The Blizzard Challenge 2012

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 2012 was the eighth annual Blizzard Challenge which was once again organised by the University of Edinburgh with assistance from the other members of the Blizzard Challenge committee – Prof. Keiichi Tokuda and Prof. Alan Black. One single-speaker English corpus was used, created from audiobook recordings on the Librivox website. Besides the main task of creating synthetic voices from these data, participants were invited to propose novel forms of evaluation.

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

1. Introduction

Since the Blizzard Challenge, conceived by Black and Tokuda in 2005 [1], is a regular event and firm fixture in the calendar, this paper only provides the specific details of the 2012 challenge. For background information, please refer to the previous summary papers for 2005 [1, 2], 2006 [3], 2007 [4], 2008 [5], 2009 [6], 2010 [7] and 2011 [8]. These, and many other useful resources, such as anonymised releases of the submitted speech, reference samples, listening test responses, scripts for running similar web-based listening tests and the statistical analysis scripts, can all be found via the Blizzard Challenge website [9].

2. Participants

The Blizzard Challenge 2005 [1, 2] had 6 participants, Blizzard 2006 had 14 [3], Blizzard 2007 had 16 [4], Blizzard 2008 had 19 [5], Blizzard 2009 had 19 [6], Blizzard 2010 had 17 participants, Blizzard 2011 had 9 participants. This year, 2012, there were 9 participants listed in Table 1 took part.

One benchmark system was included this year, to aid comparisons across the years, a Festival-based unit selection system from CSTR configured very similarly to the Festival/CSTR entry to Blizzard 2006 [10].[1].

As always, several additional groups (not listed here) registered for the Challenge, obtained the corpora, but did not submit anything for evaluation. When reporting anonymised results, the systems are identified using letters, with A denoting natural speech, B the Festival benchmark systems and C to K denoting the systems submitted by participants in the challenge.

Short name	Details	Method
NATURAL	Natural speech from the same speaker as the corpus	human
FESTIVAL	The Festival unit-selection benchmark system [10]	unit selection
USTC	University of Science and Technology of China	hybrid
HELSINKI	University of Helsinki	HMM
NTUT	National Taipei University of Technology	HMM
LESSAC A	Lessac Technologies	unit selection
I2R	Institute for Infocomm Research	unit selection
NITECH	Nagoya Institute of Technology	HMM
ILSP	Institute for Language and Speech Processing	unit selection
LESSAC B	Lessac Technologies	diphone
DFKI	Deutsche Forschungszentrum für Künstliche Intelligenz	HMM

Table 1: The participating systems and their short names. The first two rows are the benchmarks and correspond to the system identifiers A and B. The remaining rows are in alphabetical order of the system's short name and *not* in alphabetical order. Note that Lessac Technologies were permitted two entries to this year's challenge as an acknowledgement of their generosity in providing the data for last year's challenge.

3. Voices to be built

3.1. Speech databases

The English data for voice building was originally obtained from the Librivox audiobook website, but extensive preparation was carried out by Toshiba Research Europe Ltd who generously shared this work with other participants. The data are now available to non-participants and can be obtained via the Blizzard Challenge website. The speaker is John Greenman and is a male native speaker of US English. Four audiobooks, all by the same author (Mark Twain) and read by the same speaker, were made available to participants: A Tramp Abroad, Life on the Mississippi, The Adventures of Tom Sawyer, The Man That Corrupted Hadleyburg and Other Stories. Toshiba provided a processed version of the data comprising audio segmented into utterances and text automatically aligned with the segmented audio (along with a confidence measure relating to the reliability of the text transcription).

3.2. Tasks

Participants were invited to take part in the following tasks, in accordance with the rules of the challenge, published on the website.

- EH2.1: build a voice from the database.
- ES2.2: devise a method for evaluating synthetic speech for audiobook applications, and use it to evaluate task EH2.1. The evaluation can use any text you wish (but you are encouraged to consider using both 'in domain' and 'out of domain' text). It can measure any aspect of the synthetic speech that you think is relevant to its performance as an "audiobook reader". You will have to opportunity to request the participants in task EH2.1 to synthesise text provided by you. Participants in this task will be responsible for executing their own listening test: the Blizzard organisers will be running an independent test of their own in

[1] Many thanks to Rob Clark for creating the Festival benchmark

parallel

Only one participant took part in EH2.2, indicating perhaps that most participants are disappointingly uninterested in the form that the evaluation takes and are not concerned with evaluating specific aspects of their systems.

3.3. Listening test design and materials

The participants were asked to synthesise many hundreds of test sentences, of which a subset were used in the listening test. For a general overview of the listening test design and the web interface used to deliver it, again please refer to previous summary papers. Permission has been obtained from participants to distribute parts of this dataset along with the listener scores and this can be downloaded via the Blizzard website. Natural examples (denoted as 'System A' in the results) of a subset of the test sentences were available this year, allowing direct comparisons between natural and synthetic speech in some cases. Table 4 lists the types of material used in the listening test.

3.4. Listener types

Various listener types were employed in the test: letters in parenthesis below are the identifiers used for each type in the results distributed to participants. For English, the following listener types were used:

- Paid UK undergraduates, 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 (EE).
- Speech experts, recruited via participating teams and mailing lists (ES).
- Volunteers recruited via participating teams, mailing lists, blogs, word of mouth, etc. (ER).

Table 11, summarised in Table 2, shows the number of listeners of each type obtained.

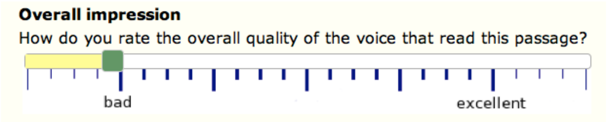
3.5. Listening tests

When using paid listeners, it is easier to employ a listening test lasting 45-60 minutes, rather than many short tests. The listening test had the following structure, comprising 9 sections, each with either 10 or 11 stimuli presented (depending on the availability of natural speech for that particular text):

1. Similarity, novel
2. Naturalness, novel
3. Naturalness, novel
4. Naturalness, news
5. Naturalness, news
6. Multiple dimensions, in-domain novel paragraphs
7. Multiple dimensions, out-of-domain novel paragraphs
8. Intelligibility, SUS, single listen only

The "Multiple dimensions"" evaluation of paragraphs was proposed in [11] and contained the following sections, in which listeners provided their response using a slider as illustrated in Figure 1:

- 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.

Within each numbered 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, something that is particularly important for testing intelligibility. The number of listeners obtained is shown in Table 2. See Table 10 for a detailed breakdown of evaluation completion rates for each listener type.

Total registered	321
of which:	
Completed all sections	225
Partially completed	55
No response at all	41

Table 2: Number of listeners obtained

4. Analysis methodology

As usual, we combined the responses from 'completed all sections' and 'partially completed' listeners together in all analyses. In this paper, we will only give the 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 via the Blizzard website. Please refer to [12] for a description of the statistical analysis techniques used and justification of the statistical significance techniques employed. In all material published by the organisers, system names are anonymised. Individual teams are free to reveal their system identifier if they wish. See Section 5.1 and Tables 5 to 31 for a summary of the responses to the questionnaire that listeners were asked to optionally 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 on task EH2.1 for all listeners combined and all 4 naturalness sections combined. Note that this ordering is intended only to make the plots more readable and *cannot 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 indicates the types of systems using colour coding. It can be seen that those systems that generate the waveform using concatenation (unit se-

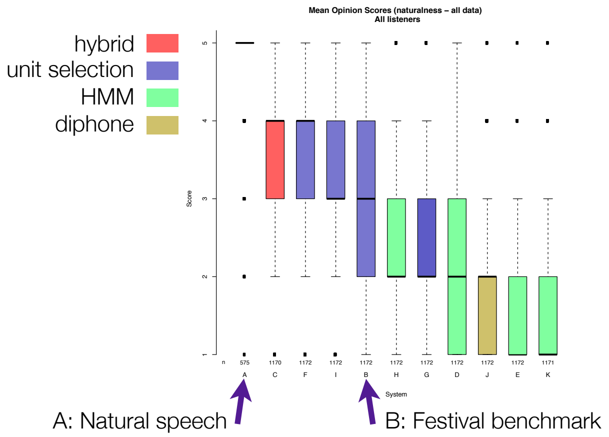


Figure 2: Indication of system types, overlaid on a plot of mean opinion scores for naturalness.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.7	0.62	575	909
B	3	1.5	3.0	1.08	1172	312
C	4	1.5	3.8	0.93	1170	314
D	2	1.5	2.2	1.02	1172	312
E	1	0.0	1.6	0.74	1172	312
F	4	1.5	3.4	0.96	1172	312
G	2	1.5	2.5	0.93	1172	312
H	2	1.5	2.6	1.01	1172	312
I	3	1.5	3.3	1.07	1172	312
J	2	1.5	1.9	0.95	1172	312
K	1	0.0	1.6	0.77	1171	313

Table 3: Mean opinion scores for naturalness on task EH2.1. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

lection or hybrid) are generally more natural-sounding than the HMM-based systems which employ a vocoder.

Naturalness results on sentence material are given in Table 3. No synthesiser is as natural as the natural speech (Figure 3 . System C is significantly more natural than all other synthesisers, with systems F and I less natural than natural speech and system C, but more natural than all the other remaining systems. System C was also judged as more similar to the original speaker than all other systems, but not as similar as the natural speech itself. Regarding intelligibility, System C was also one of the most intelligible systems, but not significantly better than systems D and H. Since we did not have natural speech available for the SUS section of the listening test, no conclusions can be drawn this year regarding the relative intelligibility of synthetic and natural speech.

The multiple dimensions of scoring for the paragraphs are reported in Figures 4 to 6. It can be seen that system C is again – along all dimensions except "emotion" – superior to all other systems, but never as good as natural speech. Systems F and I fall slightly behind system C, but ahead of the remaining systems along most dimensions. The different dimensions are (unsurprisingly) strongly related, at least in terms of system ranking but some different patterns across the systems are teased apart, especially for speech pauses and stress.

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 5 to 31.

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 and CER/PTER/PER programmes; Rob Clark built the Festival benchmark system. 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 the data, and Google provided financial 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] 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] Mark Fraser and Simon King, "The Blizzard Challenge 2007," in *Proc. Blizzard Workshop (in Proc. SSW6)*, 2007.

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

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

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

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

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

[10] 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.

[11] 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.

[12] 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 at the end of this paper, please refer to the footnotes which specify whether the numbers are based on listener feedback [2] or on the listening test results themselves. [3]

[2] 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.

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

Type	Source	Example
news	Glasgow Herald newspaper	These would still have to be ratified by member states, he added.
novel sentences	in-domain from Mark Twain novels not included in the distributed data	Let us now draw this history to a close, for little more needs to be told.
novel paragraphs	in-domain from Mark Twain novels not included in the distributed data / out-of-domain from other authors/periods/styles	The evening arrived; the boys took their places. The master, in his cook's uniform, stationed himself at the copper; his pauper assistants ranged themselves behind him; the gruel was served out; and a long grace was said over the short commons. The gruel disappeared; the boys whispered each other, and winked at Oliver; while his next neighbors nudged him. Child as he was, he was desperate with hunger, and reckless with misery. He rose from the table; and advancing to the master, basin and spoon in hand, said: somewhat alarmed at his own temerity: 'Please, sir, I want some more'.
SUS	semantically unpredictable	Why must a thumb greet the ring?

Table 4: The sentence types used in the listening test, and their sources.

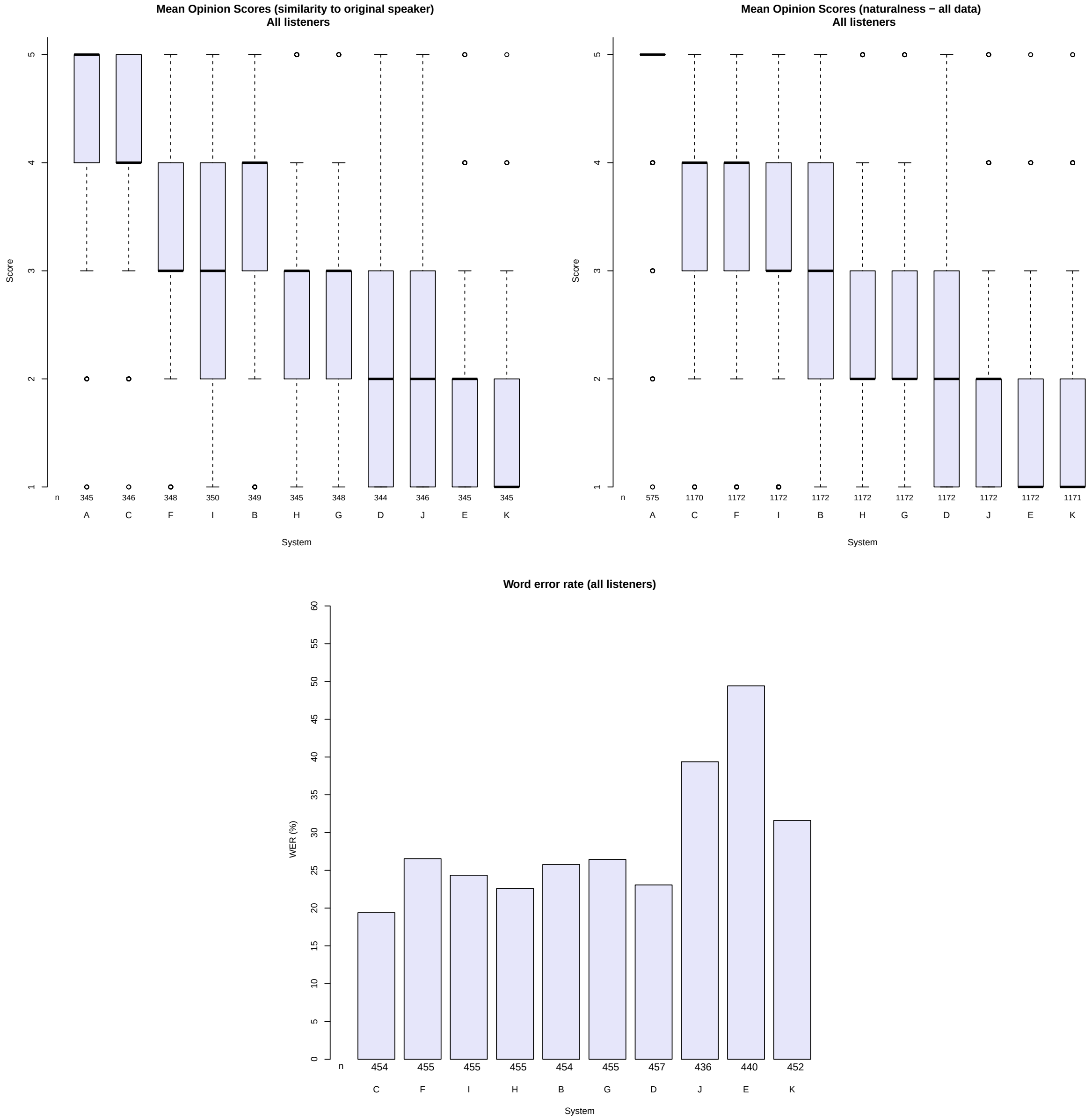


Figure 3: Results for task EH2.1 on sentence test material.

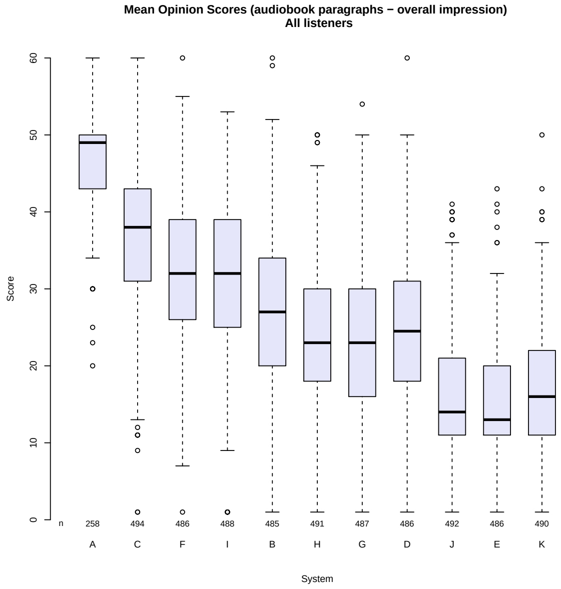
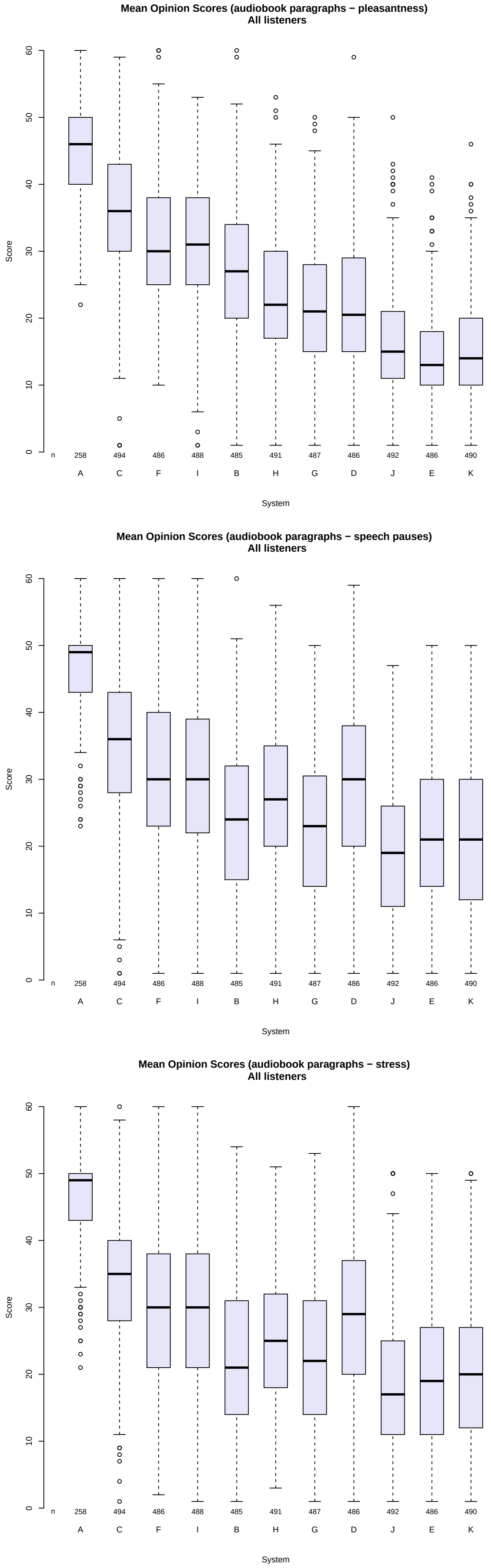


Figure 4: Results for task EH2.1 on paragraph test material, pooling both in- and out-of-domain material.



Mean Opinion Scores (audiobook paragraphs − pleasantness)
All listeners

Mean Opinion Scores (audiobook paragraphs − speech pauses)
All listeners

Mean Opinion Scores (audiobook paragraphs − stress)
All listeners

Figure 5: Results for task EH2.1 on paragraph test material, pooling both in- and out-of-domain material, continued.

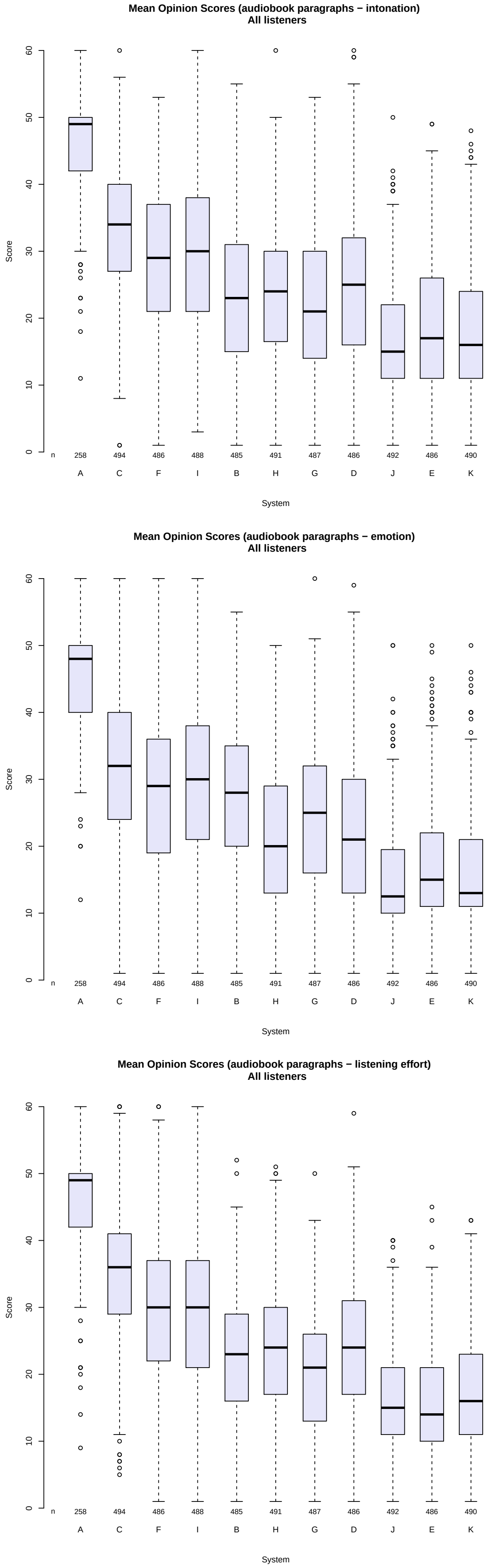


Figure 6: Results for task EH2.1 on paragraph test material, pooling both in- and out-of-domain material, continued.

Language	Total
Cantonese	1
Catalan	2
Chinese	7
Croatian	1
Dutch	2
Estonian	1
Finnish	4
French	2
German	9
Greek	5
Hindi	1
Hungarian	1
Ibibio	1
Igbo	1
Italian	1
Japanese	36
Korean	1
Nepali	1
Polish	2
Portuguese	4
Romanian	2
Slovak	2
Slovenian	1
Spanish	4
Swedish	1
Tamil	2
Telugu	1
Turkish	2
N/A	1

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

Gender	Male	Female
Total	131	91

Table 6: Gender [2]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
English total	19	171	50	24	8	6	2	0

Table 7: Age of listeners whose results were used (completed the evaluation fully or partially)

Native speaker	Yes	No
English	122	101

Table 8: Native speakers [2]

	Task EH1
EE	104
ER	52
ES	124
ALL	280

Table 9: 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) [3]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EE	104	0	0	104
ER	63	11	22	30
ES	154	30	33	91
ALL	**321**	**41**	**55**	**225**

Table 10: Listener registration and evaluation completion rates. [3]

	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
EE	8	8	8	8	8	8	8	8	8	8	8	8	8
ER	3	5	2	5	5	4	4	4	5	4	5	2	4
ES	9	10	9	8	11	11	10	11	9	10	9	10	7
ALL	20	23	19	21	24	23	22	23	22	22	22	20	19

Table 11: Listener groups - Voice EH1 (English), showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations [3]

Listener Type	EE	ER	ES	ALL
Total	104	30	91	225

Table 12: Listener type totals for submitted feedback

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate
English total	33	36	49	60	45

Table 13: Highest level of education completed [2]

CS/Engineering person?	Yes	No
English total	132	92

Table 14: Computer science / engineering person [2]

Work in speech technology?	Yes	No
English total	100	123

Table 15: Work in the field of speech technology [2]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
English total	40	37	23	44	43	9	25

Table 16: How often normally listened to speech synthesis before doing the evaluation [2]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	1	5	75	32	10	23

Table 17: Dialect of English of native speakers [2]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
English total	21	26	40	12	2

Table 18: Level of English of non-native speakers [2]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
English total	209	9	4	2

Table 19: Speaker type used to listen to the speech samples [2]

Same environment?	Yes	No
Total	220	4

Table 20: Same environment for all samples? [2]

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	162	46	12	0	3

Table 21: Kind of environment when listening to the speech samples [2]

Number of sessions	1	2-3	4 or more
Total	163	48	13

Table 22: Number of separate listening sessions to complete all the sections [2]

Browser	Firefox	IE	Chrome	Opera	Safari	Mozilla	Other
Total	52	42	17	0	110	0	3

Table 23: Web browser used (The paid listeners -type EE- all did the test on Safari.) [2]

Similarity with reference samples	Easy	Difficult
Total	145	77

Table 24: Listeners' impression of their task in section(s) about similarity with original voice. [2]

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
Total	46	1	30

Table 25: Listeners' problems in section(s) about similarity with original voice. [2]

Number of times	1-2	3-5	6 or more
Total	181	35	5

Table 26: Number of times listened to each example in section(s) about similarity with original voice. [2]

Naturalness	Easy	Difficult
Total	176	47

Table 27: Listeners' impression of their task in MOS naturalness sections [2]

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	11	23	0	13

Table 28: Listeners' problems in MOS naturalness sections [2]

Number of times	1-2	3-5	6 or more
Total	192	25	2

Table 29: How many times listened to each example in MOS naturalness sections? [2]

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	73	91	36	24

Table 30: Listeners' impressions of intelligibility task (addressess and SUS). [2]

Number of times	1-2	3-5	6 or more
Total	75	120	29

Table 31: How many times listened to each example in the intelligibility section. (SUS could only be heard once.) [2]