The Blizzard Challenge 2013

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 2013 was the ninth annual Blizzard Challenge which was once again organised by the University of Edinburgh with advice from the other members of the Blizzard Challenge committee – Prof. Keiichi Tokuda and Prof. Alan Black – joined this year by Dr. Kishore Prahallad who organised the Indian languages tasks. This summary paper only covers the English tasks; a separate paper is available which summarises those Indian language tasks. For the English tasks, a large single-speaker English corpus was used, comprising around 300 hours of audio from professionally-produced audiobooks.

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

1. Introduction

As noted in previous summary papers, the Blizzard Challenge, conceived by Black and Tokuda in 2005 [1], is now a regular event and firm fixture in the calendar. Therefore, this paper only provides the specific details of the 2013 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], 2011 [8], 2012 [8], and the Indian language tasks of 2013 [9]. A summary-of-summaries paper is also available, which attempt to find trends across all previous years of the challenge [10].

These and many other resources, from anonymised releases of the submitted speech, reference samples, listening test responses, to the scripts for running the listening test and the subsequent statistical analysis, can all be found via the Blizzard Challenge website [11].

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, Blizzard 2012 had 9 participants and this years challenge, Blizzard 2013, had 15 participants in total, of which 14 participated in the English tasks, which are listed in Table 1.

Two benchmark systems were included this year, to aid comparisons across the years. One is a Festival-based unit selection system[1] from CSTR configured very similarly to the Festival/CSTR entry to Blizzard 2006 [12]. 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`.

When reporting anonymised results, the systems are identified using letters, with A denoting natural speech, B the Festival benchmark systems, C the HTS benchmark system and the remaining letters denoting the systems submitted by participants in the challenge. The same identifiers were used across the English and Indian language tasks. Therefore, not all system identifiers will be mentioned in this paper, because we only discuss the entries to the English tasks here.

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

[2]Many thanks to Keiichi Tokuda and his team for creating the HTS benchmark

Short name	Details	Method
NATURAL	Natural speech from the same speaker as the corpus	human
FESTIVAL	Festival benchmark	unit selection
HTS	HTS benchmark	HMM
CMU	Carnegie Mellon University	HMM
DFKI	Deutsche Forschungszentrum für Künstliche Intelligenz	hybrid
I2R	Institute for Infocomm Research	unit selection
ILSP	Institute for Language & Speech Processing / Innoetics	unit selection
ISOLAR	Institute of Information Technology - Vietnam Academy of Technology / iSolar Company	HMM
LESSAC	Lessac Technologies Inc	unit selection
MERAKA	Meraka Institute	HMM
NITECH	Nagoya Institute of Technology	HMM
NTUT	National Taipei University of Technology	HMM
RACAI	Research Institute for Artificial Intelligence, Romanian Academy	unit selection
S4A	Simple4All project consortium	HMM
SHRC	Speech and Hearing Research Center, Peking University	hybrid
USTC	National Engineering Laboratory of Speech and Language Information Processing	hybrid

Table 1: The participating systems and their short names. The first three rows are the benchmarks and correspond to the system identifiers A, B and C in that order. The remaining rows are in alphabetical order of the system's short name and *not* in alphabetical order of system identifier.

3. Voices to be built

3.1. Speech databases

The English data for voice building was obtained, prepared and provided the the challenge by Lessac Technologies Inc., having originally came from the publishers Voice Factory International Inc. It comprises speech from one female professional narrator & actress, Catherine 'Bobbie' Byers, reading the text of a collection

of classic novels. These had been divided by the publishers of the original audiobooks into a number of genres, such as "Classic Novels", "Women's Classics", "Young Readers" and so on. No use was made of this categorisation in this year's challenge, other than to make sure to draw the test material from a variety of them, not just one.

In total, around 300 hours of speech was made available to participants, in the form of chapter-sized audio files that had been encoded using mp3 at a variety of bit rates. In addition, approximately 19 hours of non-compressed wav files were also made available. These wav files were segmented into sentence-sized portions and aligned with the text by Lessac Technologies. Participants were free to use this segmentation, provided text, and alignment, or to perform their own processing of the data. For the larger, 300 hour mp3-coded set, no corresponding text or segmentation of the audio was released and participants using that data were obliged to process it themselves; this included obtaining the corresponding text (e.g., from Project Gutenberg).

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

3.2. Tasks

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

- Task 2013-EH1: build a voice from the unsegmented 300 hour set, with no provided text
- Task 2013-EH2: build a voice from the segmented 19 hour set, either from the provided chapter-sized audio files, or from the segmented-into-sentences version also provided, optionally also using the provided text (aligned with the sentence-segmented audio)

All participants in the English tasks took part in task 2013-EH2, with most of them also tacking the more challenging main task 2013-EH1. Since 2013-EH1 was in fact the main focus this year (given the exceptionally large data set available), this paper focuses on the results for that task. For the natural reference (system A), test materials were manually extracted from the chapter-sized audio, for those parts of the listening test which used this type of sentence (e.g., not for the semantically-unpredictable material)

3.3. Listening test design and materials

As usual, in an attempt to preclude any manual intervention at synthesis time, participants were asked to synthesise many hundreds of test sentences, of which only a small subset were used in the listening test. 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. 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. The following listener types were used (remembering that this paper only related to the English tasks):

- 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).

Tables 11 and 12, 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 tests for the two tasks had the following structures, comprising 12 sections each with 10 or 11 stimuli for the 2013-EH1 task:

1. Similarity, novel sentences
2. Naturalness, novel sentences
3. Naturalness, novel sentences
4. Naturalness, novel sentences
5. Naturalness, news sentences
6. Naturalness, news sentences
7. Multiple dimensions, novel paragraphs
8. Multiple dimensions, novel paragraphs
9. Multiple dimensions, novel paragraphs
10. Intelligibility, SUS, single listen only
11. Intelligibility, SUS, single listen only
12. Intelligibility, SUS, single listen only

and 9 sections each with 14 or 15 stimuli for the 2013-EH2 task:

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

The variation in the number of stimuli is a consequence of the unavailability of natural speech for the news and SUS sections. The "Multiple dimensions"" evaluation of paragraphs was that proposed in [13], and which was also used in last year's challenge, that contains 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")

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.

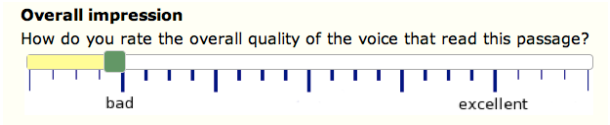


Figure 1: Example of a slider used to obtain listener responses in the paragraph sections.

	Paid	Speech experts	Volunteers
426 listeners registered:	112	110	204
of which the following percentages of listeners:			
Completed all sections	100%	43%	31%
Partially completed	0%	34%	43%

Table 2: Number of listeners obtained, with completion rates. Percentages do not sum to 100% within some listener categories due to listeners who registered but did not start the test.

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 [14] 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 5 to 34 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 on main task 2013-EH1 for all listeners combined and all 5 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 indicates the types of systems 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.

Naturalness results on sentence material in task 2013-EH1 are given in Table 3. No synthesiser is as natural as the natural speech (Figure 3). System M is more natural and more similar to the target speaker than any other system, followed by system K, then systems C, I, L as a group. Regarding intelligibility, there are

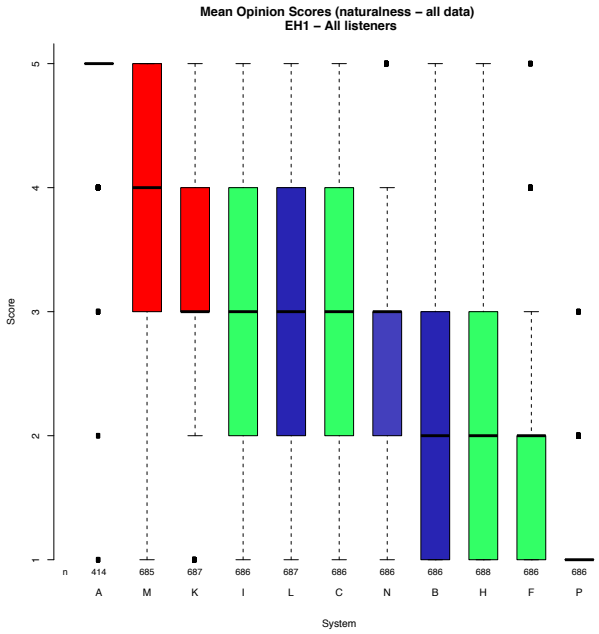


Figure 2: Indication of system types, overlaid on a plot of mean opinion scores for naturalness. Red bars correspond to hybrid systems that concatenate waveforms guided by a statistical model, green bars are for statistical parametric systems that employ some form of vocoder to generate the waveform, and blue bars are unit selection systems. System B is the Festival unit selection benchmark and system C is the HTS statistical parametric benchmark. A is natural speech.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.8	0.62	414	556
B	2	1.5	2.1	0.92	686	284
C	3	1.5	2.9	1.03	686	284
F	2	1.5	1.9	0.89	686	284
H	2	1.5	2.0	0.94	688	282
I	3	1.5	3.1	0.99	686	284
K	3	1.5	3.3	1.00	687	283
L	3	1.5	3.0	1.06	687	283
M	4	1.5	3.9	0.89	685	285
N	3	1.5	2.6	1.06	686	284
P	1	0.0	1.2	0.41	686	284

Table 3: Mean opinion scores for naturalness on task 2013-EH1. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

several equally-intelligible systems: the HTS benchmark (system C), the most natural system M, the second most natural system K, and system I. 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.

For task 2013-EH2, we do not present results here; please refer to the downloadable distribution mentioned above if you wish to see these. In summary, system M also performed well in this task, along with systems K and L. There were many equally-intelligible systems, including again the HTS benchmark (system C).

The multiple dimensions of scoring for the paragraphs are reported for task 2013-EH1 in Figures 4 to 6. As for the sentence material, please refer to the downloadable package of all results to see corresponding plots for task 2013-EH2. Unsurprisingly, no

system was judged to be as good as natural speech, along any dimension, in either the 2013-EH1 or 2013-EH2 tasks, In task 2013-EH1, system M was better than all other systems along most dimensions, followed by system K which was generally better than the remaining systems. In task 2013-EH2, fewer significant differences were found (partly due to larger number of systems which has the consequence of fewer stimulus presentations per system) although we can say that systems M and K were often better than the other systems.

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 34.

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] Kishore Prahallad, Anandaswarup Vadapalli, Naresh Elluru, Gautam Mantena, Bhargav Pulugundla, Peri Bhaskararao, Hema A. Murthy, Simon King, Vasilis Karaiskos, and Alan W. Black, "The Blizzard Challenge 2013 - Indian language tasks," in *Proc. Blizzard Workshop*, 2013.

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

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

[12] 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.

[13] 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.

[14] 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 [3] or on the listening test results themselves. [4]

[3] 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.

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

Type	Source	Example
news	Glasgow Herald newspaper	I am over the moon that people like something I have written.
novel sentences	Turn of the Screw (Henry James); Northanger Abbey (Jane Austen); The other two (Edith Wharton); The Ice Palace (FS Fitzgerald); A pair of silk stockings (Kate Chopin); Desiree's baby (Kate Chopin); Just so stories (R Kipling); Little Lord Fauntleroy (Frances Hodgson Burnett); Alice in Wonderland (Lewis Carroll); The Wizard of Oz (L. Frank Baum).	Alice looked up, and there stood the Queen in front of them, with her arms folded, frowning like a thunderstorm.
novel paragraphs		John Thorpe, who in the meantime had been giving orders about the horse, soon joined them, and from him she directly received the amends which were her due; for while he slightly and carelessly touched the hand of Isabella, on her he bestowed a whole scrape and half a short bow. He was a stout young man of middling height, who, with a plain face and ungraceful form, seemed fearful of being too handsome unless he wore the dress of a groom, and too much like a gentleman unless he were easy where he ought to be civil, and impudent where he might be allowed to be easy.
SUS	semantically unpredictable	Remember the bears and the fine steps.

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

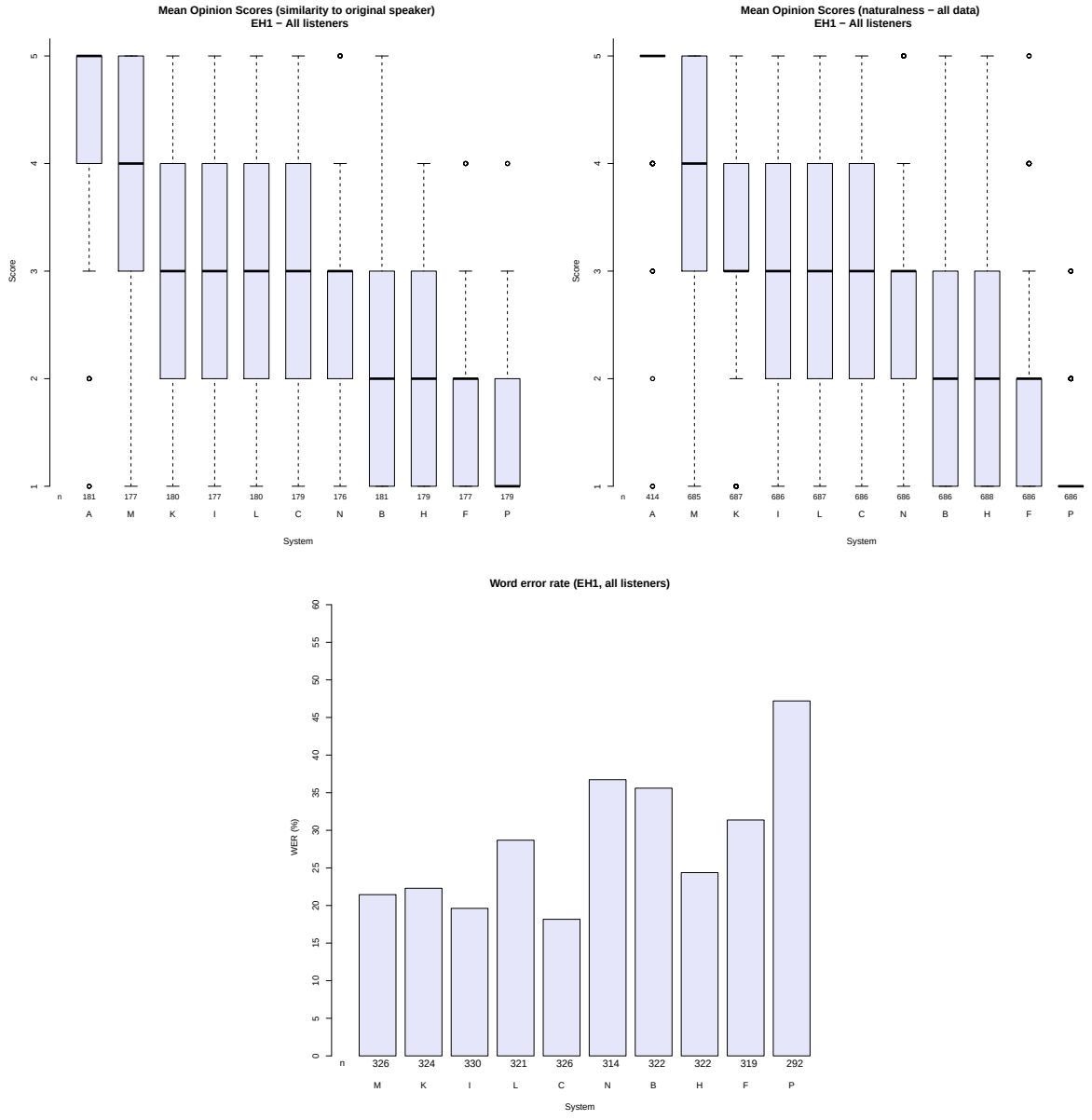


Figure 3: Results for task 2013-EH1 on sentence test material, pooling all material.

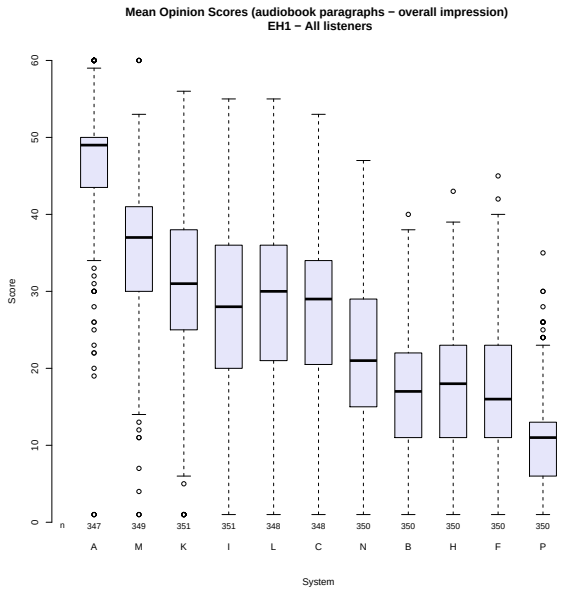
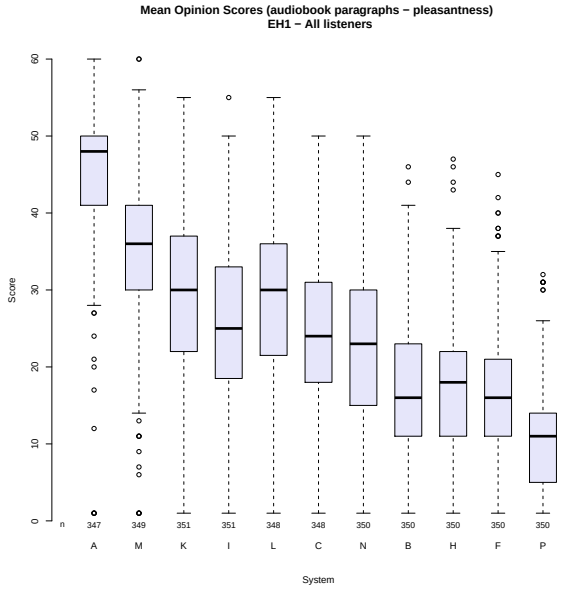
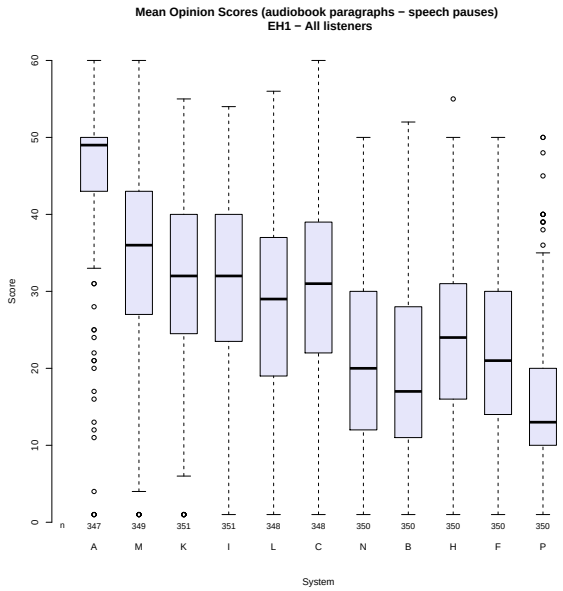


Figure 4: Results for task 2013-EH1 on paragraph test material, pooling all material.





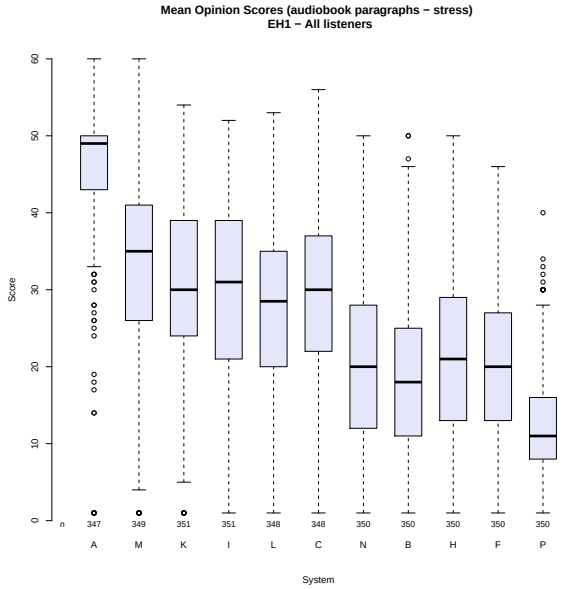
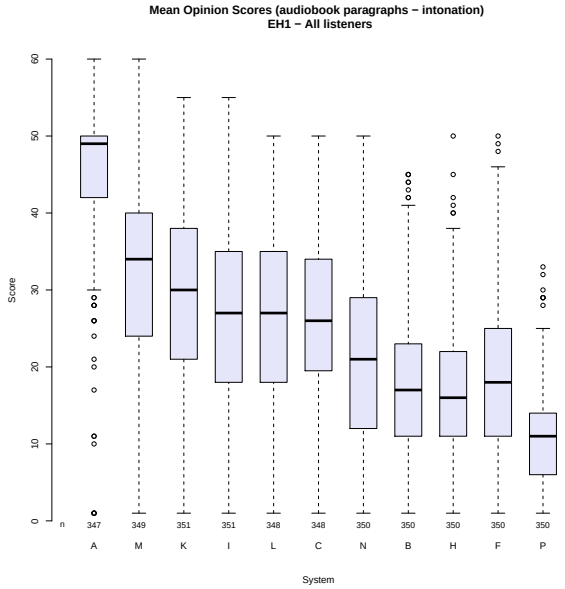
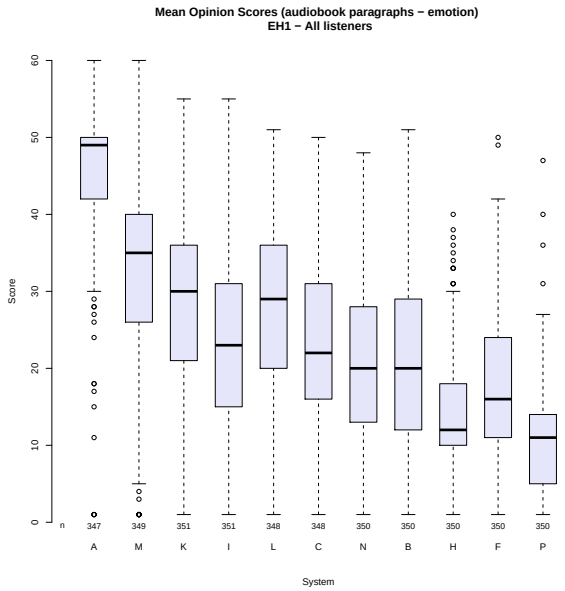


Figure 5: Results for task 2013-EH1 on paragraph test material, pooling all material, continued.





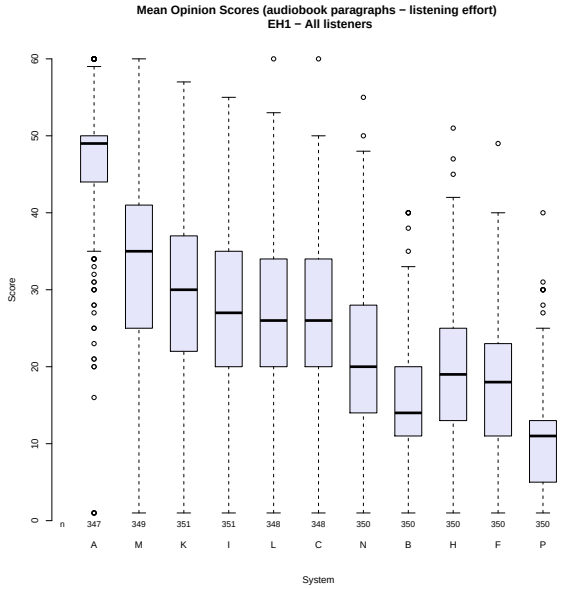


Figure 6: Results for task 2013-EH1 on paragraph test material, pooling all material, continued.

Language	Total
Catalan	1
Chinese (Mandarin)	8
Czech	2
Dutch	4
Finnish	1
French	3
German	15
Greek	5
Italian	1
Japanese	38
Norwegian	1
Romanian	2
Spanish	1
N/A	3

Table 5: First language of non-native speakers [3]

Gender	Male	Female
Total	107	133

Table 6: Gender [3]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	14	244	85	44	27	10	3	0

Table 7: Age of listeners whose results were used (completed the evaluation fully or partially) [4]

Native speaker	Yes	No
English	155	86

Table 8: Native speakers [3]

	Task 2013-EH1	Task 2013-EH2
EE	50	62
ER	92	112
ES	52	58
ALL	194	232

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) [4]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EE	112	0	0	112
ER	277	72	119	86
ES	143	33	62	48
ALL	**532**	**105**	**181**	**246**

Table 10: Listener registration and evaluation completion rates. [4]

	EH1_01	EH1_02	EH1_03	EH1_04	EH1_05	EH1_06	EH1_07	EH1_08	EH1_09	EH1_10	EH1_11
EE	5	5	5	5	5	5	4	4	4	4	4
ER	8	8	7	9	9	9	5	9	11	8	10
ES	3	4	5	4	6	4	6	4	5	6	5
ALL	16	17	17	18	20	18	15	17	20	18	19

Table 11: Listener groups for task 2013-EH1, showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations [4]

	EH2_01	EH2_02	EH2_03	EH2_04	EH2_05	EH2_06	EH2_07	EH2_08	EH2_09	EH2_10	EH2_12	EH2_13	EH2_14	EH2_15
EE	5	5	5	5	5	5	4	4	4	4	4	4	4	4
ER	8	7	8	5	9	6	6	9	10	9	10	6	9	10
ES	5	4	3	4	5	5	6	3	5	3	4	5	3	3
ALL	18	16	16	14	19	16	14	16	19	16	18	15	16	17

Table 12: Listener groups for task 2013-EH2, showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations [4] (group EH2_11 is missing: due to a typo, no listeners were assigned to it).

Listener Type	EE	ER	ES	ALL
Total	112	86	48	246

Table 13: Listener type totals for submitted feedback

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	21	35	83	66	32	2

Table 14: Highest level of education completed [3]

CS/Engineering person?	Yes	No
Total	120	118

Table 15: Computer science / engineering person [3]

Work in speech technology?	Yes	No
Total	78	163

Table 16: Work in the field of speech technology [3]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	18	37	44	60	52	13	15

Table 17: How often normally listened to speech synthesis before doing the evaluation [3]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	0	1	89	46	10	97

Table 18: Dialect of English of native speakers [3]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
Total	24	22	28	10	2

Table 19: Level of English of non-native speakers [3]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	209	17	8	5

Table 20: Speaker type used to listen to the speech samples [3]

Same environment?	Yes	No
Total	223	14

Table 21: Same environment for all samples? [3]

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	168	60	9	2	0

Table 22: Kind of environment when listening to the speech samples [3]

Number of sessions	1	2-3	4 or more
Total	158	59	23

Table 23: Number of separate listening sessions to complete all the sections [3]

Browser	Firefox	IE	Chrome	Opera	Safari	Mozilla	Other
Total	50	20	40	0	121	1	6

Table 24: Web browser used (The paid listeners -type EE- all did the test on Safari.) [3]

Similarity with reference samples	Easy	Difficult
Total	174	65

Table 25: Listeners' impression of their task in section(s) about similarity with original voice. [3]

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
Total	25	3	36

Table 26: Listeners' problems in section(s) about similarity with original voice. [3]

Number of times	1-2	3-5	6 or more
Total	210	25	1

Table 27: Number of times listened to each example in section(s) about similarity with original voice. [3]

Naturalness	Easy	Difficult
Total	199	37

Table 28: Listeners' impression of their task in MOS naturalness sections [3]

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	5	19	2	12

Table 29: Listeners' problems in MOS naturalness sections [3]

Number of times	1-2	3-5	6 or more
Total	219	14	0

Table 30: How many times listened to each example in MOS naturalness sections? [3]

Book passage	Easy	Difficult
Total	112	127

Table 31: Listeners' impression of their task in the sections involving book passages. [3]

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	86	2	41

Table 32: Listeners' problems in the sections involving book passages [3]

Number of times	1-2	3-5	6 or more
Total	217	16	0

Table 33: How many times listened to each example in the sections involving book passages? [3]

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	24	141	64	9

Table 34: Listeners' impressions of intelligibility task (SUS). [3]