The Blizzard Challenge 2008

Vasilis Karaiskos[a], Simon King[b], Robert A. J. Clark[b] and Catherine Mayo[b]

[a]School of Informatics, [b]Centre for Speech Technology Research,
University of Edinburgh
Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2008 was the fourth annual Blizzard Challenge. This year, participants were asked to build two voices from a UK English corpus and one voice from a Mandarin Chinese corpus. This is the first time that a language other than English has been included and also the first time that a large UK English corpus has been available. In addition, the English corpus contained somewhat more expressive speech than that found in corpora used in previous Blizzard Challenges.

To assist participants with limited resources or limited experience in UK-accented English or Mandarin, unaligned labels were provided for both corpora and for the test sentences. Participants could use the provided labels or create their own. An accent-specific pronunciation dictionary was also available for the English speaker.

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

The Blizzard Challenge was conceived by Black and Tokuda [1] and is the only open, international evaluation of corpus-based speech synthesisers. 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 2008 Challenge, the general structure of the listening test followed that of the 2007 Challenge.

The first two Blizzard Challenges, in 2005 and 2006, were organised by Carnegie Mellon University, USA, with the 2007 and 2008 Challenges being organised by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK.

For general details of Blizzard 2008, the rules of participation, a timeline, and information on previous and future Blizzard Challenges, see [2]. In this paper we summarise Blizzard 2008 – participants, voices to be built, evaluation design, results, and listener feedback – and consider possible designs for the next Blizzard Challenge.

2. Participants

The Blizzard Challenge 2005 [1, 3] had 6 participants, Blizzard 2006 had 14 [4]. In 2007, the number of entries increased again with 16 submitted entries. This year, there were 19 participants, 18 of who submitted samples for the English voice built using the full dataset, 16 submitted samples from a voice built using a smaller subset of the data, and 11 submitted Mandarin samples. A small number of additional groups, not listed here, registered for the Challenge and downloaded the corpora (and in some cases also the test sentences), but did not submit samples for evaluation.

- Aholab, University of the Basque Country, Spain[1]
- ATR Research Laboratories, Japan[2]
- Carnegie Mellon University, USA[1]
- CereProc Ltd, UK[2]
- DFKI GmbH, Germany[1]
- HTS working group (Nagoya Institute of Technology, Nara Institute of Science and Technology, University of Edinburgh), Japan and UK[2]
- I2R, Singapore[2]
- IBM Haifa Labs, Israel[2]
- IIIT Hyderabad, India[2]
- INESC-ID, Portugal[1]
- Institute of Automation, Chinese Academy of Sciences, China[2]
- mXac, Australia[1]
- Nokia Research Center Beijing, P.R. China[2]
- Technische Universitt Dresden, Germany[1]
- Toshiba, China[3]
- Universitat Politecnica de Catalunya, Spain[2]
- University of Science and Technology of China[2]
- University of Stellenboch, South Africa[2]
- Vrije Universiteit, Belgium[2]

Two systems from participants in previous challenges were used as benchmarks, in an attempt to calibrate the results from year to year: a Festival-based system from CSTR configured very similarly to the Festival/CSTR entry to Blizzard 2006 [5], and an HTS system configured the same as the HTS entry to Blizzard 2005 [6]. The Festival benchmark system was only used for the two English voices; the HTS system provided benchmarks for both English voices and the Mandarin voice. Although precise calibration of Mean Opinion Score (MOS) ratings is probably not possible, an approximate calibration of the relative rankings of systems may be. For example, if a participant was ranked lower than Festival in 2007 but higher in 2008, they may believe their system has improved.

[1]English only.
[2]English and Mandarin.
[3]Mandarin only.

3. Voices to be built

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. These data were previously unreleased. For Mandarin, the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, released a 6.5 hour Mandarin Chinese database of a female speaker with a standard Beijing accent. Participants were asked to build three synthetic voices from the database, using the same method, software, external data, and so on, as specified in the rules for participation [7] :

- Voice A: English full voice from the full dataset (about 15 hours)
- Voice B: English ARCTIC voice from the ARCTIC [8] subset (about 1 hour)
- Voice C: Mandarin voice from the full dataset (about 6.5 hours).

4. Listening test design

4.1. Interface

The listening evaluation was conducted online, using the design developed for Blizzard 2007 [9], which was itself developed from designs in previous challenges [1, 3, 4]. As an experiment, half of the English-speaking listeners run in Edinburgh used a version of the web-based test which employed a slider rather than a discrete scale to provide scores for sections 1, 3 & 4 of the test; the results of this experiment are presented in Section 7.

The registration page for each listener type presented an overview of the 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. As in previous Challenges, any individual listener in the English test heard stimulii exclusively from one voice (A or B). Listeners did not know which of the two voices they were listening to. Please refer to [9] for a fuller description of the listening test. The only substantive difference between the 2007 and 2008 tests was the use of a slider in place of the 5-point MOS scales for some listeners; the responses from these listeners are analysed in Section 7. The arrangement of listeners into groups and control of system orderings using Latin Squares was the same as in Blizzard 2007.

4.2. Materials

The participants were asked to synthesise several hundred test sentences, of which a subset were used in the listening test. Various sentences were gathered for use in future listening tests – such a corpus of synthetic speech from a variety of synthesisers will be a valuable resource for future research, including research on evaluation itself.

For English, participants synthesised sentences that had been held out from the corpus (so that natural speech samples were available for them) in the following genres:

- Conversational (100 sentences)
- News (100 sentences)
- Novel (200 sentences)
- Emphasis (20 sentences)

plus 200 Semantically Unpredictable Sentences (for which natural speech was recorded specially for the Blizzard Challenge). In the listening test, the Conversational and Emphasis sentences were not used. These are left for future experiments. The Conversation (or 'dialogue system') sentences may require a different listening test design which places them in an appropriate context (e.g., a simulated dialogue) for the listener – resources were not available to perform such a test within the Blizzard timescale. Since we did not know in advance which participants would employ specific techniques to realise the Emphasis sentences, we decided to leave their evaluation until a later date.

For Mandarin, 500 News sentences (the only genre available in this corpus) were held out from the corpus, to which 50 Semantically Unpredictable Sentences were added (natural speech was specially recorded for the subset of these used in the listening test). The Semantically Unpredictable Sentences (SUS) [10] for both English and Mandarin were kindly generated by Richard Sproat. In addition, participants were asked to synthesise the complete Blizzard Challenge 2007 test set, to be retained as a resource for future experimentation.

4.3. Listening test sentence selection

The sentences sent to participants were randomly selected from held out data. From these sentences, the relatively small number of sentences required in the listening test was randomly selected, after some sentences were excluded for one or more of the following reasons:

- Sentences with features that would be a test of text normalisation.
- Sentences containing foreign words.
- Sentences containing more than one sentence or, especially in Mandarin, sentences that appeared to be ungrammatical when read in isolation.
- Sentences that had multiple clearly ambiguous readings.
- Sentences where the natural example available was poorly read.

We also placed some loose restrictions on sentence length, in order to obtain a test set containing sentences of similar lengths. This was thought necessary because some sections of the listening test involve comparing pairs of sentences. In section 2 (multi-dimensional scaling) for English, the two sentences within a pair being compared were from the same genre - novel or news. The same sentences were also used in the MOS tests, so that MOS scores and position in MDS space could be compared.

4.4. Listener types

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

- 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 a quiet supervised lab using headphones. Half of them (EUL) used a conventional 5-point MOS scale and the other half (EUS) used a slider (with no numerical scale provided) to represent their scores in sections 1,3 and 4.
- Paid Indian students, who have studied in a University environment where English is the medium of instruction and is widely spoken for formal and informal communication, recruited at IIIT Hyderabad, India. These

listeners carried out the test in quiet supervised labs using headphones (EI).

- Volunteers recruited via participants, mailing lists, blogs, etc. (ER).
- Visually impaired volunteers (EVI)
- Volunteers over 50 years old (EO)

The EVI listener type was created 'on the fly' during the listening test, in response to user demand. The call for participation for the listening test was posted on a forum for blind computer users by a forum member. It was subsequently found that the embedded Quicktime player was not usable in conjunction with a screen reader, so an alternative version of the test was created for this listener type. Although many listeners started the test, few finished it. The cause for this is not known, but is likely to be linked to this problem with using a screenreader. Future Blizzard Challenges should take this into account; it is thought very likely that a sizeable number of such listeners could be obtained by advertising on appropriate mailing lists and forums, provided that a suitable version of the web-based test is available at the start of the testing period.

The EO listener type was created because of a specific interest in older listeners by a project underway at CSTR. However, lack of time prevented a concerted recruitment effort and we obtained very few listeners of this type. Again, it is thought that, with a little more effort, sufficient listeners (i.e., a minimum of one for each listener group in the Latin Square – see below) could be found for a future Blizzard Challenge. For Mandarin, the following listener types were used:

- Speech Experts, recruited via participants and mailing lists (MS).
- Paid undergraduate native speakers of Mandarin aged about 20-25, recruited in Edinburgh to do the evaluation in a quiet supervised lab using headphones (ME).
- 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).
- Volunteers, recruited via participants, mailing lists, etc. (MR).

4.5. Number of listeners

The listener responses used for the distributed results were extracted from the database on 23rd June 2008 for English, and on 25th June 2008 for Mandarin. The online evaluation had been run for approximately six weeks. The number of listeners obtained is shown in Table 1.

	English	Mandarin
Total registered	816	283
of which:		
Completed all sections	438	209
Partially completed	223	33
No response at all	155	41

Table 1: Number of listeners obtained.

See Table 19 for a detailed breakdown of evaluation completion rates for each listener type. From the Table, we can see that the reason for the much higher completion rate for Mandarin listeners is simply that the proportion of volunteer listeners is very low.

5. Analysis methodology

Since Bennett & Black [4] found that the statistics from listeners who completed the entire test and from those who completed only part of the test generally agree, we pooled 'completed all sections' and 'partially completed' listeners together in all analyses. MOS data from sections 3 and 4 is combined in the analysis presented here, although the raw data was provided to participants so that they were able to perform additional analysis, should they so desire. Word error rate (WER) for English was calculated automatically, using the same method as in 2007, which allows for typographical errors and spelling mistakes. Here, we present only results for all listener types combined, except in Section 7. Analysis by listener type was provided to participants.

For Mandarin, listener responses were first converted (where necessary) into simplified Chinese characters, and then three measures of intelligibility were computed:

- character error rate (CER)
- pinyin+tone error rate (PTER)
- pinyin error rate (PER)

The conversion from characters to pinyin+tone is a one-to-many mapping. Therefore, we considered all possible mappings from the character sequence to a pinyin+tone sequence (in the form of a lattice, for efficiency) and chose the sequence (path through the lattice) that minimised the PTER when compared to the correct transcription using a standard WER-like metric. The same procedure was used to compute PER, after stripping away the tone information from the listener response and the correct transcription. Note that CER is a rather harsh metric (it is analogous to WER for English without any spelling correction), so we recommend PTER be used as the primary measure of intelligibility. PER may be used to detect synthesisers that render tone incorrectly: such systems will have a large difference between PER and PTER, relative to other systems.

As in previous years, system names were anonymised in all distributed results. Raw listener response data for sections 1,3, 4 and 5 were also distributed to participants along with background information about each anonymised listener extracted from responses to the listener feedback questionnaire presented on completion of the evaluation. See Section 8.2 and Tables 23 to 47 for a summary of this information. Note that completion of the questionnaire was optional for listeners.

The statistical analysis for Blizzard 2008 followed that for 2007. Please refer to [11] for a complete description of the techniques used and justification of the statistical significance techniques employed. The statistical calculations regarding significant differences in WER assume that there are the same number of words in every sentence. This is not strictly true for Mandarin, where there are small variations in the sentence length (number of characters per sentence). To examine the effect of this, we computed the overall CER, PTER and PER using two methods: the correct method (which sums insertions, deletions, substitutions and reference transcription length individually across all test sentences, then makes a final error rate calculation) and an incorrect method (which computes the CER, PTER or PER on a per sentence basis, then averages these). The incorrect method will give the same result as the correct method in the special case where all sentences have the same length. We found very small differences in the results from two methods, which indicates that the variations in sentence length are negligible. Therefore, for Mandarin, we used the same statistical significance testing method as for English, which assumes a fixed sentence length.

6. Results

In the following, the results are presented first for English voice A (the voice using the full English data set), then for English voice B (using the smaller ARCTIC data set), and finally for Mandarin. Please note that, although the listener responses were collected for section 2 (in which pairs of stimulii are compared), the multi-dimensional scaling analysis is left for future work. In section 1, listeners (other than type EUL) were asked to use a 5 point scale with the endpoints labelled "1: Sounds like a totally different person" and "5: Sounds like exactly the same person" and in section 3 & 4 the scale was labelled "1: Completely Unnatural" to "5: Completely Natural".

As in Blizzard 2007, 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. The ordering of the systems in the plots is in descending order of the mean MOS for sections 3 and 4 combined – see Tables 2 and 10. 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.

6.1. English voice A

Table 2 presents descriptive statistics for the mean opinion scores for English voice A. Figure 1 displays the results of the tests for English voice A 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 4 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. System J is also significantly different from all other systems. We may therefore say that although system J is significantly less natural than the natural speech, is is significantly more natural than all other systems for English voice A – in other words, it is the most natural synthesiser for this voice. From the plot of similarity scores and by referring to Table 3, we can also say that, although system J is significantly less similar to the original speaker than natural speech, it is significantly more similar to the original speaker than all other systems, for English voice A.

6.2. English voice B

For English voice B, results are illustrated in Figure 2 with statistical significance shown in Table 7 for similarity and Table 8 for naturalness. Again, system J is more similar to the original speaker than all other synthesiers, although still significantly less similar to the original speaker than the natural speech itself. Now, systems J, S and V are all equally natural, and significantly more natural than all other systems but significantly less natural than natural speech. In other words, systems J, S and V are jointly the most natural synthesisers for this voice.

6.3. Mandarin

Table 10 and Figure 3 presents the corresponding results for Mandarin. Again, natural speech (System A) has a median MOS of 5. The significance tests illustrated in Table12 show that once again no system is as natural as the natural speech. The most natural synthesisers are U, C, F, T, V and F. There are no significant differences between systems in this group

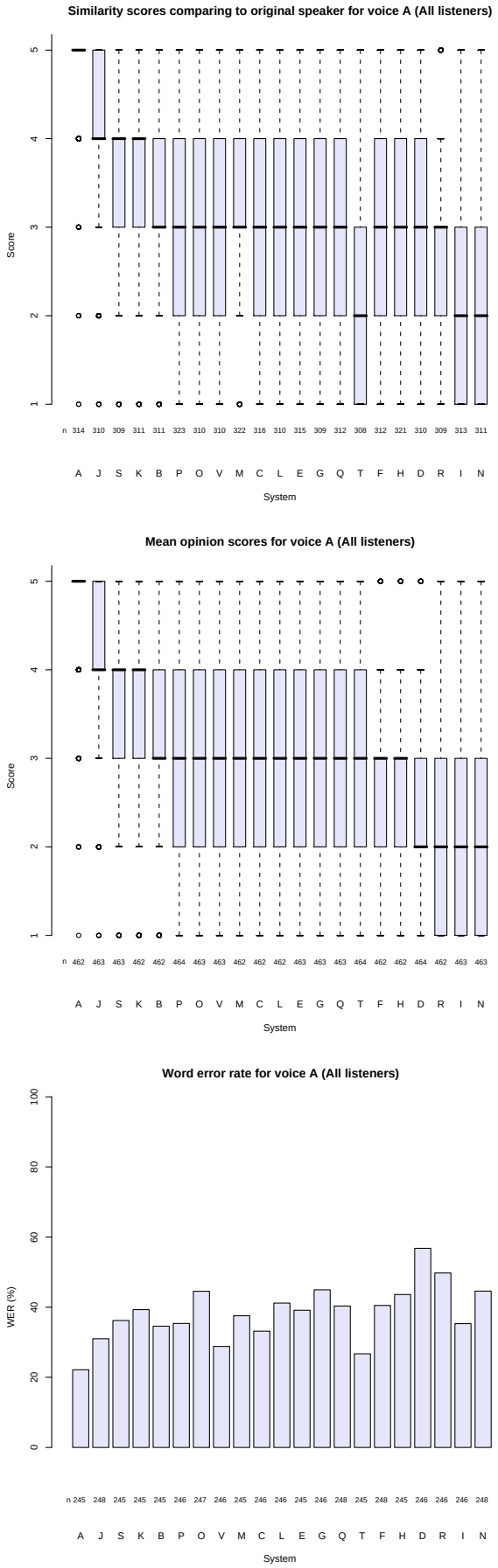


Figure 1: Results for English voice A.

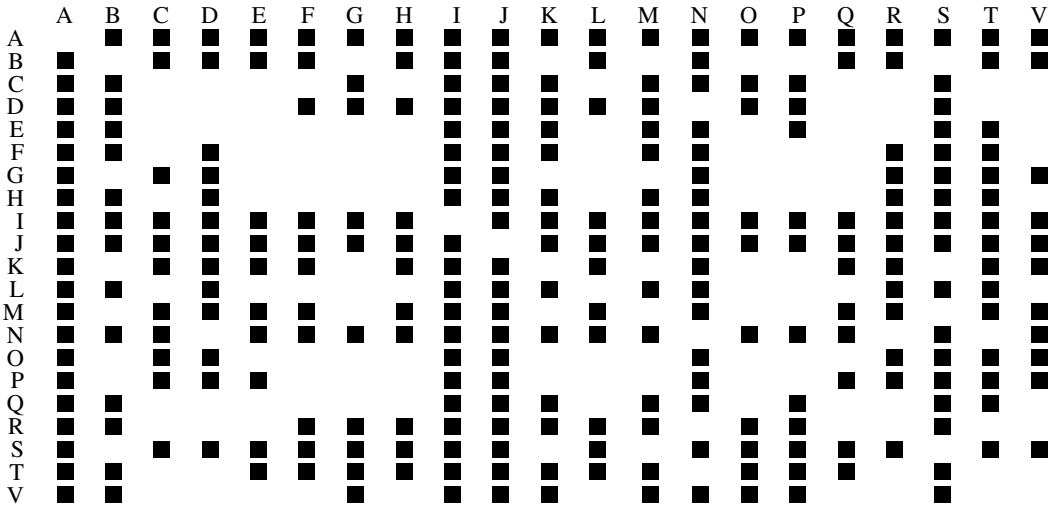


Table 3: Significant differences in similarity to the original speaker for English voice A: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

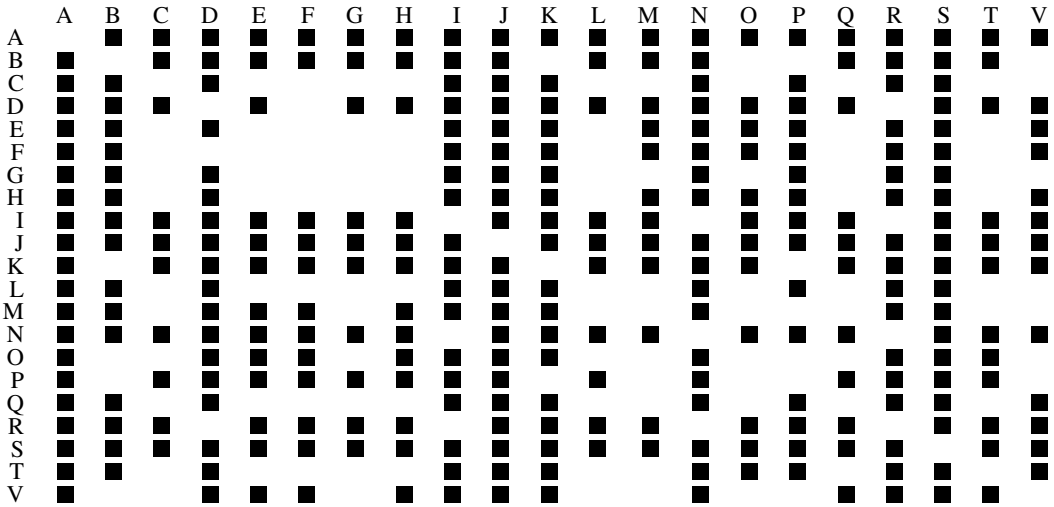


Table 4: Significant differences in naturalness for English voice A: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

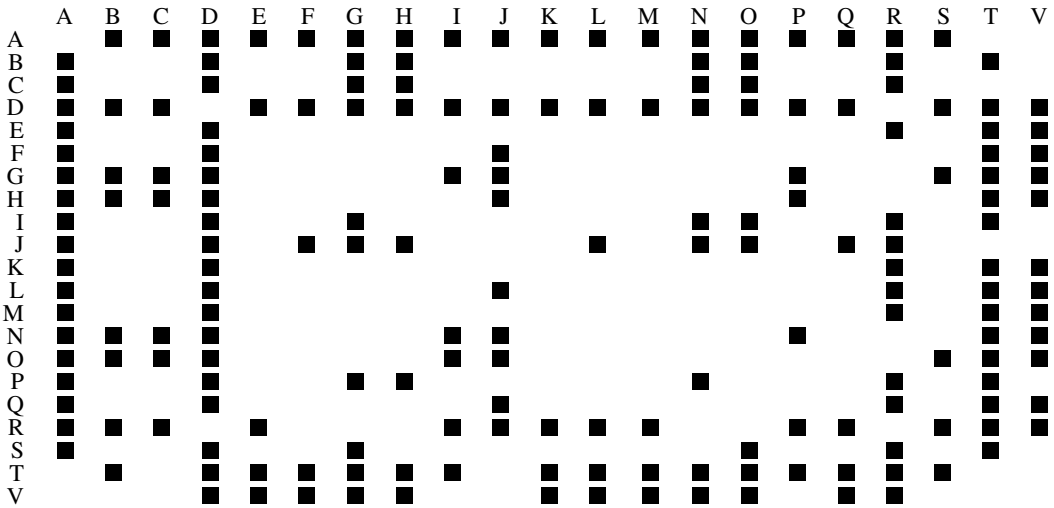


Table 5: Significant differences in intelligibility for English voice A: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.

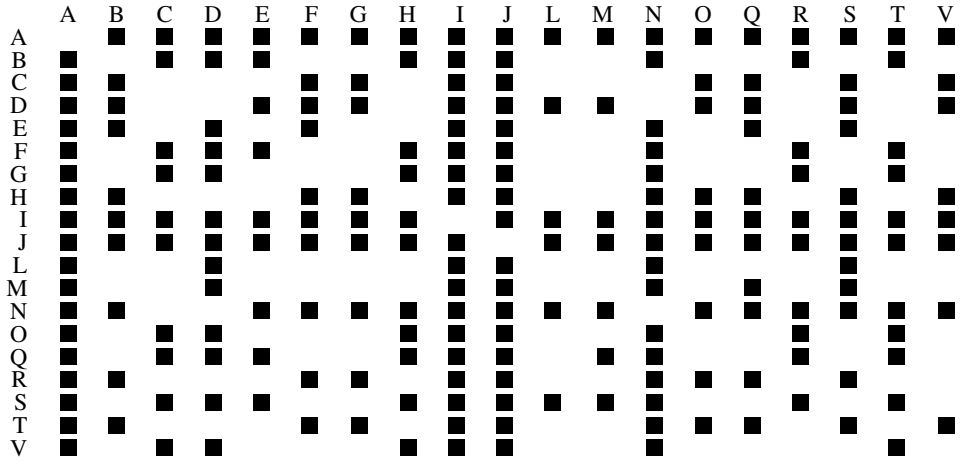


Table 7: Significant differences in similarity to the original speaker for English voice B: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

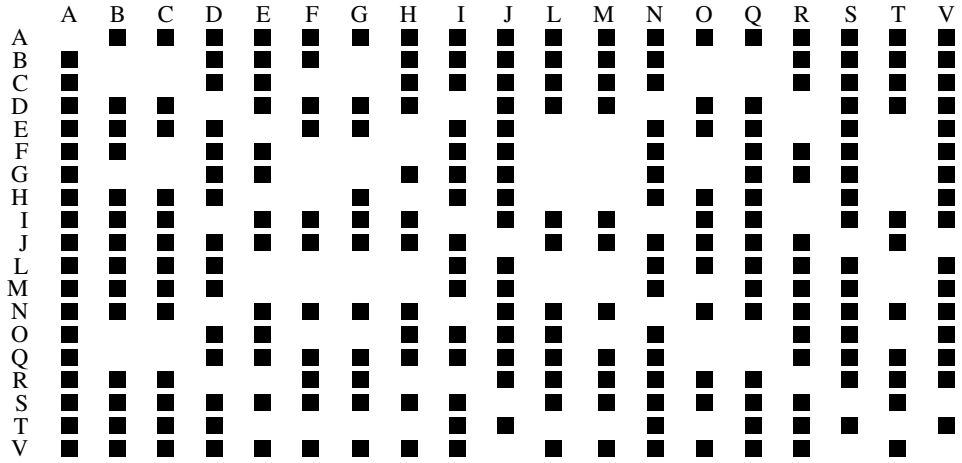


Table 8: Significant differences in naturalness for English voice B: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

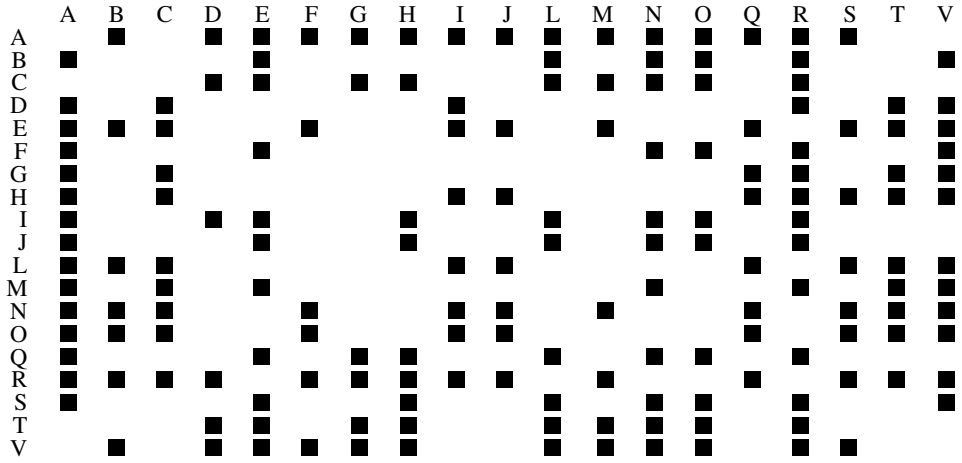
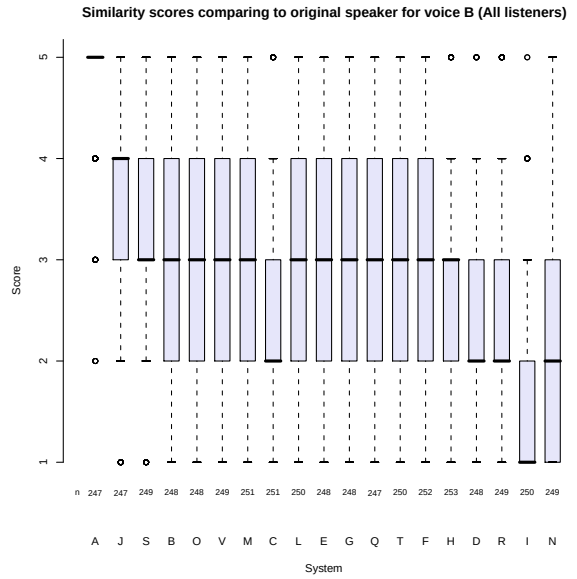
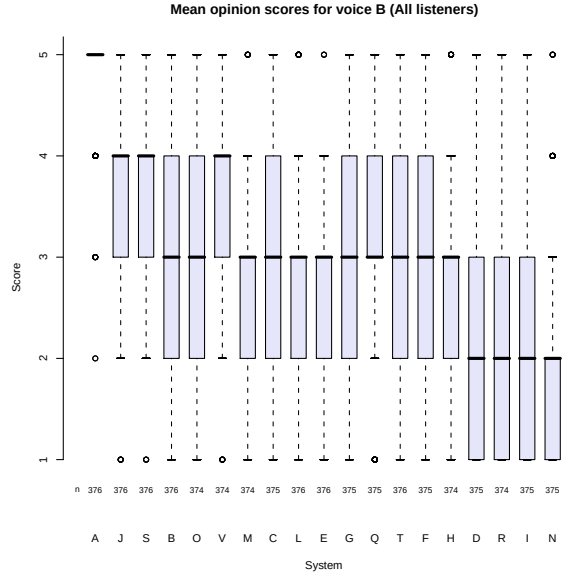


Table 9: Significant differences in intelligibility for English voice B: results of pairwise Wilcoxon signed rank tests between systems' word error rates. ■ indicates a significant difference between a pair of systems.





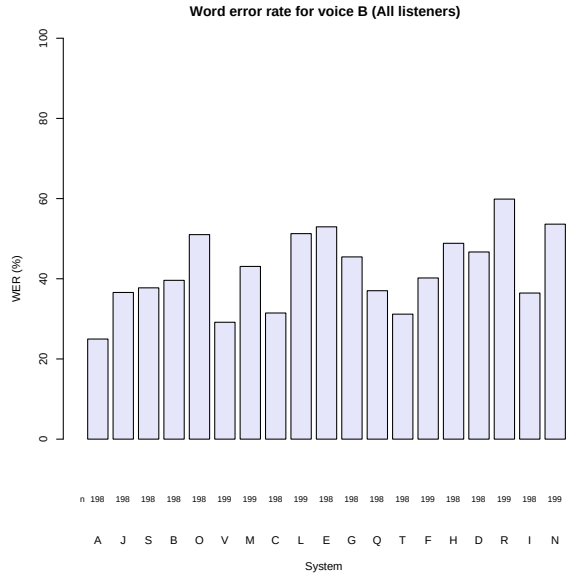


Figure 2: Results for English voice B.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.8	0.57	462	240
B	3	1.5	3.3	1.05	462	240
C	3	1.5	2.9	1.09	462	240
D	2	1.5	2.4	1.11	464	238
E	3	1.5	2.8	1.07	463	239
F	3	1.5	2.7	1.08	462	240
G	3	1.5	2.8	1.18	463	239
H	3	1.5	2.7	1.04	462	240
I	2	1.5	2.1	1.13	463	239
J	4	1.5	4.1	0.91	463	239
K	4	1.5	3.4	1.13	462	240
L	3	1.5	2.9	1.08	462	240
M	3	1.5	3.0	1.03	462	240
N	2	1.5	2.0	0.98	463	239
O	3	1.5	3.1	1.22	463	239
P	3	1.5	3.2	1.07	464	238
Q	3	1.5	2.8	1.08	463	239
R	2	1.5	2.2	1.03	462	240
S	4	1.5	3.7	1.00	463	239
T	3	1.5	2.7	1.11	464	238
V	3	1.5	3.1	1.10	463	239

Table 2: Mean opinion scores for voice A (full data set) on the combined results from sections 3 and 4 of the evaluation. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and NA (data points excluded due to missing data)

System	median	MAD	mean	sd	n	na
A	5	0.0	4.8	0.48	376	170
B	3	1.5	3.1	1.00	376	170
C	3	1.5	3.0	1.12	375	171
D	2	1.5	2.1	0.88	375	171
E	3	1.5	2.5	0.98	376	170
F	3	1.5	2.8	1.02	375	171
G	3	1.5	2.9	1.00	375	171
H	3	1.5	2.6	1.04	374	172
I	2	1.5	2.1	1.06	375	171
J	4	1.5	3.8	0.94	376	170
L	3	1.5	2.7	0.98	376	170
M	3	1.5	2.7	1.00	374	172
N	2	1.5	1.9	0.95	375	171
O	3	1.5	2.9	1.08	374	172
Q	3	1.5	3.2	1.00	375	171
R	2	1.5	2.3	1.12	374	172
S	4	1.5	3.6	0.86	376	170
T	3	1.5	2.7	1.06	376	170
V	4	1.5	3.6	0.92	374	172

Table 6: Mean opinion scores for voice B on the combined results from sections 3 and 4 of the evaluation. Table shows median, median absolute deviation (MAD), mean, standard deviation (sd), n and NA (data points excluded due to missing data)

and they are each significantly different to all other systems, with the exception that systems T,V and F are not significantly different to system O. The reduced number of statistical differences for Mandarin, compared to English, reflects the smaller number of listeners obtained for the listening test. Many differences between systems are either too small, or too inconsistent, to discern using this number of listeners.

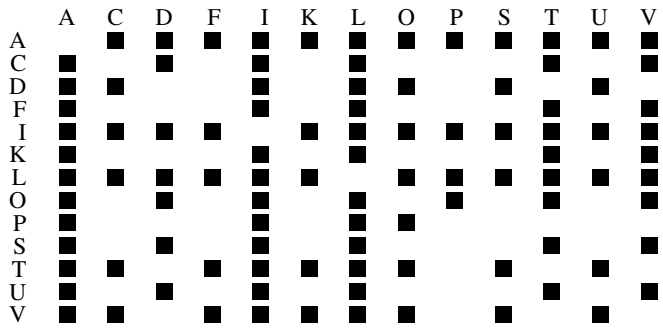


Table 11: Significant differences in similarity to the original speaker for Mandarin: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

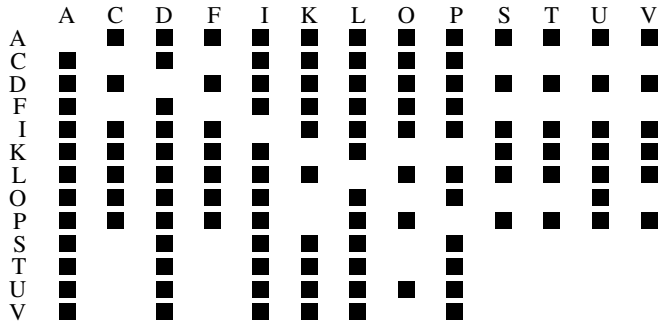


Table 12: Significant differences in naturalness for Mandarin: results of pairwise Wilcoxon signed rank tests between systems' mean opinion scores. ■ indicates a significant difference between a pair of systems.

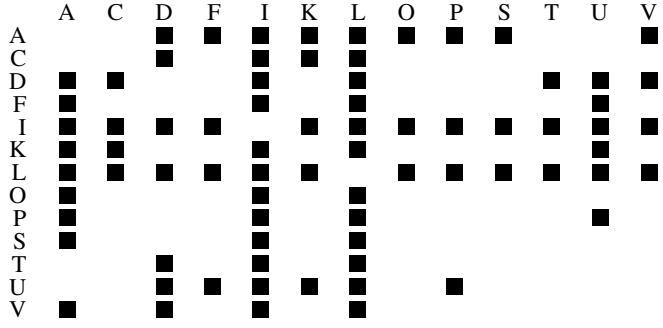


Table 13: Significant differences in intelligibility for Mandarin: results of pairwise Wilcoxon signed rank tests between systems' pinyin+tone error rate (PTER). ■ indicates a significant difference between a pair of systems.

System	median	MAD	mean	sd	n	na
A	5	0.0	4.4	1.0	428	56
C	4	1.5	3.6	1.0	428	56
D	2	1.5	2.6	1.1	429	55
F	4	1.5	3.6	1.0	429	55
I	1	0.0	1.6	1.0	430	54
K	3	1.5	3.1	1.1	430	54
L	1	0.0	1.8	1.1	428	56
O	3	1.5	3.3	1.1	427	57
P	3	1.5	3.0	1.1	428	56
S	4	1.5	3.4	1.0	428	56
T	4	1.5	3.5	1.0	427	57
U	4	1.5	3.7	1.0	428	56
V	4	1.5	3.5	1.0	427	57

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

7. The use of continuous rating scales

We now discuss the results obtained from the EUL and EUS groups of listeners. Listeners of type EUL performed the standard version of the test (using a 5 point scale), whereas listeners of type EUS were presented with an uncalibrated slider instead, for sections 1, 3 and 4. The limitations of the web interface meant that this slider was of a finite length. This is in contrast to open-ended magnitude estimation scales as used by [12]. The slider was implemented as a 101 (0–100) point scale, but the listeners were not presented with any markings or scale, and therefore see the slider as continuously variable.

The results are shown in Figure 5 where it can be seen that the slider-based results are largely consistent with those obtained using the 5-point scale. With this small group of listeners (21 in each of the EUL and EUS groups) we must interpret the results with caution. The difference in naturalness between natural speech and even the most natural of the synthesisers is quite clear (lower right plot in Figure 5). UK undergraduate listeners appear to give a lower score (whether with the continuous or Lickert scales) to systems C, T and V than the general population of listeners: these systems are all HMM-based.

We can test the hypothesis that there is a direct linear relationship between the way in which listeners use the slider scales and the 5-point scales by using linear regression. We propose the following linear model:

$$S_i = \beta_1 F_i + \beta_0$$

where S_i is the mean slider value for a system, F_i is the mean five-point scale value for system i and β_0 and β_1 are the intercept and gradient coefficients respectively. The regression line for the listener types EUL and EUS are shown in Figure 6. We can see here the clear strong positive correlation between the 5 point scale results and the slider results. The coefficients for the regression are -24.9 for the intercept and 23.5 for the gradient, both are significant at 0.001%.

If the slider, which has underlying values 0–100, and the 5-point scale, with values 1-5 were used in the same way by listeners, then the model coefficients would be $\beta_0 = -25$ and $\beta_1 = 25$. We can use t-tests to test the hypothesis that these values for the model coefficients are significantly different from the values found from our data ($\beta_0 = -24.9$ and $\beta_1 = 23.5$). For the intercept coefficient β_1, we find $t = 0.029, df = 19, p = 0.977$; for the gradient coefficient β_0 we find $t = 1.093, df = 19, p = 0.288$. As neither of these results are significant, we are unable to reject the hypothesis. We conclude that the listeners use the two types of scale in the same way. Each listener only performed ratings on one scale or the other, so it is not possible to perform an analysis of the behaviour of individual subjects.

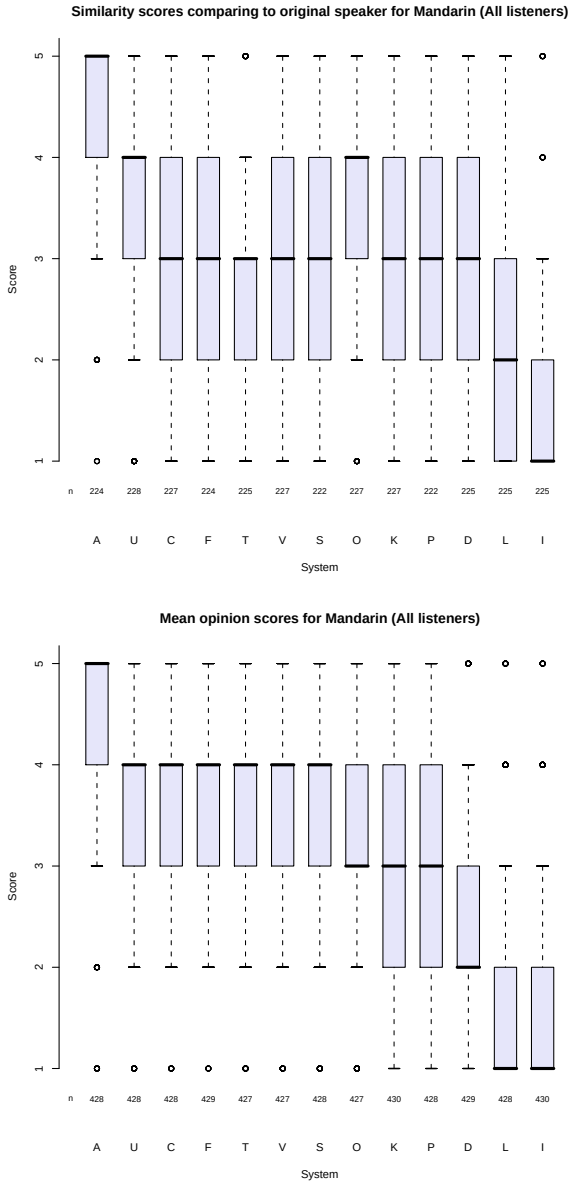


Figure 3: Results for Mandarin: similarity and naturalness

This result strongly suggests that listeners are treating the 5-point Lickert scale not as an ordinal scale but as an interval scale, where the differences between successive points on the scale are equal in magnitude.

8. Discussion

8.1. Listener recruitment and completion rates

For English, 97 out of 143 registered speech experts completed the evaluation – compared to 163 of 202 in 2007–. 23 registered but did not do any of the test. On the other hand, we had many more volunteers (types ER, EVI, EO[4], see Section

[4]To date we have not collected enough data to report separately on visually impaired (EVI), and older (EO) volunteers. At the time of

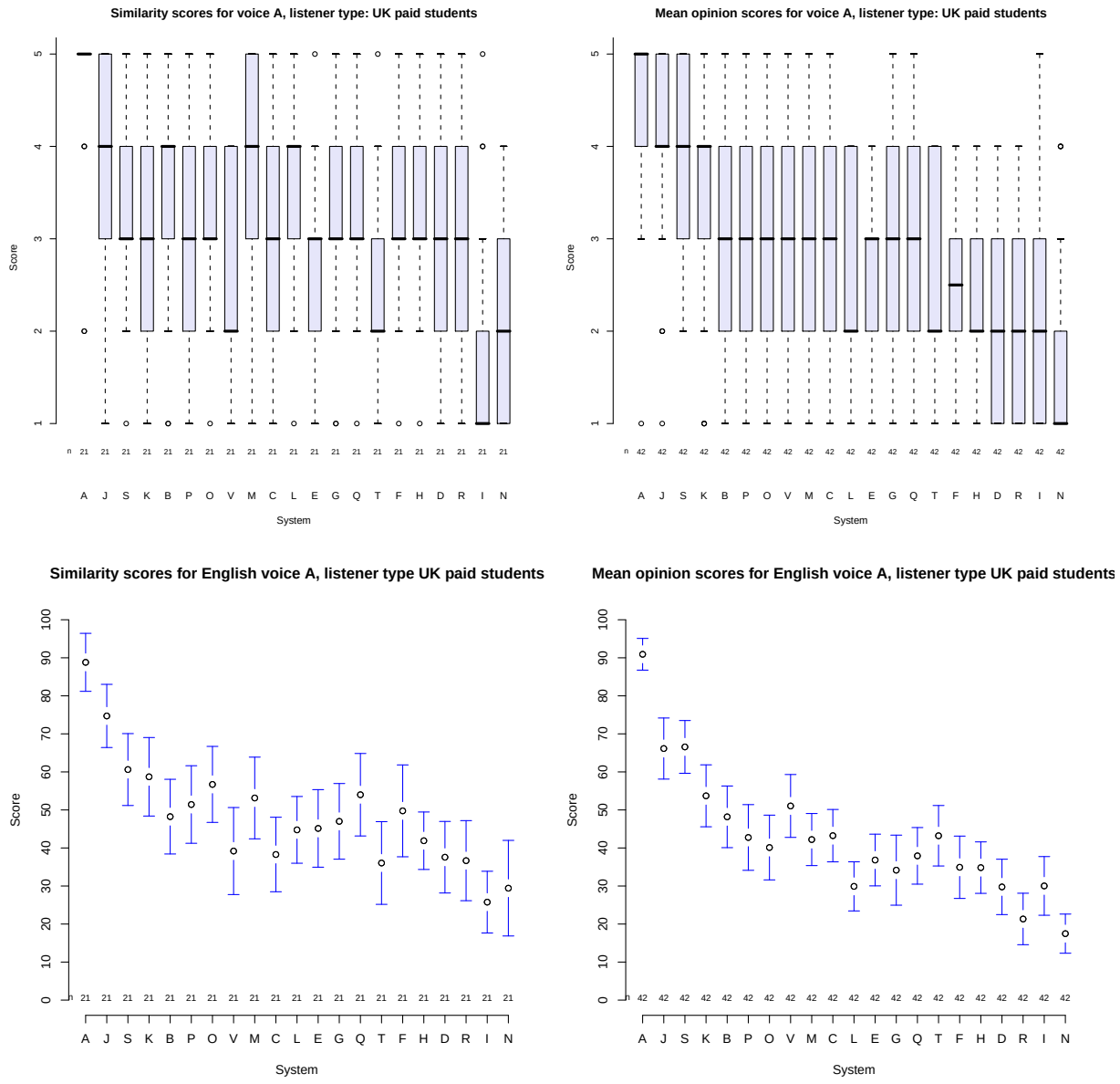
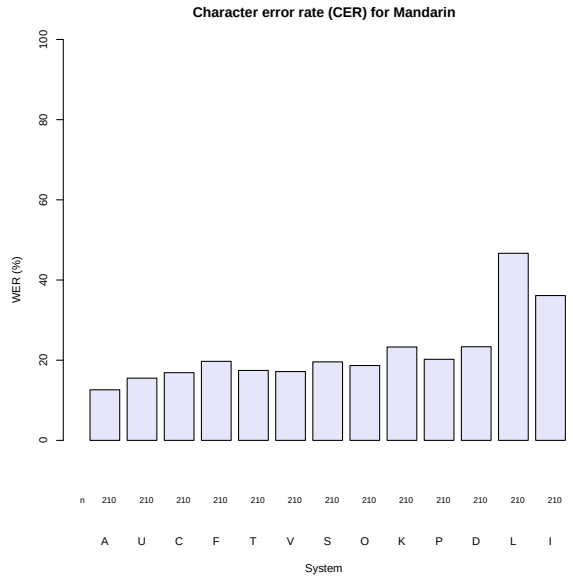
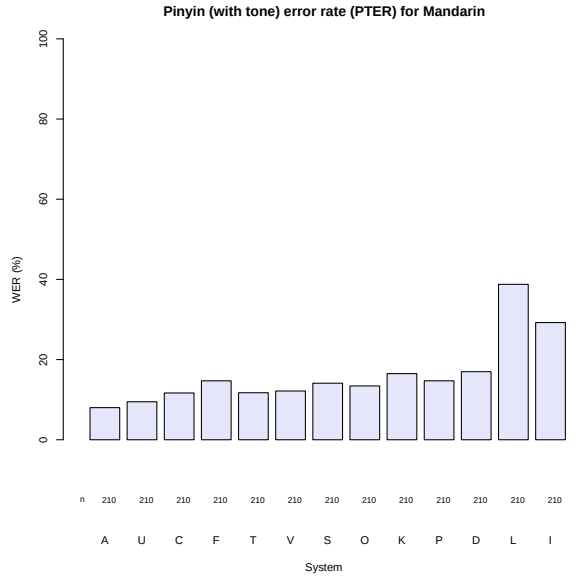


Figure 5: Results for English voice A using paid native-speaker UK undergraduate listeners, comparing the results from listener type EUL who used a 5-point Lickert scale (upper two figures) with the results from listener group EUS who used a continuous rating scale (lower two figures).





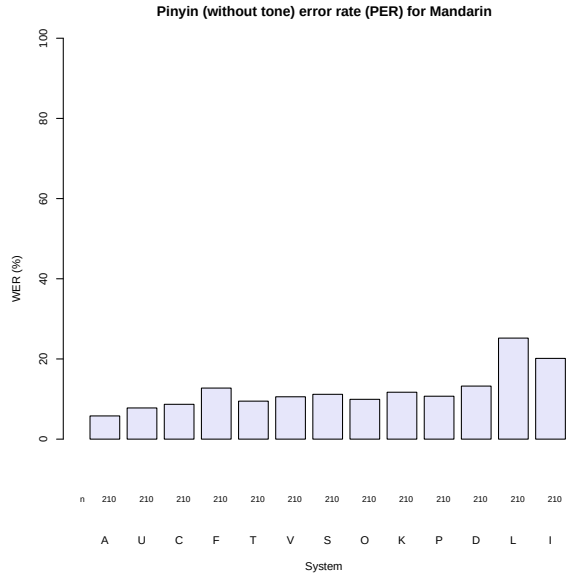


Figure 4: Results for Mandarin: intelligibility

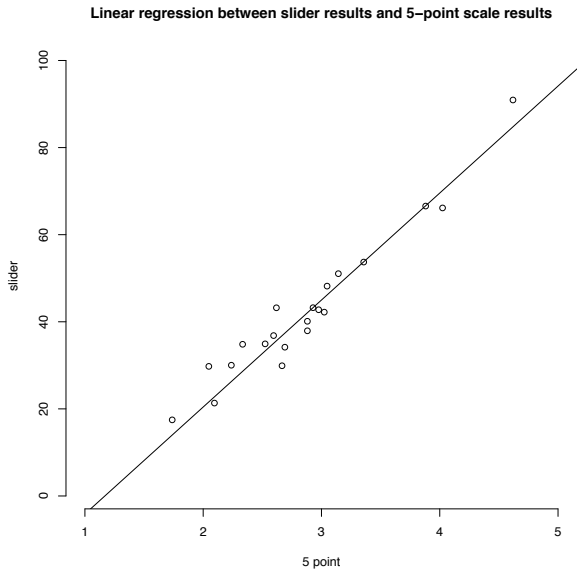


Figure 6: Linear regression for English voice A using paid native-speaker UK undergraduate listeners, predicting the results from listener type EUL who used a continuous scale using listener group EUS who used a 5-point Lickert scale.

4.4) this year (513 registered volunteers in 2008 vs. 198 in 2007), although the completion rates remained similar: 181 (35%) volunteers completed the evaluation this year, compared to 65 (33%) last year). We attribute this large increase in the number of volunteers to the fact that the (English part of the) Blizzard Challenge was mentioned on some popular blogs and technology web sites. Conducting paid evaluations enabled us to achieve 100% completion rates for the 160 listeners recruited in India and Edinburgh.

This was the first year that another language (Mandarin) was used in the Blizzard Challenge. For speech experts (type MS) the registration/completion numbers were 56/44, while for volunteers (type MR) they were 42/17. The Mandarin evaluation was also completed by 148 paid listeners (107 of them in China, 41 in Edinburgh). Clearly, future Challenges must find ways to obtain more volunteer listeners for languages other than English.

As noted in Section 6 we used all listener responses to compute the summary statistics in this year's analysis: responses from both complete and partially completed evaluations were pooled. A detailed breakdown of the number of listeners of each type whose responses were used in the results for each voice is shown in Tables 18 to 22.

8.2. 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 [9]. All responses were optional. Feedback forms were submitted by all the listeners who completed the evaluation (Tables 23 and24) and included many detailed comments and suggestions from all listener types.

Listener information and feedback is summarised in Tables 15 to 47. For English, and similar to Blizzard 2007 [9], there were more than twice as many male listeners as female (Table 15), and the number of native speakers of English and

writing, we only have 3 and and 8 completed evaluations respectively, some of them after the deadline for computing the statistics.

non-natives was almost equal (Table 17). The most frequent first languages (Table 14) of non-natives were Telugu (41)[5], Japanese (29) and Hindi (26).

In the case of Mandarin, the numbers of male and female listeners were almost equal (Table 15), while 185 out of 209 listeners who completed the feedback questionnaire stated that they were native speakers of Mandarin (only 5 listeners stated a different native language).

For both languages, the vast majority of listeners used headphones (Table 32), most were in the same environment for all samples (Table 33), mostly a quiet environment (Table 34), and most did the evaluation in one session (Table 35). This is pleasing, because these external factors are hard to control in an online evaluation; the majority of listeners reported that they conducted the test in an environment similar to that used in the laboratory-based tests. Details of the web browser[6] used can be seen in Table 36: because of the embedded audio files, and the variety of ways different browsers and operating systems allow these to be handled, we could not control the browser behaviour fully.

Listeners were asked if they found the tasks easy or difficult, and in the latter case to give reasons why. They were also asked about the average number of times they listened to samples in each section (Tables 37 to 47). For English and Mandarin respectively, about 80% and 71% of listeners found sections 1 and 2 easy, and about 86% and 83% found sections 3 and 4 easy. 47% and 30% of listeners found section 5 hard. This is reflected by the number of times samples were listened to: while only between 12% and 19% listened to the samples in sections 1-4 more than twice, these numbers increased to 69% (English) and 55% (Mandarin) in section 5. From the rest of the feedback about Sections 1-4, it seems that any difficulties regarding these tasks arise from trouble understanding how to use the scales given (Sections 1, 3 and 4), and what is meant by 'similar' (section 1) and 'natural' (Sections 2-4). This is a typical problem for listeners doing these kinds of tests. Some suggested that actual examples should have been given to illustrate the scale, but we wanted to avoid imposing our own subjective choices with respect to this, in particular because in Section 2 we wanted to identify the features that listeners themselves appeared to focus on in order to define naturalness. The comments about these issues from all listener types showed that they gave serious thought to the task. Some listeners felt confused by the instructions, although we had expended considerable effort on the wording in order to avoid ambiguity. That the task of evaluating speech quality is unfamiliar for many listeners made this more difficult.

At the end of the feedback questionnaire, listeners were asked to state what they liked most and least, one thing they would change in the evaluation, and for any additional comments. There were many positive comments about the evaluation interface, simple layout, clarity of instructions, length and variety of tasks, and many listeners found the evaluation interesting and fun to do. Concerning the samples themselves, listeners were impressed by the variety of systems and techniques and how good/convincing/natural the better samples were, but some complained about the inclusion of poor samples which they found made the task more tedious. Not surprisingly perhaps, there were also comments that the evaluation was too long and repetitive (mainly for the English version), and that a new page had to load for each part of a section.

Section 5 (SUS) was most often singled out as the favourite section by native speakers, who often found the sentences hilarious. For non-natives, it was the most difficult section however, and some suggested that this section should be for native speakers only, due to the obscure vocabulary. This year, natural speech was included in this section, so we are able to use this as a reference for intelligibility.

[5]Telugu is the main language in the state of Andhra Pradesh, India, where listeners were recruited and paid to participate in the evaluation.

[6]For all the evaluations conducted in the Edinburgh University laboratory (listener types EUL, EUS, ME), we used Firefox.

8.3. Suggestions for future Blizzard Challenges

Listeners feedback regarding changes in the evaluation included:

- A female voice
- Larger audio samples in order to assess naturalness.
- More expressive, emotional or emphatic speech for the synthetic samples (e.g. telling jokes or arguing)
- Having a first round of the evaluation in order to exclude the worst systems
- Revealing the text of the sentences from section 5 (SUS) at the end
- More information on the results, related publications, the participating systems and their availability, as well as how these subjective evaluations are useful

Participants were divided on whether to use larger databases or not. With respect to the content of the evaluation, suggestions included:

- Evaluate appropriateness of synthetic speaking style for different genres
- Use a larger scale to find degrees of similarity in section 2 (MDS)
- Measure the listening effort for intelligibility (e.g. count number of times a listener plays the sample)
- Define 'naturalness' more precisely
- Have more labeling information (for Mandarin), or, generally, provide common components like labeling and text analysis for all systems to use
- Allow collaborative efforts, and different systems for different languages
- Use more domains (navigation, dialogue)

9. Acknowledgements

Dong Wang wrote the WER and CER/PTER/PER programmes; Evia Kainada assisted in running the listening tests in Edinburgh; Kishore Prahallad ran the listening tests in India; Volker Strom and Junichi Yamagishi provided the benchmark systems. Roger Burroughes is 'roger', the English voice; Richard Sproat generated the SUS sentences; Jianhua Tao of the Chinese Academy of Sciences provided the Mandarin data. Finally, thanks to all participants and listeners.

10. 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 2008 website," http://www.synsig.org/index.php/Blizzard_Challenge_2008.

[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] 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.

[6] Heiga Zen and Tomoki Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *Proc. Blizzard Workshop*, 2007.

[7] "Blizzard Challenge 2008 rules," http://www.synsig.org/index.php/Blizzard_Challenge_2008_Rules.

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

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

[10] 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.

[11] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Workshop (in Proc. SSW6)*, August 2007.

[12] Ellen Gurman Bard, Dan Robertson, and Antonella Sorace, "Magnitude estimation of linguistic acceptability," *Language*, vol. 72, no. 1, pp. 32–68, 1996.

Language	English total	Mandarin total
Afrikaans	3	0
Bulgarian	1	0
Cantonese	2	1
Catalan	1	0
Chinese	8	0
Danish	1	0
Dutch	6	0
Finnish	2	0
Flemish	1	0
French	4	0
German	27	0
Hebrew	7	0
Hindi	26	0
Hungarian	2	0
Indonesian	1	0
Italian	2	0
Japanese	29	1
Kannada	3	0
Korean	0	1
Mandarin	2	0
Marathi	2	0
Norwegian	1	0
Polish	1	0
Portuguese	8	0
Punjabi	3	0
Russian	3	0
Slovenian	1	0
Spanish	9	0
Swedish	3	0
Telugu	41	0
Thai	1	0
Turkish	3	0
Urdu	1	0
N/A	7	2

Table 14: First language of non-native speakers for English and Mandarin versions of Blizzard

Gender	Male	Female
English total	302	129
Mandarin total	101	96

Table 15: Gender

Age	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
English total	51	359	148	59	33	13	1	0
Mandarin total	36	168	33	5	0	0	0	0

Table 16: Age of listeners whose results were used (completed the evaluation fully or partially

Native speaker	Yes	No
English	223	212
Mandarin	185	5

Table 17: Native speakers for English and Mandarin versions of Blizzard

	English A	English B	Mandarin
EI	42	38	0
EO	4	0	0
ER	200	160	0
ES	64	56	0
EUL	21	19	0
EUS	21	19	0
EVI	20	0	0
MC	0	0	124
ME	0	0	41
MR	0	0	28
MS	0	0	49
ALL	372	292	242

Table 18: Listener types per voice, showing the number of listeners whose responses were used in the results

	Registered	No response at all	Partial evaluation	Completed Evaluation
EI	80	0	0	80
EO	4	0	0	4
ER	479	121	184	174
ES	143	23	23	97
EUL	40	0	0	40
EUS	40	0	0	40
EVI	30	11	16	3
ALL ENGLISH	**816**	**155**	**223**	**438**
MC	143	19	17	107
ME	41	0	0	41
MR	42	15	10	17
MS	57	7	6	44
ALL MANDARIN	**283**	**41**	**33**	**209**

Table 19: Listener registration and evaluation completion rates

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21
EI	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
EO	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ER	7	13	12	10	9	6	10	9	10	6	9	9	7	16	4	6	17	9	6	11	13
ES	4	5	2	3	3	3	2	1	3	3	3	4	3	4	3	3	2	3	4	3	3
EUL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EUS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EVI	9	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALL	25	24	28	18	17	13	16	14	17	13	16	17	14	24	11	13	23	16	14	18	20

Table 20: Listener groups - Voice A (English), showing the number of listeners whose responses were used in the results - i.e. those with partial or completed evaluations

	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19
EI	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
EO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ER	7	9	10	5	6	12	7	4	12	17	9	3	6	11	5	14	3	9	11
ES	2	4	5	3	2	4	5	4	3	5	2	2	2	2	2	2	1	2	3
EUL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EUS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EVI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ALL	13	17	19	12	12	20	16	12	19	26	15	9	12	17	11	20	8	15	18

Table 21: Listener groups - Voice B (English), showing the number of listeners whose responses were used in the results

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13
MC	8	8	7	10	8	9	9	7	19	10	13	7	9
ME	4	3	3	3	3	3	3	3	3	3	4	3	3
MR	3	4	3	2	2	4	1	1	2	1	2	1	2
MS	4	7	3	4	6	4	3	2	4	4	2	3	3
ALL	19	22	16	19	19	20	16	13	28	18	21	14	17

Table 22: Listener groups - Mandarin, showing the number of listeners whose responses were used in the results

Listener Type	EI	EO	ER	ES	EUL	EUS	EVI	ALL ENGLISH
Total	80	4	174	97	40	40	3	438

Table 23: Listener type totals for submitted feedback (English)

Listener Type	MC	ME	MR	MS	ALL MANDARIN
Total	107	41	17	44	209

Table 24: Listener type totals for submitted feedback (Mandarin)

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate
English total	42	71	143	125	62
Mandarin total	3	12	84	76	31

Table 25: Highest level of education completed

CS/Engineering person?	Yes	No
English total	303	139
Mandarin total	102	102

Table 26: Computer science / engineering person

Work in speech technology?	Yes	No
English total	166	278
Mandarin total	60	143

Table 27: Work in the field of speech technology

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
English total	67	50	49	113	93	36	38
Mandarin total	24	22	5	23	35	50	39

Table 28: How often normally listened to speech synthesis before doing the evaluation

Dialect of English	Australian	Indian	UK	US	Other	N/A
Total	2	6	112	86	8	9

Table 29: Dialect of English of native speakers

Dialect of Mandarin	Beijing	Shanghai	Guangdong	Sichuan	Northeast	Other	N/A
Total	45	7	8	4	26	56	43

Table 30: Dialect of Mandarin of native speakers

Level of English	Elementary	Intermediate	Advanced	Bilingual	N/A
English total	24	84	85	26	1
Madarin total	1	0	1	1	2

Table 31: Level of English of non-native speakers

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
English total	374	43	19	2
Mandarin total	151	24	17	1

Table 32: Speaker type used to listen to the speech samples

Same environment?	Yes	No
English total	437	5
Mandarin total	183	14

Table 33: 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	264	153	22	1	0
Mandarin total	112	69	14	3	0

Table 34: Kind of environment when listening to the speech samples

Number of sessions	1	2-3	4 or more
English total	296	117	28
Mandarin total	134	58	6

Table 35: Number of separate listening sessions to complete all the sections

Browser	Firefox	IE	Mozilla	Netscape	Opera	Safari	Other
English total	252	118	36	1	7	23	4
Mandarin total	41	138	6	0	0	0	15

Table 36: Web browser used

Section 1	Easy	Difficult
English total	307	126
Mandarin total	148	36

Table 37: Listeners' impression of their task in Section 1

Problem	Scale too big, too small, or confusing	Bad speakers, playing files files disturbed others, connection too slow, etc	Other
English total	89	1	47
Mandrin total	26	3	6

Table 38: Listeners' problems in Section 1

Number of times	1-2	3-5	6 or more
English total	362	67	4
Mandarin total	152	30	3

Table 39: Number of times listened to each example in Section 1

Section 2	Easy	Difficult
English total	305	128
Mandarin total	142	36

Table 40: Listeners' impression of their task in Section 2

Problem	Unfamiliar task	Instructions not clear	Bad speakers, playing files disturbed others connection too slow, etc	Other
English total	57	28	0	53
Mandarin total	12	20	0	6

Table 41: Listeners' problems in Section 2

Number of times	1-2	3-5	6 or more
English total	381	50	1
Mandarin total	151	28	1

Table 42: How many times listened to each example in section 2

Section 3 and 4	Easy	Difficult
English total	351	75
Mandarin total	149	23

Table 43: Listeners' impression of their task in Sections 3 and 4

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	13	29	0	25
Mandarin total	5	10	1	2

Table 44: Listeners' problems in Sections 3 and 4

Number of times	1-2	3-5	6 or more
English total	373	57	0
Mandarin total	152	28	0

Table 45: How many times listened to each example in sections 3 and 4?

Section 5 (SUS)	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	34	197	170	31
Mandarin total	19	112	52	0

Table 46: Listeners' impressions of the task in Section 5

Number of times	1-2	3-5	6 or more
English total	134	237	63
Mandarin total	82	91	9

Table 47: How many times listened to each example in section 5