The Blizzard Challenge 2011

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 2011 was the seventh annual Blizzard Challenge which was 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 English corpus was used: the 'Nancy' corpus provided by Lessac Technologies. In commone with previous challenges, participants had the option of using labels that were provided for the corpus and for the test sentences.

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

1. Introduction

The Blizzard Challenge, originally conceived by Black and Tokuda [1], is now well established, this paper only provides the specific details of the 2011 challenge. For background information, please refer to the previous summary papers for 2005 [1, 2], 2006 [3], 2007 [4], 2008 [5], 2009 [6] and 2010 [7]. These, and 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 [8].

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] and Blizzard 2010 had 17 participants. This year, 2011, the 9 participants listed in Table 1 took part.

Three benchmark systems were included 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 [9], and two HTS speaker-dependent systems built from 16kHz and 48kHz sampling rate waveforms respectively. [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 to D denoting the three benchmark systems and E to M denoting the systems submitted by participants in the challenge.

3. Voices to be built

3.1. Speech databases

The English data for voice building was provided by Lessac Technologies, who also participated in the challenge and suggested the inclusion of task ES1. The speaker is known as 'Nancy' and is a native speaker of US English, professional female voice talent, voice coach, and singer. 16.6 hours of data was made available to participants, comprising around 12k utterances. The data were provided at a 16kHz sampling rate as individual utterances, with the original long session files available at higher sampling rates (44kHz, 96kHz) . Submitted voices could be at any sampling rate and no resampling was done before the listening test. The natural speech used in the listening test was at 96kHz.

Short name	Details	Method
NATURAL	Natural speech from the same speaker as the corpus	human
FESTIVAL	The Festival unit-selection benchmark system [9]	unit selection
HTS	HTS 16kHz benchmark	HMM
HTS48k	HTS 48kHz benchmark	HMM
BUCEADOR	Aholab (UPV) & TALP (UPC)	hybrid
HELSINKI	Helsinki University & Aalto University	HMM
I2R	Institute for Infocomm Research (IR)	hybrid
ILSP	Institute for Language and Speech Processing	unit selection
LESSAC	Lessac Technologies	unit selection
NITECH	Nagoya Institute of Technology	HMM
PUB	Politehnica University of Bucharest	HMM
UCD	University College Dublin	unit selection
USTC	University of Science and Technology of China	hybrid

Table 1: The participating systems and their short names. The first four rows are the benchmarks and correspond to the system identifiers A to D in that order. The remaining rows are in alphabetical order of the system's short name and *not* the order E to M.

3.2. Tasks

Participants were asked to build several synthetic voices from the databases, in accordance with the rules of the challenge [10]. A hub and spoke design was again adopted this year. Task names start with E (for English) followed by either H (for hub) or S (for spoke) and finishing with a number denoting the subtask within that language & task, as listed in the following sections.

- EH1: build a voice from the UK English 'Nancy' database, using any sampling rate.[2]
- ES1: build a voice designed to read names and addresses (in US format) - the evaluation of this task will focus mainly on intelligibility.

[1]Many thanks to Keiichiro Oura & Shinji Takaki for constructing the HTS benchmarks and to Rob Clark for the Festival benchmark

[2]The original rules specified that waveforms would be resampled down to 16kHz for the listening test but in the event the organisers decided not to do this since only a few high sampling rate entries were received and a separate listening test for them was not justified.

Type	Source	Example
news	Glasgow Herald newspaper	He was taken to the Western Infirmary and later released.
novel	out-of-copyright novels	It was a blow in the face to Sheldon.
reportorial	newsreader-style	Most of the new additions were barely profitable, if not outright loss makers.
SUS	semantically unpredictable	The fire turned as the capital point.
address	Lessac	Six twenty-three South Whitehead Street, Key West, Florida, six three seven two one, dash nine three eight four.

Table 2: The sentence types used in the listening test.

Very few participants constructed a specific system for task ES1, and so a combined listening test for both tasks was devised.

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. The selection of which sentences to use in the listening tests was made as in 2008 / 2009 / 2010 – please see [5, 6, 7] for details. For details 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 almost all 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 all test sentences were available this year, including for the semantically unpredictable sentences and addresses. Table 2 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 UK English and aged about 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).
- Volunteers recruited via participating teams, mailing lists, blogs, etc. (ER).
- Speech experts, recruited via participating teams and mailing lists (ES).

Table 16, summarised in Table 3, 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 combined listening test for all submitted voices and both tasks had the following structure, comprising 8 sections, each with 13 stimuli presented:

1. Similarity, novel
2. Similarity, news
3. Naturalness, novel
4. Naturalness, news
5. Naturalness, reportorial
6. Intelligibility, address (task ES1) – multiple listens allowed
7. Intelligibility, SUS (task EH1) – single listen only
8. Intelligibility, SUS (task EH1) – single listen only

Within each numbered section of the listening test, a listener heard one example from each system. Great care was taken to ensure no listener heard the same sentence more than once – this is particularly important for testing intelligibility. The number of listeners obtained is shown in Table 3. See Table 15 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 3: 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 [11] 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 5.3 and Tables 10 to 36 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 EH1 for all listeners 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.

5.1. Task EH1 – general-purpose TTS

Naturalness results are given in Table 4. No synthesiser is as natural as the natural speech (Figure 1 and Table 6). System G is significantly more natural than all other synthesisers. System C is as intelligible as natural speech, when compared using SUS material (Figure 1 and Table 7) although a number of systems are not significantly less intelligible than system C.

5.2. Task ES1 – reading out addresses

This task only concerned intelligibility. No significant differences were found between any systems or natural speech (Table 9), presumably because of the ceiling effect caused by the task material.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.8	0.63	510	48
B	3	1.5	2.9	1.06	509	49
C	3	1.5	2.6	1.04	510	48
D	2	1.5	2.4	1.10	509	49
E	3	1.5	3.1	1.07	516	42
F	2	1.5	2.4	1.06	509	49
G	3	1.5	3.3	1.08	512	46
H	3	1.5	2.9	1.05	510	48
I	1	0.0	1.4	0.69	510	48
J	2	1.5	2.5	1.04	512	46
K	3	1.5	2.8	1.05	510	48
L	3	1.5	2.8	1.01	512	46
M	3	1.5	2.7	1.06	510	48

Table 4: Mean opinion scores for naturalness on task EH1 (general-purpose TTS). Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

The ceiling of intelligibility on this task appears to be around 13% WER (the fact that this is not closer to zero warrants further investigation: a possible cause might be the automatic scoring method itself). The WER of all systems including natural speech lie in a narrow range of approximately 13% to 16% – compare this to Table 8 in which we see that SUS material is able to differentiate much better between systems, with WER ranging from 17% to 29% (Figure 1).

5.3. 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 were submitted by all the listeners who completed the evaluation and included many detailed comments and suggestions from all listener types. Listener information and feedback is summarised in Tables 10 to 36.

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, Keiichiro Oura & Shinji Takaki built the benchmark systems. Tim Bunnell of the University of Delaware provide the tool to generate the SUS sentences for English. 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] "The Blizzard Challenge website," http://www.synsig.org/index.php/Blizzard_Challenge.

[9] 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.

[10] "Blizzard Challenge 2011 rules," http://www.synsig.org/index.php/Blizzard_Challenge_2011_Rules.

[11] 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 [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.

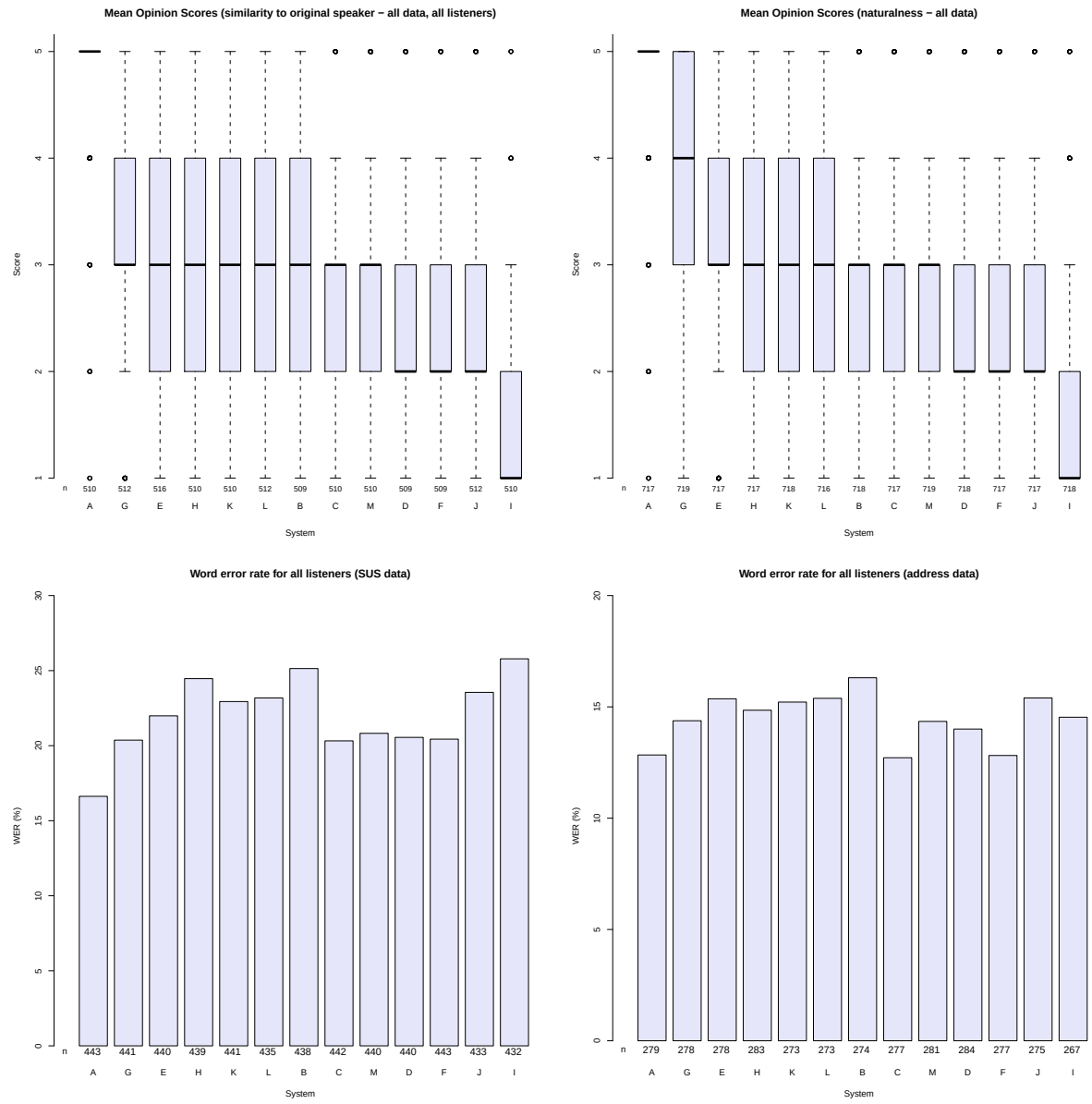

Figure 1: Results for tasks EH1 and ES1.

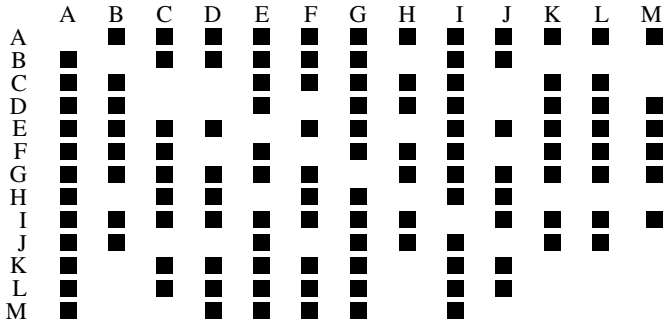

Table 5: Significant differences in similarity to the original speaker: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

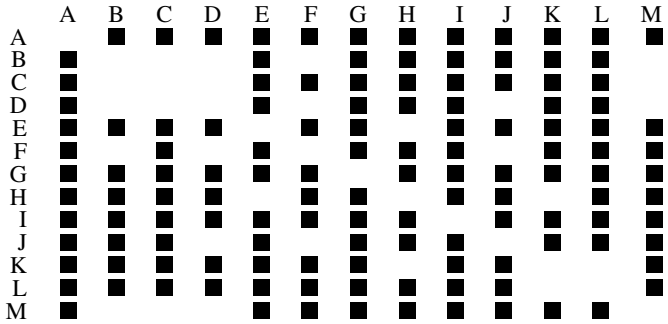

Table 6: Significant differences in naturalness: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

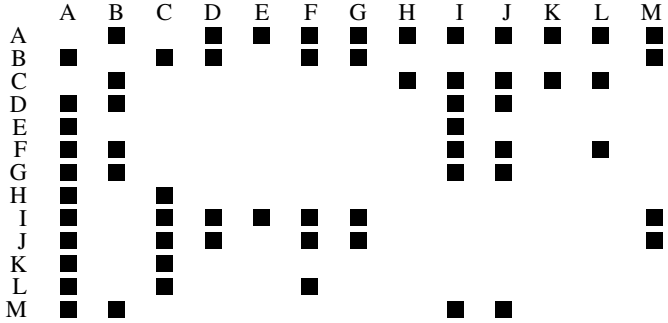

Table 7: Significant differences in intelligibility on all types of material: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

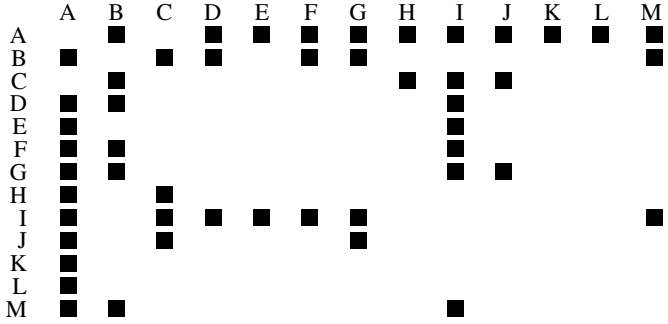

Table 8: Significant differences in intelligibility on Semantically Unpredictable Sentences: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

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

Table 9: Significant differences in intelligibility on address material: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems. Note that no significant differences were found in this section of the test.

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 10: First language of non-native speakers [3]

Gender	Male	Female
Total	131	91

Table 11: Gender [3]

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

Native speaker	Yes	No
English	122	101

Table 13: Native speakers [3]

	Task EH1
EE	104
ER	52
ES	124
ALL	280

Table 14: 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	104	0	0	104
ER	63	11	22	30
ES	154	30	33	91
ALL	**321**	**41**	**55**	**225**

Table 15: 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	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 16: 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 [4]

Listener Type	EE	ER	ES	ALL
Total	104	30	91	225

Table 17: 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 18: Highest level of education completed [3]

CS/Engineering person?	Yes	No
English total	132	92

Table 19: Computer science / engineering person [3]

Work in speech technology?	Yes	No
English total	100	123

Table 20: Work in the field of speech technology [3]

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

Table 21: How often normally listened to speech synthesis before doing the evaluation [3]

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

Table 22: Dialect of English of native speakers [3]

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

Table 23: Level of English of non-native speakers [3]

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

Table 24: Speaker type used to listen to the speech samples [3]

Same environment?	Yes	No
Total	220	4

Table 25: 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	162	46	12	0	3

Table 26: Kind of environment when listening to the speech samples [3]

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

Table 27: Number of separate listening sessions to complete all the sections [3]

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

Table 28: Web browser used (The paid listeners -type EE- all did the test on Safari.) [3]

Similarity with reference samples	Easy	Difficult
Total	145	77

Table 29: 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	46	1	30

Table 30: Listeners' problems in section(s) about similarity with original voice. [3]

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

Table 31: Number of times listened to each example in section(s) about similarity with original voice. [3]

Naturalness	Easy	Difficult
Total	176	47

Table 32: 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	11	23	0	13

Table 33: Listeners' problems in MOS naturalness sections [3]

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

Table 34: How many times listened to each example in MOS naturalness sections? [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	73	91	36	24

Table 35: Listeners' impressions of intelligibility task (addressess and SUS). [3]

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

Table 36: How many times listened to each example in the intelligibility section. (SUS could only be heard once.) [3]