The Blizzard Challenge 2021

Zhen-Hua Ling[1], Xiao Zhou[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]Centre for Speech Technology Research, University of Edinburgh, UK
zhling@ustc.edu.cn, xiaozh@mail.ustc.edu.cn, Simon.King@ed.ac.uk

Abstract

The Blizzard Challenge 2021 is the seventeenth annual Blizzard Challenge. A European Spanish dataset was provided to participants this year and two tasks were designed. The hub task is to synthesize texts containing only Spanish words. The spoke task is to synthesize Spanish texts containing a small number of English words in each sentence. Twelve and ten teams submitted their results for these two tasks, respectively. Listening tests were conducted online to evaluate the naturalness, intelligibility and speaker similarity of the synthetic speech. In addition to these conventional metrics, the subjective acceptability of the English words in Spanish sentences was measured for the spoke task. The top system in the hub task achieved comparable naturalness with, and better speaker similarity than, the natural reference speech.
Index Terms: Blizzard Challenge, speech synthesis, evaluation, listening test

1. Introduction

The first Blizzard Challenge was held in 2005 [1] and there have been annual summary papers like this one every year. 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 [2].

The Blizzard Challenge 2021 was 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. The majority of previous challenges have used English speech databases. The other languages used in previous challenges are Mandarin Chinese (Blizzard Challenges 2008-2010 and 2019-2020) and several Indian languages (Blizzard Challenges 2013-2015). The Blizzard Challenge 2021 is the first time that a non-English European language, i.e., European Spanish, has been used. In addition to the standard hub task of synthesizing Spanish texts, a spoke task was desgined to synthesize Spanish texts containing a few English words. This paper will present the details of the speech dataset, tasks, participating systems, listening tests and results of the challenge.

2. Voices to build

2.1. Speech dataset

A European Spanish speech dataset kindly provided by iFLYTEK Co., Ltd. was released for voice building this year. The dataset contains recorded speech from a professional female native European Spanish speaker together with text transcriptions. The texts were from various domains, including dialogue, custom service, daily life, travelling, etc. The speech was recorded in a studio with quiet environment; the total duration of the waveform files – sampled at 48 kHz – amounts to around 9.5 hours.

2.2. Tasks

There were two tasks in the Blizzard Challenge 2021, both using the same speech dataset (introduced above).

- Hub task 2021-SH1: Each participating team should build a voice from the provided European Spanish data to synthesize texts containing only Spanish words, following the challenge rules[1].
- Spoke task 2021-SS1: Each participant should build a voice from the provided European Spanish data to synthesize Spanish texts containing a small number of English words in each sentence.

For both tasks, the submitted synthetic speech should be single channel, 16 bit depth, and at any standard sampling rate (e.g., 16 kHz, 22.05 kHz, 44.1 kHz, or 48 kHz).

Regarding the use of external data, the Blizzard Challenge 2021 required that each participant must use no more than 100 hours of audio (including the provided data) for each task, and must choose one of the two options below.

- Option A: only use freely-available external data, and report the exact data being used.
- Option B: use any data, whether freely-available or not.

Participants were asked to report their options when submitting synthetic speech and in their paper.

For the hub task 2021-SH1, teams were required to synthesise 1310 test sentences (disjoint from the training data) that contained only Spanish words and were composed as follows.

- MOS: 510 distinct sentences, to be used for naturalness and speaker similarity evaluation. These sentences were from the same source corpus as the training data.
- INT: 800 distinct sentences for intelligibility evaluation. 700 were taken from the Sharvard corpus [3], a phonemically-balanced Spanish sentence dataset designed for for intelligibility testing. The other 100 semantically unpredictable sentences (SUS) were kindly generated by TALP-UPC[2] and Aholab-EHU[3] research laboratories using the grammatical structures proposed by Grice [4].

For the spoke task 2021-SS1, teams were required to synthesise 224 test sentences, being Spanish texts containing a few English words. For most sentences, the number of English words was no more than 5. The task definition did not specify how the English words should be pronounced (e.g., British English, North American English, fully or partially nativised

[1]https://www.synsig.org/index.php/Blizzard_Challenge_2021_Rules
[2]http://www.talp.upc.edu/
[3]https://aholab.ehu.eus/

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

Short name	Team	Acoustic Model	Vocoder
NATURAL	Natural speech	Human	Human
ADAPT-VT	Trinity College Dublin	Conformer-based FastSpeech2	Parallel WaveGAN
Aholab	University of the Basque Country (UPV/EHU) - Aholab Signal Processing Laboratory	Tacotron2	WaveGlow
CPQD-Unicamp	CPQD Foundation and Dept. of Computer Engineering and Automation, School of Electrical and Computer Engineering, University of Campinas (Unicamp)	Tacotron2	Parallel WaveGAN
CSTR	University of Edinburgh	FastPitch	WaveGlow
DelightfulTTS	Microsoft	Sequence-to-Sequence	HiFiNet
IMS	University of Stuttgart, Germany, Institute for Natural Language Processing, Digital Phonetics Research Group	FastSpeech2	MelGAN
IOA-THINKIT	Institute of Acoustics, Chinese Academy of Sciences	Sequence-to-Sequence	HiFiGAN
SCUT	South China University of Technology	Tacotron2	Multiband WaveRNN
SRCB-SL	Samsung Research China - Beijing (SRCB)	FastSpeech-based	HiFiGAN
SUTD-NUS	Singapore University of Technology and Design, Singapore National University of Singapore, Singapore	Tacotron2	MelGAN
tal_speech	TAL	BERT+GST-Tacotron	HiFiGAN
VivaVoice	XiaoyingTech (*no paper submission*)	DNN-based	HiFiGAN
VRAIN-UPV	Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV)	ForwardTacotron	HiFiGAN
wechat_ihearing	Tencent	Transformer and RNN based hybrid model	LPCNet

to Spanish, etc.) and so making that decision was part of the challenge faced by participating teams.

3. Participants

14 teams submitted results: 12 for the hub task and 10 for the spoke task: Table 1. No benchmark systems were employed this year; there seems little value in continuing with unit selection (Festival) or HMM-based (HTS) benchmarks, or indeed the Merlin DNN system. Given the rapidly-changing array of neural models available, the choice of neural benchmark model is not clear. Following previous challenges, all systems are identified using letters in these published results. This year, letter R denotes natural ('real') speech. Letters A to N were assigned randomly to the systems submitted by participants. Each participating team is free to choose whether to reveal their system identifier in their workshop paper.

In Table 1 we see that all systems this year adopted a neural approach, and the great majority employed a sequence-to-sequence acoustic model, such as Tacotron, FastSpeech2, and so on. Neural vocoders were also adopted by all teams, of which the majority (10 out of 14) were GAN-based, including HiFiGAN, Parallel WaveGAN, and MelGAN.

4. Listening tests

4.1. Listening test materials

For the hub and spoke tasks, 1310 and 224 test sentences were synthesised by participants, respectively. Similar to previous challenges, only a relatively small subset of those sentences (77 for the hub task / 66 for the spoke task) were actually used in the listening test. This means that there is a large amount of synthetic speech material available to use in future listening tests. Please refer to the summary papers of previous challenges [5] for a description of the listening test design and the web interface used to deliver it. The detailed listening test results will be distributed via the Blizzard Challenge website [6] in a package also including all submitted synthetic speech.

4.2. Listener types

Similarly to previous years, there were three types of listeners in the test.

- Paid listeners. This year, the organizers recruited paid listeners (denoted SP), through the Prolific crowdsourcing platform[4]. These listeners were all self-certified native speakers of Spanish and all instructions and other text on the listening test webpages for this listener type were in Spanish.
- Speech experts (self-declared), recruited via participating teams and mailing lists (denoted SE). All text on the listening test webpages for SE listeners was in English.
- Volunteers recruited via participating teams, mailing lists, etc. (denoted SR). Again, the text on the listening test webpages for SR listeners was in English.

[4] https://prolific.co/

Following previous challenges, the organisers asked participating teams to help recruit speech experts and volunteer listeners. 9 of the 14 teams complied with this requirement and collectively recruited 81 listeners in total of types SE and SR.

4.3. Listening test design

The listening tests for 2021-SH1 consisted of six sections, as follows:

- Similarity, MOS sentences
- Similarity, MOS sentences
- Naturalness, MOS sentences
- Naturalness, MOS sentences
- Intelligibility, INT sentences (Sharvard)
- Intelligibility, INT sentences (SUS)

In each of the above sections, one example from each system was played to the listener. For the first four sections, system R (natural recordings of the original speaker) was included. Due to the lack of natural recordings from the original speaker for the INT sentences, natural recordings of the female voice from the Sharvard corpus were used as system R in section 5; section 6 did not include system R. Therefore, there were 13 samples in each of the first 5 sections and 12 samples in the last section.

The listening tests for 2021-SS1 consisted of six sections:

- Similarity, MOS sentences
- Naturalness, MOS sentences
- Naturalness, MOS sentences
- Acceptability, Spanish sentences containing English words
- Acceptability, Spanish sentences containing English words
- Acceptability, Spanish sentences containing English words

Each of the above sections included one example from each system, including system R (natural speech). Thus, there were 11 samples in each sections.

As in previous challenges, the presentation order of systems in each section was determined by a Latin Square design. In addition, no listener heard the same sentence more than once throughout the whole test, which is especially important the intelligibility sections. Each INT sentence could be played at only once.

The methodology for scoring in the various sections of both tasks this year was the same as the previous Blizzard Challenge [7]. MOS was computed for the naturalness and speaker similarity sections and Word Error Rate (WER) for the intelligibility sections. A new metric, the acceptability of the English words (ACC), was employed for the spoke task. The instructions given to the listeners when evaluating ACC can be found in the Appendix.

The full wording of the instructions for both tasks in both English and Spanish are included in the released package of submitted samples and listening test results, linked from the Challenge website [6].

4.4. Listening test completion rate

Table 2 and Table 3 show the statistics of evaluation completion rates for different listener types in the two tasks. We can see that the overall completion rates this year (78.2% and 82.6%) were slightly lower than that in 2019 (84.8%) [7]. This is, in part, because most SE and SR listeners were non-native and therefore unable to complete the intelligibility sections.

Table 2: *Listener registration and evaluation completion rates for task 2021-SH1.*[1]

	Registered	No response at all	Partial evaluation	Completed evaluation
SP	254	35	21	198
SE	76	12	10	54
SR	87	9	4	74
ALL	**417**	**56**	**35**	**326**

Table 3: *Listener registration and evaluation completion rates for task 2021-SS1.*[1]

	Registered	No response at all	Partial evaluation	Completed evaluation
SP	132	13	10	109
SE	41	3	3	35
SR	40	5	3	32
ALL	**213**	**21**	**16**	**176**

To get the final listening test results, we further excluded some listeners from "completed all sections" if they gave uniform scores in at least one section or low-effort responses in the intelligibility sections. As shown in Table 4, 313 and 176 valid listeners were used to calculate the final results for the two tasks respectively.

Table 4: *The number of listeners whose responses were used to calculate the final results.*

	Task SH1	Task SS1
SP	187	109
SE	54	35
SR	72	32
ALL	**313**	**176**

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. A package of submitted synthetic speech and complete listening test results (eventually to include raw listener scores for each stimulus) is available via the Challenge website [6]. This allows, amongst other things, re-analysis of the listening test results by interested parties. We followed the statistical analysis techniques described in [8] to produce the listening test results. In this paper and in the listening test results distributed by the organizers, all system names are in an anonymous form. The participating teams a free to decide whether or not to reveal their system identifiers in their workshop papers. Additionally, a summary of listener questionnaire responses for task 2021-SH1 are shown in Tables 5 to 25.[2]

6. Results

According to the questionnaires returned by participating teams, system F and system G in task 2021-SH1 and system I in task 2021-SS1 adopted the Option B of using external audio data, i.e., some internal and non-freely-available data was used. Other systems either followed Option A or didn't use any

external audio data.

The listening test results are shown in Figures 1 to 9. The standard boxplots are employed to present the ordinal data, e.g., mean opinion scores (MOS). Please refer to [5, 8] for more information on how to interpret the boxplots. In all figures of each task, 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 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 task 2021-SH1, when combining the opinions of all listeners, system F achieved significantly better naturalness and similarity than all other submitted systems as shown in Fig. 2 and Fig. 3. The naturalness ratings of natural speech (system R) and system F were 4.21 and 4.17 respectively, and the difference is statistically insignificant as shown in Fig. 2. The other systems were not as natural as natural speech. Among the systems that only used freely-available external data, systems K, J and I achieved comparable naturalness and their performance was significantly better than other submitted systems. The speaker similarity scores of system R and system F were 4.07 and 4.35 respectively, and this difference is, somewhat surprisingly, statistically significant as shown in Fig. 3. One possible reason that system F achieved higher speaker similarity than natural recordings is that there existed style variations in the corpus, but we used two fixed natural utterances as references for all similarity sections of all listener groups. Additionally, system I was rated as equally similar to the target speaker as the natural recordings. No other systems were rated as similar.

Regarding intelligibility, only the results of SP listeners are reported, because there were very few native Spanish listeners of SE and SR types. In the intelligibility test using Sharvard sentences, the differences between all pairs of systems are insignificant, except that system N (the system with highest WER) is significantly less intelligible than systems G and J (the two systems with lowest WER), as shown in Fig. 4. As introduced in Section 4.3, we didn't have natural recordings of SUS sentences, so no comparisons with natural speech are possible for this sentence type. As shown in Fig. 5, the differences between most system pairs are insignificant, excluding systems L and N (the two systems with highest WER).

In this year's task 2020-SS1, when combining the opinions of all listeners, no system was as natural as natural speech, as shown in Fig. 7. System K was rated as significantly more natural than all other submitted systems, except I. System K was also rated as sounding as similar to the target speaker as natural speech and significantly different to all other submitted systems, as shown in Fig. 8. Regarding the acceptability of English words in Spanish sentences, systems K and I achieved significantly higher ratings than all other submitted systems. However, there was still a clear gap between their acceptability ratings (mean MOS of 3.41 and 3.40) and that of natural speech (4.14).

Listeners ratings of acceptability and naturalness are unlikely to be independent. Indeed, the correlation coefficient is 0.96. It is not possible to determine how much of this correlation is down to better vs. worse systems, and how much is a consequence of listeners' inability to judge the acceptability of the English words independently of other parts of the utterance. The correlation between acceptability and speaker similarity is also high, at 0.91.

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 data. Sébastien Le Maguer at Trinity College Dublin helped to anonymize and normalize the submitted data. Pilar Oplustil Gallegos translated the listening test instructions into Spanish. María Fernandez proofread the instructions and also provided lists of acceptable mis-spellings used for the WER calculations. 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] "The Blizzard Challenge website," http://www.synsig.org/index.php/Blizzard_Challenge.

[3] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology," *International journal of audiology*, vol. 53, no. 9, pp. 633–638, 2014.

[4] M. Grice, "Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[5] M. Fraser and S. King, "The Blizzard Challenge 2007," *Proc. SSW6*, 2007.

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

[7] X. Zhou, Z.-H. Ling, and S. King, "The Blizzard Challenge 2020," *Proc. Blizzard Challenge Workshop*, 2020.

[8] 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.

[1]Calculated from listening test results.

[2]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 313.

9. Appendix

When evaluating the acceptability of the English words (ACC) in task 2021-SS1, listeners were given the following instructions in English,

"In this section, after you listen to each sentence, you will choose a score for the audio file you've just heard. This score should reflect your opinion on how acceptable or unacceptable the English words in the sentence sounded. You should not judge the grammar or content of the sentence or the quality of other Spanish words in the sentence, just how the English words sound.

Listen to the example below.

Now choose a score for how acceptable or unacceptable the English words sounded. The scale is from 1 [Not Intelligible] to 5 [Perfect].

1 : Not Intelligible
2 :
3 : Acceptable
4 :
5 : Perfect"

or the following instructions in Spanish.

"En esta sección, después de que escuches cada frase, tendrás que escoger una puntuación para el audio que acabas de escuchar. Esta puntuación debe reflejar tu opinión sobre lo aceptable o inaceptable que suenan las palabras en inglés en esta frase. No tienes que evaluar la gramática o el contenido de la frase, ni tampoco la calidad de las otras palabras en español en la frase, solamente cómo suenan las palabras en inglés.

Escucha el ejemplo más abajo.

Ahora, escoge una puntuación según lo aceptable o inaceptable que suenan las palabras en inglés. La escala es de 1 [No se entienden en absoluto] a 5 [Perfecto].

1 : No se entienden
2 :
3 : Aceptables
4 :
5 : Perfectas"

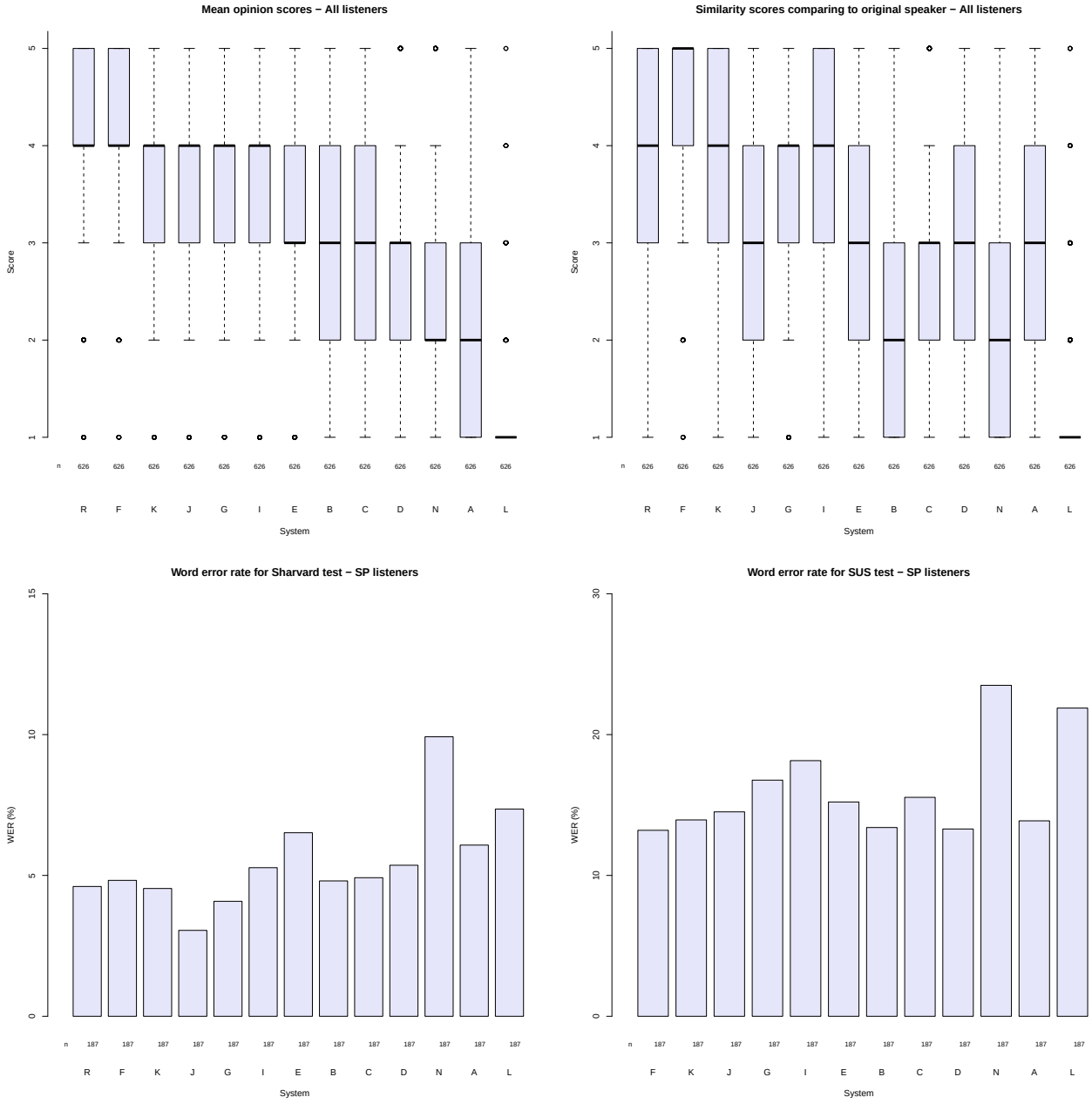

Figure 1: *Results for task 2021-SH1. R is natural speech, the remaining letters denote the systems submitted by participants. For intelligibility tests (Sharvard and SUS), only the results of SP listeners were reported considering that there were very few native Spanish listeners of SE and SR types.*

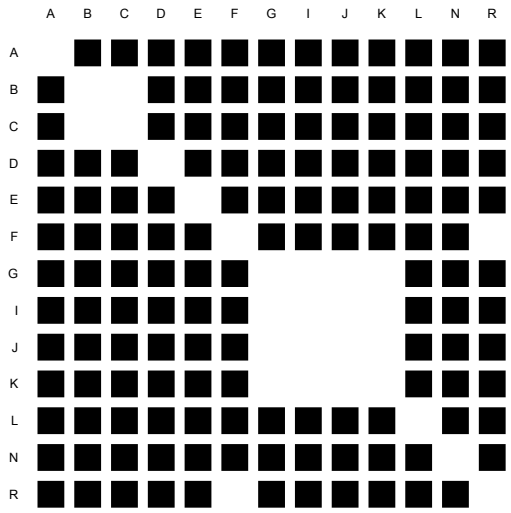

Figure 2: *Significant differences in naturalness between systems are indicated by solid black boxes for task 2021-SH1.*

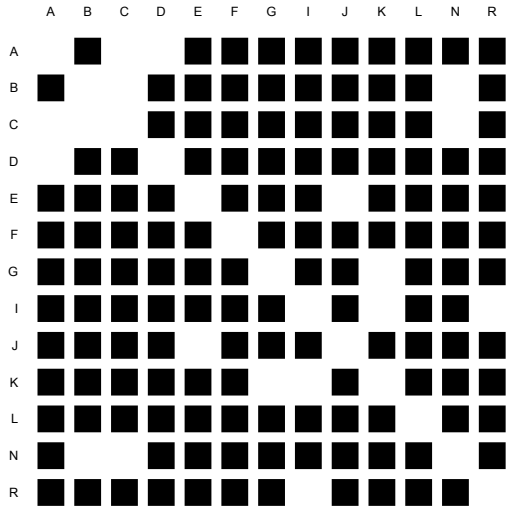

Figure 3: *Significant differences in speaker similarity between systems are indicated by solid black boxes for task 2021-SH1.*

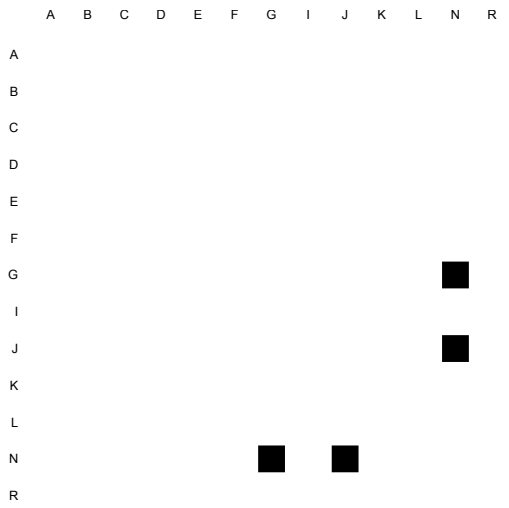

Figure 4: *Significant differences in intelligibility (INT) of Sharvard sentences between systems are indicated by solid black boxes for task 2021-SH1.*

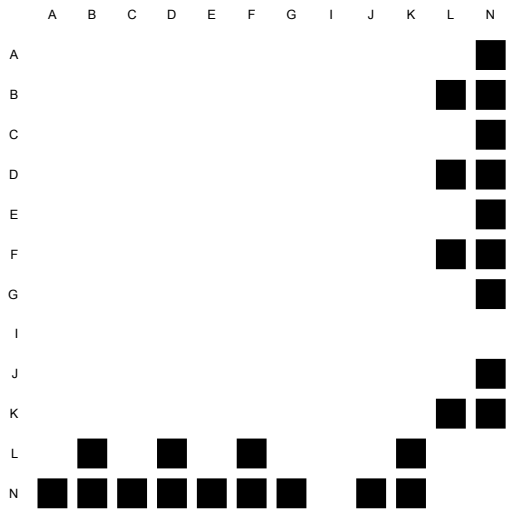

Figure 5: *Significant differences in intelligibility (INT) of SUS sentences between systems are indicated by solid black boxes for task 2021-SH1.*

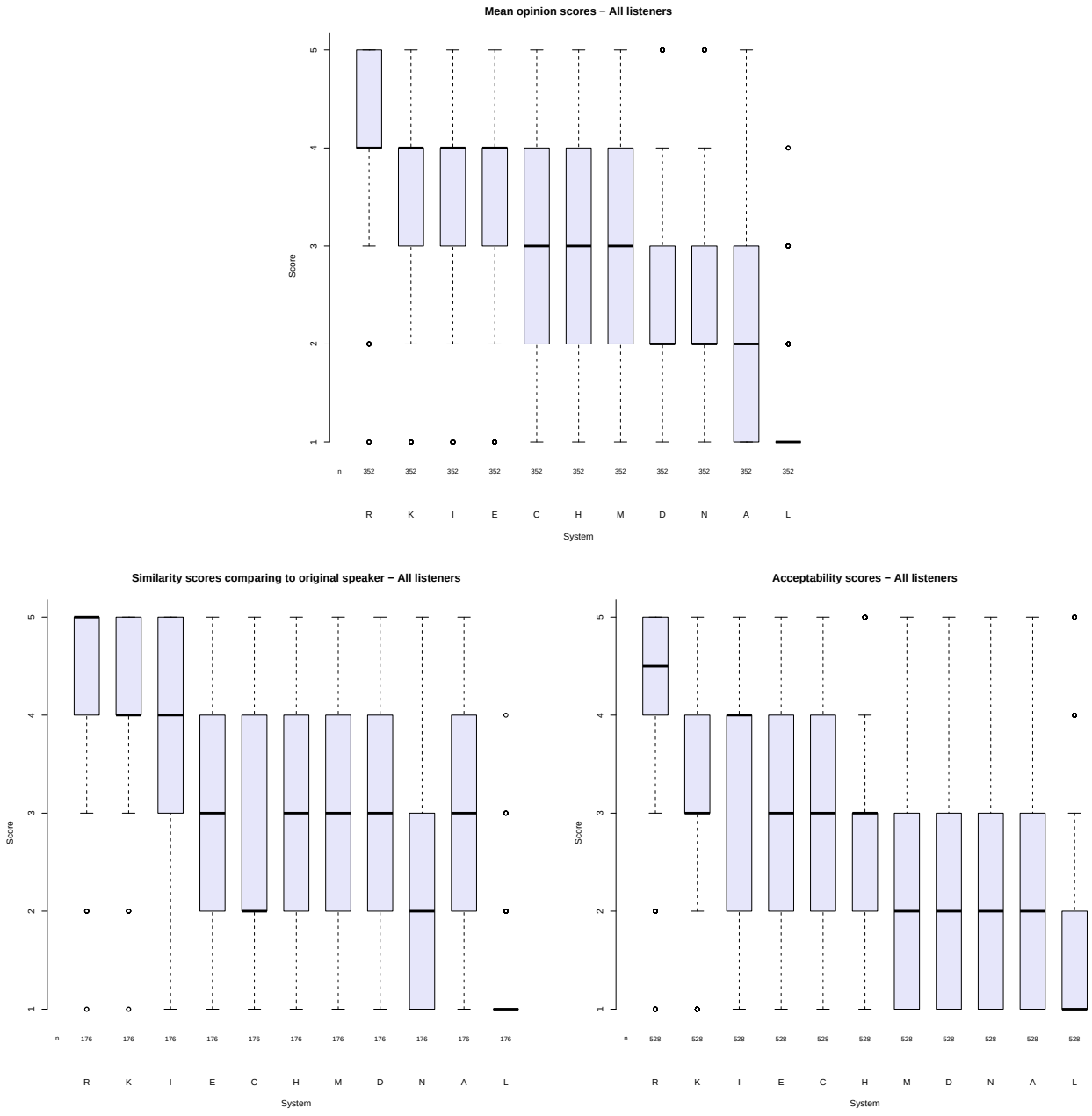

Figure 6: *Results for task 2021-SS1. R is natural speech, the remaining letters denote the systems submitted by participants.*

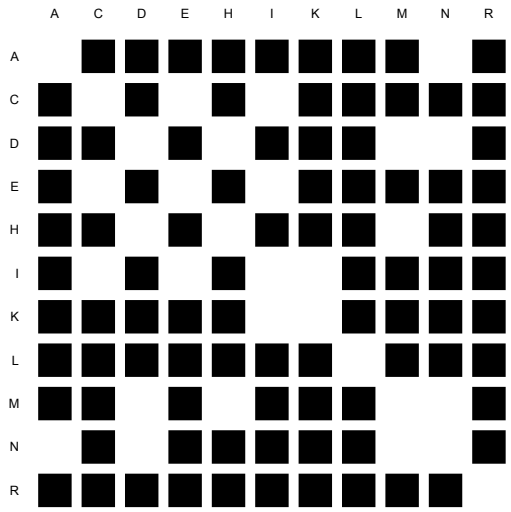

Figure 7: *Significant differences in naturalness between systems are indicated by solid black boxes for task 2021-SS1.*

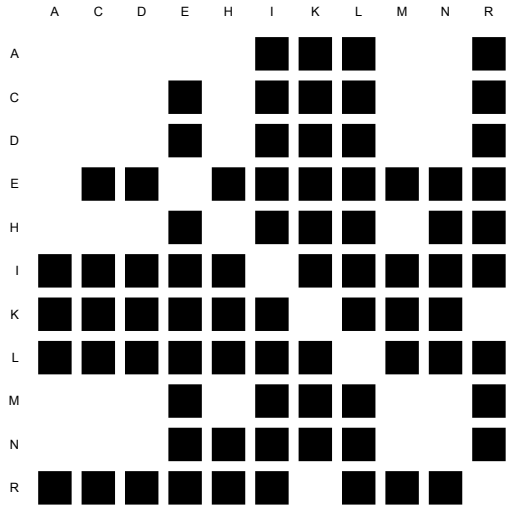

Figure 8: *Significant differences in speaker similarity between systems are indicated by solid black boxes for task 2021-SS1.*

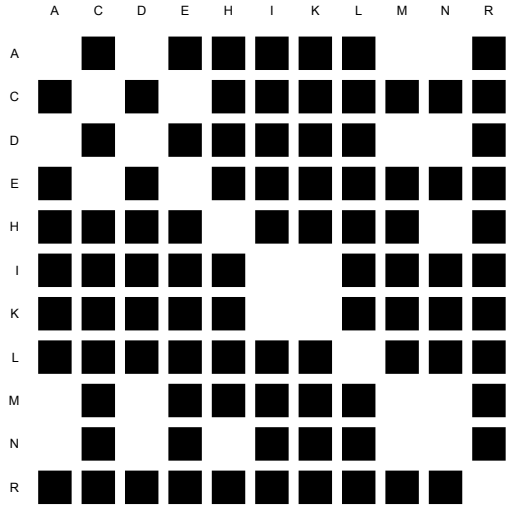

Figure 9: *Significant differences in acceptability of English words between systems are indicated by solid black boxes for task 2021-SS1.*

Group ID	01	02	03	04	05	06	07	08	09	10	11	12	13
SP	14	13	14	14	17	16	14	11	15	15	15	14	15
SE	4	6	5	5	4	3	5	5	4	5	7	2	5
SR	5	5	8	6	4	5	3	5	5	3	6	6	5
ALL	23	24	27	25	25	24	22	21	24	23	28	22	25

Table 5: *The numbers of listeners in different listener groups for task 2021-SH1 whose responses were used in the results.*[2]

Gender	Male	Female
Total	143	165

Table 6: *Gender.*[2]

	under 20	20-29	30-39	40-49	50-59	60-69	70-79	over 80
Total	34	187	48	22	9	4	0	1

Table 7: *Age of listeners whose results were used.*[2]

Native	Yes	No
Spanish	223	82

Table 8: *Native speakers.*[2]

Level	High School	Some College	Bachelor's Degree	Master's Degree	Doctorate	Other
Total	27	80	100	79	17	2

Table 9: *Highest level of education completed.*[2]

CS/Engineering person?	Yes	No
Total	181	124

Table 10: *Computer science / engineering person.*[2]

Work in speech technology?	Yes	No
Total	83	221

Table 11: *Work in the field of speech technology.*[2]

Frequency	Daily	Weekly	Monthly	Yearly	Rarely	Never	Unsure
Total	72	92	41	46	32	8	17

Table 12: *How often normally listened to speech synthesis before doing the evaluation.* [2]

Dialect of Spanish	Valenciano	Euskara	Catalán	Gallego	Asturiano	Extremeño	Aragonés	Leonés	Other
Total	15	15	13	4	4	3	0	0	124

Table 13: *Dialect of Spanish native speakers.*[2]

Speaker type	Headphones	Computer Speakers	Laptop Speakers	Other
Total	255	24	25	5

Table 14: *Speaker type used to listen to the speech samples.*[2]

Same environment?	Yes	No
Total	299	9

Table 15: *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	177	102	25	4	1

Table 16: *Kind of environment when listening to the speech samples.*[2]

Number of sessions	1	2-3	4 or more
Total	230	51	28

Table 17: *Number of separate listening sessions to complete all the sections.*[2]

Browser	Chrome	Firefox	Safari	IE	Opera	Mozilla	Other
Total	198	41	39	6	9	1	14

Table 18: *Web browser used.*[2]

Similarity with reference samples	Easy	Difficult
Total	246	58

Table 19: *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	33	10	17

Table 20: *Listeners' problems in the section(s) about similarity with original voice.*[2]

Number of times	1-2	3-5	6 or more
Total	227	58	4

Table 21: *Number of times listened to each example in the section(s) about similarity with original voice.*[2]

Naturalness	Easy	Difficult
Total	254	53

Table 22: *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	10	34	9	10

Table 23: *Listeners' problems in the MOS naturalness sections.*[2]

Number of times	1-2	3-5	6 or more
Total	244	41	5

Table 24: *Number of times listened to each example in the MOS naturalness sections.*[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	147	80	27	34

Table 25: *Listeners' impressions of the intelligibility task (INT).*[2]