The Blizzard Challenge 2014

[1]*Kishore Prahallad,* [1]*Anandaswarup Vadapalli,* [1]*Santosh Kesiraju,* [2]*Hema A. Murthy*
[3]*Swaran Lata,* [4]*T. Nagarajan,* [5]*Mahadeva Prasanna,* [6]*Hemant Patil,* [7]*Anil Kumar Sao*
[8]*Simon King,* [9]*Alan W. Black and* [10]*Keiichi Tokuda*

[1] Speech and Vision Lab, IIIT Hyderabad, India
[2] Department of CSE, IIT Madras, India
[3] Department of Electronics and Information Technology, Govt. of India
[4] Department of IT, SSN College of Engineering, India
[5] Department of EEE, IIT Guwahati, India
[6] DAIICT, India
[7] School of Computing and Electrical Engineering, IIT Mandi, India
[8] Center for Speech Technology Research, University of Edinburgh, UK
[9] Language Technologies Institute, Carnegie Mellon University, USA
[10] Department of Computer Science, Nagoya Institute of Technology, Japan

Abstract

The Blizzard challenge 2014 was the tenth annual Blizzard challenge organized by the following group of institutions : IIIT Hyderabad, IIT Madras, DAIICT, SSN College of Engineering, IIT Mandi and IIT Guwahati with support and collaboration from DeitY, Government of India. This paper describes the tasks in the Blizzard challenge 2014. The tasks consisted of data from six Indian languages : Assamese, Gujarati, Hindi, Rajasthani, Tamil and Telugu. Seven participants from around the world used the speech data provided as well as the corresponding text transcriptions in UTF-8, to build synthetic voices, which were then evaluated by means of listening tests.
Index Terms: Blizzard challenge, Speech synthesis, Evaluation of synthetic speech

1. Introduction

The Blizzard challenge, originally started by Black and Tokuda [1], is a well established challenge in the field of speech synthesis. [1–11] are summary papers which provide information about the previous challenges. These resources can be found on the Blizzard Challenge website[1]. This paper is a summary paper describing the tasks in the Blizzard 2014 challenge.

2. Blizzard 2014 tasks

2.1. Database used

Speech and text data for six Indian languages i) Assamese, ii) Gujarati, iii) Hindi, iv) Rajasthani, v) Tamil and vi) Telugu were released. The speech data for each language was 2 hours (sampled at 16 KHz), recorded by professional speakers in a high quality studio environment. Along with the speech data the corresponding text was provoded in UTF-8 format. No other information, like segment labels was provided as part of the challenge. However, there was no restriction on the particpants to learn / use information like phonesets or labels from other resources.

[1]http://www.festvox.org/blizzard/

For the nature of scripts and sounds of Indian language please refer to [11].

2.2. Tasks

Blizzard challenge 2014 consisted of two tasks, a hub task and and a spoke task.

- Hub task 2014-IH1 : Participants were asked to build one voice in each language from the provided data, in accordance of the rules of the challenge. The subtasks were numbered from IH1.1 to IH1.6 corresponding to the six languages : IH1.1 (Assamese [AS]), IH1.2 (Gujarati [GU]), IH1.3 (Hindi [HI]), IH1.4 (Rajasthani [RJ]), IH1.5 (Tamil [TA]) and IH1.6 (Telugu [TE]).
- Spoke task 2014-IH2 : Participants had to synthesize multilingual sentences containing Indian language text as well as English. The subtasks were numbered from IH2.1 to IH2.6 corresponding to the six languages : IH2.1 (Assamese [AS]), IH2.2 (Gujarati [GU]), IH2.3 (Hindi [HI]), IH2.4 (Rajasthani [RJ]), IH2.5 (Tamil [TA]) and IH2.6 (Telugu [TE]).

For the IH1 task (hub task), the synthetic voices were evaluated through listening tests on the following test data (for each Indian language)

- Read speech (RD) - 100 distinct sentences, not a part of the training data
- Semantically unpredictable sentences (SUS) - 50 distinct sentences not a part of the RD/training data

The SUS sentences were prepared in the following manner. 50 sentences in each language were randomly selected, and POS tagging was performed on these sentences. The words in each sentence were then reordered as *Subject Object Verb Conjuction Subject Object Verb* to generate the SUS sentence.

For the IH2 task (spoke task), the systems were evaluated through listening tests by synthesizing the following test data (for each Indian language + English combination)

- Multilingual sentences (ML) - 50 distinct sentences containing both Indian language as well as English words.

No language tags were provided in the ML sentences. The participants were expected to identify the language from the Unicode code point.

2.3. Participants in the challenge

The participants in the Blizzard challenge 2014 consisted of the seven participants listed in Table 1. To annonimyze the results, the systems are identified using letters, with A denoting natural speech, B denoting the baseline system and C to K denoting the systems submitted by the participants in the challenge. Each participant could submit as many systems as they wished.

Table 1: Participants in Blizzard challenge 2014

Short name	Details	Synthesis method
NATURAL	Natural speech	
BASE	Baseline system	HMM
NITECH	Nagoya Institute of Technology	HMM
USTCP	National Engineering Laboratory of Speech & Language Information Processing (Primary system)	Hybrid (IH1.3) / HMM (remaining)
CMU	Carnegie Mellon University	HMM
S4A	Simple4All project consortium	HMM + DNN
ILSP	Institute for Language and Speech Processing / Innoetics	USS
IITMS	IIT Madras (Secondary system)	HMM (IH1.3,IH1.4 and IH1.6) / USS (remaining)
IITMP	IIT Madras (Primary system)	USS (IH1.3,IH1.4 and IH1.6) / HMM (remaining)
MILE-TTS	Dept. of Electrical Engg, Indian Institute of Science	USS
USTCS	National Engineering Laboratory of Speech & Language Information Processing (Secondary system)	HMM

2.4. Baseline systems

Baseline systems were built for each language using the speaker independent HTS-2.2 + STRAIGHT scripts[2]. The data was labeled at the phone level using the HMM labeling script (EHMM) in FestVox[3] [12]. For letter to sound rules a set of simple naive first order approximations were used for each language.

3. Evaluation

The participants were asked to synthesize the complete test set, out of which a subset was used in the listening tests. The listening tests for IH1.1 - IH1.6 consisted of ten sections while the listening tests for IH2.1 - IH2.6 consisted of five sections. The different sections of the listening tests are described below.

- Listening tests for IH1.1 - IH1.6
  1. two sections for similarity (one section using RD and one section using SUS)
  2. seven sections for naturalness (four sections using RD and three sections using SUS)
  3. one section for intelligibility using SUS
- Listening tests for IH2.1 - IH2.6
  1. one section for similarity
  2. four sections for naturalness

The methodology of scoring in the various sections of the listening tests are described below.

- **Similarity** : The listener plays a few samples of the original speaker and one synthetic sample. The listener then chooses a response that represented how similar the synthetic voice sounded as compared to the original speakers voice on a scale from

  1 : Sounds like a totally different person

  to

  5 : Sounds exactly like the same person

- **Naturalness** : The listener listenes to a sample of synthetic speech and chooses a score which represents how natural or unnatural the sentence sounded on a scale of

  1 : Completely Unnatural

  to

  5 : Completely Natural

- **Intelligibility** : Listeners listen to an utterance and type in what they hear. Word Error Rate (WER) is computed in the same manner it is computed for speech recognition tasks.

For the list of changes made in the evaluation portal to enable the conduct of listening tests in Indian languages, please refer to [11]

4. Results

The following listener types were used for the listening tests :

- Paid users
- Online volunteers

Apart from these types of listeners, we also experimented with conducting listening tests on Amazon mechanical turk (AMT).

Table 2 shows the statistics of the different listener types for the tasks.

Table 2: User statistics for the Blizzard 2014 tasks

Task	Paid Users	Online volunteers	AMT users
IH1.1 + IH1.1	106	09	-
IH1.2 + IH2.1	50	0	-
IH1.3 + IH2.3	100	09	54
IH1.4 + IH2.4	101	09	-
IH1.5 + IH2.5	100	09	55
IH1.6 + IH2.6	100	06	44

4.1. Results

For the six languages in the IH1 hub task (IH1.1 - IH1.6), Figures 1 to 6 and Figures 7 to 12 show the similarity and naturalness results on RD and SUS respectively. The intelligibility results for the hub task (IH1.1 - IH1.6) are shown in Figures 13 to 18.

For the spoke task (IH2.1 - IH2.6), Figures 19 to 24 show the similarity and naturalness results on ML.

[2]http://hts.sp.nitech.ac.jp/?Download

[3]http://www.festvox.org

For a detailed discussion of the results, please refer to the papers describing each system submitted by individual participants, available on the Blizzard Challege website.

5. Conclusions

The conclusions drawn from the results of the Blizzard challege 2014 are :

- The high quality audio recordings provided decent performances by all systems
- All teams performed better than the baseline system. This can be attributed to the fact that open source toolkits typically require sufficient tuning to make them work better for new/arbitrary languages.
- There does not seem to be much utility in computing WER as a measure of intelligibility for Indian languages.
- Some teams performed better on the ML task as compared to RD and SUS.
- Scores obtained from Amazon mechanical turk listeners show too much noise and variability in the score. These listeners can not be used as an alternative to paid listeners.

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005 : Evaluating corpus-based speech synthesis on common datasets," in *Proceedings of Intespeech 2005*, Lisbon, 2005.

[2] C. L. Bennett, "Large scale evaluation of corpus-based synthesizers : Results and lessons from the Blizzard Challenge 2005," in *Proceedings of Interspeech 2005*, 2005.

[3] C. L. Bennett and A. W. Black, "The Blizzard Challenge 2006," in *Blizzard Challenge Workshop, Interspeech 2006 - ICSLP satellite event*, 2006.

[4] M. Frazer and S. King, "The Blizzard Challenge 2007," in *Proceedings Blizzard Workshop 2007 (in Proc. SSW6)*, 2007.

[5] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proceedings Blizzard Workshop 2008*, 2008.

[6] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proceedings Blizzard Workshop 2009*, 2009.

[7] ——, "The Blizzard Challenge 2010," in *Proceedings Blizzard Workshop 2010*, 2010.

[8] ——, "The Blizzard Challenge 2011," in *Proceedings Blizzard Workshop 2011*, 2011.

[9] ——, "The Blizzard Challenge 2012," in *Proceedings Blizzard Workshop 2012*, 2012.

[10] ——, "The Blizzard Challenge 2013," in *Proceedings Blizzard Workshop 2013*, 2013.

[11] K. Prahallad, A. Vadapalli, N. Elluru, G. Mantena, B. Pulugundla, P. Bhaskararao, H. A. Murthy, S. King, V. Karaiskos, and A. W. Black, "The Blizzard Challenge 2013 – Indian Language Tasks," in *Proceedings Blizzard Workshop 2013*, 2013.

[12] A. W. Black and K. Lenzo, "Building voices in the festival speech synthesis system," 2002, available Online: http://festvox.org/bsv.

[13] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proceeding Blizzard Workshop 2007 (in Proceedings SSW6)*, 2007.

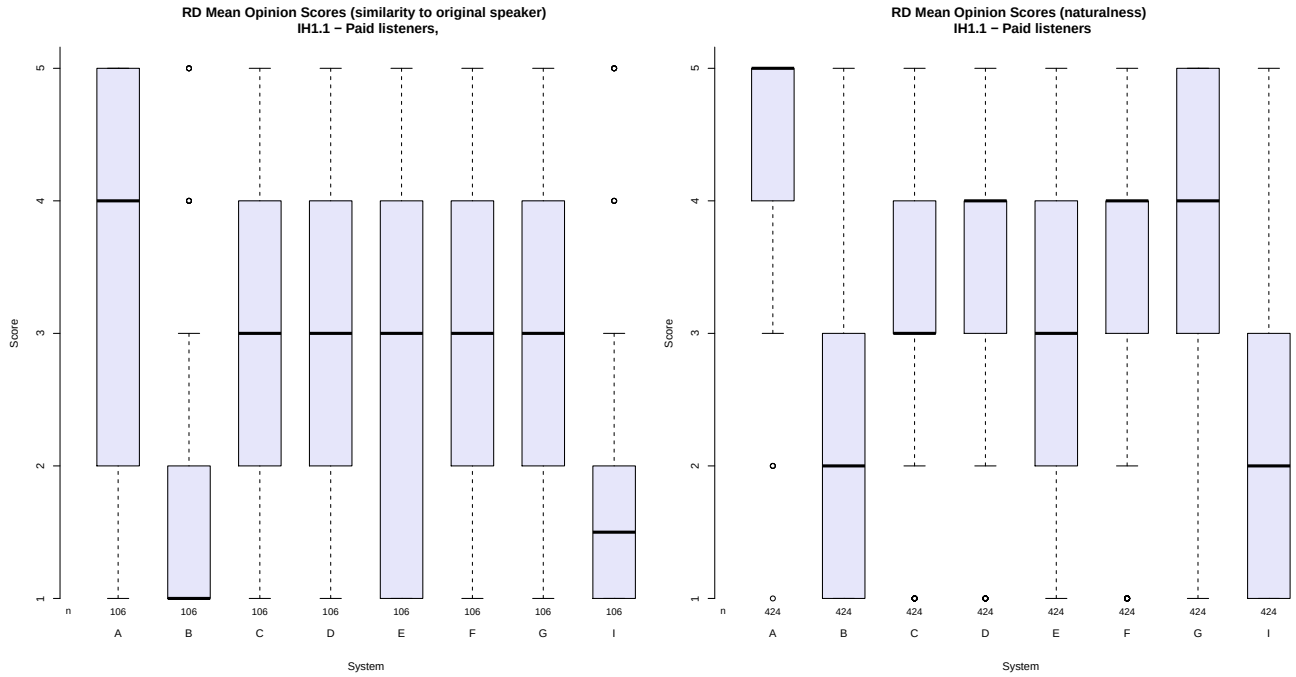

Figure 1: Similarity and Naturalness results on RD for IH1.1 (Assamese)

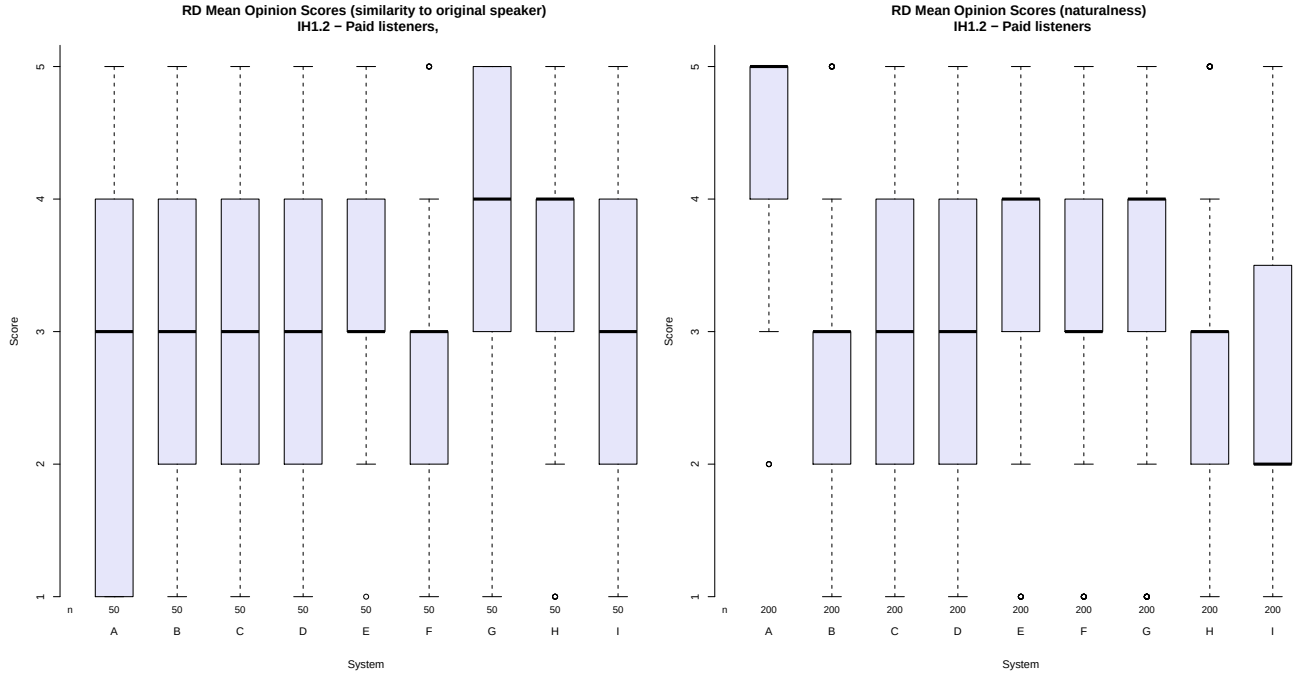

Figure 2: Similarity and Naturalness results on RD for IH1.2 (Gujarati)

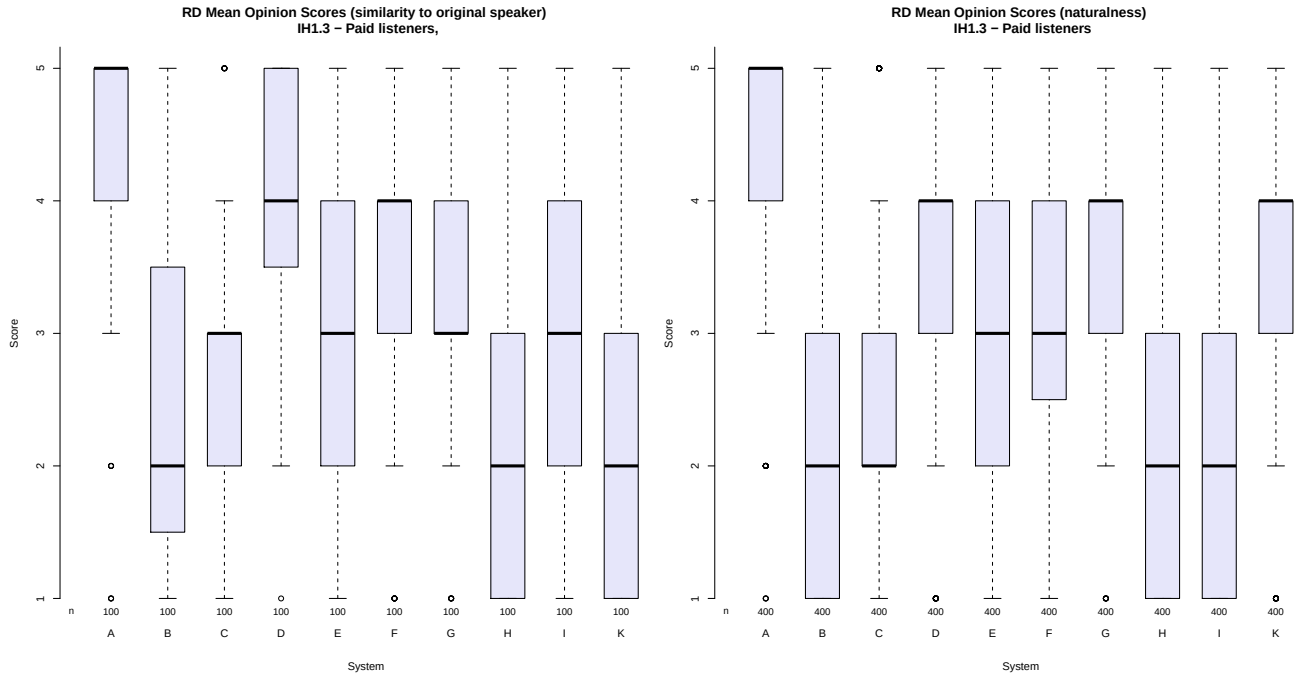

Figure 3: Similarity and Naturalness results on RD for IH1.3 (Hindi)

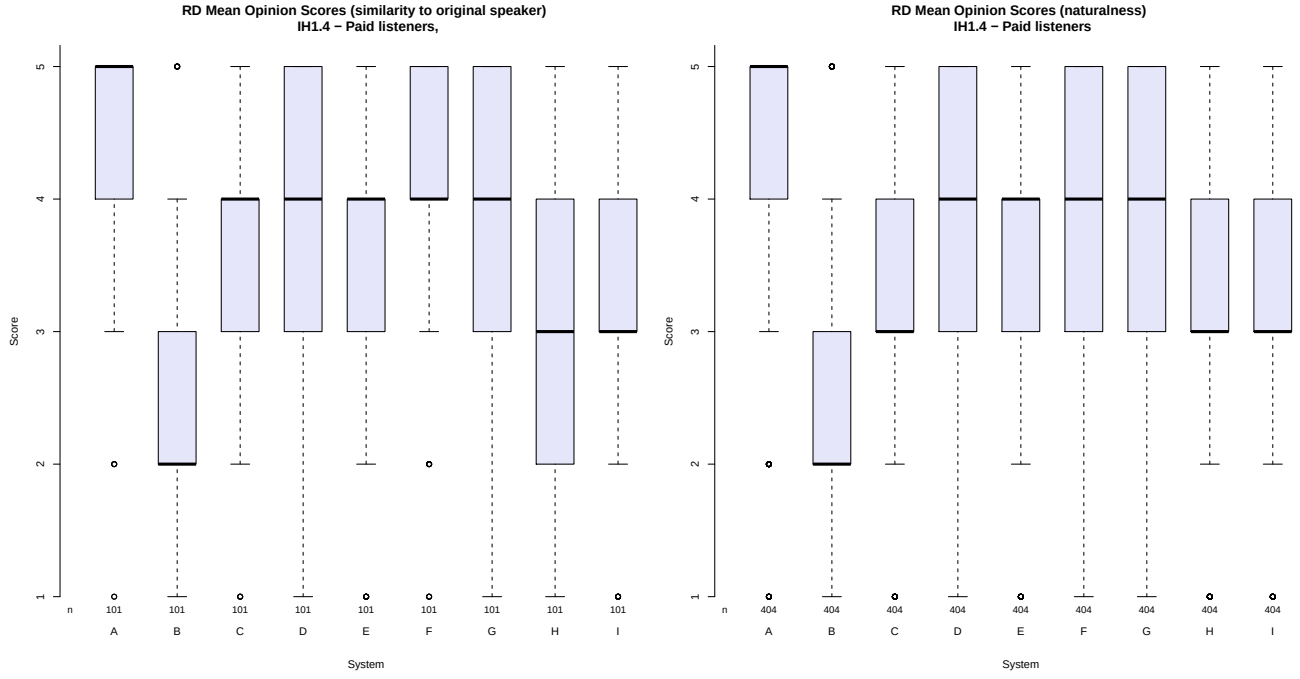

Figure 4: Similarity and Naturalness results on RD for IH1.4 (Rajasthani)

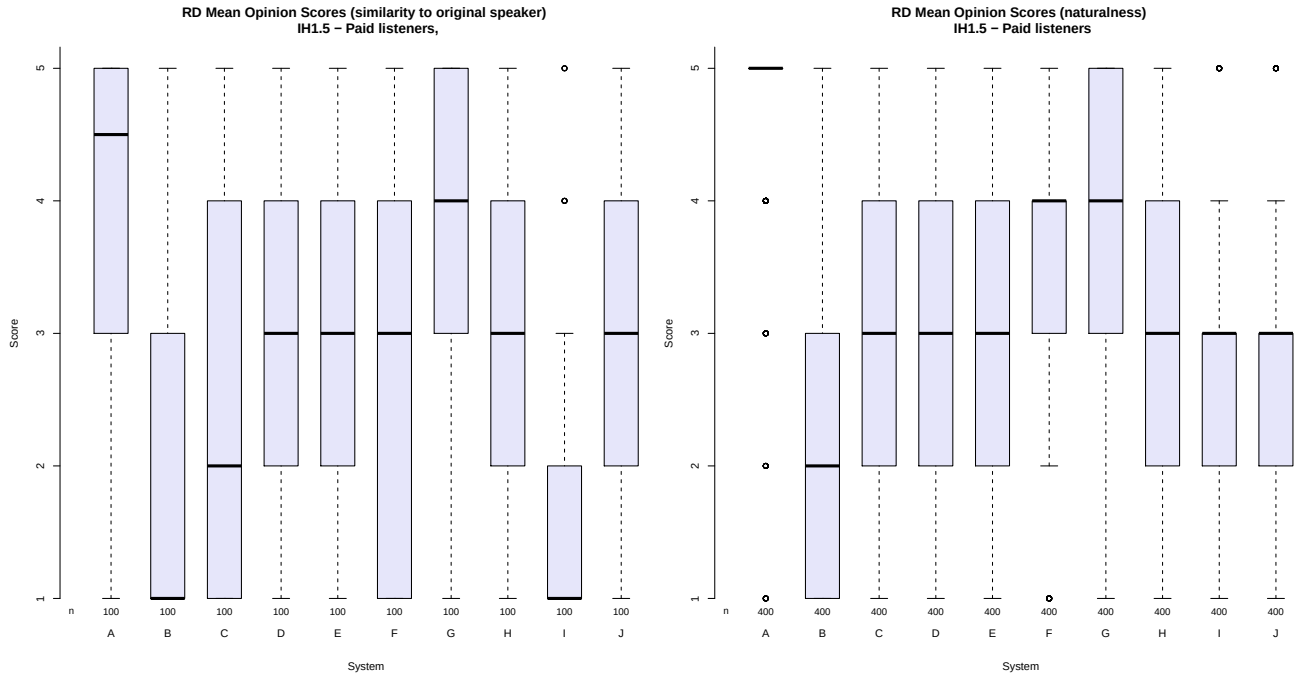

Figure 5: Similarity and Naturalness results on RD for IH1.5 (Tamil)

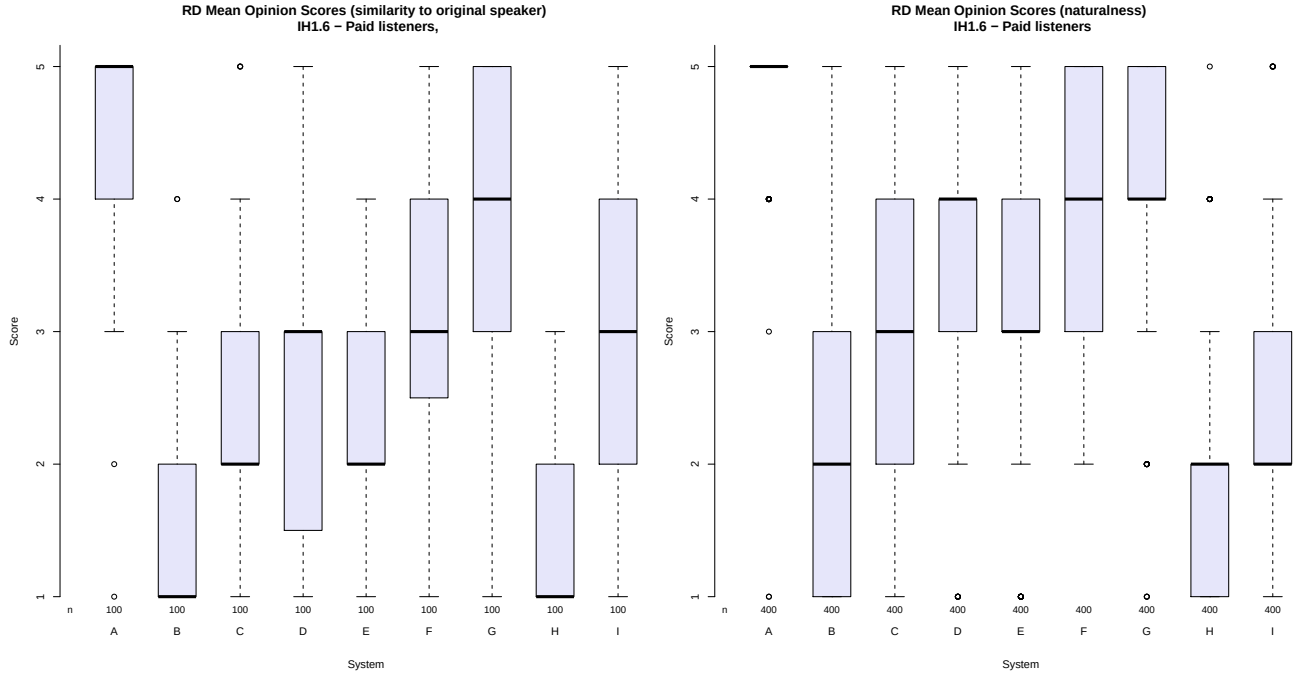

Figure 6: Similarity and Naturalness results on RD for IH1.6 (Telugu)

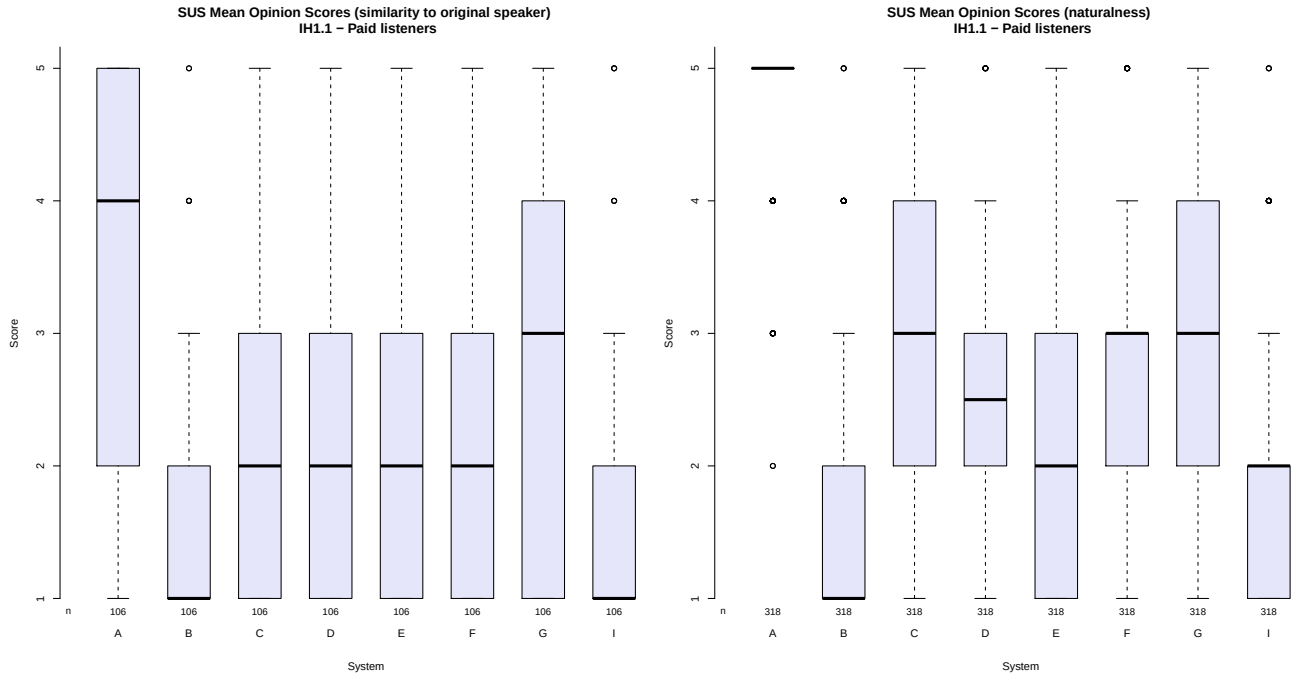

Figure 7: Similarity and Naturalness results on SUS for IH1.1 (Assamese)

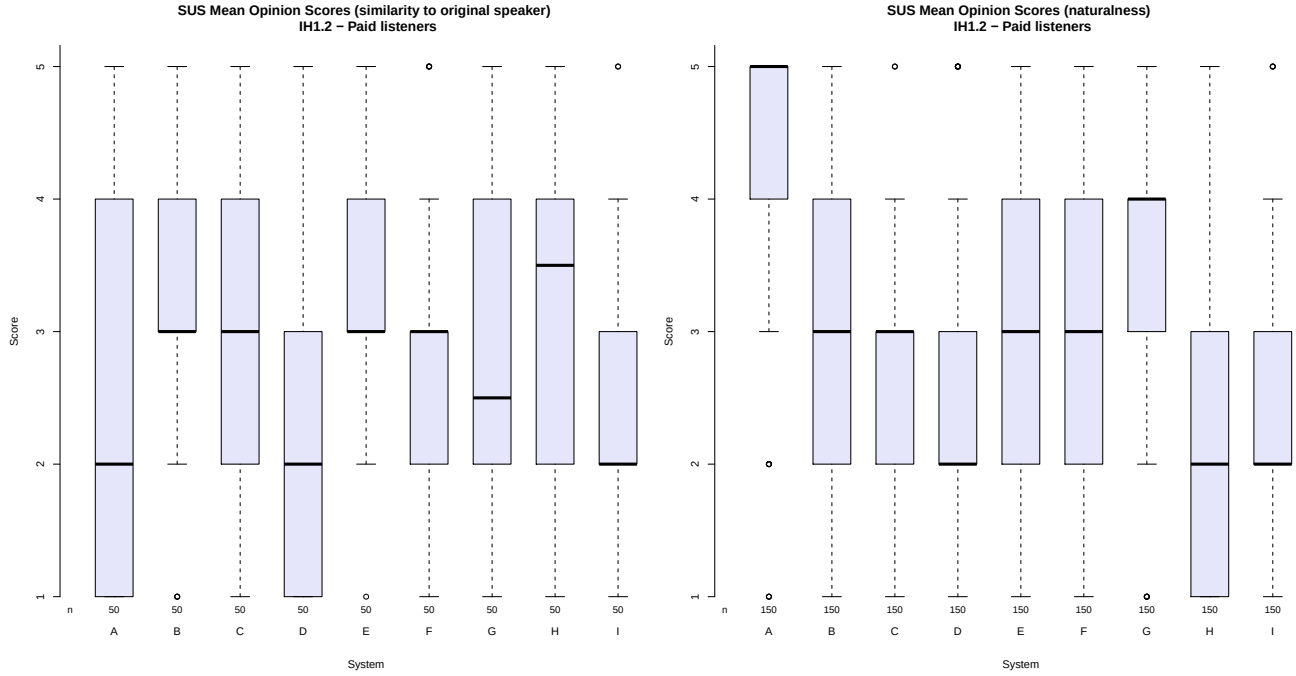

Figure 8: Similarity and Naturalness results on SUS for IH1.2 (Gujarati)

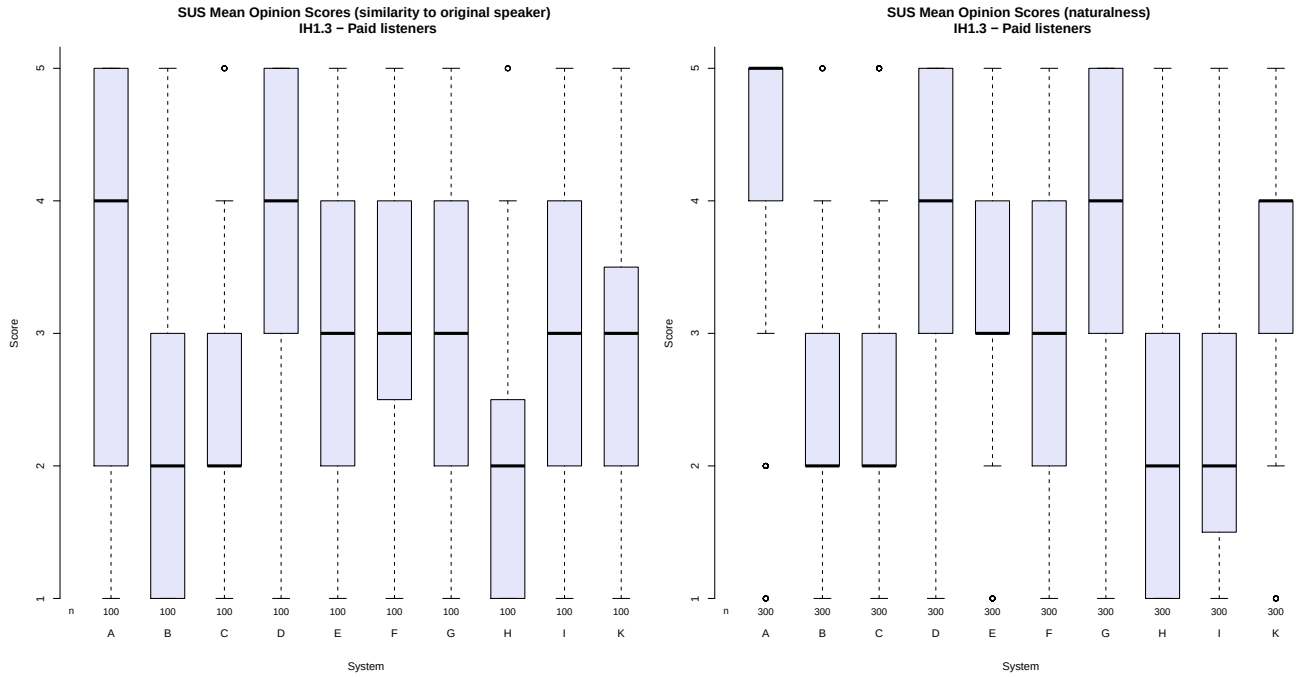

Figure 9: Similarity and Naturalness results on SUS for IH1.3 (Hindi)

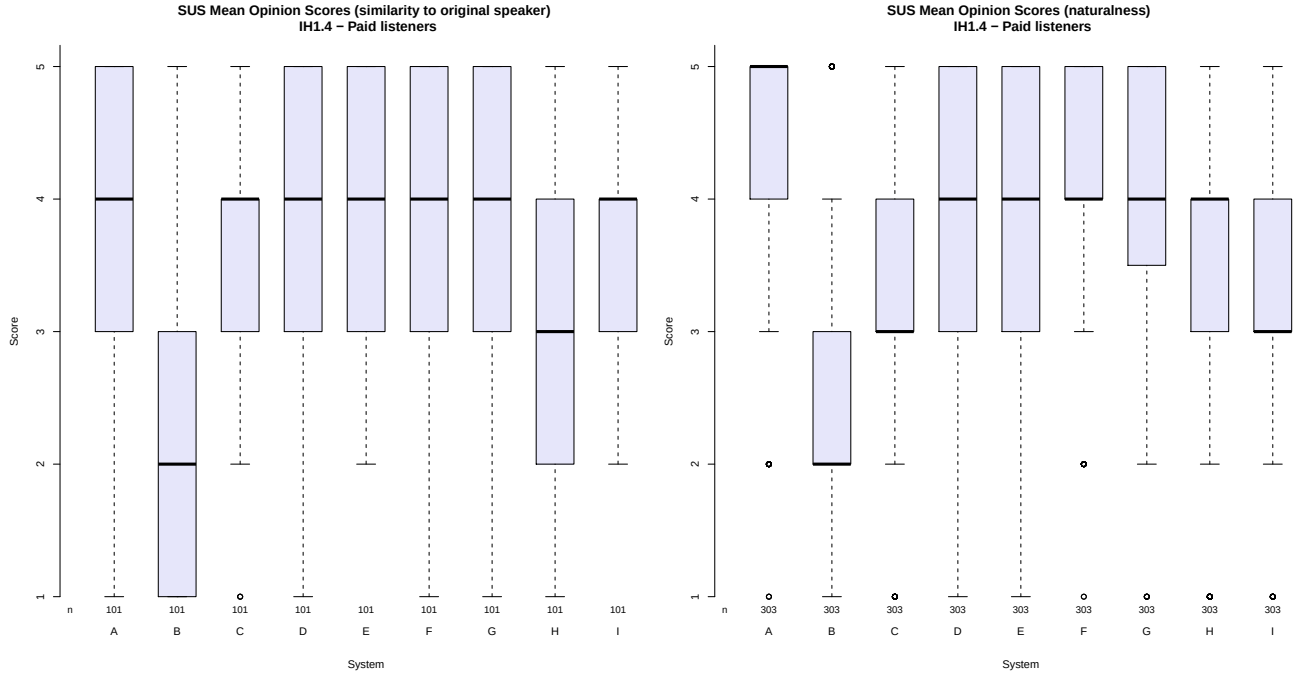

Figure 10: Similarity and Naturalness results on SUS for IH1.4 (Rajasthani)

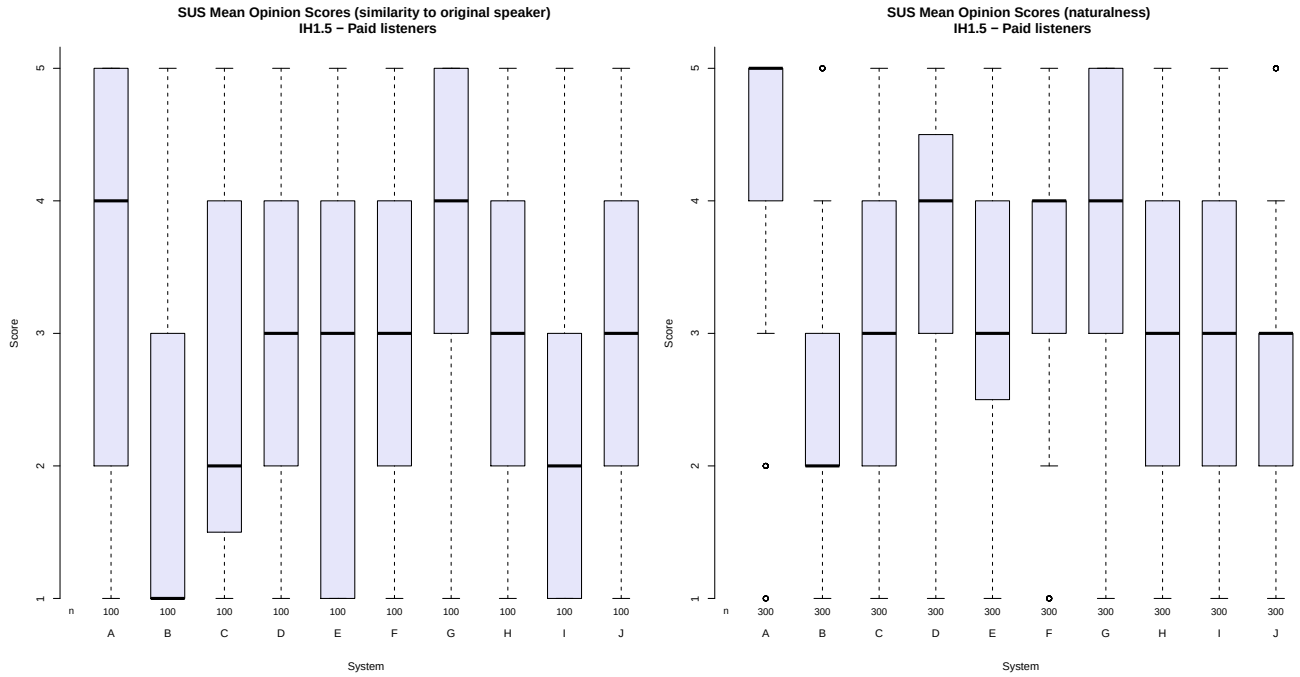

Figure 11: Similarity and Naturalness results on SUS for IH1.5 (Tamil)

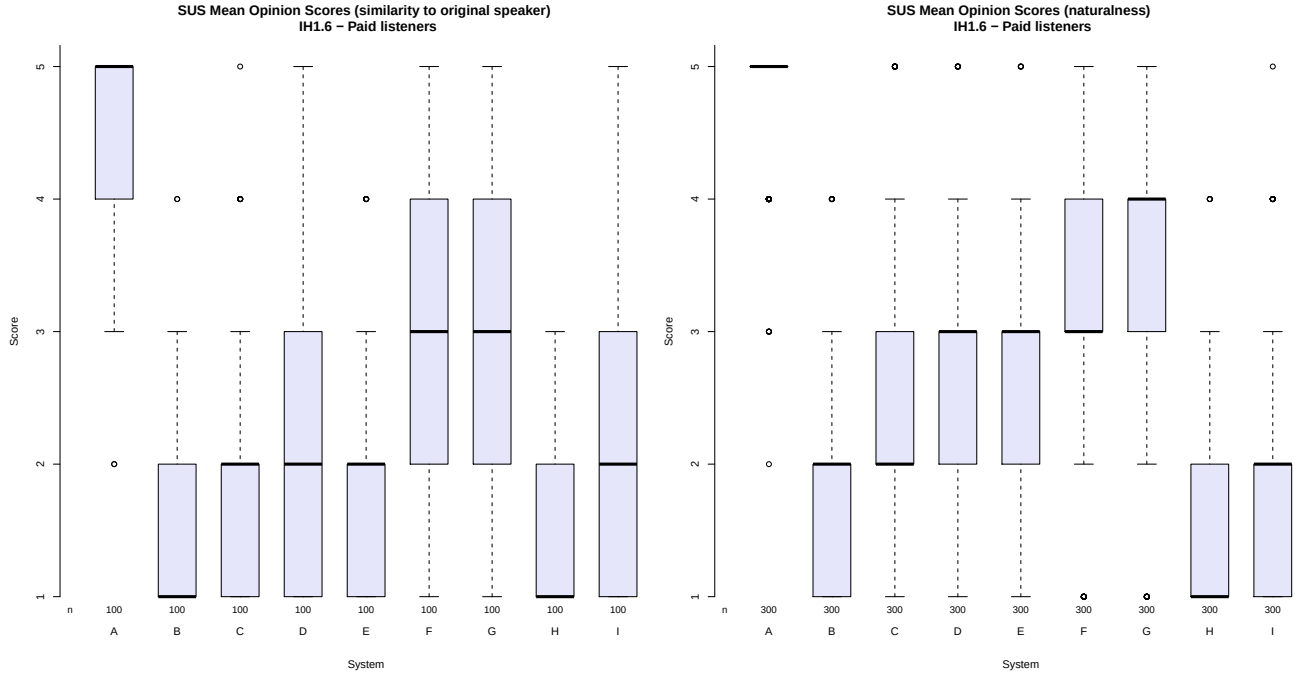

Figure 12: Similarity and Naturalness results on SUS for IH1.6 (Telugu)

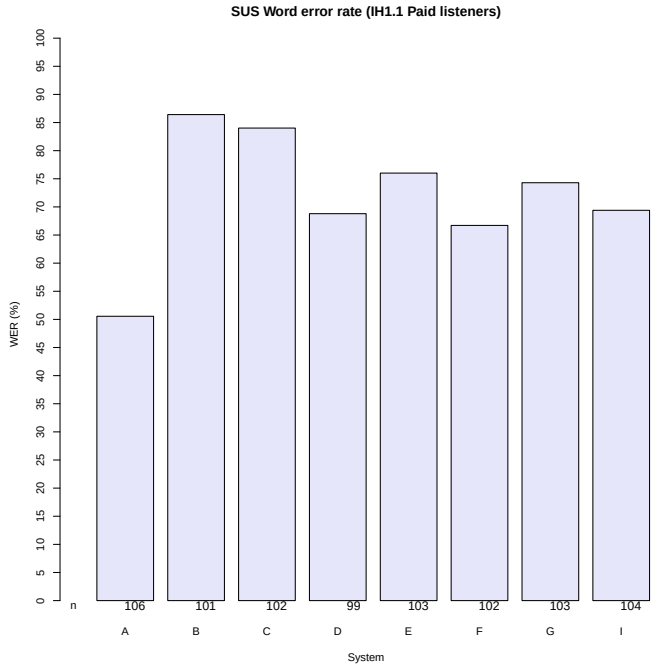

Figure 13: Intelligibility results on SUS for IH1.1 (Assamese)

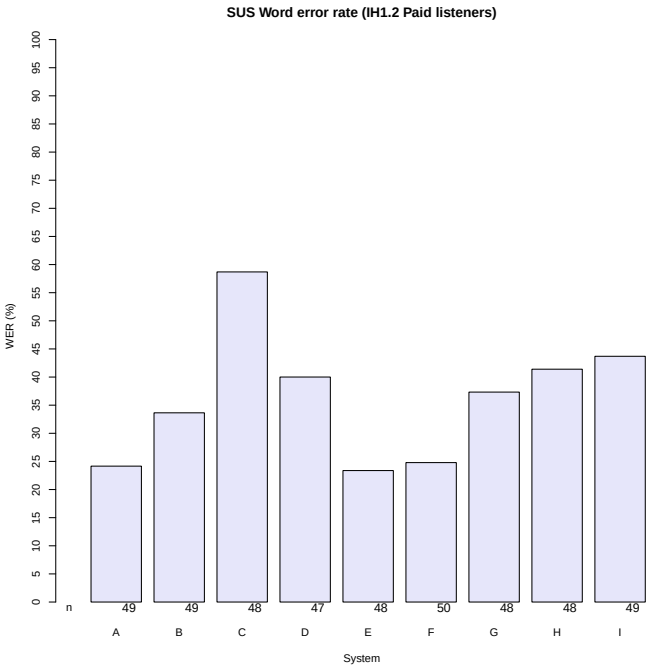

Figure 14: Intelligibility results on SUS for IH1.2 (Gujarati)

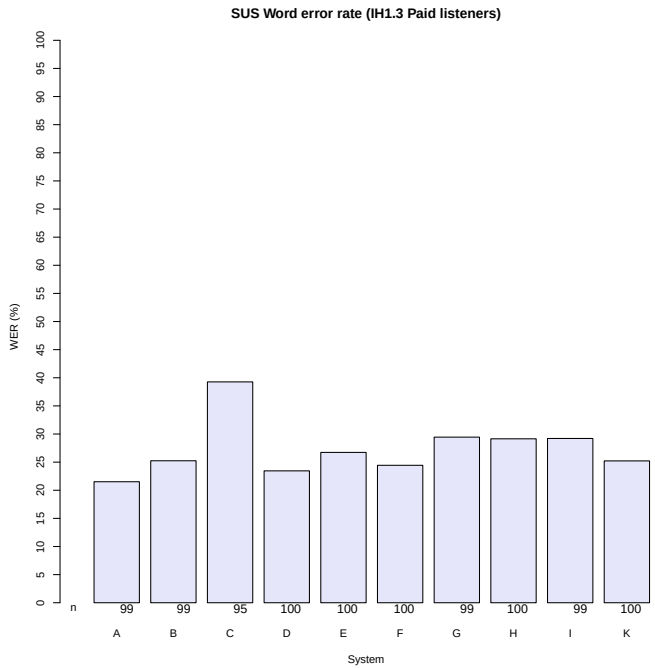

Figure 15: Intelligibility results on SUS for IH1.3 (Hindi)

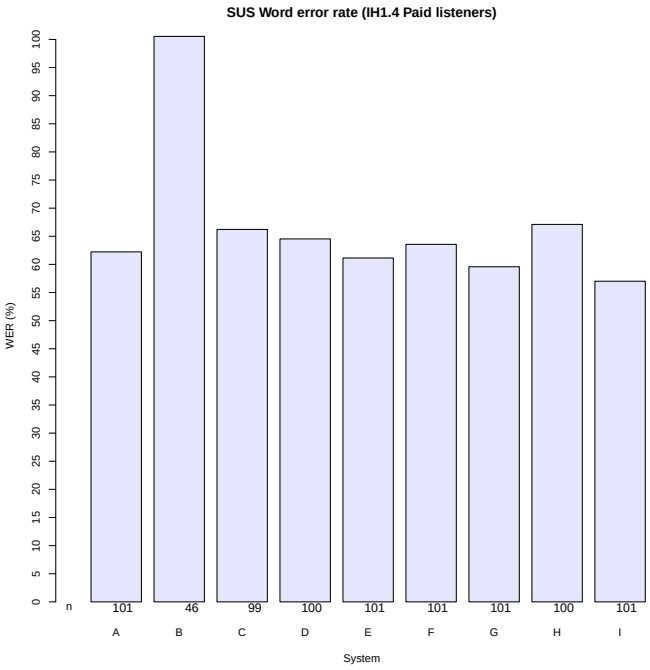

Figure 16: Intelligibility results on SUS for IH1.4 (Rajasthani)

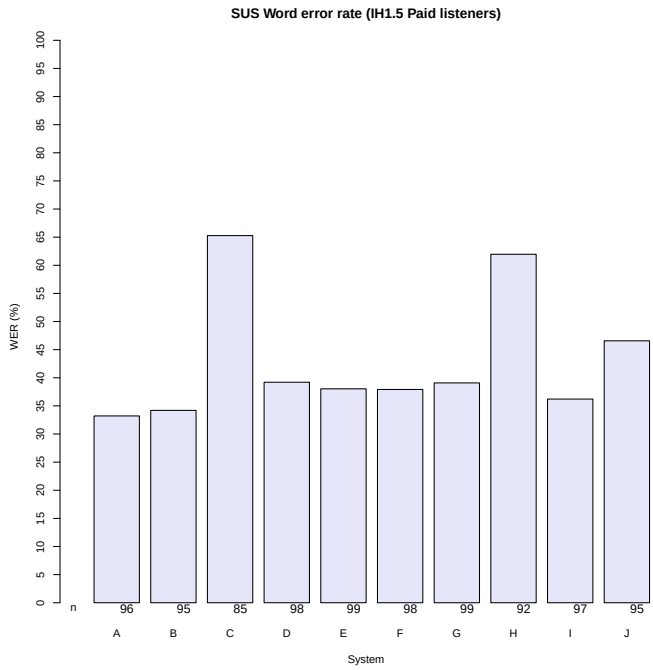

Figure 17: Intelligibility results on SUS for IH1.5 (Tamil)

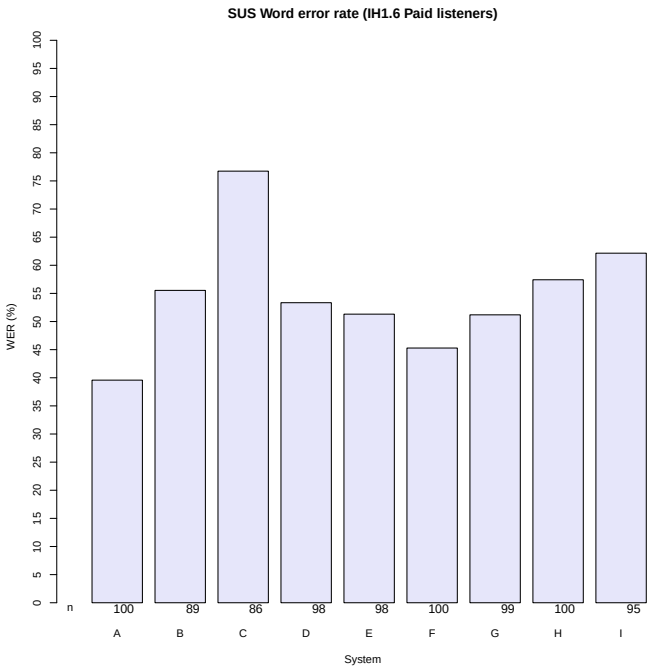

Figure 18: Intelligibility results on SUS for IH1.6 (Telugu)

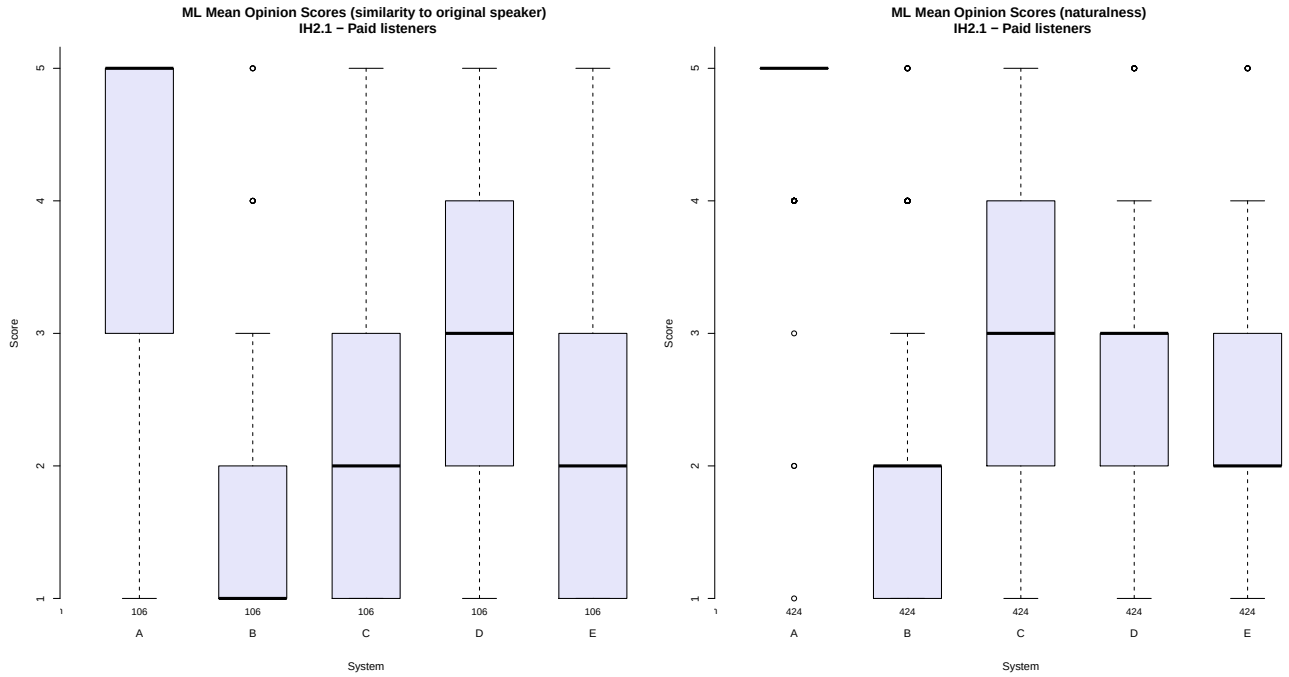

Figure 19: Similarity and Naturalness results on ML for IH2.1 (Assamese)

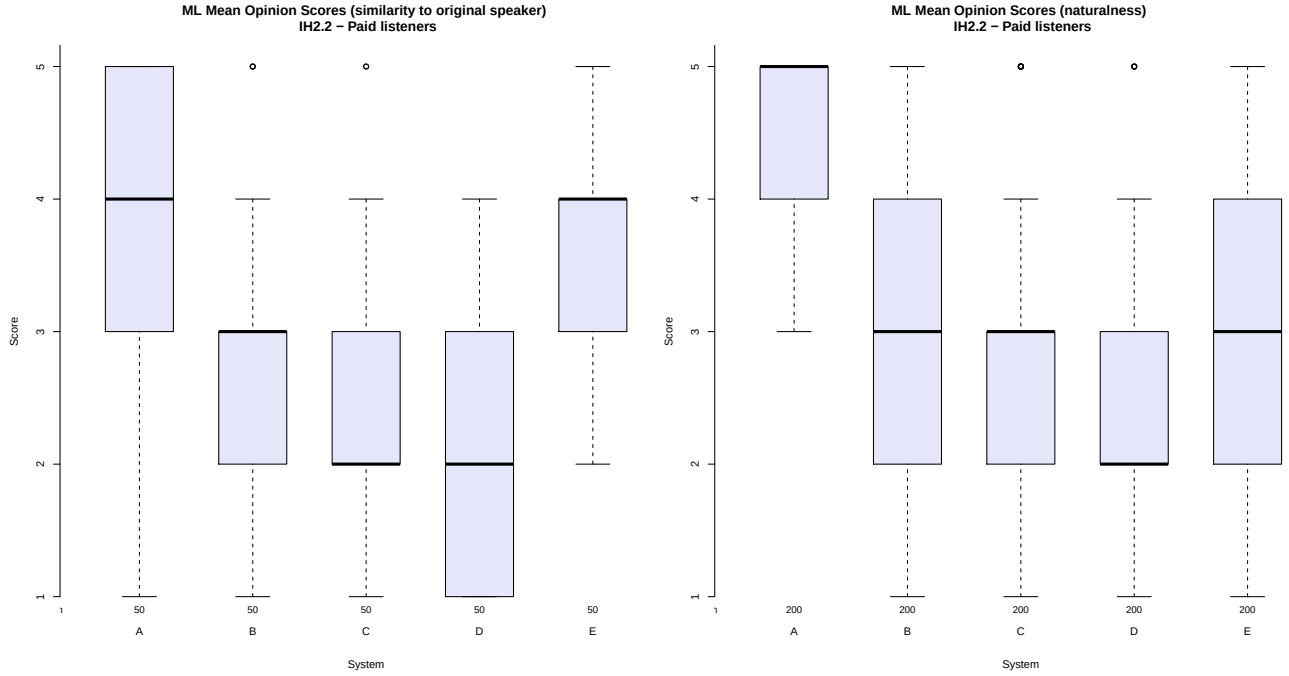

Figure 20: Similarity and Naturalness results on ML for IH2.2 (Gujarati)

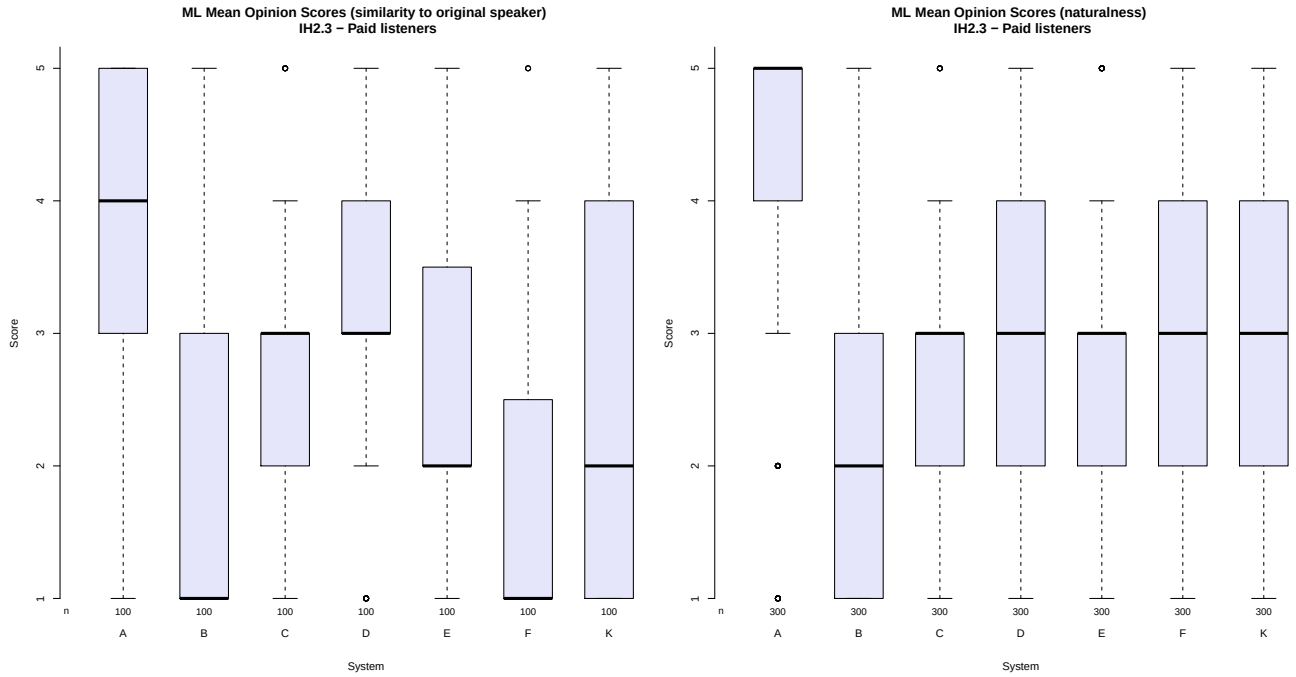

Figure 21: Similarity and Naturalness results on ML for IH2.3 (Hindi)

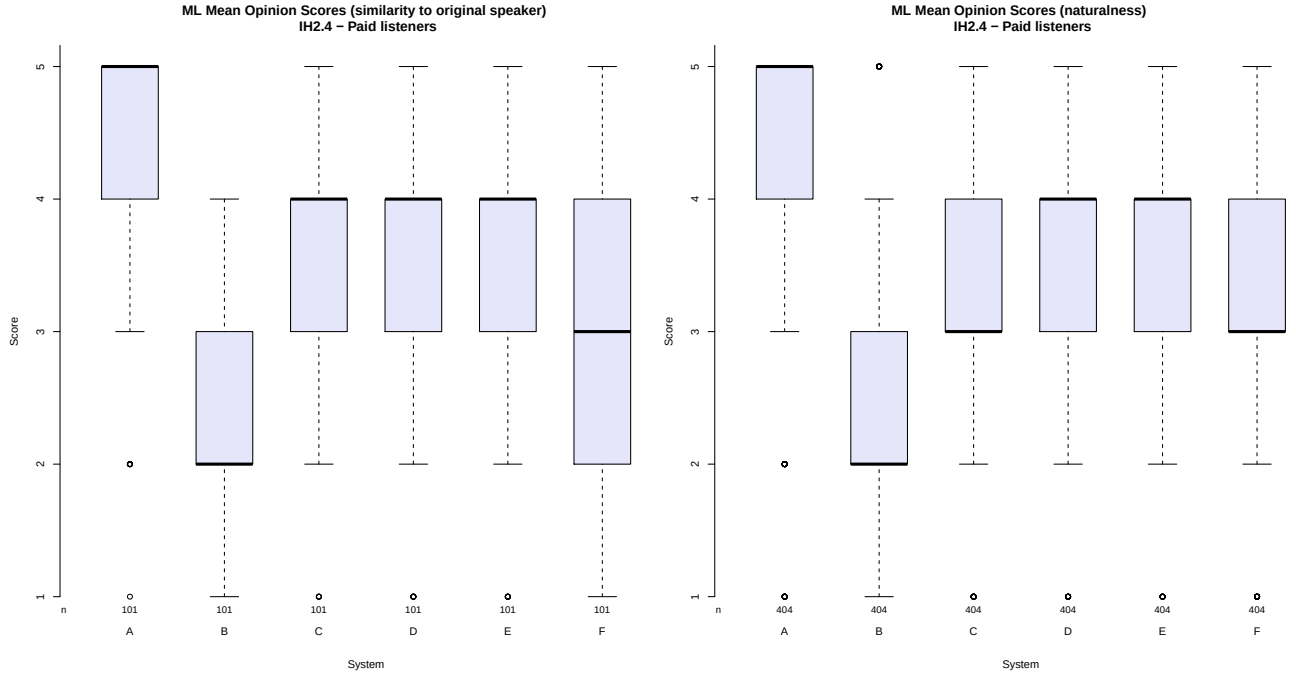

Figure 22: Similarity and Naturalness results on ML for IH2.4 (Rajasthani)

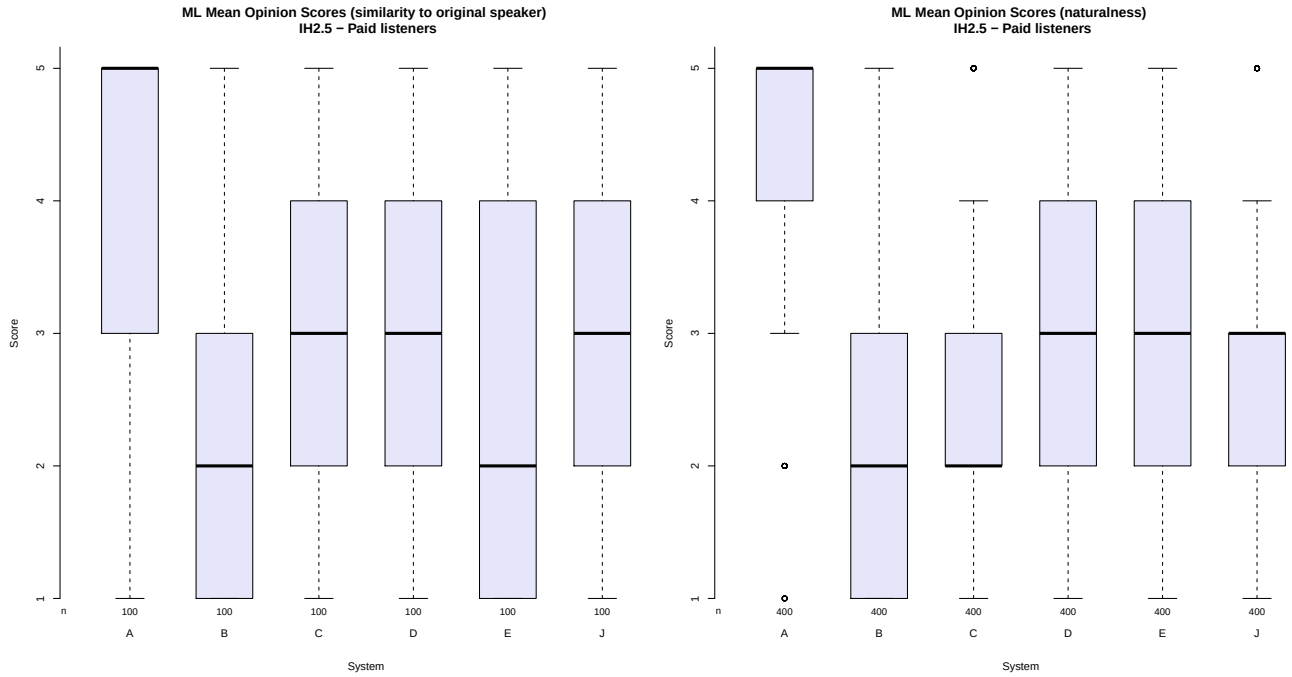

Figure 23: Similarity and Naturalness results on ML for IH2.5 (Tamil)

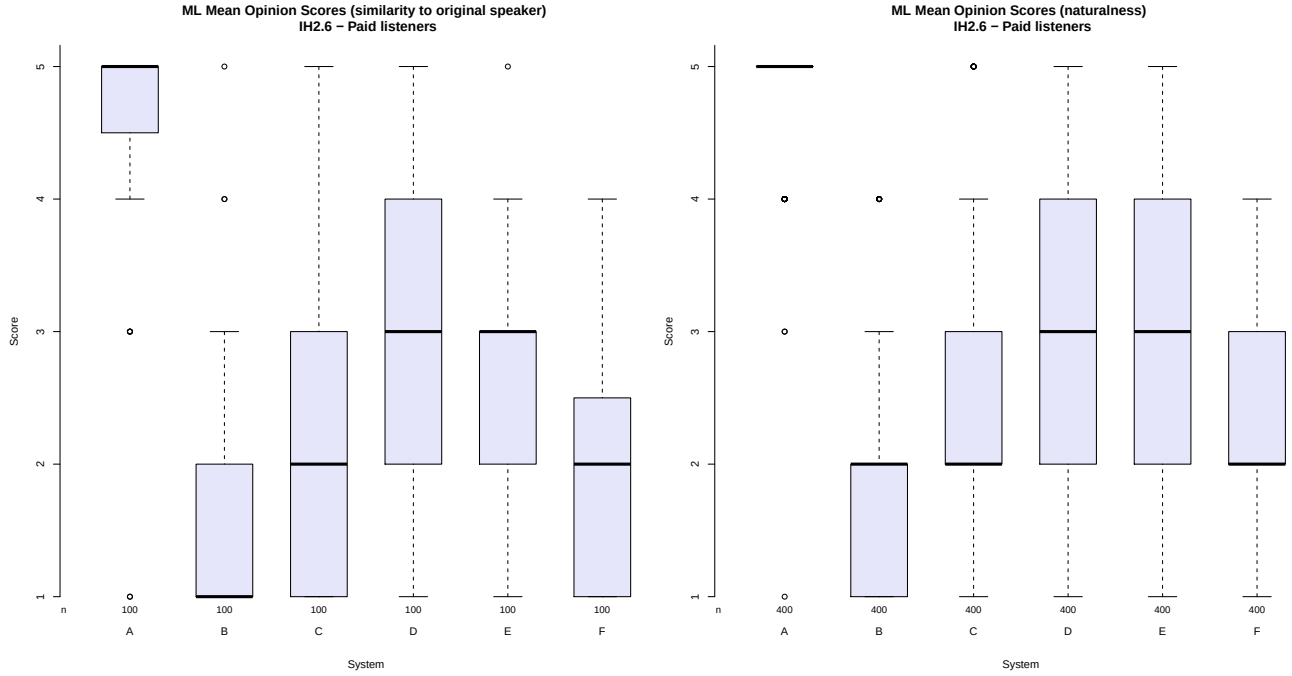

Figure 24: Similarity and Naturalness results on ML for IH2.6 (Telugu)