Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020
30 October 2020, Shanghai, China

The Blizzard Challenge 2020

Xiao Zhou[1], Zhen-Hua Ling[1], Simon King[2]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R. China
[2]Center for Speech Technology Research, University of Edinburgh, UK
xiaozh@mail.ustc.edu.cn, zhling@ustc.edu.cn, Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2020 is the sixteenth annual Blizzard Challenge. The challenge this year includes a hub task of synthesizing Mandarin speech and a spoke task of synthesizing Shanghainese speech. The speech data of these two Chinese dialects as well as corresponding text transcriptions were provided. Sixteen and eight teams participated in the two tasks respectively. Listening tests were conducted online to evaluate the performance of synthetic speech.
Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

Black and Tokuda conceived the Blizzard Challenge in 2005 [1] and there have been annual summary papers like this one every year, plus a one-off retrospective summary-of-summaries covering the first decade [2]. For many previous challenges, the submitted speech, reference natural samples, raw listening test responses, scripts for running the listening test and scripts for the statistical analysis, can be obtained from the Blizzard Challenge website [3].

The Blizzard Challenge 2020 is the first one organised by the University of Science and Technology of China (USTC), with assistance from the University of Edinburgh and the other members of the Blizzard Challenge committee. Two tasks of synthesizing Mandarin speech and Shanghainese speech were designed, which received 16 and 8 submissions respectively. This paper will present the details of speech datasets, tasks, participating systems, listening tests and results of the challenge.

2. Voices to build

2.1. Speech datasets

Speech waveforms and text transcriptions of two Chinese dialects, Mandarin and Shanghainese, were released for voice building. These two datasets were provided by iFLYTEK Co., Ltd. The speech data for Mandarin was 9.5 hours (sampled at 48 KHz) from a professional male native Mandarin speaker. The speech data for Shanghainese was 3 hours (sampled at 16 KHz) from a professional female native Shanghainese speaker. The texts of both datasets were from daily news. Both datasets were recorded in studios and quiet environments. The Mandarin data was provided with text transcriptions only. The Shanghainese data was provided with both text and phonetic transcriptions.

2.2. Tasks

There were two tasks in the Blizzard Challenge 2020 which used the two datasets respectively.

- Hub task 2020-MH1: Each participant should build one voice in Mandarin using the provided data in agreement with challenge rules[1], and synthesize a test set of 700 sentences. The synthetic speech should be single channel, 16bit depth, and can be at any standard sampling rate (16kHz, 22.05kHz, 44.1kHz or 48kHz).
- Spoke task 2020-SS1: Each participant should build one voice in Shanghainese using the provided data in agreement with challenge rules, and synthesize a test set of 391 sentences. The synthetic speech should be single channel, 16bit depth, and can be at any standard sampling rate (16kHz, 22.05kHz, 44.1kHz or 48kHz).

For the 2020-MH1 task (hub task), the 700 test sentences were composed as follows.

- news: 500 distinct sentences (68 paragraphs) from daily news, without overlapping with the training data.
- PSC: 100 distinct sentences (11 paragraphs) from the materials of the Putonghua Proficiency Test in China, without overlapping with the training data.
- INT: 100 distinct sentences for intelligibility evaluation, without overlapping with the training data. These sentences were meaningless and composed by randomly choosing words according to their POS tags.

For the 2020-SS1 task (spoke task), the 391 test sentences were composed as follows.

- news: 291 distinct sentences from daily news, without overlapping with the training data.
- chat: 100 distinct sentences from daily chat, without overlapping with the training data.

3. Participants

In the Blizzard Challenge 2020, 16 participants submitted their results for the hub task and 8 participants for the spoke task. The details can be found in Table 1. No benchmark systems were prepared this year. Following previous challenges, all systems are identified using anonymous letters when announcing results. Here, letter A denotes natural speech. The other letters are assigned randomly and denote the systems submitted by participants in the challenge. The participating teams can choose whether or not revealing their system identifiers in their workshop papers.

Unsurprisingly, the neural-network-based statistical parametric speech synthesis (SPSS) approach completely dominated this year's Blizzard Challenge as shown in Table 1. For the first time in Blizzard Challenge history, no HMM-based or unit selection systems were submitted. More than half of the submitted systems utilized sequence-to-sequence neural networks, such as Tacotron, for acoustic modeling. Neural vocoders were adopted by all systems, with WaveRNN, WaveNet and LPCNet being popular choices.

[1]https://www.synsig.org/index.php/Blizzard_Challenge_2020_Rules

10.21437/VCC_BC.2020-1

Table 1: *The participating teams and their short names. The system identifier of natural speech (the first row) is letter A. The remaining rows are in alphabetical order of the system short name and not in alphabetical order of system identifier. The method descriptions are summarized based on the questionnaires returned from participants.*

Short name	**Details**	**Method for task MH1**	**Method for task SS1**
NATURAL	Natural speech from the same speaker as the corpus	Human	Human
AI.SG	Nanyang Technological University	End-to-End + WaveNet	End-to-End + WaveNet
ALONG	Sun Yat-sen University, NetEase Games AI Lab	Tacotron + WaveNet	-
ajmide	Ajmide Media Co., Ltd.	DNN + WaveRNN	-
Duke	Duke Kunshan University	Tacotron2 + WaveRNN	-
hmlyTTS	Ximalaya FM	Seq2Seq + WaveRNN	Seq2Seq + WaveRNN
laiye	Laiye Technology	DNN + LPCNet	-
NLPR	National Laboratory of Pattern Recognition	Tacotron + LPCNet	Tacotron + LPCNet
NUS-HLT	National University of Singapore, the Human Language Technology (HLT) Laboratory	Tacotron + WaveRNN	Tacotron + WaveNet
RoyalFlush	Zhejiang Hithink RoyalFlush AI Research Institute	Tacotron + LPCNet	Tacotron + LPCNet
SCUT	South China University of Technology, Guangzhou Higher Education Mega Center	Tacotron + WaveRNN	-
SHNU	Shanghai Normal University	End-to-End + WaveNet	End-to-End + WaveNet
Sogou	Sogou Inc.	VAE-FastSpeech + WaveRNN	-
SunAtEight	Harbin Institution of Technology	DNN + Parallel WaveGAN	-
TSS	Tencent Holdings Ltd.	DNN + WaveRNN	DNN + WaveRNN
Whatever	The Chinese University of Hong Kong	Tacotron2 + WaveGlow	-
OPPO-TTS	Guangdong OPPO Mobile Telecommunications Corp., Ltd	DNN + WaveRNN	DNN + MelGAN

4. Listening tests

4.1. Listening test materials

Several hundreds of test sentences were synthesized by participants, which prevented the manual intervention made by participants during synthesis. Only a small subset of synthesized utterances were used in the listening test. This means that there are a large amount of material that might be used in future listening tests. Please refer to the summary papers of previous challenges [4] for a description of the listening test design and the web interface used to deliver it. After obtaining the permissions from participants, the detailed listening test results will be distributed via the Blizzard Challenge website [5].

4.2. Listener types

Similarly to previous years, various listener types were used in the test.

- Paid university students. For the 2020-MH1 task, the listeners were native speakers of Chinese (any accent), i.e., the MP type. For the 2020-SS1 tasks, the listeners were native speakers of Shanghainese recruited at Shanghai International Studies University, i.e., the SP type. The listeners of both MP and SP types were generally aged 18-25. Due to the impact of COVID-19, all paid listener completed the test online this year.
- Speech experts (self-declared), recruited via participating teams and mailing lists for the 2020-MH1 task, i.e., the ME type.
- Volunteers recruited via participating teams, mailing lists, WeChat groups, etc. for the 2020-MH1 task, i.e., the MR type.

Following previous challenges, organizers asked participating teams to help recruit volunteer listeners (in categories ME or MR). According to the numbers reported to the organizers, almost all teams recruited at least 10 listeners.

4.3. Listening test design

As mentioned before, only a subset of the complete test set was used in the formal listening tests. The listening tests for 2020-MH1 consisted of six sections each with 17 samples, while the listening tests for 2020-SS1 consisted of seven sections each with 9 samples. These sections are listed as follows.

- Listening tests for 2020-MH1
  1. Similarity, news sentences
  2. Similarity, PSC sentences
  3. Naturalness, news sentences
  4. Naturalness, PSC sentences
  5. Multiple dimensions, news paragraphs
  6. Intelligibility, INT sentences
- Listening tests for 2020-SS1
  1. Similarity, chat sentences
  2. Similarity, news sentences
  3. Naturalness, chat sentences
  4. Naturalness, news sentences
  5. Naturalness, news sentences
  6. Intelligibility, chat sentences
  7. Intelligibility, news sentences

Table 2: *Listener registration and evaluation completion rates for task 2020-MH1.*[1]

	Registered	No response at all	Partial evaluation	Completed evaluation
MP	258	4	4	250
ME	145	18	22	105
MR	111	13	17	81
ALL	**514**	**35**	**43**	**436**

Table 3: *The number of listeners whose responses were used in the final listening test results.*

	Task MH1	Task SS1
MP/SP	208	87
ME	93	-
MR	69	-
ALL	**370**	**87**

In each of the above sections, one example from each system, including natural speech, was played to a listener. Following previous challenges, the orders of examples were determined by a Latin Square design. Besides, no listener heard the same sentence or paragraph more than once throughout the whole test, which was especially crucial for the intelligibility sections.

The methodology of scoring in the various sections of 2020-MH1 were the same as previous Blizzard Challenge [6], except that INT sentences replaced the SUS sentences for dictation and it can be played at most twice instead of once. Pinyin Error Rate with Tones (PTER) was used as the metric for intelligibility.

Considering the complexity of input transcriptions for Shanghainese, the dictation test was not conducted for the intelligibility test of task 2020-SS1. Instead, listeners were asked to choose a response that represented how intelligible the synthetic voice was on a scale from 1 (completely not intelligible) to 5 (completely intelligible). Each synthetic sentence for intelligibility test can be played at most twice. The disadvantage of such evaluation is that the listeners' judgement may be significantly affected by other factors, such as the naturalness of synthetic speech.

4.4. Listening test completion rate

Table 2 shows the statistics of evaluation completion rates for different listener types. We can see that the overall completion rate this year (84.8%) was much better than that in 2019 (63.7%) [6]. We should appreciate all participating teams for recruiting expert and volunteer listeners.

To get the final listening test results, we further excluded some listeners from "completed all sections" listeners mainly based on two anti-cheating considerations. First, the natural speech should not be given a very low naturalness or similarity score. Second, it was abnormal if all but one systems were given very low scores by a listener. As shown in Table 3, 370 and 87 valid listeners were used by the two tasks respectively.

5. Analysis methodology

In this paper, we only show the results combining all listener types. The detailed results by listener types have been distributed to participants. After obtaining the permissions from participants, the complete listening test results, including raw listener scores for each stimulus, will be distributed via the Blizzard website [5]. Thus, anyone can re-analyze the listening test data if they are interested. We followed the statistical analysis techniques described in [7] to produce the listening test results. In this paper and the listening test results distributed by the organizers, all system names are in an anonymous form. The participating teams can decide whether or not revealing their system identifiers in their workshop papers. Besides, a summary of listener questionnaire responses for task 2020-MH1 are shown in Tables 4 to 27.[2]

6. Results

The listening test results are shown in Figures 1 to 22. The standard boxplots are employed for representing the ordinal data, e.g., mean opinion scores (MOS). More information on how to interpret the boxplots can be found in [6]. In all figures, a consistent system ordering is adopted, which is the descending order of mean naturalness. The mean naturalness is calculated from the listeners' scores on the two sentence-based naturalness sections for each task. Please note that this ordering only aims to make the plots more readable by using the same system ordering across all plots for each task and *can not be interpreted as a ranking*, because the ordering does not indicate which systems are significantly better than others.

In this year's task 2020-MH1, when combining the opinions of all listeners, no system was as natural as natural speech (Figures 1 and 2), or as similar to the target speaker (Figures 1 and 3). I and O were significantly more natural than all other systems. I was also significantly more similar to the target speaker than all other systems except O. In the intelligibility test, there was no significant difference among D, I, L, P and natural speech (Figures 1 and 4). In the evaluation results on news paragraphs, the speech synthesized by all systems still had significant difference with natural speech on all dimensions.

In this year's task 2020-SS1, no system was as natural as natural speech (Figures 19 and 20), or as similar to the target speaker (Figures 19 and 21). I was significantly more natural than all other systems. Compared with most other systems, E was significantly more similar to the target speaker except L. In the intelligibility test, there was no significant difference between I and E, and I was the only system with a median score of 5 like natural speech (Figures 19 and 22). It should be noticed that we found a significant correlation between the naturalness and intelligibility scores of all systems. The Pearson correlation coefficient was 0.7675 ($p = 0.0262$) when excluding natural speech. For comparison, the Pearson correlation coefficient between the naturalness scores and PTERs of all systems in the 2020-MH1 task was -0.4846 ($p = 0.0571$) when excluding natural speech. These results imply the disadvantage of replacing dictation with an intelligibility MOS test for 2020-SS1 as discussed in Section 4.3.

7. Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61871358). We wish to thank a number of additional contributors without whom running the challenge would not be possible. Yuan Jiang at iFlytek Co., Ltd. helped to prepare the materials of training

[1] We calculated these numbers from the database that stored the listening test results.

[2] These numbers were calculated from the feedback forms that listeners completed at the end of the test. As this was optional, many listeners decided not to fill it in. If they did, they did not always reply to all the questions in the form. Therefore, the total number of items was usually smaller than 370.

data. Ailing Zhang, Chenxuan Liu, and Yifei Sun at the Graduate Institute of Interpretation and Translation, Shanghai International Studies University organized the paid listeners of 2020-SS1 task. Vasilis Karaiskos at the University of Edinburgh provided advices of building evaluation webpages. Bo Xu at the University of Science and Technology of China helped to config the server for the webpage deployment. The listening test scripts were based on earlier versions provided by previous organizers of the Blizzard Challenge. Thanks to all participants and listeners.

8. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[2] Simon King, "Measuring a decade of progress in Text-to-Speech," *Loquens*, vol. 1, no. 1, p. 006, 2014.

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

[4] M. Fraser and S. King, "The blizzard challenge 2007," *Proc. SSW6*, 2007.

[5] "Submissions and listening test results from previous Blizzard Challenges," http://www.cstr.ed.ac.uk/projects/blizzard/data.html.

[6] Z. Wu, S. Le Maguer, J. Cabral, and S. King, "The Blizzard Challenge 2019," *Proc. Blizzard Challenge Workshop*, 2019.

[7] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," *Proc. SSW6*, 2007.

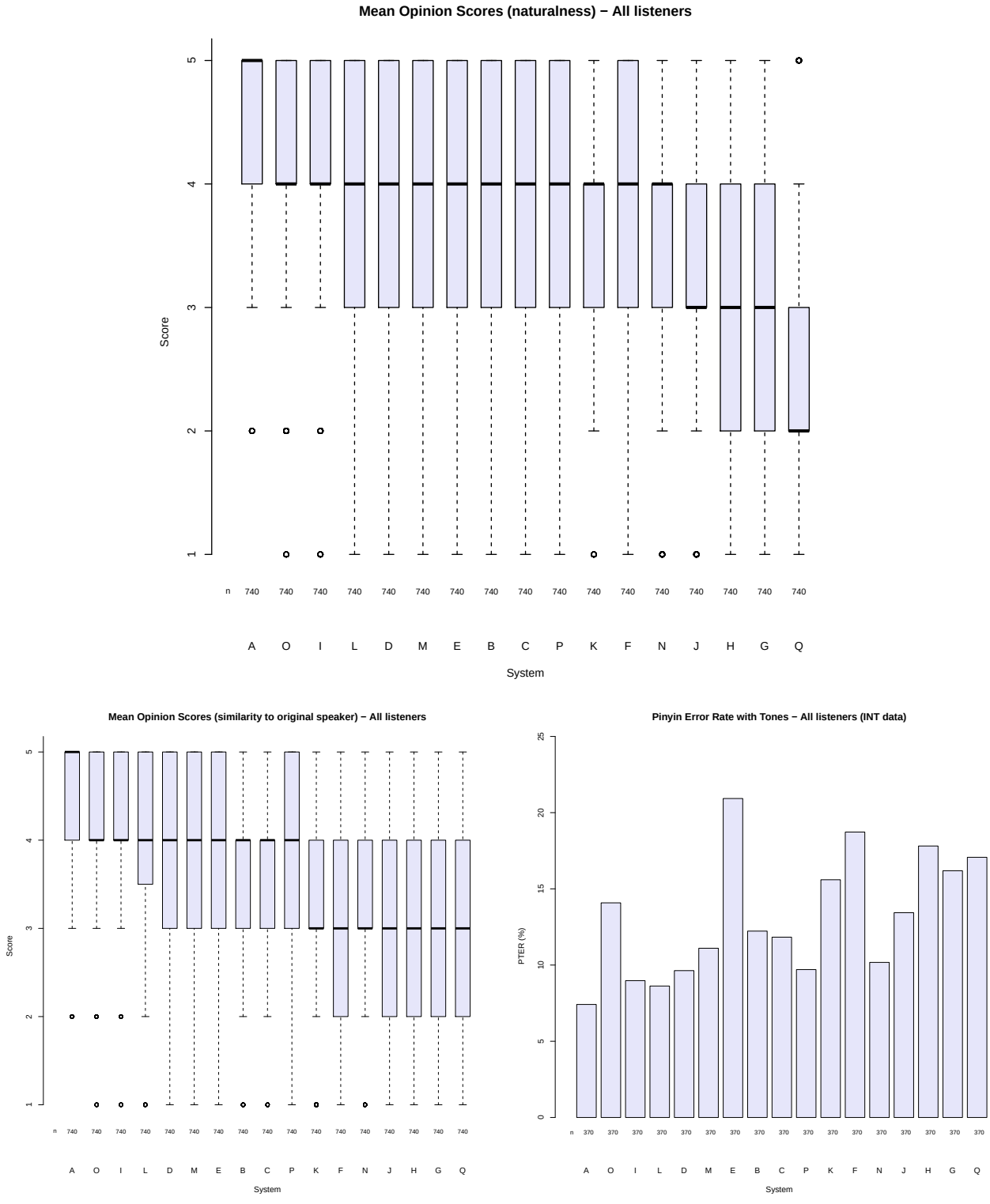

Figure 1: *Results for task 2020-MH1 on sentence test material, combining all listener types. A is natural speech, the remaining letters denote the systems submitted by participants.*

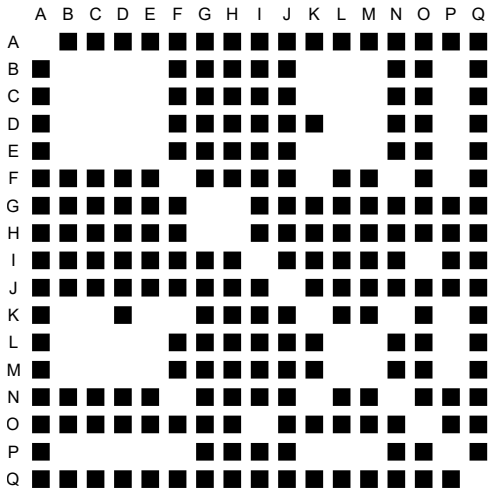

Figure 2: *Significant differences in naturalness between systems are indicated by solid black boxes for task 2020-MH1.*

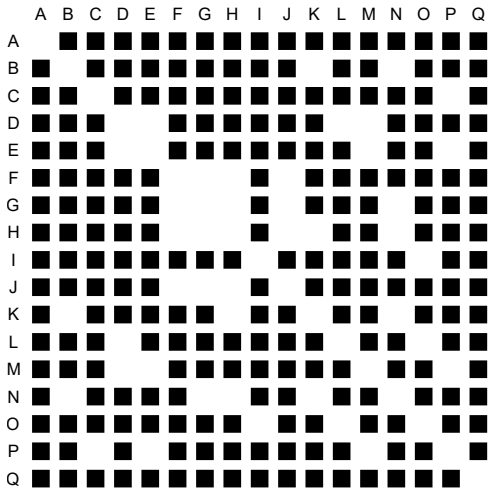

Figure 3: *Significant differences in speaker similarity between systems are indicated by solid black boxes for task 2020-MH1.*

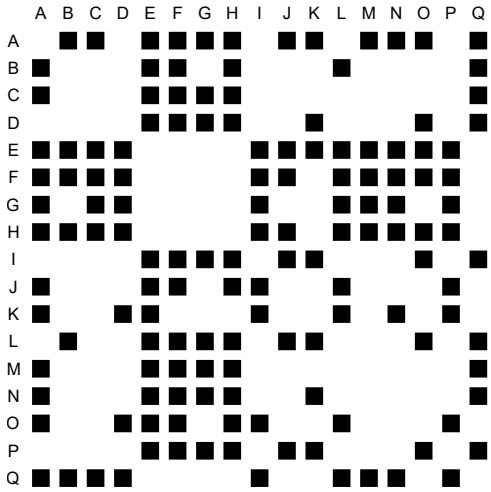

Figure 4: *Significant differences in intelligibility (INT) between systems are indicated by solid black boxes for task 2020-MH1.*

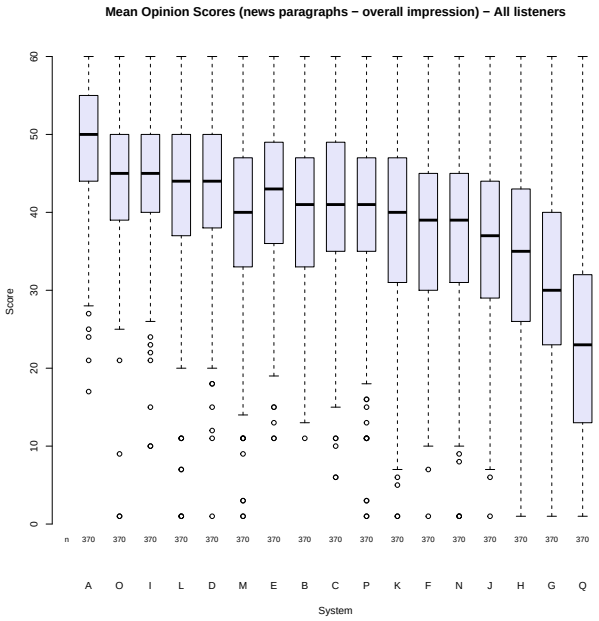

Figure 5: *Overall impression of paragraphs for task 2020-MH1.*

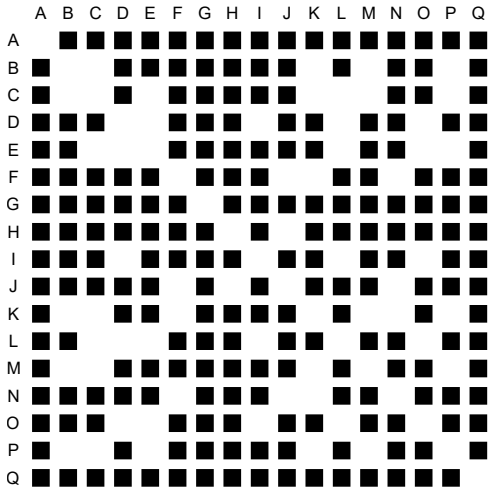

Figure 6: *Significant differences in overall impression of paragraphs by solid black boxes for task 2020-MH1.*

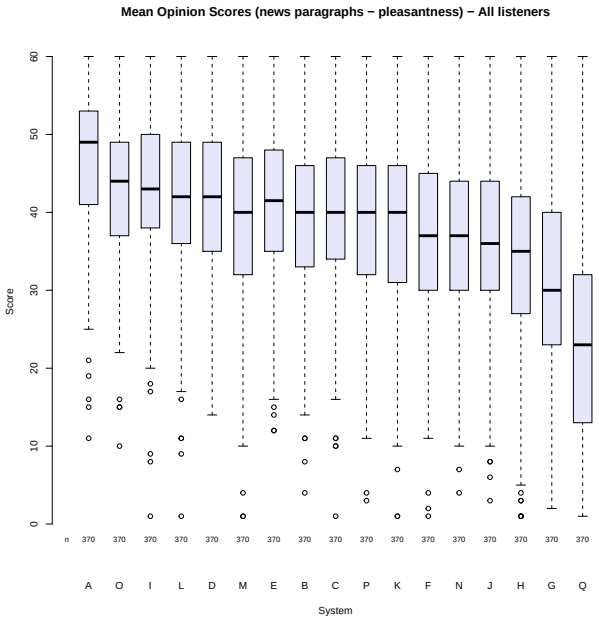

Figure 7: *Pleasantness of paragraphs for task 2020-MH1.*

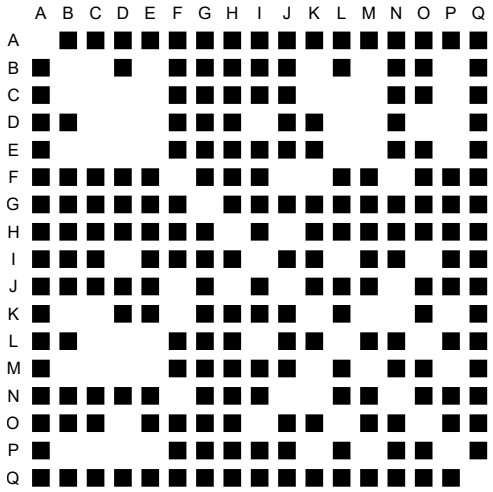

Figure 8: *Significant differences in pleasantness of paragraphs by solid black boxes for task 2020-MH1.*

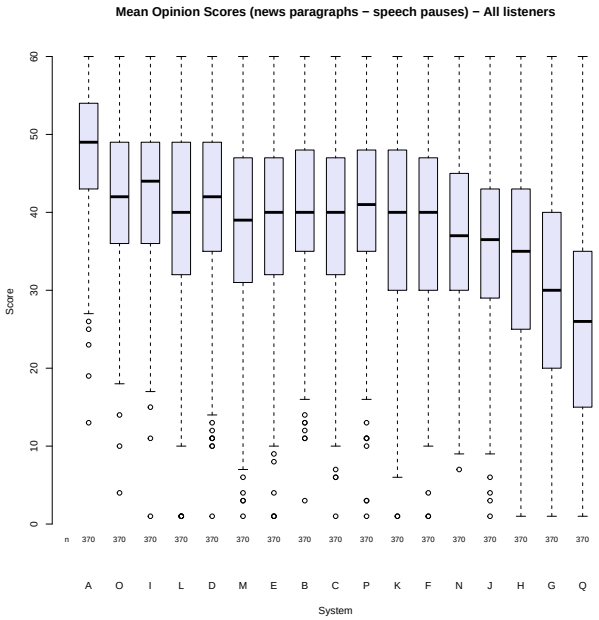

Figure 9: *Speech pauses of paragraphs for task 2020-MH1.*

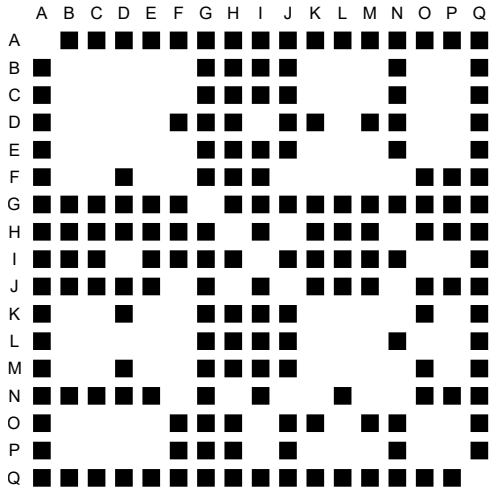

Figure 10: *Significant differences in speech pauses of paragraphs by solid black boxes for task 2020-MH1.*

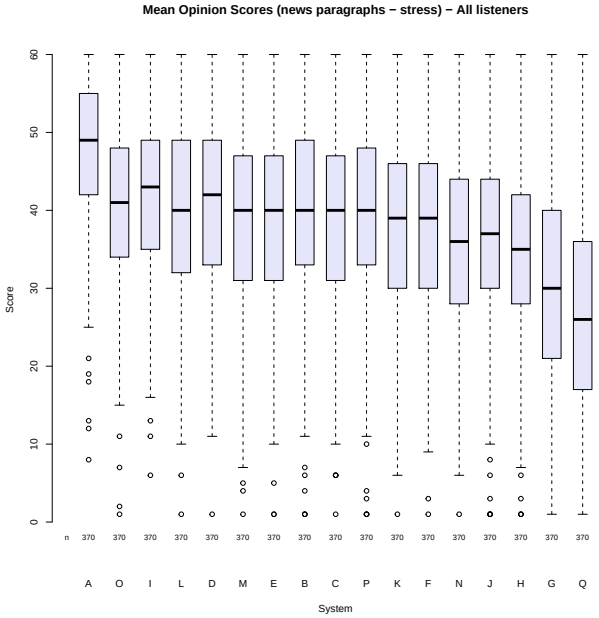

Figure 11: *Stress of paragraphs for task 2020-MH1.*

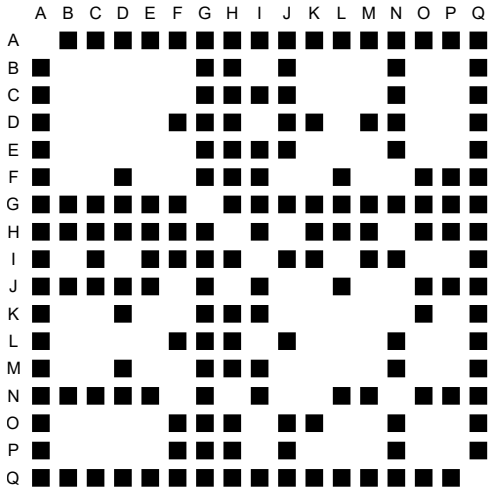

Figure 12: *Significant differences in stress of paragraphs by solid black boxes for task 2020-MH1.*

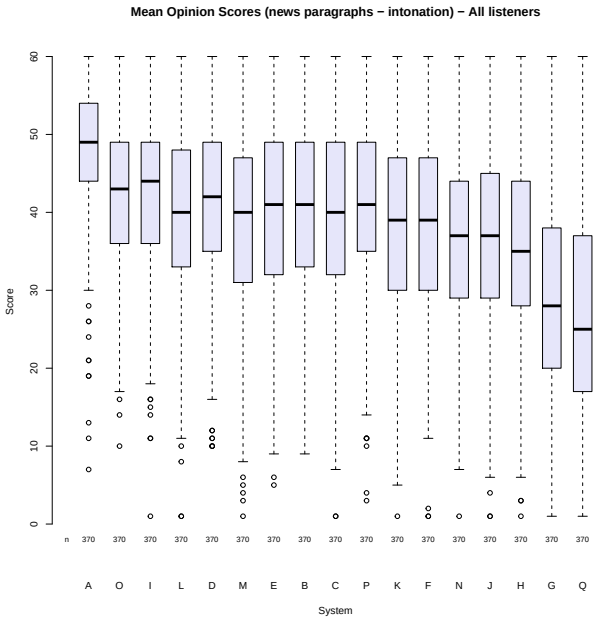

Figure 13: *Intonation of paragraphs for task 2020-MH1.*

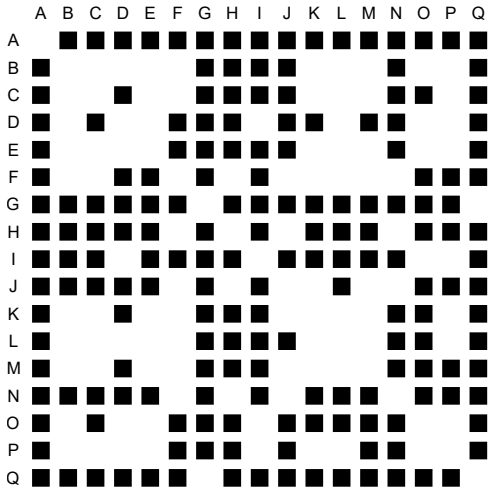

Figure 14: *Significant differences in intonation of paragraphs by solid black boxes for task 2020-MH1.*

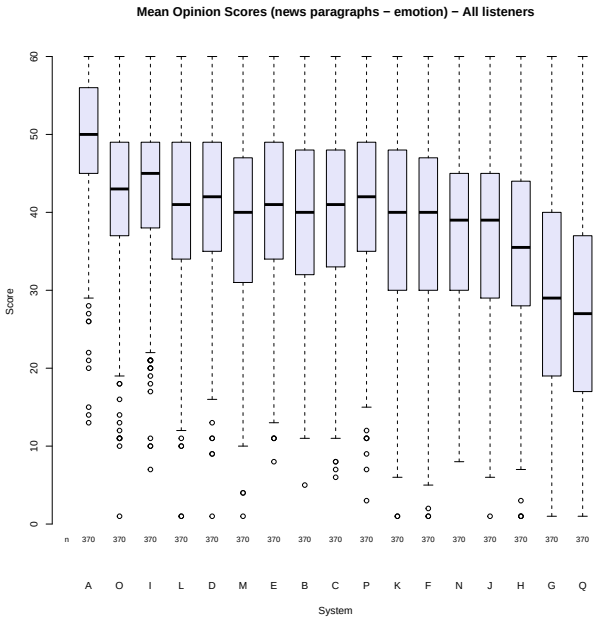

Figure 15: *Emotion of paragraphs for task 2020-MH1.*

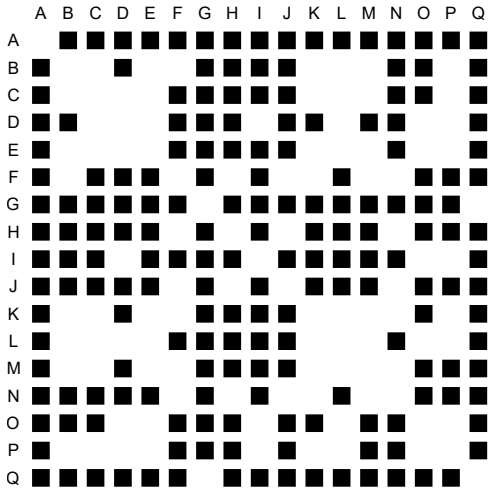

Figure 16: *Significant differences in emotion of paragraphs by solid black boxes for task 2020-MH1.*

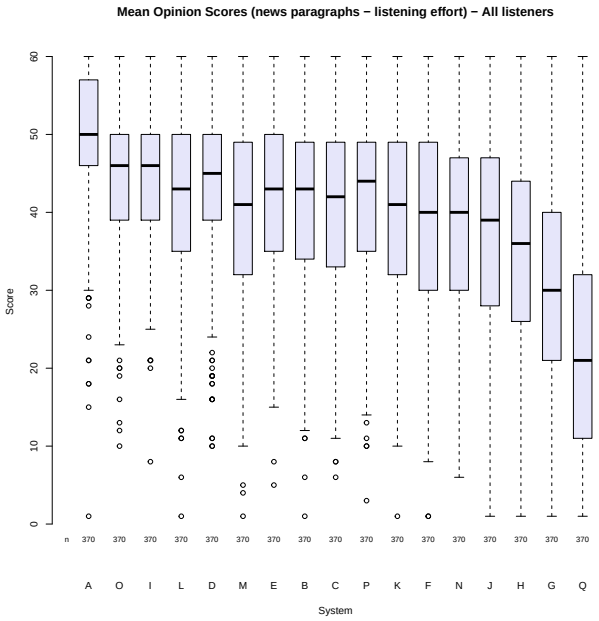

Figure 17: *Listening effort of paragraphs for task 2020-MH1.*

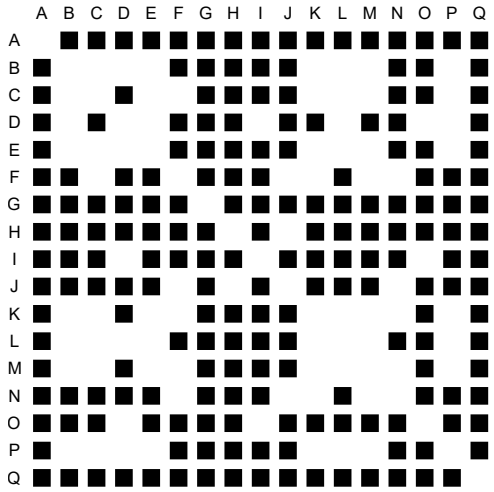

Figure 18: *Significant differences in listening effort of paragraphs by solid black boxes for task 2020-MH1.*

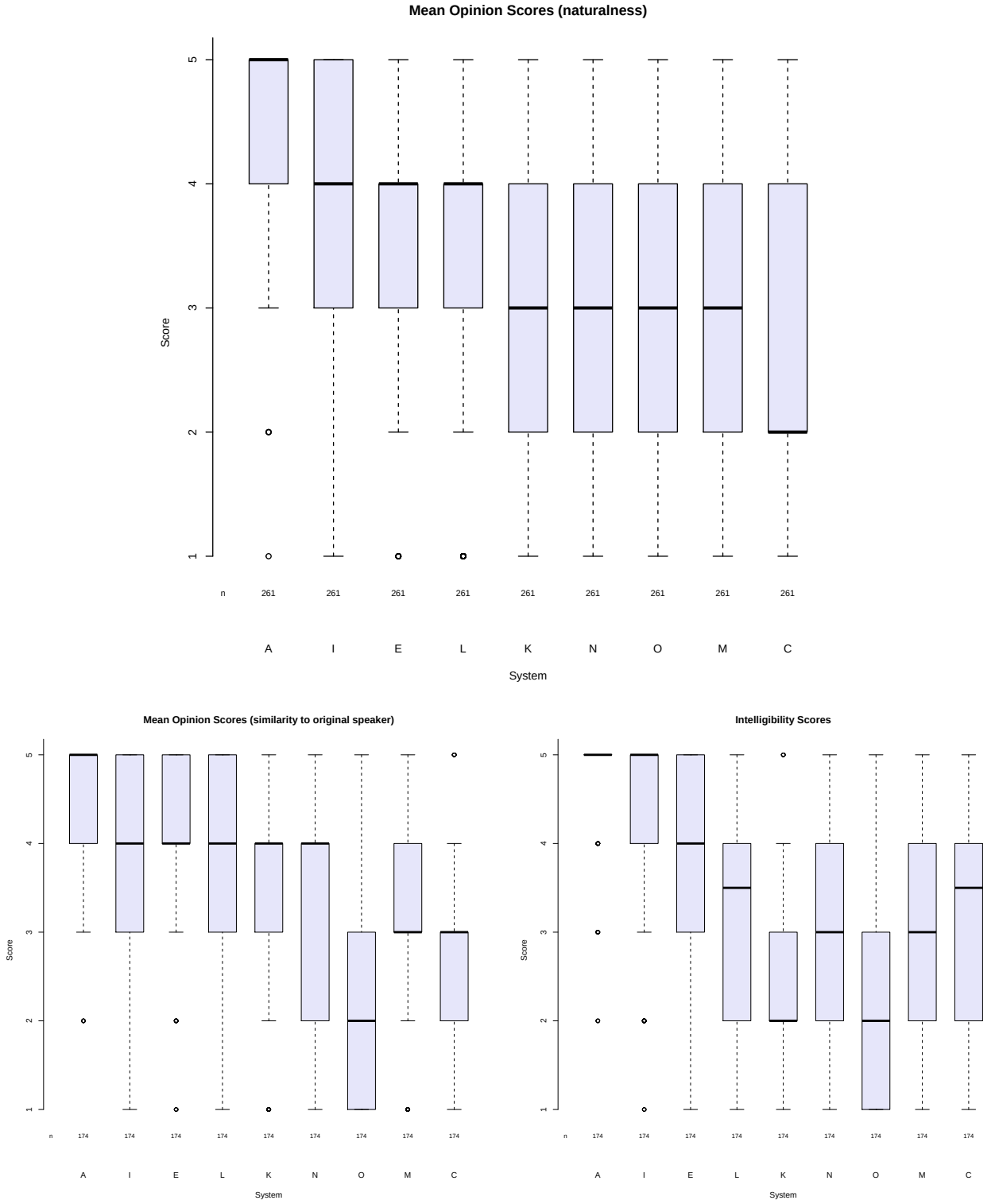

Figure 19: *Results for task 2020-SS1. A is natural speech, the remaining letters denote the systems submitted by participants.*

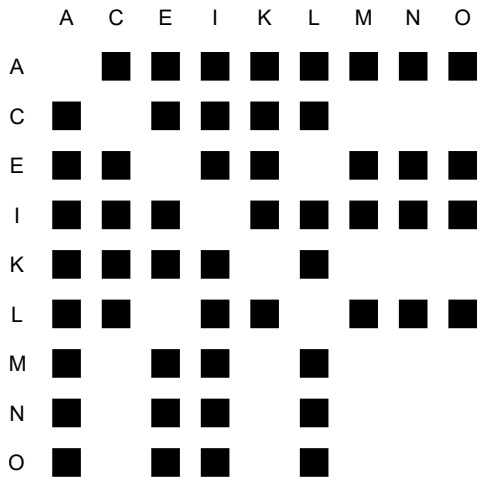

Figure 20: *Significant differences in naturalness between systems are indicated by solid black boxes for task 2020-SS1.*

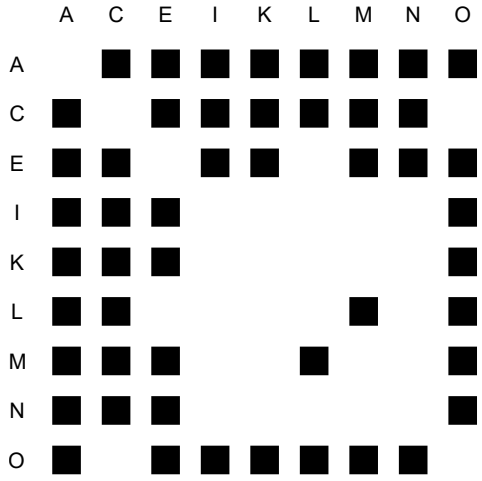

Figure 21: *Significant differences in speaker similarity between systems are indicated by solid black boxes for task 2020-SS1.*

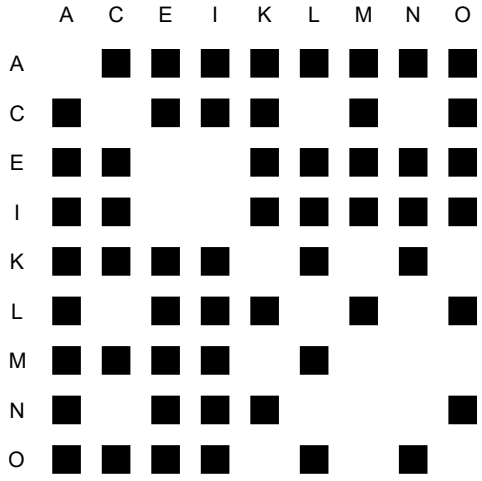

Figure 22: *Significant differences in intelligibility (INT) between systems are indicated by solid black boxes for task 2020-SS1.*

Group ID	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
MP	13	14	13	14	13	11	6	9	13	12	15	13	13	14	11	11	13
ME	4	4	6	7	8	4	8	4	7	7	5	5	4	7	5	4	4
MR	4	4	3	4	5	5	4	2	3	5	4	5	3	5	4	4	5
ALL	21	22	22	25	26	20	18	15	23	24	24	23	20	26	20	19	22

Table 4: *The numbers of listeners in different listener groups for task 2020-MH1 whose responses were used in the results.* [2]

Gender	Male	Female
Total	126	214

Table 5: *Gender.* [2]

	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	58	271	35	4	2	0	0	0

Table 6: *Age of listeners whose results were used.* [2]

Native	Yes	No
Mandarin	337	1

Table 7: *Native speakers.* [2]

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	15	10	207	92	13	0

Table 8: *Highest level of education completed.* [2]

CS/Engineering person?	Yes	No
Total	142	195

Table 9: *Computer science / engineering person.* [2]

Work in speech technology?	Yes	No
Total	123	213

Table 10: *Work in the field of speech technology.* [2]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	62	45	33	45	85	27	41

Table 11: *How often normally listened to speech synthesis before doing the evaluation.* [2]

Dialect of Chinese	Beijing dialect	Shanghainese	Cantonese	Northeast Dialect	Sichuan dialect
Total	42	32	21	21	14

Table 12: *Dialect of Chinese native speakers.* [2]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	265	14	50	8

Table 13: *Speaker type used to listen to the speech samples.* [2]

Same environment?	Yes	No
Total	329	6

Table 14: *Same environment for all samples?* [2]

Environment	Quiet all the time	Quiet most of the time	Equally quiet and noisy	Noisy most of the time	Noisy all the time
Total	210	105	15	2	0

Table 15: *Kind of environment when listening to the speech samples.* [2]

Number of sessions	1	2-3	4 or more
Total	175	107	43

Table 16: *Number of separate listening sessions to complete all the sections.* [2]

Browser	Chrome	Firefox	Safari	IE	Opera	Mozilla	Other
Total	167	24	31	51	2	0	60

Table 17: *Web browser used.* [2]

Similarity with reference samples	Easy	Difficult
Total	247	77

Table 18: *Listeners' impression of their task in the section(s) about similarity with original voice.* [2]

Problem	Scale too big, too small, or confusing	Issues with hardware	Other
Total	63	3	20

Table 19: *Listeners' problems in the section(s) about similarity with original voice.* [2]

Number of times	1-2	3-5	6 or more
Total	223	75	2

Table 20: *Number of times listened to each example in the section(s) about similarity with original voice.* [2]

Naturalness	Easy	Difficult
Total	283	41

Table 21: *Listeners' impression of their task in the MOS naturalness sections.* [2]

Problem	Difficulties with judging naturalness	Scale too big, too small, or confusing	Issues with hardware	Other
Total	24	24	2	1

Table 22: *Listeners' problems in the MOS naturalness sections.* [2]

Number of times	1-2	3-5	6 or more
Total	260	34	1

Table 23: *Number of times listened to each example in the MOS naturalness sections.* [2]

Naturalness	Easy	Difficult
Total	239	90

Table 24: *Listeners' impression of their task in the sections involving news paragraphs.* [2]

Problem	Difficulties with judging naturalness	Scale too big, too small, or confusing	Issues with hardware	Other
Total	39	55	7	12

Table 25: *Listeners' problems in the sections involving news paragraphs.* [2]

Number of times	1-2	3-5	6 or more
Total	257	48	0

Table 26: *How many times listened to each example in the sections involving news paragraphs.* [2]

INT section(s)	Usually understood all the words	Usually understood most of the words	Very hard to understood all the words	Typing problems: words too hard to spell, or too fast to type
Total	94	203	16	9

Table 27: *Listeners' impressions of the intelligibility task (INT).* [2]