A BLSTM Guided Unit Selection Synthesis System for Blizzard Challenge 2016

Jianhua Tao[1,2], *Yibin Zheng*[1], *Zhengqi Wen*[1], *Ya Li*[1], *Biu Liu*[1]

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
{jhtao, yibin.zheng, zqwen, yli, liubin }@nlpr.ia.ac.cn

Abstract

The paper introduces the speech synthesis system developed by Institute of Automation, Chinese Academy of Sciences (CASIA) for Blizzard Challenge 2016. About 5 hours of speech data from professionally-produced children's audiobooks is adopted as the training data for the construction this year. Different from our previous systems, the BLSTM guided unit selection and waveform concatenation approaches is selected to develop our speech synthesis using the provided corpus. We will describe our definitions of the acoustic, prosodic and linguistic parameters, procedure of candidate unit selection, components of cost function, etc. Finally, we will also present the results of the listening test conducted.

Index Terms: speech synthesis, phone duration modeling, BLSTM, unit selection, Blizzard Challenge 2016

1. Introduction

This paper describes our fourth participation of speech synthesis system in a Blizzard Challenge. The task in this year's challenge is to build a speech synthesis from the provided data that is suitable for reading audiobooks to children. The narration in the audiobooks is lively and expressive and the speaker impersonates or performs several characters apart from the narrator herself.

Statistical parametric speech synthesis (SPSS) systems have flexible and robust advantages [1] over unit selection [2] systems. However, during the process of extracting and modeling speech parameters, followed by re-synthesis, the naturalness of the speech is substantially reduced. As a consequence, these systems are consistently rated as less natural than unit selection, as we can see in the results of many Blizzard Challenges [3, 4, 5, 6].

In order to improve the quality of speech above the ceiling imposed by vocoding, a bi-direction long short-term memory recurrent neural network (BLSTM) guided unit selection synthesis system is built for the Blizzard Challenge 2016. This system differs in our previous Blizzard Challenge system [7] in mainly three aspects, (i) the acoustic modeling approach, which has been updated from HMM to BLSTM. Because BLSTM based acoustic modeling techniques have achieved state-of-the-art performance in SPSS due to its deep architecture and capacities to capture long-term dependencies across the linguistic inputs, which HMM doesn't possess. (ii) duration modeling approach, which has been substituted with the BLSTM. This is also the novel employment of BLSTM for duration model in existing work. (iii) target cost calculation approach, as BLSTM guided synthesis system doesn't have the concept of "state". Some target cost computation approaches are investigated and compared in our system.

The rest of the paper is organized as follows. Section 2 gives an overview of our methods used for system construction. Section 3 gives an detailed introduce of the BLSTM based phone duration prediction. Section 4 introduces the unit selection module, including the pre-selection of units, calculation of target and concatenation cost. In section 5, the evaluation results of our system in Blizzard Challenge 2016 are shown and discussed. The conclusions are presented in section 6.

2. System Overview

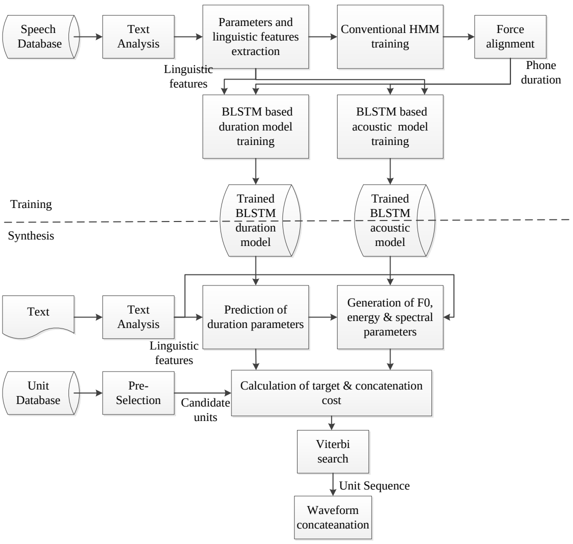
Speech Database
Text Analysis
Parameters and linguistic features extraction
Conventional HMM training
Force alignment
Linguistic features
Phone duration
BLSTM based duration model training
BLSTM based acoustic model training
Training
Synthesis
Trained BLSTM duration model
Trained BLSTM acoustic model
Text
Text Analysis
Prediction of duration parameters
Generation of F0, energy & spectral parameters
Linguistic features
Unit Database
Pre-Selection
Candidate units
Calculation of target & concatenation cost
Viterbi search
Unit Sequence
Waveform concateanation

Figure 1: *An overview of our system.*

A hybrid synthesis system that uses BLSTM statistical models (by generating speech parameter trajectories) as the basis of the target cost function [8, 9, 10] is adopted as our approach for Blizzard Challenge 2016. HMM is the preferred statistical model in hybrid system's target cost function in previous years. However, as recent but compelling evidence that BLSTM is superior to the regression tree employed in HMM systems [11, 12, 13], a hybrid synthesis system based on BLSTM is employed to synthesis the voices for Blizzard Challenge 2016. The flowchart of the BLSTM guided unit selection speech synthesis system for Blizzard Challenge 2016 is shown in Figure 1. It consists of two main stages: the training stage and the synthesis stage, to build the BLSTM guided unit selection speech synthesis system.

In the training stage, BLSTM based acoustic and duration model is trained to guide the unit selection. Before the training of BLSTM based acoustic model, a HMM based force alignment is performed first. In the HMM based force alignment part, acoustic parameters are extracted from the speech waveforms. The complete feature vector for each frame consists of static, delta and acceleration components of the spectral parameters and the logarithmized F0. With the segmental and lingustic features data from text analysis module (which is done by festival toolkit [18]), the spectral part is modeled by continuous probability HMM and F0 part is modeled by multi-space probability HMM (MSD-HMM).. Then the phone boundaries of the training utterances are determined by Viterbi alignment using the trained HMM model. Then the linguistic features, together with the phone duration from the force alignment part is made up of the input for BLSTM training. As for the output of BLSTM training part, the complete feature vector for each frame only consists of static components of the spectral parameters and the logarithmized F0, together with a flag of unvoice/voice (U/V). The BLSTM based acouctic model is used to calculate the target cost. The linguistic features and the phone duration also made up of the training corpus for the BLSTM based duration model. The BLSTM based duration model is used to predict the phone duration, target and concatenation cost in the open test. A more detailed description about the BLSTM based duration model is given in section 3.

In the synthesis stage, firstly, the contextual information of the text to be synthesized is analyzed and extracted by text analyzer (festival toolkit). Secondly, the pre-selection procedure is conducted according to the contextual information. Then the phone duration is predicted using the trained BLSTM duration model. Then the phone duration model, together with the linguistic features are fed into the BLSTM to predict the target acoustic parameters. Next, the target cost of candidate unit and the concatenation costs between each pair of adjacent candidate units can be calculated. The optimal candidate units are selected by Viterbi search. Finally, the waveform fragments of optimal units are concatenated, and the silence sections are inserted between some adjacent words based on the value predicted by silence model.

3. Phone Duration Modeling

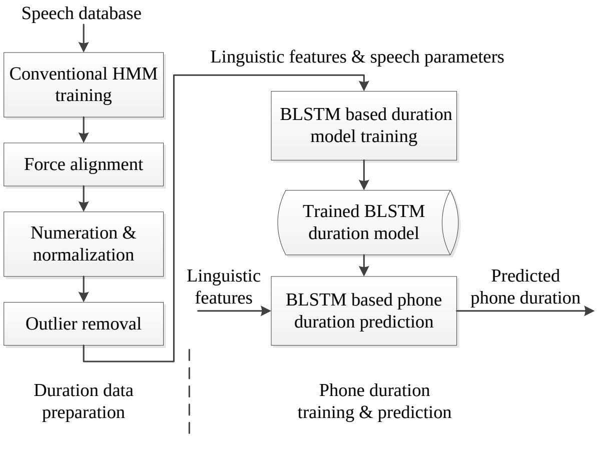
Speech database
Conventional HMM training
Force alignment
Numeration & normalization
Outlier removal
Linguistic features & speech parameters
BLSTM based duration model training
Trained BLSTM duration model
Linguistic features
BLSTM based phone duration prediction
Predicted phone duration
Duration data preparation
Phone duration training & prediction

Figure 1: *BLSTM based phone predition in our system.*

Accurate modeling and prediction of phone duration is an important components in generating more natural synthetic speech [14, 15]. There are mainly two enployments of duration model in our system. One is that the prediction of phone duration is used as the input for the BLSTM based acoustic model. Another enployment is for the calculation of the duration target cost. Therefore, we focus on statistical techniques for improved duration modeling, as a key step towards the overarching goal of more natural and appropriate synthetic speech.

As recent rise of BLSTM has brought an increase in performance in both automatic speech recognition (ASR) [16] and statistical acoustic model for speech synthesis [17]. In these tasks, its powerful sequence modeling has been proved. Here, we consider duration modeling at phone level, for the audiobook data, using BLTSM. BLSTM based duration method with outlier removal is shown in Figure 2. This framework is general, thus it is easy to replace BLSTM with other machine learning methods for duration prediction purpose. Some important steps involved are briefly discussed below.

In duration data preparation part, force alignment is carried out at phone level after conventional HMM training. This step is to segment each utterance into a sequence of shorter and simple speech units, thus each unit can be modelled independently in subsequent steps. Unlike decision tree, BLSTM can only handle numeric features, thus it is necessary to encode all nominal features to be numeric values. Normalization is immediately carried out to transfer feature values into a limited interval. Outlier removal is then carried out for all speech units.

In phone duration training and prediction part, BLSTM is trained with “cleaned” training samples and stored. After that, phone level duration are predicted by BLSTM for any given linguistic features of full context label.

3.1. Outlier removal

Few previous researches tried to remove duration outliers in duration prediction task, but outliers have dramatic degradation effect to most machine learning methods, thus we incorporated outlier removal in duration data preparation step.

Careful manual checking shows that the aligned duration boundaries are roundly acceptable, but not very accurate. Some alignment errors may also occur if the speech wave doesn’t strictly correspond to its text transcription. Theses errors are usually cause the durations of some speech units in the problematic utterance too long, ie., longer than a maximum duration based on our prior knowledge. It is clear that some of the aligned duration are unreasonably long, and mismatch between speech wave and its text transcription is verified by visual checking.

Therefore, a simple outlier removal method is designed: given any phone unit, the upper 1% of duration of this unit is regarded as the maximum duration, and any utterance containing an automatically aligned speech unit longer than the maximum duration is removed; Such removal shows its effectiveness in prelimilary experiments.

4. Unit Selection Module

4.1. Pre-selection

In a corpus based speech synthesis system, there are too many candidate for each target unit. Conducting unit selection procedure on such a large database is very time-consuming. To decrease the number of candidate units and thus improve the running speed, a contextual information difference (CID) based pre-selection is conducted. The *CID* is defined in Equation (1) as below:

$$CID = \sum_{i=1}^{N} w_i * D_i \qquad (1)$$

,where *N* is the number of contextual information category, D_i is the difference of the *i-th* contextual information between current candidate unit and the target unit and w_i is the weight of the *i-th* contextual information.

The *CID* depicts the difference of contextual information between the candidate unit and the target unit to be synthesized. The contextual information used here includes the location of the current speech unit in word, phrase and sentence, the name of the phone, the length of word, phrase and sentences, the boundary types before and after the current unit, etc.

After the pre-selection, a small number of candidate units which have the smallest *CID* will be kept for the later processing.

4.2. Target cost calculation approaches

Target cost is defined as the difference between the predicted parameters and the parameters of candidate unit. In our work, the parametrers used for target cost include F0, duration, energy and spectral parameters). The context embeddings derived from a neural network, or alternatively the actual speech parameters predicted at the output of the network, can be thought of as a non-linear projection of the input lingutic features. The projection is learned in a supervised manner, according to whatever optimization criterion is used to train the network. We suppose that these BLSTM-derived features are more powerful than the purely linguistic features or HMM-derived features. The motivation for using a BLSTM – that, crucially, has been trained to perform the start-of –the-art performance in SPSS in recent research.

The system we built for the Blizzard Challenge 2016 operates on the phone units. Different from HMM model, the BLSTM based model doesn't have the concept of "State". Therefore, it leaves an important problem to calculate the target cost effectively. For this year's Challenge, three target cost calculation methods are tested and compared for the output of speech parameters from BLSTM based acoustic model (including the F0, erengy and spectral parameters). And for the output of duration parameters from BLSTM based duration model, only the Euclidean distance is used.

4.2.1. *The Kullback Leibler divergence (KLD)*

We divided each phone into 4 sections. The features being used for the target cost (output of speech parameters from BLSTM based acoustic model) are gathered toghteher across all frames within each of these 4 regions, from which we compute the mean and variance per section. The variance is floored at 1% of the gloval variance per feature (the floor value was chosen via informal listening). This is done for both the candidate and the target units.

The Kullback Leibler divergence (KLD) is computed for each of the 4 sub-phone regions individually.

The KLD between distribution *f* of the features computed for the frames corresponding to a given section in the test sentence, and distribution *g*, is:

$$D_{KL}(f \,\|\, g) = \frac{1}{2}[\log \frac{|\sum_g|}{|\sum_f|} + Tr[\sum_g^{-1} \sum_f] - \mathrm{d} + (\mu_f - \mu_g)^T \sum_g^{-1} (\mu_f - \mu_g)] \qquad (2)$$

where μ and $\sum$ are mean and covariance and *d* is the dimensionality of the feature vector. The KLD for each of the 4 sections comprising a phone is summed toghther to give the final divergence score. The average of $D_{KL}(f \,\|\, g)$ and $D_{KL}(g \,\|\, f)$ was used in order to make the measure symmetrical.

4.2.2. *Maximum likelihood criterion (LL)*

The same as the approach mentioned in section 4.2.1, we also divided each phone into 4 sections. The mean and variance calculation approaches is also the same as section 4.2.1. The only difference is that maximum likelihood criterion is employed for the target cost.

The maximum likelihood criterion (LL) is computed for each of the 4 sub-phone regions individually.

The LL between distribution *f* of the features computed for the frames corresponding to a given section in the test sentence, and distribution *g*, is:

$$LL(f \,\|\, g) = \frac{D_f}{D_g} \sum_{i=1}^{D_g} (x_{fi} - \mu_g)^T \sum_g^{-1} (x_{fi} - \mu_g) \qquad (3)$$

where the likelihood of acoustic model is normalized by the candidate sub-phone duration D_g and predicted sub-phone duration D_f, and the x_{fi} is the speech parameters in *i-th* frame in the sub-section of the candidate phone.

4.2.3. *Relative position based Euclidean distance (ED)*

As the BLSTM based acoustic model doesn't possess the concept of "State", and we suppose that relative position of the acoustic parameters can capture the trajectory of the acoustic parameters well. As we choose the same number of relative position for the candidates and target units, then they would have the same length. As a result, Euclidean distance can easily be employed in such situation. Therefore, a relative positon based Euclidean distance (ED) is tested in our system.

The relative position based Eucidean distance (ED) is computed for each phone regions individually (we don't need to divide each phone into 4 sub-phone in this situation). The ED between candidate features X_f and target features X_g, is:

$$ED(\mathrm{X}_f \,\|\, \mathrm{X}_g) = \sum_{i=1}^{N} (\tilde{x}_{fi} - \tilde{x}_{gi})^2 \qquad (4)$$

where $\tilde{x}_{fi}$ is the speech parameters in *i-th* relative frame in the candidate phone, $\tilde{x}_{gi}$ is the speech parameters in *i-th* relative frame in the target phone, and *N* is the number of the relative position.

4.3. Concatenation cost

The concatenation cost which includes spectra, energy and F0 cost is trying to make spectra and prosody smoothing for the synthesized speech. The final concatenation cost will be the sum of the spectra, energy and F0 concatenation cost. For concatenation cost, we simply used deviation between two speech units:

$$Concatenation\_cost = w_{F0} * D_{F0} + w_{energy} * D_{energy} + w_{spec} * D_{spec} \quad (5)$$

where D_{F0}, D_{energy} and D_{spec} are the deviation of F0, energy and spectral between two speech units, and w_{F0}, w_{energy} and w_{spec} are their corresponding weight value.

4.4. Best unit series selection

All in all, our cost denifition is comprised by two parts: the concatenation cost and the target cost. The formula is as follows:

$$Cost = w_{target} * Target\_cost + w_{cat} * Concatenation\_cost \quad (6)$$

The weights are not assigned equally. For instance, the weightes related to prosody parameters like F0 are normally higher than others. Based on the cost definition in Equation (6), a Viterbi search algorithm will be used to find the best path with the minimum cost. The final unit selection results will be found from this path.

5. System Building for Blizzard 2016

5.1. Speech database

The speech database is the British English Speech Corpus for the Blizzard Challenge 2016, which is produced by Usborne Publishing. It contains about 5 hours of speech data from professionally-produced children's audiobooks, which is recorded by a single female talker. This includes the approx.. 2 hours of pilot data from last year's Blizzard Challenge. A sentence-level alignment between text and speech for some of the data is provided by Toshiba's Cambeidge Research Laboratory.

The task (**Single task 2016-EH1: UK English Children's Audiobooks**) is to build a voice from this provided data that is suitable for reading audiobooks to children.

5.2. Building system

The speech corpus consists of high quality, clean speech data under controlled recording condition. Speech signal is down sampled at 16 kHz frequency, windowed by 25-ms Blackman window for each frame with 5-ms shift, then 40 th order Linear Spectral Pair (LSP) coefficients and fundamental frequency F0 in log scale are extracted as static features. The delta and acceleration components are appended to the static features to form the observation vector for conventional HMM training. Multi-space Probability Distribution HMM (MSD-HMM) of 5 states, left-to-right with no skip topology are used to represent basic speech units. Single Gaussian with diagonal covariance matrix is used in each HMM state. Speech waves are forced aligned with its text transcription by HTS tool HSMMAlign [19]. The case 1 algorithm in [20] is used throughout our experiments for its simplicity.

Concerning the textual features used for training the BLSTM based phone duration and acoustic model, a varity of linguistic features are used, such as the phone identity, POS and etc. All features in full label is encoded to numeric values and normalized, to be exact, nominal feature such as phone identity is encoded with one hot method, and numeric feature is divided by its maximum value. All encoded values are then concatenated as predictive veators of 341 dimensions to train BLSTM based phone duration model. These predictive vectors of 341 dimensions, together with the duration position vector of 2 dimensions, is consisted of the input vectors to train the BLSTM based acoustic model.

For both BLSTM-based (including duration and acoustic model) systems, a 3-layer neural network consisting a single non-recurrent layer, followed by 2 stacks of bidirectional layers (each with 256*2 LSTM hidden units) is used. All networks are trained with a momentum of 0.9, an initial learning of 0.0005 for the first 5 epoch, and then decreases by 20% after each epoch.

5.3. Internal evaluation

We conduct an internal evaluation to validate the effective of the BLSTM based phone duration predition approach and different target cost calculation approaches.

5.3.1. BLSTM based duration model

Two duration predition methods are compared:

1) Baseline: Decision tree based duration in HTS;
2) BLSTM: BLSTM based duration prediction with outlier removal.

1000 utterances from the test set are utilized to carry out the following the objective evaluation. As silences before the strat and after the end of an utterance are not very meaningful, they are excluded.

We then demonstrate the improvement of the proposed method by RMSE between the predicted durations and automatically aligned durations, where phone duration is based on BLSTM and conventional decision tree in HTS, respectively. The result is presented in Table 1.

The error decreased by 10.90 %. Such improvement is abvious. This improvement may be brought by the powerful modeling of BLSTM for sequence modeling tasks and the outlier removal. Therefore, this BLSTM based duration model is used in our system.

Table 1. *RMSE on phone duration prediction.*

Systems	Baseline	BLSTM
RMSE (ms)	43.67	38.91

5.3.2. Target cost calculation approaches

We also conducted a small scale listening test to compare different target cost calculation approaches and their combinations. Therefore, 4 systems were compared:

1) Unit selection based on KLD target cost calcuuation
2) Unit selection based on LL target cost calculation
3) Unit selection based on ED target cost calculation. (The number of relative position is varies from 10 , 20 to 30, and the best one is chosen for comparison.)
4) Unit selection based on KLD target cost calculation combined with LL target cost calculation

Five listeners, all of them are majored in speech related field, took part in the test. For each system, 20 sentences were played to each listener. The listeners were asked to give a 5-point mean poinion score (MOS) for each sentence they had heard. The results are shown in Figure 3. It appears that the system 4, which uses the combination of KLD and LL target cost calculation approaches, outperformed the other three systems. Therefore, system 4 is employed to generate our final submission voices.

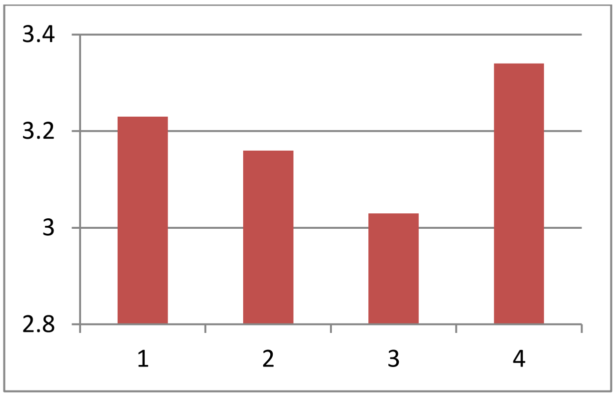

Figure 3: *MOS of the four system using different target cost calculation approaches.*

5.4. Evaluation results

16 participants attend the evaluation for **Single task 2016-EH1.** The naturalness (MOS), similarity (MOS) and intelligibility (word error rate (WER)) were calculated. The results are shown in Figure 4- Figure 6, where system A identifies natural speech and indentidy of our system is J. For all these three evaluation (naturalness, similarity and intelligibility) results, our system only ranks average level.

5.4.1. Discussion of the results

From the evaluation result, there is still a great gap between our system to the top one. There are many reasons leading this results. And the mainly one is that the text analysis module is only based on the festival toolkit, which may not quite accurate as we checked some of sentences. The unaccurate text analysis would have an undesirable consequences to the HMM training, force alignment, BLSTM based acoustic model training and BLSTM based duration model training. These results reminder us there is still many works need to be done, especially on improving the accuracy of the text analysis.

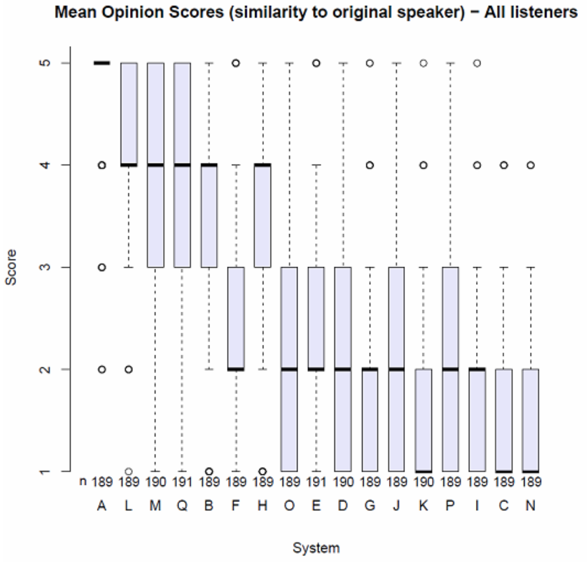

Figure 4: *Boxplot of MOS on similarity evaluation*.

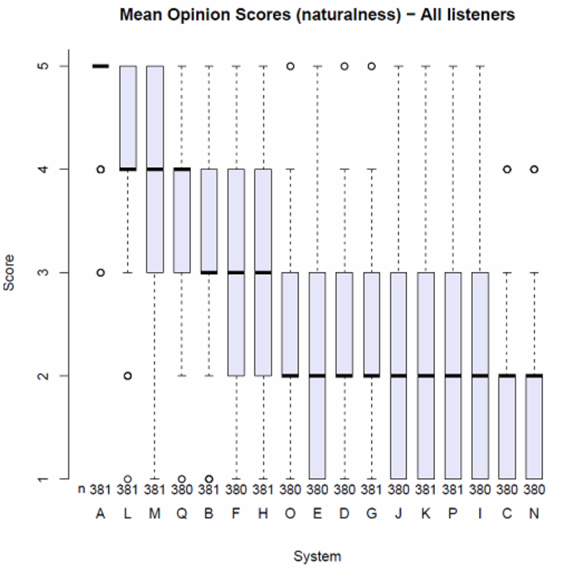

Figure 5: *Boxplot of MOS on naturalness evaluation*.

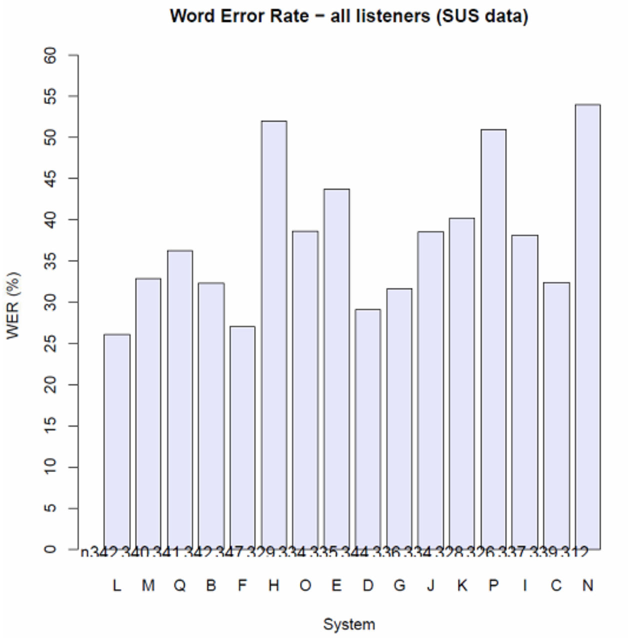

Figure 6: *Word error rates of all participants*

6. Conclusion

In this paper, the BLSTM based unit selection speech synthesis system built for Blizzard Challenge 2016 by CASIA is introduced. There are three differences from our previous Challenge system. The first one is the use of BLSTM for acoustic model. The second one is the use of BLSTM for duration predition model. The final one is the new target cost calculation approaches. The internal evaluation results show that the effectiveness of these three techniques. Also, the evaluation resuls from the Blizzard Challenge committee shows that, the naturalness, similarity and intelligibility of our system are of average level. Many works need to be done, especially on improving the accuracy of the text analysis.

7. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61305003, No.61425017，No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006) and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

8. References

[1] Heiga Zen, Keiichi Tokuda, and Alan W Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Andrew J Hunt and Alan W Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. ICASSP. IEEE, 1996*, vol. 1, pp. 373–376.

[3] Simon King and Vasilis Karaiskos, “The Blizzard Challenge 2011,” in *Proc. Blizzard Challenge*, 2011.

[4] Simon King and Vasilis Karaiskos, “The Blizzard Challenge 2012,” in *Proc. Blizzard Challenge*, 2012.

[5] Simon King and Vasilis Karaiskos, “The Blizzard Challenge 2013,” in *Proc. Blizzard Challenge*, 2013.

[6] Simon King, “Measuring a decade of progress in text-to-speech,” *Loquens*, vol. 1, no. 1, 2014.

[7] J.H. Tao, S. F. Pan, etc, “The WISTON text to speech system for Blizzard Challenge 2010,” in *Proc. Blizzard Challenge*, 2010.

[8] Zhen-Hua Ling, Heng Lu, Guo-Ping Hu, Li-Rong Dai, and Ren-Hua Wang, “The USTC system for Blizzard Challenge 2008,” in *Proc. Blizzard Challenge*, 2008.

[9] Zhi-Jie Yan, Yao Qian, and Frank K Soong, “Rich-context unit selection (RUS) approach to high quality TTS,” in *Proc. ICASSP*, 2010, pp. 4798–4801.

[10] Yao Qian, Frank K Soong, and Zhi-Jie Yan, “A unified trajectory tiling approach to high quality speech rendering,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 280–290, 2013.

[11] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, “Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis,” in *Proc. ICASSP*, 2015.

[12] Heiga Zen, “Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN,” in *Proc. MLSLP*, 2015.

[13] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li 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.

[14] Yang Wang, Minghao Yang, Zhengqi Wen, Jianhua Tao, “Combining extreme learning machine and decision tree for duration prediction in HMM based speech synthesis,” in *Proc. Interspeech*, 2015.

[15] Yamagishi Junichi, Kawai Hisashi, and Kobayashi Takao, “Phone duration modeling using gradient tree boosting,” *Speech Communication*, vol. 50, pp. 405-415, 2008.

[16] George Dahl, Dong Yu, Li Deng, and Alex Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[17] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in *Interspeech*, 2014, pp. 1964–1968.

[18] Festival [online]. Available: http://www.cstr.ed.ac.uk/projects/festival/

[19] HTS [Online]. Available: http://hts.sp.nitech.ac.jp/

[20] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM based speech synthesis," in *ICASSP*, 2000, pp. 1315-1318.