The USTC System for Blizzard Challenge 2015

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

[†]National Engineering Laboratory for 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

In this paper, the speech synthesis systems developed by USTC for Bilizzard Challenge 2015 are introduced. The constructed systems include Indian languages of Bengali, Hindi, Malayalam, Marathi, Tamil and Telugu. We accomplished two tasks for all these six languages, namely the mono-lingual task (IH1 hub task) and the multi-lingual task (IH2 spoken task). A hidden Markov model (HMM)-based unit selection system was built for Hindi language with the released large corpus. Another five HMM-based parametric speech synthesis systems were built for Bengali, Malayalam, Marathi, Tamil and Telugu repectively because of the limited corpus and lack of front-end knowledge. All these systems were submitted for evaluation. The effectiveness and robustness of the submitted systems are verified according to the results of the formal evaluation.

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

1. Introduction

The USTC team have been participating the Blizzard Challenge speech synthesis evalution for ten years since 2006. In 2006, our first HMM-based statistical parametric speech synthesis system was submitted [1]. In the coming year, to pursue a better performance of our system, an HMM-based unit selection and waveform concatenation method was adopted in the process of system construction because of the large scale of the provided corpus [2]. A new acoustic model clustering approach was used to optimize the scale of the decision tree using cross validation and minimal generation error (MGE) criterion automatically in 2009 [3]. In the process of system construction in 2010, the model covariance matrices were globally tied in order to reduce the footprint of the model as well as improving the efficiency of the model training. To further improve the quality of the synthetic system, syllable-level F0 model was introduced to evaluate the pitch combination of two adjacent syllables [4]. In Blizzard Challenge 2011 [5], a maximum log likelihood ratio (LLR) criterion was proposed to improve the unit selection criterion of the speech synthesis system. A more difficult task was emerging as the corpus in Blizzard Challenge 2012 [6] was non-standard, so the corpus filtering, channel equalization and sentence labelling for expressive sentences technique were used to deal with that kind of database. Besides the different kinds of audio-book style of the corpus, the database consist of unsegmented English recordings in Blizzard Challenge 2013. Taking these factors into consideration, a novel technique of synthesis quality prediction module was added into our existed unit selection speech synthesis in Blizzard Challenge 2013 [7]. In 2013, we also constructed our first systems for Indian languages, using a letter-to-sound (L2S) [8] approach, for which there were no available front-end text precessing module in our system. In Blizzard Challenge 2014, we focused on six Indian languages which are totally unfamiliar to our research team. We built HMM-based parametric systems for all these languages. The data-driven post-filtering techniques, the deep neural network (DNN) [9] and modulation spectrum [10] based ones, were adopted to enhance the quality of synthetic speech. We also built an HMM-based unit selection system for the Hindi language task, because we can access a front-end text processor for this language. For the IH2 task, an English to Hindi pronunciation prediction module was built using L2S approach and the English words in other languages were transliterated manually by native speakers.

This year in Blizzard Challenge 2015, the tasks were similar to that of last year. Larger size of corpus are available for three languages (Hindi, Tamil and Telugu) that has been evalated in Blizzarc Challenge 2014, and three new languages (Marathi, Bengali and Malayalam) with speech data of 2 hours for each were included [11]. Similar with our previous entries, we built HMM-based unit selection system for Hindi tasks. A new pre-selection method for non-uniform units was adopted in our system. For the other five Indian languages, we constructed the HMM-based statistical parametric speech synthesis systems to generate the test sentences.

This paper is organized as follows: In section 2, baseline system of USTC unit selection system and statistical parametric speech synthesis system was reviewed. In section 3, the system construction details in Blizzard Challenge 2015 were elaborated. In section 4, the evaluation results of our proposed system are shown and further analysis are given accordingly. Lastly in section 5, some conclusions and potential future research are given.

2. Baseline systems

Similar to Blizzard Challenge 2014, the main task is to build systems of two kind, mono-lingual and multi-lingual systems for Hindi, Bengali, Malayalam, Marathi, Tamil and Telegu. In Blizzard Challenge 2015, three Indian languages, Bengali, Malayalam and Marathi are newly added languages for constructing speech synthesis system, which further requests our synthesis system be more adaptive concerning different languages and limited knowledge on them. Two tasks are evaluated, including:

- IH1: Mono-lingual Synthesis System,
- IH2: Multi-lingual (Indian and English) Synthesis System.

Because of the availability of a front-end text processing

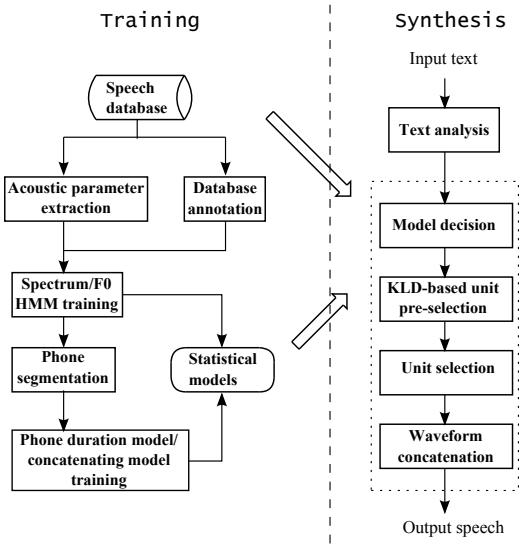

Figure 1: The flowchart of USTC unit selection system.

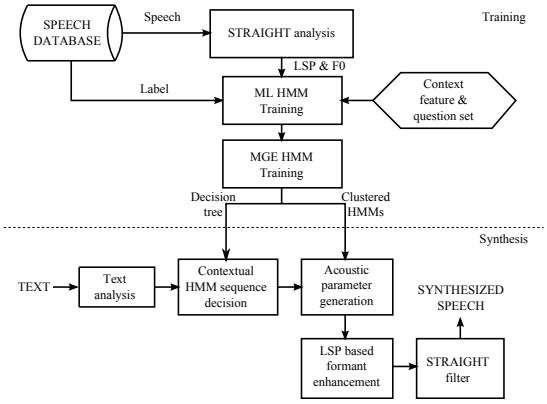

Figure 2: Flowchart of the HMM-based statistical parametric speech synthesis system.

module for Hindi language[1] and the relatively large speech corpus of 4 hours, an HMM-based unit selection speech synthesis system was built for Hindi. As for the other five Indian languages, the HMM-based statistical parametric speech synthesis systems were constructed.

A brief review of USTC HMM-based unit selection and statistical parametric methods will be given in this section respectively.

2.1. The USTC unit selection system

As indicated in Figure 1, our HMM-based unit selection system consist of two main parts, namely the training phase and the synthesis phase.

2.1.1. Training phase

Several HMMs [12] of acoustic models are firstly trained to provide some guidelines in the process of unit selection. In total, six sets of HMMs, including spectral model, F0 models, phone duration models, concatenating spectral models, concatenating F0 models, as well as syllable-level F0 models, which guides the unit selection in a larger scale. For the spectral model, acoustic features of frame-level spectrum and F0 are used. Phone-duration models are trained using the phone-duration segmented by the spectral and F0 model, Concatenation models are trained to model the acoustic distributions of spectrum and F0 at the phone boundaries, such as the delta or delta-delta of spectrum and F0. F0 features from the vowels of two adjacent syllables are extracted to train the syllable-level F0 model to provide prosody modeling in a longer range. The typical and effective approach of multi-space probability HMMs (MSD-HMMs) [13] was adopted to model the continuous probability HMMs and the F0 feature. To deal with the data sparsity and predict the related model parameters, context-depended HMM training technique was used in the process of decision tree-based model clustering. Minimum description length (MDL) [14] based model clustering is applied to control the size of the decision trees. 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, a sequence of phone units are selected under the Maximum Likelihood criterion firstly, then, these units are concatenated to form synthetic speech, considering the smoothness of two consecutive units simultaneously. Let N be the number of phonemes 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

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

where λ_m indicates the acoustic models described in the previous section, and w_m corresponds to their weights, which were manually tuned on a development set, $\boldsymbol{X}(\boldsymbol{U}, m)$ and $C(\boldsymbol{U})$ extract corresponding acoustic features and context features from the unit, $D_m()$ denotes the Kullback-Leibler divergence (KLD) [15]. 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 [16] 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 systems. 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

[1] Provided by iFLYTEK Co., Ltd.

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 [17], LSP based formant enhancement method is adopted to improve the quality and articulation of generated speech quality.

2.3. Post-filtering for HMM-based parametric speech synthesis method

Although the HMM-based parametric speech synthesis system can synthesis stable speech, the quality of synthetic speech is degraded due to the fact that fine spectral structures of natural speech are lost to some degree by statistical averaging of the model, the synthesized speech still sounds "muffled". Therefore, an post-filtering methods on synthesized speech is essentially important to improve the speech quality. During our system construction, the recently proposed DNN-based stochastic post-filter [9, 18] is adopted. A generatively trained DNN is used to model the conditional distribution of natural spectral envelopes given the corresponding synthesized spectral envelopes. The final spectral envelopes were generated according to the conditional distributions to compensate the gap between synthetic and natural speech.

3. System building

The USTC system consists of two parts: front-end text processing part and back-end acoustic modeling and waveform concatenating part. The back-end part of our system was constructed as introduced in section 2.

3.1. System construction for Hindi language

Same as it in Blizzard Challenge 2014 [19], we used the iFLYTEK Hindi TTS engine to perform phoneme transcription and prosodic information tagging from the UTF-8 format input sentence released in the IH1 and IH2 databases.

All the sentences released this year, totally 1710 sentences, are used to build the HMM-based unit selection and waveform concatenation TTS system as described in previous section, including the text normalization module, Hindi pronunciation prediction module, prosody prediction modules, etc.

In our entry this year, two more improvement were made for front-end text processing:

- Naturalness improvement.
  When a break is predicted from the input text by the prosodic module, speech unit with boundary tone, final lengthening and silence will be selected in unit selection procedure of speech synthesis. In BC2015, we pay more attention to weaker segments. The weaker segments not only influence speech segments, it also influences the prosody. We found, in Hindi speech, that the high frequently appeared central vowel /ax/, which is a restorative vowel, has two different acoustic representations: normal and weak. The weaker form of /ax/ is always short. By labeling out the two different forms of /ax/, the differentiate /ax/ can be predicted by our L2S module. The synthesis speech demonstrates the positive result of this modification of phoneme set. This fact reflects that both lengthening and weakness of this phone have their contributions to the naturalness of Hindi synthesis speech.
- Using super phonetic symbol for Hindi and English text in the multi-lingual task.
  In input Hindi and English text, the Hindi words are predicted by normal Hindi L2S module. And the English words are predicted by English L2S model with super phonetic symbols defined in iFLYTEK multi-lingual module. Due to the consistency of Hindi phonetic symbol and the super phonetic symbol, there is no extra process between the concatenation of Hindi word pronunciation and English word pronunciation. The advantage of using super phonetic symbol is the smoothness between English word and Hindi word. The correctness or naturalness of English word pronunciation in the multi-lingual text is another problem.

In addition, in the back-end module of our system, we used non-uniform units for unit selection. This is similar with the one proposed in [20], where a Hierarchical Viterbi Algorithm (HVA) was proposed in the Viterbi search process. However, in our Hindi language system, the advantage of two consecutive units in the corpus was also taken into consideration in pre-selection process.

3.2. System construction of the other five languages

Similar systems were also constructed for the other five languages. Systems were built different from the Hindi system because of the lack of front-end knowledge and limited size speech database.

As for the missing of linguistic knowledge about these languages, we built the front-end module followings the rules listed below:

1) Text Segmentation
   Text of these five languages provided in the release training data was firstly cut into small segments simply according to the space in the text. These small segments, which we name as S-Word, are assume to be similar to the prosodic word in English.
2) Unicode Transformation
   S-Words consist of several Unicode char sets which represent the vowel or consonant phonemes of the language. The Unicode char set within the S-Word is similar to phonemes. Therefore we name them as S- Phone.
3) Context Information Generation
   Once the S-words and S-Phones are got, the contextual information of the text is generated based on the type of S-Phone (vowel or consonant) and the position information of the S-Phone and S-Word, which includes: vowel/consonant tag of the current S-Phone; forward and backward position of the current S-Phone in current S-Word; forward and backward position of current S-Word within the current punctuation part, punctuation parts are defined as the text segments segmented by the punctuation; forward and backward position of current punctuation part in the sentence; previous and next S-Phone context information.

For the test sets in IH2, as the input text contains both Indian and English text, we first extract the English parts. Then the English parts in the text were substituted with transliteration results. In the transliteration process, the Google Transliteration Tool[2] was used in our system.

[2]http://google-input-tools-for-windows.en.softonic.com

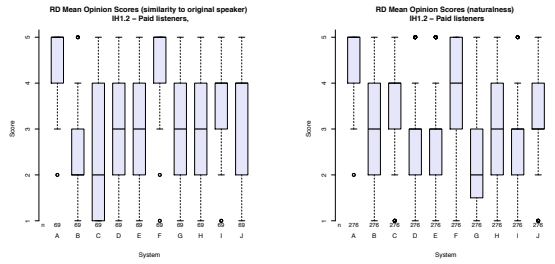

Figure 3: Evaluation results of IH1.2 (Hindi) task on similarity and naturalness.

Because of the small sizes of the released speech dataset, the stability of synthetic speech can't be guaranteed when adopting the unit selection approach in the back-end model. Instead, HMM-based parametric speech synthesis system was constructed for these five languages. To further improve the quality of the synthetic speech, post-filtering for HMM-based parametric speech synthesis method was used. In our system, the recently proposed deep neural network (DNN)-based stochastic post-filter [18] was applied to improve the naturalness of synthetic speech. The DNN is built with high-dimensional raw spectral envelopes extracted by STRAIGHT [17]. A generatively trained DNN is used to model the conditional distribution of natural spectral envelopes given the corresponding synthesized spectral envelopes. The size of input and output layers were 513, each corresponds to one frequency bin of spectral envelope. There were six hidden layers, each included 512 units. The final spectral envelope for synthesizing speech waveform were predicted by the conditional distribution given by the DNN post-filter.

Note that we simply used the Unicode as phoneme in our system for these languages, there could be errors in synthesized speech. This could affect the performance of the DNN-based postfilter, which is trained on paired-frames of synthetic and natural spectral envelopes.

4. Evaluation

In this section, the evaluation results of our submitted systems are presented and analyzed. Among the A to J systems, F represents our entry and A denotes the natural speech.

4.1. Results of IH1.2 and IH2.2 tasks (Hindi)

As being stated above, since we have a standard front-end text processing module for Hindi language, the system for Hindi language was thus built in different way from the other languages, we first discuss the evaluation results of IH1.2 and IH2.2 tasks.

As can be seen in Fig. 3, our Hindi unit selection speech synthesis system achieved the best performance in the similarity and naturalness, which is significantly better than the second better ones (F and C) considering the similarity and naturalness respectively. Comparing with the results in Blizzard Challenge 2014, the similarity of our Hindi system has improved significantly from 4.0 to 4.3. Also, the naturalness has improved from 3.6 to 3.9, which might be attributed to the new unit selection method, non-uniform unit based unit selection criteria and the improved strategy of front-end text processing in addition to the larger corpus. According to the evaluation results of IH1.2 (Hindi) task on WER, entry F achieved the lowest word error

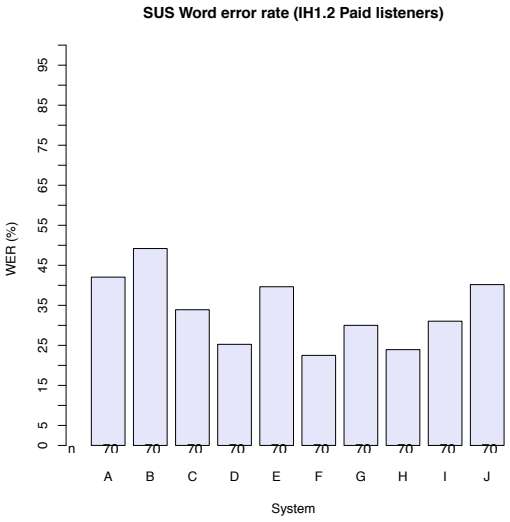

Figure 4: Evaluation results of IH1.2 (Hindi) task on WER.

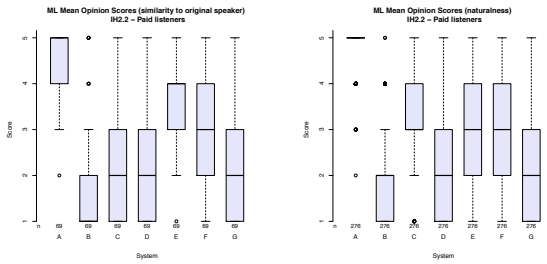

Figure 5: Evaluation results of IH2.2 task on similarity and naturalness.

rate of 23%, which further verifies that the USTC HMM-based unit selection system are of high intelligibility. However, there is an issue with the WER results that the WER of natural speech is quite high (higher than most of the submitted systems), which happened in many languages.

In IH2.2 task, our system ranks at second and third in similarity and naturalness among all the submitted systems. Comparing with the IH1.2 task, the slightly degradation in similarity and naturalness of our system may be attributed to the fact that the mixed language brings some defections in front-end prosody analysis. In IH2.2, we use the traditional method to transform the English pronunciation to Indian pronunciation, which may be different from the the real mixed language pronunciation style. The accuracy of the L2S module for predicting English pronunciation may also affect the quality of synthetic English words.

4.2. Evaluation results of the other languages

The evaluation results of the Bengali, Malayalam, Marathi, Tamil and Telugu languages are shown in Fig. 6, Fig. 7, Fig. 8, Fig. 9 and Fig. 10 respectively.

As we can see from these figures, the USTC system achieves relatively good results in Marathi tasks. On the evaluation of similarity, we achieved the highest MOS in IH1.4 and second in IH2.4. On the evaluation of naturalness, we achieved

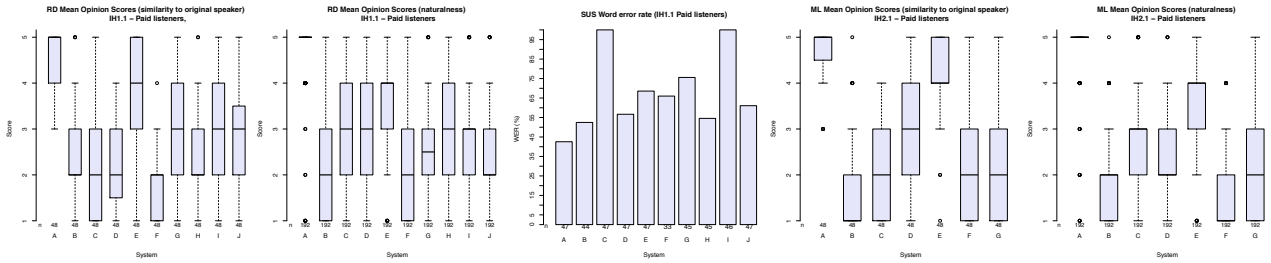

Figure 6: Boxplot of similarity, naturalness and WER test on Bengali language task.

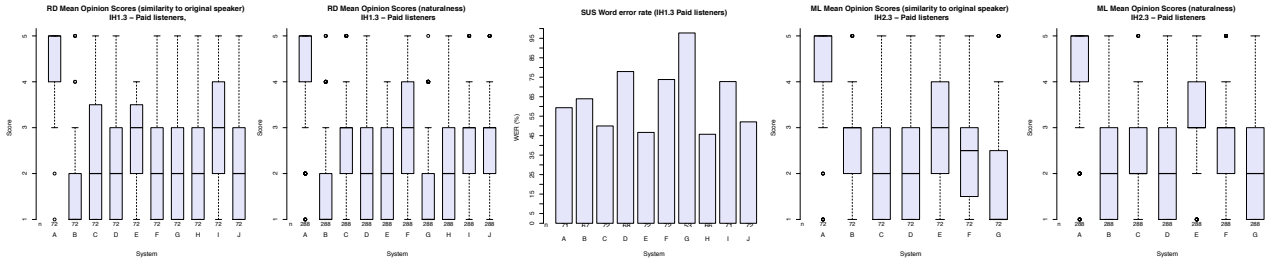

Figure 7: Boxplot of similarity, naturalness and WER test on Malayalam language task.

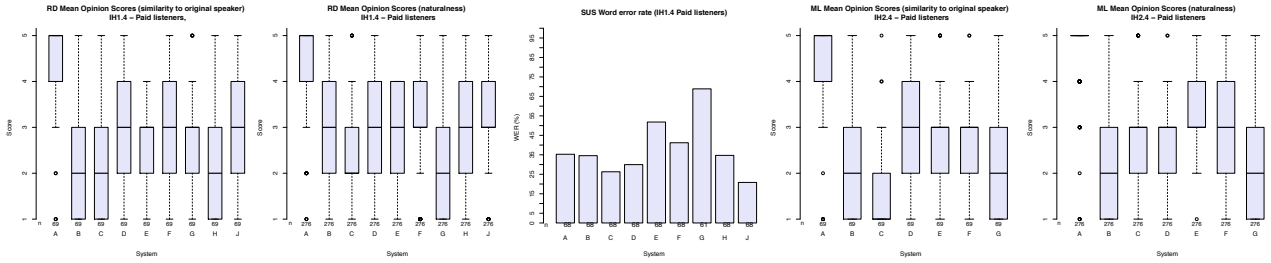

Figure 8: Boxplot of similarity, naturalness and WER test on Marathi language task.

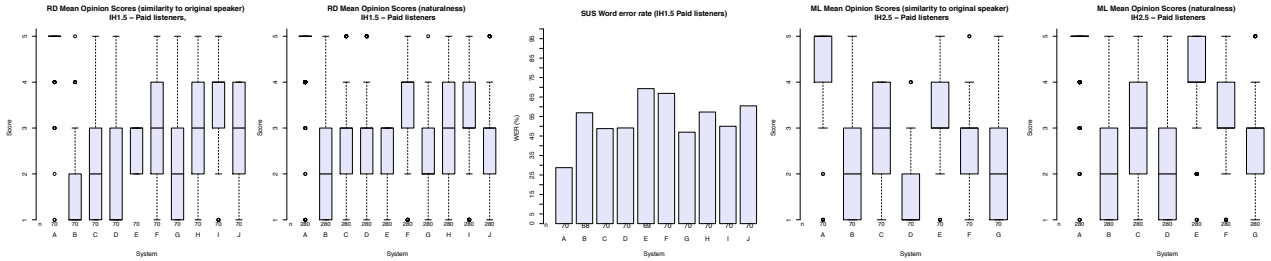

Figure 9: Boxplot of similarity, naturalness and WER test on Tamil language task.

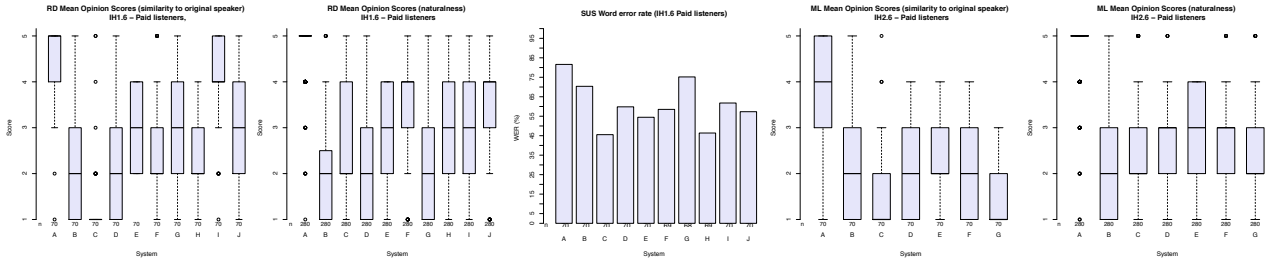

Figure 10: Boxplot of similarity, naturalness and WER test on Telugu language task.

the highest MOS on RD tasks of IH1.3, IH1.5 and IH1.5, and achieved the second highest MOS on IH1.4. There isn't significant difference between the best system and our system on WER of IH1.4 and naturalness of IH2.4. In the Malayalam, Tamil and Telugu tasks, our system achieved median performance in IH1 and IH2 tasks.

Although we built the five systems exactly in the same way, it is noticed that the performance of our system on Bengali language is very poor. We have built Bengali system in Blizzard Challenge 2013, in which the evaluation scores of our system was better than average ones. The difference in our system this year is that we used the Unicode instead of real phonemes to directly build our system. Therefore, we assume that there is a weak correlation between the Unicode and phonemes in the language of Bengali. The quality of synthetic speech is greatly degraded by our front-end strategy.

5. Conclusions

This paper gives the details of USTC's Indian languages system construction for the Blizzard Challenge 2015. Being able to access a standard front-end text processing module for Hindi language, the HMM-based unit selection approach has been adopted for IH1.2 and IH2.2 (Hindi) tasks. As for other five languages, the HMM-based statistical parametric speech synthesis method was used. The DNN based post-filtering technique was used in the spectral domain to enhance the quality of the synthesized speech. The evaluation results show the effectiveness of our submitted system, especially our HMM-based unit selection synthesis method. Still, there are still some remaining issues need further investigation in the Indian speech synthesis tasks, such as the front-end process technique for different Indian languages and the multi-lingual synthesis tasks.

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] L.-H. Chen, Z.-H. Ling, Y. Jiang, Y. Song, X.-J. Xia, Y.-Q. Zu, R.-Q. Yan, and L.-R. Dai, "The USTC system for blizzard challenge 2013," in *Blizzard Challenge Workshop*, 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] L.-H. Chen, T. Ratio, C. Valentini-Botinhao, Y. Junichi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," in *Proc. Interspeech*, 2014.

[10] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, 2014.

[11] H. Patil, T. Patel, N. Shah, H. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. Kishore, S. Prasanna, N. Adiga, S. Singh, K. Anand, P. Kumar, B. Singh, S. Binil Kumar, T. Bhadran, T. Sajini, A. Saha, T. Basu, K. Rao, N. Narendra, A. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. Murthy, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Proc. O-COCOSDA*, 2013, pp. 1–8.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 2003–2014, Nov 2015.

[19] L.-H. Chen, Z.-H. Ling, Y.-Q. Zu, R.-Q. Yan, Y. Jiang, X.-J. Xia, and Y. Wang, "The USTC system for blizzard challenge 2014," in *Blizzard Challenge Workshop*, 2014.

[20] R. Zhang, Z.-Q. Wen, J.-H. Tao, Y. Li, B. Liu, and X.-Y. Lou, "A hierarchical viterbi algorithm for mandarin hybrid speech synthesis system," in *Proc. Interspeech*, 2014, pp. 795–799.