The USTC System for Blizzard Challenge 2012

Zhen-Hua Ling, Xian-Jun Xia, Yang Song, Chen-Yu Yang, Ling-Hui Chen, Li-Rong Dai

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, P.R.China

zhling@ustc.edu

Abstract

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2012. An audiobook speech corpus is adopted as the training data for system construction this year. Similar to our previous systems, the hidden Markov model (HMM) based unit selection and waveform concatenation approach is followed to develop our speech synthesis system using this corpus. Considering the inconsistent recording conditions and the narrator's expressiveness within the corpus, we add some channel and expressiveness related labels to each sentence besides the conventional segmental and prosodic labels for system construction. The evaluation results of Blizzard Challenge 2012 show that our system performs well in all evaluation tests, which proves the effectiveness of the HMM-based unit selection approach in coping with a non-standard speech synthesis corpus.

Index Terms: Speech synthesis, unit selection, hidden Markov model

1. Introduction

USTC have been attending Blizzard Challenge since 2006. In 2006, we submitted an HMM-based statistical parametric speech synthesis system [1]. Since Blizzard Challenge 2007 [2], we started to adopt the HMM-based unit selection and waveform concatenation approach [3] to build our systems for achieving better similarity and naturalness of synthetic speech. In this method, the optimal candidate phone sequence is searched out from the speech database by optimizing a statistical criterion which is derived from a group of acoustic models. The criterion is a combination of maximum likelihood and minimum Kullback-Leibler divergence (KLD). The acoustic models are trained using different acoustic features, such as frame-level spectral and F0 features, phone durations, and so on. Furthermore, some new techniques have been developed and evaluated during the system construction of the following years. In Blizzard Challenge 2009, cross-validation (CV) and minimal generation error criterion (MGE) [4] was introduced to optimize the scale of the decision tree for model clustering automatically. State-sized concatenation units and multi-Gaussian state probability density functions (PDFs) were also employed during system construction [5]. In Blizzard Challenge 2010, a covariance tying technique was applied to improve the efficiency and reduce the footprint of the acoustic models [6] and a syllable-level F0 model was introduced to evaluation the pitch combination of two adjacent syllables [7]. In Blizzard Challenge 2011, a maximum log likelihood ratio (LLR) criterion was developed to replace the conventional maximum likelihood criterion for unit selection [8].

The same HMM-based unit selection and waveform concatenation approach is followed to build our system for Blizzard Challenge 2012. Due to the limited preparation time, we construct our system using the framework similar to Blizzard Challenge 2007. The difference is that an extra syllable-level F0 model [7] is added. Because the speech corpus for system construction is composed of audiobook recordings with automatic transcriptions, we make a sentence selection based on the confidence value of speech recognition and add some channel and expressiveness related labels to each sentence for better context-dependent model training. The evaluation results of Blizzard Challenge 2012 prove the effectiveness of our HMM-based unit selection approach in dealing with such non-standard speech synthesis corpus.

This paper is organized as follows. Section 2 introduces our methods used for system construction. In section 3, the evaluation results of our system in Blizzard Challenge 2012 are shown and discussed. The conclusions are made in section 4.

2. Methods

2.1. HMM-based unit selection method

2.1.1. Model training

At training stage, we firsly choose a group of acoustic features that can be used to evaluate the naturalness of synthetic speech. Let M denote the number of chosen features. The task of model training is to estimate a set of context-dependent statistical models $\{\lambda_1, ..., \lambda_M\}$ for these features. In our system for Blizzard Challenge 2012, phone is adopted as the basic segment for unit selection and six models, including a spectrum model, a F0 model, a phone duration model, a concatenating spectrum model, a concatenating F0 model, and a syllable-level F0 model are trained. The spectrum model and the

F0 model are used to model the frame-level spectral and F0 features. The phone duration model presents the distribution of frame numbers within a phone. The concatenating spectrum and F0 models describe the distribution of spectral and F0 transitions at phone boundaries [2]. The syllable-level F0 model is trained using the F0 features extracted from the vowels of two adjacent syllables [7].

Based on the extracted spectral and F0 parameters for each frame of the training database, HMMs are estimated under maximum likelihood criterion to get the spectrum model and the F0 model, where the spectrum is modeled by a continuous probability distribution and the F0 is modeled by a multi-space probability distribution (MSD) [9]. Then we take the state alignment results using the trained HMMs to train the phone duration model, concatenating spectrum model, concatenating F0 model, and syllable-level F0 model respectively [2, 7]. In order to present the effects of context features on the distribution of acoustic features, all the models are trained context-dependently. Decision-tree-based model clustering technique [10] is applied to deal with the data-sparsity problems and to predict the model parameters for the context features that do not exist in the training set at synthesis stage.

2.1.2. Unit selection

Assume the utterance for synthesis consists of N phones and has context feature C, which is given by text analysis on the input sentence. A candidate sequence of phone-sized units to synthesis this utterance is written as $\boldsymbol{U} = \{u_1, u_2, ..., u_N\}$. Then, the optimal sequence $\boldsymbol{U}^*$ is searched out from the database under the statistical criterion of

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{M} w_m \left[\log P_{\lambda_m}(X(\boldsymbol{U}, m)|C) - w_{KLD} D_{\lambda_m}(C(\boldsymbol{U}), C)\right] \quad (1)$$

where $X(\boldsymbol{U}, m)$ extracts the acoustic features corresponding to the m-th model from the unit sequence $\boldsymbol{U}$; $C(\boldsymbol{U})$ denotes the context feature of the unit sequence $\boldsymbol{U}$; $P_{\lambda_m}(\cdot)$ and $D_{\lambda_m}(\cdot)$ represent the likelihood and KLD calculation functions respectively;w_m and w_{KLD} denote the weights for the m-th model and the KLD components[1] in the criterion. Furthermore, we can rewrite (1) into the conventional form of a sum of *target cost* and *concatenation cost* as described in [7]. Then a dynamic programming (DP) search is applied to find the optimal candidate sequence. In order to reduce the computation complexity of DP search, a KLD-based unit pre-selection algorithm [2] is applied before the DP search.

[1]Only the KLD of spectrum model, F0 model, and phone duration model are considered in our implementation.

Finally, the waveforms of every two consecutive candidate units in the optimal sequence are concatenated to produce the synthesized speech. The cross-fade technique [11] is used here to smooth the phase discontinuity at the concatenation points of phone boundaries.

2.2. Database and annotation

This year, an audiobook database [12] is released as the speech corpus for system construction. This database consists of the recordings of four books written by Mark Twain and was pronounced by an American English narrator. The texts of this database is generated by lightly supervised speech recognition technique [12] with a confidence value for each sentence. We processed the database by the following steps.

1) *Sentence selection*. The sentences with confidence value lower than 0 were discarded. The number of remaining sentences is $26,001$ and the total duration is about 50 hours.

2) *Segmental and prosodic labelling*. We adopted an English text analysis tool provided by iFLYTEK to get the phoneme transcription and ToBI information of each sentence based on the orthographic texts provided with the speech data. The phone boundary segmentation was conducted by HMM alignment using an adapted acoustic model.

3) *Channel labelling*. The database consists of four stories and we found there is significant channel inconsistency among the recordings of different stories. Thus, we added a channel label to each sentence according to the story it belongs to. We checked several samples of each story and assigned the channel label empirically. The first and third stories were labelled as *Channel 1*. The second and the fourth stories were labelled as *Channel 2* and *Channel 3* respectively. This channel label is added to the question set for decision-tree-based model clustering. At synthesis time, the label of *Channel 1* is used for input sentences.

4) *Expressiveness labelling*. Compared to the conventional speech synthesis databases with news-reading style, this audiobook database is far more expressive. In order to get relatively *neutral* speech for system construction, we made a simple two-value expressiveness labelling according to the average F0 of each sentence in an unsupervised way. The idea is similar to [13]. Firstly, the average F0 of all sentences in the database are calculated and a threshold of 175Hz is applied empirically. The sentences with average F0 lower than this threshold were labelled as *neutral* ones. Otherwise, they were labelled as *expressive* ones. This label is also added to the questions set for decision-tree-based

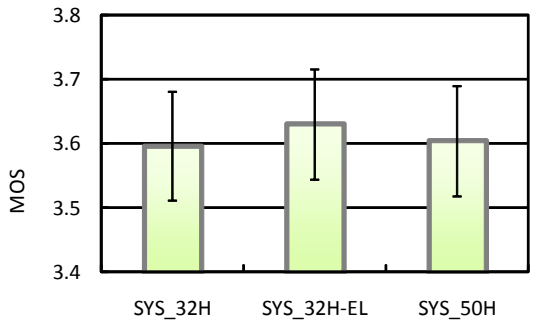

Figure 1: The mean opinion scores (MOS) with 95% confidence interval for the three systems in the internal experiment.

model clustering. At synthesis time, the *neutral* label is assigned to each input sentence.

2.3. Internal experiment

In order to evaluate the effectiveness of the sentence selection and expressiveness labelling method introduced above, an internal experiment was conducted during the system preparation. Three systems were built and compared.

- *SYS_32H*. The threshold of confidence value for sentence selection was set to 100, which leads to 32 hours of recordings for system construction. The channel labels were used and the expressiveness labels were neglected.

- *SYS_32H-EL*. The same as *SYS_32H* except that the expressiveness labels were used.

- *SYS_50H*. The same as *SYS_32H* except that the threshold of confidence value for sentence selection was set to 0.

Thirty-five sentences were synthesized using the three systems and were evaluated by five listeners. The listeners were required to give a score from 1 (very unnatural) to 5 (very natural) for each synthesized speech. The mean opinion scores (MOS) with 95% confidence interval for the three systems are shown in Fig. 1. From this figure, we see that the difference between setting the confidence threshold for sentence selection to 0 and 100 is very small. Introducing expressiveness labels improves the naturalness of synthetic speech slightly. However, the difference is also insignificant. Finally, we set the threshold of confidence value to 0 and adopt the two-value expressiveness labels in the submitted system.

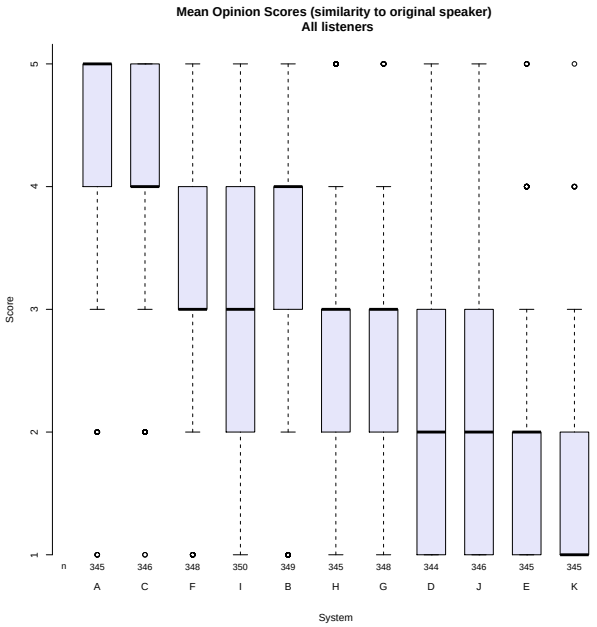

Figure 2: Boxplot of MOS on similarity.

3. Evaluation

This section introduces and discusses the evaluation results of our system in Blizzard Challenge 2012. This year, the identifier letter of our system is C. System A is the natural speech and system B is a Festival benchmark system.

3.1. Similarity test

The boxplots of MOS on similarity of all the systems are shown in Figure 2. As we can see, our system achieves the best similarity to the original speaker. The results of Wilcoxon's signed rank tests further show that the difference between system C and any other systems on similarity is significant at 1% level. 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 besides the simple waveform smoothing at phone boundaries.

3.2. Naturalness test

The boxplots of MOS on naturalness of all systems are shown in Fig. 3. The results show that our system achieved the best performance (not including the natural speech system A) on naturalness among all the participant systems. And the Wilcoxon's signed rank tests also show that the difference between C and any other participant systems on naturalness is significant.

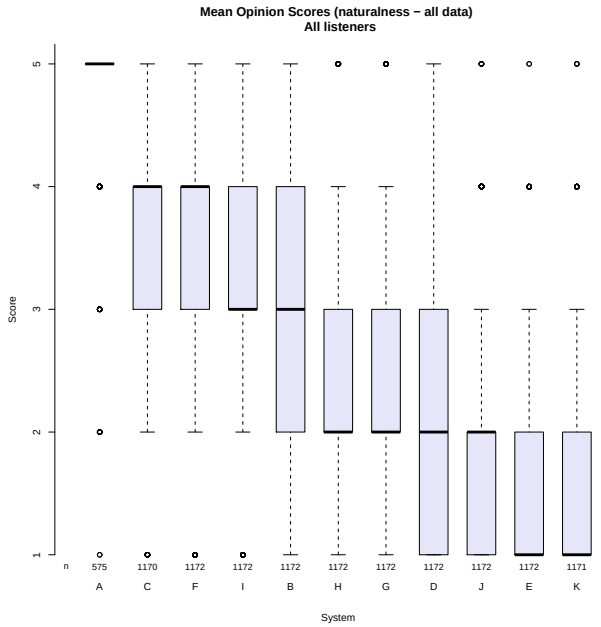

Figure 3: Boxplot of MOS on naturalness.

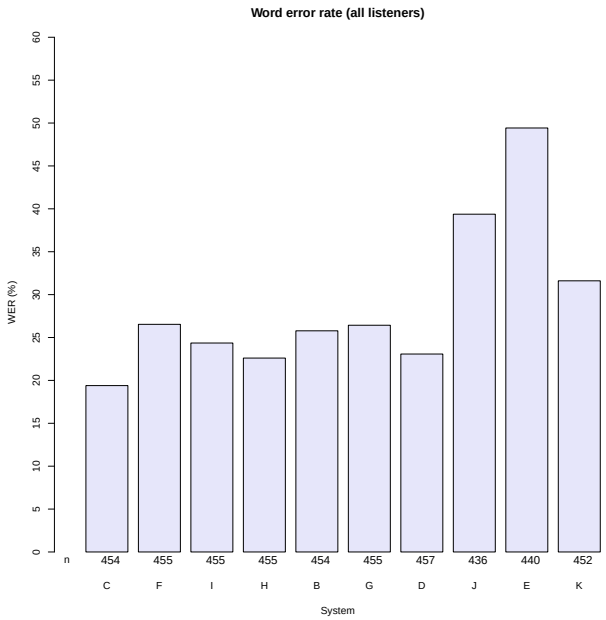

Figure 4: Word error rates of all participant systems.

3.3. Intelligibility test

Fig. 4 shows the results of the overall word error rate (WER) test of all systems. Our system achieves the lowest WER among all the systems, which is 19% for all listeners and 7.7% for the paid native English speakers. The Wilcoxon's signed rank tests shows the difference between System C and System D and H is insignificant.

3.4. Paragraph test

In this test, each listener listened to one whole paragraph from a novel and chose a score on a scale of 1 to 60 for the following seven aspects: overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening effort. Then, a mean opinion score could be calculated for each aspect. The evaluation results show that our system achieves the best performance in all the seven aspects. The mean opinion scores of our system and the natural speech are listed in Table 1. From this table, we see that the *emotion* and the *intonation* are the weakest aspects of our system. This is due to the lack of emotion and intonation related context features in current system.

System	A	C
Overall	48	37
Pleasantness	45	36
Speech Pause	47	35
Stress	47	34
Intonation	47	33
Emotion	46	32
Listening Effort	47	35

Table 1: MOS of our system (C) and the natural speech (A) in the paragraph test.

4. Conclusions

This paper introduced the USTC speech synthesis system built for the Blizzard Challenge 2012. The HMM-based unit selection approach has been adopted for system construction. The evaluation results of Blizzard Challenge 2012 has proved the effectiveness of this approach in synthesizing the texts of novel domain using a non-standard speech synthesis database. The audiobook synthesis is still a challenging task and there are still several problems need to be solved in the future work, such as channel equalization, automatic labelling for expressiveness and emotion factors, intonation modelling, and so on.

5. Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (Grant No. 60905010) and the Fundamental Research Funds for the Central Universities (Grant No. WK2100060005). The authors also thank the research division of iFlytek Co. Ltd., Hefei, China, for providing the English text analysis tools.

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] 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.

[3] Z. Ling and R. Wang, "HMM-based hierarchical unit selection combining kullback-leibler divergence with likelihood criterion," in *Proc. of ICASSP 2007*, vol. 4, april 2007, pp. 1245 – 1248.

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

[5] 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.

[6] 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.

[7] Z. Ling, Z. Wang, and L. Dai, "Statistical modeling of syllable-level f0 features for hmm-based unit selection speech synthesis," in *ISCSLP*, 2010.

[8] L. Chen, C. Yang, Z. Ling, Y. Jiang, L. Dai, Y. Hu, and R. Wang, "The USTC system for Blizzard Challenge 2011," in *Blizzard Challenge Workshop*, 2011.

[9] 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.

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

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

[12] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, 2010, pp. 2222–2225.

[13] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved HMM-TTS synthesis quality," in *Interspeech*, 2011, pp. 1821–1824.