The Blizzard Challenge 2017

Simon King, Lovisa Wihlborg, Wei Guo

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

Abstract

The Blizzard Challenge 2017 was the thirteenth annual Blizzard Challenge and was once again organised by Simon King at the University of Edinburgh, with support from the other members of the Blizzard Challenge committee – Keiichi Tokuda and Alan Black. The task this year was the same as in 2016, using a slightly expanded 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

Since Black and Tokuda conceived the Blizzard Challenge [1], there have been annual summary papers like this one, plus a 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].

2. Participants

This years challenge, Blizzard 2017, had 13 participants, which are listed in Table 1 alongside the 4 benchmarks.

As well as natural speech, three benchmark synthesis systems were included this year. The unit selection and HMM-based benchmarks are the same types as in many previous challenges and will aid comparisons with those previous years. The DNN benchmark is the same as that first introduced in 2016.

The unit selection benchmark[1] is Festival 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` and using the Unisyn dictionary [5]. 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 Festival front end (again with the Unisyn dictionary) 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 each year. Most participanting teams reveal their system identifier in their workshop paper.

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
Alibaba-IDST	Alibaba Group	hybrid
CMU	Carnegie Mellon University	clustergen
CSTR	Centre for Speech Technology Research, U Edinburgh	hybrid
CUED	University of Cambridge Engineering Department	DNN
I2R-NWPU	Institute for Infocomm Research & Northwestern Polytechnical U	hybrid
IFLYTEK	Iflytek Ltd	hybrid
IIITH	International Institute of Information Technology	DNN
IRISA	U Rennes	unit selection
MARYTTS	Deutsche Forschungszentrum für Künstliche Intelligenz	DNN
NITECH	Nagoya Institute of Technology	DNN
NLPR	National Laboratory of Pattern Recognition	hybrid
USTC-NELSLIP	U Science and Technology of China & iFLYTEK	DNN
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: HMM (Hidden Markov Model statistical parametric), DNN (Deep Neural Network statistical parametric) Clustergen (decision tree statistical parametric), unit selection (using waveform concatenation), or hybrid (waveform concatenation guided by a statistical parametric model such as a DNN).

[1] Thanks to Srikanth Ronanki, CSTR

[2] Thanks to Kei Sawada, NIT

[3] Thanks again to Srikanth Ronanki

3. Voice to be built

This section repeats, for the reader's convenience, information that was provided in the 2016 summary paper [6], the only significant difference being the modest increase in the amount of data made available to participants.

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 6.5 hours of material was made available to participants in the challenge, comprising the 5 hours used for the 2016 Blizzard Challenge, plus the 6 books that were the test set from that year.

Each of the 56 books in the 6.5 hour released set of data 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'.

For copyright reasons, participants only had access to the plain text of the books. The full versions of the books are richly illustrated on every page. One feature that was hard to replicate in the plain text was the layout of pages, and in particular how some text is associated with, wrapped around, or actually part of an illustration (e.g., in speech bubbles). Paragraphs (which are not always clearly defined in these books) and page breaks were indicated in the numbering of the lines in the plain text version.

As in all Blizzard Challenges, the organisers held out some of the material for use as a test set. The 2017 test set included 6 complete audiobooks across a range of genres and reading ages that have never been released to participants.

3.2. Tasks

Participants were invited to take part in a single task in accordance with the rules of the challenge, which were identical to those of 2016, and were published on the website: build a voice from the provided data, suitable for reading children's audiobooks. This was denoted as task 2017-EH1, following the standard Blizzard Challenge task naming scheme.

This year, there was a related (but separately organised) challenge in which participants were provided with carefully prepared features extracted from text and audio, based on the data from the main chellenge of 2016. A description of the "Blizzard Machine Learning Challenge 2017" can be found at https://www.synsig.org/index.php/Blizzard_Machine_Learning_Challenge_2017_Rules .

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 obtained from most 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 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 (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. An innovation for 2017 was to require them to report this number to the organisers. Around half of the teams recruited 10 or more listeners each.

3.5. Listening tests

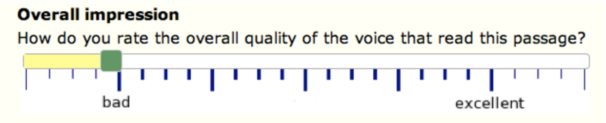
The listening test had the following structure, comprising 7 sections. There were 16 systems (3 benchmarks + 13 participating teams), plus natural speech. That leads to 17 samples being judged by each listener in each of sections 1-5, and 16 samples in each of sections 6-7 of the test:

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 [7], 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")



Overall impression

How do you rate the overall quality of the voice that read this passage?

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

3.6. Listening test completion rate

Table 7 gives a breakdown of evaluation completion rates for each listener type. It appears that completion rates are very much better than in 2016. This suggests that placing more responsibility on participating teams to recruit listeners was very effective. This should become a regular feature in future.

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 [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. 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.

We can comparing Figure 2 with the corresponding figure in the 2016 summary paper [6], and by also referring to Table 1, we observe some small but possibly important 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. The DNN benchmark (D) is also towards the lower end. These are both positive results because they indicate that most participating teams were able to create systems that are at least as natural as these simple benchmark systems.

There is only one unit selection based entry (there were 3 in 2016). The other unit selection system is the Festival benchmark (B), which once again performs quite well. This is a little surprising, given the expressive nature of the speech data.

It can be seen that those systems that generate the waveform using concatenation (unit selection in blue and hybrid in red) are – as in previous challenges – generally more natural-sounding than the systems that employ a vocoder. However, this year we can see two exceptions to that: systems G and L. System G uses the same architecture as many other DNN-based statistical parametric systems (acoustic model, post-filter, waveform generator), but the waveform is generated by a neural vocoder instead of (for example) STRAIGHT or WORLD. System L is a DNN-based system using the STRAIGHT vocoder.

No synthesiser is as natural as the natural speech (refer to the first row or column of Figure 3). System I is significantly more natural than all other systems. Systems G, L, E, P form a group that are equally natural to one another, less natural that system I, but more natural than all other systems.

For intelligibility, no comparisons with natural speech possible were possible this year. System D has the lowest Word Error Rate (WER), although this is not significantly lower (Figure 5) than the WER of systems G, I, L, M. We can therefore say that system I is not only the most natural, but also amongst the most intelligible.

Systems I and E are equally similar to the original speaker, although both are significantly less so than recordings of the speaker herself.

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 I is better than all other systems along 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 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. Apple and Google provided financial support, including funding for the second author 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, 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] Susan Fitt, "Documentation and user guide to unisyn lexicon and post-lexical rules," Tech. Rep., Centre for Speech Technology Research, Edinburgh, 2000.

[6] S. King and V. Karaiskos, "The Blizzard Challenge 2016," in *Proc. Blizzard Workshop*, 2016.

[7] 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.

[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)*, 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 [4] or on the listening test results themselves. [5]

[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.

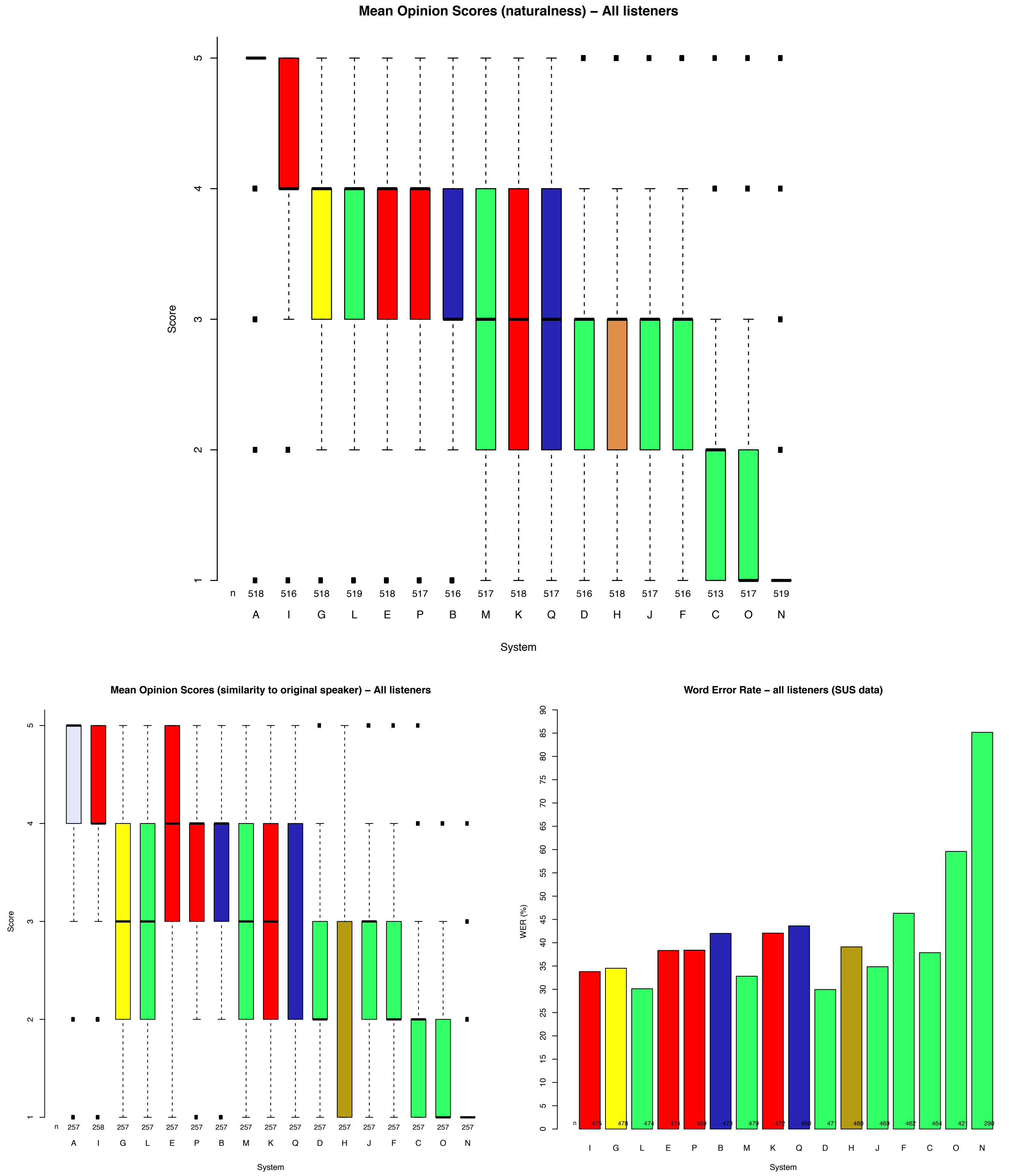


Figure 2: Results for task 2017-EH1 on sentence test material, pooling all listeners' responses. The plots are colour-coded: green for statistical parametric systems that employ some form of vocoder to generate the waveform, blue for unit selection systems and red for hybrid systems that concatenate waveforms (generally guided by a DNN). The exceptions are system A (natural speech), system G which is described by its authors as a 'hybrid parametric system of HMM, RNN, GAN and wavenet' and system H which is a statistical parametric system using decision trees. 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.

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

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.

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

Figure 4: Significant differences in speaker similarity (book sentences) between systems are indicated by a solid black box.

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

Figure 5: Significant differences in intelligibility (SUS) between systems are indicated by a solid black box.

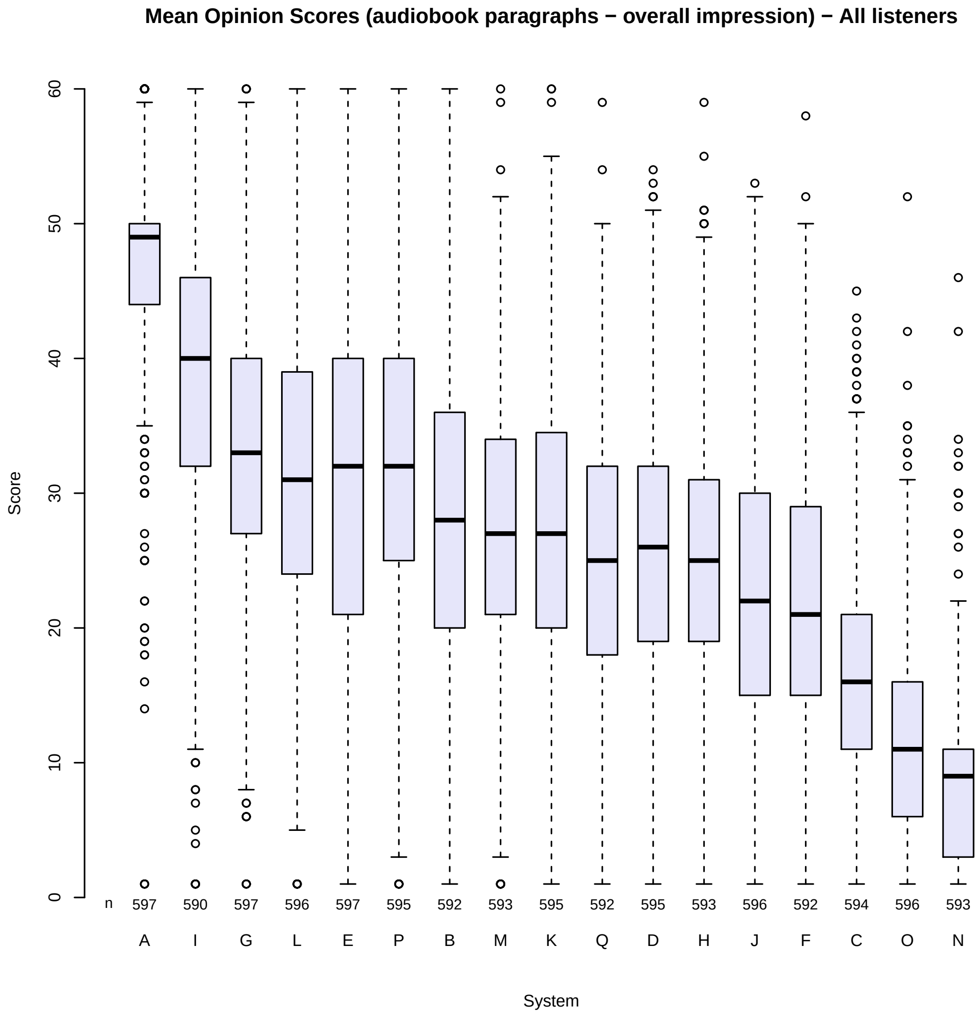


Figure 6: Overall impression of paragraphs for task 2017-EH1.

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

Figure 7: Significant differences in overall impression of paragraphs for task 2017-EH1.

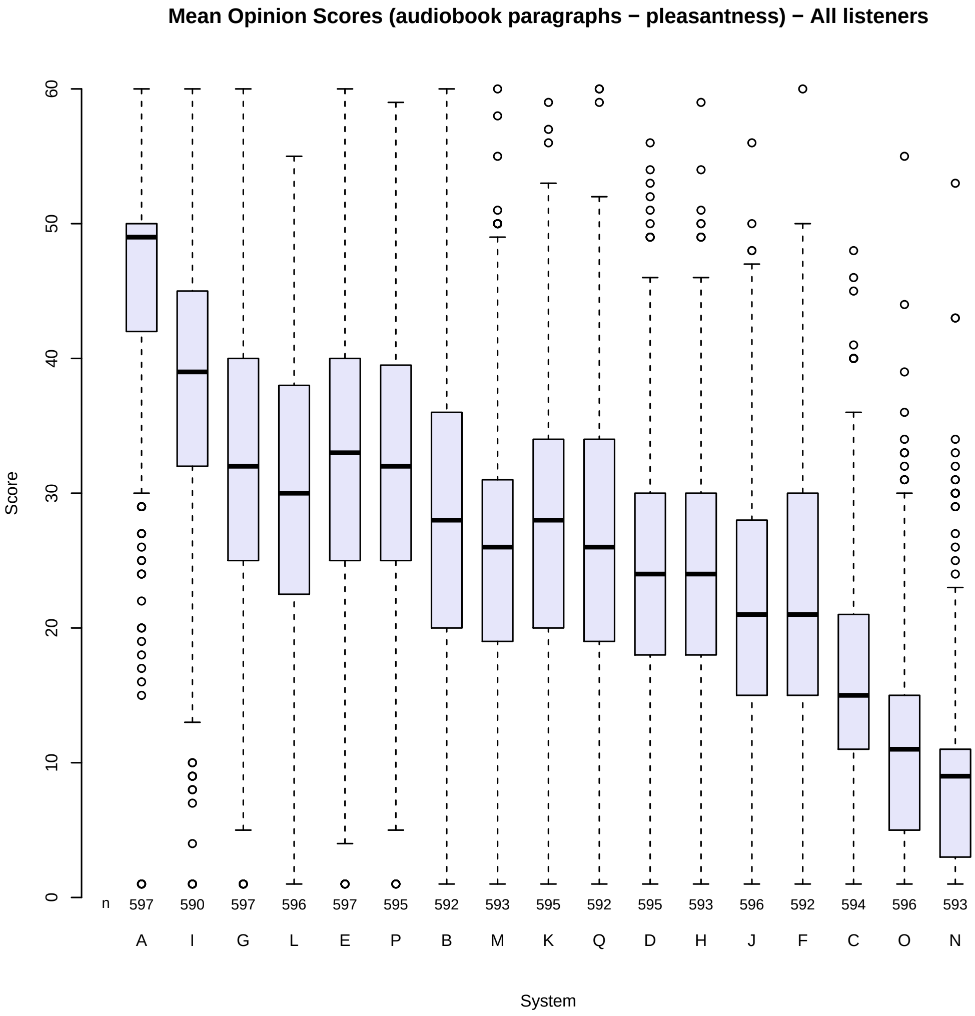


Figure 8: Pleasantness of paragraphs for task 2017-EH1.

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

Figure 9: Significant differences in pleasantness of paragraphs for task 2017-EH1.

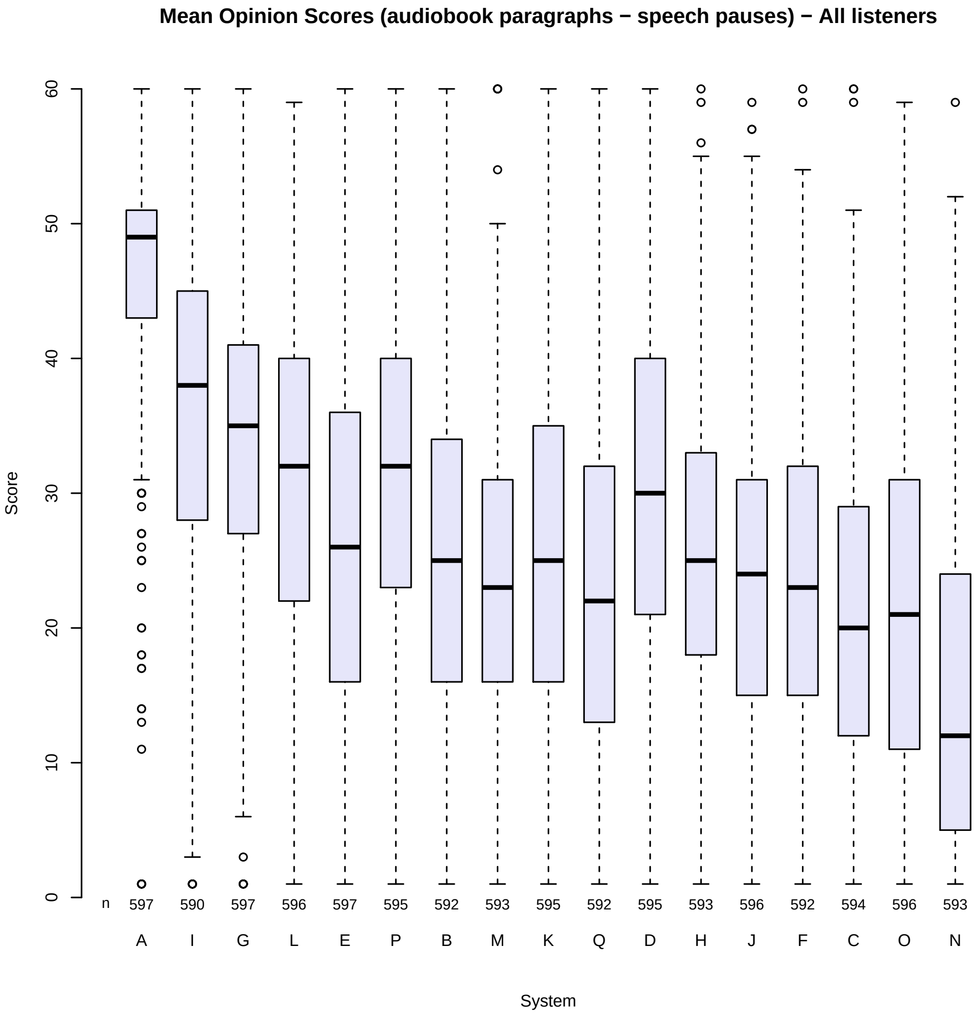


Figure 10: Speech pauses of paragraphs for task 2017-EH1.

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

Figure 11: Significant differences in speech pauses of paragraphs for task 2017-EH1.

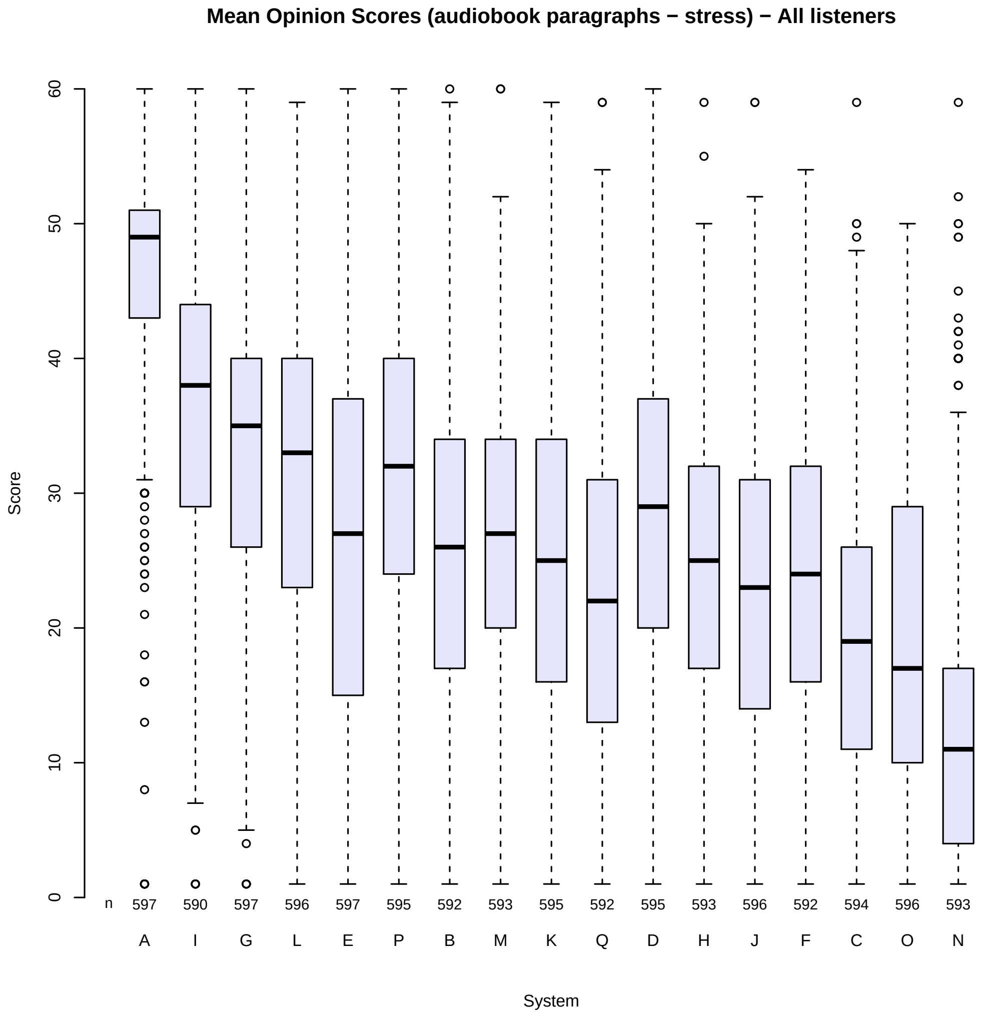


Figure 12: Stress of paragraphs for task 2017-EH1.

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

Figure 13: Significant differences in stress of paragraphs for task 2017-EH1.

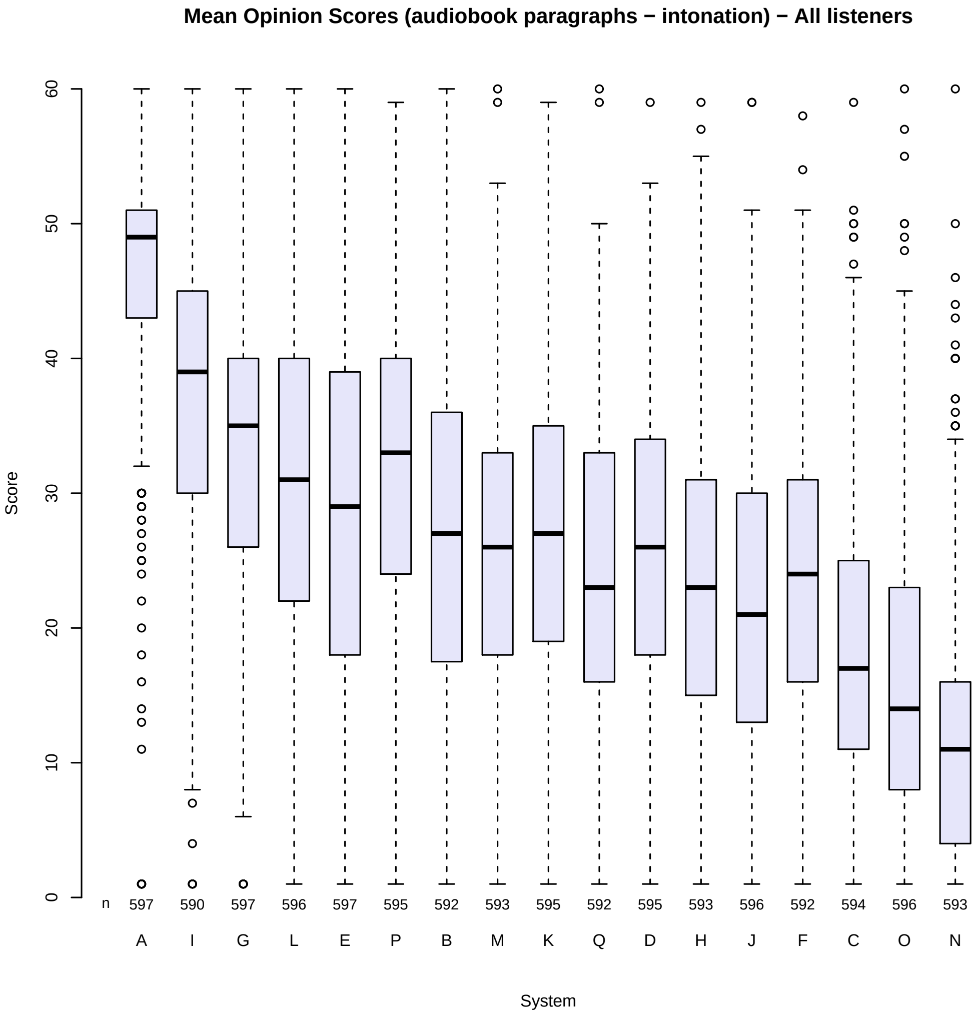


Figure 14: Intonation of paragraphs for task 2017-EH1.

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

Figure 15: Significant differences in intonation of paragraphs for task 2017-EH1.

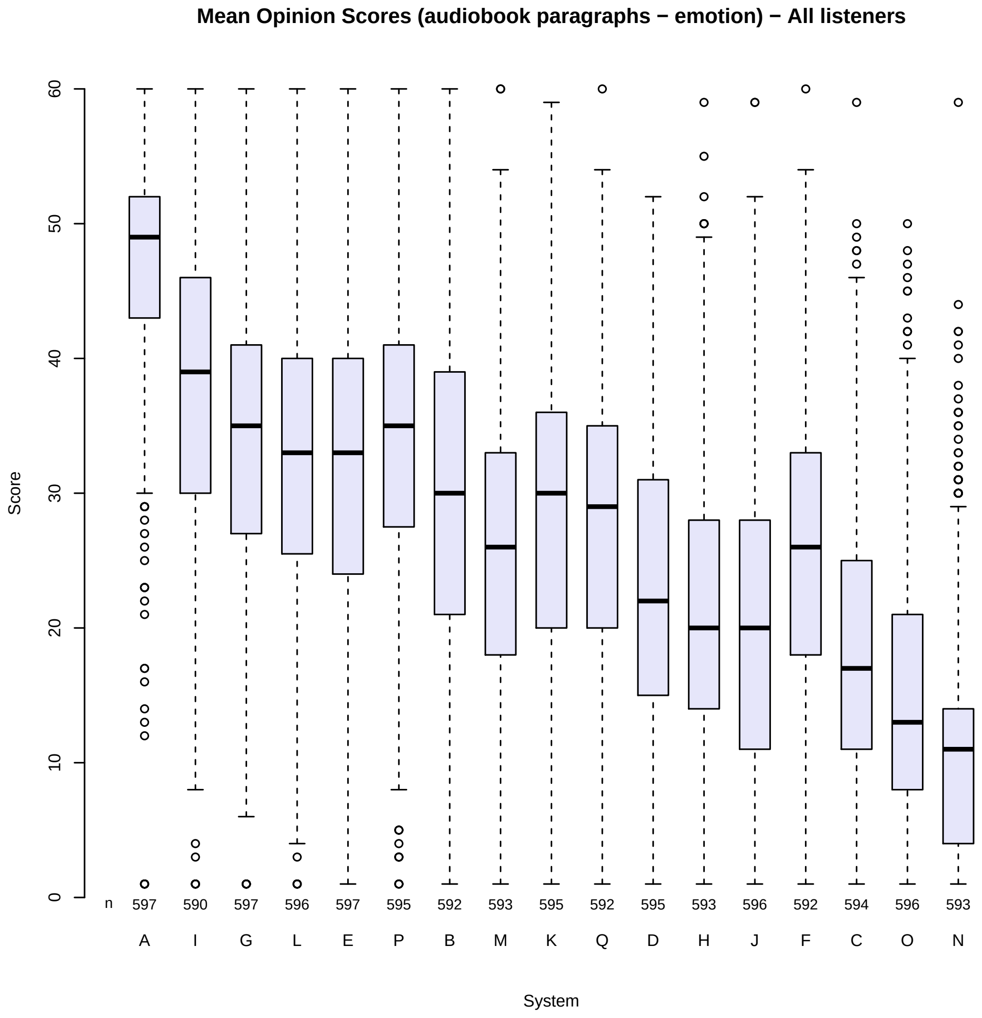


Figure 16: Emotion of paragraphs for task 2017-EH1.

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

Figure 17: Significant differences in emotion of paragraphs for task 2017-EH1.

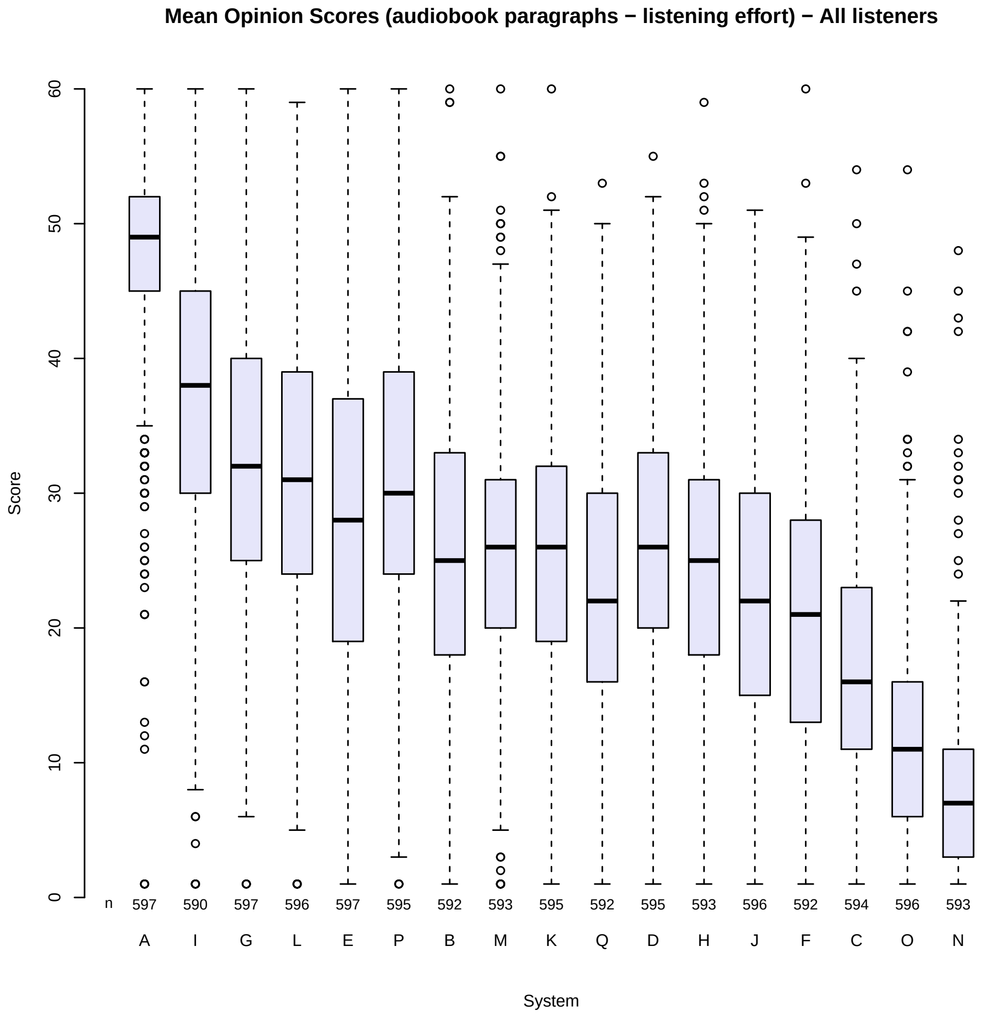


Figure 18: Listening effort of paragraphs for task 2017-EH1.

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

Figure 19: Significant differences in listening effort of paragraphs for task 2017-EH1.

Language	Total
Chinese (Mandarin)	48
Persian	1
Hebrew	1
French	10
Japanese	36
Hungarian	1
Spanish	1
Hindi	1
Marathi	1
German	1

Table 2: First language of non-native speakers. [4]

Gender	Male	Female
Total	131	114

Table 3: Gender. [4]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	14	211	76	31	19	17	4	0

Table 4: Age of listeners whose results were used (completed the evaluation fully or partially). [5]

Native speaker	Yes	No
English	141	103

Table 5: Native speakers. [4]

	Task EH1
EP	105
ER	186
EE	81

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.) [5]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EP	105	0	0	105
ER	208	22	97	89
EE	93	12	23	58
ALL	**406**	**34**	**120**	**252**

Table 7: 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	EH1_16	EH1_17
EP	7	7	7	6	6	7	6	6	6	5	5	6	6	7	6	6	6
ER	12	11	13	11	11	11	11	9	11	12	9	10	11	11	12	11	10
EE	5	5	6	6	5	5	5	5	6	4	5	4	3	5	3	4	5
ALL	24	23	26	23	22	23	22	20	23	21	19	20	20	23	21	21	21

Table 8: 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	ER	EE	ALL
Total	105	88	58	251

Table 9: Listener type totals for submitted feedback.

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	21	29	103	65	28	0

Table 10: Highest level of education completed. [4]

CS/Engineering person?	Yes	No
Total	115	131

Table 11: Computer science / engineering person. [4]

Work in speech technology?	Yes	No
Total	85	161

Table 12: Work in the field of speech technology. [4]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	36	27	31	46	61	28	16

Table 13: How often normally listened to speech synthesis before doing the evaluation. [4]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	0	5	82	26	6	22

Table 14: Dialect of English of native speakers. [4]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
Total	24	54	15	9	1

Table 15: Level of English of non-native speakers. [4]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	225	10	10	1

Table 16: Speaker type used to listen to the speech samples. [4]

Same environment?	Yes	No
Total	237	8

Table 17: 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	160	75	9	1	0

Table 18: Kind of environment when listening to the speech samples. [4]

Number of sessions	1	2-3	4 or more
Total	164	61	19

Table 19: Number of separate listening sessions to complete all the sections. [4]

Browser	Firefox	IE	Chrome	Opera	Safari	Mozilla	Other
Total	43	12	104	1	59	0	10

Table 20: Web browser used (the paid listeners - type EP - did the test on either Safari or Chrome). [4]

Similarity with reference samples	Easy	Difficult
Total	186	56

Table 21: Listeners' impression of their task in the section(s) about similarity with original voice. [4]

Problem	Scale too big, too small, or confusing	Difficulties with judging similarity	Other
Total	4	13	5

Table 22: Listeners' problems in the section(s) about similarity with original voice. [4]

Number of times	1-2	3-5	6 or more
Total	208	27	0

Table 23: Number of times listened to each example in the section(s) about similarity with original voice. [4]

Naturalness	Easy	Difficult
Total	216	23

Table 24: Listeners' impression of their task in the MOS naturalness sections. [4]

Problem	Difficulties with judging naturalness	Scale too big, too small, or confusing	Other
Total	6	1	2

Table 25: Listeners' problems in the MOS naturalness sections. [4]

Number of times	1-2	3-5	6 or more
Total	218	17	0

Table 26: Number of times listened to each example in the MOS naturalness sections. [4]

Book passage	Easy	Difficult
Total	148	97

Table 27: 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	Difficulties with separating the different variables of the voice	Bad speakers, playing files disturbed other connection too slow, etc	Issues with what to use as a reference	Other
Total	7	6	16	1	6	4

Table 28: Listeners' problems in the sections involving book passages. [4]

Number of times	1-2	3-5	6 or more
Total	204	33	0

Table 29: 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	23	111	93	16

Table 30: Listeners' impressions of the intelligibility task (SUS). [4]