The Blizzard Challenge 2013 - Indian Language Tasks

Kishore Prahallad[1], *Anandaswarup Vadapalli*[1], *Naresh Elluru*[1]
Gautam Mantena[1], *Bhargav Pulugundla*[1], *Peri Bhaskararao*[1],
Hema A. Murthy[2], *Simon King*[3], *Vasilis Karaiskos*[4], *and Alan W. Black*[5]

[1] Speech and Vision Lab, IIIT Hyderabad, India
[2] Department of CSE, IIT Madras, India
[3] Center for Speech Technology Research, University of Edinburgh, UK
[4] School of Informatics, University of Edinburgh, UK
[5] Language Technologies Institute, Carnegie Mellon University, USA

Abstract

The Blizzard challenge 2013 was the ninth annual Blizzard challenge which was organized by University of Edinburgh handling the English language tasks and IIIT Hyderabad handling the Indian language tasks. This paper decsribes the Indian language tasks in the Blizzard challenge 2013. The Indian language tasks consisted of data from four Indian languages : Hindi, Bengali, Kannada and Tamil taken from the IIIT-H Indic database. Eight participants from across the world used the speech data provided as well as the corresponding text data 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. [1–9] are summary papers which provide information about the previous challenges. These resources can be found on the Blizzard Challenge website [10]. This paper is a summary paper describing the Indian language tasks in the Blizzard 2013 challenge.

This paper is organized as follows. We first discuss the nature of scripts and sounds in Indian languages. We then describe the Indian language tasks in the Blizzard 2013 challenge. Following that we discuss the results obtained for the various tasks.

2. Nature of scripts and sounds of Indian languages

As a majority of Indian languages (IL) use Indic scripts (IS) that are derived from the ancient Brahmi script, they share several orthographic patterns. ISs are also called Brahmic scripts. The basic units of an IS are referred to as Aksharas. The properties of the Aksharas are as follows : (1) An Akshara is an orthographic representation of one or more speech sounds in the concerned IL; (2) Aksharas are mostly syllabic in nature; (3) The canonical shapes of an Akshara are V, CV, CCV, and CCCV, and thus have a generalized form of C^*V, where C stands for a consonant and V for a vowel.

2.1. Convergence and divergence

Most of the ILs (except a few such as English and Urdu) share a common phonetic base, i.e., they share a common set of speech sounds and in addition possess a few more sounds individually. This common phonetic base consists of about 50 phones, including 15 vowels and 35 consonants. While all these languages share a common phonetic base, some of the languages like Hindi, Marathi and Nepali also share a common script called Devanagari. Languages such as Gujarati, Panjabi, Oriya, Bengali, Assamese, Telugu, Kannada and Tamil have their own Brahmic scripts.

The property that separates these languages at speech level can be attributed to the phonotactics in each of these languages, rather than the scripts and speech sounds. Phonotactics are permissible combinations of phones that can co-occur in a language. This implies that the distribution of syllables in each language is different. Prosody (intonation, duration and prominence) associated with a syllable is another property that separates these ILs significantly.

2.2. Digital representation

Prior to ISCII and Unicode, there were several representations for scripts in ILs. This included several fonts for each script and several mechanisms of keying-in the scripts using the QWERTY keyboard, such as soft keyboards, keyboard layouts and transliteration schemes. With the advent of Unicode, each letter in each IS has a unique code-point. This has standardized the representation of Aksharas and their rendering on the computer screen.

However, the keying-in mechanism of these Aksharas has not yet been standardized. It is hard to remember and key-in the Unicode of these scripts directly by a lay user of computers. Thus soft keyboards and keyboard layouts on top of QWERTY letters to key-in is another popular method. Once these Aksharas are keyed-in, they are internally processed and converted to strings of Unicode code-points. Due to this non-standardization, the keying-in mechanism of ISs has to be addressed explicitly during the development of text processing modules in text-to-speech systems and user interfaces.

3. Indian Language (IH) Tasks

3.1. Participants in the Challenge

The Indian language tasks of the Blizzard challenge 2013 consisted of the eight participants listed in Table 1. To anonymize the results, the systems are identified using letters, with A denoting natural speech and D to R denoting the systems submitted by the participants in the challenge.

Table 1: Participants in the Indian language tasks of Blizzard Challenge 2013

Short Name	Details	Synthesis Method
NATURAL	Natural Speech	Human
I^{2}R	Institute for Infocomm Research	Unit selection
DFKI	Deutsche Forschungszentrum für Künstliche Intelligenz	Hybrid
CMU	Carnegie Mellon University	HMM
NITECH	Nagoya Institute of Technology	HMM
USTC	National Engineering Laboratory of Speech & Language Information Processing	Hybrid
ILSP	Institute for Language and Speech Processing / Innoetics	Unit selection
S4A	Simple4All project consortium	HMM
MILE-TTS	Dept. of Electrical Engg, Indian Institute of Science	Unit selection

3.2. Database Used

Speech and text data for four Indian languages i) Hindi, ii) Bengali, iii) Kannada and iv) Tamil were released from the IIIT-H Indic database [11]. The speech data for each language was about 1 hour, spoken by native non-professional speakers in a quiet office environment. Along with the speech data the corresponding text was provided in the UTF-8 format. Table 2 shows the statistics of the text data for the four languages. No other information, like segment labels was provided as part of the challenge. However, there was no restriction on the participants to learn / use information like phonesets or labels from other resources [11], [12].

3.3. Challenges

Participants were asked to build synthetic voices from the databases in accordance of the rules of the challenge [13]. The tasks were numbered from IH1.1 to IH1.4 corresponding to the four languages, as listed below.

- IH1.1 - Hindi
- IH1.2 - Bengali
- IH1.3 - Kannada
- IH1.4 - Tamil

For each task, the synthetic voice built by each participant was evaluated through listening tests on the following test sets.

- WPD (Wikipedia) : 100 distinct sentences, which are not a part of the IIIT-H Indic database
- SUS (Semantically Unpredictable Sentences) : 100 distinct semantically unpredictable sentences which are not a part of WPD or IIIT-H Indic database

The semantically unpredictable sentences were prepared in the following manner. 100 sentences were randomly selected from text and POS tagging was performed by running IIITH-LTRC shallow parser [14] on these sentences. The words in each sentence were then reordered as : *Subject Object Verb Conjuction Subject Object Verb*.

3.4. 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 and IH1.3 consisted of eight sections and the listening tests for IH1.2 and IH1.4 consisted of two sections. The different sections of the listening tests are described below.

- Listening tests for IH1.1 and IH1.3
  1. one section for Similarity using WPD data
  2. five sections for Naturalness (one section using WPD data and four sections using SUS data)
  3. two sections for Intelligibility using SUS data
- Listening tests for IH1.2 and IH1.4
  1. one section for Similarity using WPD data
  2. one section for Naturalness using WPD data

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.

3.5. Changes made in the evaluation portal

The following changes were made in the evaluation portal, to enable the conduct of listening tests for the Indian language tasks :

- All HTML tags in the evaluation portal were rewritten to make it compatible with the HTML5 standard.
- A test to verify whether the listener was a native speaker of that language was added. Each participant was shown an image containing a sequence of words in the script of that language. They then had to type in the English translation, which was then automatically checked for correctness. All listeners were allowed to proceed with the evaluations only after clearing this test.

Table 2: Statistics of text data for the four languages

Language	No.of sentences	No.of words		No.of Syllables		No.of Phones		Avg.Words per line
		Total	Unique	Total	Unique	Total	Unique	
Hindi	1000	8273	2145	19771	890	30723	58	8
Bengali	1000	7877	2285	25757	866	37287	47	7
Kannada	1000	6652	2125	25004	851	37651	51	6
Tamil	1000	7045	2182	23284	930	42134	35	7

- The ability to type in Indian language text was added for the sections where the listeners were required to listen to an utterance and type in what they heard. This was achieved by linking to Google Transliterate (`http://www.google.com/inputtools/cloud/features/transliteration.html`). However as discussed in Section 2.2 this is hard, due to the non-standardization of transliteration schemes.

4. Discussion and Results

The following listener types were used for the listening tests :

- Paid users
- Online volunteers

Table 3 shows the statistics of the different listener types for the tasks (IH1.1 to IH1.4).

Table 3: User statistics for the tasks IH1.1 (Hindi), IH1.2 (Bengali), IH1.3 (Kannada) and IH1.4 (Tamil)

Task	Paid Users	Online Volunteers	Total
IH1.1	55	71	126
IH1.2	62	22	84
IH1.3	84	17	101
IH1.4	47	16	63

In all the results, the mean opinion scores are presented as standard boxplots [15], where the median is represented by a solid bar across the box showing the quartiles; and the whiskers represent 1.5 times the inter quartile range and the circles are the outliers beyond this range. Word error rates are presented as bar charts.

4.1. Paid vs. Online volunteers

Figures 1 to 4 show a comparison of the results obtained using paid volunteers with results obtained using online volunteers for the IH1.1 task (Hindi).

As can be seen from the plots, there is a difference between the results obtained using paid and online volunteers for the same datasets and same task. An examination of the plot in Figure 4 shows that the SUS WER for natural speech is significantly higher in the case of online volunteers as compared to paid volunteers. This indicates that paid volunteers are more careful and attentive and that it is better to use results obtained from paid volunteers.

For the remainder of the paper, we present only results obtained from paid volunteers.

For the purpose of discussing the results obtained from the four tasks, we group the IH1.1 (Hindi) and IH1.3 (Kannada) tasks into one group and the IH1.2 (Bengali) and IH1.4 (Tamil) tasks into another group. The reasons for doing so are two fold. Firstly, the test sets for IH1.1 and IH1.3 tasks contained data from both the WPD and SUS datasets, whereas the test sets for IH1.2 and IH1.4 contained data from only the WPD dataset. Secondly, the letter to sound rules in Hindi and Kannada are less complex than the letter to sound rules in Bengali and Tamil.

The results obtained for IH1.1 (Hindi) and IH1.3 (Kannada) tasks are discussed in 4.2, while the results obtained for IH1.2 (Bengali) and IH1.4 (Tamil) tasks are discussed in 4.3.

In all the results discussed below, System A refers to natural speech.

4.2. Results obtained for IH1.1 (Hindi) and IH1.3 (Kannada) Tasks

Tables 4 and 5 show the mean opinion scores of similarity and naturalness on WPD dataset while Tables 6 and 7 show the mean opinion scores for naturalness and the word error rate on SUS dataset, for the IH1.1 (Hindi) task. Figures 5 and 6 show plots of the results of similarity, naturalness and intelligibility tests on both the WPD and SUS datasets.

For the IH1.3 task (Kannada), Tables 8 and 9 show the mean opinion scores of similarity and naturalness on WPD dataset while Tables 10 and 11 show the mean opinion scores for naturalness and the word error rate on SUS dataset. Figures 7 and 8 show plots of the results of similarity, naturalness and intelligibility tests on both the WPD and SUS datasets.

4.3. Results obtained for the IH1.2 (Bengali) and IH1.4 (Tamil) Tasks

Tables 12 and 13 show the mean opinion scores of similarity and naturalness on WPD dataset, for the IH1.2 (Bengali) task. Figure 9 shows the plots of the results of similarity and naturalness tests for the IH1.2 (Bengali) on the WPD dataset.

For the IH1.4 task (Tamil), Tables 14 and 15 show the mean opinion scores of similarity and naturalness on WPD dataset. Figure 10 shows the plots of the results of similarity and naturalness tests for the IH1.4 (Tamil) on the WPD dataset.

4.4. Discussion of results

A study of the WPD and SUS mean opinion scores for naturalness (Tables 5, 6 and Tables 9, 10), for both IH1.1 (Hindi) and IH1.3 (Kannada), shows that the scores obtained by the systems for both the WPD and SUS datasets are in similar ranges. This can be explained by the fact that Indian languages are relatively free word order, and so the word reordering during the generation of SUS sentences may not have an effect on the output of the system. As a result the outputs for both WPD and SUS sentences are scored similarily for naturalness.

An examination of the WPD mean opinion scores for similarity (Tables 4, 8, 12 and 14) shows that System L has the highest score among all systems for IH1.1 (Hindi), IH1.2 (Bengali) and IH1.4 (Tamil) and the third highest score for IH1.3 (Kannada). A similar examination of the WPD mean opinion scores for naturalness (Tables 5, 9, 13 and 15), shows that System L again has the highest score among all systems for all four languages. Analysis of the SUS mean opinion scores for naturalness (Tables 6, 10) shows that System L has the second highest score for IH1.1 (Hindi) and the highest score for IH1.3 (Kannada). This shows that System L scores high, both in similarity to original speaker as well as in naturalness of output, for both the WPD and SUS dataset. However, System L does not perform as well in terms of the SUS WER.

In terms of the SUS WER (Tables 7 and 11), System D has the best performance for both IH1.1 (Hindi) and IH1.3 (Kannada). However, it's performance in terms of the mean opinion scores for similarity and naturalness (both WPD and SUS datasets) are poor, for all four languages (IH1.1 (Hindi), IH1.2 (Bengali), IH1.3 (Kannada), IH1.4 (Tamil)).

5. Acknowledgements

Thanks to the following individuals and their students for help with the listening tests: Ms. Kasthuri Sekaran (IIT Madras), Dr. Nagarajan (SSN College of Engineering), Prof. A.G. Ramakrishnan (MILE Lab, IISc Bangalore), Dr. K.S. Rao (IIT Kharagpur), Dr. Jayanna (Siddaganga Institute of Technology), Dr. Hemanth Patil (DAIICT, Gandhinagar), Prof. S.R. Mahadeva Prasanna (IIT-Guwahati) and Dr. Suryakanth V Gangashetty (IIIT Hyderabad).

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] S. King and V. Karaiskos, "The Blizzard Challenge 2010," in *Proceedings Blizzard Workshop 2010*, 2010.

[8] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proceedings Blizzard Workshop 2011*, 2011.

[9] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proceedings Blizzard Workshop 2012*, 2012.

[10] "The Blizzard Challenge Website," http://www.synsig.org/index.php/Blizzard_Challenge.

[11] K. Prahallad, E. N. Kumar, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H Indic Speech Databases," in *Proceedings of Interspeech*, Portland, Oregon, USA, 2012.

[12] H. A. Murthy *et al.*, "Building Unit Selection Speech Synthesis in Indian Languages : An Initiative by an Indian Consortium," in *Proceedings of COCOSDA*, Kathmandu, Nepal, 2010.

[13] "The Blizzard Challenge 2013 Rules," http://www.synsig.org/index.php/Blizzard_Challenge_2013_Rules.

[14] P. Avinesh and K. Gali, "Part-of-Speech tagging and chunking using conditional random fields and transformation based learning," in *Proceedings of the IJCAI-07 workshop on Shallow Parsing in South Asian Languages*, 2007.

[15] 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.

Table 4: WPD Mean Opinion Scores (Similarity to original speakers - Paid listeners) for IH1.1 (Hindi)

System	Mean	Std. Deviation
A	4.2	0.95
D	2.3	1.20
E	2.1	1.22
F	**1.7**	0.99
I	2.5	1.29
K	**2.8**	1.16
L	**2.8**	1.25
P	2.2	1.11

Table 5: WPD Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.1 (Hindi)

System	Mean	Std. Deviation
A	4.7	0.72
D	**2.4**	0.87
E	2.7	1.15
F	2.6	1.25
I	2.8	1.16
K	3.5	0.86
L	**3.7**	1.07
P	2.8	1.08

Table 6: SUS Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.1 (Hindi)

System	Mean	Std. Deviation
A	4.6	0.83
D	**2.3**	0.97
E	2.5	0.90
F	**2.3**	0.99
I	2.8	0.97
K	**3.5**	0.98
L	3.2	1.23
P	2.7	1.10

Table 7: SUS Word Error Rate - Paid listeners for IH1.1 (Hindi)

System	Mean	Std. Deviation
A	34%	24%
D	**43**%	26%
E	47%	27%
F	53%	23%
I	**57**%	26%
K	**43**%	25%
L	53%	26%
P	50%	24%

Table 8: WPD Mean Opinion Scores (Similarity to original speakers - Paid listeners) for IH1.3 (Kannada)

System	Mean	Std. Deviation
A	4.1	1.3
D	2.0	1.3
F	1.8	1.3
I	2.2	1.4
K	**1.7**	1.2
L	2.5	1.5
P	2.8	1.5
R	**3.0**	1.5

Table 9: WPD Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.1 (Kannada)

System	Mean	Std. Deviation
A	4.5	0.81
D	2.9	1.26
F	**2.5**	1.31
I	**2.5**	1.16
K	3.0	1.18
L	**3.7**	1.15
P	3.5	1.02
R	3.4	1.01

Table 10: SUS Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.3 (Kannada)

System	Mean	Std. Deviation
A	4.4	1.0
D	2.8	1.1
F	**2.2**	1.1
I	2.4	1.1
K	3.0	1.2
L	**3.1**	1.2
P	2.8	1.2
R	2.6	1.2

Table 11: SUS Word Error Rate - Paid listeners for IH1.3 (Kannada)

System	Mean	Std. Deviation
A	48%	31%
D	**50**%	28%
F	62%	29%
I	**72**%	26%
K	55%	29%
L	57%	29%
P	57%	27%
R	67%	26%

Table 12: WPD Mean Opinion Scores (Similarity to original speakers - Paid listeners) for IH1.2 (Bengali)

System	Mean	Std. Deviation
A	4.7	0.49
D	**2.1**	0.97
F	**2.1**	1.15
I	2.3	1.01
K	2.7	1.11
L	**3.5**	1.14
P	2.9	1.19

Table 13: WPD Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.2 (Bengali)

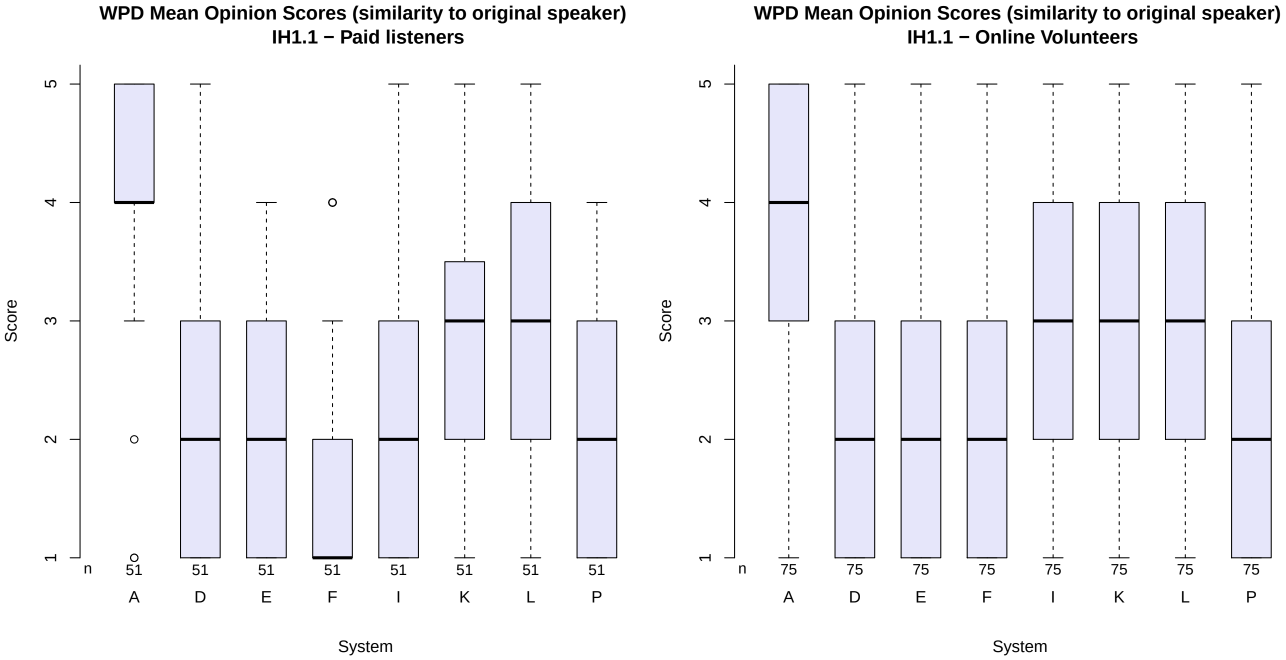
System	Mean	Std. Deviation
A	4.7	0.53
D	2.6	0.76
F	2.4	0.90
I	**2.1**	0.83
K	3.3	1.01
L	**3.8**	0.83
P	3.0	0.89

Table 14: WPD Mean Opinion Scores (Similarity to original speakers - Paid listeners) for IH1.4 (Tamil)

System	Mean	Std. Deviation
A	4.4	0.99
D	2.7	1.00
F	1.9	0.83
I	**1.8**	0.84
K	2.7	1.13
L	**3.2**	1.29
P	2.9	1.26
R	2.8	1.16

Table 15: WPD Mean Opinion Scores (naturalness - all data - Paid listeners) for IH1.4 (Tamil)

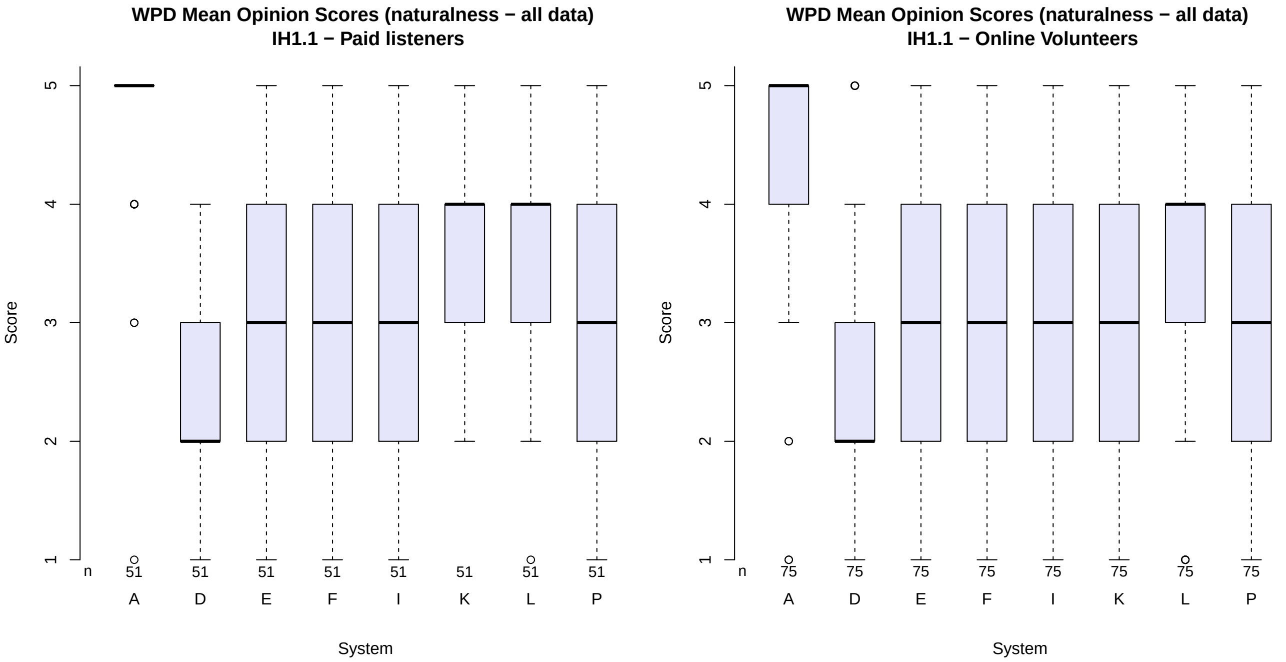
System	Mean	Std. Deviation
A	4.3	0.78
D	2.5	0.97
F	**2.2**	0.89
I	2.3	1.02
K	3.1	0.96
L	**3.9**	1.03
P	2.9	1.05
R	3.2	0.95



WPD Mean Opinion Scores (similarity to original speaker)
IH1.1 − Paid listeners

WPD Mean Opinion Scores (similarity to original speaker)
IH1.1 − Online Volunteers

Figure 1: Comparison of results of paid and online volunteers for similarity test on WPD database (IH1.1 Task)



WPD Mean Opinion Scores (naturalness − all data)
IH1.1 − Paid listeners

WPD Mean Opinion Scores (naturalness − all data)
IH1.1 − Online Volunteers

Figure 2: Comparison of results of paid and online volunteers for naturalness test on WPD database (IH1.1 Task)

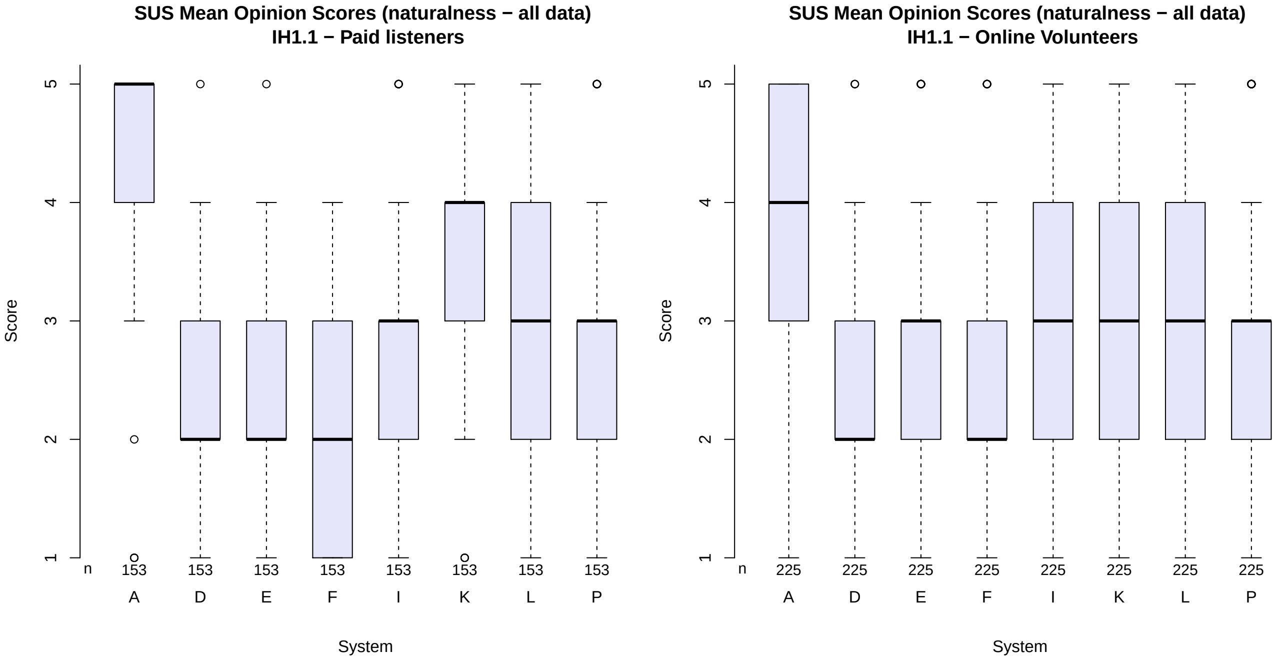


Figure 3: Comparison of results of paid and online volunteers for naturalness test on SUS database (IH1.1 Task)

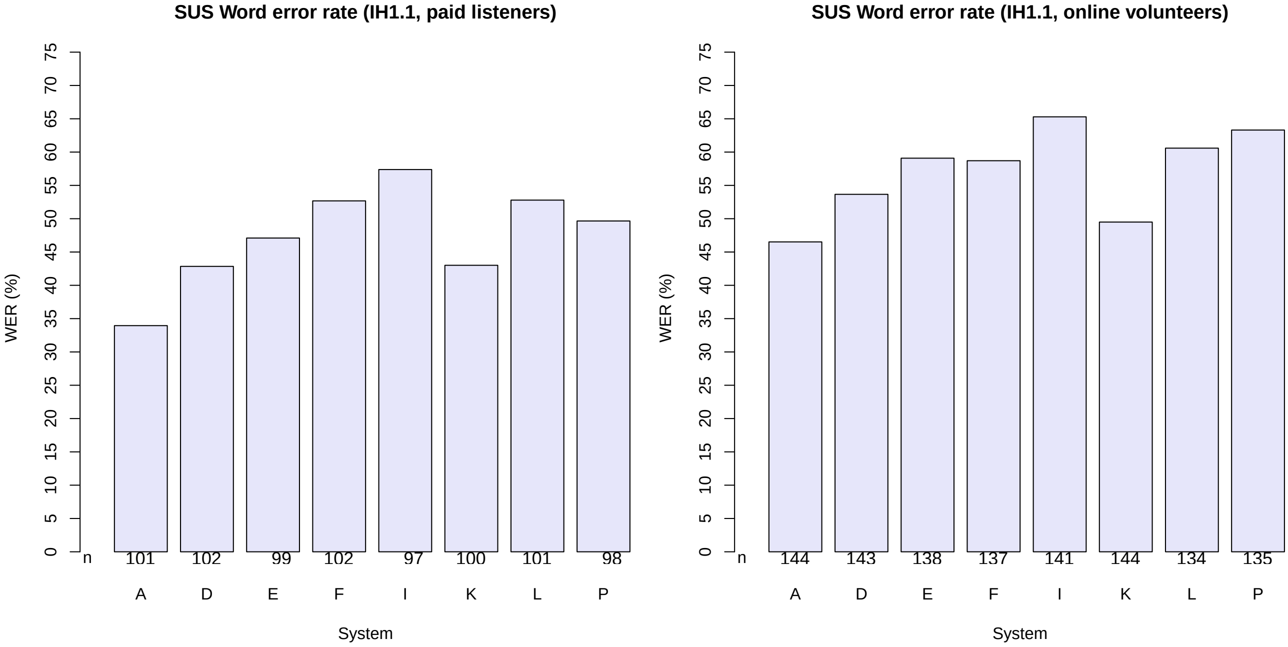


Figure 4: Comparison of results of paid and online volunteers for intelligibility test on SUS database (IH1.1 Task)

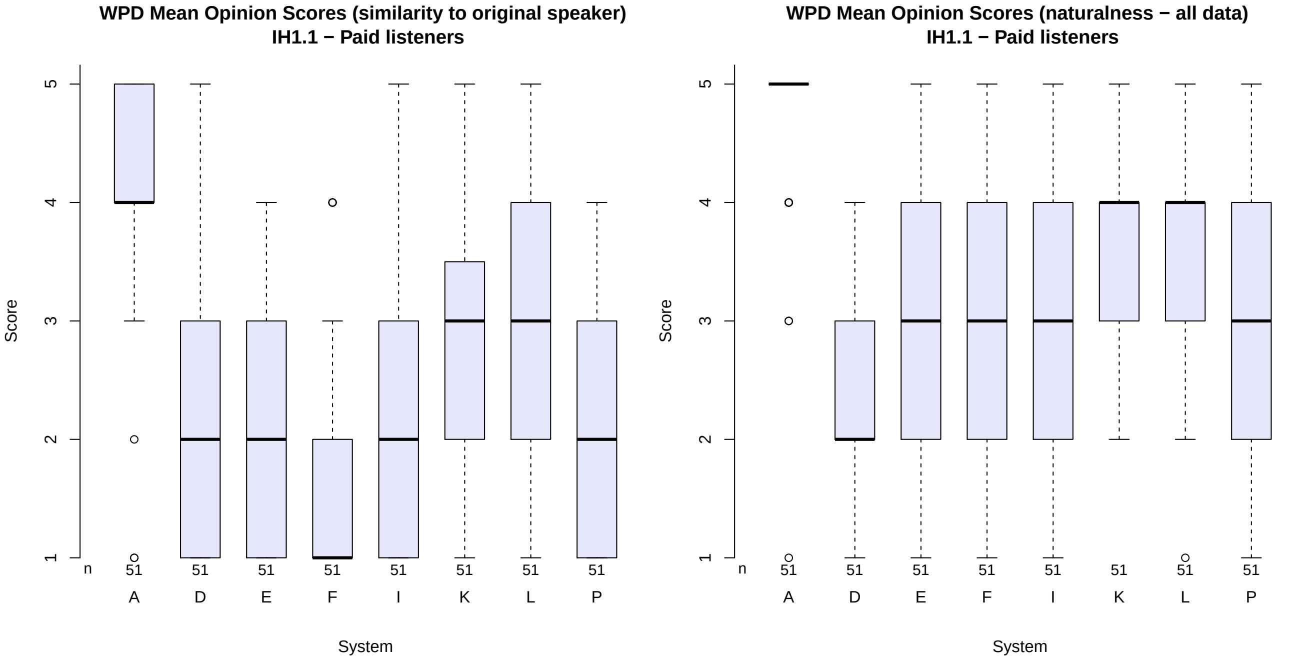


Figure 5: Similarity and Naturalness results on WPD database for IH1.1 (Hindi)

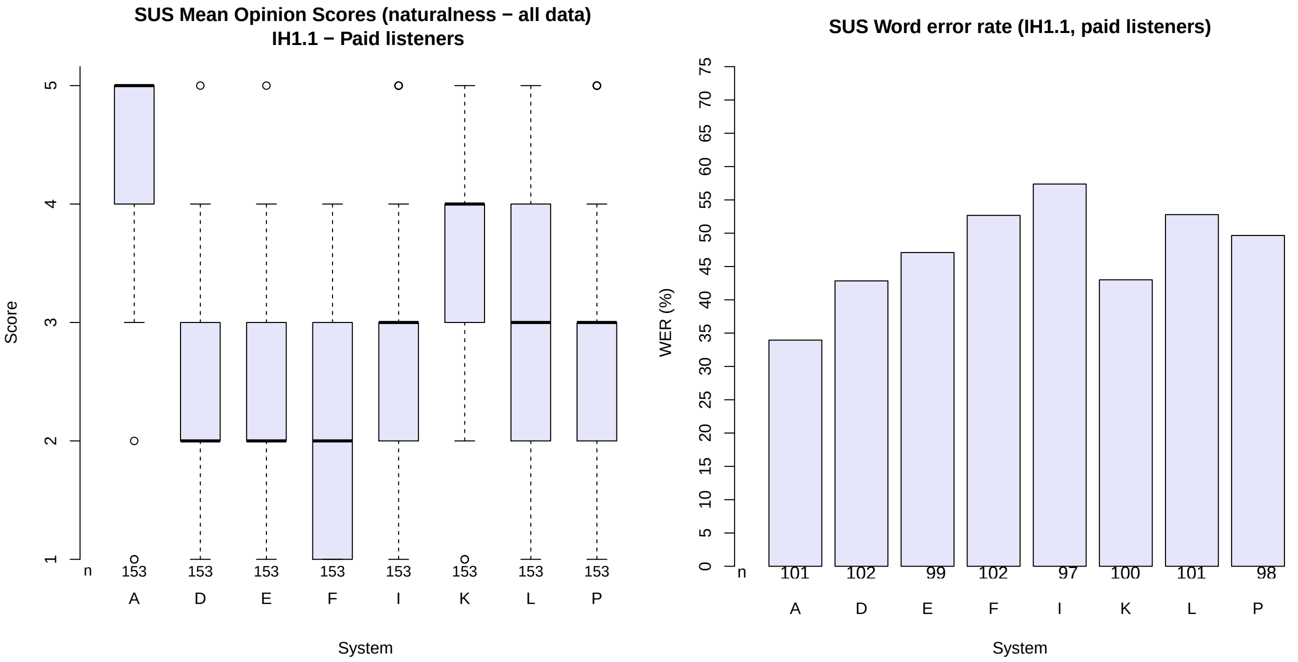


Figure 6: Naturalness and Intelligibility results on SUS database for IH1.1 (Hindi)

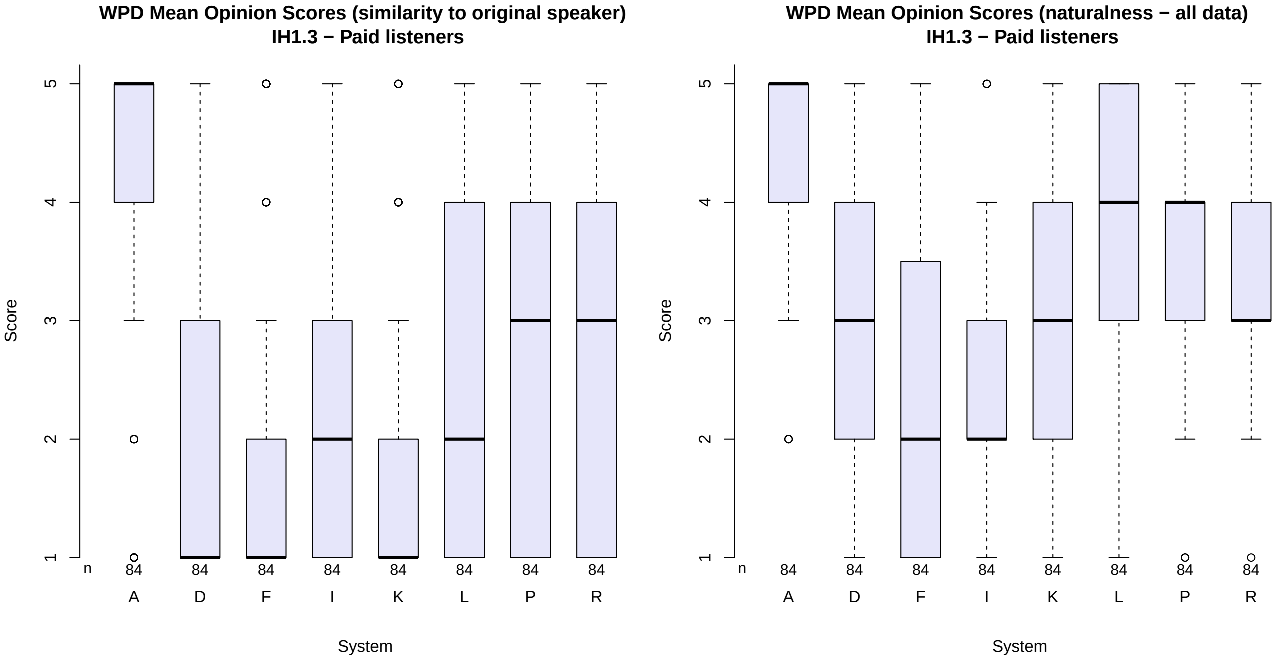


Figure 7: Similarity and Naturalness results on WPD database for IH1.3 (Kannada)

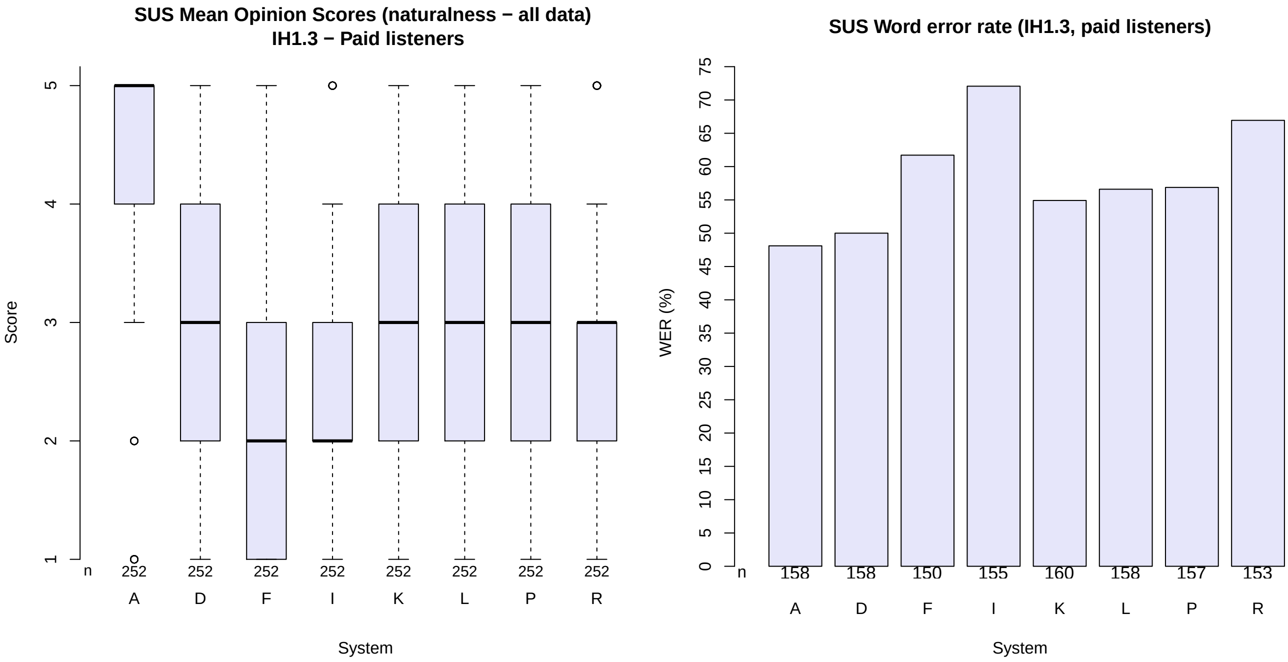


Figure 8: Naturalness and Intelligibility results on SUS database for IH1.3 (Kannada)

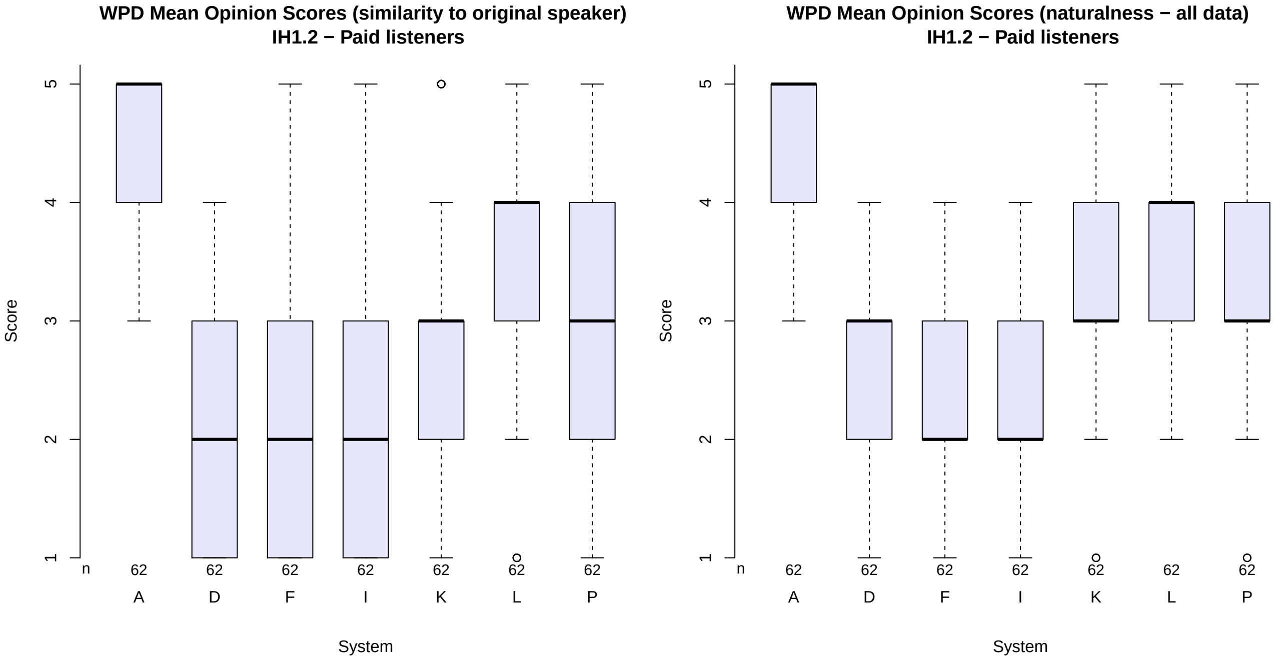


Figure 9: Similarity and Naturalness results on WPD database for IH1.2 (Bengali)

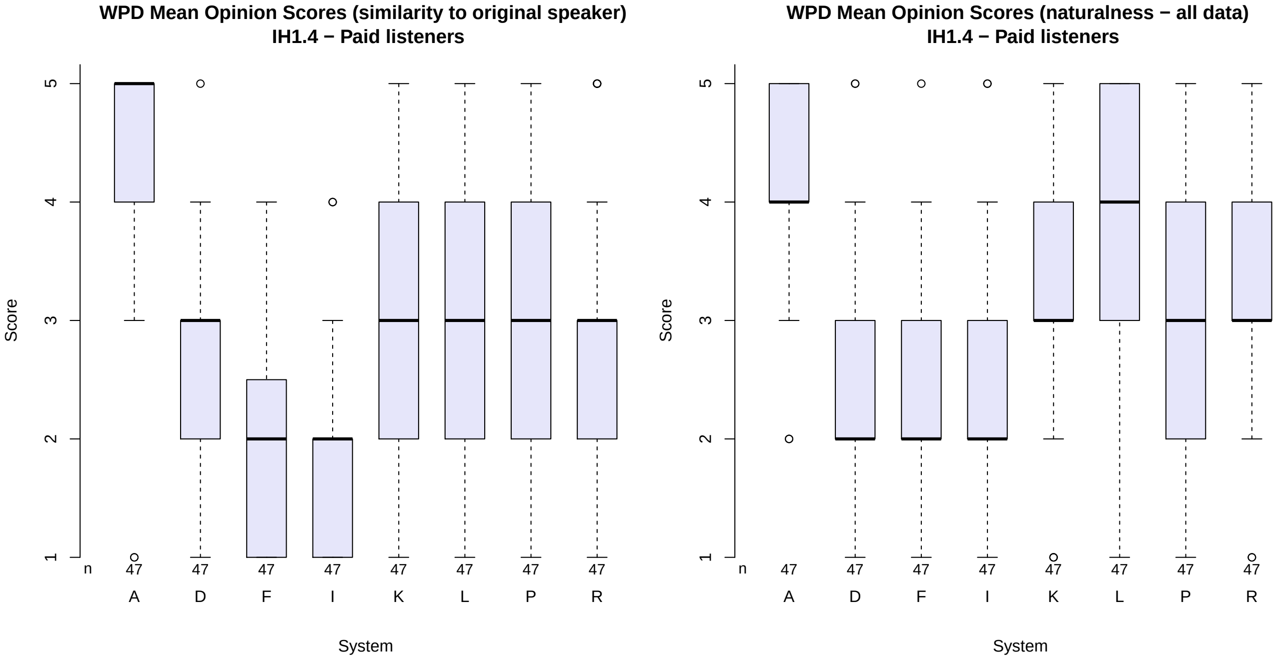


Figure 10: Similarity and Naturalness results on WPD database for IH1.4 (Tamil)