The Blizzard Challenge 2016

Simon King[a] *and Vasilis Karaiskos*[b]

[a]Centre for Speech Technology Research, [b]School of Informatics,
University of Edinburgh

Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2016 was the twelfth annual Blizzard Challenge and was once again organised by Simon King at the University of Edinburgh, with advice from the other members of the Blizzard Challenge committee – Keiichi Tokuda, Alan Black, and Kishore Prahallad. For the task this year, a medium-sized single-speaker English corpus was used, comprising around 5 hours of audio from professionally-produced children's audiobooks.

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

1. Introduction

The Blizzard Challenge, conceived by Black and Tokuda in 2005 [1], is a regular event and the many previous summery papers report previous findings. This paper concerns itself only with the 2016 challenge. A summary-of-summaries (up to 2014), which attempt to find trends across the first decade of the challenge, is available [2].

Many useful resources, such as releases of the submitted speech, reference samples, listening test responses, scripts for running the listening test and scripts for the statistical analysis, can all be found via the Blizzard Challenge website [3].

2. Participants

This years challenge, Blizzard 2016, had 13 participants, which are listed in Table 1, along with 4 benchmarks.

Three benchmark systems were included this year. The unit selection and HMM-based are he same types as in many previous challenges and will aid comparisons with those previous years. The new DNN benchmark will allow comparisons for future years. The unit selection benchmark is Festival[1] from CSTR and was configured very similarly to the Festival/CSTR entry to Blizzard 2006 [4]. This system can be replicated by following the Multisyn recipe available from `http://www.cstr.ed.ac.uk/downloads/festival/multisyn_build`. 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 and the STRAIGHT vocoder. The third benchmark[3] uses the Merlin toolkit, which is available from `https://github.com/CSTR-Edinburgh/merlin` in conjunction with the Ossian front end and the WORLD vocoder.

When reporting results, the systems are identified using letters, with A denoting natural speech, B the Festival benchmark systems, C the HTS benchmark system, D the Merlin benchmark system and the remaining letters denoting the systems submitted by participants in the challenge. The system identifiers are assigned randomly and are not the same across different years of the challenge.

[1]Thanks to Srikanth Ronanki, Oliver Watts and Tom Merritt.

[2]Thanks to Keiichi Tokuda and his team.

[3]Thanks to Oliver Watts.

Short name	Details	Method
NATURAL	Natural speech from the same speaker as the corpus	human
FESTIVAL_BM	Festival benchmark	unit selection
HTS_BM	HTS benchmark	HMM
DNN_BM	Merlin benchmark	DNN
ADAPT	Trinity College Dublin & Dublin City U	HMM
CSTR	Centre for Speech Technology Research, U Edinburgh	DNN hybrid
I2R-NWPU-NTU	Institute for Infocomm Research & Northwestern Polytechnical U & Nanyang Technological U	DNN hybrid
IIITH	International Institute of Information Technology	unit selection
INNOETICS	Innoetics & Institute for Language & Speech Processing	hybrid
IRISA	U Rennes	unit selection
MARYTTS	Deutsche Forschungszentrum für Künstliche Intelligenz	unit selection
MERAKA	Meraka Institute, CSIR	HMM
NII	Aalto U & National Institute of Informatics & Sokendai U & Naver Corporation & CSTR	DNN
NITECH	Nagoya Institute of Technology	DNN
NLPR	National Laboratory of Pattern Recognition	DNN hybrid
USTC	U Science and Technology of China & iFLYTEK	DNN hybrid
UTokyo	U Tokyo	DNN

Table 1: The participating systems and their short names. The first four rows are the benchmarks and correspond to the system identifiers A, B, C and D 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 statistical parametric based on Hidden Markov Models (HMM) or Deep Neural Networks (DNN), unit selection with waveform concatenation, or hybrid (in all cases this is DNN-guided unit selection).

3. Voice to be built

3.1. Speech database

The data was provided by Usborne Publishing Ltd (`http://www.usborne.com`) and is from their commercial product range of children's audiobooks. The British English speaker, Lesley Sims, is female. Around 5 hours of material was made available to participants in the challenge. A 2 hour subset of this material was released one year earlier, for use in pilot experiments. Each of the approximately 50 books in the 5 hour set is rated by Usborne for reading age (mainly 4,5 or 6 years, with a handful of books rated as "18 months+"). Genres include classic children's stories (e.g., The Three Little Pigs), simplified & abridged versions of Shakespeare (e.g., Romeo and Juliet), and factual books (e.g., Knights and Castles). A feature of almost all the fiction titles is the high proportion of quoted speech, and number of proper names. In general, the speaker reads in an expressive and engaging style, but without highly-dramatic 'acting' or 'character voices'.

As in all Blizzard Challenges, the organisers held out some of the material for use as a test set. This material was a few complete audiobooks across a range of genres and reading ages.

3.2. Tasks

Participants were invited to take part in the a single tasks, in accordance with the rules of the challenge, published on the website: build a voice from the provided data, suitable for reading children's audiobooks. This was denoted as task 2016-EH1, following the standard Blizzard Challenge task naming scheme.

3.3. Listening test design and materials

Participants were asked to synthesise many hundreds of test sentences, of which only a small subset were used in the listening test. This provides a large amount of material that might be used in future listening tests, and also prevents participants from manually intervening in synthesis.

For a description of the listening test design and the web interface used to deliver it, please refer to previous summary papers. Permission was been obtained from participants to distribute parts of this dataset along with the listener scores and this can be downloaded via the Blizzard website.

3.4. 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. The following listener types[4] were used:

- 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 (EP). All listeners of this type completed the entire listening test.
- Speech experts, recruited via participating teams and mailing lists (EE).
- Volunteers recruited via participating teams, mailing lists, blogs, word of mouth, etc. (ER).

Table 7 gives a breakdown of evaluation completion rates for each listener type.

[4]Experimenter error means that the letter identifiers do not correspond to those in previous years.

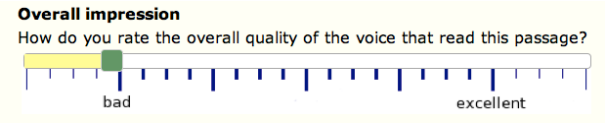

Figure 1: Example of a slider used to obtain listener responses in the paragraph sections.

3.5. Listening tests

The listening test had the following structure, comprising 7 sections each with 17 stimuli (or 16 in the case of intelligibility, since no natural recorded SUS were available this year):

1. Multiple dimensions, book paragraphs
2. Multiple dimensions, book paragraphs
3. Naturalness, book sentences
4. Naturalness, book sentences
5. Similarity, book sentences
6. Intelligibility, SUS, single listen only
7. Intelligibility, SUS, single listen only

Within each section of the listening test, a listener heard one example from each system, including natural speech where available. As always, a Latin Square design was employed to ensure that no listener heard the same sentence or paragraph more than once across the entire test, something that is particularly important for testing intelligibility.

The "Multiple dimensions" evaluation of paragraphs was that proposed in [5], and which has been used in previous challenges. For each presented spoken paragraph (hand selected to generally be no more than 30 seconds in duration), listeners were asked to provide ratings using sliders, as illustrated in Figure 1, along these dimensions:

- Overall impression ("bad" to "excellent")
- Pleasantness ("very unpleasant" to "very pleasant")
- Speech pauses ("speech pauses confusing/unpleasant" to "speech pauses appropriate/pleasant")
- Stress ("stress unnatural/confusing" to "stress natural")
- Intonation ("melody did not fit the sentence type" to "melody fitted the sentence type")
- Emotion ("no expression of emotions" to "authentic expression of emotions")
- Listening effort ("very exhausting" to "very easy")

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. 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 [6] 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. Finally, Section 5.1 and Tables 2 to 30 provide a summary of the responses to a questionnaire that listeners were asked to complete at the end of the listening test.

5. Results

Standard boxplots are presented for the ordinal data where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. Bar charts are presented for the word error rate type interval data. 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 2 shows the results for sentences and indicates the type of each system using colour coding. It can be seen that those systems that generate the waveform using concatenation (unit selection or hybrid) are – as in previous challenges – generally more natural-sounding than the systems that employ a vocoder. A striking result this year is the relatively high naturalness of the benchmark Festival unit-selection system.

No synthesiser is as natural as the natural speech (Figure 2). System L is significantly (Figure 3) more natural than all other systems except M. Systems L, M and Q form a group, and are significantly more natural than all other systems.

For intelligibility, no comparisons with natural speech possible this year; systems L, B, F, D, G are all equally (Figure 5) intelligible.

The multiple dimensions of scoring for the paragraphs are reported in Figures 6 to 18. Unsurprisingly, no system was judged to be as good as natural speech, along any dimension. System L is better than all other systems along most dimensions, except that is is not better than M in terms of stress, intonation or emotion. System M is in turn generally better than all the remaining systems. Significance tables can be found in the full results package.

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 2 to 30.

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. Rob Clark designed and implemented the 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. Toshiba Research Europe Ltd, Cambridge Research Laboratory prepared some of the data. Amazon, Apple, and Google provided support. 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, 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] 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, Sept. 2006.

[5] Florian Hinterleitner, Georgina Neitzel, Sebastian Moeller, and Christoph Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proc. Blizzard Workshop*, 2011.

[6] 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)*, August 2007.

In the tables on the following pages, the footnotes in the captions specify whether the numbers in that table are based on listener feedback [5] or on the listening test results themselves. [6]

[5] 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.

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

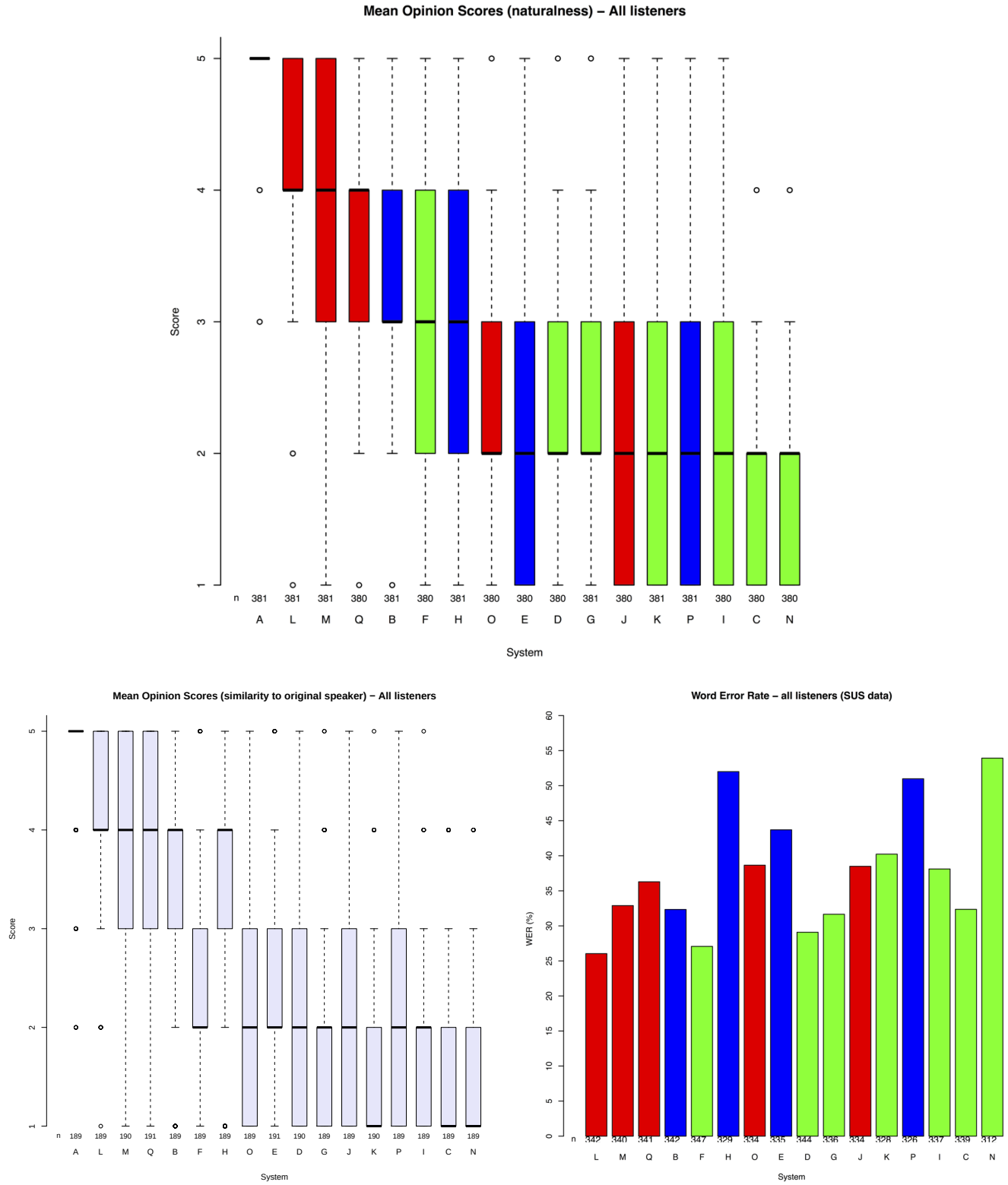

Figure 2: Results for task 2016-EH1 on sentence test material, pooling all listeners' responses. The plots for naturalness and intelligibility are colour-coded: green for statistical parametric systems that employ some form of vocoder to generate the waveform, blue for unit selection systems and red for hybrid systems that concatenate waveforms guided by a DNN. Intelligibility results are not available for A (natural speech). System B is the Festival unit selection benchmark, C is the HMM statistical parametric benchmark and D is the DNN statistical parametric benchmark.

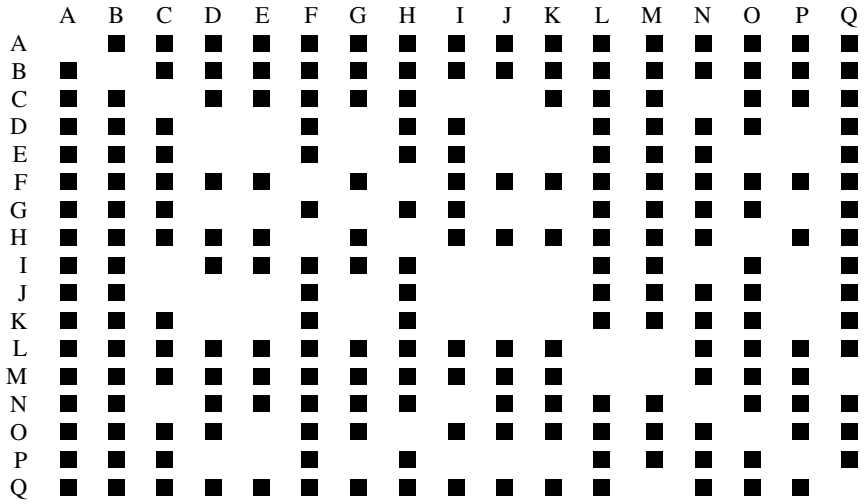

Figure 3: Significant differences in naturalness (book sentences) between systems are indicated by a solid black box. Refer to [4] for details of significance testing.

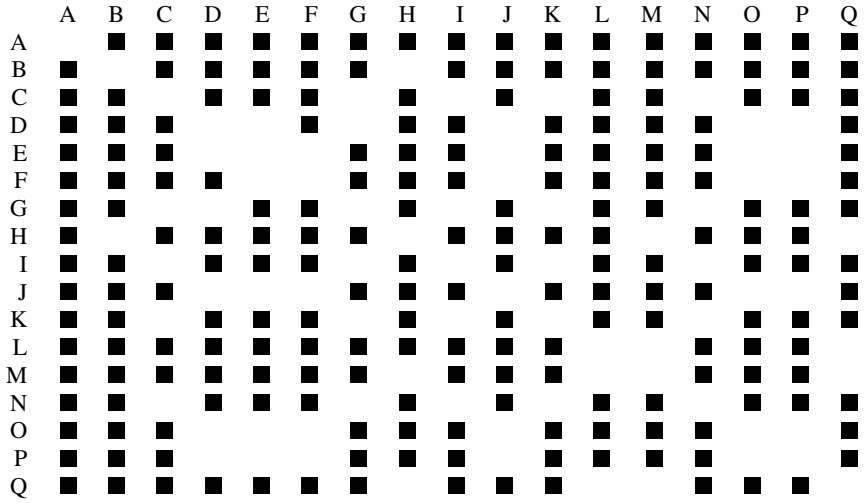

Figure 4: Significant differences in speaker similarity (book sentences) between systems are indicated by a solid black box. Refer to [4] for details of significance testing.

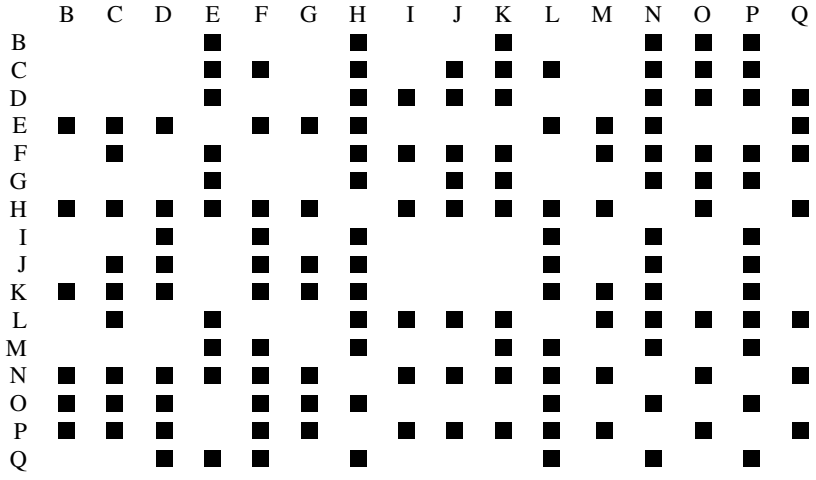

Figure 5: Significant differences in intelligibility (SUS) between systems are indicated by a solid black box. Refer to [4] for details of significance testing.

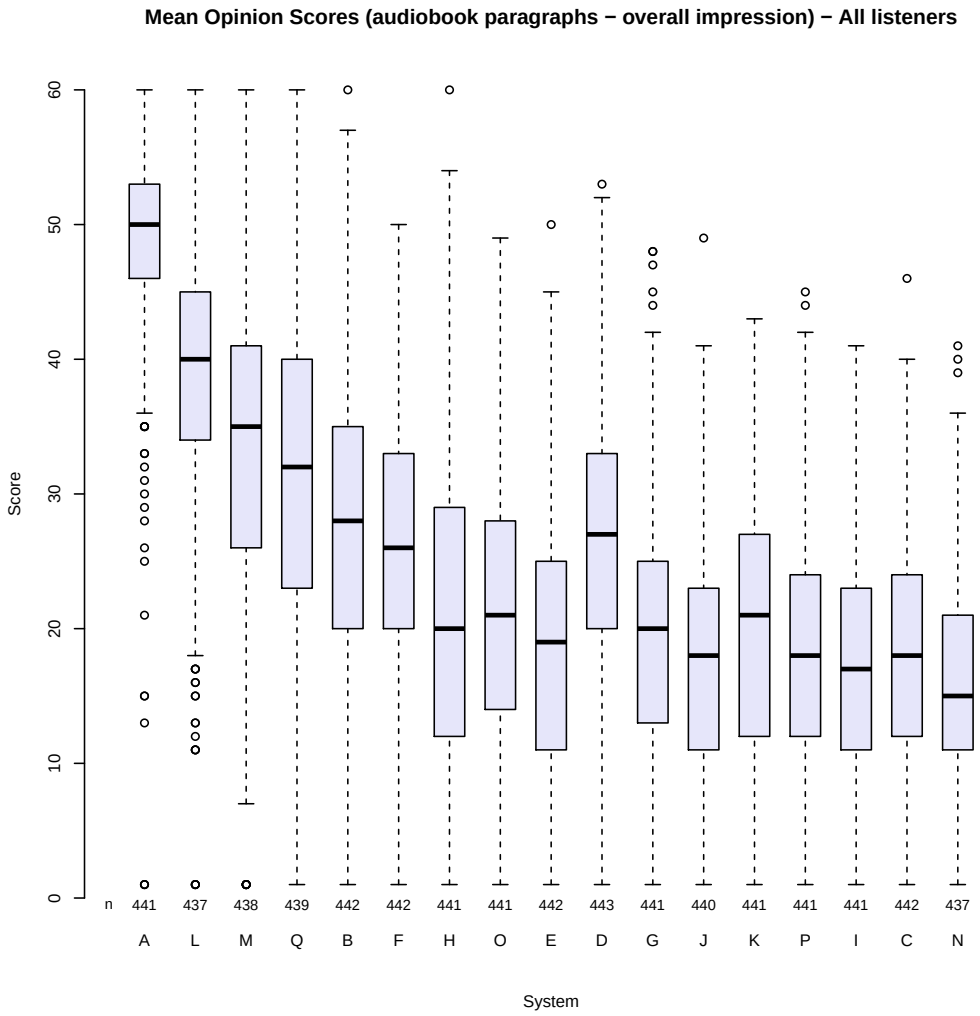

Figure 6: Overall impression of paragraphs for task 2016-EH1.

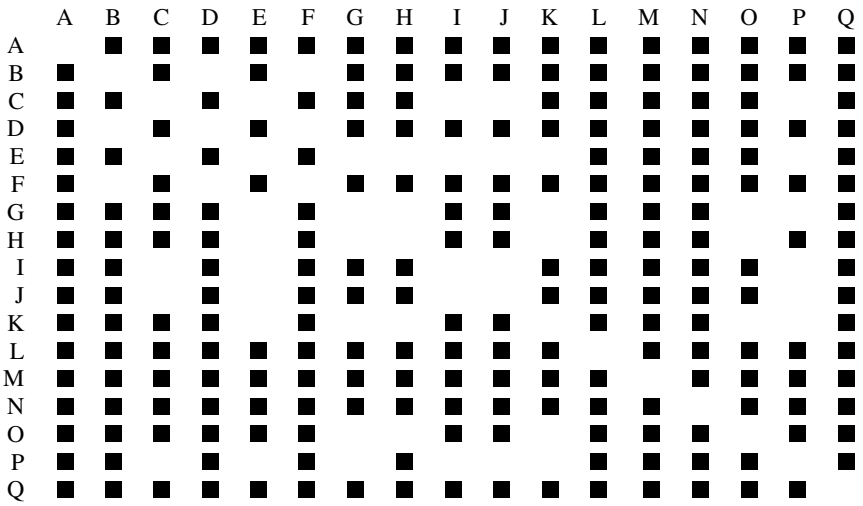

Figure 7: Significant differences in overall impression of paragraphs for task 2016-EH1.

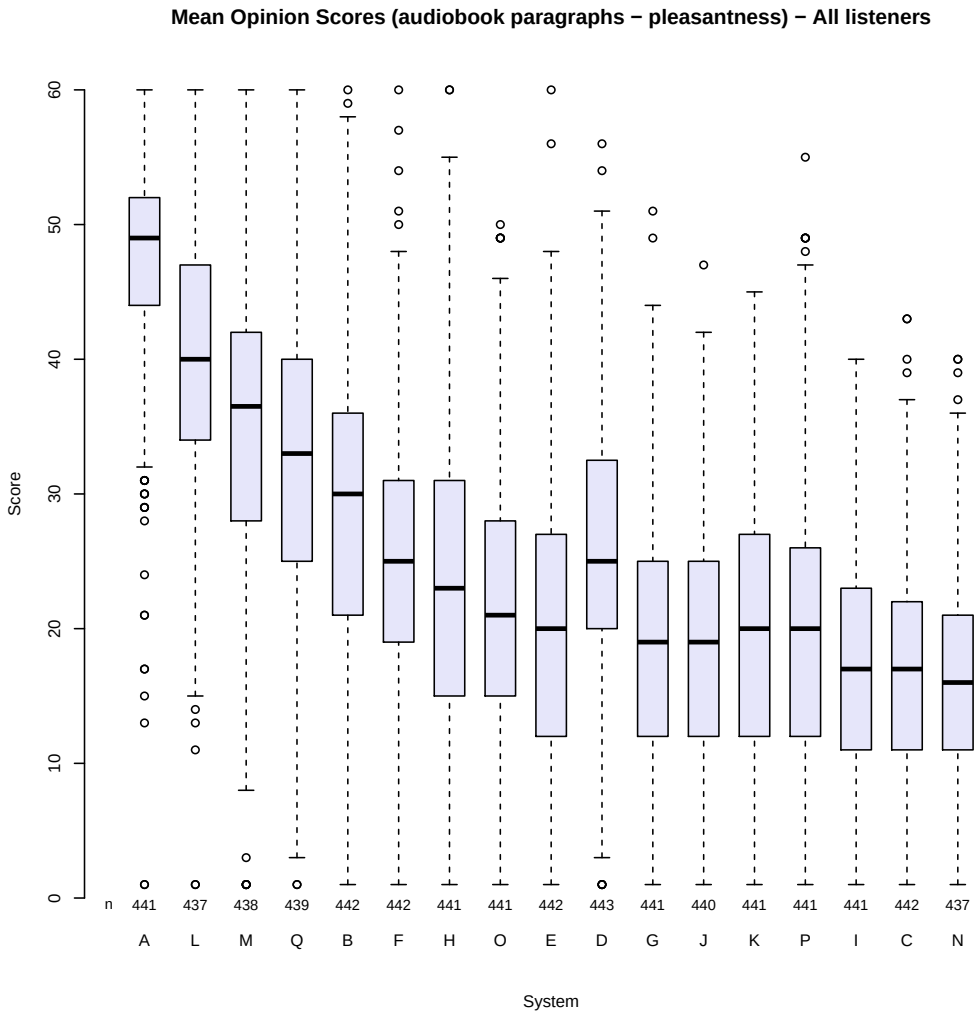

Figure 8: Pleasantness of paragraphs for task 2016-EH1.

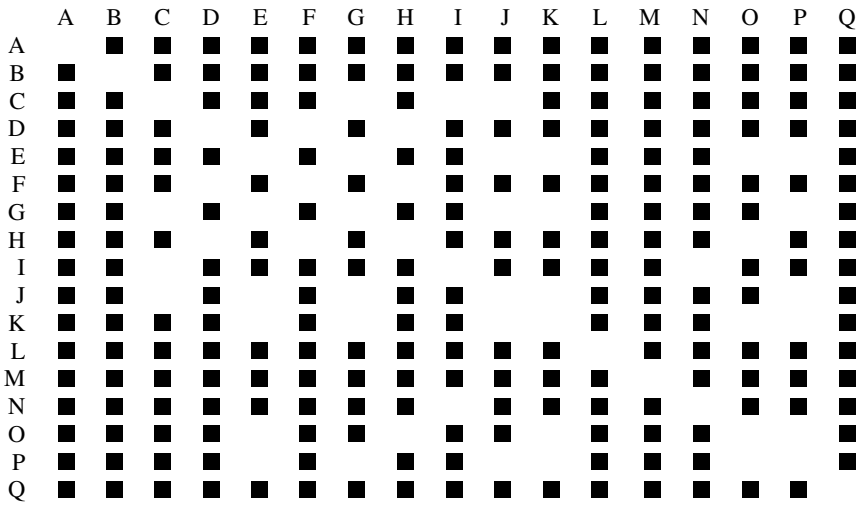

Figure 9: Significant differences in pleasantness of paragraphs for task 2016-EH1.

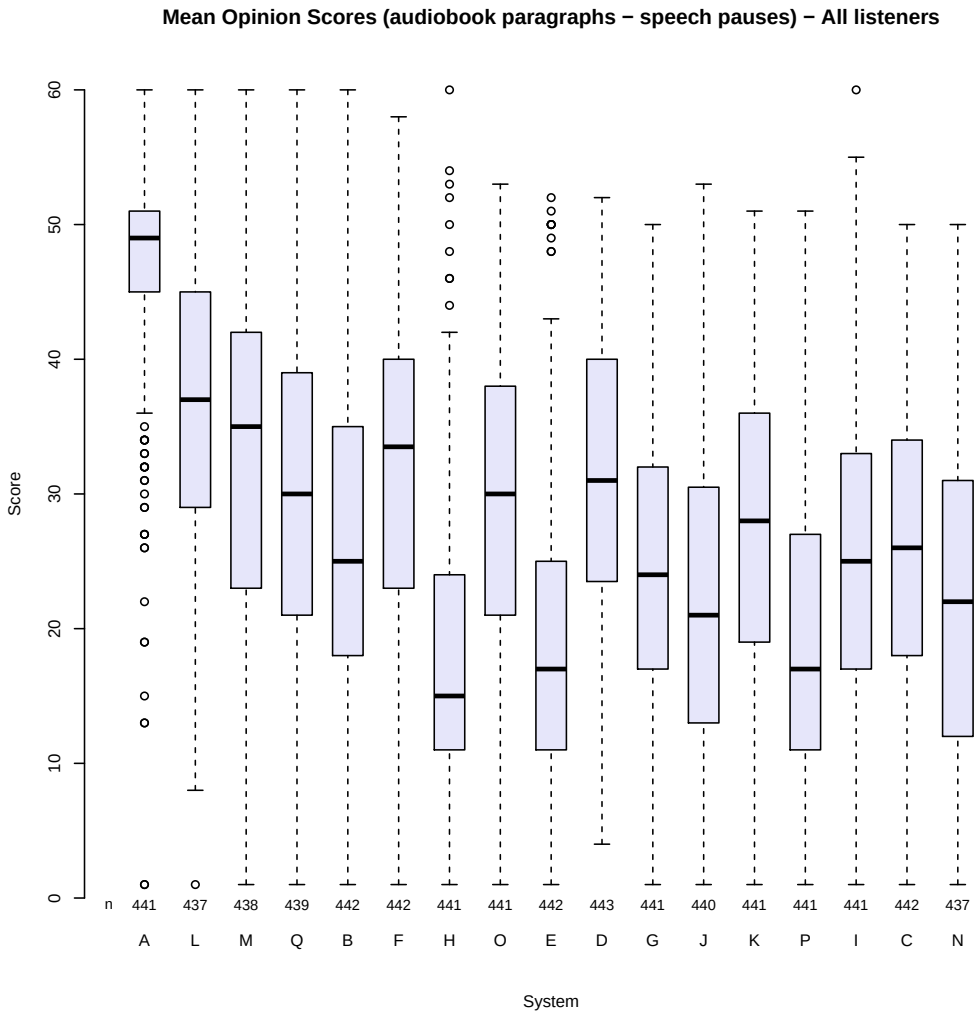

Figure 10: Speech pauses of paragraphs for task 2016-EH1.

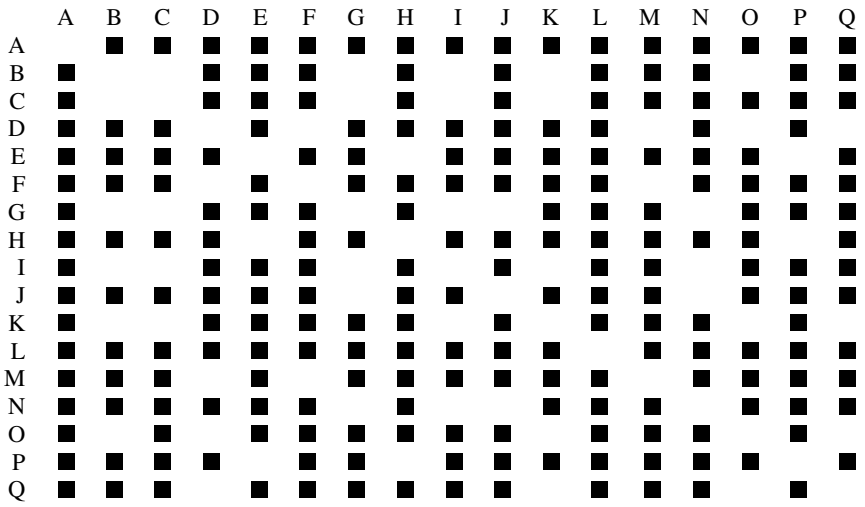

Figure 11: Significant differences in speech pauses of paragraphs for task 2016-EH1.

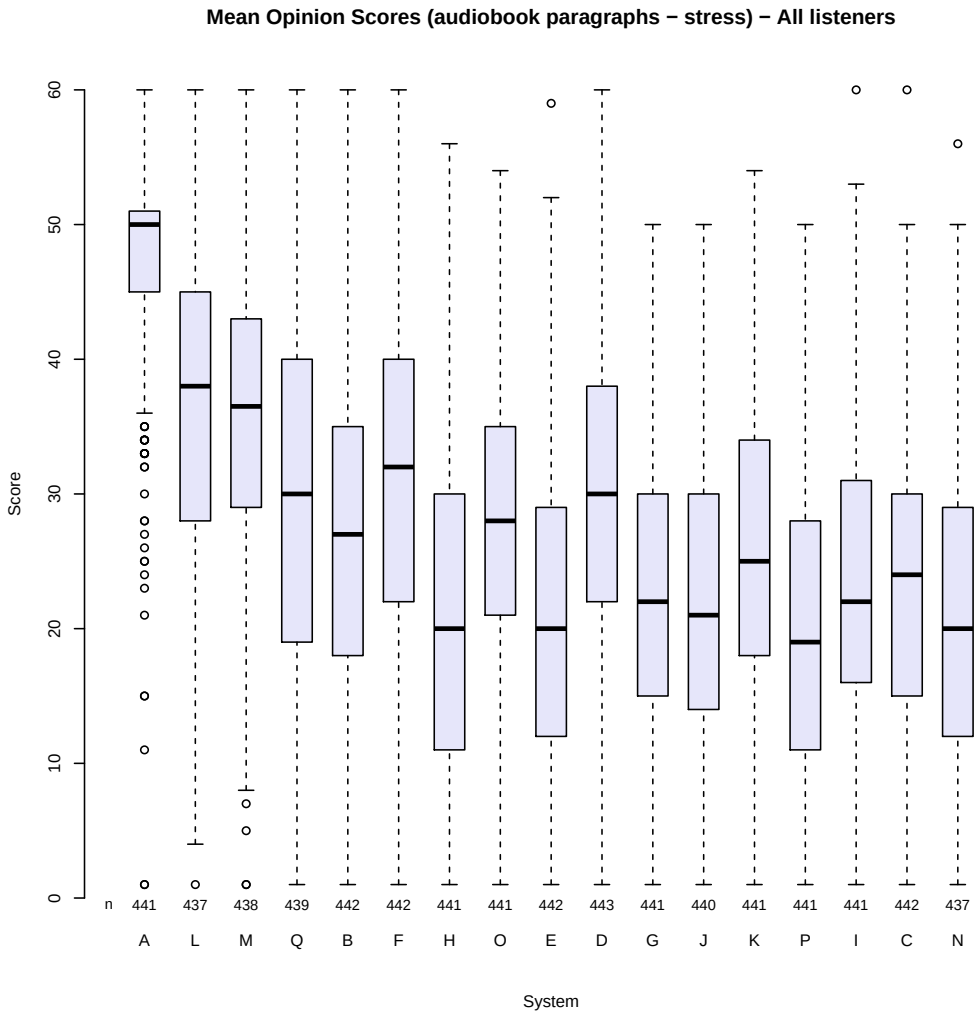

Figure 12: Stress of paragraphs for task 2016-EH1.

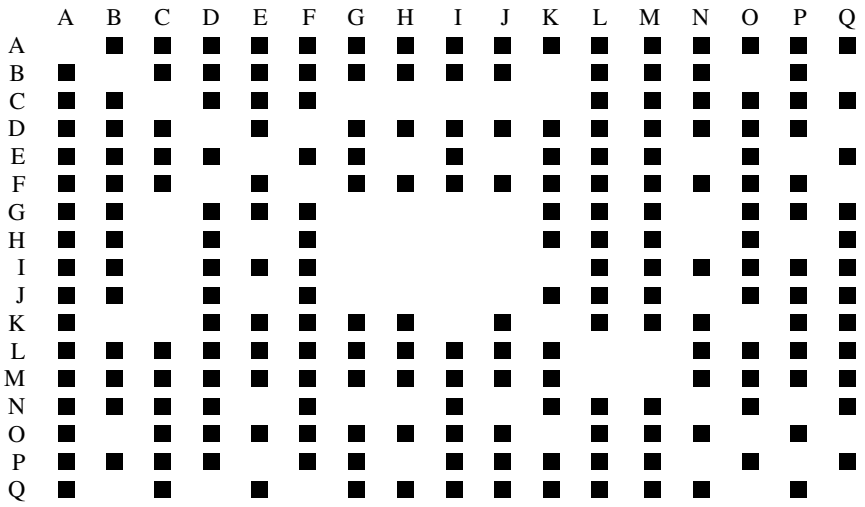

Figure 13: Significant differences in stress of paragraphs for task 2016-EH1.

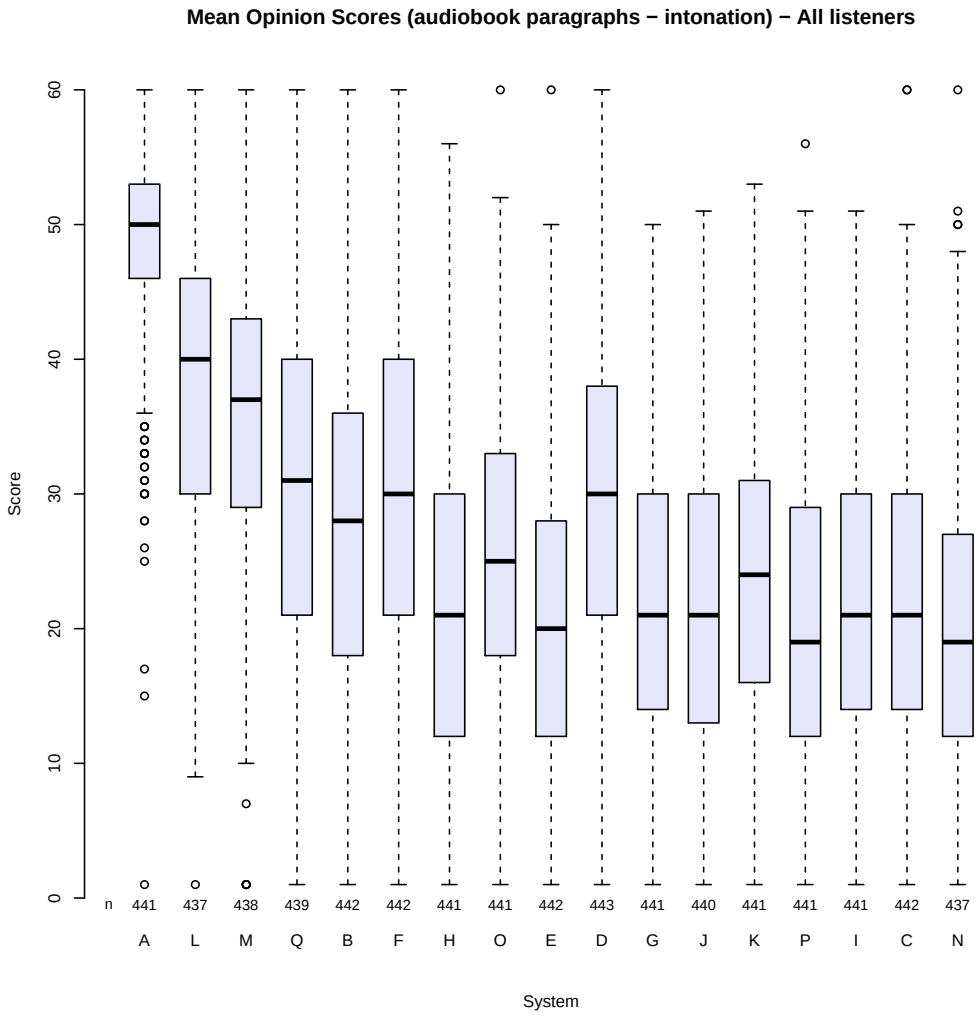

Figure 14: Intonation of paragraphs for task 2016-EH1.

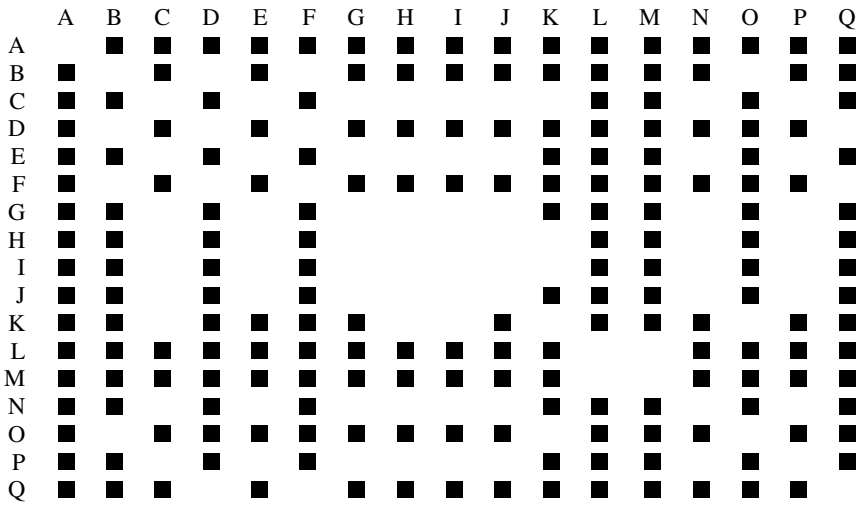

Figure 15: Significant differences in intonation of paragraphs for task 2016-EH1.

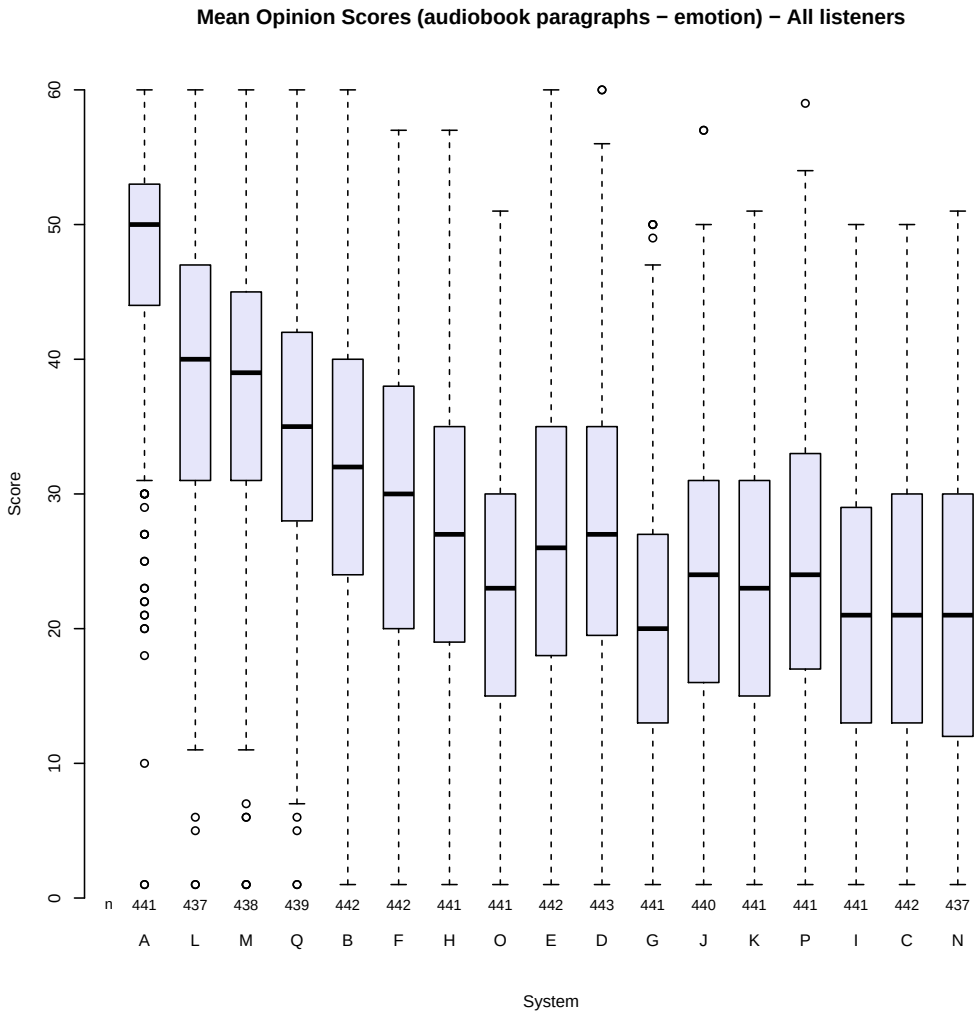

Figure 16: Emotion of paragraphs for task 2016-EH1.

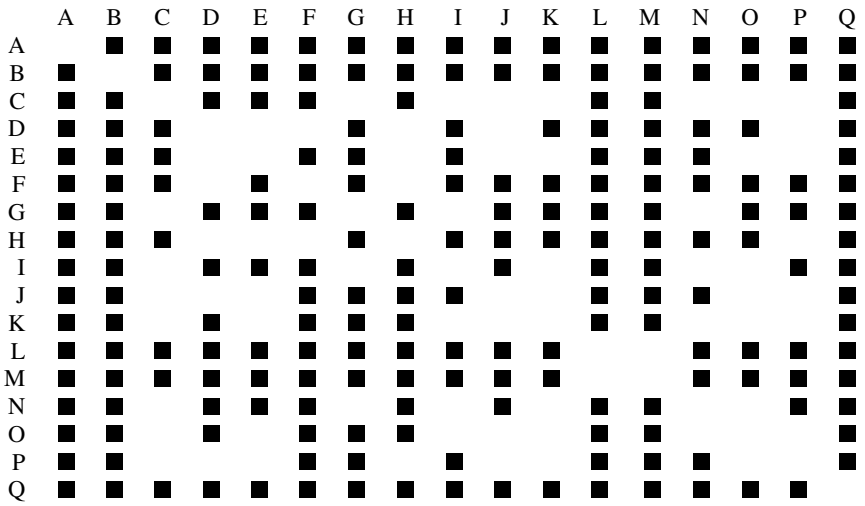

Figure 17: Significant differences in emotion of paragraphs for task 2016-EH1.

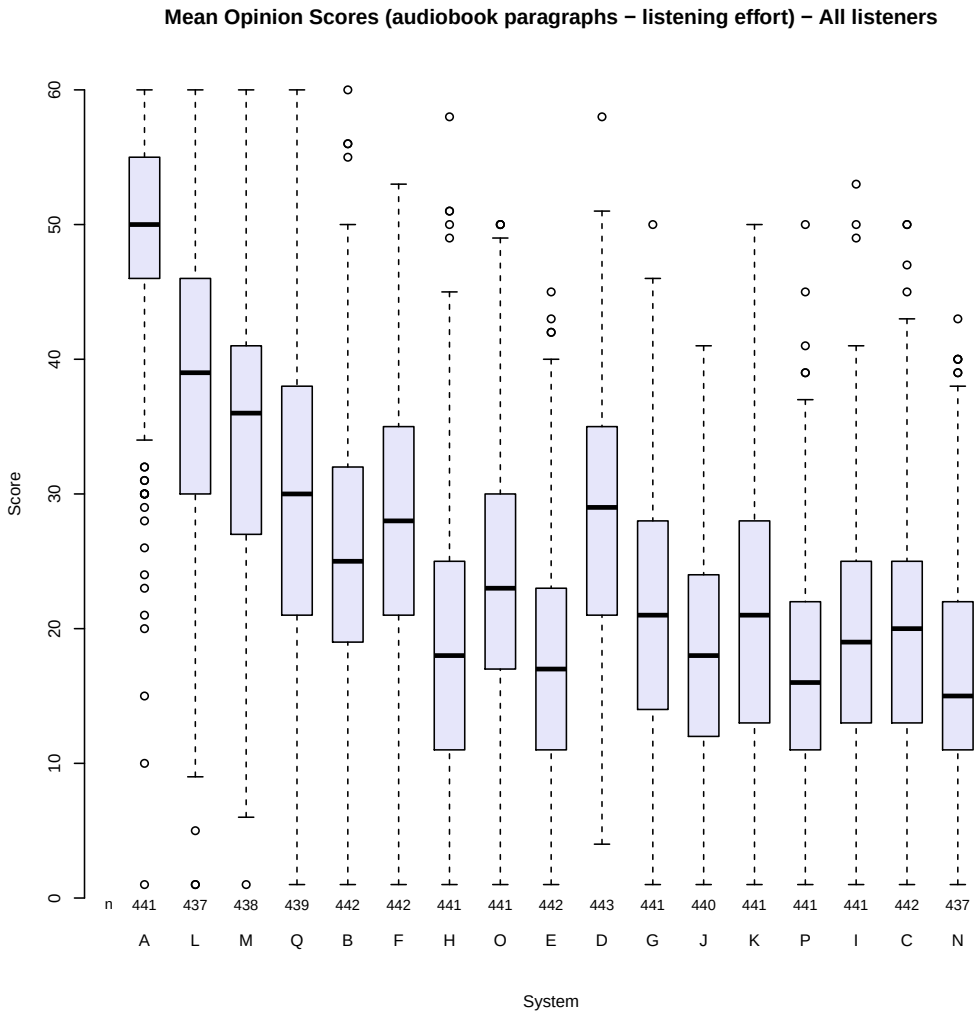

Figure 18: Listening effort of paragraphs for task 2016-EH1.

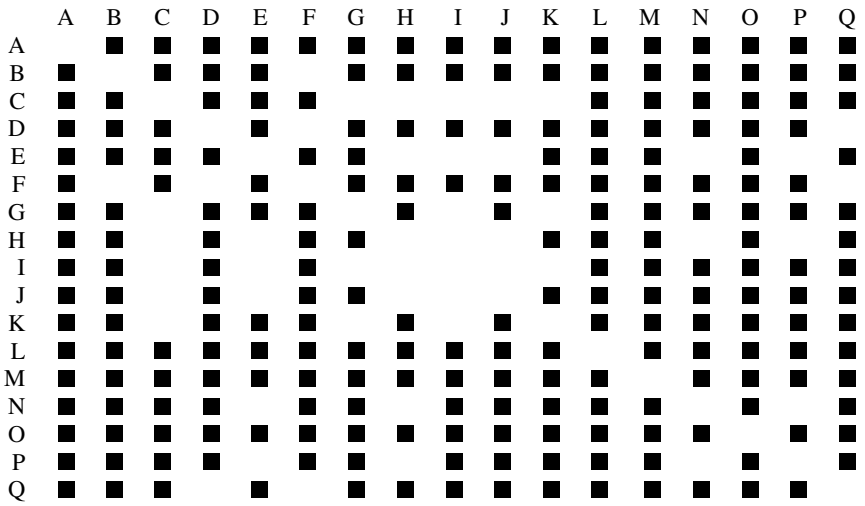

Figure 19: Significant differences in listening effort of paragraphs for task 2016-EH1.

Language	Total
Catalan	1
Chinese (Mandarin)	12
French	7
German	2
Greek	4
Hebrew	2
Hindi	1
Italian	1
Japanese	30
Korean	1
Portuguese	2
Russian	1
Telugu	3
Urdu	1

Table 2: First language of non-native speakers [5]

Gender	Male	Female
Total	93	89

Table 3: Gender [5]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	18	161	53	24	17	8	1	0

Table 4: Age of listeners whose results were used (completed the evaluation fully or partially) [6]

Native speaker	Yes	No
English	112	69

Table 5: Native speakers [5]

	Task EH1
EP	104
ER	118
EE	64
ALL	286

Table 6: 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) [6]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EE	104	0	0	104
ER	496	381	69	46
EE	72	9	29	34
ALL	**672**	**390**	**98**	**184**

Table 7: Listener registration and evaluation completion rates. [6]

	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	EH1_16	EH1_17
EP	7	7	7	6	6	6	6	6	6	6	6	6	6	6	5	6	6
ER	6	7	8	8	5	9	10	6	5	4	7	6	5	7	9	3	10
EE	5	4	5	5	3	4	4	2	4	4	3	3	4	3	4	3	3
ALL	18	18	20	19	14	19	20	14	15	14	16	15	15	16	18	12	19

Table 8: Listener groups - showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations [6]

Listener Type	EP	ER	EE	ALL
Total	104	45	34	183

Table 9: Listener type totals for submitted feedback

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	30	25	63	36	28	0

Table 10: Highest level of education completed [5]

CS/Engineering person?	Yes	No
Total	80	99

Table 11: Computer science / engineering person [5]

Work in speech technology?	Yes	No
Total	60	121

Table 12: Work in the field of speech technology [5]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	18	45	35	40	29	8	7

Table 13: How often normally listened to speech synthesis before doing the evaluation [5]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	0	3	65	47	15	76

Table 14: Dialect of English of native speakers [5]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
Total	18	28	16	7	0

Table 15: Level of English of non-native speakers [5]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	171	2	7	2

Table 16: Speaker type used to listen to the speech samples [5]

Same environment?	Yes	No
Total	173	7

Table 17: Same environment for all samples? [5]

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	131	40	8	1	0

Table 18: Kind of environment when listening to the speech samples [5]

Number of sessions	1	2-3	4 or more
Total	124	39	18

Table 19: Number of separate listening sessions to complete all the sections [5]

Browser	Firefox	IE	Chrome	Opera	Safari	Mozilla	Other
Total	104	3	38	0	30	3	3

Table 20: Web browser used (The paid listeners -type EE- all did the test on Safari.) [5]

Similarity with reference samples	Easy	Difficult
Total	138	34

Table 21: Listeners' impression of their task in section(s) about similarity with original voice. [5]

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
Total	19	1	16

Table 22: Listeners' problems in section(s) about similarity with original voice. [5]

Number of times	1-2	3-5	6 or more
Total	144	32	0

Table 23: Number of times listened to each example in section(s) about similarity with original voice. [5]

Naturalness	Easy	Difficult
Total	168	8

Table 24: Listeners' impression of their task in MOS naturalness sections [5]

Problem	All sounded same and/or too hard to understand	Scale too big, too small, or confusing	Bad speakers, playing files disturbed others connection too slow, etc	Other
Total	2	5	1	4

Table 25: Listeners' problems in MOS naturalness sections [5]

Number of times	1-2	3-5	6 or more
Total	158	12	0

Table 26: How many times listened to each example in MOS naturalness sections? [5]

Book passage	Easy	Difficult
Total	101	82

Table 27: Listeners' impression of their task in the sections involving book paragraphs. [5]

Problem	All sounded same and/or too hard to understand	Scale too big, too small, or confusing	Bad speakers, playing files disturbed others connection too slow, etc	Other
Total	14	47	1	27

Table 28: Listeners' problems in the sections involving book paragraphs [5]

Number of times	1-2	3-5	6 or more
Total	154	20	0

Table 29: How many times listened to each example in the sections involving book passages? [5]

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	5	85	7	85

Table 30: Listeners' impressions of intelligibility task (SUS). [5]