The Blizzard Challenge 2009

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 2009 was the fifth annual Blizzard Challenge. As in 2008, UK English and Mandarin Chinese were the chosen languages for the 2009 Challenge. The English corpus was the same one used in 2008. The Mandarin corpus was provided by iFLYTEK. As usual, participants with limited resources or limited experience in these languages had the option of using unaligned labels that were provided for both corpora and for the test sentences. An accent-specific pronunciation dictionary was also available for the English speaker. This year, 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 application.

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, degree of similarity to the original speaker and, for one of the spoke tasks, "appropriateness."

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

1. Introduction

The Blizzard Challenge, conceived by Black and Tokuda [1], is the international evaluation of corpus-based speech synthesisers open to any participant. Blizzard Challenges are scientific research exercises, not competitions, in which participants use a common corpus to build speech synthesisers. A common test set is then synthesised and a large listening test is used to obtain listeners' judgements regarding the overall naturalness of the speech, its intelligibility and how similar it sounds to the original speaker. In this, the 2009 Challenge, we used the same general setup as in recent challenges, but with the tasks organised into a hub and spoke structure, as explained in this paper.

The first two Blizzard Challenges, in 2005 and 2006, were organised by Carnegie Mellon University, USA, with the 2007, 2008 and 2009 Challenges being organised by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. For general details of Blizzard, the rules of participation, a timeline, and information on previous and future Blizzard Challenges, see the website [2].

2. Participants

The Blizzard Challenge 2005 [1, 3] had 6 participants, Blizzard 2006 had 14 [4], Blizzard 2007 had 16 [5] and Blizzard 2008 had 19 (18 for English, 11 for Mandarin) [6]. This year, there were again 19 participants, listed in Table 1. One participant requested to withdraw from the Challenge after the listening test was completed. The results for this system ('ANON' in Table 1) have been retained in the tables and plots presented here and in the complete set of results distributed to participants. This is important, because listener scores obtained using 5-point scales are effectively internally normalised by listeners with respect to the range of stimuli they are presented with. In other words, the similarity and naturalness ratings of any individual system are relative to the scores of all the other systems present in the listening test. The upper end of the 5-point scale can be fixed by the inclusion of natural speech, but the remainder of the scale is calibrated only by the other systems present in the test. Proper interpretation of the results therefore requires presentation of the scores from all systems together. In future Blizzard Challenges, we may explicitly disallow withdrawal after distribution of results.

Three systems from previous challenges were used as benchmarks, in an attempt to facilitate comparisons between the results from one year to another: a Festival-based system from CSTR configured very similarly to the Festival/CSTR entry to Blizzard 2006 [7], an HTS speaker-dependent system configured the same as the HTS entry to Blizzard 2005 [8] and the HTS speaker-adaptive system from Blizzard 2007 [9]. Whilst precise calibration of Mean Opinion Score (MOS) ratings across different listening tests (with different participating systems and different listeners) is almost certainly not possible, the *ranking* of a system relative to these benchmarks may possibly be meaningfully compared from one year to another. Comparisons of the absolute scores across different years should be avoided, noting both the point made above about the relative nature of such scores and also that each year different sentences and a different pool of listeners is used.

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 D denoting the three benchmark systems and E to W denoting the nineteen 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 the Centre for Speech Technology Research, University of Edinburgh, UK. Participants who had signed a user agreement were able to download about 15 hours of recordings of a UK English male speaker with a fairly standard RP accent. An accent-specific pronunciation dictionary, and Festival utterance files created using this dictionary, were also available for the English speaker, under a separate licence. This is exactly the same data used for the 2008 Challenge.

For Mandarin, the ANHUI USTC iFLYTEK Company, Ltd. (iFLYTEK) released a 10 hour / 6000 utterance Mandarin Chinese database of a young female professional radio broadcaster with a standard Beijing accent, reading news sentences. The first 1000 sentences were manually phonetically segmented and prosodically labelled, with the remainder being segmented or labelled automatically. Because it was not possible to make additional recordings of this speaker, no natural semantically unpredictable sentences were available this year. However, we took the

view that this was a reasonable price to pay, given the opportunity to use a commercially-produced corpus.

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 adopted this year. Hub tasks contain 'H' in the task name, spoke tasks contain 'S' and each are described in the following sections.

3.2.1. English tasks

- EH1: English full voice from the full dataset (about 15 hours)
- EH2: English ARCTIC voice from the ARCTIC [11] subset (about 1 hour)
- ES1: build voices from the specified 'E_SMALL10', 'E_SMALL50' and 'E_SMALL100' datasets, which consist of the first 10, 50 and 100 sentences respectively of the 'ARCTIC' subset. Participants could use voice conversion, speaker adaptation techniques or any other technique.
- ES2: build a voice from the full UK English database suitable for synthesising speech to be transmitted via a telephone channel. The telephone channel simulation tool described in Section 3.3 was made available to assist participants in system development. It was permissible to enter the same voice as task EH1 or EH2, but specially-designed voices were strongly encouraged.
- ES3: build a voice from the full UK English database suitable for synthesising the computer role in a human-computer dialogue. A set of development dialogues were provided, from the same domain as the test dialogues. Participants could enter the same voice as task EH1 or EH2, but again specially-designed voices were strongly encouraged. Participants were allowed to add simple markup to the text, either automatically or manually, if they wished. The markup had to be of a type that could conceivably be provided by a text-generation system (e.g. emphasis tags were acceptable, but a handcrafted F0 contour was not).

3.2.2. Mandarin tasks

- MH: Mandarin voice from the full dataset (about 10 hours / 6000 utterances / 130000 Chinese characters)
- MS1: build voices from each of the specified 'M_SMALL10', 'M_SMALL50' and 'M_SMALL100' datasets, which consist of the first 10, 50 and 100 sentences respectively of the full Mandarin database. Same rules as ES1.
- MS2: build a voice from the full Mandarin database suitable for synthesising speech to be transmitted via a telephone channel. Same rules as ES2.

System short name	Details
NATURAL	Natural speech from the same speaker as the corpus
FESTIVAL	The Festival unit-selection benchmark system [7]
HTS2005	A speaker-dependent HMM-based benchmark system [8]
HTS2007	A speaker-adaptive HMM-based benchmark system [9]
AHOLAB	Aholab, University of the Basque Country, Spain
ANON	Identity withheld
CASIA	National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
CEREPROC	CereProc Ltd, UK
CMU	Carnegie Mellon University, USA
CSTR	The Centre for Speech Technology Research, University of Edinburgh, UK
DFKI	DFKI GmbH, Germany
EMIME	The EMIME project consortium
I2R	Institute for Infocomm Research (I^2R), Singapore
ITRI	Industrial Technology Research Institute, Taiwan
IVO	IVO Software Sp. z o. o.
MXAC	μXac, Australia
NICT	National Institute of Information and Communications Technology, Japan
NIT	Nagoya Institute of Technology, Japan
NTUT	National Taipei University of Technology, Taiwan
SHRC	Speech and Hearing Research Center, Peking University, China
TOSHIBA	Research and Development Center, Toshiba (China)
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 four rows are the benchmark systems 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 W.

System	EH1	EH2	ES1	ES2	ES3	MH	MS1	MS2
NATURAL	X	X	X	X	X	X	X	X
FESTIVAL	X	X		X				
HTS2005	X	X		X		X		X
HTS2007	X	X	X	X		X	X	X
AHOLAB	X	X	X					
ANON	X	X						
CASIA	X	X				X	X	
CEREPROC	X	X						
CMU	X	X	X	X				
CSTR		X			X			
DFKI	X	X		X		X		
EMIME	X	X	X	X		X	X	X
HTS	X	X	X	X		X	X	X
I2R	X	X		X		X	X	X
ITRI						X		
IVO	X	X		X				
MXAC	X	X	X					
NICT	X	X		X				
NTUT						X	X	X
SHRC						X		X
TOSHIBA						X		X
USTC	X	X	X	X	X			
VUB	X	X		X				

Table 2: The tasks completed by each participating system. The first four rows are the benchmark systems 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 W

```
# Set active speech level of source signal to
# -26 dBov
sv56demo -q -lev -26 -sf 8000 in.pcm tmp1.pcm

# The level-normalized source speech signal is
# then filtered according to the
# "telephone bandpass" defined in
# ITU-T Rec. G.712
c712demo tmp1.pcm tmp2.pcm

# The G.712-filtered version is successively
# G.711-encoded, encoded and decoded according
# to G.726 at 16 kbit/s, and decoded by
# G.711 (A-law)
g711demo A lilo tmp2.pcm tmp3a.pcm 160
g726demo A lolo 16 tmp3a.pcm tmp3b.pcm 160
g711demo A loli tmp3b.pcm tmp3.pcm 160

# The decoded signal is filtered according to
# the (modified) Intermediate Reference System
# in receive direction, as defined in
# ITU-T Rec. P.830
filter -q RXIRS8 tmp3.pcm tmp4.pcm 160

# Set active speech level of output signal to
# -26 dBov
sv56demo -q -lev -26 -sf 8000 tmp4.pcm out.pcm
```

Figure 1: The pipeline of processes used to simulate the telephone channel. Input and output are headerless PCM files at 8kHz sampling rate and 16 bit sample depth.

3.3. Telephone channel simulation for tasks ES2 and MS2

In order to investigate the effects of telephone channels on the intelligibility of the submitted synthetic speech, a simulated telephone channel was used. Although it would have added more realisim to present listeners with the stimuli monaurally using a telephone handset, this was not practical for the large numbers of listeners required by the Blizzard Challenge.

The simulated channel was implemented using the "G.191: Software tools for speech and audio coding standardization" software freely available from the ITU[1] with a pipeline of processes kindly suggested by Telekom Laboratories & The Quality and Usability Lab at TU Berlin, shown in Figure 1. We elected to implement a relatively low quality channel with a 16kbps transmission rate.

Participants were provided with this pipeline, should they wish to use it during development of their ES2 and MS2 voices. They were encouraged to build special voices for these tasks, but were allowed to enter their EH1, EH2 or MH1 voices instead.

3.4. Appropriateness (task ES3)

At previous Blizzard workshops there was a clear desire to evaluate more than just naturalness and intelligibility; specifically, participants wished to evaluate synthetic speech in a particular usage context. Therefore, we conceived task ES3 in which the synthetic speech was evaluated in a simulated human-computer dialogue. A real-time dialogue system, which dynamically generates the computer response, would require participants to submit run-time synthesisers. It was decided that this would be unattractive for some participants, and impractical for the organisers. Therefore, we used pairs of dialogue utterances comprising one user's query to the system followed by the system's response. These were kindly provided by the CLASSIC project[2] and were in a restaurant recommendation domain. The sentences were manually adjusted by the Blizzard organisers in order to remove difficult-to-pronounce restaurant names (e.g. French words). Since these dialogue pairs were static, participants could pre-synthesise all the system utterances and submit them for evaluation. For the test sentences, the texts of both the user query and the corresponding system response were provided to participants.

[1] http://www.itu.int/rec/T-REC-G.191-200509-I/en

[2] www.classic-project.org

4. Listening test design

4.1. Interface

The listening evaluation was conducted online, using the design developed for Blizzard 2007 [5] and refined in Blizzard 2008 [6], which was itself developed from designs in previous challenges [1, 3, 4]. The registration page for each listener type presented an overview of the listening test and the tasks to be completed. It was possible for a listener to register for both the English and Mandarin listening tests separately, if they wished. Please refer to [5] for a complete description of the listening test interface.

4.2. Materials

The participants were asked to synthesise several hundred test sentences (including the complete Blizzard Challenge 2007 and 2008 test sets, to be retained as a resource for future experimentation), 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 – please see [6] for details. Permission has been obtained from almost all participants to distribute parts of this dataset along with the listener scores; we hope to find the resources to do this shortly. For English, participants synthesised sentences that had been held out from the corpus (so that natural speech samples were available for them) plus Semantically Unpredictable Sentences (SUS) [12] generated using a tool provided by Tim Bunnell of the University of Delaware and recorded by us specially for the Challenge with the same speaker as the distributed corpus. These SUS conform more closely to the original specification [12] and use simpler words than the SUS used in previous Blizzard Challenges. In order to mitigate this and avoid ceiling effects, listeners were only permitted to play each such sentence once. For Mandarin, held out sentences were also used. The SUS for Mandarin were generated using the same tool as in 2008. Natural SUS were not available for Mandarin, since the original speaker was not available.

4.3. 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 participants, mailing lists, blogs, etc. (ER).
- Speech experts, recruited via participants and mailing lists (ES).
- Paid UK undergraduates, native speakers of UK English, 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).

Section number	Tasks being evaluated	Type (see Section 4.4.1)
Test name: EH1 + ES3		
1	EH1	SIM
2	EH1	Multidimensional scaling (MDS)
3	EH1	MOSnews
4	EH1,ES3	MOSconv
5	EH1	SUS
6	ES3	MOSapp
Test name: EH2 + ES3		
1	EH2	SIM
2	EH2	MDS
3	EH2	MOSnews
4	EH1,ES3	MOSconv
5	EH2	SUS
6	ES3	MOSapp
Test name: ES1 + ES2		
1	ES1	SIM
2	ES1	SIM
3	ES1	MOSnews
4	ES1	MOSconv
5	ES1	SUS
6	ES2	SIM
7	ES2	MOSnews
8	ES2	MOSconv
9	ES2	SUS
10	ES2	SUS

Table 3: The three listening tests conducted for English.

- Volunteers, recruited via participants, mailing lists, etc. (MR).
- Speech Experts, recruited via participants and mailing lists (MS).

Tables 29 to 35, 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.

4.4. Listening tests

Since the tests for tasks ES1, ES2, MS1 and MS2 were relatively short, they were combined into pairs in order to make the best use of available listeners. Only two participants entered ES3 voices, so the listening test for this task was handled differently. Rather than simply performing a comparison between these two systems, they were included in two sections of the main EH1 and EH2 listening tests, as described in Section 4.4.1. Tables 3 and 4 show the five independent listening tests that were 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, with the exception of the MDS sections (which involved pairwise comparisons) and the MOSconv/MOSapp sections in tests EH1+ES3 and EH2+ES3. 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. Samples of the original speaker were included in all sections, except for Mandarin SUS.

Section number	Tasks being evaluated	Type (see Section 4.4.1)
Test name: MH		
1	MH	SIM
2	MH	MDS
3	MH	MOSnews
4	MH	MOSnews
5	MH	SUS
6	MH	SUS
7	MH	SUS
Test name: MS1 + MS2		
1	MS1	SIM
2	MS1	SIM
3	MS1	MDS
4	MS1	MOSnews
5	MS1	MOSnews
6	MS2	SUS
7	MS2	SIM
8	MS2	MDS
9	MS2	MOSnews
10	MS2	MOSnews
11	MS2	SUS
12	MS2	SUS

Table 4: The two listening tests conducted for Mandarin.

4.4.1. Description of each type of section in the listening test

SIM In each part listeners could play 4 reference samples of the original speaker and one synthetic sample. They chose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person].
MDS In each part listeners heard one sample from each of two of the participating systems, (or, in the case of one system ordering for each dataset, two samples from the same system). Listeners were asked to ignore the meanings of the sentences and instead concentrate on how natural or unnatural each one sounded. They then chose whether in their opinion the two sentences were similar or different in terms of their overall naturalness. The results of this section are intended for analysis using Multidimensional Scaling (not presented here).
MOSnews Mean Opinion Score (MOS - naturalness), news domain. In each part listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].
MOSconv Mean Opinion Score (MOS - naturalness), conversational domain. In each part listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].

There were only two entries to the ES3 task, so we devised a listening test design in which the listening tests EH1+ES3 and EH2+ES3 included sections in which samples from the two ES3 systems that were submitted, plus samples from all systems for voice EH1 or EH2. However, due to a small error in the listening test scripts, these two sections actually contained samples from all EH1 systems *except* the EH1 samples from the two teams that submitted an ES3 voice, but *including* samples from the ES3 voice of those two teams. The consequence of this is that the EH1 samples from those two teams were evaluated by fewer listeners than intended. We used the results for all EH1 samples from both MOSconv sections (the ones from test EH1+ES3, and the one from test EH2+ES3) to compute the MOS scores for voice

EH1. The results from the ES3 samples are presented separately.

SUS Semantically Unpredictable Sentences (SUS) designed to test the intelligibility of the synthetic speech. Listeners heard one utterance in each part and typed in what they heard. The error rates were computed as in previous years [5, 6].

MOSapp Mean opinion scores (MOS - appropriateness), conversational domain. In each part, listeners saw a question (provided in text form only) of the type that a human user might ask a restaurant enquiry service, and then listened to one spoken sample that represented the response to that question. Listeners chose a score which represented how appropriate or not the response sounded in that dialogue context on a scale of 1 [Completely Inappropriate] to 5 [Completely Appropriate]. For this section we used the samples from the two teams that submitted a separate voice for ES3; we decided to also add EH1 samples from all the other teams. The results are presented together.

4.4.2. Number of listeners

The listener responses used for the distributed results were extracted from the database on 26th June 2009 after the online evaluation had been running for approximately six weeks. The number of listeners obtained is shown in Table 5.

	English	Mandarin
Total registered	482	334
of which:		
Completed all sections	365	311
Partially completed	59	14
No response at all	58	9

Table 5: Number of listeners obtained

See Table 28 for a detailed breakdown of evaluation completion rates for each listener type. As in last year's challenge, the higher completion rate for Mandarin listeners is a consequence of the higher proportion of paid listeners.

5. 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. Please refer to [13] for a complete 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 7.3 and Tables 23 to 63 for a summary of the responses to the questionnaire that listeners were asked to optionally complete at the end of the listening test.

6. 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 for a particular language. This ordering is in descending order of the mean MOS (combining MOSnews and MOSconv) for the main task (EH1 or MH) – see Tables 6 and 8. 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.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.9	0.38	463	43
B	3	1.5	2.9	1.06	457	49
C	3	1.5	2.7	1.07	463	43
D	2	1.5	2.5	1.02	456	50
E	2	1.5	2.1	1.01	462	44
H	3	1.5	2.8	1.01	463	43
I	3	1.5	3.1	1.02	462	44
J	2	1.5	2.4	0.98	463	43
K	4	1.5	3.8	0.88	457	49
L	3	1.5	2.8	0.97	457	49
M	2	1.5	1.9	0.92	462	44
O	3	1.5	2.6	0.98	463	43
P	2	1.5	2.0	0.98	457	49
Q	2	1.5	2.1	0.93	463	43
R	2	1.5	2.1	0.97	463	43
S	4	1.5	4.2	0.71	163	343
T	2	1.5	2.0	0.97	463	43
W	2	1.5	2.1	0.94	456	50

Table 6: Mean opinion scores for task EH1 (full data set) on the combined results from sections 3 and 4 of the EH1+ES3 listening test, excluding the ES3 samples. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded). Note the high value of na for system S – this is due to the error in the setup of section 4 of this listening test.

6.1. Task EH1

Table 6 presents descriptive statistics for the mean opinion scores for English task EH1. Figure 2 displays the results of the tests graphically. As expected, we see that natural speech (system A) has a MOS naturalness of 5. Inspecting the Bonferoni-corrected pairwise Wilcoxon signed rank significance tests ($\alpha = 0.01$) for naturalness presented in Table 11 reveals that system A is significantly different from all other systems. We can therefore say that no synthesiser is as natural as the natural speech. Systems S and K, whilst not as natural as the natural speech, are both significantly more natural than all other systems.

From the plot of similarity scores and by referring to Table 10, we can also say that, although systems K and S are significantly less similar to the original speaker than natural speech, they are both significantly more similar to the original speaker than all other systems, for English task EH1. Likewise, from Table 11, systems S and K are equally natural and significantly more natural than all other systems, although significantly less natural than natural speech.

System S is as intelligible as natural speech (Table 12). However, there is no significant difference between system S and a number of other systems (B,C,K,L,O,P), so we cannot state that system S is more intelligible than other systems.

6.2. Task EH2

For English task EH2, results are given in Table 7 and Figure 3 with statistical significances shown in Table 13 for similarity, Table 14 for naturalness and Table 15 for intelligibility. Again, no system is as natural as the natural speech, or as similar to the original speaker. There is no system that is clearly more natural than the rest. Although it was as intelligible as natural speech on task EH1, system S is no longer as intelligible as natural speech on task EH2.

6.3. Task ES1

In both English and Mandarin, we chose to evaluate just one of the three voices built for this task. For task ES1, we selected the

System	median	MAD	mean	sd	n	na
A	5	0.0	4.8	0.47	139	13
B	3	1.5	2.9	1.08	140	12
C	3	1.5	3.2	0.94	140	12
D	3	1.5	2.6	0.97	141	11
E	2	1.5	1.7	0.83	140	12
H	3	1.5	2.6	0.94	140	12
I	3	1.5	3.3	0.98	141	11
J	3	1.5	2.6	1.03	140	12
K	4	0.0	3.6	0.89	140	12
L	3	1.5	3.3	1.01	139	13
M	2	1.5	1.8	0.89	140	12
O	2	1.5	2.5	0.92	141	11
P	2	1.5	2.4	0.99	141	11
Q	3	1.5	2.5	1.03	140	12
R	2	1.5	2.3	0.88	140	12
S	4	1.5	3.7	0.92	141	11
T	2	1.5	2.2	1.01	140	12
U	2	1.5	1.7	0.88	141	11
W	2	1.5	2.3	0.99	140	12

Table 7: Mean opinion scores for task EH2 (ARCTIC data set) on the combined results from sections 3 and 4 of the EH2+ES3 listening test, excluding the ES3 samples. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded).

E_SMALL100 voice, based on preferences expressed by participants who submitted entries for task ES1. For English task ES1 (building a voice from very small amounts of speech), results are given in Figure 4 and significance tests are shown in Table 16.

All systems are rated as unnatural and not similar to the original speaker. Systems J and P and significantly less similar to the original speaker than the other systems. Systems W and D are somewhat more natural than other systems, although this is not significant in all cases.

The systems fall neatly into three groups for intelligibility: natural speech is significantly more intelligible than all synthesisers, systems P, S, D, W and L are equally intelligible, followed by systems H and J.

6.4. Task ES2

For English task ES2 (building a voice for use over the telephone), results are given in Figure 5. Significance tests are shown for naturalness and intelligibility in Table 17. Now there is no system that is as intelligible as natural speech – it appears that synthetic speech may be generally more degraded by the telephone channel than natural speech in terms of intelligibility.

6.5. Task ES3

For English task ES3 (building a voice for a dialogue system), results are given in Figure 6. System S is rated as significantly more appropriate than system U (using the same type of pairwise Wilcoxon signed rank tests as in other tasks), although this may be simply because system U is significantly less natural than system S.

6.6. Task MH

Table 8 and Figure 7 presents the results for the Mandarin hub task MH. The significance tests illustrated in Table19 show that again, as for English, no system is as natural as the natural speech. The most natural synthesiser is system L which, although less natural than natural speech, is significantly more natural than all other systems.

Since natural SUS were not available for Mandarin this year, we are unable to test whether any system was as intelligible as natural speech. We can say, from the significance tests illustrated in Table20, that systems L, F and C are equally intelligible, although only system L is significantly more intelligible than the remaining systems.

With regards to similarity to the original speaker, Table18 shows that no system was regarded as being as similar to the original speaker as the natural speech. Systems L, F, C and R form a group of systems that appear to be most similar to the original speaker, although only systems L and F are significantly more similar than the remaining systems. Note that system F is actually significantly different to system R within this approximate grouping.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.6	0.79	370	26
C	4	1.5	3.6	0.96	370	26
D	3	1.5	2.9	1.06	370	26
F	4	1.5	3.8	1.07	370	26
G	3	1.5	2.8	1.13	371	25
I	3	1.5	3.3	1.24	370	26
L	4	1.5	4.1	0.93	370	26
M	3	1.5	3.1	1.14	370	26
N	3	1.5	2.8	1.25	370	26
R	4	1.5	3.5	1.04	371	25
V	3	1.5	3.0	1.18	370	26
W	3	1.5	3.1	1.04	370	26

Table 8: Mean opinion scores for task MH. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and na (data points excluded due to missing data)

6.7. Task MS1

For Mandarin task MS1 (building a voice from very small amounts of speech), results are given in Figure 8 and significance tests are shown in Table 21. We selected the M_SMALL100 voice for evaluation, based on preferences expressed by participants who submitted entries for this task. No system was found to be as natural or as similar to the original speaker as the natural speech. Systems L, R, W and D form a group of systems which sound most similar to the original speaker (although there is a significant difference between system R and system D). System L is significantly more natural than all other systems except W. There are few significant differences in intelligibility between most systems, in terms of PTER.

6.8. Task MS2

For Mandarin task MS2 (building a voice for use over the telephone), results are given in Figure 9. Significance tests are shown for naturalness and intelligibility in Table 22. The natural speech (system A) is no longer rated by listeners as being very similar to the original speaker, although it is still found to be highly natural and significantly more so than any other system. System L is significantly more natural than all other systems, except natural speech. There are relatively few significant differences in intelligibility, with systems C, F, L, V and W forming a group of roughly equally intelligible systems (although there are some significant differences between systems within this group, and also some insignificant differences between some members of this group and the remaining systems).

7. Discussion

There is continued interest in the Blizzard Challenge, with 19 teams participating this year. We therefore propose to organise another Challenge in 2010. In 2009, we made several additions

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

Table 9: Comparing the results of the benchmark systems for English (main voice, large database) across three 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 were simpler than those in 2007 and 2008

to the challenge, with varying degrees of success. Both the 'very small amounts of data' and 'speech for transmission by telephone' tasks seemed popular with participants. The dialogue speech task was not popular, with only two entries, even though from our discussions with past participants this type of application for TTS is widely thought to be important and interesting. Entries to this task probably required considerably more effort, and perhaps needed more expert knowledge of the language (English). We would welcome suggestions for ways to evaluate 'appropriateness' or any other measure of how good synthetic speech is in particular usage situations or applications. Task-based scenarios are attractive, since they allow objective measures of task success (e.g. completion rate or time taken). However, they also tend to be lengthy and may require on-line generation of synthetic speech; neither of these are practical for the Blizzard Challenge.

7.1. Benchmark systems

The inclusion of the benchmark systems is intended to provide reference points for comparison between different years of the Challenge. If this is to be possible, then the relative ranking of the benchmark systems should be constant from year to year. Table 9 presents the key results for the English benchmark systems for 2007, 2008 and 2009. These results do seem to be consistent year-on-year. WER decreases uniformly by about one third for all systems from 2008 to 2009, due to the simpler SUS used this year. The relative MOS and WER of the three systems is consistent: for MOS, the ranking is Natural–Festival–HTS 2005; for WER, the ranking is Natural–HTS 2005–Festival.

7.2. Limitations of the listening test design

The current listening test design has many advantages, including the ability to perform evaluations for quite large number of systems (perhaps up to 25) with a fully balanced design which controls for possible effects of sentence and order of presentation by using a Latin Square design.

We consider this year's hub and spoke design a success, because it allowed participants to enter whichever tasks they desired. The disappointing number of entries to task ES3 necessitated special treatment in the listening test, which created considerable additional complexity in the design which in turn lead to a small error being made in this part of the test.

However, there are two significant weaknesses which should be considered when designing future listening tests for the Blizzard Challenge:

- The listening tests for each hub and spoke task are conducted independently, making cross-task comparisons impossible. In particular, this year's test does not allow direct calculation of the difference in intelligibility for a single system between a hub task and the telephone channel spoke task.
- Each new task added increases the number of listeners required. This year, we were able to use the same listener pool for some pairs of tasks, but this necessitated the use of different sentences in each test (particularly important for SUS) which only increases the difficulty in comparing results across tasks for a single system.

7.3. Listener feedback

On completing the evaluation, listeners were given the opportunity to tell us what they thought through an online feedback form. This was the same as in Blizzard 2007 and 2008. 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 23 to 63.

8. Acknowledgements

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. Roger Burroughes is 'roger', the English voice; Tim Bunnell of the University of Delaware generated the 2009 SUS sentences; iFLYTEK provided the Mandarin data; the listening test scripts are based on earlier versions provided by previous organisers of the Blizzard Challenge. Thanks to all participants and listeners.

9. 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] "Blizzard Challenge 2009 website," http://www.synsig.org/index.php/Blizzard_Challenge_2009.

[3] C.L. Bennett, "Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005," in *Proceedings of Interspeech 2005*, 2005.

[4] C.L. Bennett and A. W. Black, "The Blizzard Challenge 2006," in *Blizzard Challenge Workshop, Interspeech 2006 - ICSLP satellite event*, 2006.

[5] Mark Fraser and Simon King, "The Blizzard Challenge 2007," in *Proc. Blizzard Workshop (in Proc. SSW6)*, 2007.

[6] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Workshop (in Proc. SSW7)*, 2008.

[7] 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.

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

[9] 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.

[10] "Blizzard Challenge 2009 rules," http://www.synsig.org/index.php/Blizzard_Challenge_2009_Rules.

[11] J. Kominek, NewAuthor1, and A. W. Black, "The CMU Arctic speech databases," in *SSW5-2004*, 2004, pp. 223–224.

[12] C. Benoit and M. Grice, "The SUS test: a method for the assessment of text-to-speech intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, pp. 381–392, 1996.

[13] 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.

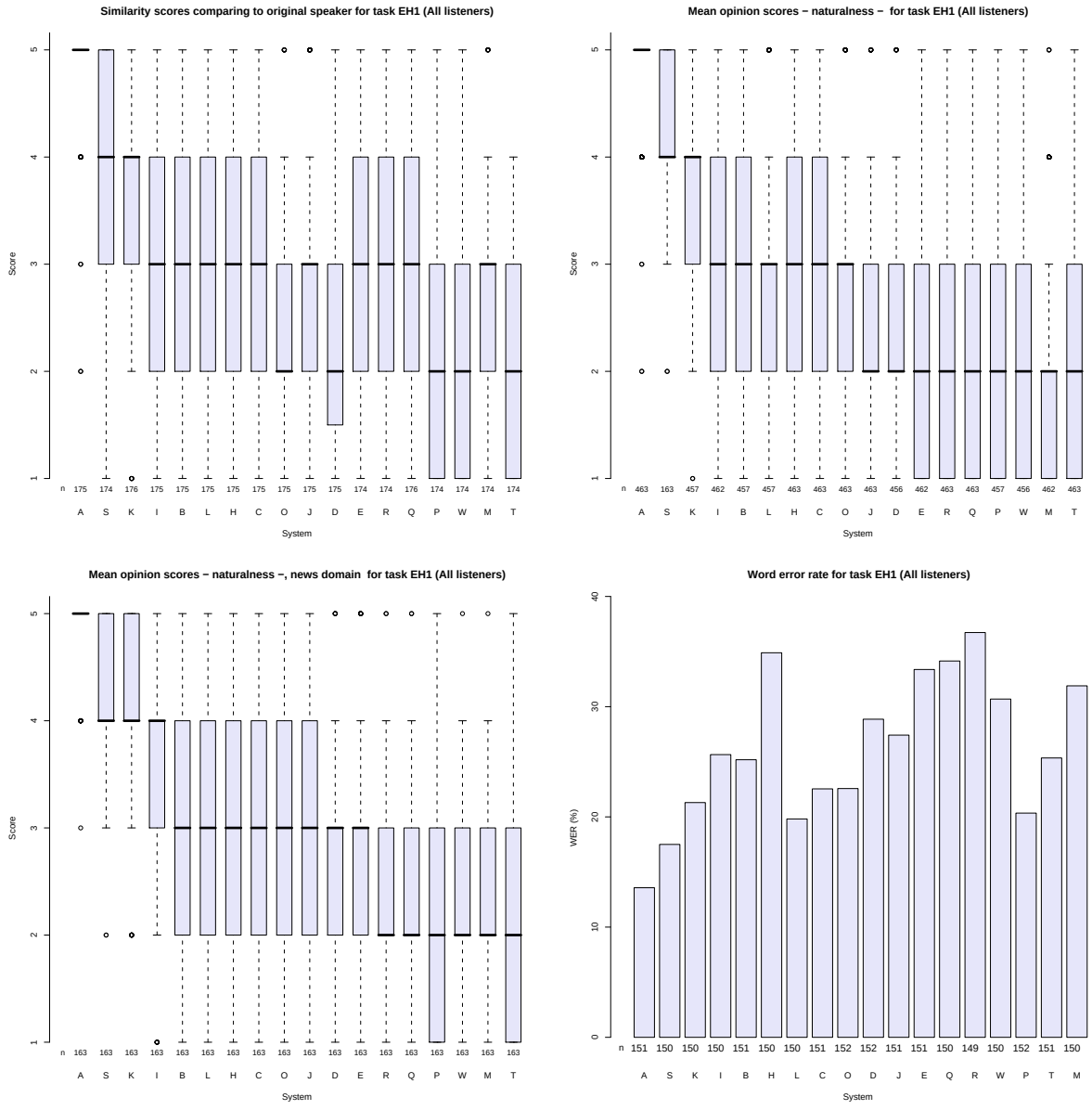


Figure 2: Results for task EH1.

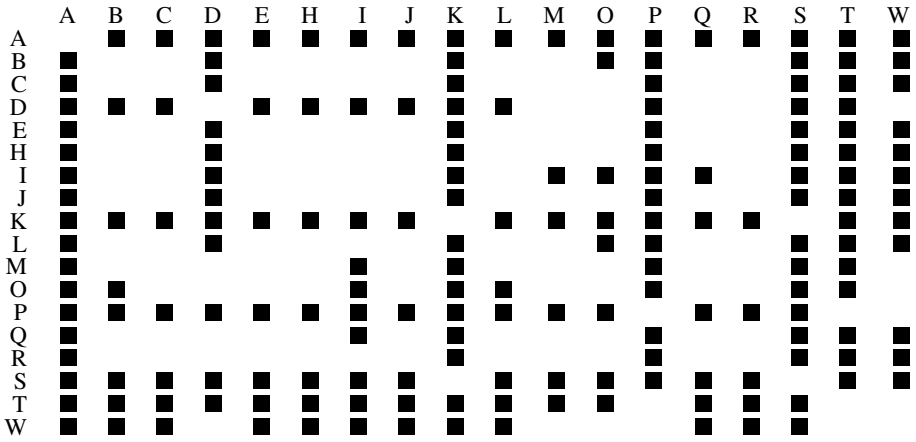


Table 10: Significant differences in similarity to the original speaker for task EH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

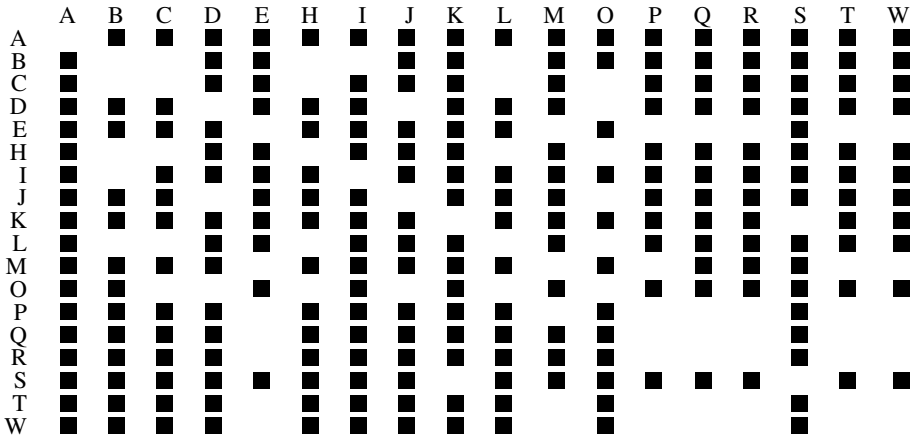


Table 11: Significant differences in naturalness for task EH1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

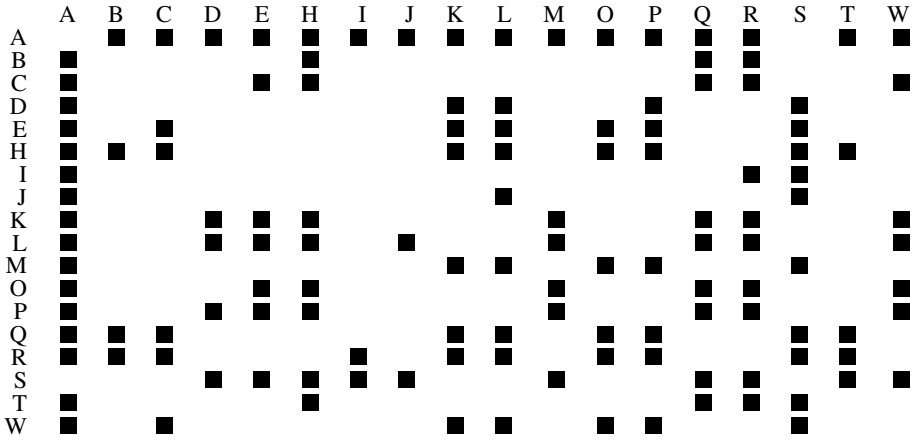
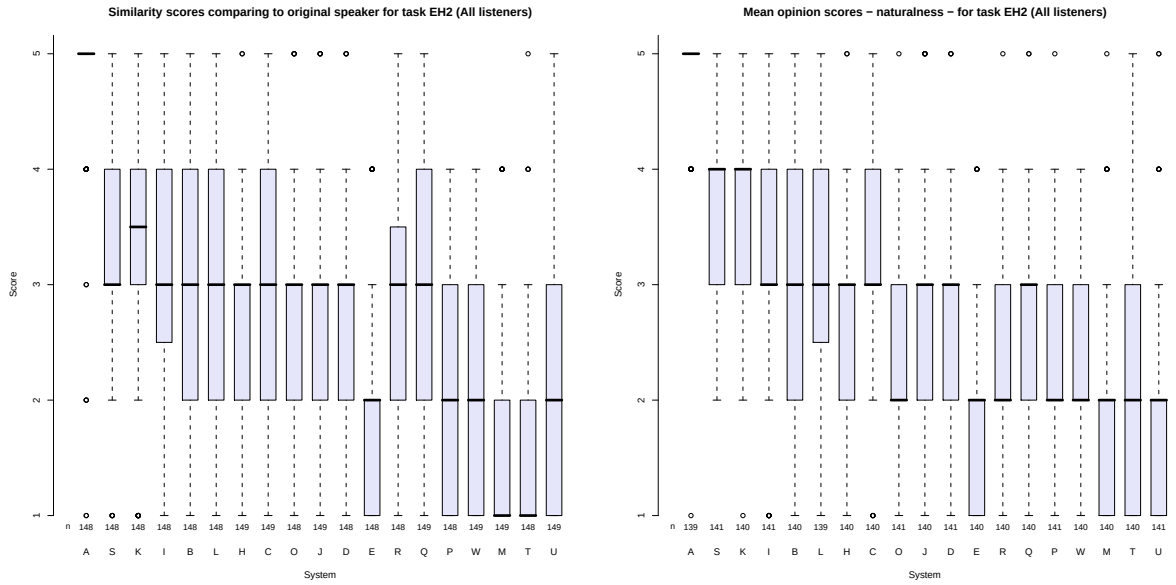


Table 12: Significant differences in intelligibility for task EH1: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.



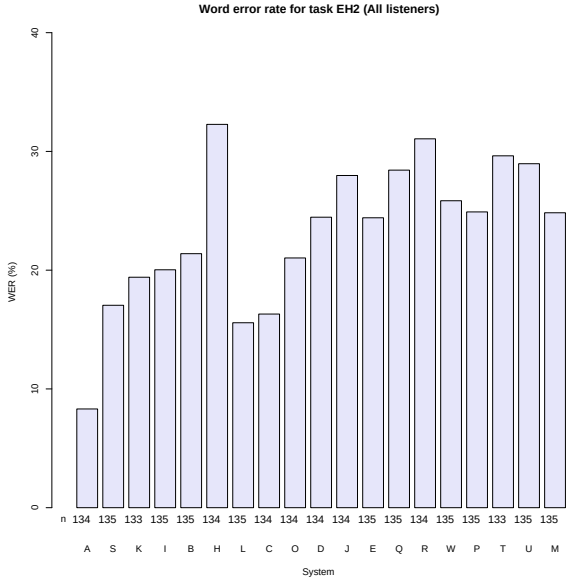


Figure 3: Results for task EH2.

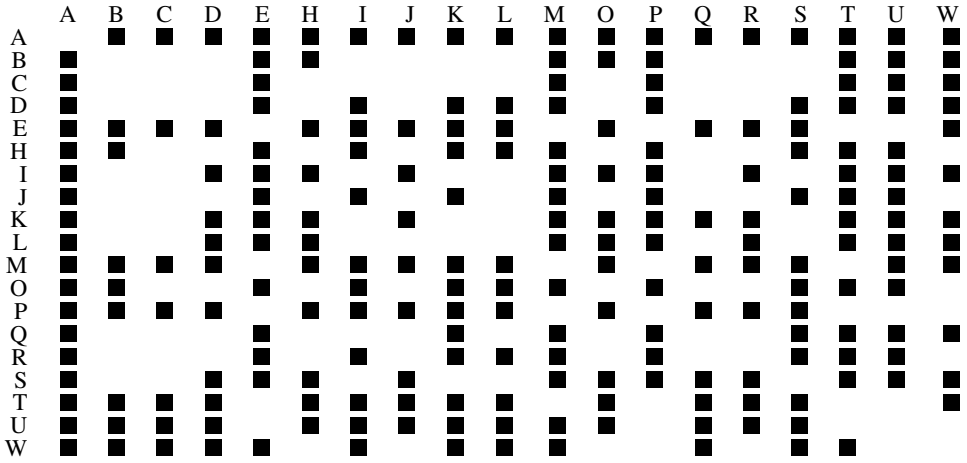


Table 13: Significant differences in similarity to the original speaker for task EH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

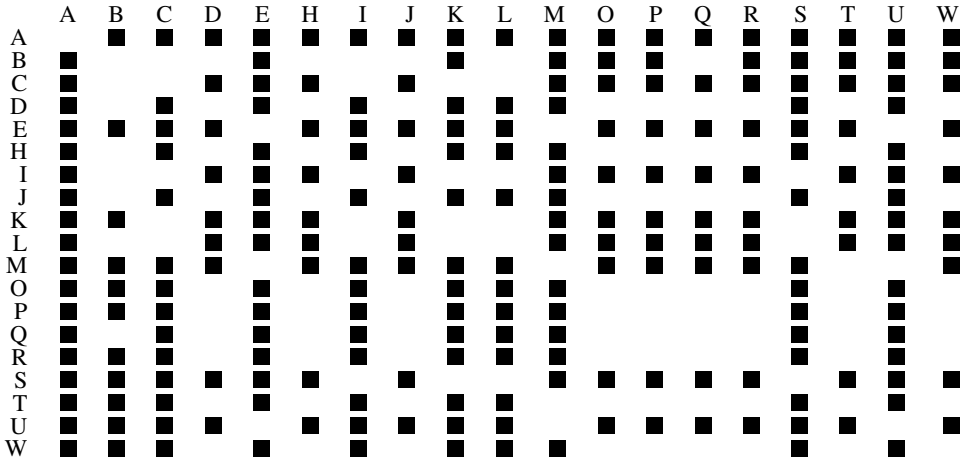


Table 14: Significant differences in naturalness for task EH2: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

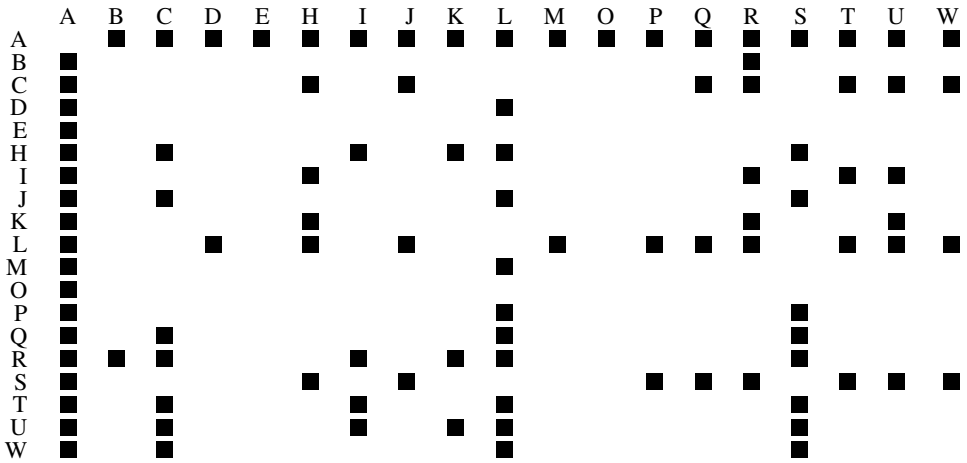


Table 15: Significant differences in intelligibility for task EH2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

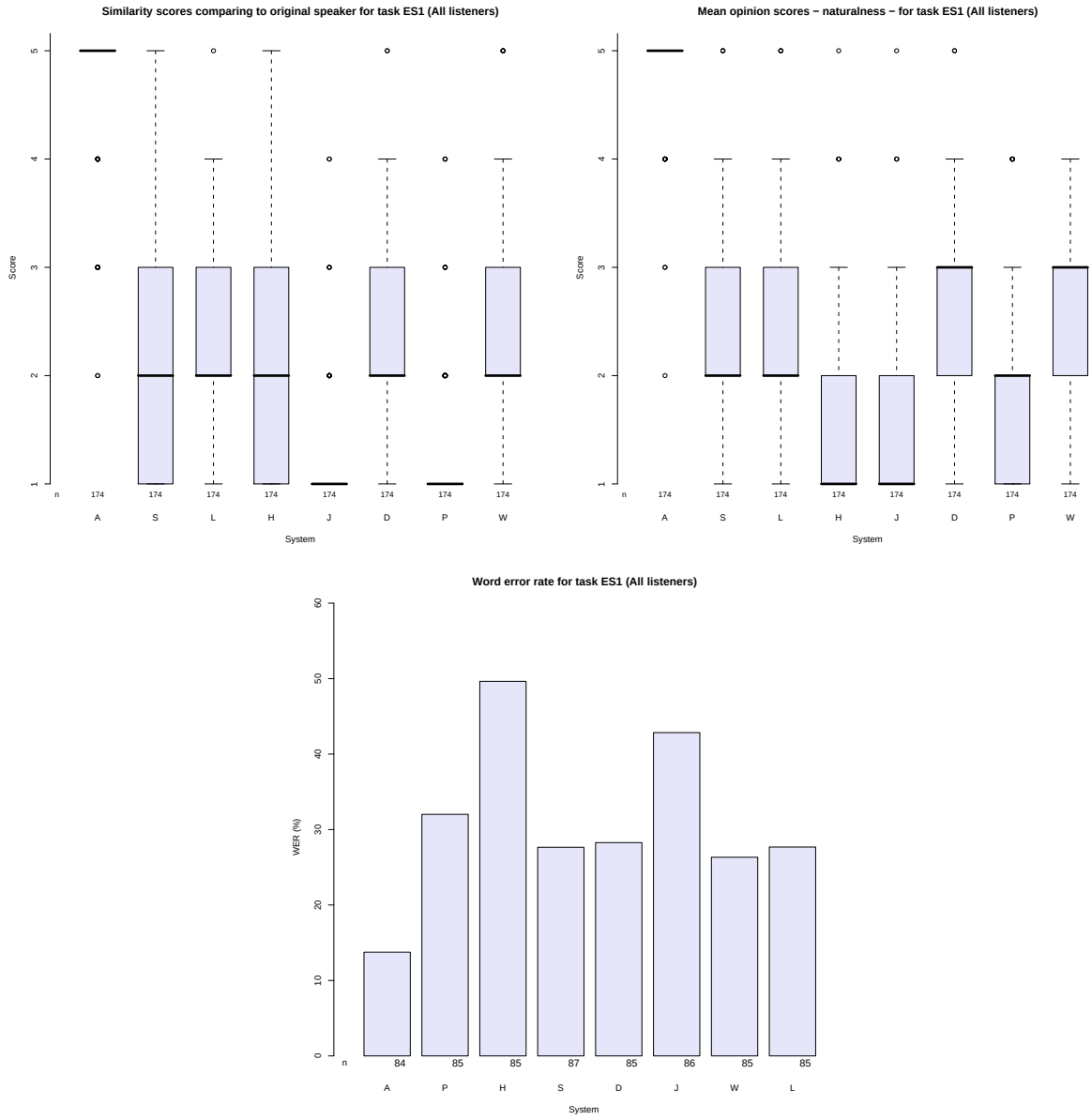


Figure 4: Results for task ES1.

	A	D	H	J	L	P	S	W
A		■	■	■	■	■	■	■
D	■		■	■		■		
H	■	■		■		■		■
J	■	■	■		■		■	■
L	■			■		■		
P	■	■	■		■		■	■
S	■			■		■		■
W	■		■	■		■	■	

	A	D	H	J	L	P	S	W
A		■	■	■	■	■	■	■
D	■		■	■		■		
H	■	■			■		■	■
J	■	■			■	■	■	■
L	■		■	■		■		■
P	■	■		■	■		■	■
S	■		■	■		■		
W	■		■	■	■	■		

	A	D	H	J	L	P	S	W
A		■	■	■	■	■	■	■
D	■		■	■				
H	■	■			■	■	■	■
J	■	■			■	■	■	■
L	■		■	■				
P	■		■	■				
S	■		■	■				
W	■		■	■				

Table 16: Significant differences in similarity to the original speaker (left table) and naturalness (middle table) and intelligibility (right table) for task ES1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

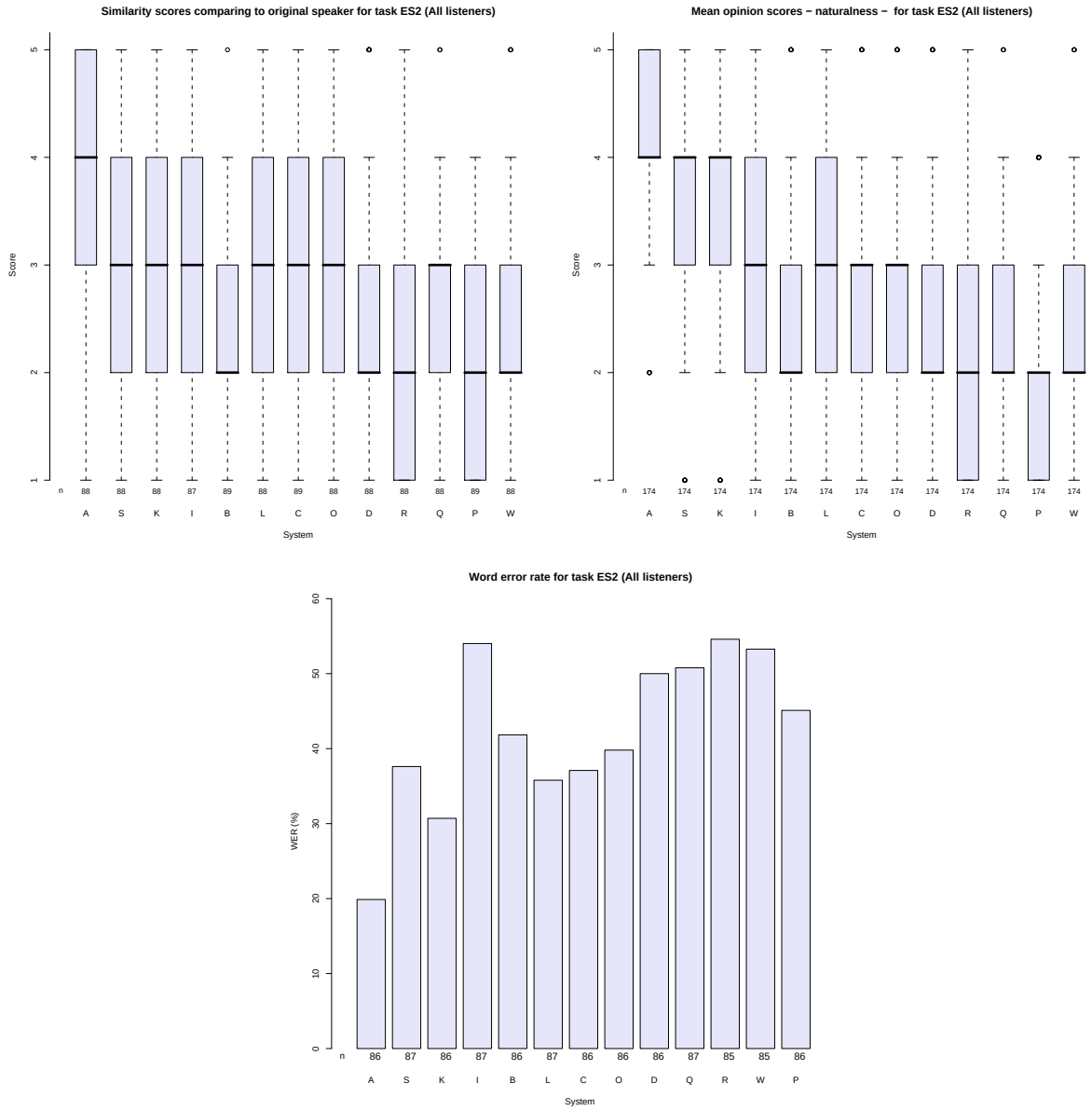


Figure 5: Results for task ES2.

	A	B	C	D	I	K	L	O	P	Q	R	S	W
A		■	■	■	■	■	■	■	■	■	■	■	■
B	■				■	■	■	■	■			■	
C	■					■	■		■		■	■	
D	■					■	■		■		■	■	
I	■	■				■			■	■	■	■	■
K	■	■	■	■	■		■	■	■	■	■		■
L	■	■	■	■		■			■	■	■	■	■
O	■	■				■			■	■	■	■	■
P	■	■	■	■	■	■	■	■		■		■	■
Q	■				■	■	■	■	■			■	
R	■		■	■	■	■	■	■				■	
S	■	■	■	■	■		■	■	■	■	■		■
W	■				■	■	■	■	■			■	

	A	B	C	D	I	K	L	O	P	Q	R	S	W
A		■	■	■	■	■	■	■	■	■	■	■	■
B	■					■					■		
C	■				■						■		■
D	■					■	■						
I	■		■			■	■					■	
K	■			■	■				■	■	■		■
L	■			■	■					■	■		■
O	■										■		
P	■					■							
Q	■					■	■						
R	■	■	■			■	■	■				■	
S	■				■						■		
W	■		■			■	■						

Table 17: Significant differences in naturalness (left table) and intelligibility (right table) for task ES2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

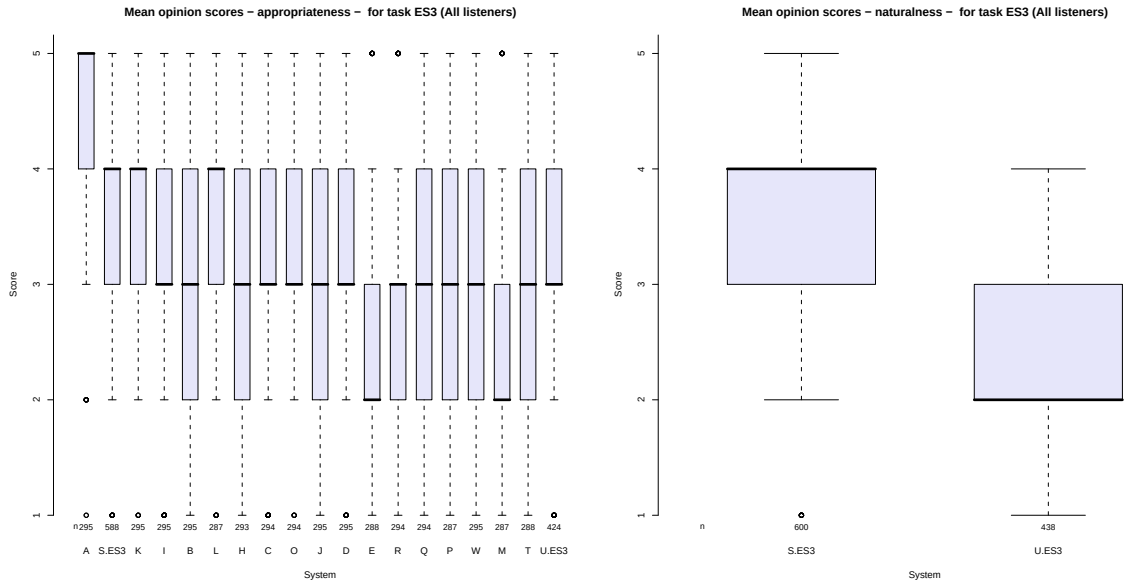


Figure 6: Results for task ES3.

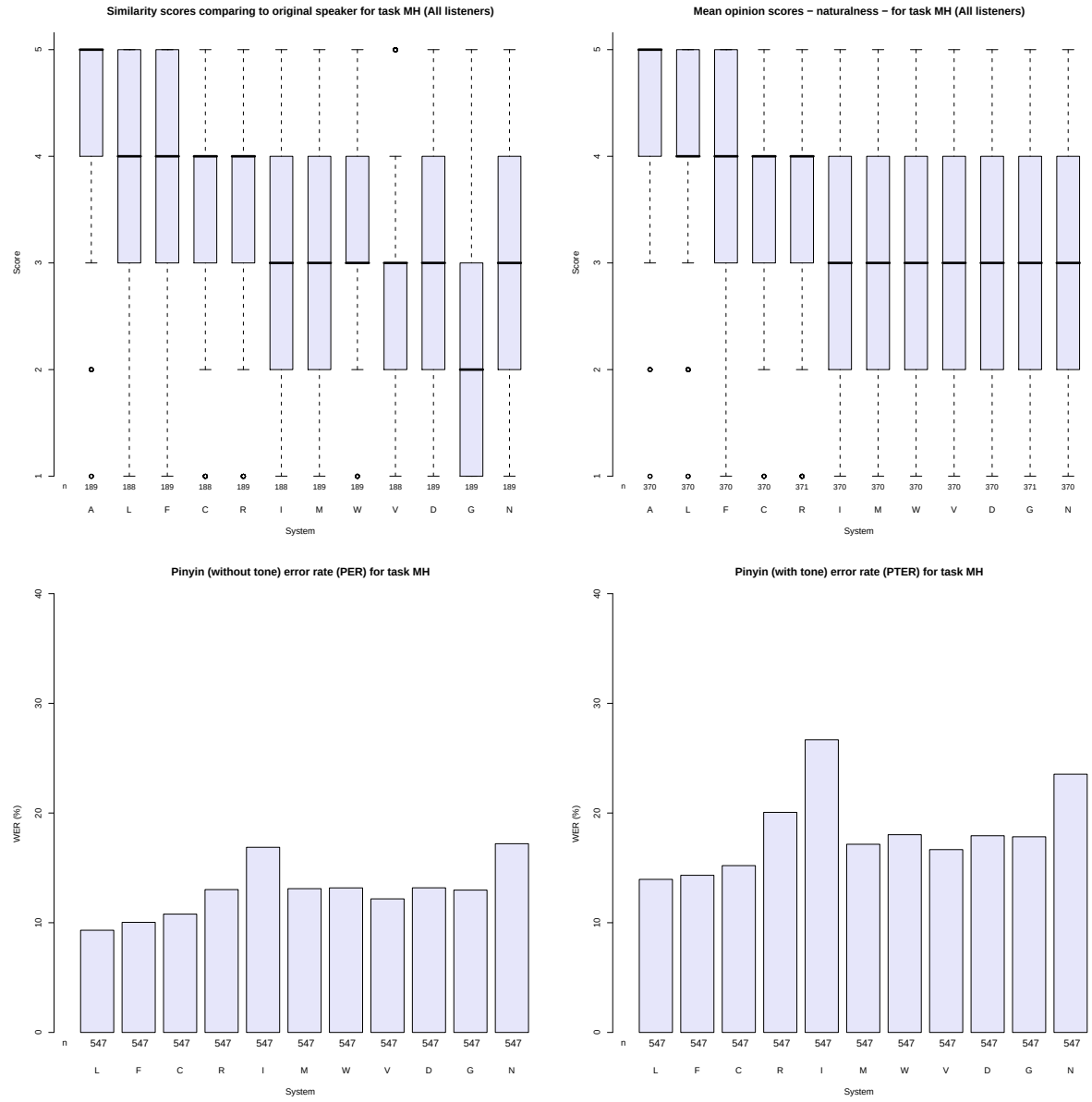


Figure 7: Results for task MH

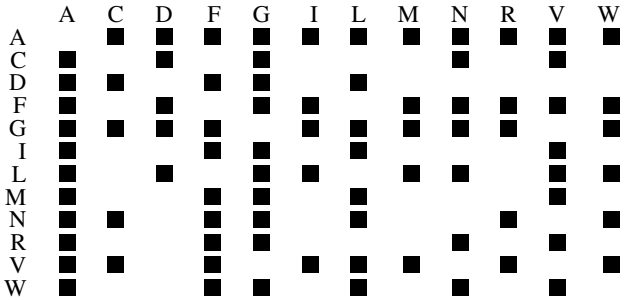


Table 18: Significant differences in similarity to the original speaker for task MH: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

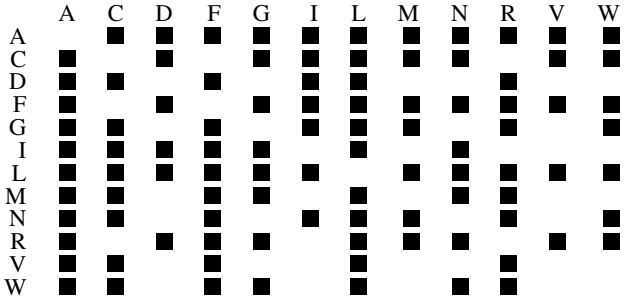


Table 19: Significant differences in naturalness for task MH: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

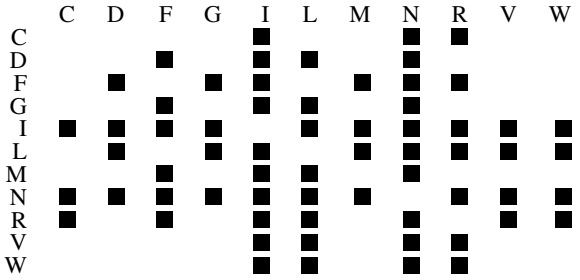


Table 20: Significant differences in intelligibility for task MH: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rate (PTER). ■ indicates a significant difference between a pair of systems.

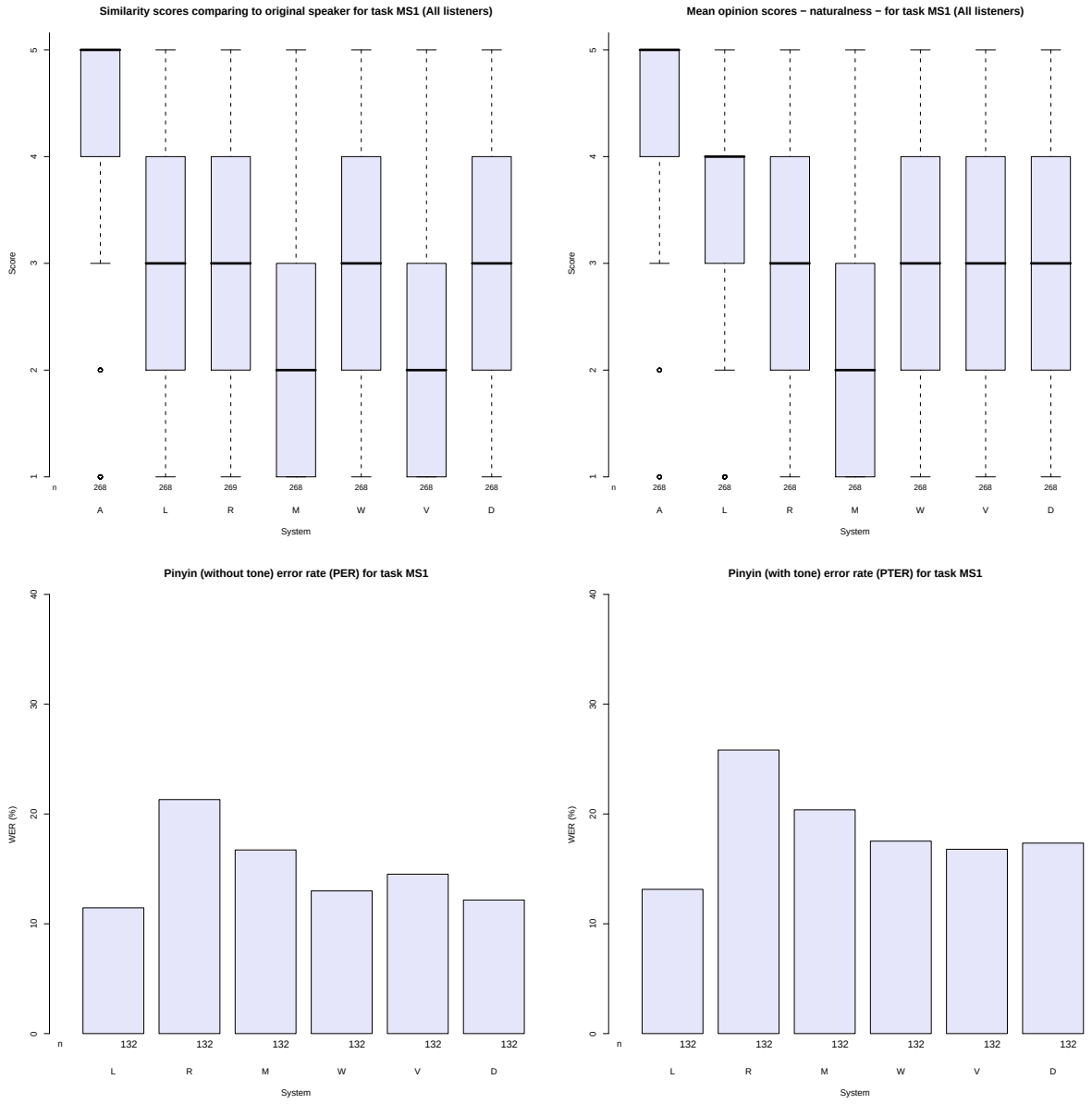


Figure 8: Results for task MS1

	A	D	L	M	R	V	W
A		■	■	■	■	■	■
D	■			■		■	
L	■			■		■	
M	■	■	■		■		■
R	■			■		■	■
V	■	■	■		■		■
W	■			■	■	■	

	A	D	L	M	R	V	W
A		■	■	■	■	■	■
D	■		■	■	■		
L	■	■		■	■	■	
M	■	■	■		■	■	■
R	■	■	■	■			■
V	■		■	■			■
W	■			■	■	■	

	D	L	M	R	V	W
D				■		
L			■	■		
M		■				
R	■	■			■	■
V				■		
W				■		

Table 21: Significant differences in similarity to the original speaker (left table) and naturalness (middle table) and intelligibility in terms of PTER (right table) for task MS1: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

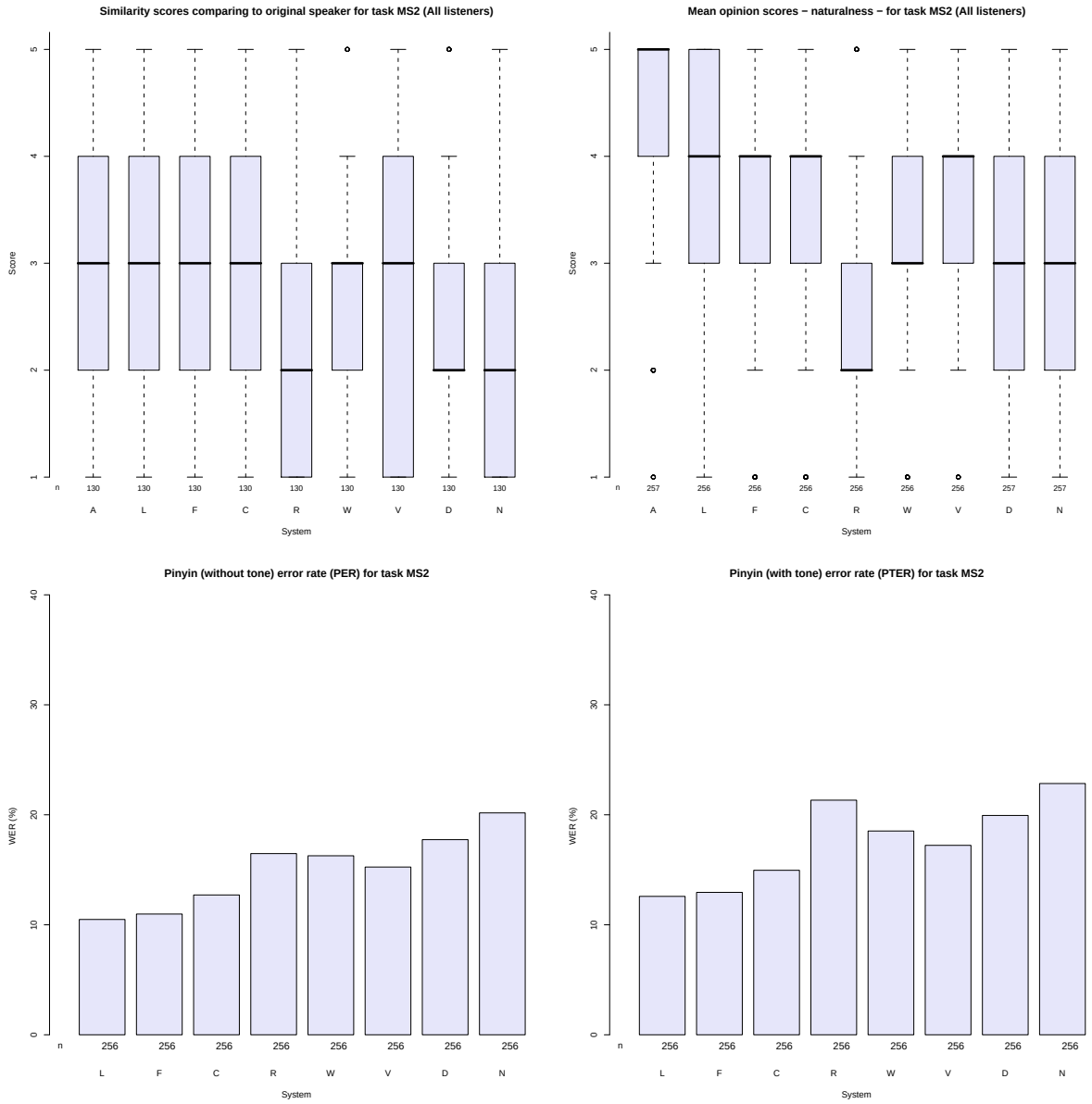


Figure 9: Results for task MS2

	A	C	D	F	L	N	R	V	W
A		■	■	■	■	■	■	■	■
C	■		■		■	■	■		
D	■	■		■	■		■	■	
F	■		■		■	■	■		
L	■	■	■	■		■	■	■	■
N	■	■		■	■		■	■	■
R	■	■	■	■	■	■		■	■
V	■		■		■	■	■		
W	■				■	■	■		

	C	D	F	L	N	R	V	W
C		■			■	■		
D	■		■	■				■
F		■			■	■	■	
L		■			■	■	■	■
N	■		■	■			■	■
R	■		■	■			■	■
V			■	■	■	■		
W		■		■	■	■		

Table 22: Significant differences in naturalness (left table) and intelligibility in terms of PTER (right table) for task MS2: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

Language	English total	Mandarin total
Amharic	1	0
Basque	1	0
Cantonese	1	0
Chinese	14	1
Czech	1	0
Danish	1	0
Dutch	3	0
Estonian	1	0
Finnish	4	0
French	5	0
German	9	0
Hebrew	2	0
Hindi	2	0
Hungarian	2	0
Japanese	35	0
Kannada	1	0
Korean	1	3 0
Mandarin	6	0
Norwegian	1	0
Polish	6	0
Portuguese	2	0
Russian	2	0
Slovak	1	0
Spanish	12	0
Swedish	2	0
Telugu	1	0
Turkish	1	0
Uighur	0	1
N/A	8	0

Table 23: First language of non-native speakers for English and Mandarin versions of Blizzard

Gender	Male	Female
English total	192	176
Mandarin total	160	144

Table 24: Gender

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
English total	39	273	79	23	7	7	0	0
Mandarin total	64	226	27	7	1	0	1	0

Table 25: Age of listeners whose results were used (completed the evaluation fully or partially

Native speaker	Yes	No
English	239	128
Mandarin	299	5

Table 26: Native speakers for English and Mandarin versions of Blizzard

	EH1	EH2	ES1	ES2	MH	MS1	MS2
ER	39	27	15	18	0	0	0
ES	58	41	21	22	0	0	0
EU	80	84	51	51	0	0	0
MC	0	0	0	0	117	86	86
ME	0	0	0	0	36	20	20
MR	0	0	0	0	15	12	8
MS	0	0	0	0	22	18	16
ALL	177	152	87	91	190	50	44

Table 27: Listener types per voice, showing the number of listeners whose responses were used in the results. Tasks ES1/ES2 and MS1/MS2 were bundled together, so most, but not all, of their respective listeners overlap.

	Registered	No response at all	Partial evaluation	Completed Evaluation
ER	125	39	38	48
ES	142	19	21	102
EU	215	0	0	215
ALL ENGLISH	**482**	**58**	**59**	**365**
MC	204	1	0	203
ME	56	0	0	56
MR	31	4	7	20
MS	43	4	7	32
ALL MANDARIN	**334**	**9**	**14**	**311**

Table 28: Listener registration and evaluation completion rates. For listeners assigned to do the ES1/ES2 and MS1/MS2 tests, finishing one but not both of the tests was included as partial completion.

	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	EH1_19	EH1_20
ER	3	3	3	3	3	3	3	1	3	1	1	1	2	3	2	2	1	0	1	0
ES	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
EU	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
ALL	9	10	10	10	10	10	10	8	10	8	8	8	9	10	10	8	8	7	8	7

Table 29: 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

	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	EH2_19	EH2_20	EH2_21
ER	1	1	1	0	1	1	1	1	1	1	2	1	2	1	2	1	2	2	2	2	1
ES	2	3	3	2	2	2	2	1	3	3	2	2	2	2	1	2	1	2	2	1	1
EU	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
ALL	7	8	8	6	7	7	7	6	8	8	8	7	8	7	7	7	7	8	8	7	6

Table 30: Listener groups - Voice EH2 (English), showing the number of listeners whose responses were used in the results

	ES1_01	ES1_02	ES1_03	ES1_04	ES1_05	ES1_06	ES1_07	ES1_08
ER	2	3	0	2	2	2	2	2
ES	4	3	3	2	2	2	3	2
EU	7	7	7	6	6	6	6	6
ALL	13	13	10	10	10	10	11	10

Table 31: Listener groups - Voice ES1 (English), showing the number of listeners whose responses were used in the results

	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
ER	1	2	2	0	1	2	1	1	2	2	0	2	2
ES	2	2	1	2	2	1	2	2	1	2	2	1	2
EU	4	4	4	4	4	4	4	4	4	4	4	4	3
ALL	7	8	7	6	7	7	7	7	7	8	6	7	7

Table 32: Listener groups - Voice ES2 (English), showing the number of listeners whose responses were used in the results

	MH_01	MH_02	MH_03	MH_04	MH_05	MH_06	MH_07	MH_08	MH_09	MH_10	MH_11	MH_12
MC	10	10	9	10	10	10	9	10	10	10	9	10
ME	3	3	3	3	3	3	3	3	3	3	3	3
MR	2	2	1	2	2	1	1	0	1	1	1	1
MS	3	2	2	1	1	2	1	2	2	2	1	2
ALL	18	17	15	16	16	16	14	15	16	16	14	16

Table 33: Listener groups - Voice MH (Mandarin), showing the number of listeners whose responses were used in the results

	MS1_01	MS1_02	MS1_03	MS1_04	MS1_05	MS1_06	MS1_07
MC	11	13	13	13	12	12	12
ME	3	3	3	3	3	3	2
MR	2	2	2	2	2	1	1
MS	3	3	3	3	2	2	2
ALL	19	21	21	21	19	18	17

Table 34: Listener groups - Voice MS1 (Mandarin), showing the number of listeners whose responses were used in the results

	MS2_01	MS2_02	MS2_03	MS2_04	MS2_05	MS2_06	MS2_07	MS2_08	MS2_09
MC	9	10	10	10	9	10	10	9	9
ME	3	3	2	2	2	2	2	2	2
MR	1	1	1	1	0	1	1	1	1
MS	2	2	2	2	1	2	2	1	1
ALL	15	16	15	15	12	15	15	13	13

Table 35: Listener groups - Voice MS2 (Mandarin), showing the number of listeners whose responses were used in the results

Listener Type	ER	ES	EU	ALL ENGLISH
Total	51	102	215	368

Table 36: Listener type totals for submitted feedback (English)

Listener Type	MC	ME	MR	MS	ALL MANDARIN
Total	201	44	18	33	296

Table 37: Listener type totals for submitted feedback (Mandarin)

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate
English total	48	65	94	104	50
Mandarin total	6	6	204	64	32

Table 38: Highest level of education completed

CS/Engineering person?	Yes	No
English total	149	215
Mandarin total	89	214

Table 39: Computer science / engineering person

Work in speech technology?	Yes	No
English total	131	234
Mandarin total	61	240

Table 40: Work in the field of speech technology

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
English total	58	54	44	74	81	26	30
Mandarin total	20	19	14	36	82	83	50

Table 41: How often normally listened to speech synthesis before doing the evaluation

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	1	1	169	33	13	22

Table 42: Dialect of English of native speakers

Dialect of Mandarin	Beijing	Shanghai	Guangdong	Sichuan	Northeast	Other	N/A
Total	47	7	8	17	11	156	53

Table 43: Dialect of Mandarin of native speakers

Level	Elementary	Intermediate	Advanced	Bilingual	N/A
English total	15	49	52	11	1
Madarin total	0	1	0	4	0

Table 44: Level of English/Mandarin of non-native speakers

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
English total	346	11	6	0
Mandarin total	263	36	5	0

Table 45: Speaker type used to listen to the speech samples

Same environment?	Yes	No
English total	359	4
Mandarin total	294	7

Table 46: Same environment for all samples?

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	281	71	13	0	0
Mandarin total	141	111	43	7	1

Table 47: Kind of environment when listening to the speech samples

Number of sessions	1	2-3	4 or more
English total	267	71	0
Mandarin total	208	75	0

Table 48: Number of separate listening sessions to complete all the sections

Browser	Firefox	IE	Mozilla	Netscape	Opera	Safari	Other
English total	61	78	1	5	0	207	0
Mandarin total	233	15	0	1	0	40	0

Table 49: Web browser used

Similarity with reference samples	Easy	Difficult
English total	266	100
Mandarin total	223	74

Table 50: Listeners' impression of their task in section(s) about similarity with original voice.

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
English total	43	4	49
Mandrin total	53	11	12

Table 51: Listeners' problems in section(s) about similarity with original voice.

Number of times	1-2	3-5	6 or more
English total	327	37	2
Mandarin total	177	116	1

Table 52: Number of times listened to each example in section(s) about similarity with original voice.

MDS section	Easy	Difficult
English total	269	91
Mandarin total	237	54

Table 53: Listeners' impression of their task in section about similarity of voice between two samples.

Problem	Unfamiliar task	Instructions not clear	Bad speakers, playing files disturbed others connection too slow, etc	Other
English total	33	8	1	41
Mandarin total	26	16	6	3

Table 54: Listeners' problems in section about similarity of voice between two samples.

Number of times	1-2	3-5	6 or more
English total	323	32	1
Mandarin total	193	95	0

Table 55: How many times listened to each example in section section about similarity of voice between two samples.

MOS naturalness sections	Easy	Difficult
English total	341	142
Mandarin total	275	92

Table 56: Listeners' impression of their task in MOS naturalness sections

Problem	All sounded same and/or too hard to understand	1 to 5 scale too big, too small, or confusing	Bad speakers, playing files disturbed others, connection too slow, etc	Other
English total	12	66	4	70
Mandarin total	22	63	12	13

Table 57: Listeners' problems in MOS naturalness sections

Number of times	1-2	3-5	6 or more
English total	355	53	3
Mandarin total	234	145	4

Table 58: How many times listened to each example in MOS naturalness sections?

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	30	203	112	19
Mandarin total	31	196	57	10

Table 59: Listeners' impressions of the task in SUS section(s)

Number of times	1-2	3-5	6 or more
English total	357	4	1
Mandarin total	81	202	11

Table 60: How many times listened to each example in SUS section(s)

MOS appropriateness sections	Easy	Difficult
English total	149	161

Table 61: Listeners' impression of their task in MOS appropriateness sections

Problem	All sounded same and/or too hard to understand	1 to 5 scale too big, too small, or confusing	Bad speakers, playing files disturbed others, connection too slow, etc	Other
English total	24	54	1	80

Table 62: Listeners' problems in MOS appropriateness sections

Number of times	1-2	3-5	6 or more
English total	292	19	1

Table 63: How many times listened to each example in MOS appropriateness sections?