The USTC System for Blizzard Challenge 2009

Heng Lu, Zhen-Hua Ling, Ming Lei, Cheng-Cheng Wang, Huan-Huan Zhao, Ling-Hui Chen, Yu Hu, Li-Rong Dai, Ren-Hua Wang

iFlytek Speech Lab, University of Science and Technology of China, Hefei, China

luhenglh@mail.ustc.edu.cn

Abstract

This paper introduces the USTC's speech synthesis system for Blizzard Challenge 2009. USTC attended all English tasks including the hub tasks and the spoke tasks. According to the various conditions for different tasks, different versions of HMM based unit-selection systems are constructed based on the USTC Blizzard Challenge 2008 system. 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 2009 prove that our system has good quality in all the naturalness, similarity and intelligibility of the synthetic speech.

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. This year in 2009, many new techniques are introduced. Firstly, other than tuning the decision tree scale manually, a method using cross-validation (CV) and minimal generation error criterion (MGE) [7] is introduced to optimize the scale of the decision tree automatically. Secondly, in order to solve the lack of the suitable phone unit problem in 1-hour speech synthesis system building task, states in HMMs other than phones are used as the basic unit for selection and concatenation. Thirdly, in order to further capture the variable prosody in UK English, multi-Gaussian HMMs are employed in the 15-hour speech synthesis system building. At last, GMM-based voice conversion and HMM adaptation method, speech intelligible index (SII) feature [8], and the emphasis automatic labeling are introduced in Spoke Tasks 1,2 and 3 separately.

This paper is organized as follows. Section 2 reviews the USTC 2008 unit-selection system. Section 3 introduces the speech synthesis systems built for EH1, EH2, ES1, ES2, ES3 tasks in Blizzard Challenge 2009, including the new techniques added to the 2008 system and the internal experiments conducted in 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 USTC Blizzard Challenge 08 system

The USTC Blizzard Challenge 2008 system consists of two main parts, the HMM model training part [9] and the speech synthesis part. In the HMM model training part, acoustic parameters are extracted from the training data, including spectral and prosody features. Then the spectrum part is modeled by a continuous probability HMM and the F0 part is modeled by a multi-space probability HMM (MSD-HMM) [10]. As there are enormous combinations of context features, minimum description length (MDL) [6] based HMM model clustering is conducted to avoid data sparse problem and, at the same time, to predict models for the text to be synthesized. Phone duration model is also trained to model the duration of phone. Apart from the acoustic models mentioned above, spectral and F0 concatenating models are introduced to measure the smoothness at the concatenated phone boundaries in the synthesized speech. In the synthesis part, five statistical models, including the spectrum model, F0 model, phone duration model, concatenating spectrum model and concatenating F0 model are employed to choose the most suitable unit from the training database. Then phone unit are concatenated to synthesize speech waves. The framework of USTC 2008 system is shown in Fig. 1.

In the unit selection process, Kullback-Leibler divergence (KLD) [11] between the model of the candidate unit and the

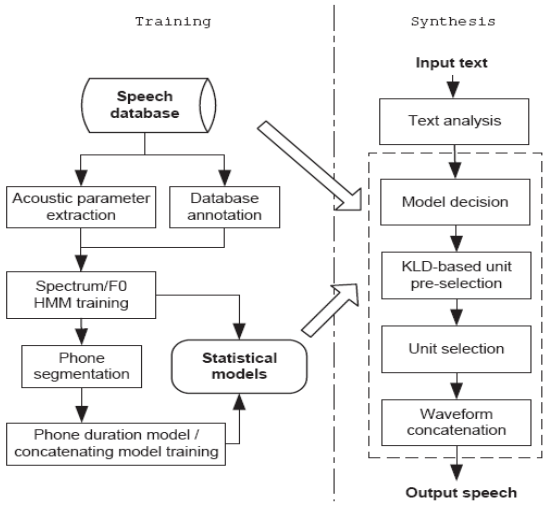

Figure 1: *Framework of USTC 2008 system*

10.21437/Blizzard.2009-16

target model is used to conduct the unit pre-selection to reduce computational cost in the Viterbi unit-selection process. After pre-selection, a function consisting of "target" cost and "concatenation" cost are maximized to select the optimum unit sequence.

3. The USTC Blizzard Challenge 09 system

Blizzard Challenge 2009 English Evaluation consists of 5 sub-evaluations.

- EH1. 15-hour speech synthesis system building.
- EH2. 1-hour speech synthesis system building.
- ES1. 10, 50, 100 sentences voice conversion.
- ES2. Telephone transmitted system building.
- ES3. Man machine conversation system building.

The EH1, EH2 evaluation is the same with Blizzard Challenge 2008, but ES1, ES2 and ES3 sub-evaluations are new ones. In Blizzard Challenge 2009, we build systems for these 5 sub-evaluations separately. USTC EH1, EH2, ES2, ES3 systems are the modified versions on the base of USTC 2008 system, including new methods and techniques. And USTC ES1 system is a voice conversion system especially built for the task, with different framework from USTC 2008 system.

Systems for each sub-evaluation are introduced as follows, including construction method and experiments in the system building process.

3.1. EH1 task

We built 3 systems for EH1 task. (1) the USTC 2008 system as our baseline system. (2) a system with optimized HMM clustering decision tree that is pruned using minimum "cross" generation error criterion in the HMM clustering process on the base of the baseline system. (3) the multi-Gaussian HMM system, which is our submitted system in EH1 task. The USTC 2008 system has been introduced in section 2. And we introduce (2) and (3) in the following part.

3.1.1. Decision tree pruning system

In our HMM based unit-selection system, context dependent phone is used as the basic unit for model training and unit-selection. As there are enormous combinations for the context features, minimum description length (MDL) criterion based HMM model clustering is conducted to avoid the data sparse problem and, at the same time, to predict models for the synthesis voice. However, in actual speech synthesis system construction, a larger decision tree is always observed to give better performance than the MDL criterion generated one. So in our USTC 2008 system, compared with USTC 2007 unit-selection system, MDL factor on the spectral and F0 HMM clustering was set to 0.1 other the default value 1 to generate larger decision tree. In 2008, MDL factor was set according to our subjective listening tests, and in 2009, we propose to choose the MDL threshold parameters automatically according to objective distortion.

As there is mismatch between HMM model training or clustering and parameter generation, minimum generation error (MGE) criterion is proposed in parameter generation, HMM model training [7] and model clustering [12]. However, MGE can't control the scale of the decision tree, and it only employs the minimum generation error other than the maximum likelihood (ML) criterion to choose the optimum question for the splitting of each tree node. But the scale of the MGE generated one is controlled to as large as the ML based decision tree as in [8]. Other than MGE, Cross-validation (CV) is proved to be an effective way to avoid the model over-training and less-training problem. In the system construction, we proposed a CV based MGE criterion, the minimum "cross" generation error (MCGE) criterion to prune the decision tree.

In the "cross" generation error computation, training data set is divided into M sub-sets, here we set $M=10$. And for each time in a total of M times, $M-1$ sub-sets are employed to train the HMM parametric speech synthesis system, and the other one sub-set is utilized for synthesizing. By computing the average generation error for all the sentences over the whole training database, the "cross" generation error is obtained. The computation process is shown in figure 2.

Other than the two commonly used methods for decision tree generation: from root to leaves or from the bottom to root, we generating the decision tree from the middle. By first tuning the MDL threshold parameter using the MCGE criterion, we are able to initialize an optimal decision tree with minimal "cross" generation error. Then, "cross" generation error for each decision tree leaf of the initialized decision tree is inspected for further pruning.

In MDL based HMM clustering, ML criterion is used to choose the optimal question for each tree node, and MDL is used as the stopping criterion. For one node, for example node S_m, if all questions in the question set fulfill the following equation (1)

$$\frac{1}{2}\Gamma_m \log\left|\Sigma_m\right| - \frac{1}{2}\Gamma_{mqy} \log\left|\Sigma_{mqy}\right| - \frac{1}{2}\Gamma_{mqn} \log\left|\Sigma_{mqn}\right| < \alpha K \log W \quad (1)$$

then the splitting of node S_m stops. In equation (1), Γ_m, Σ_m, Γ_{mqy}, Σ_{mqy}, Γ_{mqn}, Σ_{mqn} indicate the occupation probability and the covariance matrix for node S_m, S_{mqy} and S_{mqn} separately, where S_{mqy} and S_{mqn} are two child nodes for the possible splitting of S_m. MDL factor α is parameter controls the scale of the decision tree. A larger α leads to smaller tree, and a smaller α leads to a larger one. Our

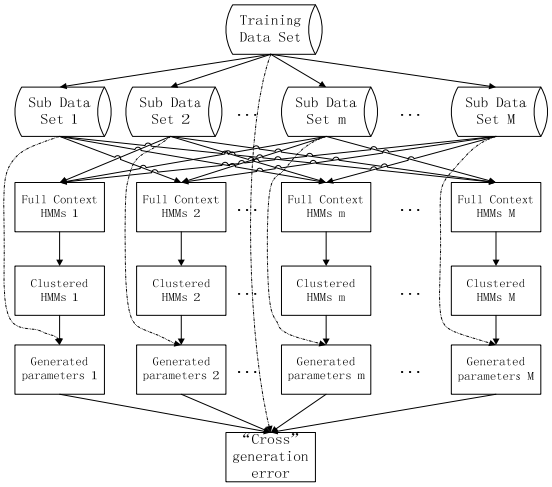

Fig. 2: *"Cross" generation error computation.*

experiments by tuning α on spectral decision tree is shown in table 1.

Table 1. *Spectral feature "cross" generation error change according to MDL factor* α

MDL factor α	"cross" generation error
2	0.21414
1	0.19613
0.5	0.18899
0.1	0.18424
0.05	0.18424
0.01	0.18424
0.005	0.18424

From table 1, we can see the "cross" generation error changes according to α . As α becomes smaller, decision tree becomes larger. But there are other factors that control the scale of the decision tree, for example, the minimum sample number per node. However, we tune only α . So the scale of the decision tree stops becoming larger when α reaches 0.1, and the "cross" generation error stops changing too. At last, in system construction we choose both spectral and F0 MDL factor $\alpha = 0.1$, which is the same with the MDL factors in our 2008 system.

After deciding the optimum MDL factor α , we start to prune the decision tree initialized by the optimum MDL factor α . Only the spectrum decision tree is pruned. Using $\alpha = 0.1$, a decision tree is generated by MDL based context dependent HMM clustering. By inspecting the "cross" generation error for each node of the initialized decision tree, we are able to further back-off or split the tree node to reduce "cross" generation error. The decision tree pruning steps are:

- **Step 1.** Other than using the same MDL factor α in the M times of full context dependent HMM clustering in the MDL factor tuning cross validation process, the same initialized decision tree given in the MDL factor tuning step is applied to M times HMM model clustering. No ML or MDL criterion but the initialized decision tree is employed in the CV process. It is possible that certain $M-1$ sub-databases lack the samples necessary in the initialized decision tree based HMM clustering. Then the initialized decision leaves which lack samples in certain $M-1$ sub-databases are backed-off with its brother leaves until there are enough samples for these leaves. If the initialized decision tree is called $Tree_0$, then decision tree $Tree_1$ is obtained from $Tree_0$ by the pre-process described above. Then $Tree_1$ is applied to M times HMM model clustering. Set $i = 1$.
- **Step 2.** Backing-off all the leaves in $Tree_i$ to their father node by one tree level to get $Tree_i'$. And for each node in decision tree $Tree_i'$, take node n for example, we compare the "cross" generation error C_n' of node n from $Tree_i'$ and the average "cross" generation error C_n by its two according child leaves in $Tree_i$. If C_n is larger than C_n' , then the two according child leaves in $Tree_i$ is backed-off to their father node. Otherwise they are kept stable. Then $Tree_{i+1}$ is obtained from $Tree_i$. Set $i = i+1$.
- **Step 3.** Step 2 continues until little leaf in $Tree_i$ can be backed-off.
- **Step 4.** Splitting is conducted the similar way after the backing-off process is done.

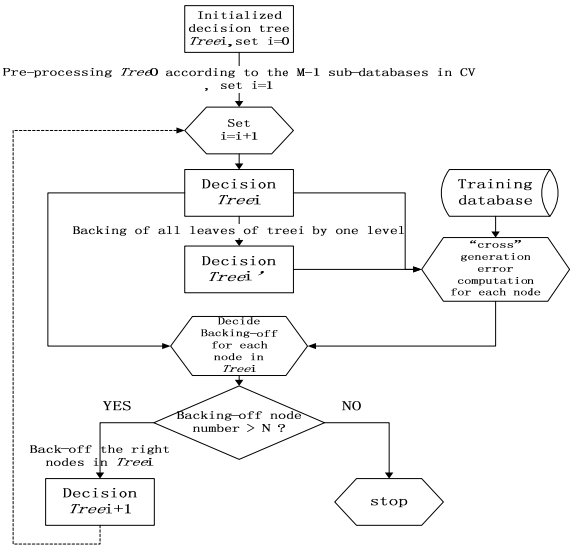

Fig. 3. *Flowchart for backing-off each leaf in initialized decision tree.*

The flowchart for decision tree backing-off and splitting is shown in figure 3. By MCGE based decision tree pruning, we are able to generate more accurate HMM model, then we use these HMM models to conduct unit-selection based speech synthesis.

3.1.2. Multi-Gaussian HMM based unit-selection system

With the same purpose of building more accurate HMMs to model the variances in the UK English prosody as the decision pruning system, we build a multi-Gaussian HMM model based unit-selection speech synthesis system based on our 2008 system. Compared with the USTC 2008 single Gaussian HMM based unit-selection system, in 2009, we use the 4-Gaussian HMMs to model spectral and F0 feature, but the duration model and the spectral and F0 concatenation models remain single Gaussian HMMs.

3.1.3. Experiments

We employ these three systems to synthesis Blizzard Challenge 2008 test sentences. A total of 24 sentences, including 8 sentences for news, conversation, and novel each, are synthesized and tested. Two native English speakers are asked to give MOS score. The results are listed in table 2.

From the result, we can conclude that both the decision tree pruning system and multi-Gaussian system out-perform the USTC 2008 baseline system. At last, we choose multi-Gaussian system as our EH1 system, but we use decision tree pruning in our submitted EH2 task system.

Table 2. *MOS score for the three EH1 task systems*

System	MOS
USTC 2008 system	3.24
Decision tree pruning system	3.45
Multi-gaussian system	3.45

3.2. EH2 task

The main reason for the 15-hour system out-performing the 1-hour system is that, there are more suitable samples to select in a larger training database. So other than using context dependent phone as our basic unit in the unit selection system, we use context dependent HMM state as our basic unit. The advantage for smaller basic unit is the comparatively larger numbers of candidate unit in the training database. And smaller basic unit also can lead to larger overall observation to HMM model likelihood. However, the disadvantage is obvious too. More concatenating point may harm the fluency of the synthesis speech. In our EH2 system, we employ the combination of state concatenation and decision tree pruning technique on the base of the USTC 2008 unit-selection system. At first, full context dependent HMMs are trained, then decision tree pruning technique is employed to optimize the decision tree and the clustered HMM model to better model the training data. In the synthesis stage, HMM models are used to conduct the HMM state based unit-selection and concatenation. The same multi-Gaussian HMM system as in EH1 task is also built for EH2 tasks and tested.

Three systems are built for EH2 task and tested. (1) The USTC 2008 system as the baseline system. (2) multi-Gaussian HMM unit-selection system the same with EH1 task. (3) Decision tree combining with HMM state selection and concatenation system. Five native English speakers are asked to give MOS score on the 24 sentences synthesized by the three systems. The result is shown in table 3.

Table 3. *MOS score for the three EH2 task systems*

System	MOS
USTC 2008 system	3.52
Multi-Gaussian HMM system	3.44
Decision tree pruning combining state selection	3.61

The result indicates that, the multi-Gaussian HMM system may not be a good choice in the 1-hour EH2 task. Because there are not enough sample for large HMM model, the multi-Gaussian model may be over-training. We choose the decision tree pruning combining HMM state unit selection system as our EH2 system.

3.3. ES1 task

Unfortunately, in the ES1 task, one big problem for us is that we do not have another male UK English training database to train the models of the source speaker. What we have is only a female, American style English database. So we use the female, American English speaker Catherine as our source speaker. In the voice transformation system building, we first transform the American English labels to the UK English labels by mapping the American English phones to Unilex style ones. Then voice conversion system is trained for each 10, 50 and 100 sentences.

For 10 sentences voice conversion, since the target sentence number is small, we use the GMM voice conversion method [13]. And for 50, 100 sentences voice conversion, maximum likelihood linear regression (MLLR), structure maximum a posteriori probability (SMAP) [14], and maximum a posteriori probability (MAP) [15] voice conversion method are conducted all together one by one to increase the similarity of the transformed speech.

3.4. ES2 task

In the ES2 task, our analysis shows that, after transmitting through phone channel, the speech quality degenerates in the several aspects, including the more buzzy noise and the narrowed band spectrum. And these changes harm people's understanding of the voice. In order to increase the intelligibility of the telephone channel transmitted voice, we include the Speech Intelligibility Index (SII) in our unit-selection speech synthesis system. SII is a measure of the speech intelligibility in different noise environment and circumstances standardized by ANSI [8]. Before the SII calculation, the speech and noise are analyzed by 1/3 octave filter-bank with the central frequency at 160, 200, 250, 315, 400, 800, 1000, 1250, 1600, 2000, 2500, 3150, 4000, 5000, 6300 and 8000 HZ. From the filter-bank, 18 speech parameters and 18 noise parameters are generated to be used as the input of SII calculation. The SII calculation utilizes the configuration "standard speech, 1/3 octave critical frequency and default hearing threshold". The result of the SII calculation is a real number in range of 0.0 to 1.0. All the context dependent phones in the training set have their according SII calculated with the actual noise generated by the phone channel simulation tools. And the SII calculation is offline. After SII for each phone is calculated, we use minus SII as a cost added to our USTC 2008 15-hour unit-selection system's cost function. Thus the ES2 system is a SII modified version of our EH1 system. The SII system considers not only target cost, concatenation cost, but also the SII. Unit with higher SII score is preferred on the unit-selection process [16][17].

We have 2 native English speakers to take a subjective listening test to give the preference score (According to the overall speech quality, including naturalness and intelligibility). Three systems are tested, (1) the USTC 2008 15-hour baseline system, (2) SII modified version on the base of USTC 2008 system with SII weight 1, (3) SII modified version on the base of USTC 2008 system with SII weight 5. 30 sentences synthesized by each of the three systems are transmitted by the developing telephone channel tools and tested. For every sentence, the most preferable system gets 3 points, the second 2 and the least 1 point, the same score for different systems is permitted. The test result is as follows in table 4.

Table 4. *Preference score for the three ES2 task systems*

System	Preference Score
USTC 2008 system	1.68
SII weight 1 system	1.99
SII weight 5 system	2.12

From the result, we can conclude that both the two SII modified systems outperform the 2008 USTC baseline system in the overall speech quality. Comparing the speech voices by SII system and the baseline system, one can easily find that, SII modified system tends to choose the unit with higher pitch, which may cause the voices to be more understandable in the telephone channel noisy environment. Though the SII weight 5 system get higher preference score than the SII weight 1

system, we submit the SII weight 1 system. Because we find too high a SII weight degrades the naturalness of the synthesis speech.

3.5. ES3 task

Labels of the synthesis speech are automatically added with emphasis on the base of our EH1 system. We decide where to add emphasis according to the part of speech, and the old or new information. In common, words describing position, numbers, and the adjectives are emphasized. If the identical emphasized words appear in the questions of the man-machine conversation, then it will no longer be emphasized. 20 sentences labels are automatically labeled with emphasis and they are, at the same time, labeled by speech experts manually. Table 5 describes the statistical results by comparing these two emphasis labels.

Table 5. *Statistical result for automatic emphasis labeling*

Total words number	310
Expert labeled emphasis	76
Automatic labeled emphasis	49
Accuracy rate	82%
Recall rate	53%

Though the recall rate is not high, the accuracy rate is good. In order to make the emphasis more prominent, we delete the automatic predicted accent label for the phones where there are no automatic added emphasis labeling. And add accent label to the words where emphasis is labeled.

Subjective preference score test on synthesis speech appropriation is conduct for the baseline EH1 system and the ES3 system. The result shows that 56% people prefer the ES3 system and 44% prefer the EH1 system.

Though our ES3 system may be less natural compared with our EH1 system, people are able to hear the information of the answer more clearly with the answer by ES3 system.

4. Blizzard Challenge Evaluation result

This section discusses the evaluation result of Blizzard Challenge 2009. The identifier for USTC 2009 system is “S”. And in the evaluation, system “A” indicates the natural speech used for comparing. System “B”, “C”, “D” represents the Festival Benchmark system, HTS 2005 Benchmark system, and HTS 2007 Benchmark system separately. And system “E” to “W” are the participants.

4.1. EH1

MOS (naturalness), Similarity, and Word error rate (WER) evaluations are conducted for the EH1 systems. Our system performs well in all the three evaluations. Compared with the USTC 2008 system, we have a promotion in the rank of the WER evaluation. We believe that two factors made this promotion. The first one is the more accurate model we are using to model acoustic features this year. And the second one is the style change of the Semantically Unpredictable Sentences (SUS). In this year, easier and common used words are tested, other than the hard and rarely used ones. This is a progress in the evaluation rules. The evaluation result are listed in figure 4, 5, 6 and 7.

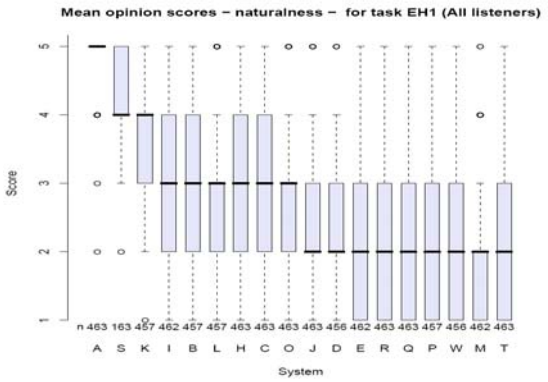

Fig. 4. *MOS score (naturalness) by all listeners*

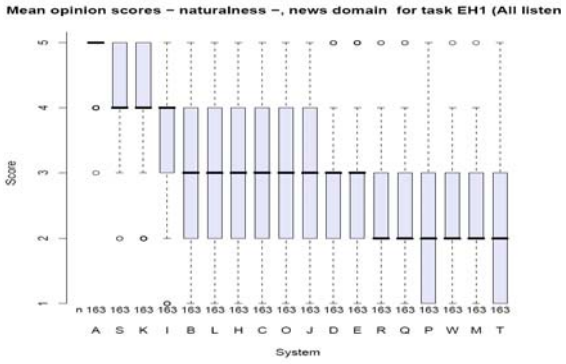

Fig. 5. *MOS score (naturalness) in news domain by all listeners*

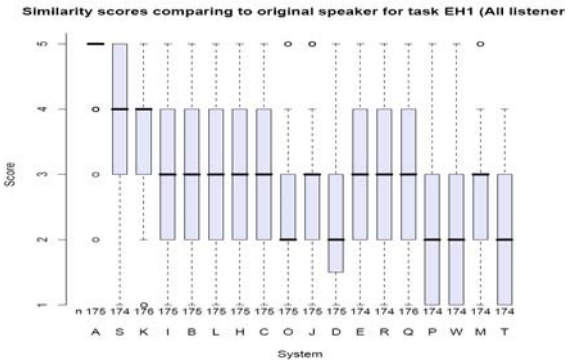

Fig. 6. *Similarity score by all listeners*

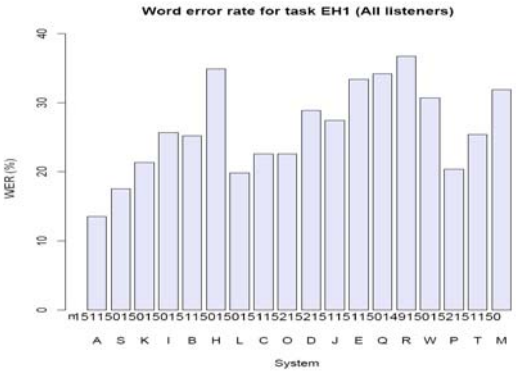

Fig. 7. *WER by all listeners*

4.2. EH2, ES1, ES2, ES3 result

The evaluation results for EH2, ES1, ES2, ES3 tasks are selectively listed in figure 8, 9, 10, 11. The evaluation results show the UTSC 2009 system also performs well in these tasks.

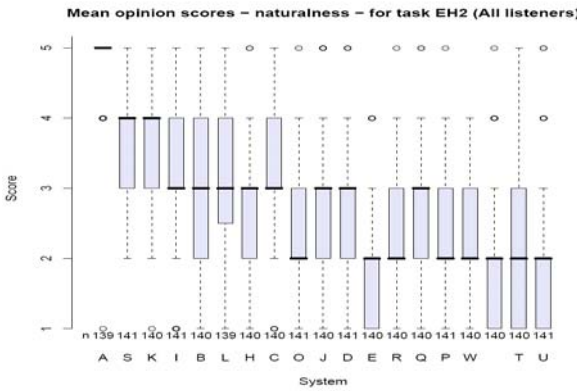

Fig. 8. *MOS score (naturalness) by all listeners for EH2 task*

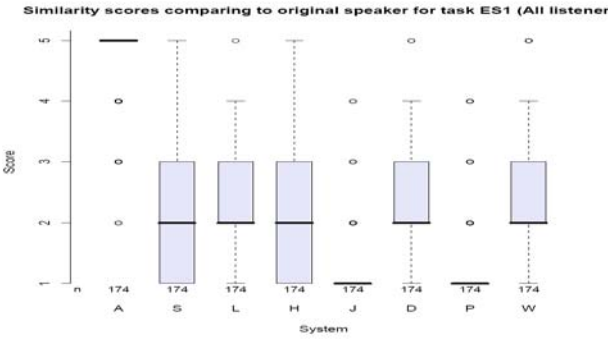

Fig. 9. *Similarity score by all listeners for ES1 task*

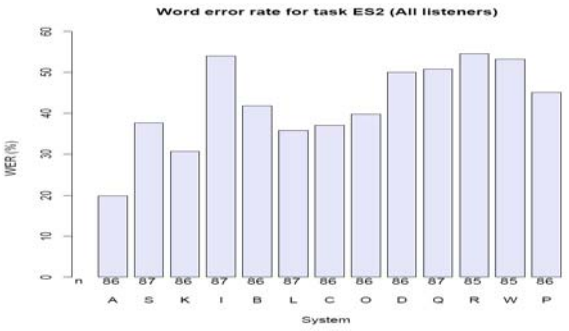

Fig. 10. *WER by all listeners for ES2 task*

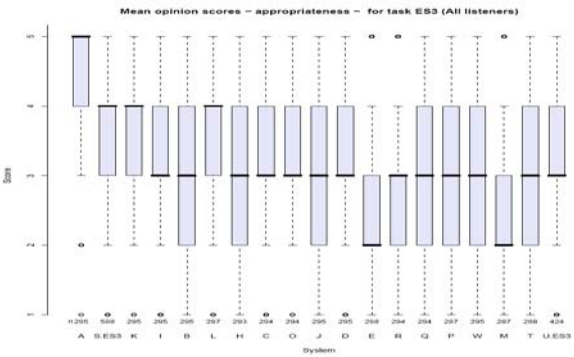

Fig. 11. *Appropriateness score by all listeners for ES3 task*

5. Conclusions

This paper introduced the USTC speech synthesis system built for the Blizzard Challenge 2009. Comparing the UTSC 2008 system, new techniques are introduced to train the acoustic model more accurately as to better model the variable UK English speaking style. Voice conversion system is built for ES1 task. And different modifications are made on the base of the baseline system to fulfill the commands of the ES2 and ES3 tasks. The evaluation results show that, the USTC 2009 system performs well in all the MOS (naturalness), Similarity, and WER evaluations.

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 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 HMMbased speech synthesis. In: Proc. ICASSP. pp. 89–92.

[8] ANSI-S3.5. American National Standard, Methods for Calculation of the Speech Intelligiblity Index. ANSI, 1997.

[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.