *Boxplot of similarity scores of each submitted system for all listeners.*

can not be modeled. LSTM-RNN is a kind of RNN models, the structure of which is specially designed that it can capture the long-term dependencies in sequential data[18]. The superiorities of LSTM-RNN for acoustic modeling have been demonstrated in [5]. Therefore, we proposed to use LSTM-RNN based acoustic model for context embedding feature extraction.

Despite the conventional phone-level and frame-level linguistic features used for neural network based acoustic model training, dialogue mark and sentence type label were also used in our system in order to enrich the input linguistic features for prosody modeling. Dialogue mark was determined based on the fact that whether the current phoneme was in a dialogue. Sentence type was determined based on the punctuation in the raw text. The embedding vectors of those features were added as part of the input vector for LSTM-RNN based acoustic model training.

#### *2.1.3. DNN based unit embedding model*

A DNN named *Unit2Vec* is designed to learn a fixed-length vector for each phone unit in the corpus for unit selection from scratch[15]. The flowchart of the Unit2Vec model is shown in Fig. 2.

The dimension of the unit embedding matrix is $R \times D$, where $R$ is the total number of candidates in the corpus and $D$ is the length of the embedding vector for each candidate. All unit vectors are stored in the weight of the embedding layer as an embedding matrix. Given a row index, we can extract corresponding unit vector from the matrix. The phone boundaries in the corpus are given by HMM-based force-alignment. For each frame in the corpus, a unit vector is determined by selecting the row index of the embedding matrix corresponding to the phone unit that the frame belongs to. Then the unit vector is concatenated with frame position information to predict the acoustic features (i.e., MCCs and F0s) of this frame. The unit embedding matrix is learnt by minimizing the mean square error (MSE) between the predicted and the natural acoustic features. Shuffling the frame-level training data is necessary, which makes the data corresponding to a same unit randomly distribute in the training set and helps to improve the estimation of unit vectors. The learnt unit vector of each phone unit is expected to describe the overall acoustic characteristics of the unit, which will be further modeled to derive the cost functions for unit selection.

### 2.2. Synthesis phase

#### *2.2.1. Unit selection considering context embedding features*

Distance between context embedding features of the target context and context of candidate units was used to guide unit selection process in the synthesis phase.

Supposing $N$ phonemes are included in the text to be synthesized and the context feature sequence is $C$. The optimal phone unit sequence $\boldsymbol{U} = \{u_1, u_2, ..., u_N\}$ is searched out from the pre-stored database by maximizing the criterion below:

$$\begin{aligned} \boldsymbol{U}^* = \arg\min_{\boldsymbol{U}} \sum_{m=1}^{6} \omega_m [logP(\boldsymbol{X}(\boldsymbol{U}, m)|C, \lambda_m) \\ - \omega_{KLD} D_m(C(\boldsymbol{U}), C)] \\ - \omega_{ce} D_e(H(C(\boldsymbol{U}), H(C)) \\ - \omega_{cu} D_u(V(C(\boldsymbol{U}), V(C)), \end{aligned} \tag{1}$$

where $\{\lambda_m\}_{m=1}^{6}$ are the HMM-based acoustic statistical models obtained in the training stage, $\boldsymbol{X}(\boldsymbol{U}, m)$ denotes the acoustic features extracted from the $m$-th model corresponding to the candidate unit sequence $\boldsymbol{U}$, and $C(\boldsymbol{U})$ denotes the context feature sequence of the candidate unit sequence $\boldsymbol{U}$, $P(\cdot)$ and $D_m(\cdot)$ represent the likelihood and KLD calculation functions separately, $\omega_m$ and $\omega_{KLD}$ are the weight coefficients for the $m$-th model and the KLD component in the criterion, which need to be manually tuned. Besides, $H(C)$ and $H(C(\boldsymbol{U}))$ are the frame level context embedding feature sequence of $C$ and that of context feature sequence $C(\boldsymbol{U})$, $D_e$ denotes the Euclidean

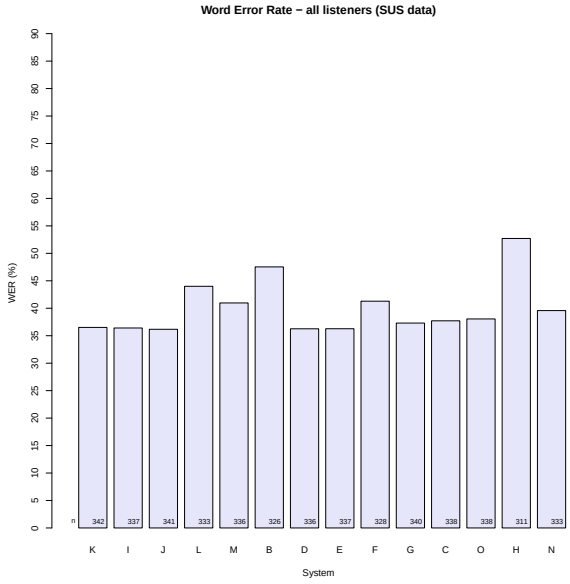


Figure 6: *Word error rate score of each submitted system for all listeners.*

distance between the two context embedding feature sequences, $\omega_{ce}$ is the tunable weight coefficient of this component. Phone-level linear interpolation was applied to normalize the lengths of phone units with duration. $V(C)$ and $V(C(\boldsymbol{U}))$ are the phone level unit embedding feature sequence of $C$ and that of unit feature sequence $C(\boldsymbol{U})$, $D_e$ denotes the Euclidean distance between the two context embedding feature sequences, $\omega_{cu}$ is the tunable weight coefficient of phone level component. Dynamic programming(DP) algorithm was used to search the optimal sequence out. Before that, a KLD-based unit pre-selection method[1] was used in order to reduce computational complexity.

#### *2.2.2. Waveform concatenation*

After the optimal unit sequence was searched out, the corresponding waveforms of these units were concatenated. The cross-fade technique[19] was used in order to smooth the phase discontinuity at the concatenation points of unit boundaries.

## 3. System Building and Evaluations

### 3.1. System building

#### *3.1.1. Database annotation and text analysis*

We used the iFLYTEK English text analysis tool to get the phoneme transcriptions of the texts in the provided dataset. The accent, phrase boundary and boundary tone in the token and breaking indices(ToBI) set were predicted using the same method as that used before years[20].

#### *3.1.2. Statistical models training*

Recurrent layer with long short-term memory projected(LSTMP) architecture[21] was used in our network with the aim to reduce computational cost. The network structure of LSTM-RNN based acoustic model included 1 feed-forward layer, 3 LSTMP layers and 1 linear output layer. As shown in figure

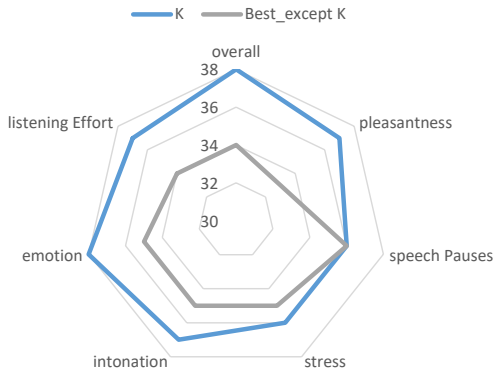


Figure 7: *Comparison between our system K and the best system except K on all the listening aspects in paragraph test.*

3, the time delay strategy was used in this model. So that the extracted context embedding features can integrate the linguistic information in both past and future. In our system, time delay $d$ is set to 10 frames. A mini-batch stochastic gradient descent (SGD) algorithm was adopted for model optimization. After some informal comparison listening tests, we chose the output of the second hidden layer as the context embedding vector at last.

Acoustic features extracted from 16kHz waveforms were used for all statistical models training and unit selection. 48kHz waveforms were generated at last by concatenating waveform segments corresponding to the optimal candidate unit sequence.

### 3.2. Evaluation results

We present the listening tests results of our system in Blizzard Challenge 2018. 14 systems, including 4 benchmarks and 10 submitted systems, plus the natural speech were evaluated. The identifiers for the benchmark systems and our system are:

- A: Natural speech
- B: Festival benchmark
- C: HTS benchmark
- D: DNN benchmark
- E: DNN trajectory benchmark
- K: Our system

#### *3.2.1. Naturalness test*

Figure 4 shows the boxplot of evaluation results of all systems on naturalness. 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.

#### *3.2.2. Similarity test*

The boxplot of similarity evaluation results is presented in figure 5. Our system achieves the equal highest speaker similarity among all the submitted systems. The difference between our system and any other participant system is significant except for system I and J.

#### 3.2.3. Intelligibility test

The word error rate(WER)s of all participant systems are presented in figure 6. When evaluated by all listeners, the WER of our system is 37%. The score of our system is one percent higher than benchmark system D and E, and the participant systems I and J, but Wilcoxon signed rank tests indicate that the difference is not significant. When evaluated by the paid native English listeners, the WER of our system is 16% . The difference is insignificant toobetween our system and other systems with better score.

#### 3.2.4. Paragraph performance

In this test, seven aspects of speech, including overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening efforts are evaluated separately. Our system achieves almost the best performance on all aspects as shown in figure 7. The mean opinion scores of our system are listed in table1. Considering the full mark is 60, the performance of our system is not good enough. More efforts have to be made to promote performance improvements all aspects, especially on speech pause and stress.

Table 1: *Paragraph listening test scores of our system.*

| | MOS |
|---|---|
| overall impression | 38 |
| pleasantness | 37 |
| speech pauses | 36 |
| stress | 36 |
| intonation | 37 |
| emotion | 38 |
| listening effort | 37 |

## 4. Conclusions

We submitted an unit selection system this year. The system was constructed following the previous framework we developed for Blizzard Challenge 2017, and a new DNN based unit embedding model was used in this year. Three techniques were proposed to improve the performance of HMM-based unit selection system. These techniques were: 1)the context embedding features extracted from a LSTM-RNN based acoustic model were used to improve the unit selection results; 2)LSTM-RNN based models were trained for ToBI prediction in order to improve the prosody; 3)Frame level and phone level embedding vectors were included in the linguistic features for LSTM-RNN based acoustic model and DNN based unit embedding model training in order to enrich the expressiveness of synthetic speech. Evaluation results in Blizzard Challenge 2018 demonstrated the effectiveness of our system.

## 5. References


[1] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen *et al.*, "The USTC and IFLYTEK speech synthesis systems for Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007.

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

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

[4] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

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

[6] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[7] R. Maia, H. Zen, and M. J. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters." in *SSW*, 2010, pp. 88–93.

[8] K. Tokuday and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4215–4219.

[9] K. Tokuda and H. Zen, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5640–5644.

[10] 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.

[11] X. Gonzalvo, S. Tazari, C.-a. Chan, M. Becker, A. Gutkin, and H. Silen, "Recent advances in google real-time HMM-driven unit selection synthesizer." in *INTERSPEECH*, 2016, pp. 2238–2242.

[12] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices," *arXiv preprint arXiv:1606.06061*, 2016.

[13] Y.-J. Hu, C. Ding, L.-J. Liu, Z.-H. Ling, and L.-R. Dai, "The USTC system for Blizzard Challenge 2017," in *Blizzard Challenge Workshop*, 2017.

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

[15] X. Zhou, Z.-H. Ling, Z.-P. Zhou, and L.-R. Dai, "Learning and modeling unit embeddings for improving HMM-based unit selection speech synthesis," in *Proc. INTERSPEECH*, 2018.

[16] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1. IEEE, 1999, pp. 229–232.

[17] K. Shinoda and T. Watanabe, "Mdl-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology*, vol. 21, no. 2, pp. 79–86, 2001.

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

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

[20] 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.

[21] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.