MILE TTS for Tamil and Kannada for blizzard challenge 2013

[1]Shiva Kumar H R, [1]Ashwini J K, [1]Rajaram B S R, [1]A G Ramakrishnan

[1]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
ashwinijk,rajaram,{@mile.ee.iisc.ernet.in}, shivahr,ramkaig {@ee.iisc.ernet.in}

Abstract

We participated in the Blizzard Challenge 2013 for Tamil and Kannada, using our unit selection based concatenative speech synthesis system. Sentence level viterbi search is used to select the reliable speech units among a set of candidate units. The same Text-to-Speech synthesis framework is used to synthesize speech in both the Indian languages. The given Wikipedia and semantically unpredictable test sentences are synthesized using IIIT-H Indic corpus and the listening test results reported by the blizzard evaluation team is discussed. The letter code for MILE TTS is “R”.

Index Terms: speech synthesis, unit selection, joint costs, blizzard challenge.

1. Introduction

MILE lab is making its maiden entry for the annual blizzard challenge. This year’s participation is on the Indian languages tasks 2013-IH1.3 Kannada and 2013-IH1.4 Tamil using the given one hour of speech data and corresponding text in UTF-8 format. Two types of test sentences were given for synthesizing task; sentences from Wikipedia (wpd) and semantically unpredictable sentences (sus). To build the voice, the test sentence is first converted to its phoneme equivalent and then split into the required target units. The target units are searched using a set of search rules from the synthesis database, which is created using wave data, pitch and label files. Word level and sentence level viterbi search were explored to build voice but there was no significant difference in the performance between those two approaches and the later approach is used to synthesize the test sentences.

The main intended application of Text-to-Speech synthesis (TTS) system developed at MILE lab is to build automated book reader to assist visually challenged people to access material printed in Kannada and Tamil languages, including interspersed English words and preferably extending to other Indic languages. Users would take a snap of the printed material using their mobile camera, and the ABR installed on the users’ mobile would perform optical character recognition and read aloud the recognized text using a Text-to-Speech synthesizer.

Section.2 gives an overview of MILE TTS engine. Section.3 briefly describes the steps followed to build voices. Results of blizzard listening test are discussed in section.4.

2. Description of MILE TTS Engine

MILE TTS engine is built on concatenative speech synthesis [1] framework. The optimal units are selected by Viterbi search considering the lowest total join cost for a sentence. Viterbi search was carried out initially at the word level; later, it was implemented at the sentence level but with no apparent improvement in the performance. The block diagram of MILE TTS is shown in Fig.1.

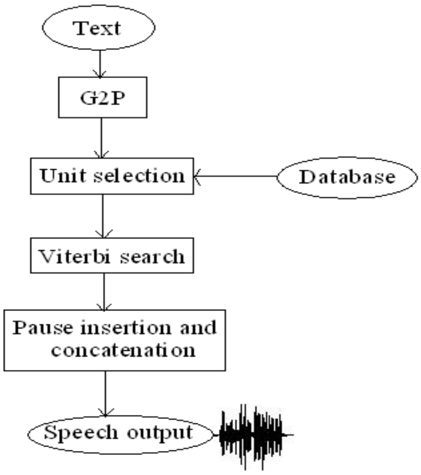

Fig.1. Block Diagram of MILE TTS

2.1. Database creation

MILE TTS engine utilizes a database having duration information of each phoneme and corresponding wave data information. Using only duration information, it is able to synthesize reasonably good speech due to the appropriate selection of units from the database.

2.2. Viterbi search

Among the selected candidate units, the one that best matches the target unit is selected by viterbi search for candidate units with minimum total cost. Total cost is the sum of concatenation and target costs. Target cost is neglected in the current version of the MILE TTS.

The concatenation cost, $C^c(u_{i-1},u_i)$ is determined by the weighted sum of q concatenation sub-costs, $C_j^c(u_{i-1},u_i)(j=1,....,q)$. The sub-cost of concatenation cost can be broadly grouped under spectral and pitch based features. Here only the pitch based feature is used to compute concatenation cost. The continuity metric method proposed in [2] is used to derive pitch based feature.

$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q} w^c_{j(=1pitch)} C^c_{j(=1pitch)}(u_{i-1},u_i) \qquad (1)$$

$$C^c_{j(=1pitch)}(u_{i-1},u_i)=\sqrt{\sum_{k=-K}^{K} |p_{i-1}(k)-p_i(k)|^2} \qquad (2)$$

where, $p_i(k)$ is the average pitch value of the k^{th} frame from the i^{th} unit concatenation boundary. K is the number of frames employed on either side of the concatenation boundary. The value K=0 represents the matching based only on the frames at the concatenation boundary. Value of K is limited by the duration of the sub-word unit and it has been experimentally found in [2] that K=4 is sufficient for this application.

Word level and sentence level viterbi search:

For word level viterbi, the candidate units are the selected units from the database for a word. For sentence level viterbi, candidate units are the selected units from the database for a word and across words. The MOS of ten listeners for word level viterbi and sentence level viterbi for synthesized Wikipedia test sentences did not show appreciable difference.

2.3. Pause insertion and speech output

After selection of units using viterbi search, a fixed pause is inserted between the end of previous word and beginning of the current word. POS tagging is under work and hence not included in this blizzard challenge test sentence synthesis. After selecting reliable units for all the test sentences, wave data is loaded to create wave file.

3. Steps followed to build voices

About an hour of speech data, which is constituted by 1000 wave files and corresponding text files were released for Blizzard challenge Indian language task. In addition to these, we also used the label files which were made available online by Blizzard team. Following modifications were made to the label files:

- Segmentation was cross verified and re-segmented wherever required manually.
- Corrected mistakes in phoneme transcription to match the sound recording in both languages.
- Multiple silences were labeled in silence region and they are replaced as a single silence region.

3.1. Adaptation to MILE conventions

We have used different phoneme transcription conventions and hence all the phoneme transcription conventions used in IIIT-H Indic corpus were mapped to MILE conventions. In addition, the following change was also made:

Special Alphabets in Kannada: Alphabets ತ್ರ (tra) and ಟ್ರ (t:ra) in Kannada were labeled as separate units in IIIT-H Indic corpus, but we are handling them similar to ಕ್ರ, ದ್ರ, ಪ್ರ.

3.2. Voice building

With all the modifications discussed in Sec. 3.1, database is created with wave files, label files and pitch files and voices synthesized for the given set of test sentences.

4. Results and Discussion

The Blizzard challenge results are discussed in this section. MILE TTS identifier letter is R and A corresponds to the natural speech.

4.1. Similarity test

The boxplot of MOS on similarity to original speaker obtained by the Blizzard evaluation committee for Kannada (IH1.3) and Tamil (IH1.4) engines is shown in Fig.2. The left and right column in Fig.2 shows the MOS of All listeners, paid listeners and online volunteers for Kannada and Tamil, respectively. For both the languages, our system performance is one among the best and is consistent for all types of listeners. Especially, the MOS of online volunteers in Kannada is encouraging. The system performance can be attributed to the unit selection based concatenative speech synthesis approach.

4.2. Naturalness test

The boxplot of MOS on naturalness obtained by the Blizzard evaluation committee for the different Kannada (IH1.3) and Tamil (IH1.4) engines is shown in Fig.3. The left and right column in Fig.3 shows the MOS of All listeners, paid listeners and online volunteers for Kannada and Tamil, respectively. For both the languages, our system performance is among the second best and is consistent for all types of listeners. Since not much of signal processing is included in the current version of MILE TTS, the naturalness is lower than others. At the same time, we have maintained a minimum MOS value of around 3 for all types of listeners, for both the languages in naturalness and similarity tests.

The web demo of MILE TTS for both Tamil and Kannada are available at http://mile.ee.iisc.ernet.in/tts. A link for Indic Keyboard interface, an open source Indic script input software developed by MILE Lab is also provided at the demo site, which enables the users to input Tamil and/or Kannada text in Unicode. The text can also be copied and pasted from any website supporting Unicode Tamil and/or Kannada text.

5. Acknowledgements

We thank Simon King and Kishore Prahallad for conducting this Indic Blizzard Challenge, thus creating an opportunity for us to participate and present our TTS results.

6. References

[1] Andrew *J.* Hunt and Alan *W.* Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, IEEE International Conference on Acoustic, Speech and Signal processing (ICASSP’96) Atlanta, Georgia, May 7-8, 1996.

[2] Vikram Ramesh Lrakkavalli, Arulmozhi P and A G Ramakrishnan, “Continuity Metric for Unit Selection based Text-to-Speech Synthesis,” IEEE International Conference on Signal Processing & Communications (SPCOM 2010), 2010, Bangalore, India.

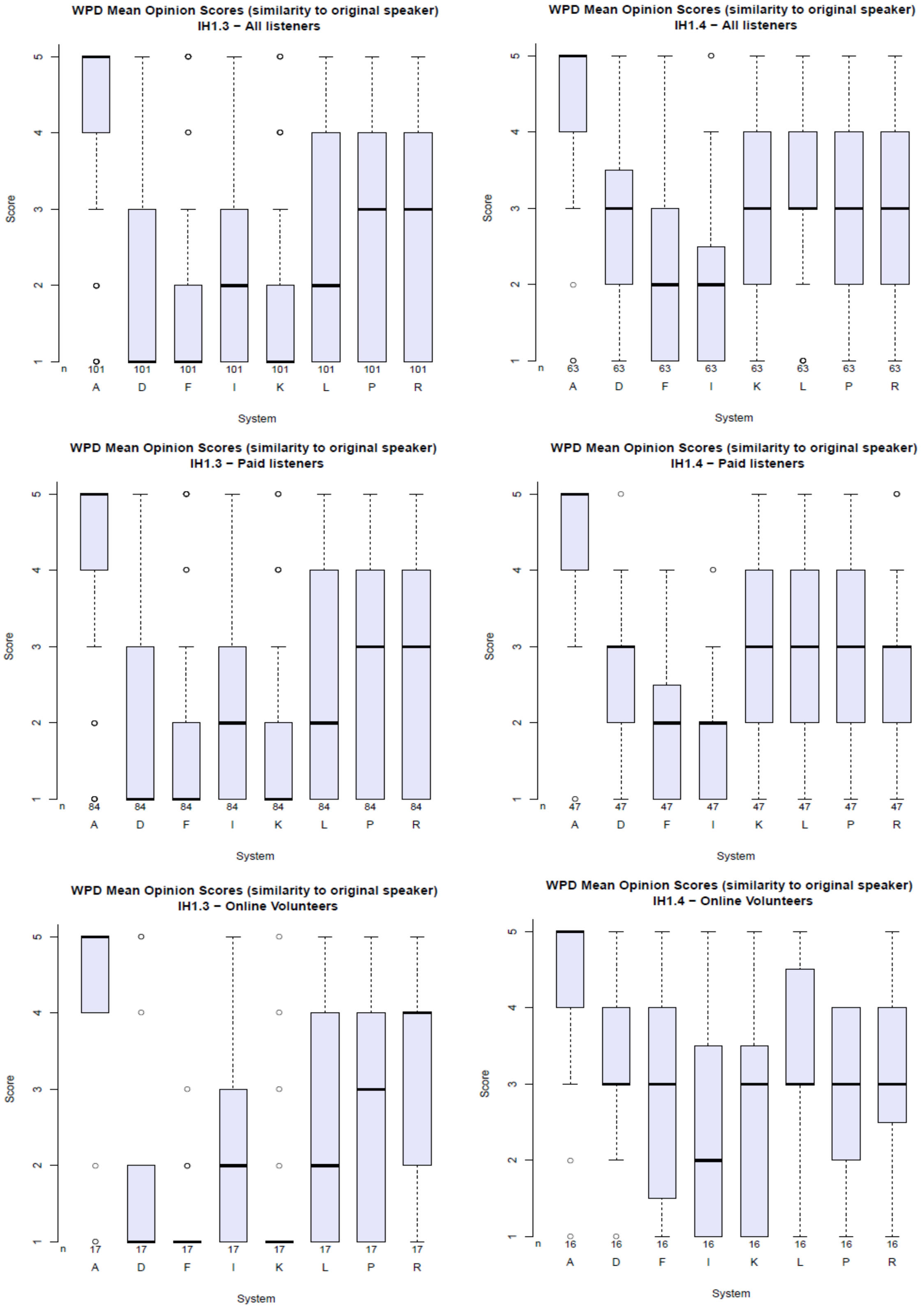

Fig.2. Boxplot of MOS on Similarity to original speaker in Kannada (IH1.3) and Tamil (IH1.4) for wpd sentences.

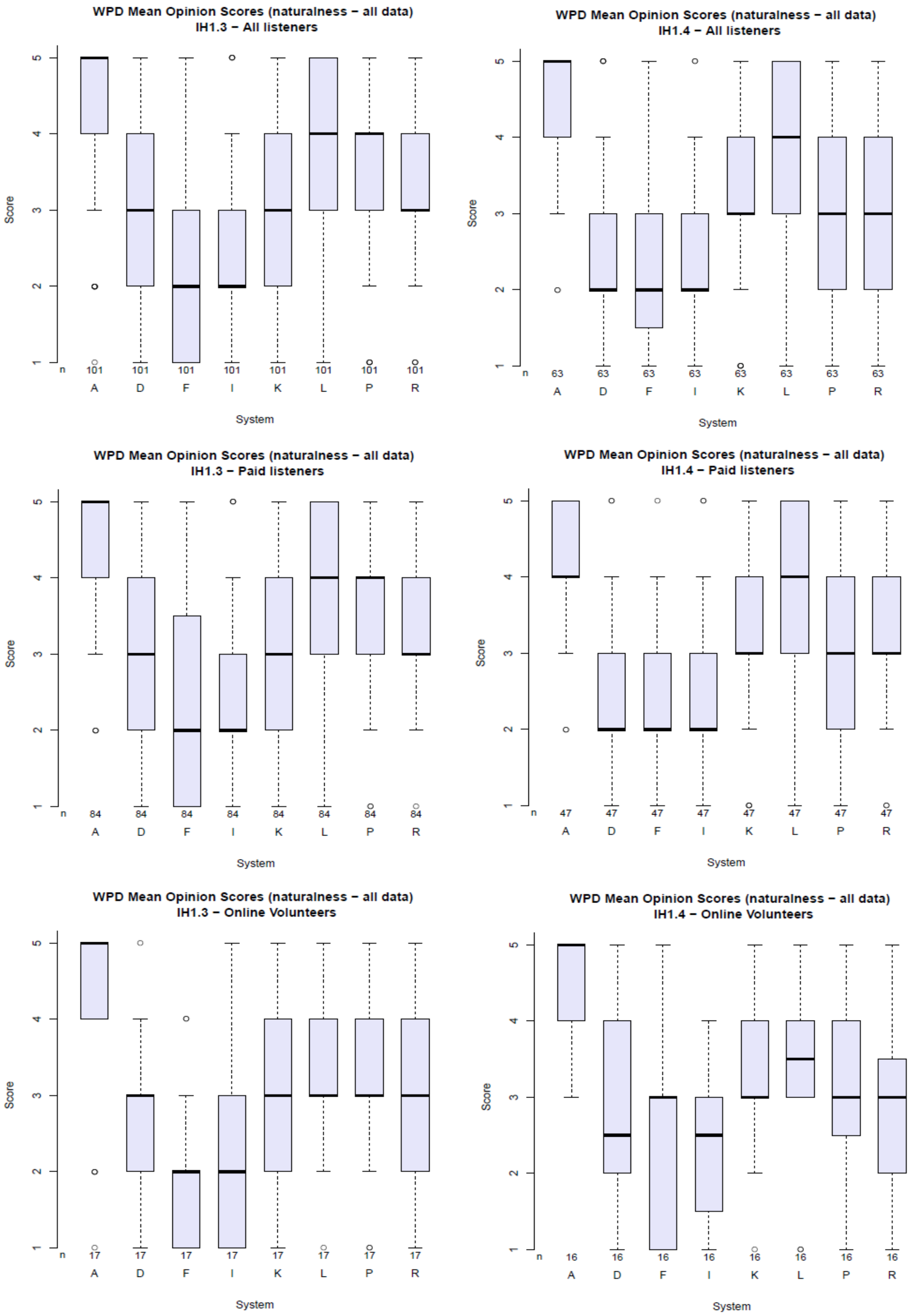

Fig.3. Boxplot of MOS on naturalness in Kannada (IH1.3) and Tamil (IH1.4) for wpd sentences.