MILE TTS for Tamil for blizzard challenge 2014

[1]*Rajaram B S R,* [1]*Shiva Kumar H R,* [1]*A G Ramakrishnan*

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

Abstract

Our participation in the Blizzard Challenge 2014 is only for the Tamil language. We have a 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 given RD (reading), SUS (semantically unpredictable sentences) and ML (multi-lingual) test sentences are synthesized using the corpus made available for the participants. The listening test results reported by the blizzard evaluation team are discussed. The letter code for MILE TTS is "J".

Index Terms: speech synthesis, unit selection, join cost, blizzard challenge.

1. Introduction

In this year's annual blizzard challenge 2014, MILE lab participated in the Indian languages tasks 2014-IH1.5-Tamil and 2014-IH2.5-Tamil using the given one hour of speech data and corresponding text in UTF-8 format. Unlike last year, the given test sentences for synthesizing were of 3 kinds; reading sentences (RD), semantically unpredictable sentences (SUS) and sentences (ML) containing interspersed English words. To build the voice, firstly the automatic segmentation of the given data is performed through forced alignment using HTK. Following this, a text normalization module is used to handle the multi-lingual sentences, where required. Subsequently, test sentences are converted to their phoneme equivalents 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. Viterbi search is performed at the sentence level and the selected units are concatenated.

The main intended application of Text-to-Speech synthesis (TTS) system [1,2] 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 user's mobile would perform optical character recognition and read aloud the recognized text using a Text-to-Speech synthesizer.

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

2. Description of MILE TTS Engine

MILE TTS engine is built on concatenative speech synthesis [3] framework. The optimal units are selected by Viterbi search considering the lowest total cost for a sentence. The block diagram of MILE TTS is shown in Fig.1.

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.

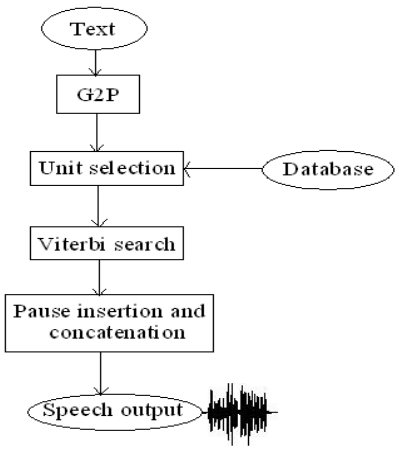


Fig.1. Block diagram of MILE TTS

2.2. Text Normalization

Two exception lists are used to handle the multi-lingual data in the form of interspersed English words. The first list has the English words transliterated to Tamil, which will be properly handled by the Tamil G2P [4] and the second list has the phoneme outputs directly replacing the English words.

2.3. 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. The total cost is the sum of concatenation and target costs.

The concatenation cost, Cc(ui-1, ui) is determined by the weighted sum of q concatenation sub-costs, $C_j^c(u_{i-1},u_i)(j=1,....,q)$. The sub-costs of concatenation cost arise broadly from spectral and pitch based features. Here, only the pitch based feature is used to compute concatenation cost. The continuity metric method proposed in [5] is used to derive pitch based feature.

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

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

where pi(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 available sub-word unit and it has been experimentally found in [5] that K=4 is sufficient for this application.

A rudimentary target cost is employed based on the duration of the vowels in the current version of the MILE TTS. The mean duration for each of the vowels is computed using all the occurrences of the respective vowels in the database. Thus, the units with minimum distance from this mean value have a higher probability in getting selected when the total cost is obtained.

2.4. Pause insertion and speech output

After selecting the units using Viterbi search, a fixed pause is inserted between the end of the previous word and the beginning of the current word. POS tagging is currently being developed and hence not included in this Blizzard challenge test sentence synthesis. After selecting reliable units for all the test sentences, the wave data is loaded to create the output wave file.

3. Steps followed to build the voice

The Tamil speech data, which is constituted by 822 wave files and corresponding text were released for Blizzard challenge Indian language task. Unlike the last time, there were no label files made available by Blizzard team. HTK was used to auto segment the wav files at the phone level. The following modifications were made to the label files:

1. Segmentation was cross verified by quick scanning through the files and re-segmented

manually, wherever it was felt absolutely required.

2. Obvious mistakes in pronunciation by the speaker were handled by corresponding corrections of the phoneme transcription to match the sound recording. For example, in `text_01062` **அந்தப் புன்முறுவலின் காந்தி நமது இளம் வீரனைத் திக்குமுக்காடித் திணறச் செய்தது**," the word **காந்தி** is pronounced as /Gaandhi/, rather than the correct pronunciation of /kaanthi/.

3.1. Voice building

With all the modifications discussed above, the synthesis database is created with wave files, label files and pitch files [6]. Then, the given sets of test sentences were synthesized.

4. Results and Discussion

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

4.1. Observations on the Tamil database

The following are some of the observations on the synthesis database [8] supplied for the challenge:

1. The quality of the voice and the pronunciation of the speaker were quite good.

2. However, the pronunciation of some of the phones were not consistent.

3. Interestingly, the so-called SUS sentences were the best Tamil sentences, because they were fairly predictable and nice. For example, SUS 33 is: **காக்கையின் குரல் கேட்பதற்கு மிகவும் இனிமையாக உள்ளது**. This means “The voice of the crow is very sweet.” Except for the fact that it is not true, it is a grammatically correct and logically predictable and complete sentence.

3. On the other hand, each so-called sentence in RD had multiple sentences (up to 4 in many cases). For example, the sentence 67 in RD is: **என்னது லெட்டரா இங்கிலீஷ்ல இருக்கு உங்க அப்பாவுக்கு இங்கிலீஷ் கூடத்தெரியுமா ஒன்ன விட எங்கப்பாவோட வொக்காப்லரி க்ரேட்தான் உங்க ஊர்லலாம் டிவி உண்டா உண்டே தூர்தர்ஷன் மட்டும் வரும்**

This actually contains 8 sentences and is a conversation between two people. So, with the proper punctuation marks, it must have been like this: **என்னது? லெட்டரா? இங்கிலீஷ்ல இருக்கு! உங்க அப்பாவுக்கு இங்கிலீஷ் கூடத் தெரியுமா? ஒன்ன விட எங்கப்பாவோட வொக்காப்லரி க்ரேட்தான். உங்க ஊர்லலாம் டிவி உண்டா? உண்டே. தூர்தர்ஷன் மட்டும் வரும்.**

Since the sentence-ending full stops (periods) had been deliberately removed from them, it was odd to synthesize. Since we were asked not to edit these target sentences, we refrained from editing them. However, one wonders as to why multiple sentences were combined by removing the delineating periods. It is not an easy NLP task to automatically segment them into individual sentences and manual editing was prohibited by instructions.

4.2. Naturalness test

Figures 1, 2 and 5 are the box-plots of mean opinion score (MOS) on the *naturalness* metric obtained by the Blizzard evaluation committee for Tamil (2014-IH1.5 and 2014-IH2.5) language. Also the figures show the MOS plots separately for all listeners and paid listeners under the categories of RD, SUS and ML sentences. For the RD & SUS test sentence category, our system performance is comparable with the second best and is consistent for all types of listeners. Since no signal processing or prosody modification is

included in the current version of MILE TTS, the naturalness is lower than the others.

4.3. Similarity test

Figures 3, 4 and 6 are the box-plots of MOS on the similarity to the original speaker metric obtained by the Blizzard evaluation committee for Tamil (2014-IH1.5 and 2014-IH2.5) language. Once again, the figures show the MOS plots separately for all listeners and paid listeners for each of the categories of RD, SUS and ML sentences. For the RD & SUS test sentence category, the performance of our system is the second best and is consistent for all types of listeners. However, the system performance is among the best for the ML sentences. This can be attributed to efficient handling of the interspersed English words.

The overall system performance can be attributed to the unit selection based concatenative speech synthesis approach. At the same time, we have maintained a minimum MOS value of around 3 for all types of listeners, under all types of test sentences for naturalness and similarity metrics. This performance is the same as last year [7].

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.

References

1. G L Jayavardhana Rama, A G Ramakrishnan, R Muralishankar and R Prathibha, “A complete text-to-speech synthesis system in Tamil,” Proc. IEEE Speech Synthesis Workshop 2002, pp. 191 – 194.

2. Konakanchi Partha Sarathy and A G Ramakrishnan, “A research bed for unit selection based text to speech synthesis,” Proc. IEEE Spoken Language Technology Workshop, 2008 (SLT 2008), Goa, pp. 229 – 232.

3. 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.

4. A G Ramakrishnan, Lakshmish N Kaushik and M Laxmi Narayana, “Natural Language Processing for Tamil TTS,” Proc. 3rd Language and Technology Conf., Poznan, Poland, 2007, pp. 192 - 196.

5. Vikram Ramesh Lakkavalli, Arulmozhi P and A G Ramakrishnan, “Continuity Metric for Unit Selection based Text-to-Speech Synthesis,” IEEE Intern. Conf. on Signal Proc. and Communications (SPCOM 2010), 2010, Bangalore, India.

6. A P Prathosh, T V Ananthapadmanabha and A G Ramakrishnan, “Epoch extraction based on integrated linear prediction residual using plosion index,” IEEE Trans. Audio, Speech and Language Processing, 2013, Vol. 21 (12), pp. 2471-2480.

7. H R Shiva Kumar, J K Ashwini, B S R Rajaram and A G Ramakrishnan, “MILE TTS for Tamil and Kannada for blizzard challenge 2013,” Proc. Blizzard Challenge 2013 workshop, Barcelona, Catalonia, 2013.

8. Patil H A, Patel T B, Shah N J, Sailor H B, Krishnan R, Kasthuri GR, Nagarajan T, Christina L, Kumar N, Raghavendra V, Kishore S P, Prasanna S R M, Adiga N, Singh S R, Anand K, Kumar P, Singh B C, Binil Kumar S L, Bhadran T G, Sajini T, Saha A, Basu T, Rao K S, Narendra N P, Sao A K, Kumar R, Talukdar P, Acharyaa P, Chandra S, Lata S and Murthy H A, “A syllable-based framework for unit selection synthesis in 13 Indian languages,” Proc. Oriental COCOSDA held jointly with 2013 Conf. on Asian Spoken Language Res. and Evaluation (O-COCOSDA/CASLRE 2013), pp. 1-8, doi:10.1109/ICSDA.2013.6709851

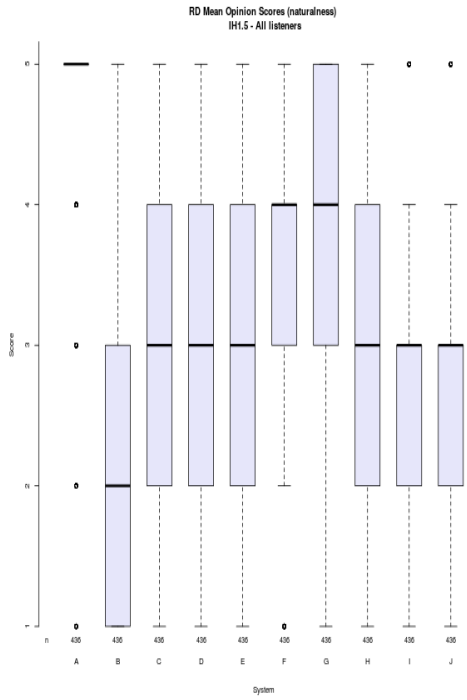


Fig 1: Box plot (All listeners) of RD and SUS sentences for the *naturalness* metric.

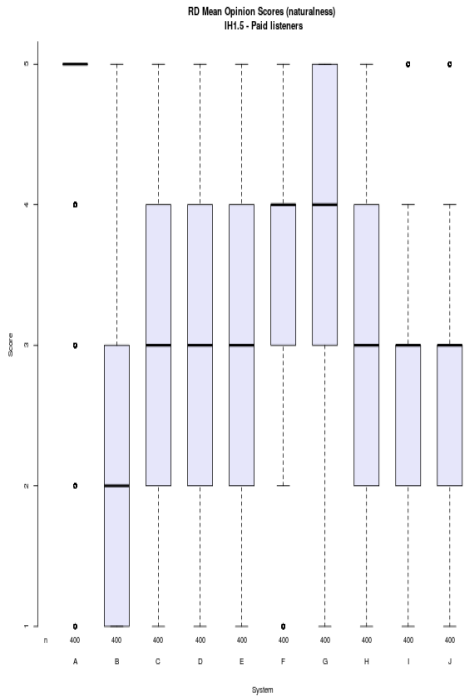
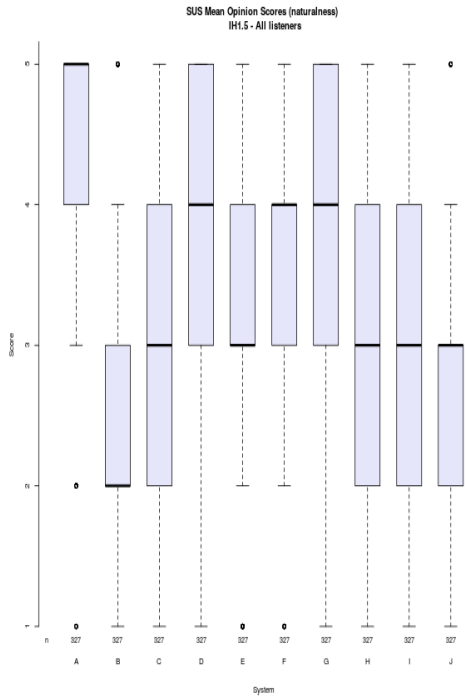
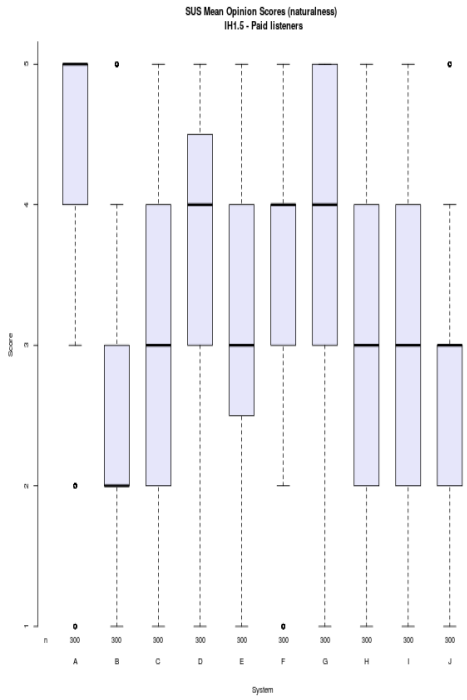


Fig 2: Box plot (paid listeners) of RD and SUS sentences for the naturalness metric.





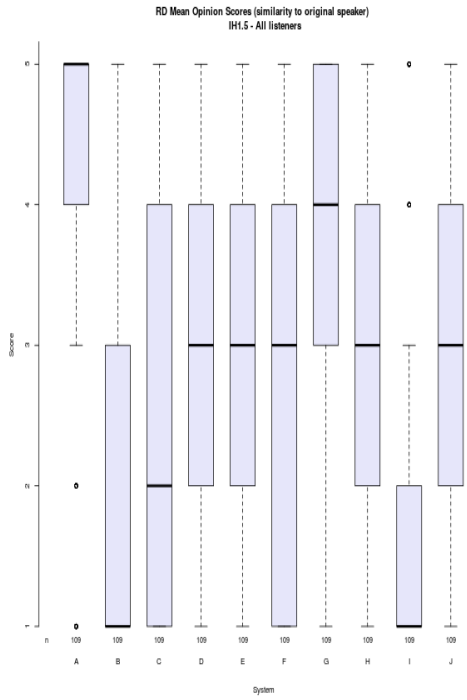
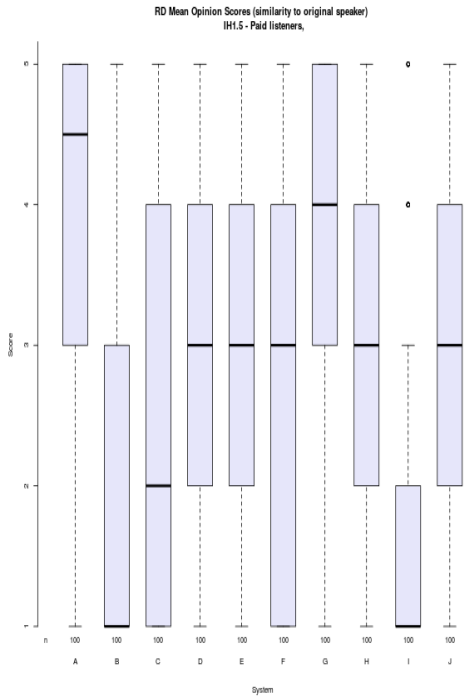


Fig 3: Box plot of RD and SUS sentences for the *similarity to original speaker* metric



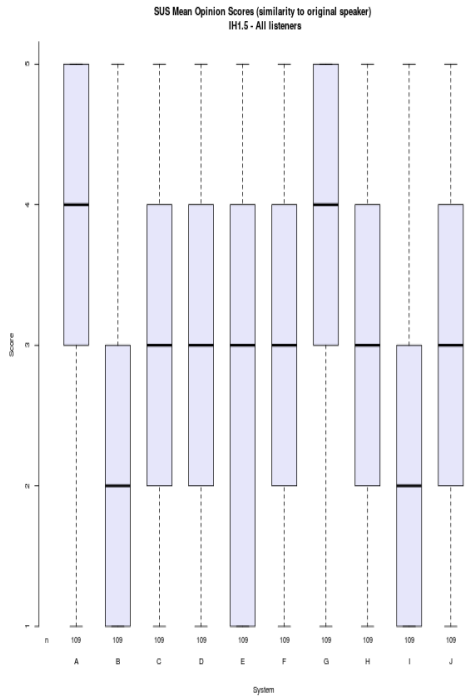
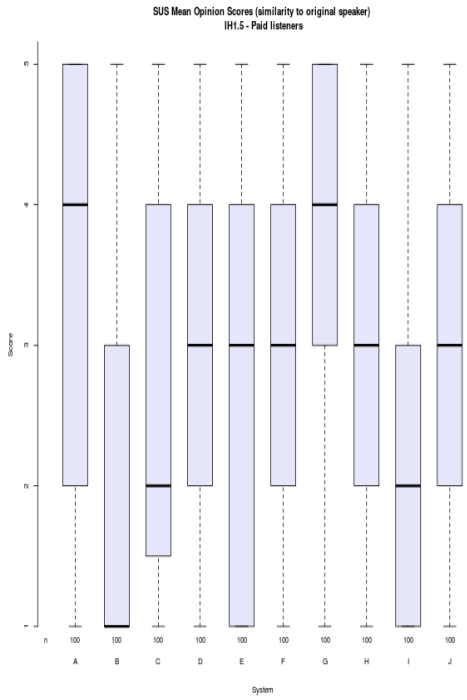
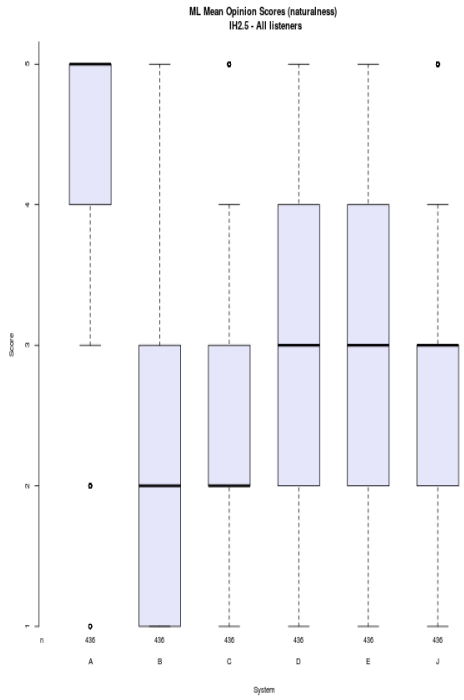
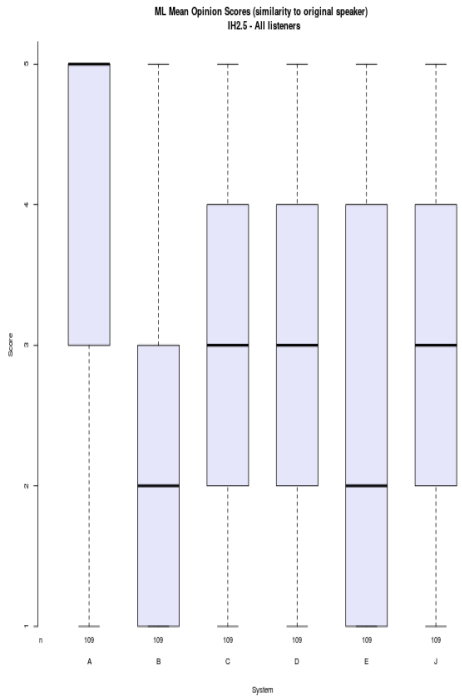
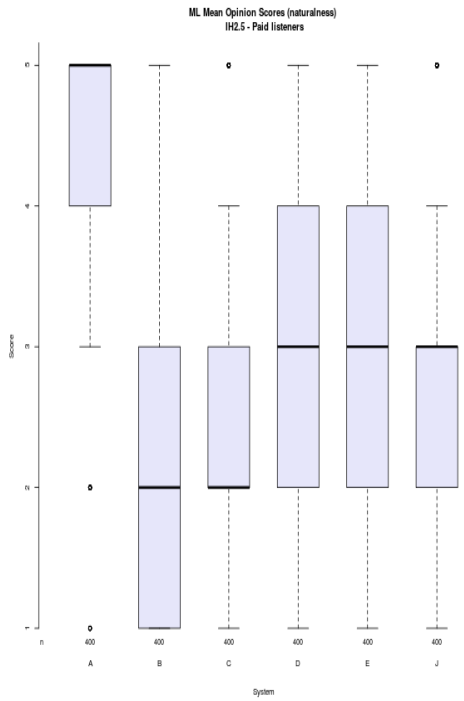


Fig 4: Box plot of RD and SUS sentences for the *similarity to original speaker* metric









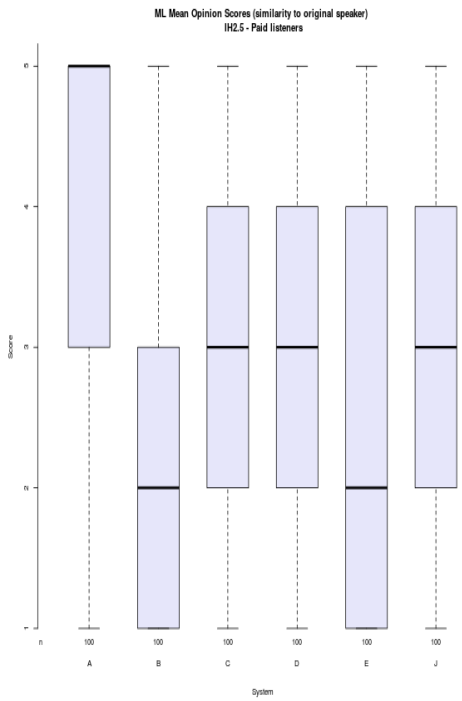


Fig 5: Box plot (both paid and all listeners) of ML sentences for the *naturalness* metric

Fig 6: Box plot (both paid and all listeners) of ML sentences for the *similarity to original speaker* metric