The USTC System for Blizzard Challenge 2013

Ling-Hui Chen[†], *Zhen-Hua Ling*[†], *Yuan Jiang*[‡] *Yang Song*[†],
Xian-Jun Xia[†], *Yi-Qing Zu*[‡], *Run-Qiang Yan*[‡], *Li-Rong Dai*[†]

[†]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R. China
[‡]iFLYTEK Research, Hefei, P.R. China

chenlh@mail.ustc.edu.cn

Abstract

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2013. There are two evaluation tasks in this year: the English audiobook tasks and the pilot tasks on 4 Indian languages. According to the various amount of training data, different speech synthesis systems are constructed. The hidden Markov model (HMM) based unit selection and waveform concatenation system is built for the English audiobook tasks, and the HMM based statistical parametric speech synthesis system is built for the Indian language tasks. A synthesis quality prediction based method is applied to automatically optimize the weights of costs for unit selection in EH2 task. For Indian languages which we don't have front-end text processing system, letter-to-sound models are built. The evaluation results shows the effectiveness of our submitted system.

Index Terms: Statistical parametric speech synthesis, unit selection, hidden Markov models

1. Introduction

USTC have been attending Blizzard Challenge since 2006. We submitted our HMM-based HMM-based statistical parametric speech synthesis system in 2006 [1]. Since Blizzard Challenge 2007, when larger scale of corpus was provided, we started to adopt the HMM-based unit selection and waveform concatenation approach to build our systems in order to achieve better similarity and naturalness in synthetic speech [2]. And this approach is further developed in the Blizzard Challenge of the following years. In Blizzard Challenge 2009 [3], a new acoustic model clustering method was introduced to automatically optimize the scale of decision tree using cross-validation (CV) and minimal generation error (MGE) criterion. In Blizzard Challenge 2010 [4], a covariance tying approach was adopted to reduce the footprint of model and improve the efficiency of model training. Besides, syllable-level F0 model was introduced to evaluate the pitch combination of two adjacent syllables. In Blizzard Challenge 2011 [5], a maximum log likelihood ratio (LLR) criterion was adopted instead of conventional maximum likelihood (ML) criterion to guide the unit selection. In Blizzard Challenge 2012 [6], we built a system to dealing with the released non-standard speech synthesis database by sentence selection and adding channel and expressiveness related labels.

New challenges were proposed in Blizzard Challenge 2013. In addition to a standard English evaluation task (EH2), a large scale of unsegmented English audiobook (300 hours) speech synthesis database (EH1) and 4 small scale Indian language speech databases (IH1 - IH4) were released. Due to the limitation of preparing time, we construct our system for EH1 using a similar framework in our Blizzard Challenge 2012 system. The audiobook waveforms are segmented using voice activity detect (VAD) technique, and the texts are automatically split and manually adjusted. A similar system is also constructed for EH2, and we automatically optimize the weights for search unit sequence using a synthesis quality prediction (SQP) based method [7]. An HMM based statistical parametric speech synthesis system is construct for the Indian language following the USTC system for Blizzard Challenge 2006. For Bengali, Kannada and Tamil, since we don't have the front-end text processing system for them, L2S [8] models are built by simply using the text information released in the database.

This paper is organized as follows: Section 2 reviews the basic USTC unit selection system and statistical parametric speech synthesis system. The details of building USTC system for Blizzard Challenge 2013 will be given in section 3. In section 4, the Blizzard Challenge evaluation results for our system are shown and analysed. Conclusions are made in section 5.

2. Baseline systems

The Blizzard Challenge 2013 evaluation consists of 6 sub-evaluations:

- EH1: build system on the 300-hours unsegmented audiobook data;
- EH2: build system on the provided 19-hours segmented audiobook data;
- IH tasks: build systems in Hindi, Bengali, Kannada and Tamil languages.

The EH2 evaluation is the same with the Blizzard Challenge 2012, but the EH1, IH1, IH2, IH3 and IH4 are new. According to the different scale of training database, we built two different systems for English tasks and Indian tasks. HMM-based unit selection system is adopted for English tasks with audiobook training data, Since there is only about 1 hour training data for each Indian tasks, the HMM-based statistical parameter speech synthesis system is adopted.

2.1. The USTC unit selection system

Figure 1 shows the flowchart of the USTC unit selection system. The system consists of two main phases: the training phase and the synthesis phase.

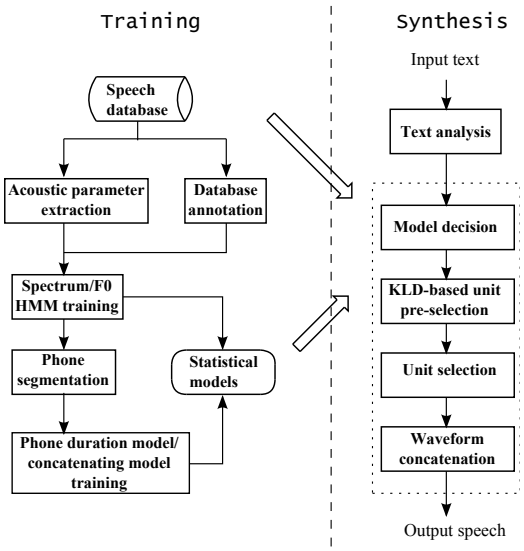

Figure 1: The flowchart of USTC unit selection system.

2.1.1. Training phase

First, at the training phase, HMMs [9] is trained as acoustic models to guide the unit selection. Six sets of HMMs are trained, including a set of spectrum models, a set of F0 models, a set of phone duration models, a set of concatenating spectrum models, a set of concatenating F0 models and a set of syllable-level F0 models. The spectrum models are trained using frame-level spectral and F0 features. The phone duration models are training using the durations (number of frames) in the phone segments. The concatenating spectral and concatenating F0 models are trained to model the distributions of spectral and F0 transitions at phone boundaries (e.g. delta spectra and delta F0s). The syllable-level F0 model is trained using the F0 features extracted from the vowels of two adjacent syllables. Spectral features are modeled by continuous probability HMMs and the F0 features are modeled by multi-space probability HMMs (MSD-HMMs) [10]. A decision-tree-based model clustering method is applied after context-dependent HMM training to deal with the data sparseness problem and predict the model parameters for the unseen context at the synthesis phase. Minimum description length (MDL) [11] based model clustering is applied to control the size of the decision tree. The phone durations, concatenating spectral features, concatenating F0 features and syllable-level F0 features are extracted using state-frame alignment information.

2.1.2. Synthesis phase

At synthesis phase, firstly, a sequence of phone units are selected under a criterion, then, these units are concatenated to form synthetic speech. Let N be the number of phones in the utterance to be synthesized with context feature C. In our system, a sequence of phone unit candidates $\boldsymbol{U} = \{u_1, u_2, \cdots, u_N\}$ are search out from the database under a statistical criterion of

$$\boldsymbol{U}^* = \arg\max_{\boldsymbol{U}} \sum_{m=1}^{6} w_m [\log P(\boldsymbol{X}(\boldsymbol{U}, m)|C, \lambda_m) - w_{KLD} D_m(C(\boldsymbol{U}), C),] \quad (1)$$

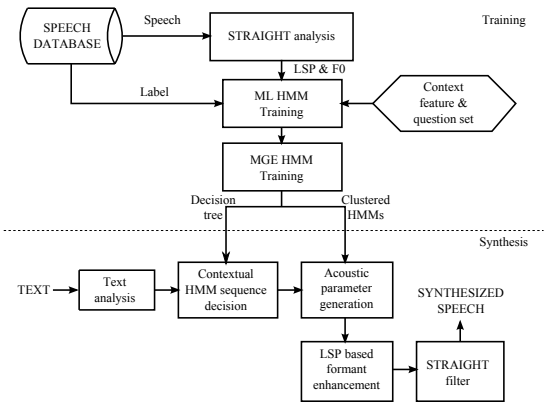

Figure 2: Flowchart of the HMM-based statistical parametric speech synthesis system.

where λ_m indicates the acoustic models described in the previous section, and w_m corresponds to their weights, $\boldsymbol{X}(\boldsymbol{U}, m)$ and $C(\boldsymbol{U})$ extract corresponding acoustic features and context features from the unit, $D_m()$ denotes the Kullback-Leibler divergence (KLD) [12]. A dynamic programming (DP) search algorithm is applied to find the optimal unit sequence, and a KLD-based unit pre-selection method is adopted to reduce the computational complexity in the DP based search.

Finally, in the concatenation step, the waveforms of every two consecutive candidate units in the optimal unit sequence are concatenated to produce the synthetic speech. The cross-fade technique [13] is used here to smooth the phase discontinuity at the concatenation points of unit boundaries.

2.2. HMM-based parameter speech synthesis method

The USTC system for Blizzard Challenge 2006 is followed to build the baseline system for IH1 $\sim$ IH4. As shown in Figure 2, in the training stage, a set of HMMs are estimated as acoustic models. First, acoustic models (including spectral, F0, phone duration and state duration models) are trained using maximizing likelihood criterion in the same manner as that in our unit selection system. Line spectral pair (LSP) is adopted as spectral feature for model training. Then, minimum generation error (MGE) training is applied to further refine the model parameters of spectral and F0 models. In the synthesis stage, firstly, state duration is determined jointly by phone duration models and state duration models. secondly, maximizing output probability parameter generation algorithm is adopted to generate static LSP sequence. Finally, before synthesizing using STRAIGHT, LSP based formant enhancement method is adopted to improve the quality and articulation of generated speech quality.

3. System building

3.1. EH1 task

In EH1 task, a 300-hours unsegmented English audiobook database is released for system building. Each waveform file in this database contains a chapter, which may be up to hours and cannot be processed by our system. Therefore, before model training, these data is segmented and aligned to corresponding text, which is downloaded from the Internet. About 200-hours data, whose corresponding text can be found from the Internet, is used for system building. The data segmentation is performed

semi-automatically in two steps:

- **Waveform segmentation** Firstly, VAD technique is adopted to detect the silent segments in the waveform, then the waveform is segmented at the silent segments whose duration is longer than a threshold we set in advance;
- **Text segmentation** Firstly, the text of a chapter is divided into sentences automatically. Then, the text segment boundaries are manually checked and adjusted.

Since the size of the database is too large, it is not necessary to train the acoustic models using all the data. Therefore, we selected about 100-hours context balanced data to train acoustic models. But in synthesis step, all 200 hours data was segmented by the acoustic models for unit selection. In order to improve the efficiency of model clustering, which is the most time consuming part during model training, decision trees were built phone-dependently, each tree corresponds to the models belong to one phoneme. Additionally, we marked the dialogue as a context attribute, because when reading those parts, the reader always attempt change his voice to imitate the character who is speaking.

3.2. EH2 task

As shown in section 2.1.2, the criterion for unit selection is a weighted summation of several component, including target costs and join costs. These weights are super-parameters of the model, and they are vital to the quality of synthetic speech. They can be optimized under minimum unit selection error (MUSE) criterion [14], but it is too computationally complex. Therefore, manual tuning according the subjective listening on a development set is usually adopted, but manual tuning is also difficult because there are too many combinations. In USTC system for Blizzard Challenge 2013, we adopted a SQP based method to automatically optimized these weights

Figure 3 shows the flowchart of constructing the synthesis quality predictor. Firstly, sentences in development set are re-synthesized using several weight candidates. Then subjective listening tests are conducted through the Amazon Mechanical Turk (AMT) to obtain the mean opinion score (MOS) for each sentence. On the other hand, a sentence-level feature for predictint synthetic quality is extracted from synthetic and reference speech. The reference speech is generated by the HMM bases statistical parametric approach. This feature is composed by unit selection costs and acoustic distances between synthetic and reference speech. 8 kinds of distances are evaluated as acoustic distances, including

- Spectral distance: average distance between Mel-cepstra of synthetic and reference speech, dynamic time warping (DTW) is applied to perform frame alignment;
- F0 distance: average distance between frame F0s, calculated in the same manner as spectral distance;
- Power distance: average distance between frame powers, calculated in the same manner as spectral distance;
- Spectral transition distance: average distance of spectral transitions at phone boundaries in two sentences;
- F0 transition distance: average distance of F0 transitions at phone boundaries in two sentences;
- Power transition distance: average distance of frame power transitions at phone boundaries in two sentences;

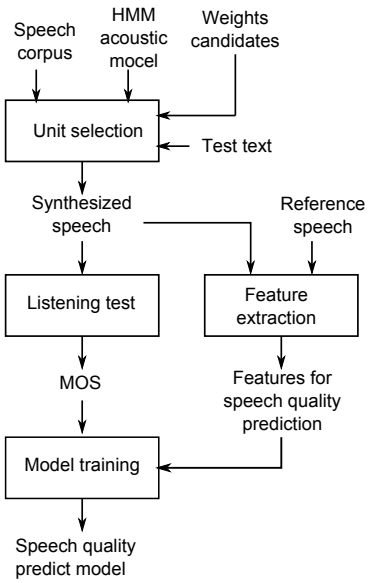

Figure 3: The flowchart of constructing a synthesis quality predictor.

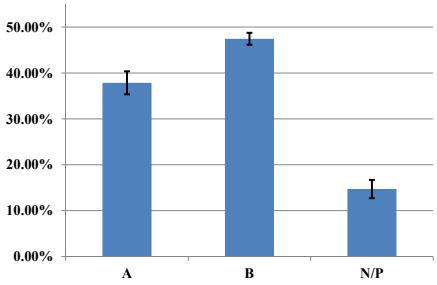

Figure 4: Preference test between synthesized speech by using traditionally optimized weights and manually optimized weights, N/P means no preference.

- Silent duration distance: distance of total silent durations in two sentences;
- Non-silent duration distance: distance of total non-silent durations in two sentences;

At last, the synthesis quality predictor is trained based on these data, we adopted multivariate adaptive regression splines (MARS) to model the mapping relation from the feature to predicted MOS.

The estimated synthesis quality predictor is used to optimize the weights. Firstly we set initial weights to re-synthesize the sentences in the development set, and the MOS of synthetic sentences are predicted to update the weights. The weights are iteratively updated by using the pattern search method until the weights that can generate the highest MOS are found.

We conduct an internal experiment to derive the optimized weights for our system. We selected 30 sentences as an development set, and set 24 weights candidates according to our experience. 183 native English listeners participated in the subjective tests through the AMT platform. A preference test was took to compare the naturalness of speech synthesized by using the weights in the traditional way and the new way. The results are shown in Figure 4, in which,

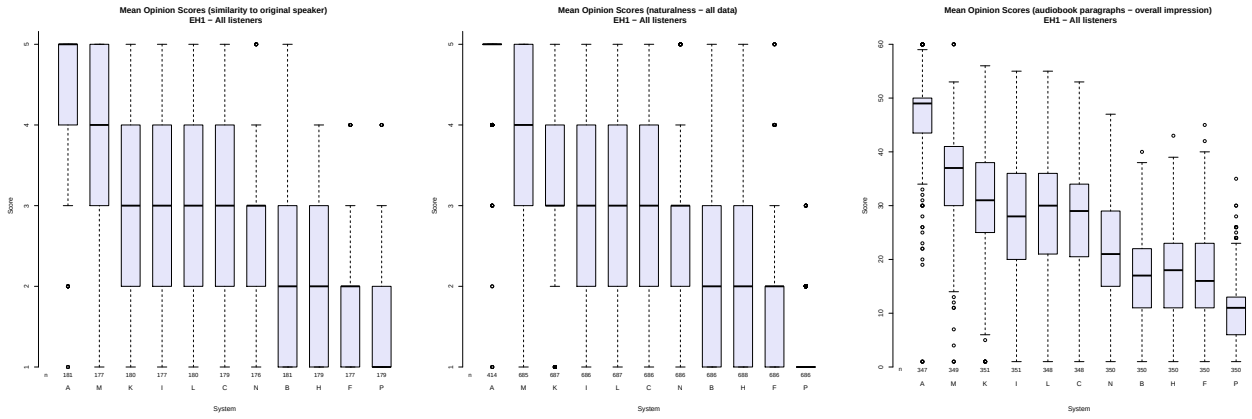

Figure 5: Boxplot of similarity, naturalness and paragraph overall impression tests in EH1 task.

A the best among the 24 sets of speech synthesized by the weights candidates;

B speech synthesized using the automatically optimized weights.

We can see that the new weights optimizing method presented above can effectively improve the naturalness of synthetic speech.

3.3. IH tasks

Blizzard Challenge 2013 released four Indian languages (Hindi, Bengali, Kannada and Tamil) speech databases for speech synthesis system building.

We used an Hindi TTS engine to perform phoneme transcription and prosodic information from each of the UTF-8 format sentence released in the IH1 database. For the other three languages, L2S models were constructed by simply using the textual materials in the released database.

In Bengali, Kannada and Tamil speech databases, Phonemic level segmental annotation, UTF-8 format text and its Roman transliteration are provided. Matching these three materials can create pronunciation dictionary for each language. Then the L2S rules is built by decision tree based algorithm. Through analysing the phonemes sequence of the words, each letter has possible pronunciations. The question set for decision tree can be roughly divided into two categories: 1) Letter category: the character of current and surrounding letters, 2) Phoneme category: the phonetic identity of preceding phonemes. In this procedure, the unmatched sentences are deleted. Table 1 show the number of the words in each dictionary and the accuracy of ten pumping test for each L2S model.

Table 1: the number of deleted sentences, words in the dictionary and the accuracy of ten pumping test for L2S models.

Language	Bangli	Kannada	Tamil
deleted sentences	5	5	0
words in dictionary	2262	2105	2123
ten pumping test accuracy	94.7%	94.3%	96.7%

The quality of database annotation is a key step to speech synthesis system building. A selective test on the released phonemic level annotation for each language shows that it is not accurate enough for speech synthesis. Therefore we adjust database annotations as the following steps for the four languages:

1) Initial models training for force alignment to acquire units boundary and recognize pause position.
2) Verify only pause position manually.
3) Models retraining based on the modified annotation.

4. Evaluation

This section discusses the evaluation results of our system on the tasks. Among all the systems, our system label is K, A is the natural speech, B and C are the benchmark systems built by Festival and HTS respectively.

4.1. EH1 task

Figure 5 shows the boxplot of the evaluation results of similarity to the original speaker, naturalness and overall impression of paragraph in EH1 evaluation task. As we can see, in similarity evaluation, our system doesn't perform as we expected, our system ranks at 4th place. One main reason may be that the training and reference speech waveforms are sampled at 44kHz sampling rate, but our system is built on data down-sampled to 16kHz. And due to the miss-match between the text and speech data, naturalness of our system is also affected, as shown that the MOS of our system is lower than system M. For the same reason, the our system performs 2nd best in the paragraph test, and the difference between our system and the 3rd one (L) is insignificant. In intelligibility test in Fig. 6, our system gets the 4th lowest word error rate (WER), but the Wilcoxon's signed rank tests show that the differences between our system and any of the top 3 best systems (C, I and M) are insignificant.

4.2. EH2 task

Figure 8 shows the boxplot of the evaluation results of similarity to the original speaker, naturalness and overall impression of paragraph in the EH2 evaluation task. The training data in this task is sampled at 16kHz, no down-sampling process was made. Therefore, our system performed best in similarity test (the same as system M), there is no significant difference among system K, M, L and N. This is attribute to the unit selection and waveform concatenation approach, the synthesized waveforms

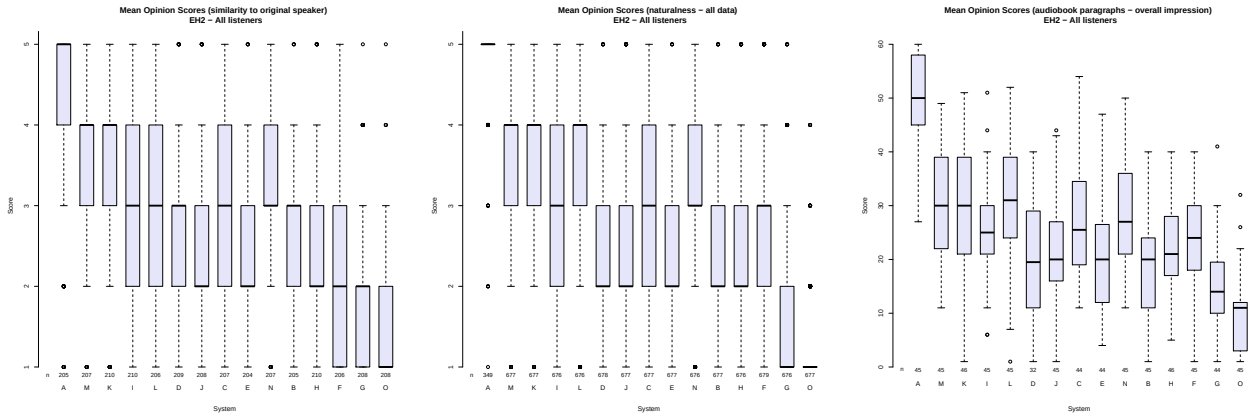

Figure 8: Boxplot of similarity, naturalness and paragraph overall impression test in EH2 task.

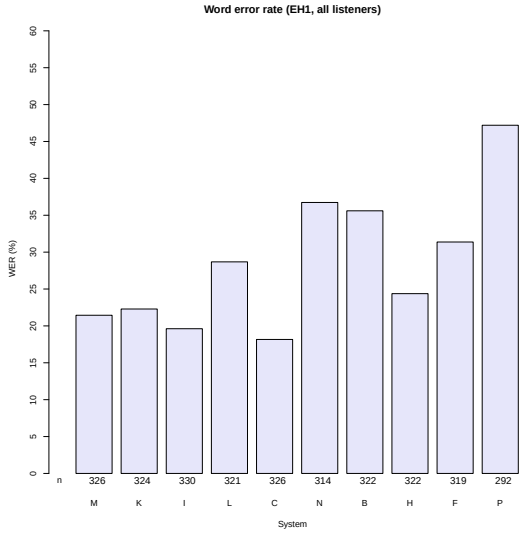

Figure 6: WER of all participants in EH1 task.

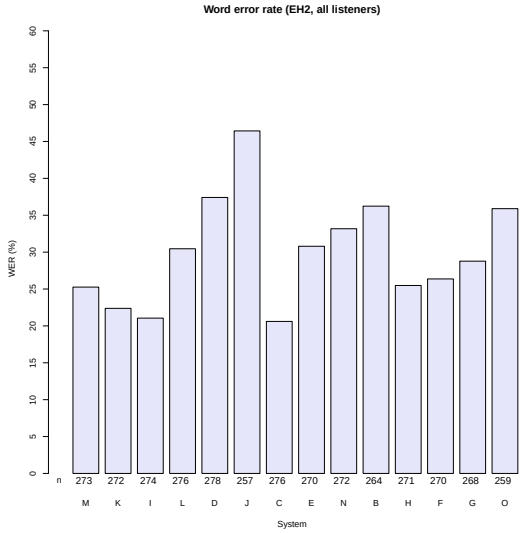

Figure 7: WER of all participants in EH2 task.

are composed by the original units from the training database, which can retain high similarity to the original speaker. Our system also outperforms all the other systems significantly in naturalness. In paragraph overall impression test, the score of our system is slightly lower than system K and M, but the difference is insignificant. In intelligibility test results shown in Fig. 7, the WER of our system is the 3rd lowest, and the differences between any two of the 5-lowest systems are insignificant.

4.3. IH tasks

Figure 9 shows the similarity evaluation results of all participants. The naturalness evaluation results are shown in Figure 10. WER results of IH1 and IH3 tasks are shown in Figure 11. As introduced in the previous section, a front-end text analyser is constructed for IH1 task, and for the rest tasks, the systems are built simply by L2S rule. Therefore, the performance of our system on IH1 tasks is better than that on the other tasks, Although we got the second highest MOS score on both similarity and naturalness, the pairwise Wilcoxon signed rank tests show that the difference between our system and the best one (L) is insignificant. In the intelligibility test of IH1 task, we got the lowest WER, insignificantly better than the second (D) and third (E) system.

As shown in the figure, our system doesn't perform well enough in IH2, IH3 and IH4 tasks as that in IH1 task, the main cause is the post-filtering process to the training waveforms before feature extraction in our system. Since the quality in the released speech is not quite good, the post-filtering process is used to enhance the speech quality, this process affected the similarity to the original speaker and naturalness of synthetic speech. Especially in IH3 task (Kannada), in which an additional post-filtering process is taken to the synthetic speech waveforms, and the similarity and naturalness are further degraded. In the intelligibility test of IH3 task, there isn't significant difference among most systems.

5. Conclusions

This paper presents the details of constructing the USTC system for the Blizzard Challenge 2013. The HMM based unit selection approach has been adopted for English tasks with large

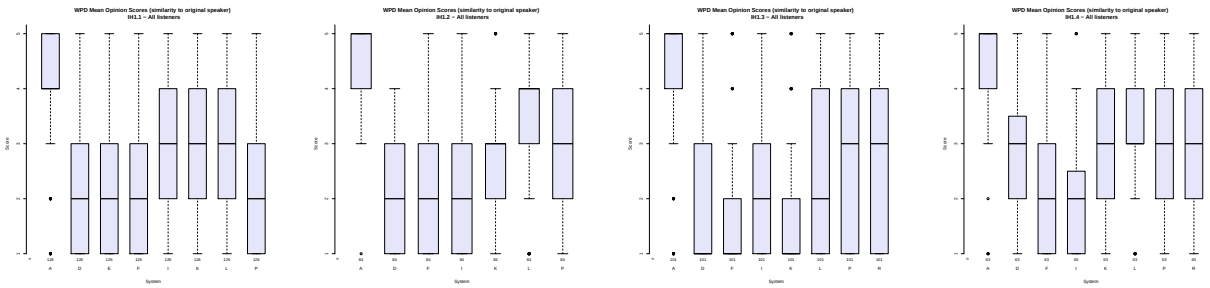

Figure 9: Boxplot of similarity evaluation results in IH tasks.

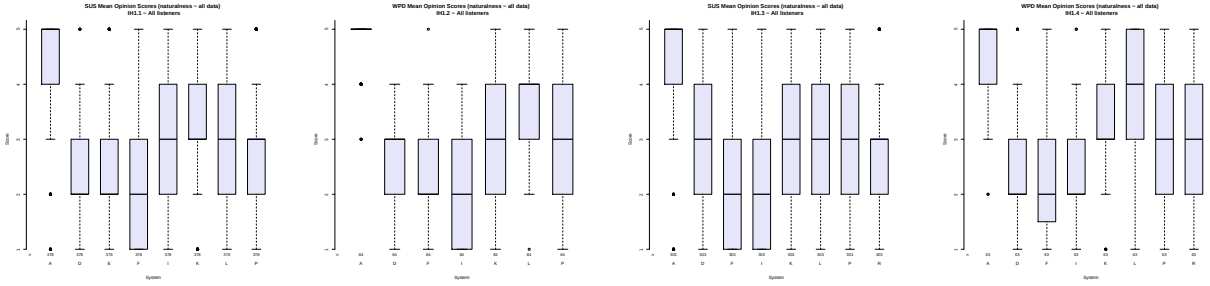

Figure 10: Boxplot of naturalness evaluation results in IH tasks.

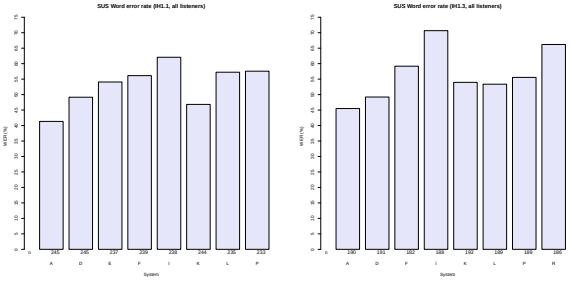

Figure 11: WER of all participants in IH1 and IH3 task.

scale training data, and the HMM based statistical parametric speech synthesis approach has been adopted for Indian tasks with small training set. A synthetic speech quality prediction based method is adopted to automatically optimized the weights for costs in unit selection. Indian language systems without text analyser are built simply using L2S rules. The evaluation results show the effectiveness of our system in some aspects. There are still many problems in the audiobook tasks to be solved in the future, such as channel equalization, automatic alignment of chapter waveform and text and so on.

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] 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.

[4] 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.

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

[6] Z.-H. Ling, X.-J. Xia, Y. Song, C.-Y. Yang, L.-H. Chen, and L.-R. Dai, "The USTC system for blizzard challenge 2012," in *Blizzard Challenge Workshop*, 2012.

[7] Y. Song, Z.-H. Ling, and L.-R. Dai, "Optimization method for unit selection speech synthesis based on synthesis quality prediction," in *National Conference on Man-Machine Speech Communication*, 2013.

[8] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules." 3rd ESCA Workshop on Speech Synthesis, 1998, pp. 77–80.

[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 *Proc. Eurospeech.*, vol. 5, 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 *Proc. of ICASSP*, 1999, pp. 229–232.

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

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

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

[14] Z.-H. Ling and R.-H. Wang, "Minimum unit selection error training for hmm-based unit selection speech synthesis system," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3949–3952.