The Blizzard Challenge 2010

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 2010 was the sixth annual Blizzard Challenge. As in 2008 and 2009, UK English and Mandarin Chinese were the chosen languages for the 2010 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. Two English corpora were used: the 'rjs' corpus provided by Phonetic Arts, and the 'roger' corpus from the University of Edinburgh. The Mandarin corpus was provided by the National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences. As usual, all participants (including those with limited resources or limited experience in these languages) had the option of using labels that were provided for both corpora and for the test sentences. The tasks were organised in the form of 'hubs' and 'spokes' where each hub task involved building a general-purpose voice and each spoke task involved building a voice for a specific situation or under specified conditions.

A set of test sentences was released to participants, who were given a limited time in which to synthesise them and submit the synthetic speech. An online listening test was conducted to evaluate naturalness, intelligibility and degree of similarity to the original speaker.

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

1. Introduction

Now that the Blizzard Challenge, originally conceived by Black and Tokuda [1], has been established for a number of years, we will confine ourselves in this paper to the specific details of the 2010 challenge. For more extensive details of the general arrangements of the Blizzard Challenge and how the listening test is conducted, please refer to the previous summary papers for 2005 [1, 2], 2006 [3], 2007 [4] 2008 [5] and 2009 [6]. Links to all of these papers, and to other useful Blizzard resources, including the rules of participation, anonymised releases of the submitted synthetic speech, the natural reference samples, raw 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 [7].

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] and Blizzard 2009 had 19 [6]. This year, there were 17 participants, listed in Table 1.

As usual, two types of systems were used as benchmarks, in an attempt to facilitate comparisons between the results from one year to another: a Festival-based unit selection system from CSTR configured very similarly to the Festival/CSTR entry to Blizzard 2006 [8], an HTS speaker-dependent system configured the same as the HTS entry to Blizzard 2005 [9] and the HTS speaker-adaptive system from Blizzard 2007 [10]. Note that the HTS entries were carefully constructed to be as similar to the respective 2005 and 2007 systems as possible, which involved reproducing bugs and all! Since this entailed a substantial amount of effort[1], it is planned that future challenges will employ as benchmarks a speaker-dependent and a speaker-adaptive system, each con-

System short name	Details
NATURAL	Natural speech from the same speaker as the corpus
FESTIVAL	The Festival unit-selection benchmark system [8]
HTS2005-auto	A speaker-dependent HMM-based benchmark system with automatically produced labels [9]
HTS2005-hand	A speaker-dependent HMM-based benchmark system with hand-corrected labels [9]
HTS2007	A speaker-adaptive HMM-based benchmark system [10]
CASIA	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
CMU	Carnegie Mellon University, USA
CSTR	The Centre for Speech Technology Research, University of Edinburgh, UK
HELSINKI	Helsinki University, Finland
I2R	Institute for Infocomm Research (I^2R), Singapore
ILSP	Institute for Language and Speech Processing, Greece
LESSAC	Lessac Technologies, USA
MERAKA	Meraka Institute, South Africa
MODELTALKER	University of Delaware, USA
MSRA	Microsoft Research Asia, China
NICT	National Institute of Information and Communications Technology, Japan
NIT	Nagoya Institute of Technology, Japan
NOKIA	Nokia Research Center, China
NTNU	Norwegian University of Science and Technology, Norway
NTUT	National Taipei University of Technology, Taiwan
USTC	iFlytek Speech Lab, University of Science and Technology of China
VUB	Vrije Universiteit, Belgium

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

[1]Many thanks to Junichi Yamagishi of CSTR for constructing the HTS benchmarks and to Volker Strom of CSTR for the Festival benchmarks

System	EH1	EH2	ES1	ES2	ES3	MH1	MH2	MS1	MS2
NATURAL	X	X	X	X	X	X	X	X	X
FESTIVAL	X	X		X					
HTS2005-auto	X	X	X	X	X	X	X	X	X
HTS2005-hand		X							
HTS2007			X		X				
CASIA		X				X			
CMU	X	X	X	X		X	X		X
CSTR	X	X	X	X	X				
HELSINKI	X	X	X	X		X	X		X
I2R	X	X		X	X	X	X	X	X
ILSP	X	X		X	X	X	X	X	X
LESSAC	X								
MERAKA	X	X		X					
MODELTALKER	X	X		X					
MSRA	X	X				X	X		
NICT	X	X							
NIT	X	X	X	X	X	X	X	X	X
NOKIA	X	X				X	X		
NTNU	X								
NTUT						X	X	X	X
USTC	X	X	X	X	X				
VUB	X	X		X	X	X	X		X

Table 2: The tasks completed by each participating system. The first five systems are the benchmarks and correspond to the system identifiers A to E in that order. The remaining rows are in alphabetical order of the system's short name and *not* the order F to V

structed with whatever is then the current version of HTS rather than with a recreated historical version.

The tasks completed by each participant are shown in Table 2. As in previous years, a number of additional groups (not listed here) registered for the Challenge and obtained the corpora, but did not submit samples for evaluation. When reporting anonymised results, the systems are identified using letters, with A denoting natural speech, B to E denoting the four benchmark systems and F to V 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 from two sources. Phonetic Arts released 4014 utterances of speech from their 'rjs' speaker, a professional 50-year old male speaker with an RP accent. The recordings were made in a commercial voiceover studio, and microphone used was an Audio Technika AT4033A, into a TL Audio PA5001 preamp, with A/D conversion performed using a Yamaha 03/D into a Digidesign 002 / ProTools 8.0 recording system. An accent-specific pronunciation dictionary, and Festival utterance files created using this dictionary, were available under a separate licence. These data were used in the main hub task for English (EH1) and two spoke tasks (ES2 and ES3).

The Centre for Speech Technology Research, University of Edinburgh, UK released the ARCTIC set from their 'roger' corpus – a speaker who has been used in previous challenges. The recordings were made several years ago, in the university's older recording facility and the microphone used was an AKG CK98 hypercardoid powered by a SE300B power module. Participants were forbidden from using any other 'roger' speech data (e.g., obtained through participation in an earlier challenge) in the construction of their entry. Participants were able to download 1132 sentences (i.e., about one hour) of recordings of this UK English male speaker with a fairly standard RP accent. These data were used in tasks EH2 and ES1. An accent-specific pronunciation dictionary, and Festival utterance files created using this dictionary, were also available under a separate licence. In addition, hand-corrected labels supplied by iFLYTEK were released for the ARCTIC subset of corpus, for optional use in the second hub task for English, EH2.

The National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences, released a corpus of recordings of a female professional radio broadcaster with a standard Beijing dialect of Mandarin Chinese. This 5884 sentence dataset is a superset of the Mandarin corpus used in the Blizzard Challenge 2008. These data were used in all tasks for Mandarin this year.

3.2. Tasks

Participants were asked to build several synthetic voices from the databases, in accordance with the rules of the challenge [11]. A hub and spoke design was again adopted this year. Task names start with either E (for English) or M (for Mandarin), 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.

3.2.1. English tasks

- EH1: build a voice from the UK English 'rjs' database. You may use either the 16kHz or 48kHz versions, but the submitted wav files must be at 16kHz sampling rate. (4014 utterances)
- EH2: build a voice from the ARCTIC portion of the UK English 'roger' database, optionally using the provided hand-corrected labels. You may use either the 16kHz or 48kHz versions, but the submitted wav files must be at 16kHz sampling rate. (1132 utterances)
- ES1: build voices from the first 100 utterances of the 'roger' database. You may use voice conversion, speaker adaptation techniques or any other technique you like. You may use either the 16kHz or 48kHz versions, but the submitted wav files must be at 16kHz sampling rate.
- ES2: build a voice from the 'rjs' database suitable for synthesising speech to be heard in the presence of additive noise. The evaluation of this task will focus on intelligibility only. We will not consider naturalness or speaker similarity. You may enter the same voice as task EH1 if you wish, although specially-designed voices are strongly encouraged. You may use either the 16kHz or 48kHz versions, but the submitted wav files must be at 16kHz sampling rate.
- ES3: the same as EH1, but you must submit 48kHz sampling rate wav files.

3.2.2. Mandarin tasks

- MH1: build a voice from the full Mandarin database (5884 utterances)
- MH2: build a voice from utterances 5085 to 5884 of the full Mandarin database (800 utterances)
- MS1: build a voice from the first 100 of the utterances used in MH2, i.e. utterances 5085 to 5184.
- MS2: build a voice from the full Mandarin database suitable for synthesising speech to be heard in the presence of additive noise. The evaluation of this task will focus on intelligibility only. We will not consider naturalness or speaker similarity. You may enter the same voice as task MH1 or MH2 if you wish, although specially-designed voices are strongly encouraged.

3.3. Additive noise for tasks ES2 and MS2

For the ES2 and MS2 tasks, participants were *not* informed in advance about the type or SNR of the noise, but merely instructed to build voices which were designed to maintain intelligibility in the presence of additive noise. They submitted clean samples and noise was added by the organisers prior to the listening test.

The organisers elected to use 6-speaker speech babble-shaped noise, which was obtained from the International Collegium of Rehabilitative Audiology (ICRA) noise CD [12]. We used track 9, as described at `http://www.icra.nu/Prod_Noise.html`, which contains aperiodic noise (with no harmonic content) that follows the modulations of speech babble both spectrally and temporally. A single channel was selected from the original stereo file and downsampled to 16kHz. Sections of the same duration as each test sentence were extracted from a variety of points in the long file provided on the CD, ensuring that the same sentence (regardless of system) always had the same section of noise added. This was necessary because of the temporal variation in the noise, to be sure of fair comparisons across systems.

The "active speech level" of the submitted files was normalised using the method in ITU P.56, implemented in a Matlab script from [13]. The noise level was set using the same ITU P.56 method, and the noise was added to speech to achieve SNRs of 0dB, -5dB, -10dB. Note that these are all fairly challenging SNRs, ranging from "equal level" speech and noise at 0dB to "much higher level noise than speech" at -10dB. Informal pilot testing by the organisers was used to arrive at this choice of SNRs, avoiding floor or ceiling effects on intelligibility, whilst providing a sufficiently challenging setting that would reveal differences between systems.

The creator of the ICRA noise CD, Prof. W. Dreschler, has generously allowed the redistribution of this noise signal, and therefore it is included in a distribution of the submitted synthetic speech and the raw listening test results, which is publicly available via the Blizzard website. The mixed speech+noise signals are also included in this distribution.

3.4. Listening test design and materials

The participants were asked to synthesise several hundred 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 and 2009 – please see [5, 6] 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 used this year, for both languages and all speakers, including semantically unpredictable sentences.

3.5. 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:

- Volunteers recruited via participating teams, mailing lists, blogs, etc. (ER).
- Speech experts, recruited via participating teams and mailing lists (ES).
- 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 (EU).

For Mandarin, the following listener types were used:

- Paid native speakers of Mandarin, aged 18-25, recruited in China using a commercial testing organisation, who carried out the test in a quiet supervised lab using headphones (MC).
- Paid undergraduate native speakers of Mandarin aged about 20-25. These were recruited in Edinburgh and carried out the test in purpose-built soundproof listening booths using good quality audio interfaces and headphones (ME).
- Volunteers, recruited via participating teams, mailing lists, etc. (MR).
- Speech Experts, recruited via participating teams and mailing lists (MS).

Tables 37 to 45, summarised in Table 5, show the number of listeners of each type obtained for each of the listening tests listed in Tables 3 and 4.

3.6. Listening tests

When using paid listeners, it is convenient to construct a listening test designed to take 45-60 minutes, rather than many short tests. Therefore, tasks were combined into listening tests, whilst ensuring that no listener ever heard the same sentence twice. This was fairly challenging this year, given the large number of tasks and systems. Tables 3 and 4 show the five independent listening tests that were created and run in parallel for this year's Blizzard Challenge. Each listener performed one of the three English tests or one of the two Mandarin tests (or, possibly one English test *and* one Mandarin test). Each test followed the same general design, although the number and type of sections varied, as described in the tables. Within each numbered section of a listening test, the listener generally heard one example from each system. Note that the number of systems involved in each task varies; where there were more systems, and therefore larger Latin Squares, fewer sections could be included in the corresponding listening test. Great care was taken to ensure no listener heard the same sentence more than once – this is particularly important for testing intelligibility.

3.6.1. Number of listeners

The number of listeners obtained is shown in Table 5. See Table 46 for a detailed breakdown of evaluation completion rates for each listener type. As in previous years, the higher completion rate for Mandarin listeners is a consequence of the higher proportion of paid listeners (i.e., the difficulty of obtaining large numbers of volunteer listeners for this language).

4. Analysis methodology

As in previous years, we pooled 'completed all sections' and 'partially completed' listeners together in all analyses. Here, we present only results for all listener types combined. Analysis by listener type was provided to participants and can now be obtained by non-participants by downloading the complete listening test results via the Blizzard website. Please refer to [14] for a description of the statistical analysis techniques used and justification of the statistical significance techniques employed. As usual, system names are anonymised in all distributed results. See Section 6.1 and Tables 32 to 68 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

Section number	Tasks being evaluated	Type
Test name: EH1 + ES2		
1	EH1	SIMnews
2	EH1	MOSnews
3	EH1	MOSnovel
4	EH1	WERsus (clean)
5	ES2	WERbroadcast0dB
6	ES2	WERbroadcast-5dB
7	ES2	WERbroadcast-10dB
8	ES2	WERsus0dB
9	ES2	WERsus-5dB
10	ES2	WERsus-10dB
Test name: EH2 + ES2		
1	EH2	SIMnews
2	EH2	MOSnews
3	EH2	MOSnovel
4	EH2	WERsus (clean)
5	ES2	WERbroadcast0dB
6	ES2	WERbroadcast-5dB
7	ES2	WERbroadcast-10dB
8	ES2	WERsus0dB
9	ES2	WERsus-5dB
10	ES2	WERsus-10dB
Test name: ES3 + ES1 + ES2		
1	ES3	SIMnews
2	ES3	SIMnews
3	ES3	MOSnews
4	ES3	MOSnovel
5	ES3	WERsus (clean)
6	ES1	SIMnews
7	ES1	SIMnews
8	ES1	MOSnews
9	ES1	MOSnovel
10	ES1	WERsus (clean)
11	ES2	WERbroadcast0dB
12	ES2	WERbroadcast-5dB
13	ES2	WERbroadcast-10dB
14	ES2	WERsus0dB
15	ES2	WERsus-5dB
16	ES2	WERsus-10dB

Table 3: The three listening tests conducted for English.

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 for a particular task. This ordering is in descending order of mean naturalness. 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 anything about which systems are significantly better than other systems.

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. We will instead highlight only those results we think are noteworthy.

5.1. Task EH1 – full database

Naturalness results are given in Table 6. No synthesiser is as natural as the natural speech (Figure 1 and Table 12). System M is significantly more natural than all other synthesisers. Systems J and R are as intelligible as natural speech (Figure 2 and Table 13), although not significantly more natural than many other systems.

Section number	Tasks being evaluated	Type
Test name: MH1		
1	MH1	SIM
2	MH1	MOSnews
3	MH1	MOSnews
4	MH1	WERsus (clean)
5	MH1	WERsus (clean)
6	MH1	WERsus (clean)
Test name: MH2		
1	MH2	SIM
2	MH2	MOSnews
3	MH2	MOSnews
4	MH2	MOSnews
5	MH2	WERsus (clean)
6	MH2	WERsus (clean)
Test name: MS2 + MS1		
1	MS2	WERnews0dB
2	MS2	WERnews-5dB
3	MS2	WERnews-10dB
4	MS2	WERsus0dB
5	MS2	WERsus-5dB
6	MS2	WERsus-10dB
7	MS1	SIM
8	MS1	SIM
9	MS1	MOSnews
10	MS1	MOSnews
11	MS1	WERsus (clean)
12	MS1	WERsus (clean)

Table 4: The two listening tests conducted for Mandarin.

	English	Mandarin
Total registered	495	311
of which:		
Completed all sections	363	261
Partially completed	74	28
No response at all	58	22

Table 5: Number of listeners obtained

5.2. Task EH2 – modest size (ARCTIC) database

Naturalness results are given in Table 7. Again, no synthesiser is as natural as the natural speech (Figure 2 and Table 15). System M is significantly more natural than all other synthesisers. No system is as intelligible as natural speech (Figure 2 and Table 13) and there is no clear leader amongst the synthesisers.

Comparing systems C (HTS 2005 using automatic labels) and D (otherwise identical to C but using the supplied hand-corrected labels) allows some insight into the benefits of hand-correcting labels. There is no difference in naturalness or speaker similarity between these systems. There is a small, but insignificant, improvement in intelligibility for system D over system C. The overall benefits of hand-corrected labels appear to be small in this case.

5.3. Task ES1 – very small database

No synthesiser is as natural as the natural speech (Figure 1 and Table 12) but systems M & V are significantly more natural than all other synthesisers. System R is as intelligible as natural speech (Figure 2 and Table 13), although not significantly more natural than most other systems.

System	median	MAD	mean	sd	n	na
A	5	0	4.8	0.45	326	28
B	3	1.5	3	1.09	325	29
C	2	1.5	2.5	1.05	326	28
F	3	1.5	3.3	1.08	325	29
G	3	1.5	2.6	1.07	326	28
H	3	1.5	2.6	1.14	326	28
J	4	1.5	3.8	0.94	325	29
L	2	1.5	2.1	0.91	326	28
M	4	1.5	4.2	0.86	325	29
N	3	1.5	2.6	1.04	327	27
O	2	1.5	1.9	0.93	326	28
P	3	1.5	3	1.09	326	28
Q	1	0	1.6	0.78	325	29
R	3	1.5	2.7	1.05	325	29
S	3	1.5	3.1	1.16	326	28
T	4	1.5	3.7	1.07	327	27
U	3	1.5	2.7	1.11	325	29
V	3	1.5	3.3	1.09	326	28

Table 6: Mean opinion scores for task EH1 (full data set) on the combined results from sections 2 and 3 of the EH1+ES2 listening test. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

System	median	MAD	mean	sd	n	na
A	5	0	4.8	0.49	338	24
B	3	1.5	2.9	1.11	338	24
C	3	1.5	2.7	0.97	338	24
D	3	1.5	2.6	1.01	338	24
G	3	1.5	2.8	1.02	339	23
H	3	1.5	2.8	1.13	338	24
J	3	1.5	3.4	0.98	338	24
K	2	1.5	1.8	0.93	338	24
L	2	1.5	2.1	0.93	338	24
M	4	1.5	3.9	0.93	338	24
N	3	1.5	2.7	1.07	338	24
O	2	1.5	2	1	338	24
P	3	1.5	3.1	1.07	338	24
Q	1	0	1.7	0.86	339	23
R	3	1.5	2.9	1.02	338	24
S	3	1.5	3.1	1.08	338	24
U	3	1.5	2.8	1.12	339	23
V	4	1.5	3.5	0.97	339	23

Table 7: Mean opinion scores for task EH2 (ARCTIC data set) on the combined results from sections 2 and 3 of the EH2+ES2 listening test. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

5.4. Task ES2 – speech in noise

Results are presented broken down by sentence type in Figure 4 and by SNR in Figure 5. There is a clear leader in terms of intelligibility (which was the only evaluation metric of concern for this task) – system N is significantly better than all other synthesisers and natural speech at SNRs of -5dB and -10dB (Tables 18 and 19). In the least challenging condition of 0dB SNR, system N is as intelligible as natural speech (Table 20), but now not significantly better than system V.

5.5. Task ES3 – higher sampling rate

No synthesiser is as natural as the natural speech (Figure 6 and Table 21) but system M is significantly more natural than all other synthesisers. Several systems appear to be intelligible as natural speech, although the sample size (number of listeners) is relatively small so caution should be exercised about this finding.

5.6. Task MH1 – full database

Naturalness results are given in Table 9. No synthesiser is as natural as natural speech (Figure 7 and Table 23), and no single system stands above the rest. System C & J appear to be as intelligible as natural speech, but not significantly more intelligible than several other systems.

System	median	MAD	mean	sd	n	na
A	5	0	4.5	0.84	128	6
C	4	0.74	3.9	0.86	128	6
H	4	1.48	3.5	0.97	128	6
I	2	1.48	2.4	1.12	128	6
J	4	0	3.9	0.86	128	6
K	3	1.48	3.1	1.01	128	6
L	3	1.48	2.9	0.89	128	6
N	4	1.48	3.5	0.9	128	6
P	3	1.48	3	1.11	128	6
Q	1	0	1.4	0.73	128	6
R	4	1.48	3.9	0.93	128	6
S	2	1.48	2.1	0.97	128	6

Table 9: Mean opinion scores for task MH1 on the combined results from sections 2 and 3 of the MH1 listening test. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded due to missing data)

5.7. Task MH2 – modest size database

Naturalness results are given in Table 10. No system is as natural as natural speech (Figure 8 and Table 26). As in MH1, systems C & J appear to be as intelligible as natural speech, but not significantly more intelligible than several other systems.

5.8. Task MS1 – very small database

No system is as natural as natural speech (Figure 9 and Table 28). Systems C & R are significantly more intelligible than all other synthesisers, although not as intelligible as natural speech.

5.9. Task MS2 – speech in noise

Intelligibility results are given as pinyin+tone error rate (PTER) and are presented broken down by sentence type in Figure 10 and by SNR in Figure 11.

As for English, there is a clear leader in terms of intelligibility (which was the only evaluation metric of concern for this task) – system N is significantly better than all synthesisers and natural speech, at SNRs of -5dB and -10dB (Tables 29 and 30). In the least challenging condition of 0dB SNR, system N is as intelligible as natural speech (Table 31).

	Year							
	2007		2008		2009		2010	
System	MOS	WER	MOS	WER	MOS	WER	MOS	WER
Natural	4.7	–	4.8	22	4.9	14	4.8	12
Festival	3.0	25	3.3	35	2.9	25	3.0	23
HTS 2005	–	–	2.9	33	2.7	23	2.5	18

Table 8: Comparing the results of some of the benchmark systems for English (main hub task, large database) across recent years of the Blizzard Challenge. MOS means mean naturalness score and WER means word error rate in percent using semantically unpredictable sentences (SUS). Note that the SUS in 2009 and 2010 were simpler than those in 2007 and 2008; for 2010, HTS 2005 used automatically produced labels and not the hand-corrected ones.

System	median	MAD	mean	sd	n	na
A	5	0	4.6	0.86	194	7
C	4	1.5	3.6	0.96	194	7
H	3	1.5	3.2	1.06	194	7
I	2	1.5	2	1.01	194	7
J	3	1.5	3.3	1.02	194	7
L	3	1.5	2.9	1.04	194	7
N	3	1.5	3.1	1.02	194	7
P	3	1.5	2.7	0.93	194	7
Q	1	0	1.5	0.81	194	7
R	4	1.5	3.4	1.05	194	7
S	2	1.5	1.9	1.01	194	7

Table 10: Mean opinion scores for task MH2 on the combined results from sections 2, 3 and 4 of the MH2 listening test. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded due to missing data)

6. Discussion

Finding a synthesisers that is as intelligible as natural speech is a milestone that was already passed in a previous challenge. We are now routinely finding one or two systems of this type. A new milestone that was passed for 2010 is finding a system that is *more* intelligible than natural speech: this was the case for system N this year, in the two more challenging noise conditions of ES2 and MS2.

Table 8 provides a comparison of the scores of the natural speech and the two principal benchmark synthesisers over several recent years. Without wanting to draw firm conclusions from this limited data, there appears to be a tendency for the naturalness (MOS) scores of the benchmark synthesisers to be slowly drifting downwards, which would indicate that the other participating systems are, on average, gradually improving year on year (MOS scales are relative and listeners normalise their use of the scale according to the systems present). This could be due to a gradual increase in the quality of individual entries to the challenge, or perhaps simply the reduction in the number of poor quality entries. The naturalness rating of natural speech remains fairly constant (at close to 5, obviously). The ratio in WER between the benchmark synthesisers and the natural speech is also stable (at around 1.5 to 2), as would be expected.

6.1. 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 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 32 to 68.

7. 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; Volker Strom and Junichi Yamagishi provided the benchmark systems. Tim Bunnell of the University of Delaware provide the tool to generate the SUS sentences for English. Richard Sproat provided the Mandarin SUS. The listening test scripts are based on earlier versions provided by previous organisers of the Blizzard Challenge. Thanks to all participants and listeners.

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

[8] 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.

[9] Heiga Zen and Tomoki Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Proc. Blizzard Workshop*, 2005.

[10] Junichi Yamagishi, Heiga Zen, Tomoki Toda, and Keiichi Tokuda, "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the blizzard challenge 2007," in *Proc. Blizzard Workshop*, 2007.

[11] "Blizzard Challenge 2010 rules," http://www.synsig.org/index.php/Blizzard_Challenge_2010_Rules.

[12] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment," *Audiology*, vol. 40, pp. 148–157, 2001.

[13] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[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.

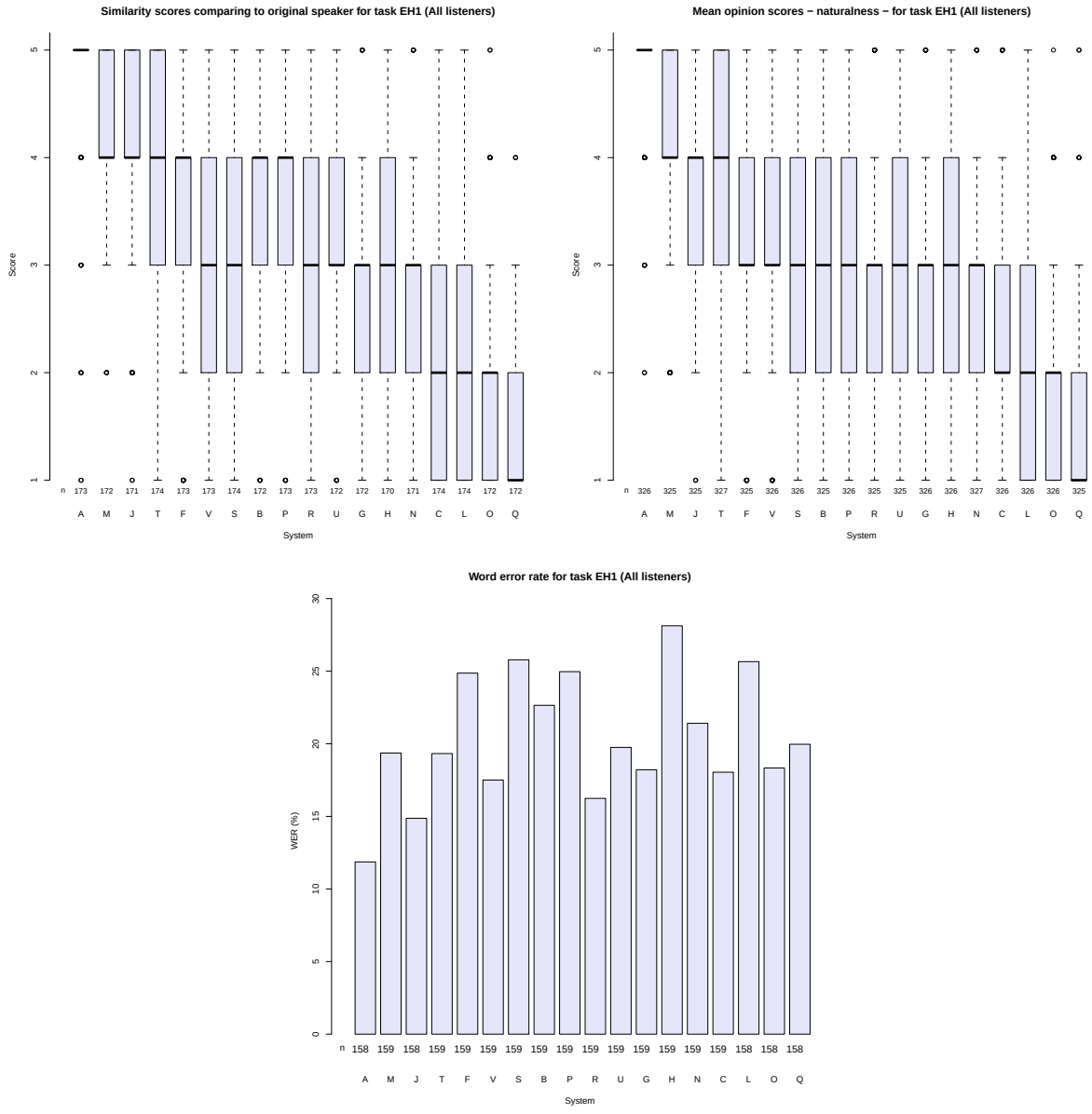

Figure 1: Results for task EH1.

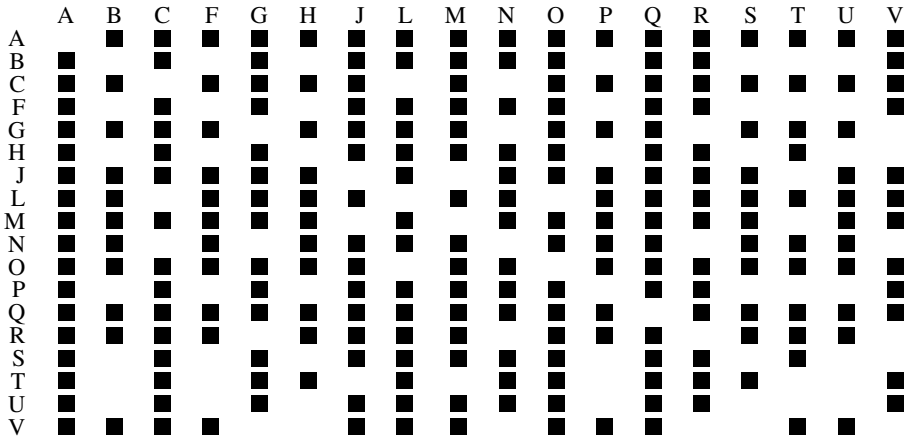

Table 11: Significant differences in similarity to the original speaker for task EH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

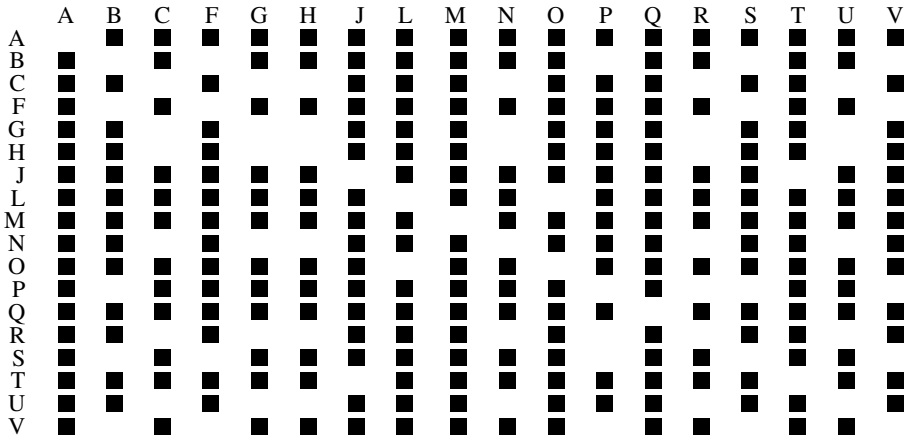

Table 12: Significant differences in naturalness for task EH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

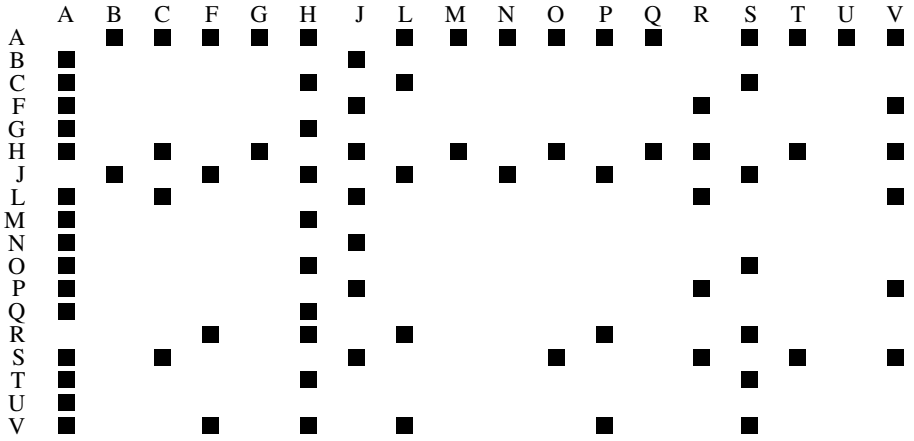

Table 13: Significant differences in intelligibility for task EH1: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

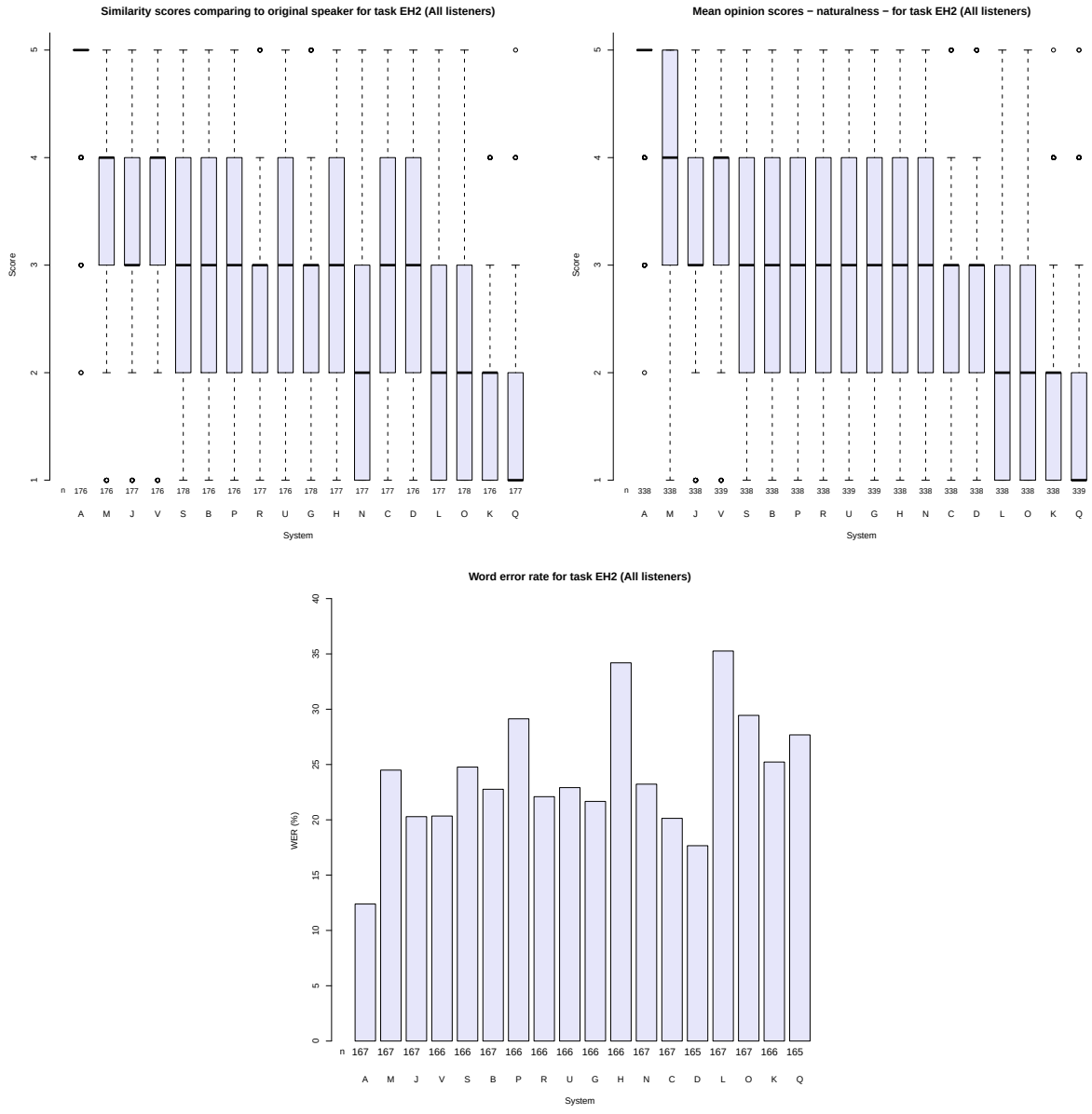

Figure 2: Results for task EH2.

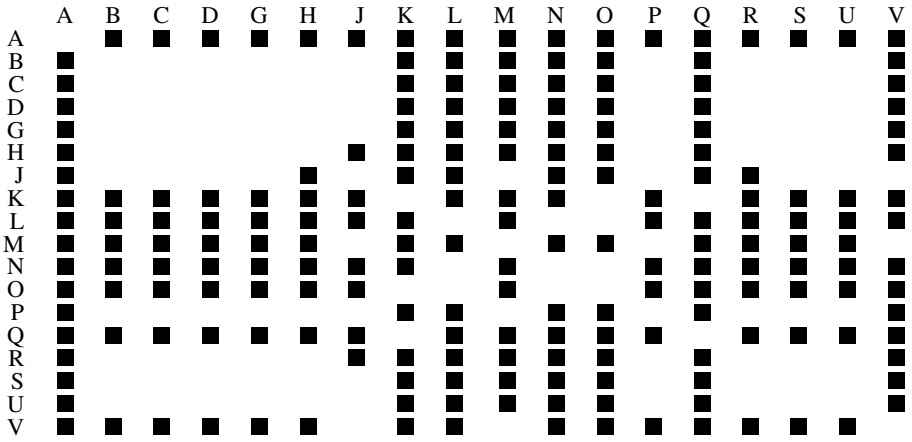

Table 14: Significant differences in similarity to the original speaker for task EH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

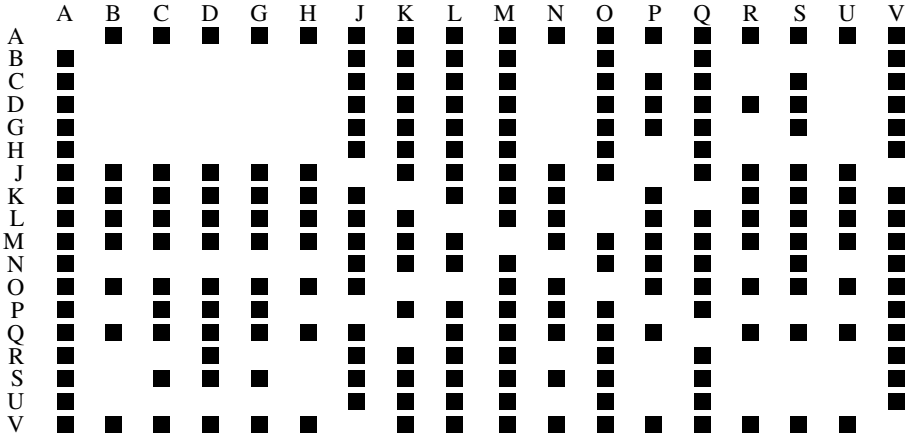

Table 15: Significant differences in naturalness for task EH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

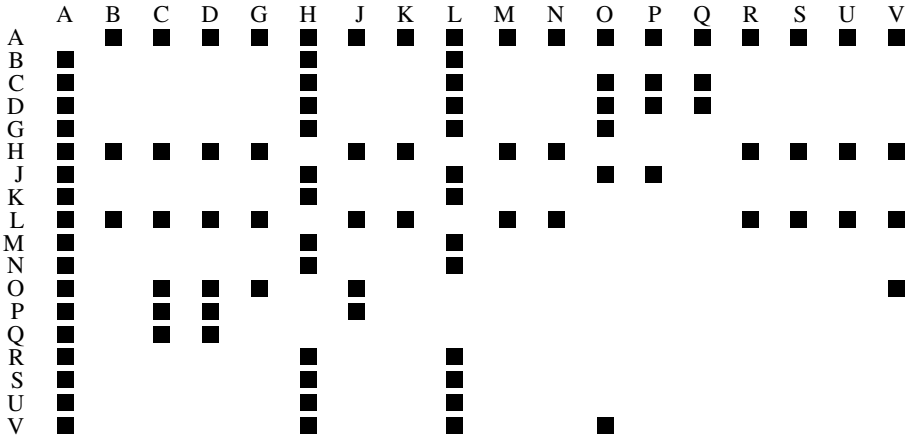

Table 16: Significant differences in intelligibility for task EH2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

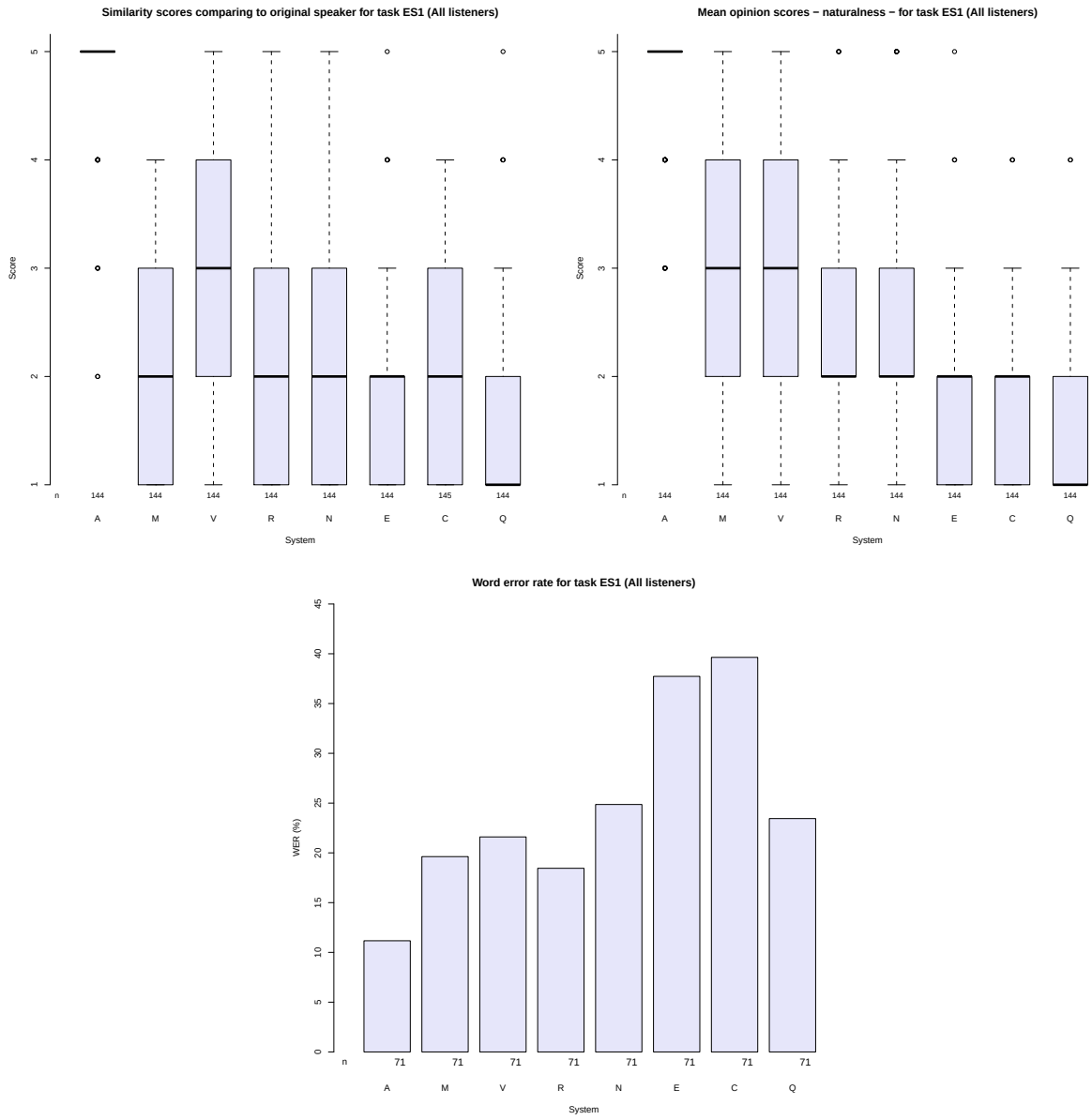

Figure 3: Results for task ES1.

	A	C	E	M	N	Q	R	V
A		■	■	■	■	■	■	■
C	■							■
E	■							■
M	■					■		■
N	■					■		■
Q	■			■	■		■	■
R	■					■		■
V	■	■	■	■	■	■	■	

	A	C	E	M	N	Q	R	V
A		■	■	■	■		■	■
C	■			■	■		■	■
E	■			■	■	■	■	■
M	■	■	■		■	■	■	
N	■	■	■	■		■		■
Q	■		■	■	■		■	■
R	■	■	■	■		■		■
V	■				■	■	■	

	A	C	E	M	N	Q	R	V
A		■	■	■	■	■		■
C	■			■	■	■	■	■
E	■			■		■	■	■
M	■	■	■					
N	■	■						
Q	■	■	■					
R		■	■					
V	■	■	■					

Table 17: Significant differences in (from left to right) similarity to the original speaker, naturalness and intelligibility for task ES1: results of pairwise Wilcoxon signed rank tests between systems' scores. ■ indicates a significant difference between a pair of systems.

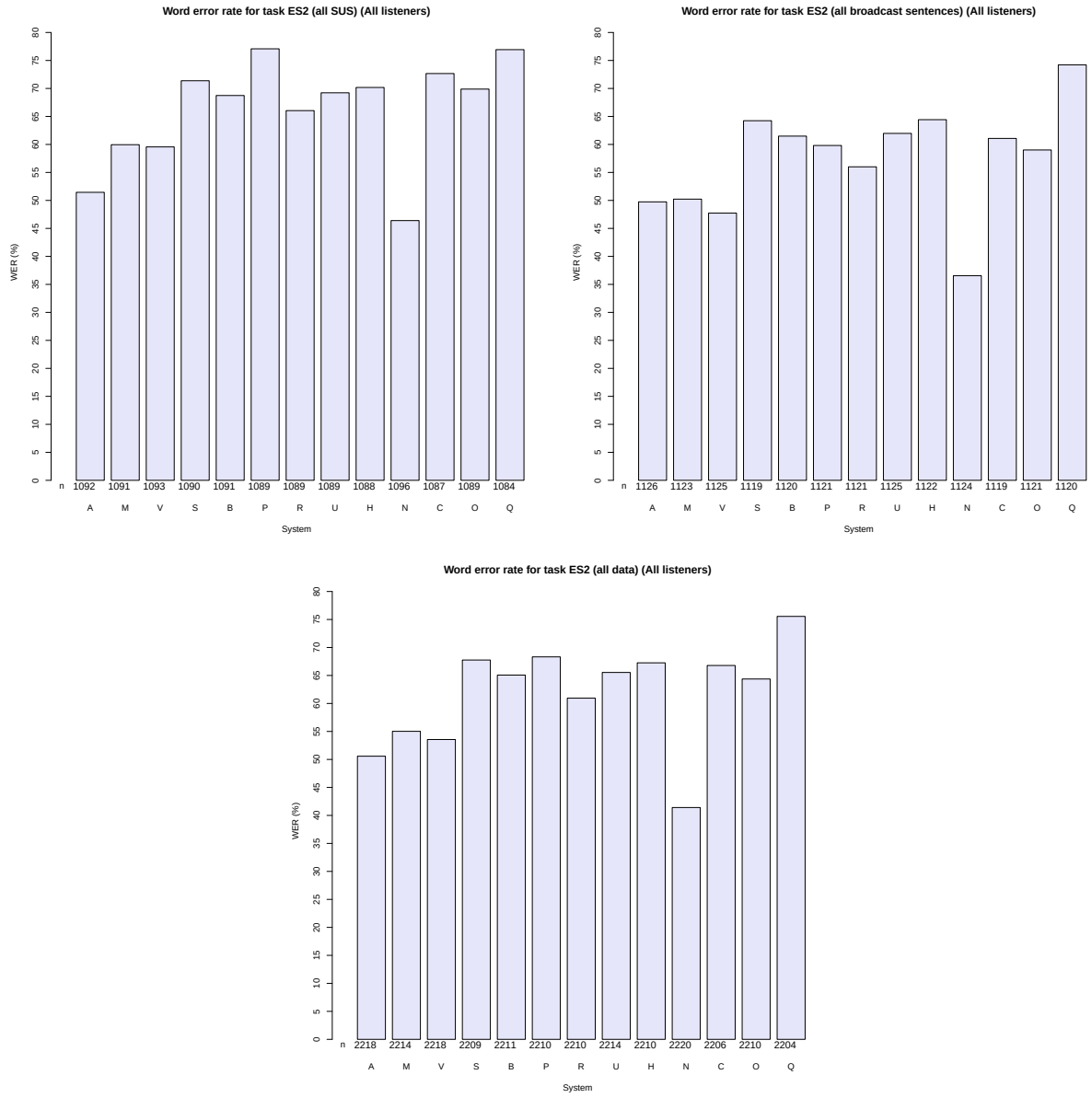

Figure 4: Intelligibility results for task ES2 per sentence type, pooled across all noise conditions.

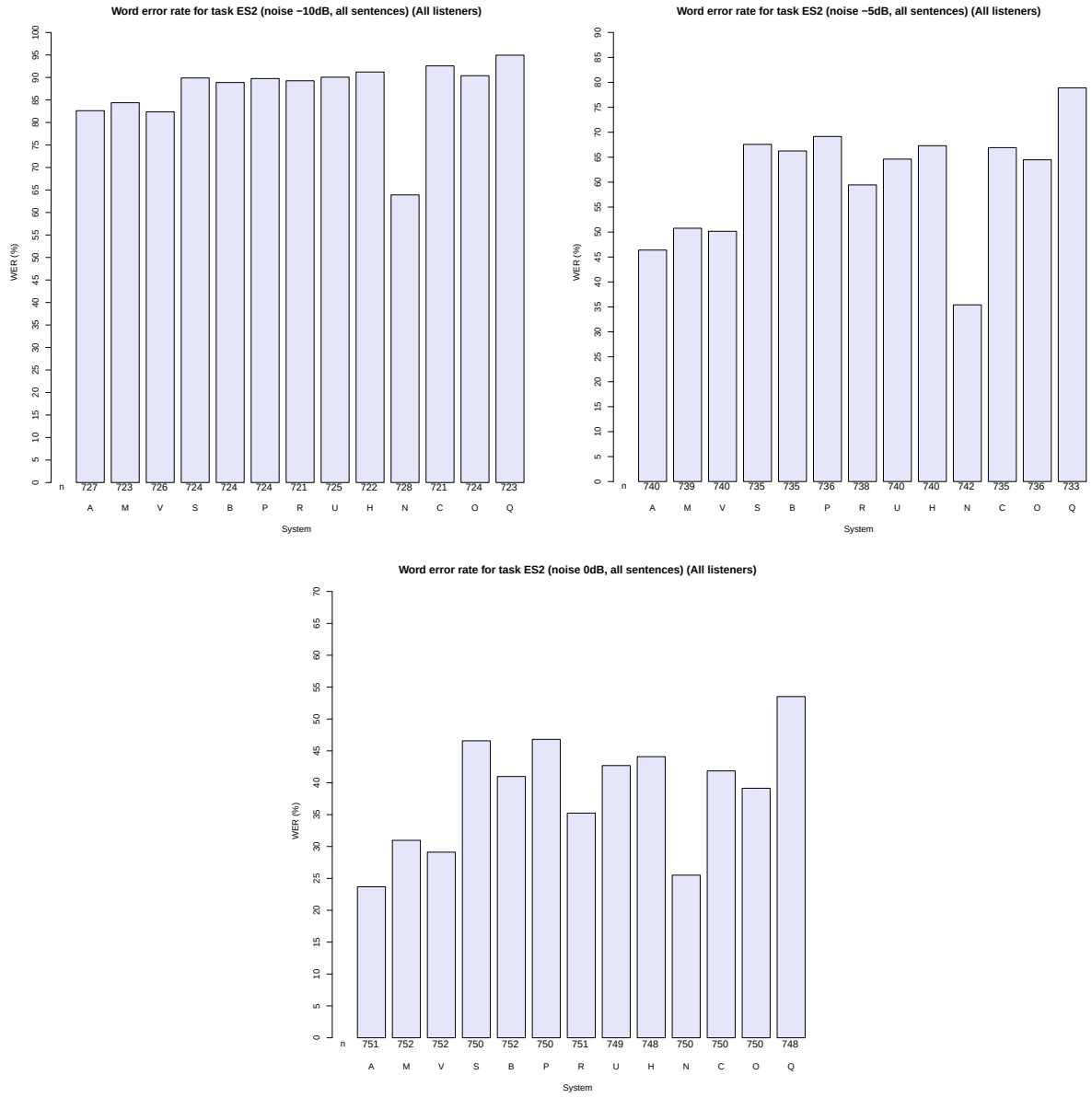

Figure 5: Intelligibility results for task ES2 per noise condition, pooled across both sentence types.

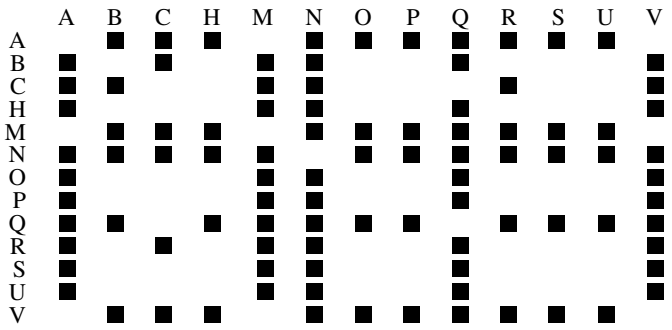

Table 18: Significant differences in intelligibility pooled across both sentence types at SNR of -10dB for task ES2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

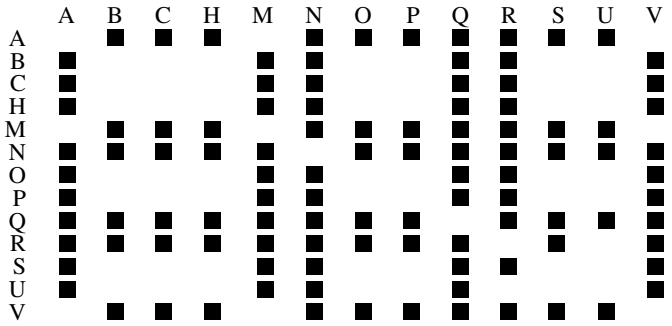

Table 19: Significant differences in intelligibility pooled across both sentence types at SNR of -5dB for task ES2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

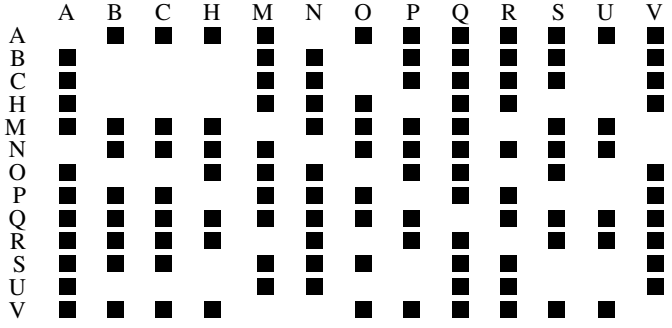

Table 20: Significant differences in intelligibility pooled across both sentence types at SNR of 0dB for task ES2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

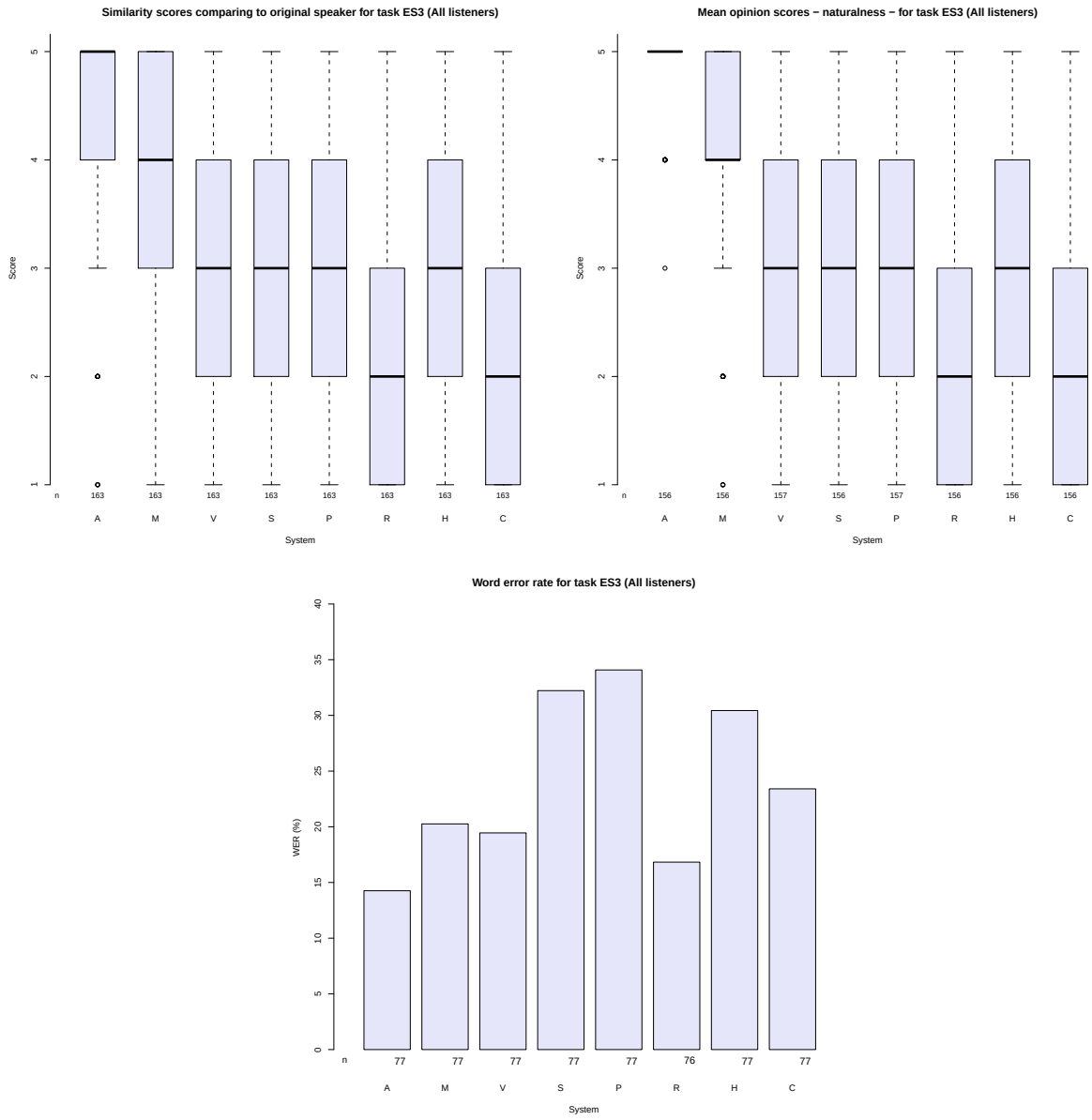

Figure 6: Results for task ES3.

	A	C	H	M	P	R	S	V
A		■	■	■	■	■	■	■
C	■		■	■	■		■	■
H	■	■		■	■	■		
M	■	■	■		■	■	■	■
P	■	■	■	■		■		■
R	■		■	■	■		■	■
S	■	■		■		■		
V	■	■		■	■	■		

	A	C	H	M	P	R	S	V
A		■	■	■	■	■	■	■
C	■		■	■	■		■	■
H	■	■		■		■		
M	■	■	■		■	■	■	■
P	■	■		■		■		
R	■		■	■	■		■	■
S	■	■		■		■		
V	■	■		■		■		

	A	C	H	M	P	R	S	V
A			■		■		■	
C								
H	■			■		■		■
M	■		■		■		■	
P	■			■		■		■
R	■		■		■		■	
S	■			■		■		■
V			■		■		■	

Table 21: Significant differences in (from left to right) similarity to the original speaker, naturalness and intelligibility for task ES3: results of pairwise Wilcoxon signed rank tests between systems' scores. ■ indicates a significant difference between a pair of systems.

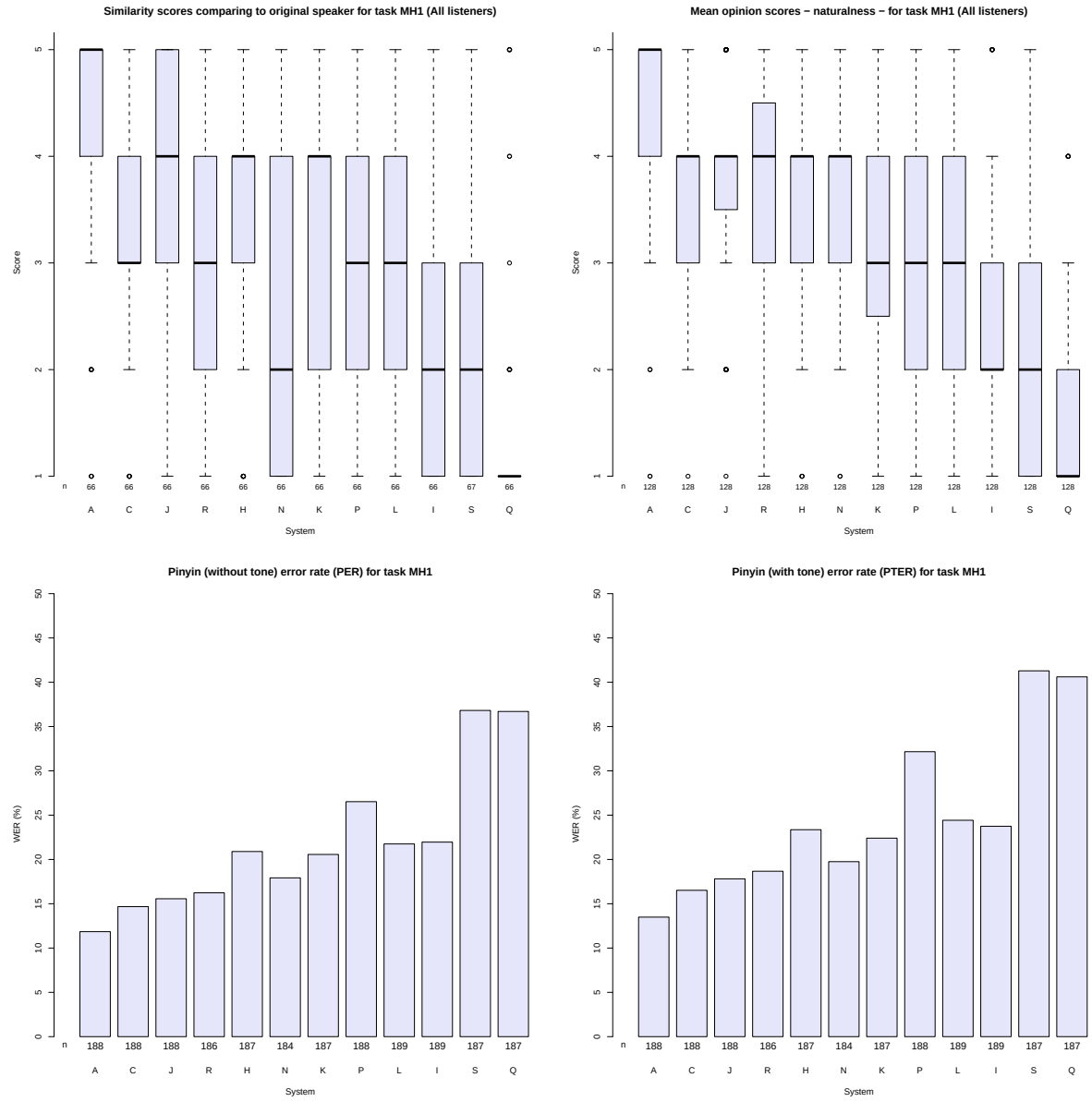

Figure 7: Results for task MH1

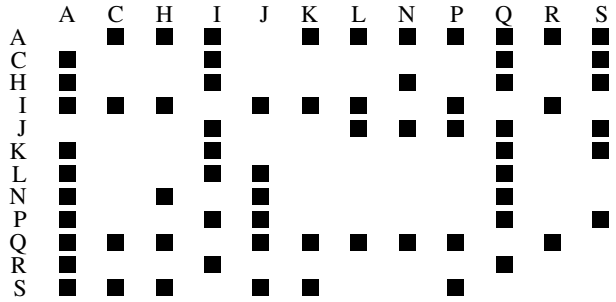

Table 22: Significant differences in similarity to the original speaker for task MH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

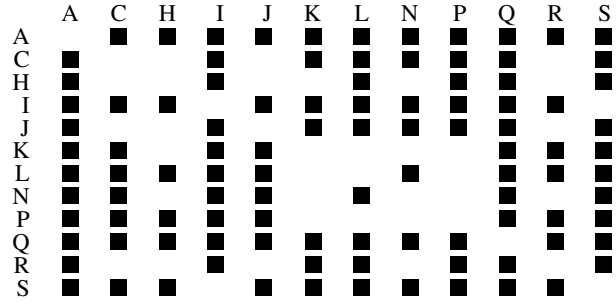

Table 23: Significant differences in naturalness for task MH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

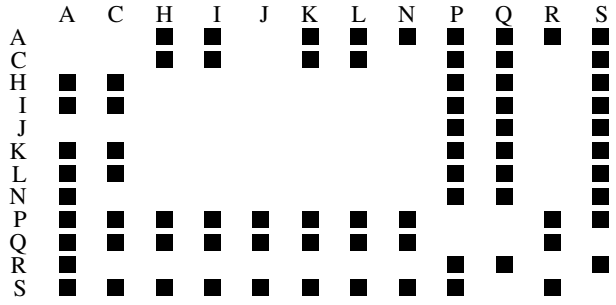

Table 24: Significant differences in intelligibility for task MH1: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rate (PTER). ■ indicates a significant difference between a pair of systems.

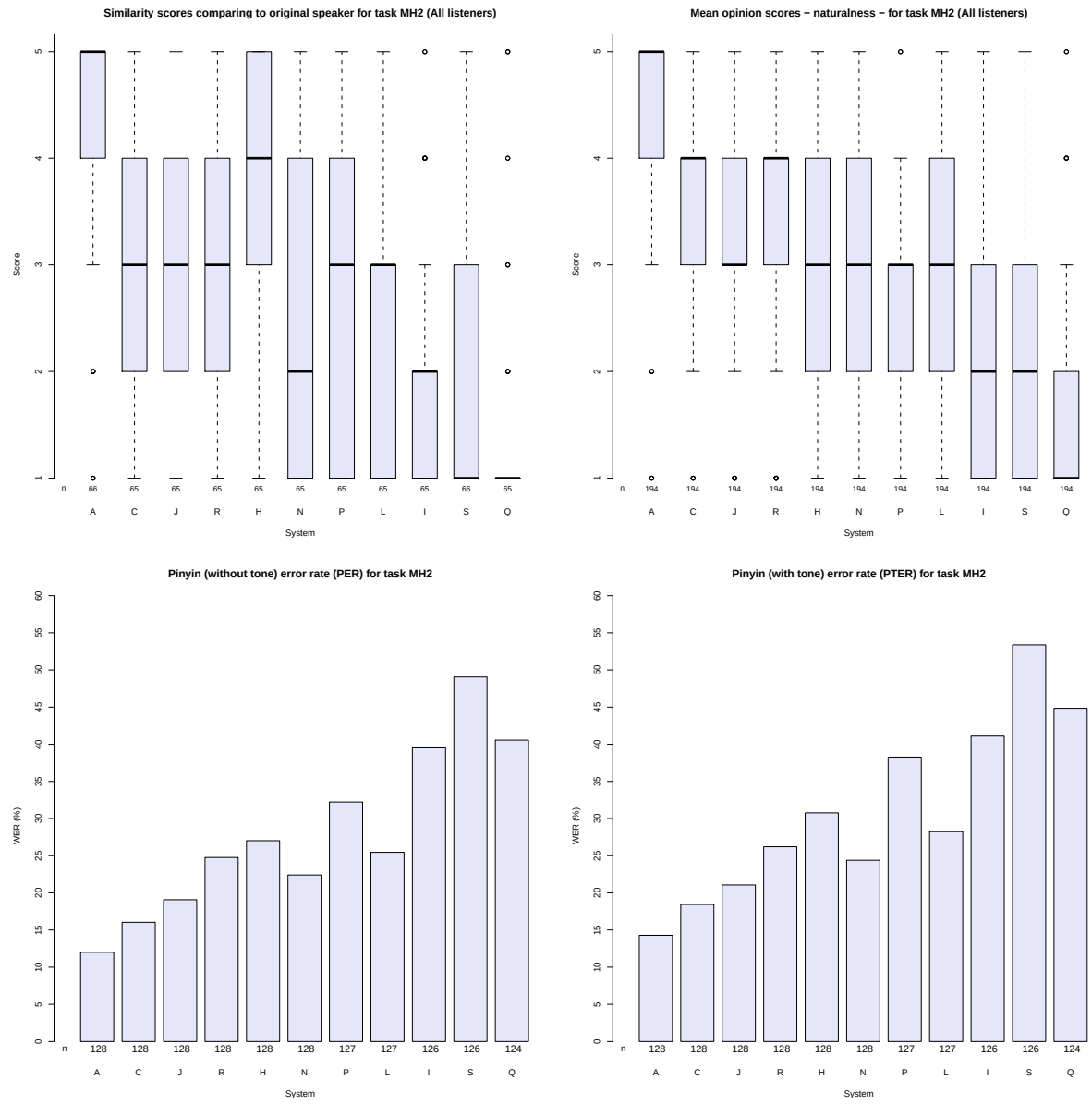

Figure 8: Results for task MH2

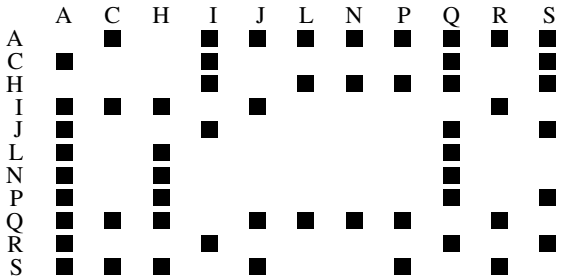

Table 25: Significant differences in similarity to the original speaker for task MH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

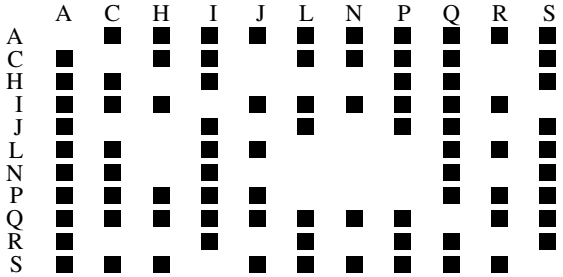

Table 26: Significant differences in naturalness for task MH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

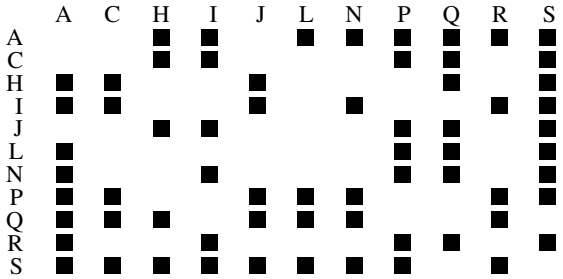

Table 27: Significant differences in intelligibility for task MH2: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rate (PTER). ■ indicates a significant difference between a pair of systems.

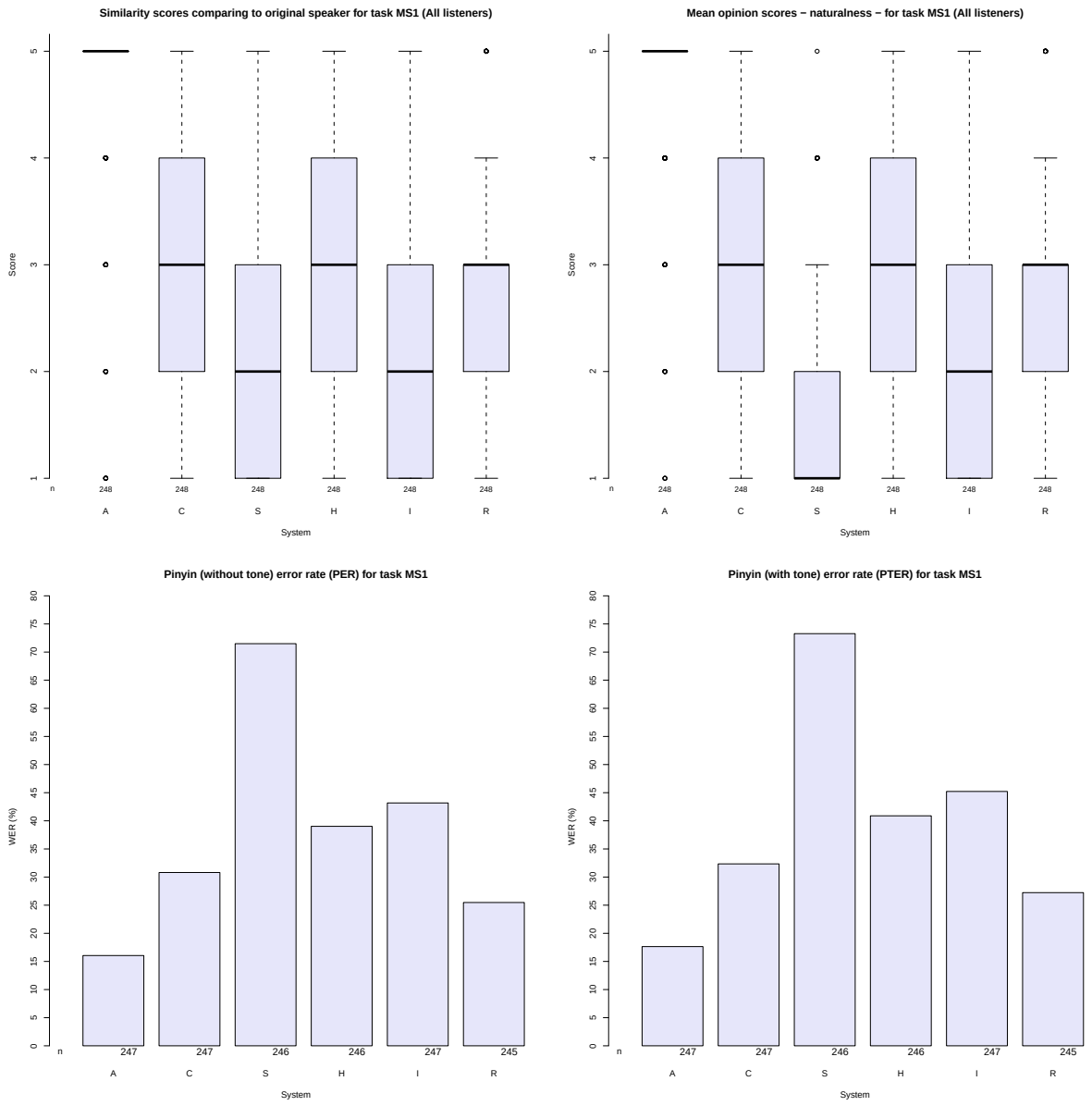

Figure 9: Results for task MS1

	A	C	H	I	R	S
A		■	■	■	■	■
C	■		■	■		■
H	■	■		■	■	■
I	■	■	■		■	
R	■		■	■		■
S	■	■	■		■	

	A	C	H	I	R	S
A		■	■	■	■	■
C	■			■		■
H	■			■		■
I	■	■	■		■	
R	■			■		■
S	■	■	■		■	

	A	C	H	I	R	S
A		■	■	■	■	■
C	■		■	■		■
H	■	■			■	■
I	■	■			■	■
R	■		■	■		■
S	■	■	■	■	■	

Table 28: Significant differences in (from left to right) similarity to the original speaker, naturalness and intelligibility (PTER) for task MS1: results of pairwise Wilcoxon signed rank tests between systems' scores. ■ indicates a significant difference between a pair of systems.

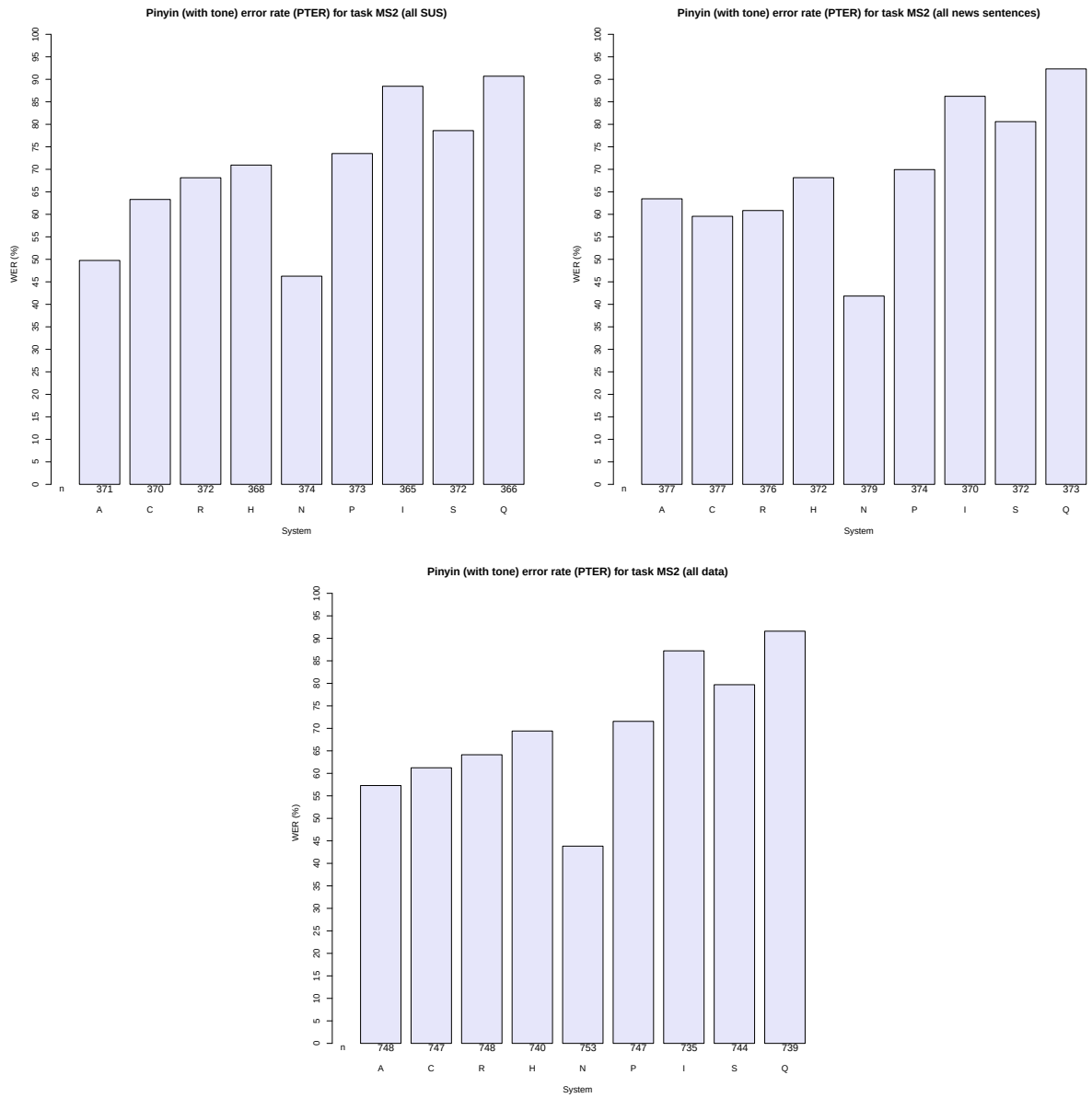

Figure 10: Intelligibility results for task MS2 per sentence type, pooled across all noise conditions.

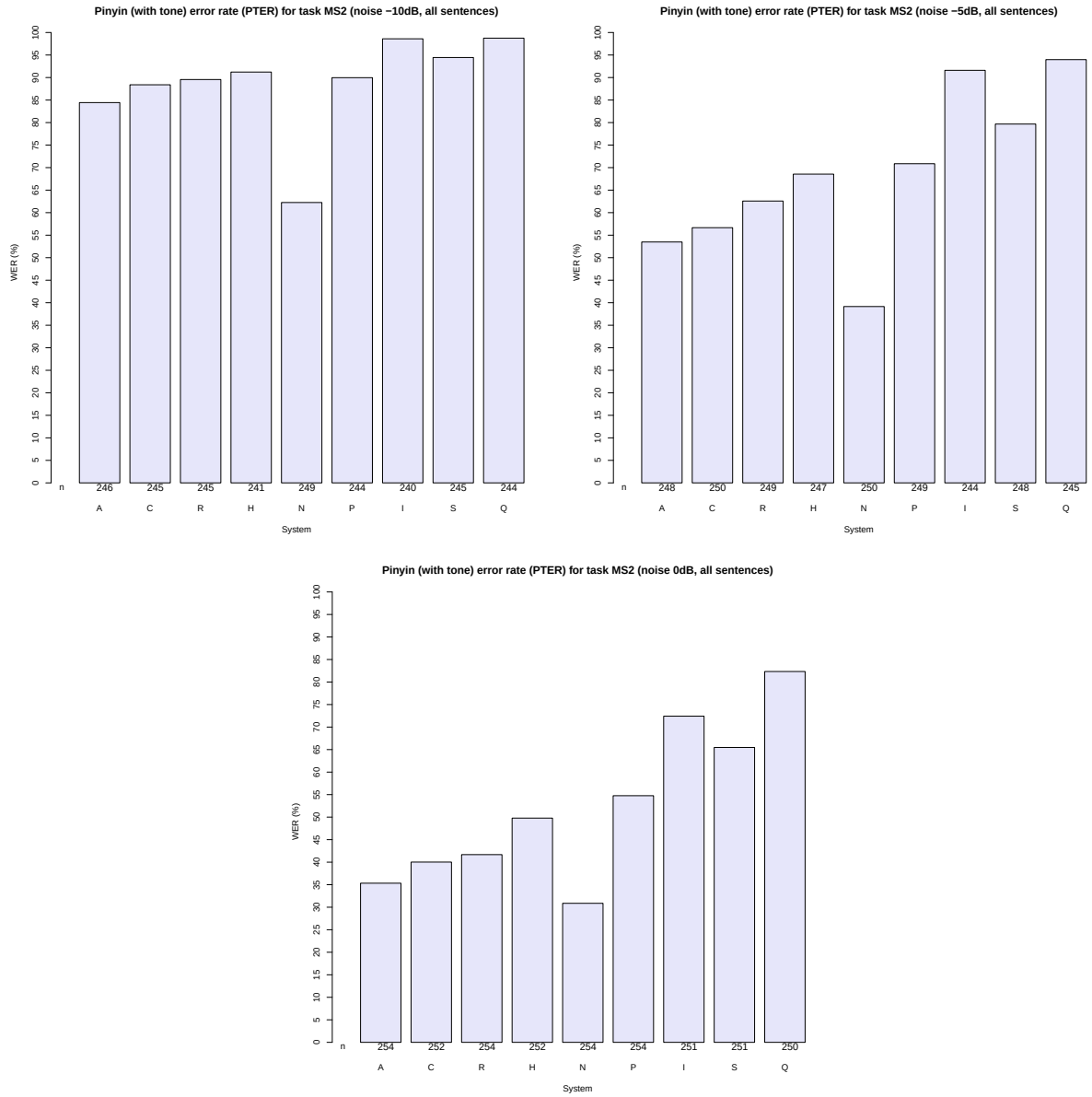

Figure 11: Intelligibility results for task MS2 per noise condition, pooled across both sentence types.

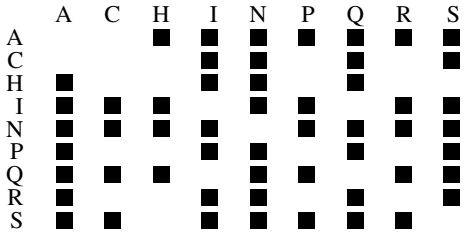

Table 29: Significant differences in intelligibility (PTER) pooled across both sentence types at SNR of -10dB for task MS2: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rates. ■ indicates a significant difference between a pair of systems.

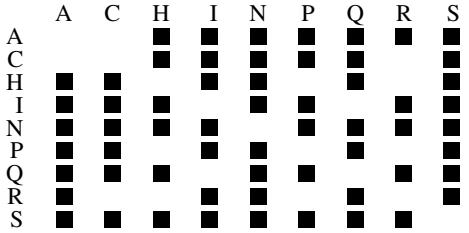

Table 30: Significant differences in intelligibility pooled across both sentence types at SNR of -5dB for task MS2: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rates. ■ indicates a significant difference between a pair of systems.

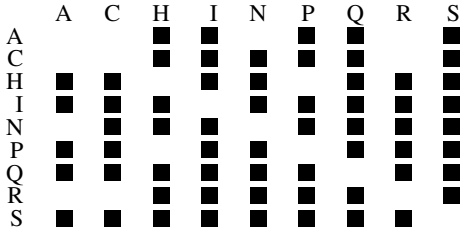

Table 31: Significant differences in intelligibility pooled across both sentence types at SNR of 0dB for task MS2: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rates. ■ indicates a significant difference between a pair of systems.

Language	English total	Mandarin total
Afrikaans	7	0
Arabic	1	0
Catalan	1	0
Chinese	11	0
Croatian	1	0
Dutch	4	0
Estonian	1	0
Finnish	6	0
French	1	0
German	14	0
Greek	11	0
Hindi	2	0
Hungarian	1	0
Italian	1	0
Japanese	38	1
Korean	2	0
Norwegian	2	0
Russian	1	0
Sesotho	1	0
Spanish	3	0
Swedish	2	0
Tamil	2	0
Telugu	1	0
Thai	1	0
N/A	4	2

Table 32: First language of non-native speakers for English and Mandarin versions of Blizzard [2]

Gender	Male	Female
English total	178	184
Mandarin total	93	88

Table 33: Gender [2]

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
English total	47	268	77	32	11	6	1	0
Mandarin total	63	179	17	4	0	1	0	0

Table 34: Age of listeners whose results were used (completed the evaluation fully or partially) [3]

[2] These numbers are calculated from the feedback forms that listeners completed at the end of the test. Since this is optional, many listeners decided not to fill it in. If they did, they did not always reply to all the questions in the form. (Listeners who did one of the bundled tests —EH1/ES2, EH2/ES2, ES3/ES1/ES2, MS1/MS2— are counted once.)

[3] These numbers are calculated from the database where the results of the listening tests are stored. (Listeners who did one of the bundled tests —EH1/ES2, EH2/ES2, ES3/ES1/ES2, MS1/MS2— are counted once.)

Native speaker	Yes	No
English	243	121
Mandarin	169	4

Table 35: Native speakers for English and Mandarin versions of Blizzard [2]

	EH1	EH2	ES1	ES2	ES3	MH	MH2	MS1	MS2
EE	88	88	40	214	40	0	0	0	0
ER	30	34	10	58	14	0	0	0	0
ES	59	59	23	114	30	0	0	0	0
MC	0	0	0	0	0	26	28	63	64
ME	0	0	0	0	0	24	22	45	45
MR	0	0	0	0	0	5	8	2	3
MS	0	0	0	0	0	12	9	14	18
ALL	177	181	73	386	84	67	67	124	130

Table 36: Listener types per voice, showing the number of listeners whose responses were used in the results. Tasks were bundled together as follows: EH1/ES2, EH2/ES2, ES1/ES2/ES3, MH1, MH2, MS1/MS2; therefore, there is a big -but not total- overlap between listeners in some voices. [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	EH1_14	EH1_15	EH1_16	EH1_17	EH1_18
EE	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4
ER	2	2	3	3	0	1	2	2	2	2	1	2	0	1	1	2	2	2
ES	5	3	4	4	3	2	3	4	3	3	3	4	3	3	3	4	2	3
ALL	12	10	12	12	8	8	10	11	10	10	9	11	8	9	9	11	8	9

Table 37: 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]

	EH2_01	EH2_02	EH2_03	EH2_04	EH2_05	EH2_06	EH2_07	EH2_08	EH2_09	EH2_10	EH2_11	EH2_12	EH2_13	EH2_14	EH2_15	EH2_16	EH2_17	EH2_18
EE	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4
ER	2	3	1	2	3	3	1	2	0	1	2	2	2	2	2	2	2	2
ES	4	4	3	4	3	3	3	3	4	4	3	3	4	3	2	4	3	2
ALL	11	12	9	11	11	11	9	10	9	10	10	10	11	10	9	11	9	8

Table 38: Listener groups - Voice EH2 (English), showing the number of listeners whose responses were used in the results [3]

	ES1_01	ES1_02	ES1_03	ES1_04	ES1_05	ES1_06	ES1_07	ES1_08
EE	5	5	5	5	5	5	5	5
ER	1	2	2	1	1	1	1	1
ES	3	4	2	3	3	3	3	2
ALL	9	11	9	9	9	9	9	8

Table 39: Listener groups - Voice ES1 (English), showing the number of listeners whose responses were used in the results [3]

	ES2_01	ES2_02	ES2_03	ES2_04	ES2_05	ES2_06	ES2_07	ES2_08	ES2_09	ES2_10	ES2_11	ES2_12	ES2_13
EE	6	6	6	6	6	6	6	6	6	6	6	6	6
ER	1	1	1	1	2	2	2	2	2	2	2	2	2
ES	5	4	4	4	4	3	3	2	3	3	3	3	2
ALL	12	11	11	11	12	11	11	10	11	11	11	11	10
	ES2_14	ES2_15	ES2_16	ES2_17	ES2_18	ES2_19	ES2_20	ES2_21	ES2_22	ES2_23	ES2_24	ES2_25	ES2_26
EE	6	6	6	6	6	6	5	5	5	5	5	5	5
ER	2	2	2	2	1	2	1	2	2	2	2	1	2
ES	3	2	2	2	3	3	3	3	2	3	3	3	2
ALL	11	10	10	10	10	11	9	10	9	10	10	9	9
	ES2_27	ES2_28	ES2_29	ES2_30	ES2_31	ES2_32	ES2_33	ES2_34	ES2_35	ES2_36	ES2_37	ES2_38	ES2_39
EE	5	5	5	5	5	5	5	5	5	5	5	5	5
ER	1	1	1	1	1	1	1	1	1	1	1	1	1
ES	3	3	3	3	3	3	2	3	2	3	3	3	3
ALL	9	9	9	9	9	9	8	9	8	9	9	9	9

Table 40: Listener groups - Voice ES2 (English), showing the number of listeners whose responses were used in the results [3]

	ES3_01	ES3_02	ES3_03	ES3_04	ES3_05	ES3_06	ES3_07	ES3_08
EE	5	5	5	5	5	5	5	5
ER	2	2	2	1	2	1	2	2
ES	3	4	4	3	4	4	4	4
ALL	10	11	11	9	11	10	11	11

Table 41: Listener groups - Voice ES3 (English), showing the number of listeners whose responses were used in the results [3]

	MH1_01	MH1_02	MH1_03	MH1_04	MH1_05	MH1_06	MH1_07	MH1_08	MH1_09	MH1_10	MH1_11	MH1_12
MC	3	2	2	2	3	1	2	2	2	2	2	3
ME	2	2	2	2	2	2	2	2	2	2	2	2
MR	1	1	1	1	1	0	0	0	0	0	0	0
MS	2	1	1	1	1	1	1	1	0	1	1	1
ALL	8	6	6	6	7	4	5	5	4	5	5	6

Table 42: Listener groups - Voice MH1 (Mandarin), showing the number of listeners whose responses were used in the results [3]

	MH2_01	MH2_02	MH2_03	MH2_04	MH2_05	MH2_06	MH2_07	MH2_08	MH2_09	MH2_10	MH2_11
MC	3	2	2	2	3	3	3	3	2	3	2
ME	2	2	2	2	2	2	2	2	2	2	2
MR	1	1	1	1	1	0	1	1	1	0	0
MS	1	1	1	0	1	0	1	1	1	1	1
ALL	7	6	6	5	7	5	7	7	6	6	5

Table 43: Listener groups - Voice MH2 (Mandarin), showing the number of listeners whose responses were used in the results [3]

	MS1_01	MS1_02	MS1_03	MS1_04	MS1_05	MS1_06
MC	9	11	11	10	10	12
ME	8	8	8	7	7	7
MR	1	1	0	0	0	0
MS	3	3	2	2	2	2
ALL	21	23	21	19	19	21

Table 44: Listener groups - Voice MS1 (Mandarin), showing the number of listeners whose responses were used in the results [3]

	MS2_01	MS2_02	MS2_03	MS2_04	MS2_05	MS2_06	MS2_07	MS2_08	MS2_09	MS2_10	MS2_11	MS2_12	MS2_13	
MC	2	2	2	3	3	3	3	3	2	3	3	3	2	
ME	2	2	2	2	2	2	2	2	2	2	2	2	2	
MR	0	1	0	0	0	0	1	0	1	0	0	0	0	
MS	2	1	1	1	0	1	1	1	0	1	1	1	1	
ALL	6	6	5	6	5	6	7	6	5	6	6	6	5	

	MS2_14	MS2_15	MS2_16	MS2_17	MS2_18	MS2_19	MS2_20	MS2_21	MS2_22	MS2_23	MS2_24	MS2_25	MS2_26	MS2_27
	3	3	1	2	3	2	1	3	2	2	2	2	2	2
	2	2	2	2	2	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	6	6	4	5	6	4	2	4	3	3	3	3	3	3

Table 45: Listener groups - Voice MS2 (Mandarin), showing the number of listeners whose responses were used in the results [3]

	Registered	No response at all	Partial evaluation	Completed Evaluation
EE	216	0	3	213
ER	108	31	25	52
ES	171	27	46	98
ALL ENGLISH	**495**	**58**	**74**	**363**
MC	151	5	17	128
ME	91	0	1	90
MR	27	11	6	10
MS	42	5	4	33
ALL MANDARIN	**311**	**22**	**28**	**261**

Table 46: Listener registration and evaluation completion rates. For listeners assigned to do a bundled test (EH1/ES2, EH2/ES2, ES1/ES2/ES3, MS1/MS2), finishing one but not both of the tests was included as partial completion. [3]

Listener Type	EE	ER	ES	ALL ENGLISH
Total	213	52	98	363

Table 47: Listener type totals for submitted feedback (English)

Listener Type	MC	ME	MR	MS	ALL MANDARIN
Total	117	90	9	33	249

Table 48: Listener type totals for submitted feedback (Mandarin)

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate
English total	62	64	99	85	51
Mandarin total	1	4	67	83	18

Table 49: Highest level of education completed [2]

CS/Engineering person?	Yes	No
English total	165	196
Mandarin total	74	100

Table 50: Computer science / engineering person [2]

Work in speech technology?	Yes	No
English total	115	243
Mandarin total	40	133

Table 51: Work in the field of speech technology [2]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
English total	47	43	50	69	82	31	37
Mandarin total	20	15	11	23	28	51	21

Table 52: How often normally listened to speech synthesis before doing the evaluation [2]

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	1	8	166	32	19	20

Table 53: Dialect of English of native speakers [2]

Dialect of Mandarin	Beijing	Shanghai	Guangdong	Sichuan	Northeast	Other	N/A
Total	37	9	9	9	10	64	18

Table 54: Dialect of Mandarin of native speakers [2]

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
English total	23	35	42	18	0
Madarin total	0	1	0	0	2

Table 55: Level of English/Mandarin of non-native speakers [2]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
English total	349	8	5	1
Mandarin total	151	1	4	1

Table 56: Speaker type used to listen to the speech samples [2]

Same environment?	Yes	No
English total	355	5
Mandarin total	150	5

Table 57: 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
English total	274	67	13	2	0
Mandarin total	131	18	5	2	0

Table 58: Kind of environment when listening to the speech samples [2]

Number of sessions	1	2-3	4 or more
English total	269	67	19
Mandarin total	112	36	7

Table 59: Number of separate listening sessions to complete all the sections [2]

Browser	Firefox	IE	Mozilla	Chrome	Opera	Safari	Other
English total	82	40	2	17	1	212	4
Mandarin total	9	59	1	1	1	73	9

Table 60: Web browser used [2]

Similarity with reference samples	Easy	Difficult
English total	277	85
Mandarin total	129	27

Table 61: 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
English total	48	2	37
Mandrin total	16	3	6

Table 62: Listeners' problems in section(s) about similarity with original voice. [2]

Number of times	1-2	3-5	6 or more
English total	322	33	2
Mandarin total	125	29	0

Table 63: Number of times listened to each example in section(s) about similarity with original voice. [2]

MDS section	Easy	Difficult
English total	303	59
Mandarin total	138	17

Table 64: 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
English total	9	33	0	16
Mandarin total	4	11	1	2

Table 65: Listeners' problems in MOS naturalness sections [2]

Number of times	1-2	3-5	6 or more
English total	326	25	0
Mandarin total	140	14	0

Table 66: 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
English total	0	39	310	11
Mandarin MH1, MH2 total	11	57	13	2
Mandarin MS1/MS2 total	1	40	29	3

Table 67: Listeners' impressions of the task in SUS section(s). (All English listeners had to do ES2, but only a subset of Mandarin listeners did MS2.) [2]

Number of times	1-2	3-5	6 or more
English ES2 total	31	44	10
Mandarin MH1, MH2 total	45	32	1
Mandarin MS1/MS2 total	33	31	2

Table 68: How many times listened to each example in SUS section(s). (For EH1, EH2, ES1, ES3 sentences could only be heard once.) [2]