The USTC System for Blizzard Challenge 2008

Zhen-Hua Ling, Heng Lu, Guo-Ping Hu, Li-Rong Dai, Ren-Hua Wang

iFlytek Speech Lab, University of Science and Technology of China, Hefei, China
zhling@ustc.edu

Abstract

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2008. Two synthetic voices from the released UK English database are built using the HMM-based unit selection synthesis method, which is a hybrid of statistical parametric synthesis and unit-selection techniques. In this method, the optimal sequence of phone-sized candidate units is selected from the database following the statistical criterions derived from a set of trained HMMs for different acoustic features. Then the waveforms of selected units are concatenated to generate the synthesized speech. The evaluation results of Blizzard Challenge 2008 show that our system has good performance on similarity, naturalness and intelligibility for both English voices.

Index Terms: speech synthesis, Blizzard Challenge, unit selection, hidden Markov model

1. Introduction

The hidden Markov model (HMM)-based statistical approach had been widely used in speech recognition field. In recent years, it also made significant progress in speech synthesis [1][2]. In the Blizzard Challenge 2007 event [3], we developed two HMM-based speech synthesis systems. These two systems employed similar model training algorithms. At synthesis stage, one of them adopted parametric synthesizer to reconstruct speech waveform and the other one followed the unit selection and waveform concatenation approach [4]. The evaluation results indicated that the latter system can achieve better similarity and naturalness especially when the 8-hour full database was used, while the former system had better intelligibility [4]. The database released this year is a 15-hour UK English database. Considering the unit selection and waveform concatenation method can benefit more from the increasing of database size, we adopt the HMM-based unit selection method to build our voices this year.

This paper is organized as follows. Section 2 introduces the HMM-based unit selection synthesis method that we used to build the voices. In section 3, some details and experiments in system construction are described. Section 4 presents the evaluation results of our system in Blizzard Challenge 2008 and section 5 is the conclusion.

2. Method

2.1. Model training

The flowchart of UTSC system for Blizzard Challenge 2008 is shown in Fig.1. It can be divided into training stage and synthesis stage. At training stage, acoustic parameters, including spectrum and F0, are extracted from the speech waveforms of training database at first. Together with the segmental and prosodic annotations of the database, a set of context-dependent

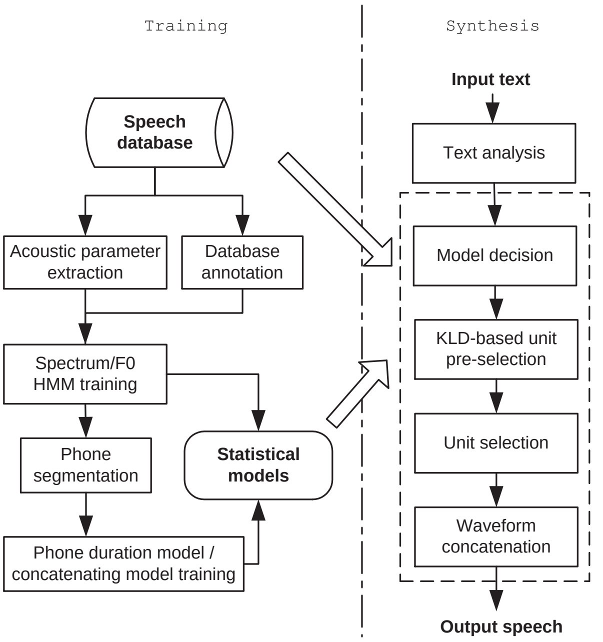

Figure 1: *Flowchart of the USTC system for Blizzard Challenge 2008.*

HMMs are estimated using the extracted acoustic features under maximum likelihood criterion [5]. The complete feature vector for each frame consists of static, delta and acceleration components of spectral parameters and logarithmized F0. The spectrum part is modeled by a continuous probability distribution and the F0 part is modeled by a multi-space probability distribution (MSD) [6]. The state transition probability matrices for all context-dependent HMMs with the same monophone label are tied. A decision-tree-based model clustering method is applied after context-dependent HMM training to deal with the data sparseness problem and predict the context-dependent model outside the training set. Minimum description length (MDL) [7] criterion is popularly used to 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 model and concatenating F0 model are trained. These models are also context-dependent and clustered using decision trees. The two concatenating models are introduced to measure the smoothness at concatenated phone boundaries in the synthesized speech. The features of these two models are defined as the differential of spectral parameters and F0 between the first frame of current phone and the last frame of previous phone [4]. Finally, five statistical models are estimated after model training: spectrum model, F0 model, phone duration model, concatenating spectrum model, and concatenating F0 model. These models are used to guide the selection of phone-sized candidate units at synthesis stage.

2.2. Unit selection

Several HMM-based unit selection speech synthesis methods have been proposed in our previous work [8][9][4]. In our system built for Blizzard Challenge 2008, phone-sized concatenation unit is used and the criterion combining likelihood and Kullback-Leibler divergence (KLD) [10] is employed . The optimal phone-sized 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.

Assume the number of phones in the utterance for synthesis is N and the contextual information of the target sentence is c. A candidate sequence of phone-sized units to synthesis this sentence is written as $\boldsymbol{U} = \{u_1, u_2, ..., u_N\}$. The contextual information of this candidate unit sequence is $c(\boldsymbol{U})$. Then the optimal sequence $\boldsymbol{U}^*$ is given by

$$\begin{aligned}\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{M} w_m[&\log P_{\Lambda_m}(X(\boldsymbol{U}, m)|c) \\ &- w_{KLD} D_{\Lambda_m}(c(\boldsymbol{U}), c)]\end{aligned} \quad (1)$$

where M denotes the number of trained models, which is 5 in our system according to the model training process discussed above; w_m means the weight for each model; w_{KLD} means the weight for KLD component; $X(\boldsymbol{U}, m)$ extracts the m-th features of the unit sequence $\boldsymbol{U}$; $\log P_{\Lambda_m}(\boldsymbol{X}|c)$ is the log likelihood function for observed feature $\boldsymbol{X}$ given model set Λ_m and contextual information c; function $D_{\Lambda_m}(c, c')$ calculates the KLD between two HMMs with contextual information c and c' given model set Λ_m. In practical implementation, the KLD function is approximated by its upper bound to simplify the calculation [11]. Eq.(1) can be rewritten into the conventional format of a sum of "target cost" and "concatenation cost" [4][9]. Then a dynamic programming search can be applied to find the optimal sequence conveniently.

In order to reduce the computation cost of dynamic programming search, a KLD-based unit pre-selection algorithm is applied [4]. For each candidate unit, we calculate KLD between the HMM it belongs to and the HMM of the target unit for synthesis. This KLD describes the similarity between the contextual factors of these two units, and does not rely on the acoustic features of candidate unit. Then the K-best candidate units with minimum KLD are selected for target cost calculation. Because the state observation PDFs of all context-dependent HMMs are clustered using decision tree in our system, the KLDs can be calculated in advance for every two leaf nodes in the decision tree. Therefore the unit pre-selection can be implemented efficiently at synthesis stage.

Finally, the waveforms of every two consecutive candidate units in the optimal phone 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 phone boundaries.

Table 1: *The number of phone transcription errors.*

Subset	*Detected by GPP*	*Verified manually*
ARCTIC	540	54
NEWS	1864	270

3. System construction

3.1. Database annotation

The quality of database annotation is important to the performance of a speech synthesis system, especially for the system using unit selection and waveform concatenation method. In Blizzard Challenge 2008, a 15-hour UK English speech database with segmental and prosodic annotation is released. However, the released annotations are not accurate enough because all of them are generated automatically and the prosodic features are given only based on the text. Therefore, we tried to improve the original annotation from three aspects: phone transcription, unit segmentation and prosodic annotation. Considering the limited time for voice building, we can only check and re-annotate part of the database. Besides, some automatic techniques were applied to assist the human labeling.

3.1.1. Phone transcription

The phone transcription is checked by two steps. At first, the original transcription is verified using a generalized posterior probability (GPP)-based confidence measure method [13]. Then the detected transcription errors are inspected manually. For each phone in the database with label p_0 and observation feature $\boldsymbol{O}$, we calculated the log likelihood ratio as a simplified version of GPP:

$$LLR = \log P(\boldsymbol{O}|p_0) - \max_{i=1,...,N} \log P(\boldsymbol{O}|p_i) \quad (2)$$

where $p_i, i = 1, ..., N$ are N hypotheses for phone p_0. Here we use all phones in the phone list that are different from p_0 as the hypotheses. The likelihood function in Eq.2 are calculated using context-dependent HMMs for spectrum and F0 features that are trained based on the original transcription. Then a threshold is applied and the phones with smaller log likelihood ratio than the threshold are detected as transcription errors for further manual inspection. Since time is limited, only the phone transcriptions in ARCTIC and NEWS subsets are checked through these two steps. The number of transcription errors detected by GPP and verified manually for the ARCTIC and NEWS subsets are shown in Table 1.

3.1.2. Unit segmentation

In the HMM-based unit selection synthesis method discussed above, the boundaries of candidate phone units used for concatenation are segmented by the trained HMMs at training stage. However, an accurate initial segmentation can help the performance of trained models. As shown in Fig.2, an iterative method is followed to generate the initial segmentation. At first, 200 utterances were selected from the ARCTIC subset and segmented manually. A seed model was trained using these utterances following the same context-dependent model training method discussed in section 2.1. Then we used the seed model to segment all utterances in the database by Viterbi alignment and re-estimated the HMMs using the entire database. Two iterations were carried out for the released database.

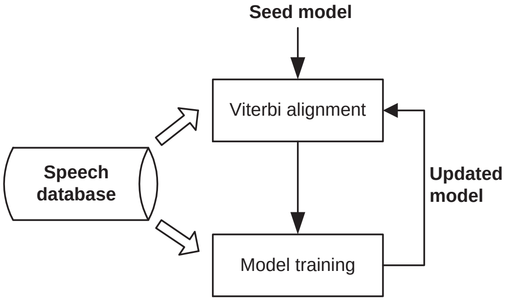

Figure 2: *Iterative unit segmentation.*

3.1.3. Prosodic annotation

Accurate prosodic annotation is essential for training a reliable prosody model. The released prosodic annotations, such as the phrase boundary, boundary tone and pitch accent, are predicted from the texts, not based on the recordings. Therefore we checked these prosodic labels of the database manually. Because of the limited time, only the corrections on the ARCTIC subset and a part of the NEWS subset are finished before the final model training.

3.2. Model training

Two UK English voices are required to submit in Blizzard Challenge 2008. For voice A, all utterances in the database were used to build our system. For voice B, only the ARCTIC subset was used. In model training, STRAIGHT [14] was used to extract F0 and spectrum from waveforms at 5ms frame shift. Mel-cepstrum was adopted to present the spectrum and the order was set to 13 (including 0-order). 5-state left-to-right without skip HMM structure was used to train context-dependent models. The question set in the decision-tree-based model clustering was composed of the following layers according to the available contextual information, which is similar to our previous systems [15][4]:

- Phone layer: the name and type of current and surrounding phones; the current vowel is reducible or not; the number and position of phones in syllable.
- Syllable layer: the stress, accent, and emphasis type of current and surrounding syllables; the number and position of syllables in word.
- Word layer: the POS of current and surrounding words; the number and position of words in phrase.
- Phrase layer: the number and position of phrases in utterance; the boundary tone of current and surrounding phrases.
- Utterance layer: the number of syllables, words and phrases in utterance.
- Subset layer: only for voice A, to indicate the subset which the sentence belongs to.

3.3. Test sentence synthesis

In order to synthesize the test sentences, contextual descriptions of these sentences are required as a precondition. In our system for Blizzard Challenge 2008, the phone sequences released with the test sentences were used. A English text analysis module developed by iFlytek company was adopted to predict the phrase boundary, boundary tone and pitch accent for the test sentences. Because we can not predict the subset layer feature from the text of test sentences, this feature was set as NEWS subset for all test sentences. In Eq.1, w_{kld} was set to 5, the model weights w_m for spectrum model, F0 model, phone duration model, concatenating spectrum model and concatenating F0 model were set as 1/39, 1/6, 2.5, 9.0 and 4.5 respectively after a few manual tuning. In unit selection, the 200-best candidates were preserved after KLD-based unit pre-selection and the 50-best candidates with smallest “target cost” were used for the “concatenation cost” calculation and dynamic programming search.

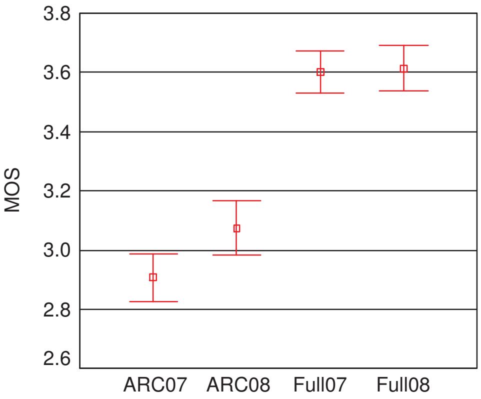

Figure 3: *The MOS (central square) and standard error of the mean (extended line) of the USTC BC07 and BC08 systems using ARCTIC subset and full database.*

3.4. A comparison between the BC07 and BC08 systems

In order to compare the performance of our system developed last year for Blizzard Challenge 2007 and the system we developed this year, a subjective evaluation was carried out. Both the voices from ARCTIC subset and the full database were tested. 20 sentences from last year’s test set were synthesized by each system and evaluated by six English-native listeners. The mean opinion score (MOS) and the standard error of the mean for each voices are shown in Fig.3. It is worth noting that the “Full08” voice was built using only the ARCTIC and NEWS subsets of this year’s database, which was smaller than the database of “Full07”. From this figure, we can see that the BC08 system is slightly better than the BC07 system for both voices. The difference between the voices from ARCTIC subset is more significant. Actually, the same approach is followed for building these two systems. After a discussion with the listeners, we think this difference may be attributed to the reading style and UK accent of the database this year, which are preferred by most listeners.

3.5. Experiments on model clustering

Decision-tree-based model clustering is an important component in the model training algorithm. During the procedure of voice building for Blizzard Challenge 2008, two experiments are carried out to test different model clustering strategies.

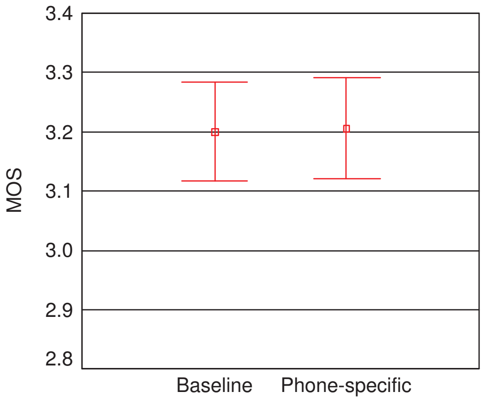

Figure 4: *The MOS (central square) and standard error of the mean (extended line) of the baseline system and phone-specific clustering system for voice B.*

3.5.1. *Phone-specific model clustering*

In the model training method discussed in section 2.1, a unique decision tree is built to cluster the PDFs of one state for all context-dependent HMMs. While, at synthesis stage the selected optimal candidate unit must have the same phone label as the target unit. That means we can also train phone-dependent models for unit selection. A system with phone-specific model clustering was constructed for voice B of Blizzard Challenge 2008. In this system, separated decision trees were trained for different phones. It was compared with the baseline system by a subjective evaluation. 15 sentences were synthesized using both systems and presented to six English-native listeners. The results are shown in Fig.4. The baseline system is the same as “ARC08” in Fig.3, while they have different scores because the stimuli and subjects are different in these two experiments. Fig.4 indicates that there is not significant difference between the baseline system and the phone-specific model clustering system. Finally, the phone-specific model clustering method is not adopted in our submitted system. However, this method has other advantages. For example, it can decrease the size of each trained decision tree effectively and is potential in reducing the memory space for pre-calculated KLDs and speeding up the calculation of model likelihood combining some vector quantization techniques.

3.5.2. *The effect of MDL factor*

Conventionally, MDL criterion is followed in model clustering to control the size of trained decision tree. The MDL factor is set as 1.0 by default and decreasing this value leads to a larger decision tree [7]. A subjective evaluation was made to test the influence of modifying the MDL factor. 15 sentences were synthesized by the voice B system with a MDL factor of 1.0, 0.5 and 0.1 for the clustering of spectrum and F0 models respectively. These sentences were evaluated by six English-native listeners. The evaluation results are shown in Fig.5. From this figure, we can see the naturalness of synthesized speech can be improved significantly when increasing the size of the decision tree. The leaf node number of the decision tree for clustering spectrum models increased from about 2,000 to more than 6,000 when decreasing MDL factor from 1.0 to 0.1 in this experiment. Finally, the MDL factor was set as 0.1 for the training of context-dependent spectrum and F0 models in our submitted system. How to optimize the size of the decision tree will be a topic of our future work.

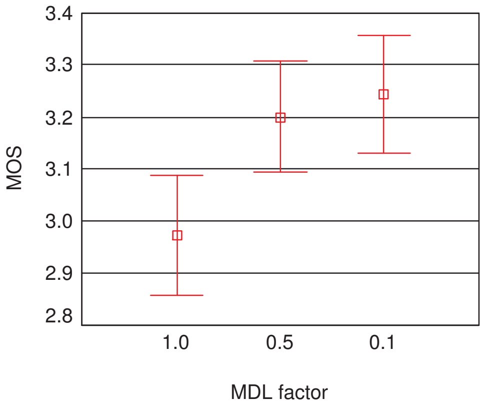

Figure 5: *The MOS (central square) and standard error of the mean (extended line) of the systems with different MDL factor for voice B.*

Table 2: *Results of Wilcoxon's signed rank tests. The system identifiers listed here are the systems that are NOT significantly different from system J.*

Test	*Voice*	*Entries*
Similarity	A	N/A
	B	N/A
MOS	A	N/A
	B	S V
WER	A	B C E I K M P S T V
	B	B C D F G I M Q S T V

4. Evaluation

This section discusses the evaluation results of our system in Blizzard Challenge 2008. The identifier of USTC system assigned by the event organizer is “J”. System A is the natural speech used for reference. System B and C are the Festival and HTS benchmark systems.

4.1. Similarity test

The boxplots of similarity scores of all systems for voice A and B are shown in Fig.6 and 7. From these figures, we can see that system J achieves the best similarity to original speaker for both voices. Table 2 gives the results of Wilcoxon's signed rank tests to determine whether the difference between two systems is significant. It can be found that the difference between system J and any other participant systems on similarity is significant for both voice A and B. The high similarity score of our system can be attributed to the unit selection and waveform concatenation synthesis approach where no signal processing is applied and the statistical criterion for unit selection which employs models of different acoustic features trained on the specific speaker's database.

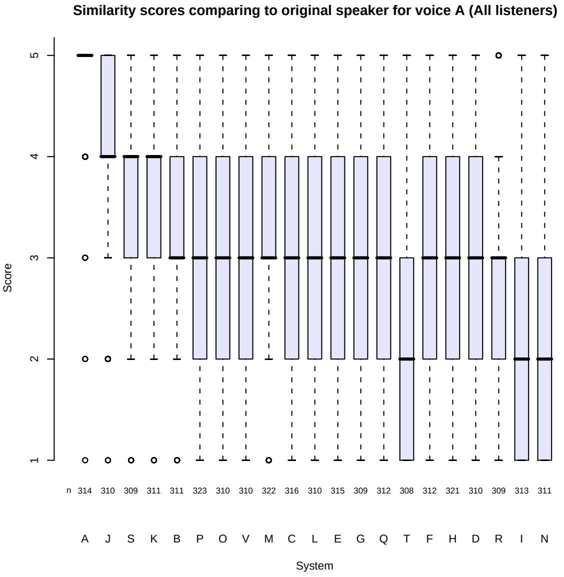

Figure 6: *Boxplot of similarity scores for voice A.*

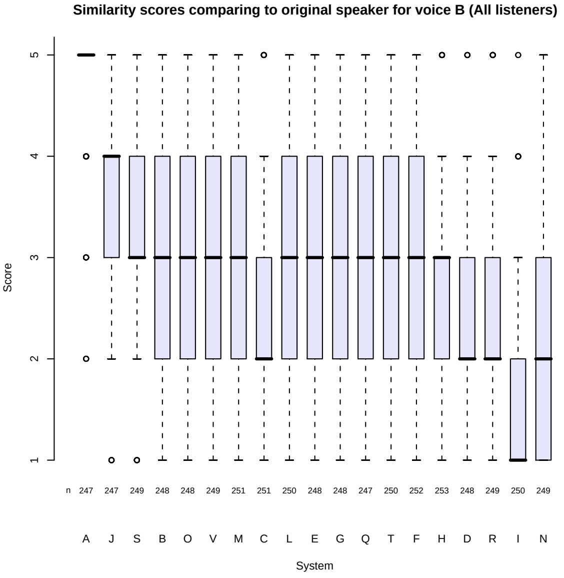

Figure 7: *Boxplot of similarity scores for voice B.*

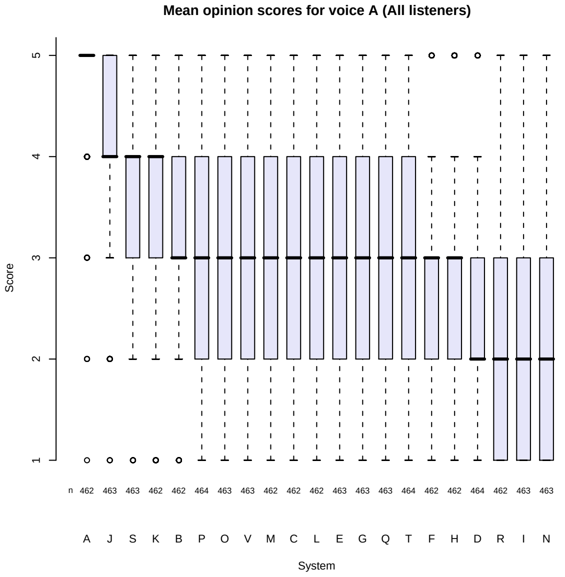

Figure 8: *Boxplot of mean opinion scores for voice A.*

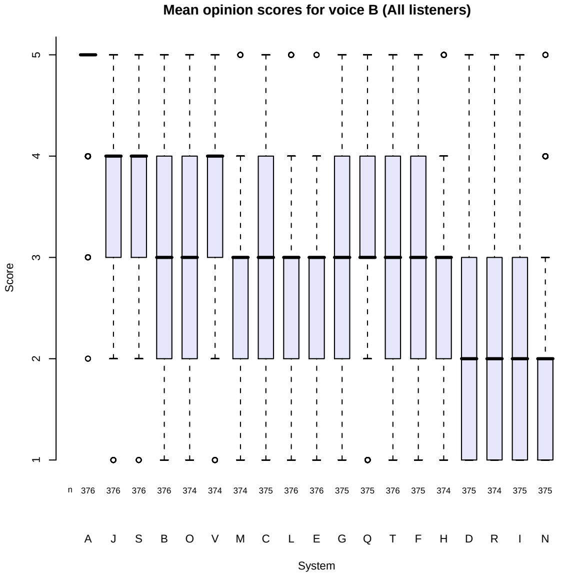

Figure 9: *Boxplot of mean opinion scores for voice B.*

4.2. MOS test

The boxplots of mean opinion scores (MOS) of all systems for voice A and B are shown in Fig.8 and 9. Combining these two figures with Table 2, we can find that our system is the best system on naturalness for voice A and one of the best systems for voice B. This proves the effectiveness of proposed HMM-based unit selection synthesis method. Considering the fact that the released UK English database this year contains richer reading styles and more prosodic fluctuation than previous years, it is difficult to design the cost functions for unit selection based on only phonetic rules or simple prediction of prosodic features and considering no or little acoustic features of the candidate units. The HMM-based unit selection method evaluates the suitability of each candidate unit sequence by measuring their acoustic features under statistical framework. Hence, the cost functions derived from Eq.1 are more accurate and compatible with the given database.

4.3. Intelligibility test

Fig.10 and 11 draw the results of word error rate (WER) test of all systems. The WERs of system J for both voices are not the lowest, but Table 2 tells us that the differences between system J and the best systems are not significant. In the last Blizzard Challenge event, we compared the intelligibility of HMM-

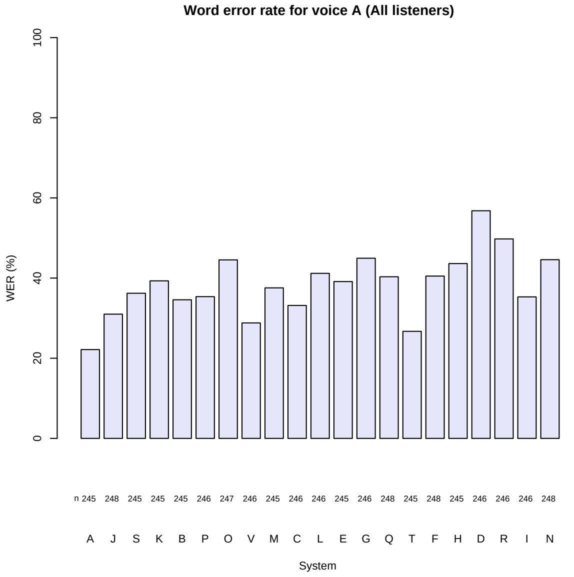

Figure 10: *Mean WER for voice A.*

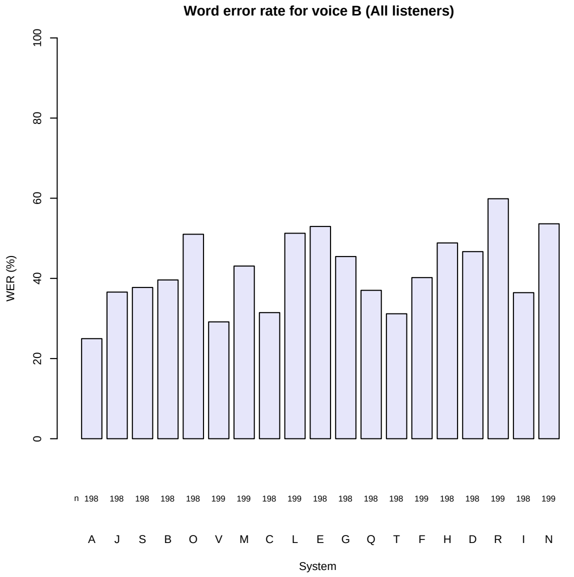

Figure 11: *Mean WER for voice B.*

based parametric synthesis method and HMM-based unit selection method. We found that the former one can achieve better intelligibility for the semantically unpredictable sentences used in the WER test of the Blizzard Challenge events [4].

5. Conclusions

This paper introduces the USTC system for Blizzard Challenge 2008. We built and submitted two UK English voices using the HMM-based unit selection synthesis method. Before model training, the annotation of released database were partly checked and corrected combining automatic and manual approaches. During the procedure of voice building, some experiments were made to test different model clustering techniques. Our internal evaluation indicates that the naturalness of synthesized speech can be improved if the sizes of the decision trees for model clustering are increased. The evaluation results of Blizzard Challenge 2008 shows that our system is one of the best systems on similarity, naturalness and intelligibility for both English voices.

6. Acknowledgements

This work was partially supported by Hi-Tech Research and Development Program of China (Grant No.: 2006AA01Z137, 2006AA010104) and National Natural Science Foundation of China (Grand No.: 60475015). The authors 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] K. Tokuda, H. Zen, and A. W. Black, “HMM-based approach to multilingual speech synthesis,” in *Text to speech synthesis: New paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004.

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

[3] M. Fraser and S. King, “The Blizzard Challenge 2007,” in *Blizzard Challenge Workshop*, 2007.

[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] 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.

[6] 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.

[7] 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.

[8] Z. Ling and R. Wang, “HMM-based unit selection using frame sized speech segments,” in *Interspeech*, 2006, pp. 2034–2037.

[9] ——, “HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion,” in *ICASSP*, 2007, pp. 1245–1248.

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

[11] P. Liu and F. K. Soong, “Kullback-Leibler divergence between two hidden Markov models,” Microsoft Research Asia, Tech. Rep., 2005.

[12] T. Hirai and S. Tenpaku, “Using 5 ms segments in concatenative speech synthesis,” in *5th ISCA Speech Synthesis Workshop*, 2004, pp. 37–42.

[13] L. Wang, Y. Zhao, M. Chu, F. K. Soong, and Z. Cao, “Phonetic transcription verification with generalized posterior probability,” in *Eurospeech*, 2005, pp. 1949–1952.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[15] 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.