The Blizzard Challenge 2018

Simon King, Jane Crumlish, Amy Martin, Lovisa Wihlborg

The Centre for Speech Technology Research
University of Edinburgh, UK
Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2018 was the fourteenth annual Blizzard Challenge and is the twelfth consecutive one organised at the University of Edinburgh, with support from the other members of the Blizzard Challenge committee. The task this year was the same as in 2016 and 2017, and used identical data to 2017: a single-speaker English corpus, comprising around 6.5 hours of audio from 56 professionally-produced children's audiobooks.

Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

Black and Tokuda conceived the Blizzard Challenge in 2005 [1] and there have been annual summary papers like this one every year, plus one-off retrospective summary-of-summaries covering the first decade [2]. For the current and many previous Challenges, the submitted speech, reference natural samples, raw listening test responses, scripts for running the listening test and scripts for the statistical analysis, can be obtained from the Blizzard Challenge website [3].

To minimise duplication, this paper will not repeat the descriptions of the speech database, voice building task, or listening test design, since these are identical to 2017. Please read [4] before continuing!

2. Participants

29 teams registered for this year's challenge and obtained the data. Complete working DNN baselines (systems D and E in Table 1) along with example generated output, were made available to teams in advance of the submission deadline. A new element of the rules this year asked teams only to submit entries that they judged (e.g., via internal listening tests) would be better than one of these baseline systems.

Of the 29 registered teams, 10 submitted entries, as listed in Table 1 alongside human speech and the 4 benchmark systems.

The unit selection benchmark[1] is Festival configured very similarly to the Festival/CSTR entry to Blizzard 2006 [5]. This system can be replicated by following the Multisyn recipe available from `http://www.cstr.ed.ac.uk/downloads/festival/multisyn_build` and using the Unisyn dictionary [6] with the Received Pronunciation setting ('rpx'). The second benchmark[2] uses the current public release of the HTS toolkit which is available from `http://hts.sp.nitech.ac.jp`, in conjunction with the Festival front end (configured the same as the unit selection benchmark) and the STRAIGHT vocoder. The third and fourth benchmarks[3] again use the HTS toolkit, this time with a DNN acoustic model, and the same front end and vocoder as the HMM benchmark. The unit selection and HMM-based benchmarks are the same types as in many previous challenges and will aid comparisons with those previous years.

Short name	Details	Method
NATURAL	Natural speech from the same speaker as the corpus	human
FESTIVAL_BM	Festival benchmark	unit selection
HTS_BM	HTS HMM benchmark	HMM
DNN_BM	HTS DNN benchmark	DNN
DNNTrj_BM	HTS DNN benchmark with trajectory training	DNN
CMU	Carnegie Mellon University	clustergen
CSTR	Centre for Speech Technology Research, U Edinburgh	hybrid
I2R	Institute for Infocomm Research	DNN
IRISA	U Rennes	hybrid
MARY	Deutsche Forschungszentrum für Künstliche Intelligenz	DNN
NITECH	Nagoya Institute of Technology	DNN
NTUT	National Taipei University of Technology	DNN
SOGOU	Sogou Inc.	DNN
TL-NTU	Xinjiang University & Nanyang Technological University & Institute for Infocomm Research	DNN
USTC	U Science and Technology of China	hybrid

Table 1: *The participating systems and their short names. The first row is natural speech (system identifier A) and the subsequent four rows are the benchmarks (system identifiers B, C, D, E, in that order). The remaining rows are in alphabetical order of the system's short name and not in alphabetical order of system identifier. Systems are categorised as: HMM (Hidden Markov Model statistical parametric), DNN (Deep Neural Network statistical parametric, including architectures such as BLSTM), clustergen (decision tree statistical parametric), unit selection (using waveform concatenation), or hybrid (waveform concatenation guided by a statistical parametric model such as a DNN).*

When reporting results, the systems are identified using letters, with A denoting natural speech, B the Festival benchmark systems, C the HMM benchmark system, D and E the DNN benchmark systems and the remaining letters denoting the systems submitted by participants in the challenge. The system identifiers are assigned randomly each year. Most participating teams reveal their system identifier in their workshop paper.

[1] Thanks to Oliver Watts, CSTR

[2] Thanks to Kei Sawada, NIT

[3] Thanks again to Kei Sawada, NIT

3. Differences to the 2017 challenge

The 2018 test set included two complete children's books that have never been released to participants, plus previously unreleased semantically-unpredictable sentences (SUS), newspaper sentences and the Harvard/IEEE sentences [7]. Teams were asked to synthesise a substantial amount of test material, not only the 2018 test set but also the 2017 and 2016 sets.

Not only is the amount of children's book content in the 2018 test set rather small, some of the corresponding recorded natural speech is problematic because it has background music or other non-speech content. We have now exhausted the supply of audio material for this speaker, and it will not be possible to use this dataset in any further challenges.

Stimuli selected for the listening test this year were taken from the 2018 test set to the greatest extent possible, then from the 2017 set where necessary. Only book sentences, book paragraphs and SUS were employed in the listening test. The newspaper sentences and the Harvard/IEEE sentences synthesised by participants remain available for other uses.

The listening test had the same structure as 2017, with the only difference being the number of systems involved. In 2017 there were 16 systems (3 benchmarks + 13 participating teams), plus natural speech. In 2018 there were 14 systems (4 benchmarks + 10 participating teams), plus natural speech. As in 2016 and 2017, natural speech is only available for book sentences and book paragraphs, and is not available for SUS.

3.1. Listener types

Various listener types were used in the test. Letters in parenthesis below are the identifiers used for each type in the results distributed to participants:

- Paid Edinburgh University students, all native speakers of English (any accent) and generally aged 18-25. These were recruited in Edinburgh and carried out the test in purpose-built soundproof listening booths using good quality audio interfaces and headphones. All listeners of this type completed the entire listening test. (EP)
- Speech experts (self-declared), recruited via participating teams and mailing lists. (EE)
- Volunteers recruited via participating teams, mailing lists, blogs, word of mouth, etc. (ER)

As in all previous challenges, participating teams were asked to help recruit volunteer participants (in categories EE or ER) for the listening test. Table 7 summarises the listeners who participated this year.

3.2. Listening test completion rate

Table 8 gives a breakdown of evaluation completion rates for each listener type. It appears that completion rates are, as in 2017, very good with 249 listeners completing the test this year (of which 150 were paid). This supports last year's conclusion that placing responsibility on participating teams to recruit listeners is very effective. This should remain a regular feature in future challenges. On the other hand, there is variability in teams' compliance with the rule to recruit at least 10 listeners, as can be seen in Table 2. Further work is required to bring all participants into compliance!

4. Analysis methodology

As usual, for the statistical analysis presented here and at the workshop, we combined the responses from 'completed all sections' and 'partially completed' listeners together in all analyses.

Short name	Number of listeners recruited
CMU	10
CSTR	150
I2R	0
IRISA	5
MARY	10
NITECH	24
NTUT	11
SOGOU	11
TL-NTU	0
USTC	11

Table 2: *The number of listeners recruited by each participating team. The rules stipulated a minimum of 10 per team. The number for team CSTR is equal to the number of paid participants run at the University of Edinburgh; no attempt was made to recruit further volunteer listeners at this location, since the pool of available listeners was already drained.*

We only give results for all listener types combined. Analysis by listener type was provided to participants and can be obtained by non-participants by downloading the complete listening test results distribution package via the Blizzard website. Since complete raw listeners scores for every stimulus presented in the listening test are included in this distribution, re-analysis of the data is possible by anyone who wishes to do so.The organisers of the challenge would be interested to hear of any such re-analysis.

Please refer to [8] for a description of the statistical analysis techniques used and justification of the statistical significance techniques employed to produce the results presented here. In all material published by the organisers, system names are anonymised. Individual teams are free to reveal their system identifier if they wish.

5. Results

Standard boxplots are presented for the ordinal data. Please refer to [4] for information on how to interpret these. A single ordering of the systems is employed in all plots. This ordering is in descending order of mean naturalness calculated from the responses of all listeners combined and both sentence-based naturalness sections combined. Note that this ordering is intended only to make the plots more readable by using the same system ordering across all plots for both tasks and *can not be interpreted as a ranking*. In other words, the ordering does not tell us which systems are significantly better than others. Given that the presentation of results as tables, significance matrices, boxplots and bar-charts is now well established, we will not provide a detailed commentary for every result. Figure 1 shows the results for sentences.

We can compare Table 1 with the corresponding tables in the 2016 and 2017 summary papers [9, 4], to observe some changes in the pattern of systems entered into the challenge.

The only purely-HMM-based system remaining is the HTS benchmark (C), whose performance is now well towards the lower end of the field. This is a positive results because it indicates that most participating teams were able to create systems that are at least as natural as that simple, and easily-reproducible, benchmark systems. The DNN benchmarks (D and E) are around the middle, which indicates that the HTS toolkit (which was used this year, as an alternative to Merlin used in 2017) creates solid benchmark systems, and would be a good choice for anyone wishing to benchmark their own in-house system.

There are no longer any "pure" unit selection entries, except for the Festival unit selection benchmark (B), which once again performs surprisingly well on naturalness. This is remarkable for a piece of software essentially unchanged for a decade. However, the Festival benchmark has relatively poor intelligibility. There are three hybrid systems (K, L, M), which generally have above average naturalness (e.g., Figure 1).

No synthesiser is as natural as the natural speech (refer to the first row or column of Figure 2). System K is significantly more natural than all other systems, and amongst the most intelligible, although a rather large number of systems have equally low WER (Figure 4).

The multiple dimensions of scoring for the paragraphs are reported in Figures 5 to 17. Unsurprisingly, no system was judged to be as good as natural speech, along any dimension. System K is better than all other systems along most (but not quite all) dimensions.

5.1. Listener feedback

On completing the evaluation, listeners were given the opportunity to tell us what they thought through an online feedback form. All responses were optional. Feedback forms included many detailed comments and suggestions from all listener types. Listener information and feedback is summarised in Tables 3 to 31.

In the tables on the following pages, the footnotes in the captions specify whether the numbers in that table are based on listener feedback [4] or on the listening test results themselves. [5]

6. Acknowledgements

In addition to those people already acknowledged in the text, we wish to thank a number of additional contributors without whom running the challenge would not be possible. Vasilis Karaiskos provided essential advice and wisdom, accumulated over many previous years of the challenge, and sanity-checked the listening test implementation. Rob Clark designed and implemented the scripts used to perform statistical analysis; Dong Wang wrote the WER program. Tim Bunnell of the University of Delaware provide the tool to generate the SUS sentences for English. Apple and Google have provided sustained financial support, including funding for the second, third and fourth authors and for payments to listening test subjects. The listening test scripts are based on earlier versions provided by previous organisers of the Blizzard Challenge. Thanks to all participants and listeners.

7. References

[1] Alan W. Black and Keiichi Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc Interspeech 2005*, Lisbon, Portugal, September 2005.

[2] Simon King, "Measuring a decade of progress in Text-to-Speech," *Loquens*, vol. 1, no. 1, 2014.

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

[4] S. King, L Wihlborg, and W. Guo, "The Blizzard Challenge 2017," in *Proc. Blizzard Workshop*, Stockholm, Sweden, August 2017.

[5] R. Clark, K. Richmond, V. Strom, and S. King, "Multisyn voices for the Blizzard Challenge 2006," in *Proc. Blizzard Challenge Workshop (Interspeech Satellite)*, Pittsburgh, USA, September 2006.

[6] Susan Fitt, "Documentation and user guide to Unisyn lexicon and post-lexical rules," Tech. Rep., Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK, 2000.

[7] "IEEE recommended practice for speech quality measurements," *IEEE No 297-1969*, pp. 1–24, June 1969.

[8] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Workshop (in Proc. SSW6)*, Bonn, Germany, August 2007.

[9] S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Proc. Blizzard Workshop*, Cupertino, USA, September 2016.

[4] These numbers are calculated from the feedback forms that listeners complete at the end of the test. As this is optional, many listeners decide not to fill it in. If they do, they do not always reply to all the questions in the form.

[5] These numbers are calculated from the database where the results of the listening tests are stored.

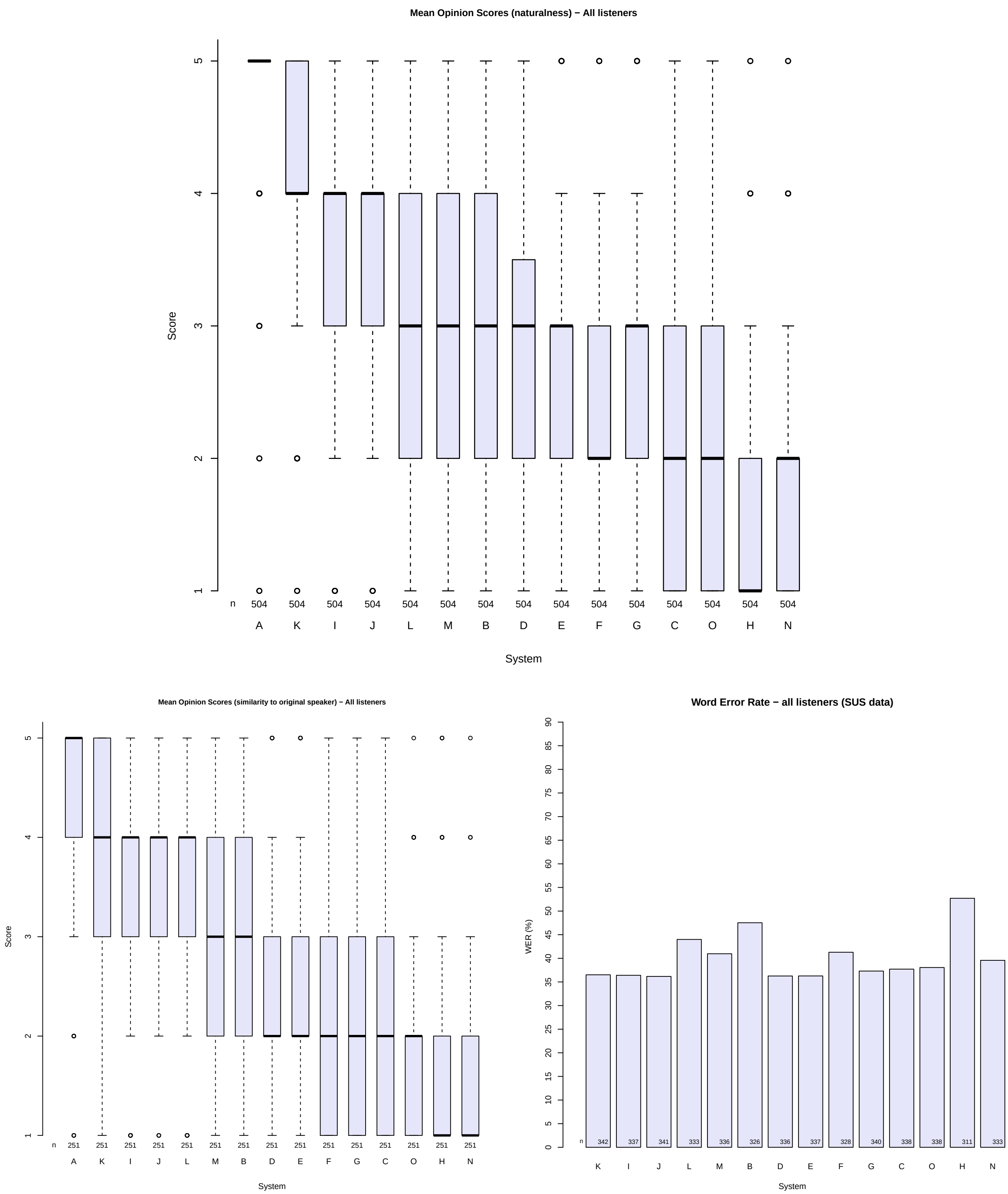

Figure 1: *Results for task 2018-EH1 on sentence test material, pooling all listeners' responses. A is natural speech, for which intelligibility results are not available. System B is the Festival unit selection benchmark, C is the HMM statistical parametric benchmark and D and E are the DNN statistical parametric benchmarks.*

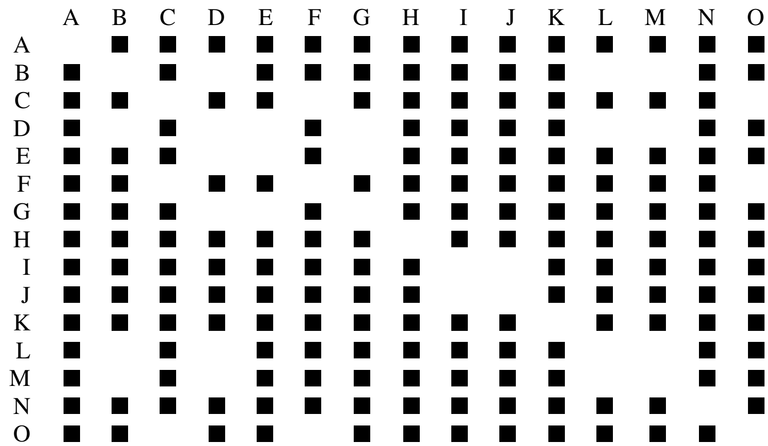

Figure 2: *Significant differences in naturalness (book sentences) between systems are indicated by a solid black box. Refer to [5] for details of significance testing.*

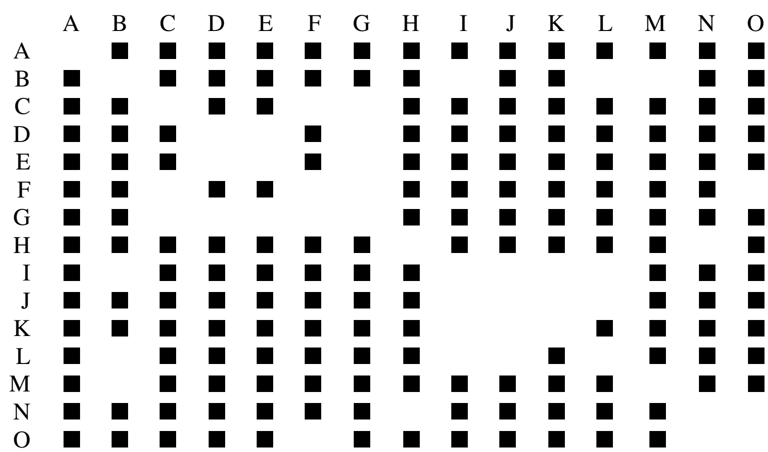

Figure 3: *Significant differences in speaker similarity (book sentences) between systems are indicated by a solid black box.*

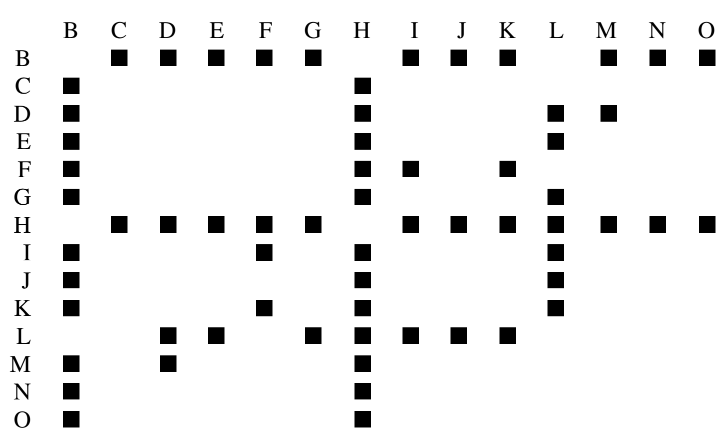

Figure 4: *Significant differences in intelligibility (SUS) between systems are indicated by a solid black box. Most systems approximately group together with equally good intelligibility. Systems B (unit selection baseline), L and H are generally of lower intelligibility than this group.*

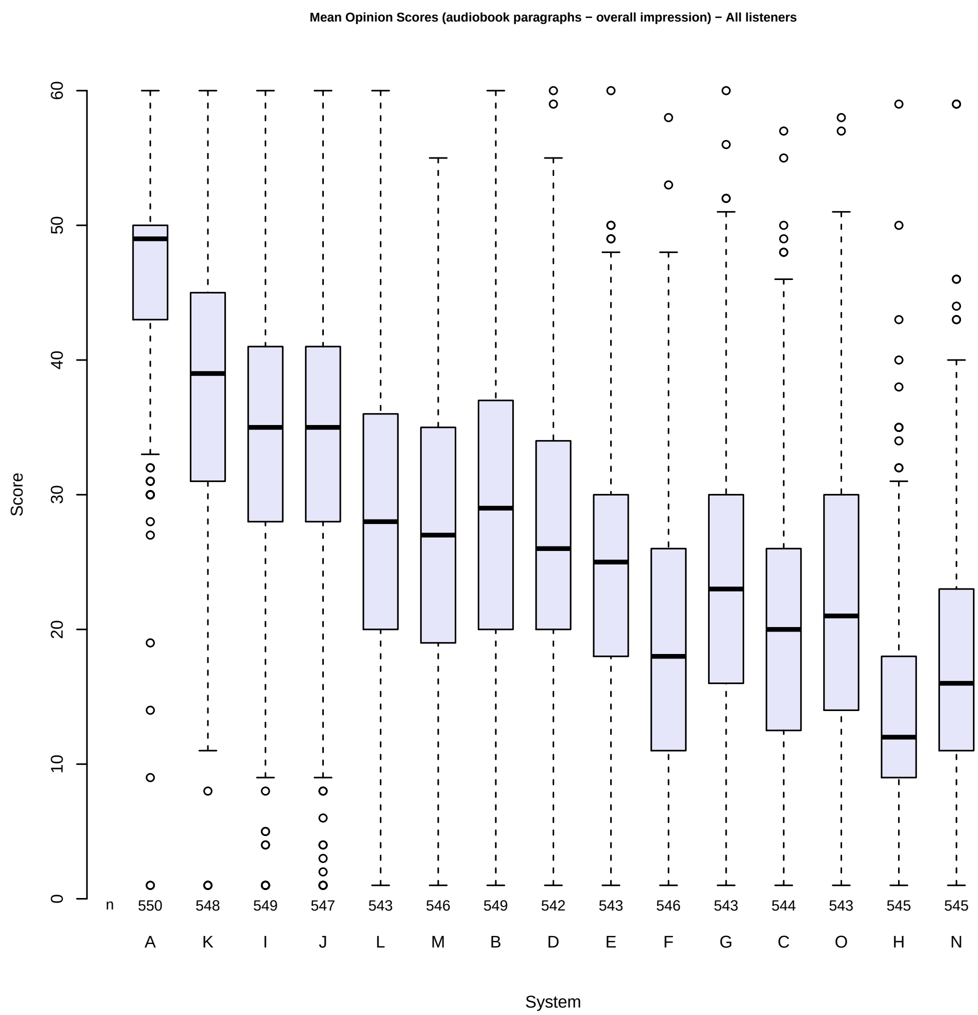

Figure 5: *Overall impression of paragraphs for task 2018-EH1.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■		■	■	■	■	■	■	■			■	■
C	■	■		■	■		■	■	■	■	■	■	■	■	
D	■		■		■	■	■	■	■	■	■			■	■
E	■	■	■	■		■		■	■	■	■	■	■	■	■
F	■	■		■	■		■	■	■	■	■	■	■	■	■
G	■	■	■	■		■		■	■	■	■	■	■	■	■
H	■	■	■	■	■	■	■		■	■	■	■	■	■	■
I	■	■	■	■	■	■	■	■			■	■	■	■	■
J	■	■	■	■	■	■	■	■			■	■	■	■	■
K	■	■	■	■	■	■	■	■	■	■		■	■	■	■
L	■		■		■	■	■	■	■	■	■			■	■
M	■		■		■	■	■	■	■	■	■			■	■
N	■	■	■	■	■	■	■	■	■	■	■	■	■		■
O	■	■		■	■	■	■	■	■	■	■	■	■	■	

Figure 6: *Significant differences in overall impression of paragraphs for task 2018-EH1.*

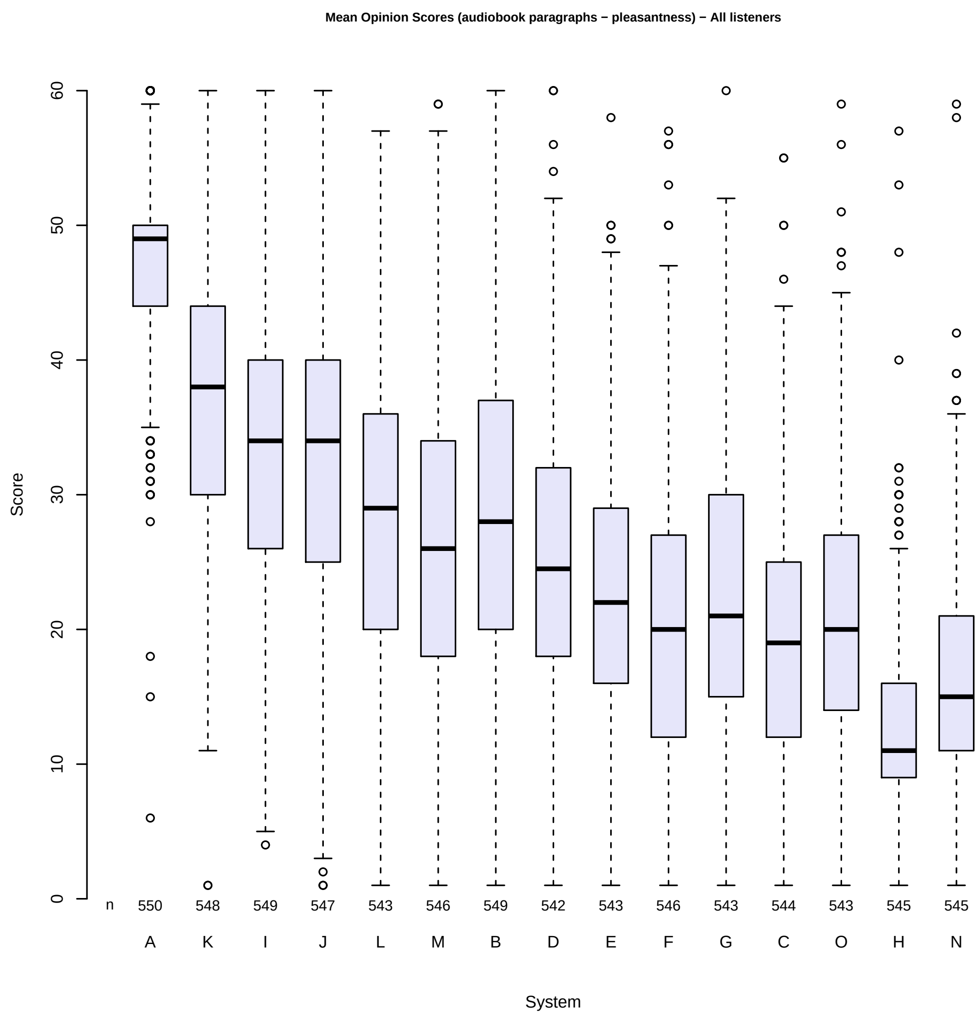

Figure 7: *Pleasantness of paragraphs for task 2018-EH1.*

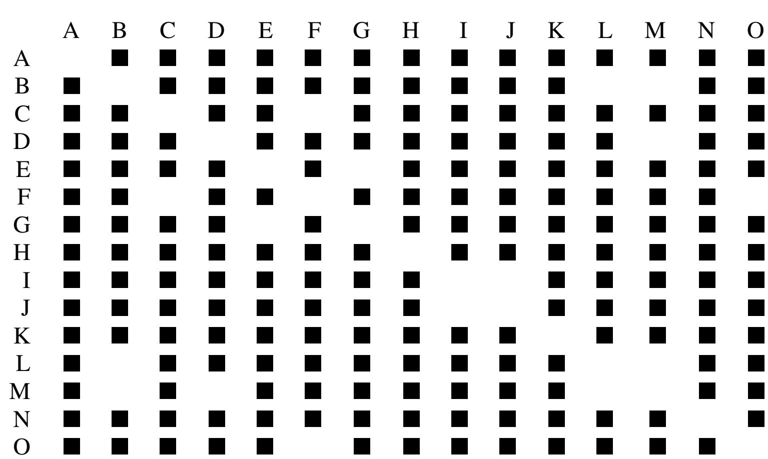

Figure 8: *Significant differences in pleasantness of paragraphs for task 2018-EH1.*

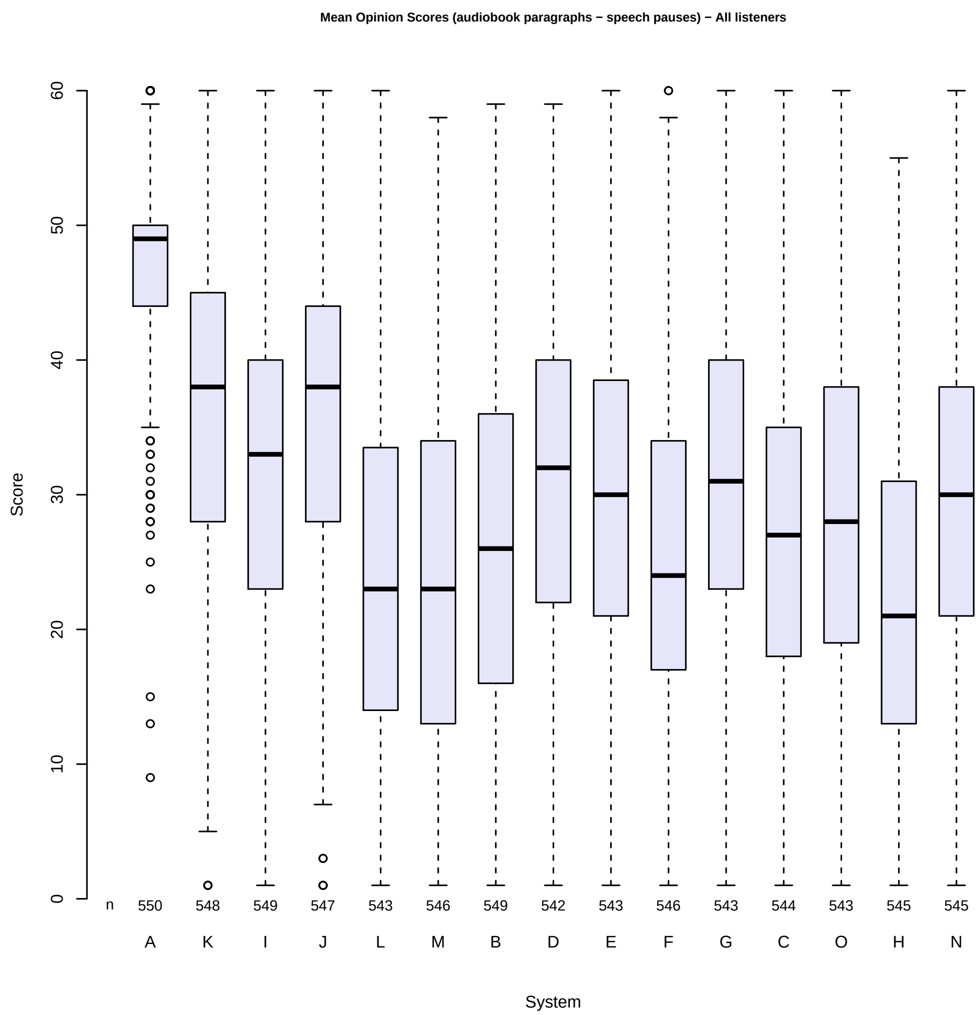

Figure 9: *Speech pauses of paragraphs for task 2018-EH1.*

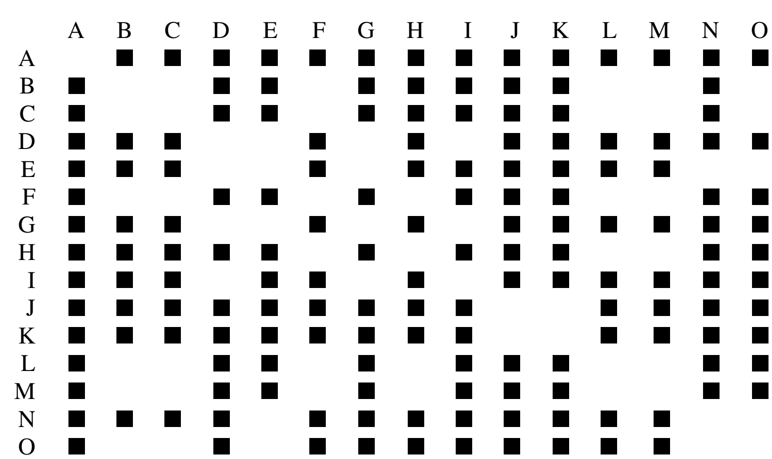

Figure 10: *Significant differences in speech pauses of paragraphs for task 2018-EH1.*

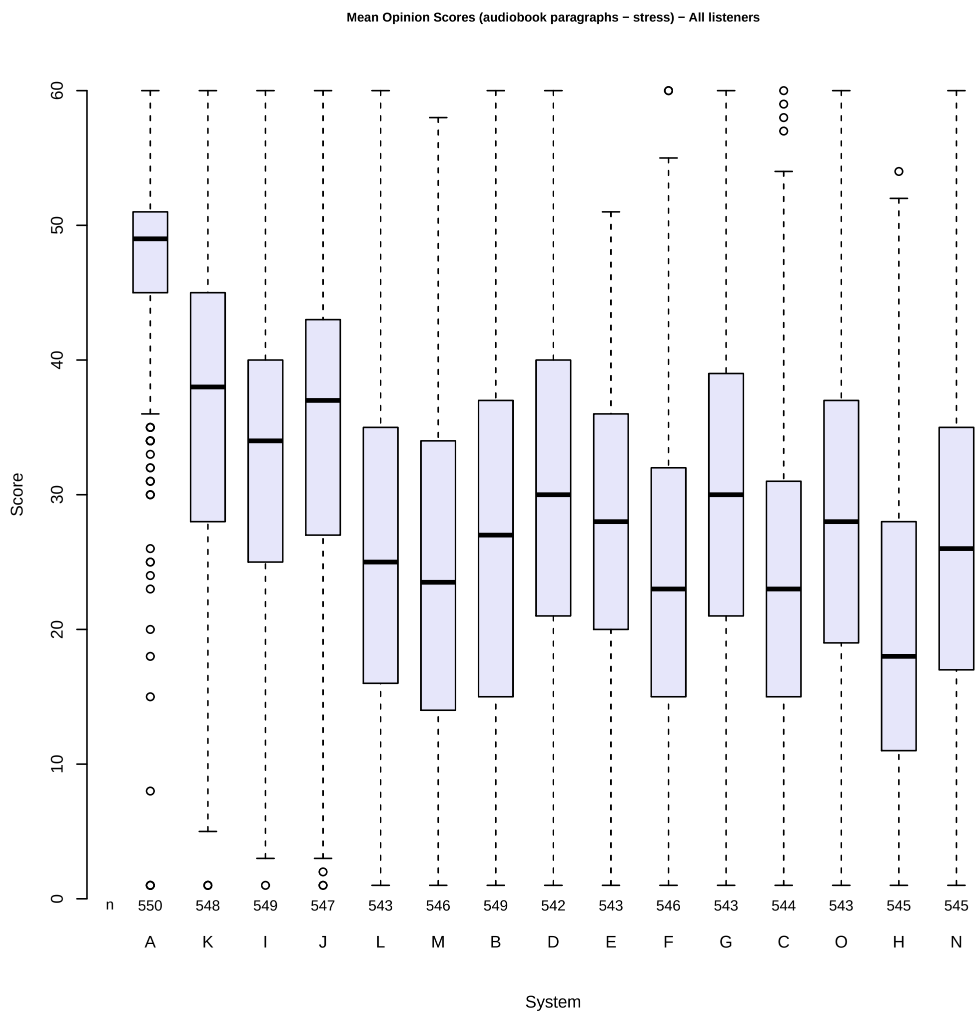

Figure 11: *Stress of paragraphs for task 2018-EH1.*

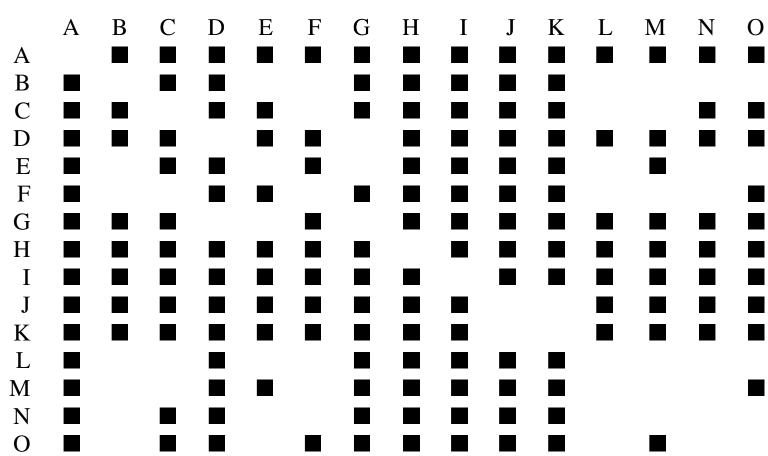

Figure 12: *Significant differences in stress of paragraphs for task 2018-EH1.*

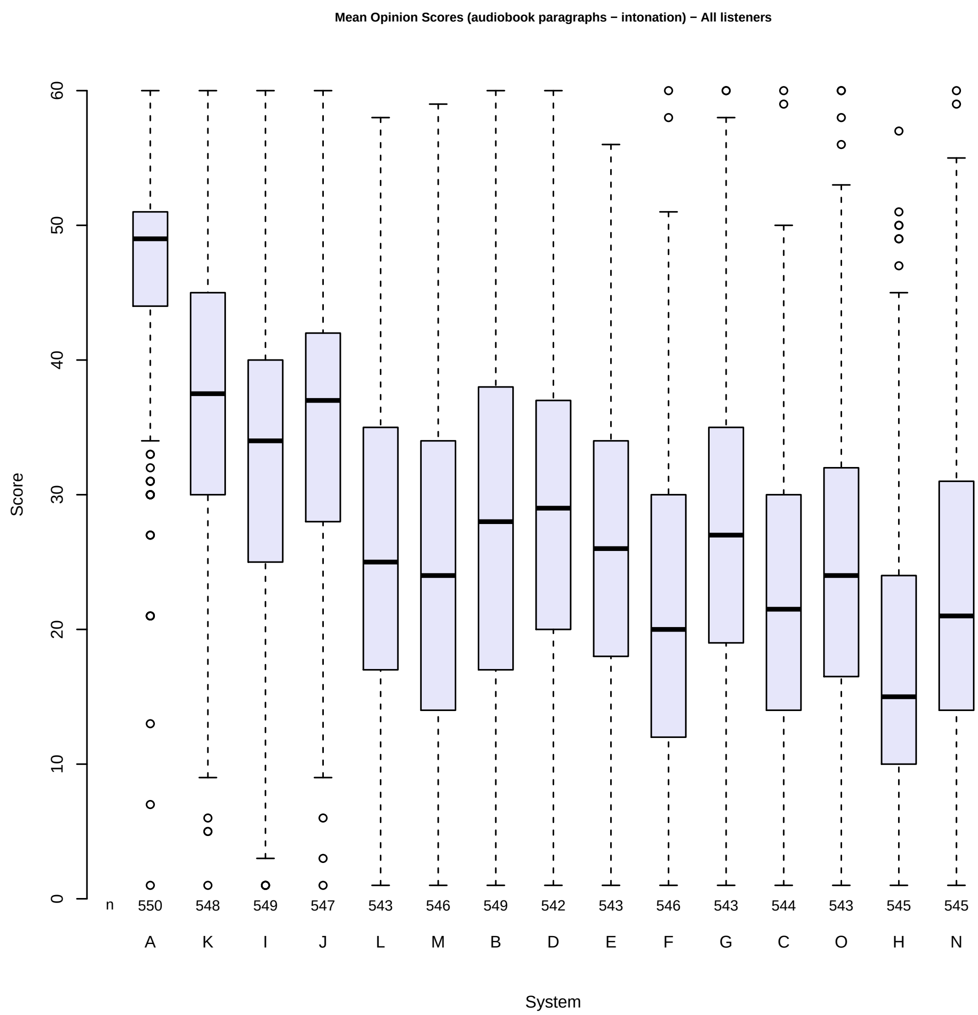

Figure 13: *Intonation of paragraphs for task 2018-EH1.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■			■		■	■	■	■		■	■	
C	■	■		■	■		■	■	■	■	■	■	■		■
D	■		■			■		■	■	■	■		■	■	■
E	■		■			■		■	■	■	■			■	
F	■	■		■	■		■	■	■	■	■	■	■		■
G	■		■			■		■	■	■	■		■	■	■
H	■	■	■	■	■	■	■		■	■	■	■	■	■	■
I	■	■	■	■	■	■	■	■			■	■	■	■	■
J	■	■	■	■	■	■	■	■				■	■	■	■
K	■	■	■	■	■	■	■	■	■			■	■	■	■
L	■		■			■		■	■	■	■			■	
M	■	■	■	■		■	■	■	■	■	■				
N	■	■		■	■		■	■	■	■	■	■			
O	■		■	■		■	■	■	■	■	■				

Figure 14: *Significant differences in intonation of paragraphs for task 2018-EH1.*

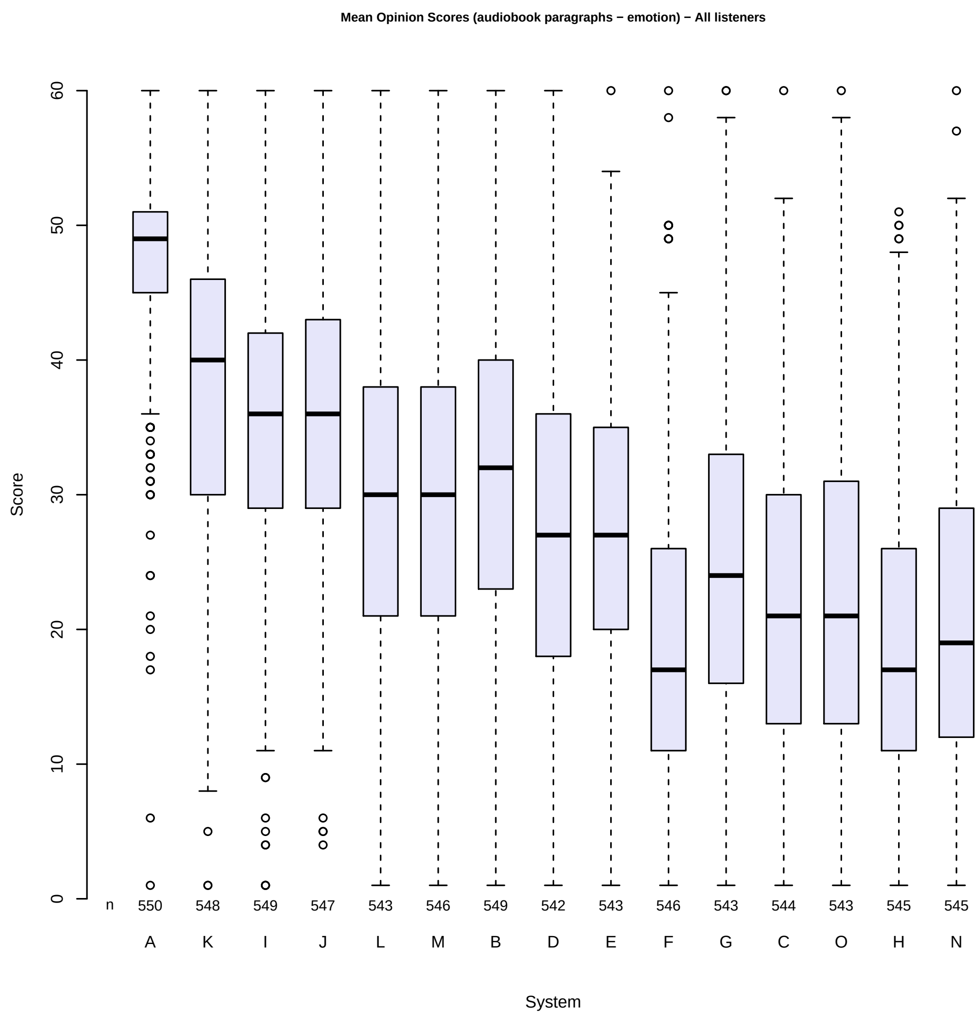

Figure 15: *Emotion of paragraphs for task 2018-EH1.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■	■	■	■	■	■	■	■			■	■
C	■	■		■	■	■	■	■	■	■	■	■	■		
D	■	■	■			■		■	■	■	■		■	■	■
E	■	■	■			■		■	■	■	■	■	■	■	■
F	■	■	■	■	■		■		■	■	■	■	■		■
G	■	■	■			■		■	■	■	■	■	■	■	■
H	■	■	■	■	■		■		■	■	■	■	■	■	■
I	■	■	■	■	■	■	■	■			■	■	■	■	■
J	■	■	■	■	■	■	■	■			■	■	■	■	■
K	■	■	■	■	■	■	■	■	■	■		■	■	■	■
L	■		■		■	■	■	■	■	■	■			■	■
M	■		■	■	■	■	■	■	■	■	■			■	■
N	■	■		■	■		■	■	■	■	■	■	■		■
O	■	■		■	■	■	■	■	■	■	■	■	■	■	

Figure 16: *Significant differences in emotion of paragraphs for task 2018-EH1.*

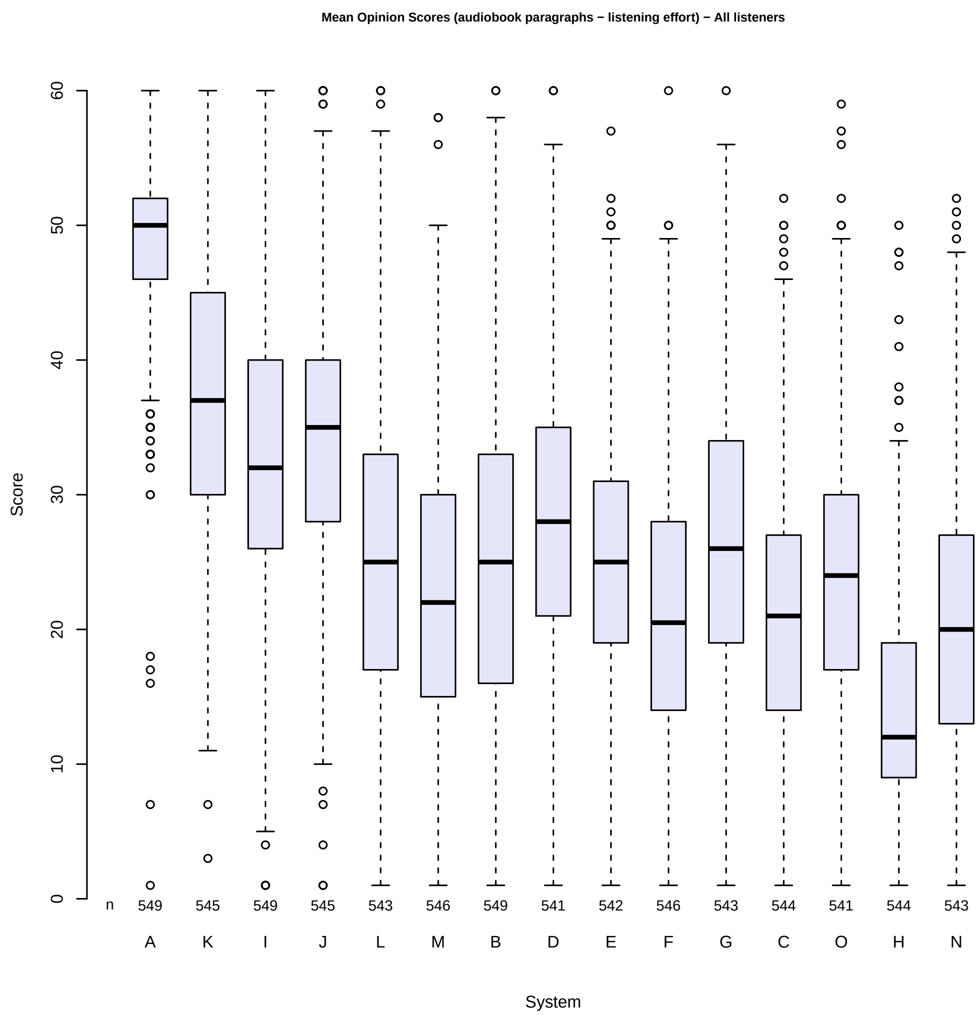

Figure 17: *Listening effort of paragraphs for task 2018-EH1.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A		■	■	■	■	■	■	■	■	■	■	■	■	■	■
B	■		■	■		■		■	■	■	■			■	
C	■	■		■	■		■	■	■	■	■	■	■		■
D	■	■	■		■	■		■	■	■	■	■	■	■	■
E	■		■	■		■		■	■	■	■			■	
F	■	■		■	■		■	■	■	■	■	■	■		■
G	■		■			■		■	■	■	■		■	■	■
H	■	■	■	■	■	■	■		■	■	■	■	■	■	■
I	■	■	■	■	■	■	■	■			■	■	■	■	■
J	■	■	■	■	■	■	■	■			■	■	■	■	■
K	■	■	■	■	■	■	■	■	■	■		■	■	■	■
L	■		■	■		■		■	■	■	■		■	■	
M	■		■	■		■	■	■	■	■	■	■		■	
N	■	■		■	■		■	■	■	■	■	■	■		■
O	■		■	■		■	■	■	■	■	■			■	

Figure 18: *Significant differences in listening effort of paragraphs for task 2018-EH1.*

Language	Total
Bengali	3
Bulgarian	2
Cantonese	1
Chinese (Mandarin)	33
Czech	1
Finnish	1
French	6
German	5
Greek	3
Hebrew	2
Hindi	3
Italian	1
Japanese	26
Persian	1
Portuguese	1
Romanian	2
Spanish	2
Tamil	1
Telugu	2
Thai	3
Urdu	1

Table 3: *First language of non-native speakers.* [4]

Gender	Male	Female
Total	129	131

Table 4: *Gender.* [4]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	18	234	36	10	8	1	0	0

Table 5: *Age of listeners whose results were used (completed the evaluation fully or partially).* [5]

Native speaker	Yes	No
English	155	105

Table 6: *Native speakers.* [4]

	Task EH1
EP	150
ES	89
ER	67
ALL	306

Table 7: *Listener types, showing the number of listeners whose responses were used in the results for similarity and naturalness. (We have counted in listeners who did some of the test, but have not necessarily completed it; therefore, numbers may be slightly different for intelligibility.)* [5]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EP	150	0	0	150
ES	111	21	33	57
ER	82	13	27	42
ALL	**343**	**34**	**60**	**249**

Table 8: *Listener registration and evaluation completion rates.* [5]

	EH1_01	EH1_02	EH1_03	EH1_04	EH1_05	EH1_06	EH1_07	EH1_08	EH1_09	EH1_10	EH1_11	EH1_12	EH1_13	EH1_14	EH1_15
EP	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
ES	5	8	7	6	7	7	7	5	5	5	6	6	5	6	4
ER	5	6	4	4	5	5	6	6	3	3	5	2	5	4	5
ALL	20	24	21	20	22	22	23	21	18	18	21	18	20	20	19

Table 9: *Listener groups - showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations.* [5]

Listener Type	EP	ES	ER	ALL
Total	150	65	46	261

Table 10: *Listener type totals for submitted feedback.*

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	33	37	109	62	20	0

Table 11: *Highest level of education completed.* [4]

CS/Engineering person?	Yes	No
Total	109	150

Table 12: *Computer science / engineering person.* [4]

Work in speech technology?	Yes	No
Total	87	173

Table 13: *Work in the field of speech technology.* [4]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	27	47	27	66	61	13	18

Table 14: *How often normally listened to speech synthesis before doing the evaluation.* [4]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	1	10	102	19	13	116

Table 15: *Dialect of English of native speakers.* [4]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
Total	19	43	29	14	0

Table 16: *Level of English of non-native speakers.* [4]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	242	11	6	2

Table 17: *Speaker type used to listen to the speech samples.* [4]

Same environment?	Yes	No
Total	248	10

Table 18: *Same environment for all samples?* [4]

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	200	45	11	0	2

Table 19: *Kind of environment when listening to the speech samples.* [4]

Number of sessions	1	2-3	4 or more
Total	195	49	15

Table 20: *Number of separate listening sessions to complete all the sections.* [4]

Browser	Firefox	IE	Chrome	Opera	Safari	Mozilla	Other
Total	31	8	72	0	139	0	6

Table 21: *Web browser used (the paid listeners - type EP - did the test on either Safari or Chrome).* [4]

Similarity with reference samples	Easy	Difficult
Total	199	50

Table 22: *Listeners' impression of their task in the section(s) about similarity with original voice.* [4]

Problem	Scale too big, too small, or confusing	Issues with hardware	Other
Total	29	12	21

Table 23: *Listeners' problems in the section(s) about similarity with original voice.* [4]

Number of times	1-2	3-5	6 or more
Total	209	39	1

Table 24: *Number of times listened to each example in the section(s) about similarity with original voice.* [4]

Naturalness	Easy	Difficult
Total	228	25

Table 25: *Listeners' impression of their task in the MOS naturalness sections.* [4]

Problem	Difficulties with judging naturalness	Scale too big, too small, or confusing	Issues with hardware	Other
Total	7	16	6	11

Table 26: *Listeners' problems in the MOS naturalness sections.* [4]

Number of times	1-2	3-5	6 or more
Total	215	23	1

Table 27: *Number of times listened to each example in the MOS naturalness sections.* [4]

Book passage	Easy	Difficult
Total	154	106

Table 28: *Listeners' impression of their task in the sections involving book passages.* [4]

Problem	Scale too big, too small, or confusing	Quality of samples too bad	Bad speakers, playing files disturbed other connection too slow, etc	Other
Total	64	23	6	26

Table 29: *Listeners' problems in the sections involving book passages.* [4]

Number of times	1-2	3-5	6 or more
Total	224	24	1

Table 30: *How many times listened to each example in the sections involving book passages.* [4]

SUS section(s)	Usually understood all the words	Usually understood most of the words	Very hard to understand the words	Typing problems: words too hard to spell, or too fast to type
Total	19	124	100	15

Table 31: *Listeners' impressions of the intelligibility task (SUS).* [4]