The USTC System for Blizzard Challenge 2011

Ling-Hui Chen, Chen-Yu Yang, Zhen-Hua Ling, Yuan Jiang
Li-Rong Dai, Yu Hu, Ren-Hua Wang

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

Abstract

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2011. USTC attended all the English tasks including a hub task and a spoke task. We developed a hidden Markov model (HMM) based unit selection system for both the tasks. And also some new techniques are employed in our speech synthesis system construction. Results of some internal experiments comparing these techniques are shown and analyzed. The evaluation results of Blizzard Challenge 2011 prove that our system performed well in the similarity and naturalness evaluation, and the differences in intelligibility between our system and the better systems are not significant.

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

1. Introduction

USTC have been attending Blizzard Challenge since 2006. In 2006, we submitted a statistical parametric speech synthesis system [1]. And since the speech quality of the statistical parametric speech synthesis [2] method suffers from the unnatural output of parametric synthesizer even if some high quality speech vocoder, such as STRAIGHT [3], has been used, we started to develop an HMM based unit-selection system since Blizzard Challenge 2007 [4]. In the Blizzard Challenge 2007, a baseline HMM based unit selection speech synthesis system using HMMs trained on acoustic features for phone unit selection was developed by USTC. The system performed well in both naturalness and similarity evaluations. In the Blizzard Challenge 2008 event, as a larger 15-hour UK database released, on the basis of the USTC unit selection system, the decision tree scale is tuned manually according to the scale of the training database to capture the variability of the speaking style of UK English [5]. Internal experiments showed 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 [7], Cross-validation (CV) and minimal generation error criterion (MGE) [8] was introduced to optimize the scale of the decision tree automatically. States of the HMMs, instead of the phone-unit, were adopted as the basic unit for selection and concatenation in the 1-hour speech synthesis system building task, and multi-Gaussian HMMs were employed in the 15-hour speech synthesis system building. In the Blizzard Challenge 2010 [9], a covariance tying technique was invited to improve the efficiency and reduce the footprint of the model, further more, we tried the sub-band waveform fusion with selected unit and the parameter synthesized speech.

This year in 2011, we used new labels generated together with both the text and its Lesseme symbols [10] released accompanied with the speech database to build an HMM based unit selection system. Besides, in the synthesis phase, we adopted the log likelihood ratio (LLR), instead of the conventional likelihood, as the target cost in the unit selection step to improve the unit-selection accuracy.

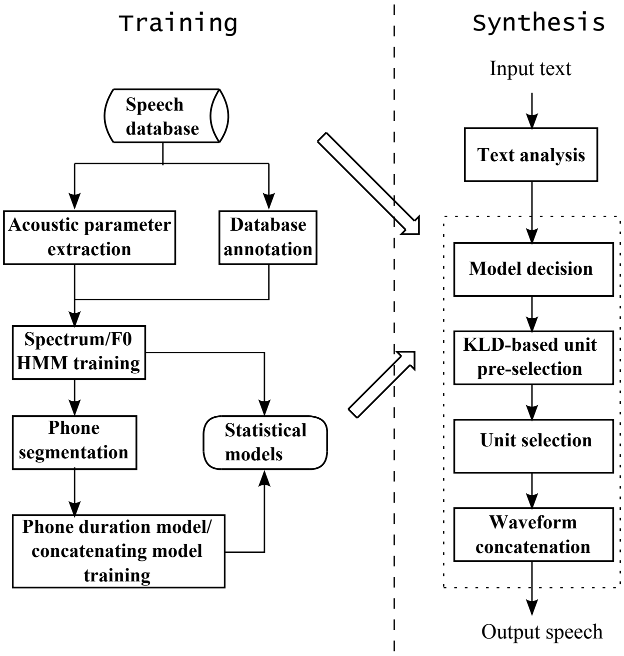

Figure 1: The framework of the USTC unit selection system.

This paper is organized as follows: Section 2 firstly reviews the basic USTC unit selection system, then introduces the techniques used in our system for Blizzard Challange 2011, and some results of the internal experiments conducted during the system building are also shown. And in section 4, the Blizzard Challenge evaluation results for our system are listed and analyzed. At last, the conculsions are made in section 5.

2. Method

2.1. The tasks in the Blizzard Challenge 2011

The Blizzard Challenge 2011 evaluation consists of two English sub-evaluations:

EH1 build a voice from the full 'Nancy' database;

ES2 build a voice designed to read names and addresses.

We built a phone-unit selection system for both the EH1 and ES1 task on the released 16 hours American English database of a speaker “Nancy”.

2.2. The USTC unit selection system

The flowchart of the USTC unit selection system is shown in Figure 1. It includes two main phases: the training phase and the synthesis phase, to build the USTC HMM-based unit selection system.

2.2.1. *Training phase*

First, at the training phase, HMMs [11] is trained to guide the unit selection. In the HMMs 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 the spectral parameters and the logarithmized F0. With the segmental and prosodic data, the spectrum part is modeled by continuous probability HMMs and the F0 part is modeled by multi-space probability HMMs (MSD-HMMs) [12]. 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) [6] based model clustering is applied 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 and F0 models are build to measure the smoothness at concatenated phone boundaries. 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]. Further, Another long time pitch model is trained to constrain the prosody between syllable units.

2.2.2. *Synthesis phase*

The synthesis phase can be divided into two steps: unit selection step and waveform concatenation step.

In the unit selection step, the phone-sized units for concatenation are selected from the speech database using a dynamic search algorithm to maximize the target function, which is a combination of likelihood and Kullback-Leibler divergence (KLD) [13]. That means, the optimal unit sequence is searched out by a criterion which is the trade-off between maximizing the likelihood of of candidate feature sequences towards the target models and minimizing the KLD between target and candidate models. The target models include spectral model, f0 model, duration model, concatenating model and long time pitch model. The weights of each model should be adjusted by manual operation patiently.

In order to reduce the computation cost of dynamic programming search, a KLD-based unit pre-selection algorithm is applied. The KLD between the target unit and the each candidate unit is measured to select the K-best units with minimum KLD for 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 implemented efficiently.

Finally, in the concatenation step, the waveforms of every two consecutive candidate units in the optimal unit sequence are concatenated to produce the synthesized speech. The cross-fade technique [14] is used here to smooth the phase discontinuity at the concatenation points of unit boundaries.

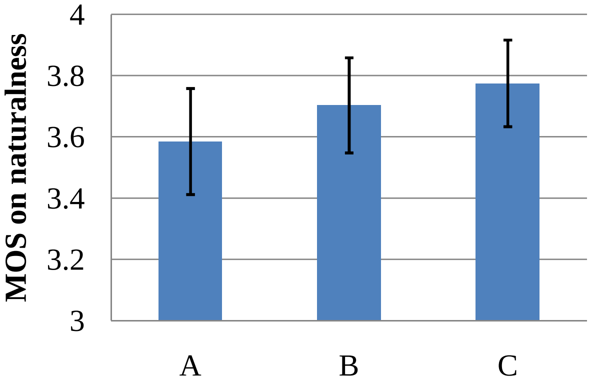

Figure 2: MOSs of the three system using different symbol systems on naturalness, error bars show the 95% confidence interval.

2.3. The use of Lesseme symbols

Accompany with the released waveform database for Blizzard Challenge 2011, a set of lesseme symbols [10] for each sentences were given. Comparing with most of the traditional symbolic representations, the Lessemes carries the musicality of speech. It consists of the traditional phonetic symbols (basic symbols) and descriptors for co-articulation and supra-segmental infomation, and the number of the basic phonetic symbols is 53, which is larger than that of the set in conventional USTC Blizzard Challenge systems, say 42.

We conducted an experiment to determine which kind of labels to use for our system. Three systems using different labels were built for comparison:

A Traditional labels: we segmented the speech database using an adapted HMM set, and the labels were generated using the iFLYTEK front-end text analyzer;

B The basic phonetic symbol set of Lessme was used. It was directly extracted and mapped from the Lesseme symbols that provided. The other parts of the labels were the same with **A**

C The co-articulation and supra-segmental infomation of the Lessemes were added into the labels used in **B**. They were converted to the format of the commonly used labels of HMM based speech synthesis systems. And some additional questions were also designed according to these information for model clustering.

All the systems were built in a similar way as we built USTC system for Blizzard Challenge 2010, except that the covariance tying technique was not used. The MDL factors for model clustering in these systems were set as 1.0; The result of a listening test was shown in Figure 2. Four listener, all native English spakers, participated in the test. 120 sentences synthesized by the systems (40 each) were played to each speaker. And the listeners were asked to give a 5-scale opinon score for each sentence. The result shows that the system **C** outperformed both the other two systems. The Lessemes can provide much more prosodic information to help us building more accurate acoustic models. Therefore, system **C** was used as the USTC Blizzard Challenge 2011 system.

2.4. LLR based target cost for unit selection

2.4.1. Conventional likelihood based target cost

As introduced in 2.2.2, in our previous entries for Blizzard Challenge, we adopted ML based criterion for unit selection. Assume that the number of phones in the utterance to be synthesized is N and the contextual information of the target sentence is C. A candidate sequence of phone-sized units which are used to synthesize this sentence can be written as $\boldsymbol{U} = \{u_1, u_2, \cdots, u_N\}$. The contextual information of the candidate unit sequence is $C(\boldsymbol{U})$. Then the optimal sequence $\boldsymbol{U}^*$ is obtained by:

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{M} w_m \Big[\Big(\log P_{\boldsymbol{\Lambda}_m}(\boldsymbol{X}(\boldsymbol{U},m)|C)\Big) - w_{KLD} D_{\boldsymbol{\Lambda}_m}(C(\boldsymbol{U}),C)\Big] \tag{1}$$

where M denotes the number of trained models, w_m and w_{KLD} is the manually set weights for likelihood and KLD component respectively, $\boldsymbol{X}(U,m)$ represents the extracted m-th feature of the unit sequence $\boldsymbol{U}$, $\log P_{\boldsymbol{\Lambda}_m}(\boldsymbol{X}(U,m)|C)$ is the log likelihood function for the observed feature $\boldsymbol{X}$ given model set $\boldsymbol{\Lambda}_m$ and contextual information C; $D_{\boldsymbol{\Lambda}_m}(C,C')$ calculates the KLD between two HMMs with contextual information C and C' given model set $\boldsymbol{\Lambda}_m$.

For more detail, equation (1) can also be written as:

$$\boldsymbol{U}^* = \arg\min_{\boldsymbol{U}} \Big[\sum_{n=1}^{N} CC(u_{n-1},u_n) + \sum_{k=1}^{K} SC(u_{m_{k-1}},u_{m_k}) + \sum_{n=1}^{N} TC(u_n)\Big] \tag{2}$$

where K is the number of syllables in the sentence for synthesis; $\{m_k\}_{k=1,,K}$ indicate the phone index of each syllables vowel in the sentence; $CC(u_{n-1},u_n)$ and $SC(u_{m_{k-1}},u_{m_k})$ denotes the concatenation costs which are introduced by the log likelihood function of spectrum/F0 concatenating models and syllable-level F0 model respectively; $TC(u_n)$ stands for the target cost function of the candidate unit u_n, which has been introduced in 2.2.2, that is,

$$TC(u_n) = -\sum_{m=1}^{M} w_m \Big[\mathcal{L}(u_n) - w_{KLD} D_{\boldsymbol{\Lambda}}(C(u_n),C)\Big] \tag{3}$$

2.4.2. LLR based target cost calculation method

The purpose of unit selection is to select the unit sequence that sounds most natural among all sequence candidates. Comparing with the ML-based unit selection criterion given in equation (1), it is more reasonable to select the unit sequence by maximizing the posterior probability given by

$$P(\boldsymbol{\Lambda}_N|\boldsymbol{X}(\boldsymbol{U})) = \frac{P(\boldsymbol{X}(\boldsymbol{U})|\boldsymbol{\Lambda}_N)P(\boldsymbol{\Lambda}_N)}{P(\boldsymbol{X}(\boldsymbol{U})|\boldsymbol{\Lambda}_N)P(\boldsymbol{\Lambda}_N) + P(\boldsymbol{X}(\boldsymbol{U})|\boldsymbol{\Lambda}_{UN})P(\boldsymbol{\Lambda}_{UN})} \tag{4}$$

where $\boldsymbol{\Lambda}_N$ and $\boldsymbol{\Lambda}_{UN}$ represent the models of natural synthesis speech and unnatural synthesis speech respectively. It is a posterior probability which invites the unnatural models as a competition mechanism for the natural models, helping select more natural units.

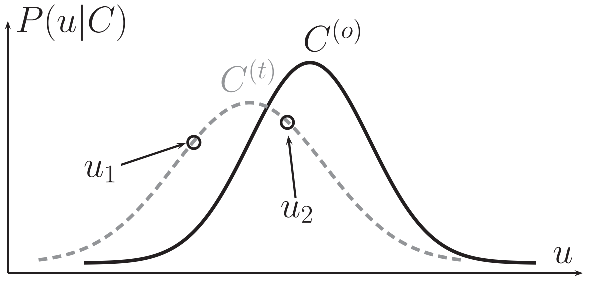

Figure 3: Illustration of the problem of ML-based unit selection, where $C^{(t)}$ denotes the target context, while $C^{(o)}$ denotes the context where the units u_1 and u_2 come from.

Therefore, a log likelihood ratio based target cost calculation method is proposed to integrate (4) into the unit selection criterion in a simplified form. At first, the log likelihood ratio between $P(\boldsymbol{X}(\boldsymbol{U})|\boldsymbol{\Lambda}_N)$ and $P(\boldsymbol{X}(\boldsymbol{U})|\boldsymbol{\Lambda}_{UN})$ for each feature and each phone unit is adopted to replace the posterior probability of unit sequence $P(\boldsymbol{\Lambda}_N|\boldsymbol{X}(\boldsymbol{U}))$; Then, since it is difficult to get the real natural and unnatural synthesis speech's models, we use the models of the target context $C^{(t)}$ and the original context $C^{(o)}$ the units belong to as an approximation of the natural and unnatural models respectively.

In this method, the target cost is given by:

$$TC(u_n) = -\sum_{m=1}^{M} w_m \Big[LLR(u_n) - w_{KLD} D_{\boldsymbol{\Lambda}}(C^{(o)}_{u_n}, C^{(t)})\Big] \tag{5}$$

in which,

$$\begin{aligned} LLR(u_n) = \\ & w_s\Big(\log P_{\boldsymbol{\Lambda}^{(s)}}(\boldsymbol{O}_n^{(s)}|C^{(t)}) - \log P_{\boldsymbol{\Lambda}^{(s)}}(\boldsymbol{O}_n^{(s)}|C^{(o)}_{u_n})\Big) \\ & + w_f\Big(\log P_{\boldsymbol{\Lambda}^{(f)}}(\boldsymbol{O}_n^{(f)}|C^{(t)}) - \log P_{\boldsymbol{\Lambda}^{(f)}}(\boldsymbol{O}_n^{(f)}|C^{(o)}_{u_n})\Big) \\ & + w_d\Big(\log P_{\boldsymbol{\Lambda}^{(d)}}(\boldsymbol{O}_n^{(d)}|C^{(t)}) - \log P_{\boldsymbol{\Lambda}^{(d)}}(\boldsymbol{O}_n^{(d)}|C^{(o)}_{u_n})\Big) \end{aligned} \tag{6}$$

where w_s, w_f, w_d are the weight for the log likelihood ratios of spectrum, F0 and phone duration respectively. $\boldsymbol{O}_n^{(s)}$, $\boldsymbol{O}_n^{(f)}$ and $\boldsymbol{O}_n^{(d)}$ stand for the acoustic features of spectrum, F0 and phone duration extracted from the current candidate u_n respectively; $\boldsymbol{\Lambda}^{(s)}$, $\boldsymbol{\Lambda}^{(f)}$ and $\boldsymbol{\Lambda}^{(d)}$ are the clustered context-dependent models for spectrum, F0 and phone duration, respectively.

Assume that there are two candidates for the target context $C^{(t)}$: u_1 and u_2, both from context $C^{(o)}$, as illustrated in Figure 3. The conventional criterion will choose u_2 which is more "like" $C^{(t)}$, that is, the one that is closer to the center of the probobility distribution function of $C^{(t)}$. However, u_2 may be also closer to the center of the PDF of $C^{(o)}$, which we assumed to be an unnatural model for the target model. This kind of problem can be avoided by using the log likelihood ratio based target cost calculation method.It will select the units whose acoustic parameters are not only close the center of target models but also far away from the center of the original models.

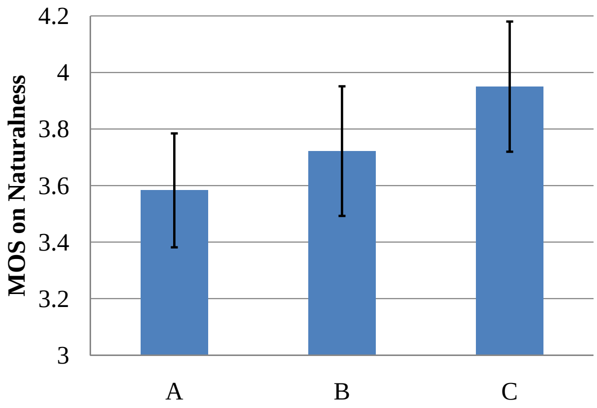

Figure 4: MOSs of three systems using different unit selection criterion on naturalness, error bars show the 95% confidence intervals.

We conducted a small scale listening test to validate the effective of the LLR-based criterion, comparing with the conventional ML-based criterion. Besides, we also built an additional system for comparison: we tried to use the correlation between the two units to be concatenated as a part of the join cost [15]. Therefore, three methods were compared:

A Unit selection based on ML criterion

B Similar with A, but the correlation between units to be contenated was used as a part of the join cost.

C Unit selection based on LLR criterion

Three listeners, all of them are native English speakers, took apart in the test. For each method, 30 sentences were play to each listener. The listeners were asked to give a 5-point mean opinion score (MOS) for each sentence they had heard. The results are shown in Figure 2.4.2. It appears that the new method is very effective, it outperformed the other two methods.

3. Evaluation

This section discusses the evaluation results of our system in Blizzard Challenge 2011. Table 1 shows the system identifying letters for some known systems. B, C and D are the benchmark systems.

Identifying letter	System
A	Natural speech
B	Festival
C	HTS
D	HTS, with 48 KHz sampling rate
G	USTC system

Table 1: Identifying letter for some known systems.

3.1. Similarity test

The boxplots of MOS on similarity of all the systems are shown in Figure 5. As we can see, the system G achieves the best similarity to the original speaker. Table 2 gives the results of Wilcoxon's signed rank tests to determine whether the difference between the two systems is significant. It can be founed that the difference between G and any other participant systems on similarity test is significant. 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.

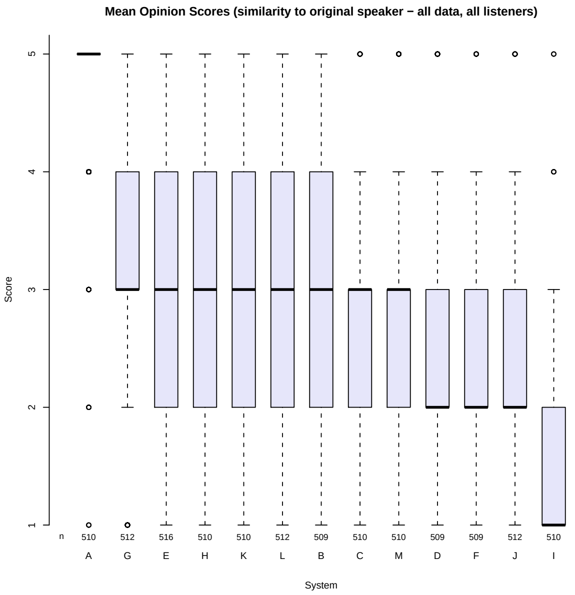

Figure 5: Boxplot of MOS on similarity evaluation.

ID	A	B	C	D	E	F	G	H	I	J	K	L	M
A	-	T	T	T	T	T	T	T	T	T	T	T	T
B	T	-	T	T	T	T	T	F	T	T	F	F	F
C	T	T	-	F	T	T	T	T	T	F	T	T	F
D	T	T	F	-	T	F	T	T	T	F	T	T	T
E	T	T	T	T	-	T	T	F	T	T	T	T	T
F	T	T	T	F	T	-	T	T	T	F	T	T	T
G	T	T	T	T	T	T	-	T	T	T	T	T	T
H	T	F	T	T	F	T	T	-	T	T	F	F	F
I	T	T	T	T	T	T	T	T	-	T	T	T	T
J	T	T	F	F	T	F	T	T	T	-	T	T	F
K	T	F	T	T	T	T	T	F	T	T	-	F	F
L	T	F	T	T	T	T	T	F	T	T	F	-	F
M	T	F	F	T	T	T	T	F	T	F	F	F	-

Table 2: Wilcoxon's signed rank tests of all participant on similarity evaluation.

The MOSs on similarity were much smaller that of the previous evaluations. A reason may be that, the refercence speech of target speaker is sampled at 96 KHz. But the sampling rate of most the entries were 16 KHz, except a few ones with 48 KHz sampling rate, including the benchmark system D.

3.2. Naturalness test

The boxplots of MOS on naturalness of all systems are shown in Figure 6. The results show that our system achieved the best

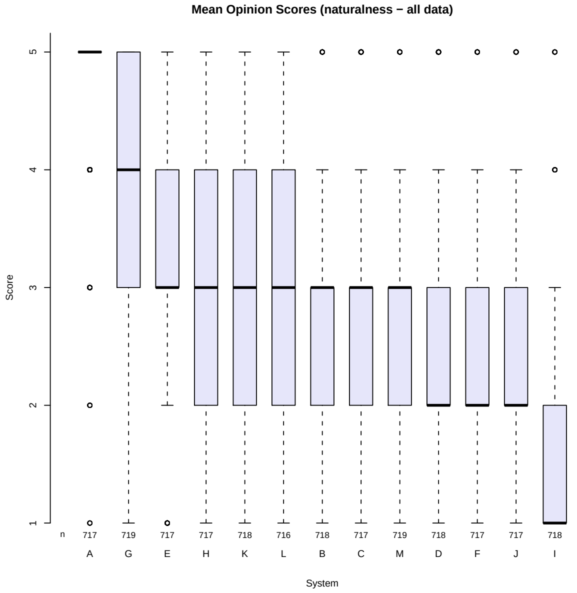

Figure 6: Boxplot of MOS on naturalness evaluation.

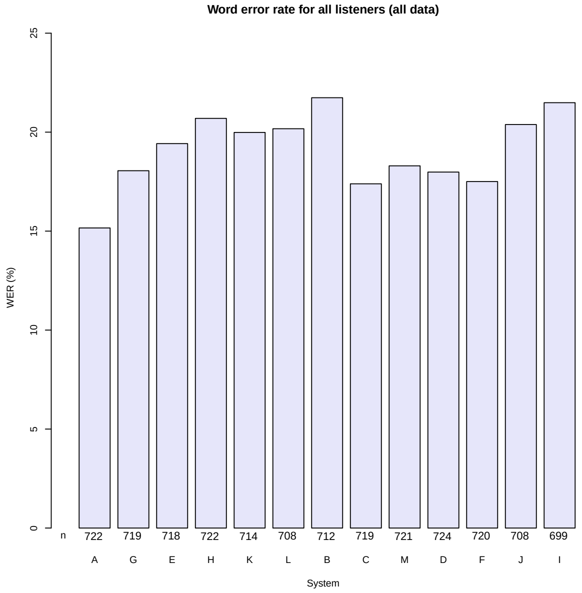

Figure 7: Word error rates of all participant on all task.

performance (not including the natural speech system A) on naturalness among all the participant systems. And the Wilcoxon's signed rank tests shown in Table 3 also shows the significance of our system comparing with all the other systems.

ID	A	B	C	D	E	F	G	H	I	J	K	L	M
A	-	T	T	T	T	T	T	T	T	T	T	T	T
B	T	-	F	F	T	F	T	T	T	T	T	T	F
C	T	F	-	F	T	T	T	T	T	T	T	T	F
D	T	F	F	-	T	F	T	T	T	F	T	T	F
E	T	T	T	T	-	T	T	F	T	T	T	T	T
F	T	F	T	F	T	-	T	T	T	F	T	T	T
G	T	T	T	T	T	T	-	T	T	T	T	T	T
H	T	T	T	T	F	T	T	-	T	T	F	T	T
I	T	T	T	T	T	T	T	T	-	T	T	T	T
J	T	T	T	F	T	F	T	T	T	-	T	T	T
K	T	T	T	T	T	T	T	F	T	T	-	F	T
L	T	T	T	T	T	T	T	T	T	T	F	-	T
M	T	F	F	F	T	T	T	T	T	T	T	T	-

Table 3: Wilcoxon's signed rank tests of all participant on naturalness evaluation.

3.3. Intelligibility test

Figure 7 shows the results of the overall word error rate (WER) test of all systems. Our system G got the 4th lowest WER among all the systems (not including A). As we found in previous Blizzard Challenge evaluation, the intelligibility of HMM based parametric synthesis method usually can achieve better intelligibility results than unit selection methods. But Table 4 tells us that the differences between our system and each of the systems that better than ours (C, D, F) are not significant. In the evaluation of this year, two tasks with different kind of sentences were conducted for the WER test: ES1 (synthesized address sentences) and semantically unpredictable sentences (SUS). The results of these two evaluations are given in Figure 8 and Figure 9. The USTC system performed the 3rd and 4th lowest WER these two tasks respectively. And there still is no significent difference between our system and any of the better systems.

ID	A	B	C	D	E	F	G	H	I	J	K	L	M
A	-	T	F	T	T	T	T	T	T	T	T	T	T
B	T	-	T	T	F	T	T	F	F	F	F	F	T
C	F	T	-	F	F	F	F	T	T	T	T	T	F
D	T	T	F	-	F	F	F	F	T	T	F	F	F
E	T	F	F	F	-	F	F	F	T	F	F	F	F
F	T	T	F	F	F	-	F	F	T	T	F	T	F
G	T	T	F	F	F	F	-	F	T	T	F	F	F
H	T	F	T	F	F	F	F	-	F	F	F	F	F
I	T	F	T	T	T	T	T	F	-	F	F	F	T
J	T	F	T	T	F	T	T	F	F	-	F	F	T
K	T	F	T	F	F	F	F	F	F	F	-	F	F
L	T	F	T	F	F	T	F	F	F	F	F	-	F
M	T	T	F	F	F	F	F	F	T	T	F	F	-

Table 4: Wilcoxon's signed rank tests of all participant on WERs.

4. Conclusions

This paper introduced the USTC speech synthesis system built for the Blizzard Challenge 2011. Comparing with the previous USTC unit selection and waveform concatenation system, a new label set were used for training HMMs, and the log likelihood ratio was adopted, instead of likelihood, as the target cost into the unit-selection phase. Some internal experiments indicated that the new system outperformed our previous sytems. The evaluation results show that, the USTC 2011 system per-

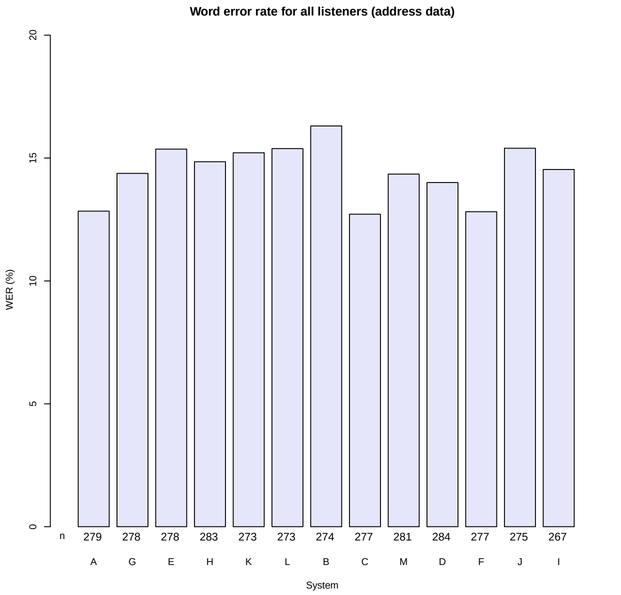

Figure 8: WERs on address task.

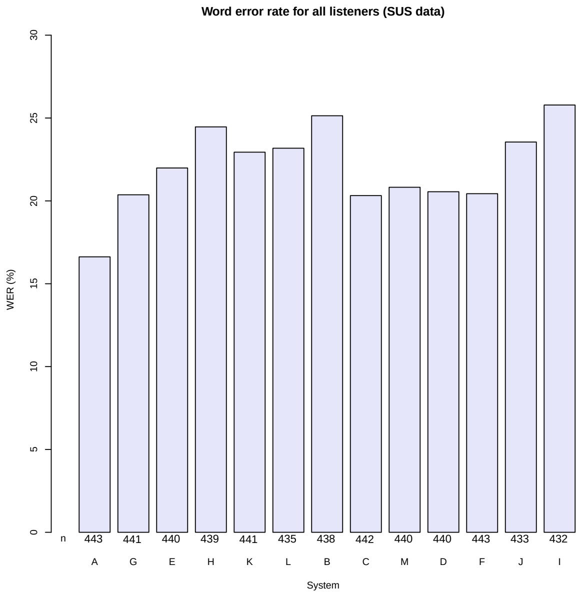

Figure 9: WERs on SUS task.

forms well in the naturalness, similarity evaluations. Though our system didn't performed best in intelligibility test, the differences between our system and any of the systems that better than ours were not significant.

5. Acknowledgements

This work was partially supported by National Nature Science Foundation of China (Grant No. : 60905010). The authors also thank the research division of iFlytek Co. Ltd., Hefei, China, for providing the English text analyzers.

6. 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] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[3] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–208, 1999.

[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, and R. Wang, "The ustc system for blizzard challenge 2008," in *Blizzard Challenge Workshop*, 2008.

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

[7] H. Lu, Z. Ling, M. Lei, C. Wang, H. Zhao, L. Chen, Y. Hu, L. Dai, and R. Wang, "The ustc system for blizzard challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[8] Y-J. Wu and R-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, May. 2006, vol. 1, pp. 89 –92.

[9] Y. Jiang, Z. Ling, M. Lei, C. Wang, H. Lu, Y. Hu, L. Dai, and R. Wang, "The ustc system for blizzard challenge 2010," in *Blizzard Challenge Workshop*, 2010.

[10] R. Nitisaroj and G. A Marple, "Use of lessemes in text-tospeech synthesis," in *M. Munro, S. Turner, A. Munro, and K. Campbell [Eds], Collective Writings on the Lessac Voice and Body Work: A Festschrift*. 2010, Llumina Press.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech.*, 1999, vol. 5, pp. 2347–2350.

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

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist*, vol. 22, no. 1, 1951.

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

[15] Y. Qian, Z-J. Yan, Y-J. Wu, F. K. Soong, G. Zhang, and L. Wang, "An hmm trajectory tiling (HTT) approach to high quality TTS - microsoft entry to blizzard challenge 2010," in *Blizzard Challenge Workshop*, 2010.