The USTC System for Blizzard Challenge 2010

Yuan Jiang, Zhen-Hua Ling, Ming Lei, Cheng-Cheng Wang, Lu Heng,
Yu Hu, Li-Rong Dai, Ren-Hua Wang

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, China
sunlion@mail.ustc.edu.cn

Abstract

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2010. USTC attended all English tasks including the hub tasks and the spoke tasks. According to the various conditions for different tasks, different versions of synthesis systems are constructed. Many new techniques are employed in our speech synthesis system construction. Results of internal experiments comparing these techniques are shown, and analyzed. The evaluation results of Blizzard Challenge 2010 prove that our system has good quality in the naturalness, similarity. But in the intelligibility of the synthetic speech, the results are not good enough.

1. Introduction

USTC have been attending Blizzard Challenge since 2006. In 2006, we submit a statistical parametric speech synthesis system [1]. And as statistical parametric system [2] can't generate synthesis speech as natural as the best sentences synthesized by unit-selection systems [3], we start to develop HMM based unit-selection system since 2007 [4]. In the Blizzard Challenge 2007, a baseline HMM based unit-selection speech synthesis system using HMMs trained by acoustic features for phone unit selection is developed by USTC. The system performs well both in naturalness and similarity. In the Blizzard Challenge 2008 event, as a larger 15-hour UK database used, on the basis of the USTC unit-selection system, the decision tree scale is tuned manually according to the scale of the training database and to capture the variable speaking style of UK English [5]. Internal experiments show that a larger decision tree compared with the MDL [6] generated one leads to better synthesis speech quality, especially in prosody. In the Blizzard Challenge 2009, Cross-validation (CV) and minimal generation error criterion (MGE) [7] is introduced to optimize the scale of the decision tree automatically. States in HMMs other than phones are used as the basic unit for selection and concatenation in 1-hour speech synthesis system building task, and multi-Gaussian HMMs are employed in the 15-hour speech synthesis system building. This year in 2010, we use the same HMM based phone unit selection system for 5-hour database and state unit selection system for 1-hour database as last year, but more new techniques are tried during the task building. First, in order to improve the efficiency and reduce the footprint, tying all model covariance was used in spectrum model, F0 model. [8]. Secondly, we are trying sub-band waveform fusion with selected unit and the parameter synthesized speech.

This paper is organized as follows. Section 2 introduces the speech synthesis systems built for English tasks in Blizzard Challenge 2010, Section 3 includes the techniques used in the synthesis systems and the internal experiments conducted during the system building. And in Section 4, the Blizzard Challenge evaluation results for our system are listed and analyzed. At last, in Section 5, conclusions are made.

2. The English Tasks in Blizzard 2010

Blizzard Challenge 2010 English Evaluation consists of 5 sub-evaluations.

- ➢ EH1. 5-hour 'rjs' database synthesis system.
- ➢ EH2. 1-hour 'Roger' database synthesis system.
- ➢ ES1. 100 sentences from 'Roger' data synthesis system.
- ➢ ES2. speech in noise task, the same data as EH1.
- ➢ ES3. 48kHz synthesis system, the same data as EH1

'rjs' is new speaker for our system building, we spent 80 person-hours to check its ToBI break. Considering the database size (5 hours) is not big as usual to preceding year, its performance is very impressive. The subset with ARCTIC utterance from 'Roger' database (1 hour) has been used in EH2 for 3 years, we still want to improve the performance of our synthesis system in small size data at the level of 1-hour.

We built a phone-unit selection system for EH1, and a state-Unit selection system for EH2. For ES1, a HMM based adaptation system was built with the model trained by the data from 'rjs'. Because we do not have any information about the noise type, and it seems that the higher sampling rate data do not provide more benefit in system building. We offered the same result for ES2 as EH1, and for ES3, the 48k waveform data were concatenated by 16k system selection result.

Model covariance tying technique was used in unit selection system for EH1 and EH2. Sub-band waveform fusion as a new method was has been tried for EH2 and ES1. They will be introduced in detail below

3. Method

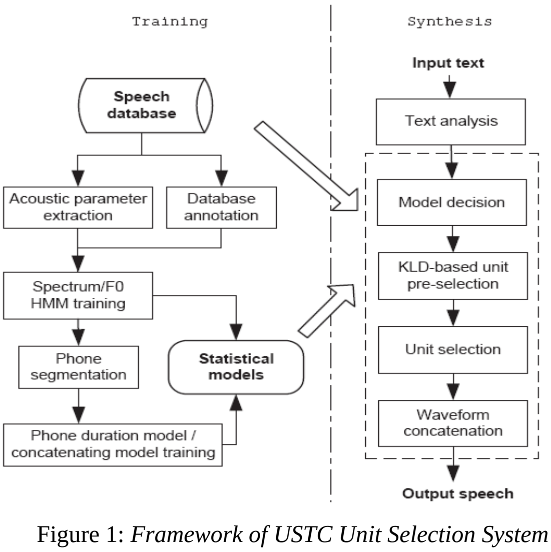

Figure 1: *Framework of USTC Unit Selection System*

3.1. Model training

The Framework of USTC Unit Selection system is shown in Fig.1. It includes two main steps to build a USTC HMM-based unit selection system. First, we should train HMM models[9] to guide unit selection. In the HMM model training part, acoustic parameters are extracted from the speech waveforms, The complete feature vector for each frame consists of static, delta and acceleration components of spectral parameters and logarithmized F0. With the segmental and prosodic data, the spectrum part is modeled by a continuous probability HMM and the F0 part is multi-space probability HMM (MSD-HMM) [10]. With the segmental and prosodic annotations of the database, the HMM models are context-dependent. Minimum description length (MDL) [6] based HMM model clustering is control the size of the decision tree. Then the phone boundaries of training utterances are determined by Viterbi alignment using the trained acoustic HMMs. Based on the phone segmentation, phone duration model, concatenating spectrum and F0 models are build to measure the smoothness at concatenated phone boundaries. For state-size unit system, concatenating models must be state concatenated level, too. Another long time pitch model is trained to guide the prosody between syllable unit.

3.2. Unit Selection

In the unit selection step, Kullback-Leibler divergence (KLD) [11] between the model of the candidate unit and the target model is used to conduct the unit pre-selection to reduce computational cost of dynamic programming search, we measure the KLD between the target unit and the each candidate unit to select the *K*-best units with minimum KLD before the calculation of target cost. Because the state observation PDFs of all contextual dependent HMMs are clustered using decision tree in our system, it can be calculated offline as a matrix for every two leaf nodes in the decision tree. Therefore the unit pre-selection step can be realized efficiently.

The optimal unit sequence is searched out from the speech database to maximize the likelihood of candidate feature sequences towards the target models and minimize the KLD between target and candidate models at the same time. the target models include spectral model, f0 model, duration model, concatenating model and long time pitch model. The weights between each model must be adjusted by manual operation patiently.

Finally, the waveforms of every two consecutive candidate units in the optimal unit sequence are concatenated to produce the synthesized speech. The cross-fade technique [12] is used here to smooth the phase discontinuity at the concatenation points of unit boundaries.

3.3. HMM based adaptation

In order to develop a 100 sentences voice speech synthesis system for ES1, we build a HMM based adaptation system with the model trained by the data from ‘rjs’. maximum likelihood linear regression (MLLR), and maximum a posteriori probability (MAP) [15] voice conversion method are used in this system.

3.4. Model covariance tying

Generally, bigger decision tree by context HMM clustering provide more precise target for unit selection, but it usually brings over-training and less-training problem. The difference between models could be three orders of magnitude of the value of covariance. That tuning a suitable MDL factor for HMM context clustering is difficult.

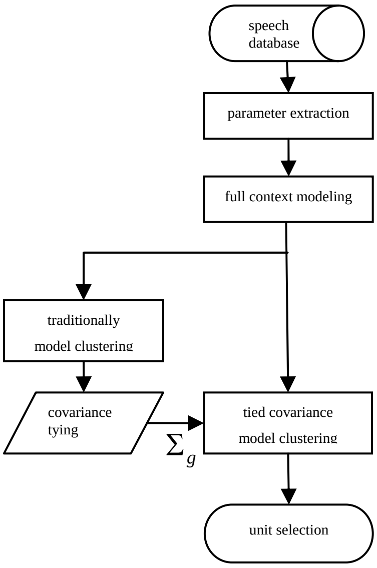

Figure 2: Framework of Convariance Tying Training

Enlightened by tying HMM covariance for parameter synthesis system [8], it could be supposed that the best target models should have the similar Gaussian distribution form in acoustic parameter vector space. If we tie the HMM covariance during context clustering, the target model for unit selection will be more reliable and robust.

As shown in Fig.2, the step to train a covariance tying model need twice clustering. In traditional ML criterion clustering, The total log likelihood of the Gaussian distribution of node S to the associated training data is calculated as

$$L(S) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in M_s}\gamma_m(t)\left\{n+\log(2\pi\left|\hat{\Sigma}_s\right|)\right\}$$

where T is the frames number of the training data, M_s is a set of HMM states (or streams), and $\gamma_m(t)$ is the posterior probability of an HMM state at the frame of t, Σ_s is the covariance matrix.

For covariance tying Training, the total log likelihood of the leaf node is calculated as follows:

$$L(S) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{m\in M_s}\gamma_m(t)\left\{Tr(\hat{\Sigma}_s\Sigma_g^{-1})+\log(2\pi\left|\hat{\Sigma}_s\right|)\right\}$$

Σ_g is a globally tied covariance. After the first step traditional clustering, we could tie a global variance and re-estimate the clustering model. It is provided as Σ_g at the second step covariance tying model clustering, make sure the twice size of decision tree context HMM clustering at the same level.

At the internal testing, covariance tying system play better than baseline system, The results are listed in table 1.

Although the promotion in the MOS score is small, on the other hand, model covariance tying could remove many processes in the algorithm calculation, and reduce the footprint at runtime significantly. It will be valuable in engineering application.

System	MOS
Phone unit Baseline	3.72
covariance tying	3.80

Table 1. *MOS score the baseline and proposed system for EH1*

3.5. sub-band waveform fusion

We used a simple method try to alleviate discontinuous issue in concatenative speech synthesis system. We observed that the main discontinuous issue among concatenative units existed in low frequency band and high frequency band, but the frequency band in middle range sounds continuous. Considering the waveforms generated from statistical parametric speech synthesis (SPSS) system always contains smooth parameter contour, we tried to replace the dis-continuous part in concatenatived speech with the continuous part in generated speech from SPSS.

We used simple filters to combine these two kinds of speech, in which pass band for concatenatived speech is middle frequency band, pass bands for generated speech are low and high frequency band, transition band of two filters are linear. After filtered these two kinds of speech with their own filter, the final speech is constructed by combining these two filtered speech on frequency scale. Note that generated speech contains same state alignment as concatenatived speech.

We conducted several experiments with different filter parameters and the conclusion is although the discontinuous issue can be solved by this method, the naturalness would be damaged by combination of two waveforms on frequency scale. Therefore, we did not use this method in final system. The internal test results are listed in table 2 and table 3.

System	MOS
state unit Baseline	3.36
sub-band waveform fusion	3.19

Table 2. *MOS score the baseline and proposed system for EH2*

System	MOS
adaptation Baseline	3.31
sub-band waveform fusion	2.67

Table 3. *MOS score the baseline and proposed system for ES1*

4. Evaluation

This section discusses the evaluation results of our system in Blizzard Challenge 2010. The identifier of USTC system assigned by the event organizer is “M”. System “A” is the natural speech used for reference. System “B”, ”C”, “D”, “E” are the benchmark systems provide by Festival and HTS.

4.1. Similarity test

The boxplots of similarity scores of all systems for EH1 and EH2 are shown in Fig.3 and 4. We can see that system “M” achieves the best similarity to original speaker for EH1, for EH2 it is the second by mean score rank. They are both the top group in similarity, have no significantly different from other system just like “J” and “V”. Our system got the high similarity score rank in ES1 and ES3, too.

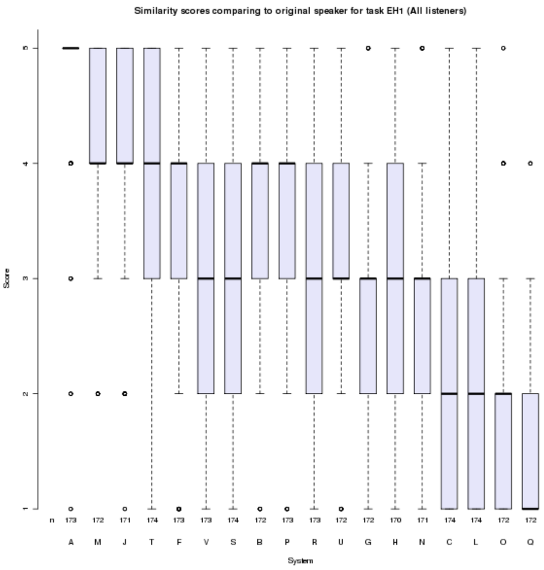

Figure 3: Boxplot of similarity scores for EH1

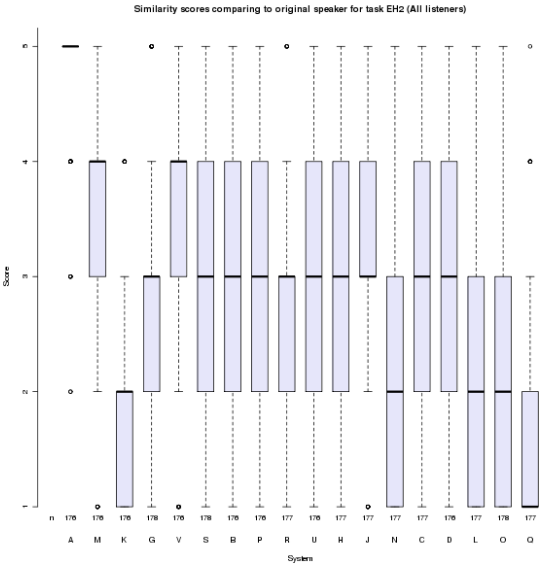

Figure 4: Boxplot of similarity scores for EH2

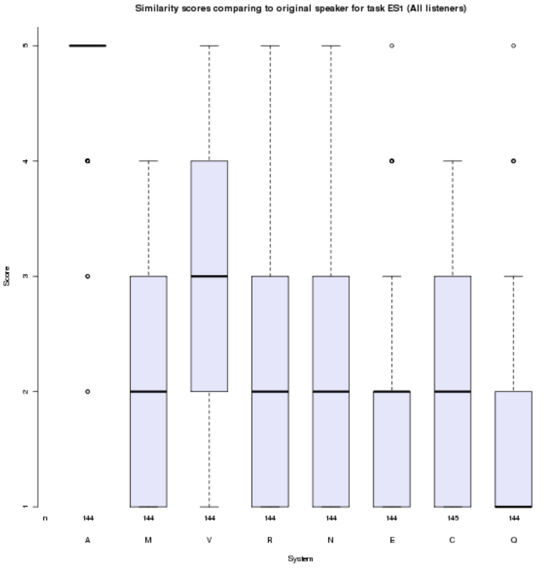

Figure 5: Boxplot of similarity scores for ES1

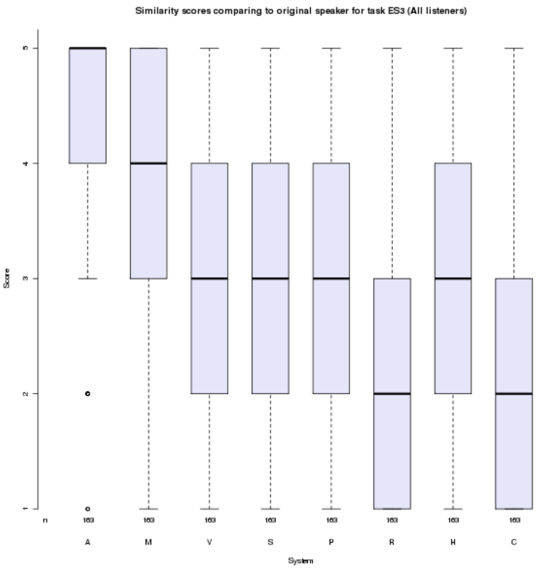

Figure 6: Boxplot of similarity scores for ES3

4.2. MOS test

The boxplots of mean opinion scores (MOS) of all systems for EH1 and EH2 are shown in Fig.7 and 8. We could find that our system is the best system on naturalness for both size of database, and we have significantly different from other synthesis system in Wilcoxon's signed rank tests for EH1 and EH2. The HMM based unit-selection systems built by USTC, are excellent with consistent performance in both size of database task. We got the first rank in ES1 and ES3 task, too.

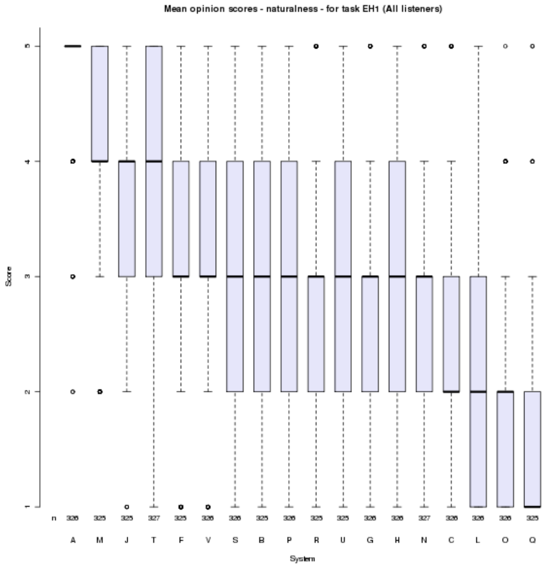

Figure 7: Boxplot of mean opinion scores for EH1

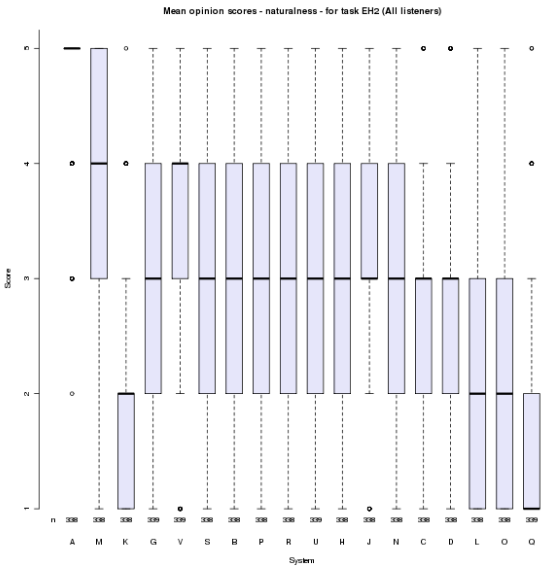

Figure 8: Boxplot of mean opinion scores for EH2

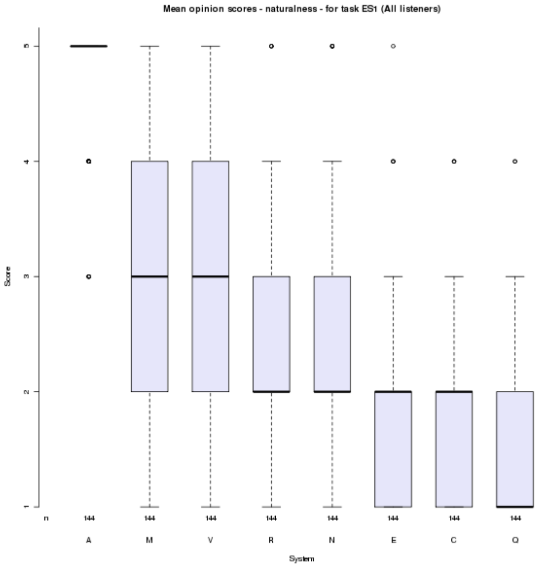

Figure 9: Boxplot of mean opinion scores for ES1

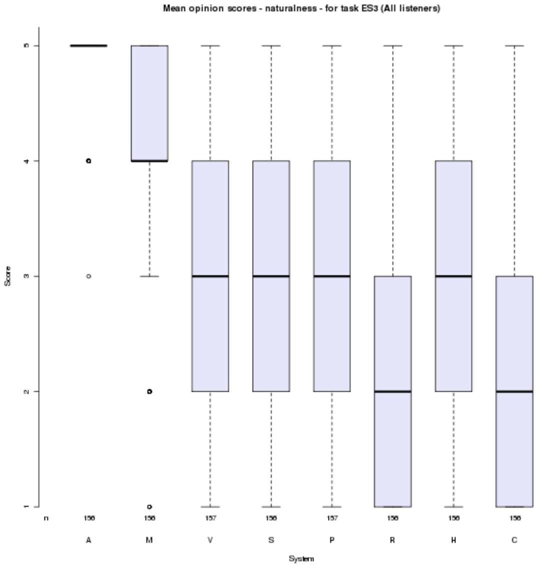

Figure 10: Boxplot of mean opinion scores for ES3

4.3. Intelligibility test

Fig.11 and 12 draw the results of word error rate (WER) test of all systems. The WERs of system “M” for EH1 and EH2 are at the middle rank by mean score. Although there are no significant differences between system “M” and the best systems, it is an unexpected result to us specially we have achieved much better rank before with the same database. It thinks like that the listening test for the semantically unpredictable sentences in BC2010 are restricted, everyone could plays the waveform just once. therefore, the score of WERs for all system are much worse than last year. For ES1, ES2 and ES3, The system “M” play better in rank of WERs by less systems.

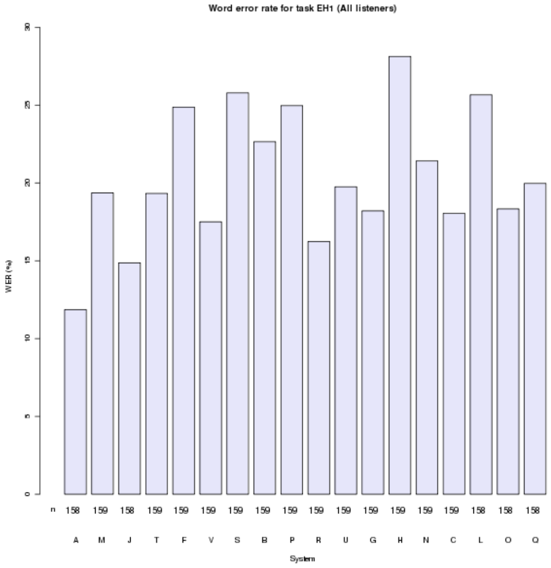

Figure 11: Mean WER for voice EH1

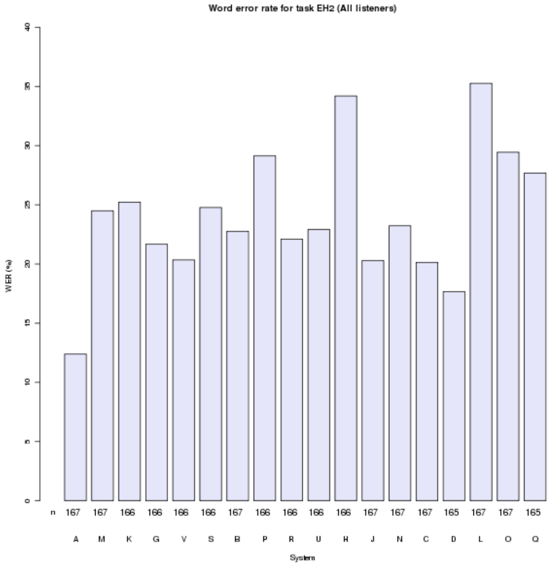

Figure 12: Mean WER for voice EH2

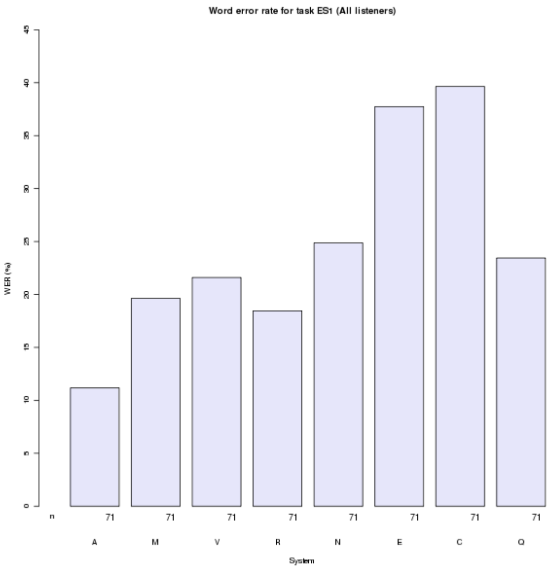

Figure 13: Mean WER for voice ES1

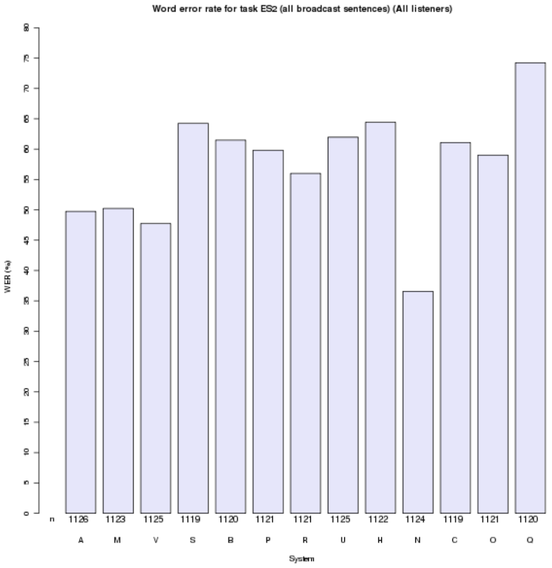

Figure 14: Mean WER for voice ES2.

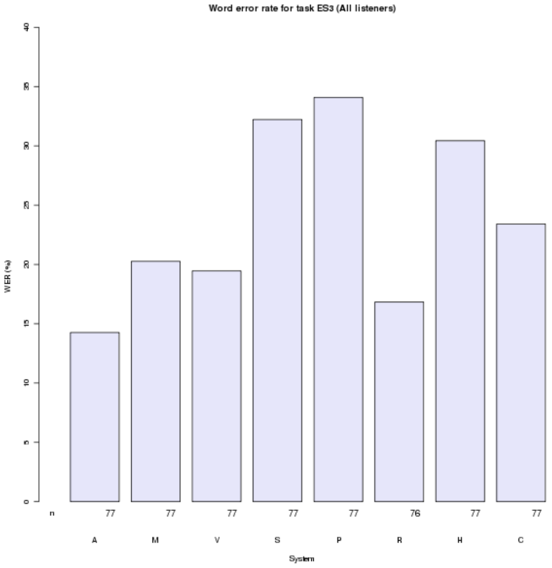

Figure 15: Mean WER for voice ES3.

5. Conclusions

This paper introduced the USTC speech synthesis system built for the Blizzard Challenge 2010. Comparing with the UTSC unit selection and waveform concatenation system, same new techniques are introduced to train the acoustic model for better performance and less footprint at runtime. The evaluation results show that, the USTC 2010 system performs well in the Naturalness, Similarity evaluations.

6. Acknowledgements

This work was partially supported by Hi-Tech Research and Development Program of China (Grant No. : 2006AA01Z137, 2006AA010104) and National Nature Science Foundation of China (Grant No. : 60905010). The authors also thank the research division of iFlytek Co. Ltd., Hefei, China, for their help in corpus annotation and providing the English text analysis tools.

7. 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] A. W. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” in *ICASSP*, vol. 4, 2007, pp. 1229–1232.

[3] Z. Ling and R. Wang,, “HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion,” in *ICASSP*, 2007, pp. 1245–1248.

[4] 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.

[5] Z. Ling, H. Lu, G. Hu, L. Dai, R.Wang, “The USTC system for Blizzard Challenge 2008,” in *Blizzard Challenge Workshop*, 2008.

[6] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” *J. Acoust. Soc. Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[7] Y. Wu, R. Wang , 2006b. Minimum generation error training for HMM based speech synthesis. In: Proc. ICASSP. pp. 89–92.

[8] Keiichiro Oura, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda, “Tying covariance matrices to reduce the footprint of HMM-based speech” in Interspeech, 2009, pp. 1759-1762.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis,” in Eurospeech, 1999, pp. 2347–2350.

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

[11] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Stat., vol. 22, pp. 79–86, 1951.

[12] Y. Wu, W. Guo , R. Wang , 2006. Minimum generation error criterion for tree-based clustering of context dependent HMMs. In: Proc. Interspeech. pp. 2046–2049.

[13] T. Toda, A.W. Black, K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2222-2235, Nov. 2007.

[14] K. Shinoda and C.-H. Lee, “Structural MAP speaker adaptation using hierarchical priors. ” In Proc. of IEEE Workshop on Speech Recognition and Understanding, 1997.

[15] O. Siohan, C. Chesta, and C.-H. Lee, “ Hidden Markov model adaptation using maximum a posteriori linear regression. ” In Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999.

[16] B. Langner and A. W. Black. Creating a database of speech in noise for unit selection synthesis. 5th ISCA Speech Synthesis Workshop - Pittsburgh, 2004.

[17] M. Cernak. Unit selection speech synthesis in noise. ICASSP 2006, 2006.