The USTC System for Blizzard Challenge 2014

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

[†]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

This paper introduces the speech synthesis system developed by USTC for Blizzard Challenge 2014. Six Indian languages were evaluated this year, including Assamese, Gujarati, Hindi, Rajasthani, Tamil and Telugu. Two tasks were built for these languages: the mono-lingual task (IH1 hub task) and the multi-lingual task (IH2 spoken task). We submitted entries to both tasks in all languages. We submitted two entries for evaluation: the primary entry and the secondary entry. In our primary entry, a hidden Markov model (HMM)-based unit selection system was built for Hindi language and HMM-based parametric speech synthesis systems were built for the remaining five languages. In the secondary entry, only an HMM-based parametric speech synthesis system was built for Hindi language. The evaluation results show the effectiveness of our submitted systems.
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. In Blizzard Challenge 2013 [7], we built a system on a large scale of unsegmented English audiobook database using unit selection approach with a synthesis quality prediction method. We also constructed our first system on Indian languages using letter-to-sound (L2S) [8] approach since there were no available front-end text precessing module in our system.

New challenges were proposed this year in Blizzard Challenge 2014. We have to construct speech synthesis systems for 6 Indian languages. In additional, we have to construct multi-lingual systems using mono-lingual dataset. This quite challenging because none of our team members is familiar with these languages. We built HMM-based parametric systems for all these languages. Some post-filtering techniques, such as stochastic deep neural network (DNN) [9] and modulation spectrum [10] based post-filtering methods, were adopted to enhance the quality of synthesized 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 tasks, 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 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 2014 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

Different from the previous challenges, the aim of the Blizzard Challenge 2014 is to built speech synthesis systems on six Indian dialects: Assamese, Gujarati, Hindi, Rajasthani, Tamil and Telugu. Two tasks were evaluated:

- IH1: Building mono-lingual speech synthesis system;
- IH2: Building multi-lingual system (Indian and English) tasks on the same dataset as IH1 task;

Since our USTC team can access an front-end text processing module for Hindi language from iFLYTEK Co., Ltd., we can built an HMM-based unit selection system for this language. The systems for the remaining 5 languages were constructed using HMM-based statistical parametric speech synthesis method. Therefore, in this section, we will firstly have a brief review on USTC HMM-based unit selection and statistical parametric approaches respectively.

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.

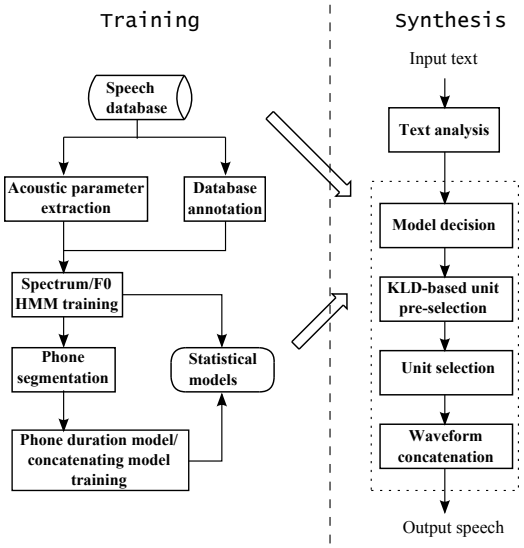

Figure 1: The flowchart of USTC unit selection system.

2.1.1. Training phase

First, at the training phase, HMMs [11] 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) [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 model parameters for the unseen context at the synthesis phase. Minimum description length (MDL) [13] 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

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

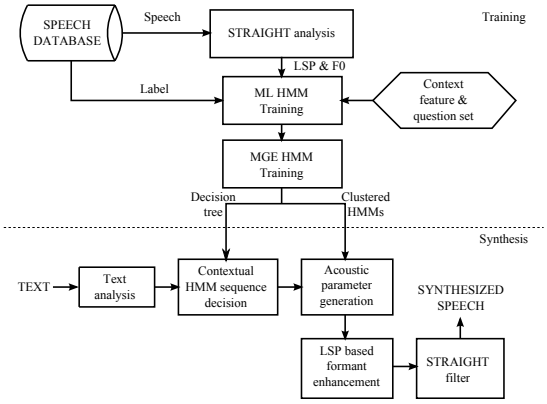

Figure 2: Flowchart of the HMM-based statistical parametric speech synthesis system.

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) [14]. 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 [15] 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 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 [16], 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

The HMM-based parametric speech synthesis method can generate speech stably. However, the synthesized speech still sounds "muffled" due to the fact that fine spectral structures of natural speech are partly lost by statistical averaging of the model. Therefore, we adopted two post-filtering methods on synthesized speech in order to improve the its quality. The first one is the deep neural network (DNN)-based stochastic post-filter [9]. The DNN is built on the high-dimensional raw spectral envelopes extracted by the STRAIGHT vocoder. A gener-

Table 1: Details of the submitted systems for each Language.

Language	Entry	System	spectral feature	Post-filter
Assamese	D	parametric	mel-cepstra	MS
Gujarati	D	parametric	LSP	DNN
Hindi	D	unit selection	mel-cepstra	–
	K	parametric	LSP	DNN
Rajasthani	D	parametric	mel-cepstra	MS
Tamil	D	parametric	LSF	DNN
Telugu	D	parametric	LSF	DNN

atively trained DNN is used to model the conditional distribution of natural spectral envelopes given the corresponding synthesized spectral envelopes. The second one is the modulation spectrum (MS)-based speech enhancement [10] which aims to enhance the natural frequency modulation in the spectral parameter trajectories. The enhancement is constructed based on the empirical findings of acoustic differences between synthesized and natural speech trajectories. Note that the MS-based approach was applied on HMM-based systems with mel-cepstra as acoustic features, because it didn't work effectively on LSP features in our internal experiments.

3. System building

The USTC system consists of two parts: front-end text processing part and back-end acoustic modeling and waveform generation part. The back-end part of our system was constructed as introduced in section 2. Table. 1 shows some details of our system for each language. Note that two post-filtering methods were adopted for different languages. The DNN approach is a frame-level method. It performs well if the sentences in the training set are correctly synthesized. Because of the simple text processing, there was a mismatch between the synthesized and natural speech in Assamese and Rajasthani. Therefore, the sentence level trajectory-based post-filtering approach, modulation spectrum, was adopted for these two languages. The rest of this session introduces the text processing methods in our system.

3.1. System construction of Hindi language

3.1.1. Hindi Front-End Processing For IH1 and IH2 Tasks

We used the iFLYTEK Hindi text-to-speech (TTS) engine to perform phoneme transcription and prosodic information tagging from the UTF-8 format input sentence released from the IH1 and IH2 databases. The Hindi TTS engine includes a text normalization module to transfer abbreviation, digits, time, data and etc. to Hindi orthographic texts, and an English to Hindi pronunciation prediction module to process mixed English words. For the multi-lingual task, an English to Hindi pronunciation prediction module was constructed in the following steps:

1) Extract high frequency English words based on large Hindi-English mixed text;
2) Acquire English word pronunciation by an English L2S model with super phonetic symbols;
3) Mix Hindi and English words in lexicon;
4) Retrain the L2S models by the decision tree based algorithm;

Table 2: Internal evaluation results by twenty native speakers.

	IH1.3			IH2.3
	Nat.	Sim.	WER	Nat.
statistical parametric	3.04	2.4	6.92%	2.83
unit selection	4	3.9	4.96%	3.14

5) Use the optimized L2S models to predict the pronunciation for not-in English words.

3.1.2. Hindi Speech Database Annotation

High quality speech database annotation is a key component to the performance of a speech synthesis system. We establish database annotations as the following steps:

1) Extract pitch from waveform and conduct energy normalization;
2) Segment waveform initially via the force alignment technology;
3) Automatically label prosody including break and accent, and verify the results manually;
4) Retrain the acoustic models based on fine tuned annotation.

3.1.3. Acoustic modeling

In Blizzard challenge 2013, we have built a Hindi TTS system using HMM-based parametric speech synthesis method, because of the small amount of training data and our lack of knowledge about this language. This time in Blizzard challenge 2014, we built two kind of synthesis system for Hindi, HMM-based unit selection speech synthesis method and HMM-based statistical parametric speech synthesis method, and compared their performance in our internal experiments. It is very difficult to build a robust unit selection system on a two hours Hindi dataset. Therefore we established some strict expert rules in the unit pre-selection and target cost calculation, and empirically increased the number of candidate units for Viterbi search.

In our internal subjective evaluation, twenty native students were invited to participate in a simulation test. The results are shown in Table 2. Unit selection system significantly outperformed the statistical parametric system in similarity and naturalness in mono-lingual test. However, its performance decreased greatly when it comes to multi-lingual test. Because the context of English are missing at the training stage of target model and concatenation model. The context of multi-lingual input text is inconsistent with the acoustic models. The word error rate (WER) of the unit selection system is lower than that of the statistical parametric system. This can be attribute to the meaningful testing text we used to generate speech samples for this evaluation.

3.2. System construction of the other five languages

The systems for the other five languages were built in almost the same way, in which the acoustic model training and back-end speech synthesis process were the same as introduced in the previous section. The front-end process is built in the following steps:

1) The first step is text segmentation. Text of these five languages provided in the release training data was firstly

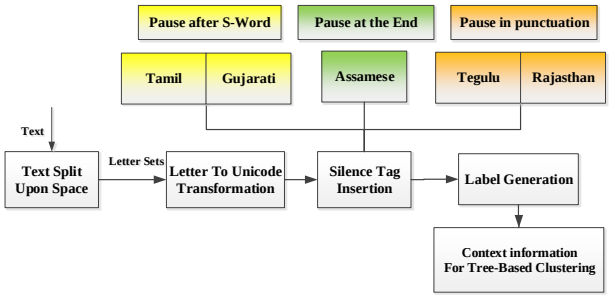

Figure 3: The front-end context information generation process for Tamil, Gujarati, Telugu, Rajasthani and Assamese languages.

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) In the second step, all the characters of the text are converted into Unicode format. S-Words consist of several Unicode char sets which represent the vowel or consonant letter of the language. The Unicode char set within the S-Word is similar to phonemes. Therefore we name them as S-Phone.
3) In the last step, labels, which contain context information, for each sentence were generated. The main context information include the followings:
   - 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.

Note that for different language, speech pause may occur differently. According to the punctuations in the training text and speech pause frequency in the training speech data, we inserted silence tags, which can be regarded as an S-Phones, in different position of different languages. In details, silence tags were inserted after each S-Word, each S-Word, position of punctuation, position of punctuation, end of sentence for Tamil, Gujarati, Telugu, Rajasthani and Assamese respectively. The front-end context information generation process is shown in Fig. 3.

For the IH2 tasks, since the front-end text process were built on rules without any knowledge about these languages, the English text in the test sentences were substituted with transliteration text by native speakers.

4. Evaluation

This section discusses the evaluation results of our systems. Among all the systems, D is the identifier of our primary entry and K is our secondary entry. A is the natural speech. We submitted entries of all 6 languages. In this section, we will present and discuss the evaluation results of our systems.

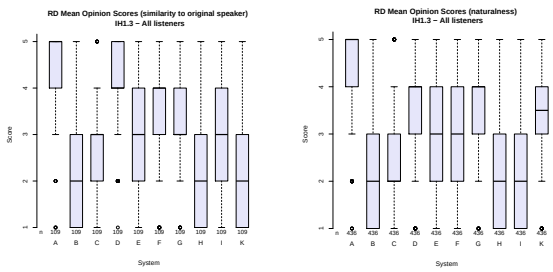

Figure 4: Evaluation results of IH1.3 task on similarity and naturalness.

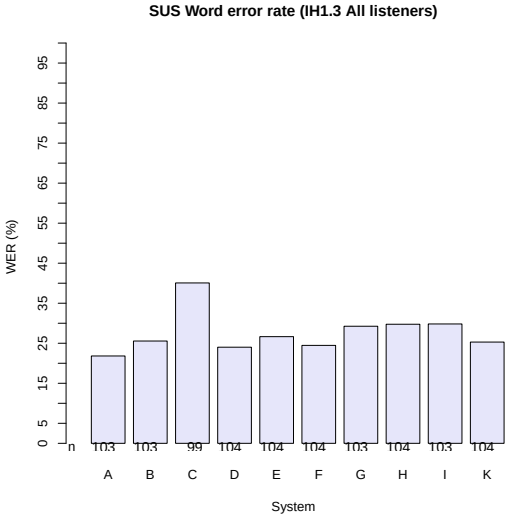

Figure 5: Evaluation results of IH1.3 task on WER.

4.1. Results of IH1.3 and IH2.3 tasks (Hindi)

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 IH1.3 and IH2.3 tasks.

Fig. 4 shows the evaluation results of similarity and naturalness. We can see that our unit selection system (D) achieved the best perform in similarity and it is significantly better than the other systems. This benefits from the effective unit selection algorithm and waveform concatenation of our system. Our HMM-based parametric system (K) does not performed well in the similarity test. This can be attributed to the post-filtering technique we used to enhance speech quality. This process makes the generated speech unlike the original speaker. In the naturalness test, our D and K system also perform well, they achieved the second and third highest MOS among all systems. And there is no significant difference between our D system and the best system (G) in the pairwise Wilcoxon signed rank tests. In the intelligibility evaluation, as shown in Fig. 5, our D system achieved the lowest word error rate (WER), which is very close to the natural speech, although there isn't significant difference among all the systems except system C. It is interesting to see that the our unit selection approach outperformed our statistical parametric approach in WER. This is different from the results of the previous challenges because meaningful

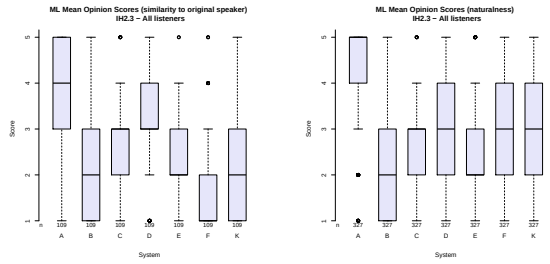

Figure 6: Evaluation results of IH2.3 task on similarity and naturalness.

sentences were evaluated instead of semantically unpredictable sentences (SUS) this year. The evaluation results of IH2.3 tasks on similarity and naturalness are shown in Fig. 6. We see that our systems still performs well on these tests. But the scores in IH2.3 tasks are much lower than those in IH1.3 tasks. These results are consistent with our internal evaluation results as shown in Table 2.

4.2. Evaluation results of the other languages

The evaluation results of the Assamese, Rajasthani, Tamil and Telugu languages are shown in Fig. 7, Fig. 9, Fig. 10 and Fig. 11 respectively. As we can see from these figures, in Assamese tasks, the USTC system achieves quite good results. We achieved the highest MOS on similarity on both IH1.1 and IIH2.1 evaluation and second highest MOS on naturalness of both evaluations. There isn't significant difference between the best system and our system on WER of IH1.1 and naturalness of IH2.1. Our system also performs well on Rajasthani tasks, it performed the best on the similarity of IH4.2. In the other evaluations, our system is not significantly worse than the best systems. In the Tamil and Telugu tasks, our system achieved median performance in IH1 tasks. It can be seen that the similarities of the systems using DNN post-filtering are not good because of our spectral enhancement processing on natural speech, which is the training data for the output of DNN post-filter. The performance of Gujarati task is not satisfying as other languages. Our system performed the best in IH2 tasks as we did in IH2 tasks in all languages except the IH2.2 task. The performance of our system on these five languages are generally worse than that of Hindi because of our simple front-end text processing strategy.

5. Conclusions

This paper presents the details of constructing the USTC system for the Blizzard Challenge 2014. The HMM based unit selection approach has been adopted for Hindi tasks and the HMM based statistical parametric speech synthesis approach has been adopted for tasks in all languages. Since we can access a standard front-end text processing module only for Hindi languages, the system for other languages were built with a front-end text processing module manually designed from text of training data. Two post-filtering methods, DNN and MS based approach, were adopted to enhance the quality of generated speech in spectral domain. The evaluation results show the effectiveness of our system in some aspects.

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*, 2012.

[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] 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.

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

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

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

[16] 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.

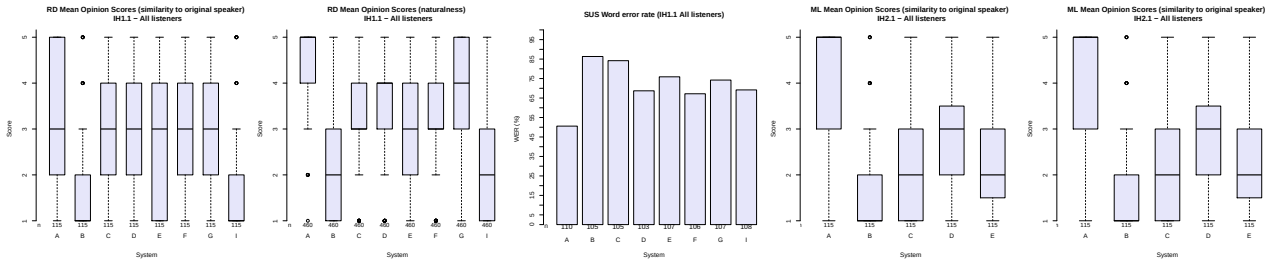

Figure 7: Boxplot of similarity, naturalness and WER test on Assamese language task.

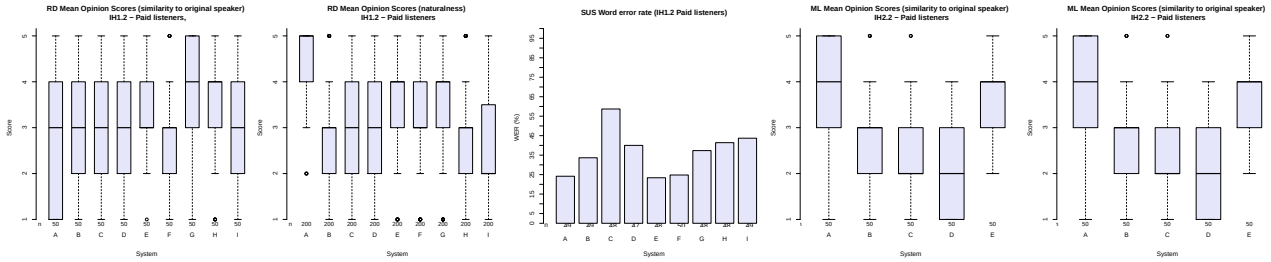

Figure 8: Boxplot of similarity, naturalness and WER test on Gujarati language task.

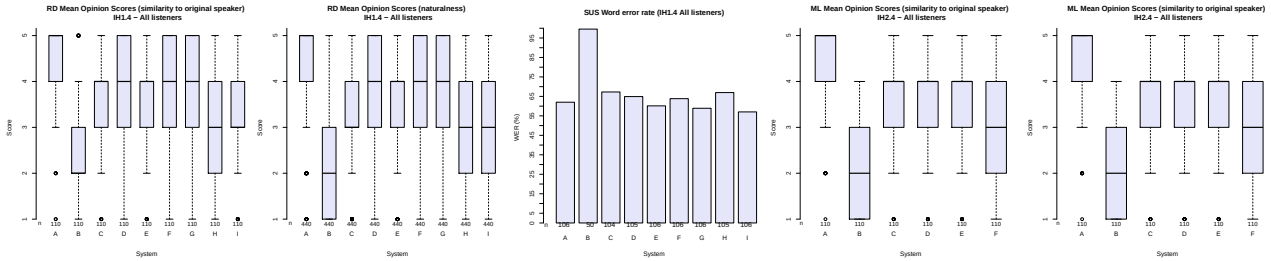

Figure 9: Boxplot of similarity, naturalness and WER test on Rajasthani language task.

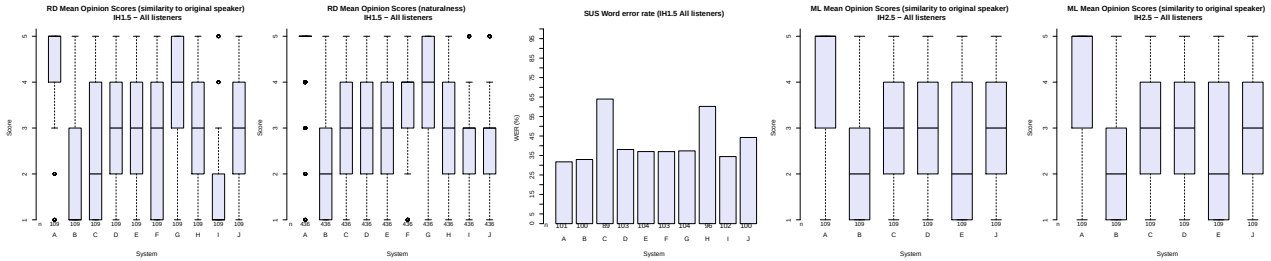

Figure 10: Boxplot of similarity, naturalness and WER test on Tamil language task.

Figure 11: Boxplot of similarity, naturalness and WER test on Telugu language task.