The Blizzard Challenge 2023

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly

18th Blizzard Challenge Workshop
29 August 2023, Grenoble, France

The Blizzard Challenge 2023

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France

firstname.lastname@grenoble-inp.fr

Abstract

The Blizzard Challenge 2023 is the eighteenth edition of the text-to-speech synthesis Blizzard Challenge. This year, two French datasets were provided to participants and two tasks were designed. The Hub task was to build a voice from a 51-hour single speaker dataset, restricted to using only publicly-available data. The Spoke task consisted of building a voice from a 2-hour single speaker dataset that sounds as close as possible to that speaker. There were no restrictions on the use of data for the spoke task. 18 teams participated in the hub task and 14 in the spoke task. All teams used neural-based systems. Synthesised samples were evaluated in terms of speech quality, speaker similarity and intelligibility.

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

1. Introduction

In order to better understand and compare research techniques in building corpus-based speech synthesisers on the same data, Blizzard Challenges have been held yearly from 2005 to 2021, each accompanied by a summary paper [1, 2, 3, for example]. Each Challenge requires its participants to complete broadly similar tasks: take the released training speech data, build synthetic voices, and synthesise a prescribed set of test sentences. The output from each synthesiser is then evaluated through extensive listening tests. Previous Challenges have tackled: English (2005-2013; 2015-2018); Asian languages such as Mandarin (2008-2010; 2019-2020) and Shanghainese (2020); Indian languages (2013-2015). In the most recent Challenge, a non-English European language was introduced for the first time: European Spanish (2021).

This year, the Challenge was organised by the Université Grenoble Alpes and used Metropolitan French (French from France). This summary paper presents the details of the speech dataset, tasks, participating systems, evaluation methods and results of the challenge. For the current Challenge, and many previous ones, 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 archive (Table 1).

2. Tasks

There were two tasks in the Blizzard Challenge 2023, each using a dedicated freely-available dataset (Table 1). For the complete task specification given to participants, see the challenge website (Table 1). In the following, we define "External data" as data, of any type, that is not part of the provided training data, and "External model" as a model, of any type, that has not been trained by the team (e.g., pre-trained wav2vec, BERT, etc.).

2.1. Hub task FH1: French TTS

The Hub dataset is a subset of the M-AILABS French dataset [4], comprising $51\,\mathrm{h}$ of five audiobooks read by the female speaker Nadine Eckert-Boulet (NEB) in Metropolitan French. The full list of books and chapters included in the Hub dataset can be found online (Table 1). All recordings come from the free public domain audiobook LibriVox project [5], and the texts are from the Gutenberg Project [6]. Audio files each corresponds to one chapter. They were originally at $44.1\,\mathrm{kHz}$ sample rate, but we downsampled them to $22.05\,\mathrm{kHz}$. The text processing detailed in [7] and performed on part of the books at that time was subsequently applied to the full Hub dataset for the Blizzard Challenge. We first restored the original chapter structure by aligning the text from the Gutenberg Project with the recordings from LibriVox. We normalised all texts and numbers, spelled out annotations, and manually corrected misspellings and omissions. End of paragraphs were annotated with the punctuation mark "§", which was introduced after the last punctuation mark preceding each carriage return. Text transcriptions were then segmented into utterances each time we found pauses of at least $400\,\mathrm{ms}$ in the corresponding audio. The audio of each chapter is provided together with the text segmentation: participants are free to use any parts of the room tone in-between utterances. In total, the Hub dataset includes $64\,000$ utterances from 289 original audio chapters, each with an orthographic transcription. $44\,000$ of these utterances were semi-automatically aligned with phonetic transcriptions. The phonetic alphabet is described on the datasets website (Table 1).

The objective of the Hub task was to build a voice from the provided French Hub dataset. Two reproducibility requirements were imposed on participants: 1) the use of external models was allowed only if they are publicly-available off-the-shelf pre-trained models, and references are given; 2) all audio data used for training models (including for fine-tuning pre-trained models) should be publicly available and reported. Participants were required to synthesise 3 test sets:

- $\mathsf{MOS}_{\mathsf{FH1}}$: 1000 distinct sentences read by NEB but from another audiobook not included in the Hub dataset (*"Le Vingtième Siècle : La Vie électrique" – Albert Robida)*, to be used for quality and speaker similarity evaluation.
- $\mathsf{INT}_{\mathsf{FH1}}$: 326 distinct sentences for intelligibility evaluation. 216 utterances include heterophonic homographs, which are "one of two or more words spelled alike but different in meaning or pronunciation" [8] (such as 'fils' which is either

This manuscript has been proof-edited by Simon King (Papercup / University of Edinburgh). This version has not been peer-reviewed.

10.21437/Blizzard.2023-1

Table 1: *Blizzard Challenge 2023 web resources.*

Name	Link
Challenge website (call, rules)	https://www.synsig.org/index.php/Blizzard_Challenge_2023
Challenge datasets (FH1, FS1)	https://zenodo.org/record/7560290
Challenge archive (syntheses, tools for analysis)	https://www.cstr.ed.ac.uk/projects/blizzard/
eSpeak	https://espeak.sourceforge.net
NVIDIA Tacotron 2 implementation	https://github.com/NVIDIA/tacotron2.git
Fairseq FastSpeech 2 implementation	https://github.com/facebookresearch/fairseq/tree/main/fairseq/models/text_to_speech
HiFi-GAN implementation, models and weights	https://github.com/jik876/hifi-gan
Our Tacotron 2 full list of hyperparameter values	https://tinyurl.com/y59k3jb7
Prolific crowdsourcing platform	https://www.prolific.com

'son', pronounced /fis/, or the plural of 'fil', a thread or wire, pronounced /fil/). From [9], we identified 36 pairs of homographs and generated three sentences for each member of the pair ($3 \times 2 \times 36 = 216$ utterances). The remaining 110 sentences are semantically unpredictable sentences (SUS), generated using the method described in [10].

- $\mathsf{EXP}_{\mathsf{FH1}}$: sentences for evaluating expressivity and comprehensibility [11]. $\mathsf{EXP}_{\mathsf{FH1}}$ includes 100 sentences that are enumerations of 4 objects, in the form: "Dans mon panier, il y a: *det1 obj1 col1*, *det2 obj2 col2*, *det3 obj3 col3* et *det4 obj4 col4*" (translation: "In my basket, there are: …") where *det*, *obj* and *col* stand for determiner, object, and colour, which were randomly picked from lists of 7 objects and 5 colours. $\mathsf{EXP}_{\mathsf{FH1}}$ also includes 213 paragraphs, extracted from the same audiobook than $\mathsf{MOS}_{\mathsf{FH1}}$.

All test sets were provided to participants as orthographic text only.

2.2. Spoke task FS1: Speaker adaptation

The Spoke dataset, recorded at GIPSA-lab, comprises recordings of Aurélie Derbier, a professional theatre actress speaking Metropolitan French, whose voice is not available in the public domain. Sentences are taken from the SIWIS database [12], which is composed of isolated sentences from French novels and French parliamentary debates. Similar audio and text post-processing to the Hub dataset was performed, and we selected 2 h of this corpus to release as the Spoke dataset. In total, it includes 2515 utterances from 13 continuous audio recordings, all with aligned phonetic transcriptions. Again, participants are free to use any parts of the room tone in-between utterances.

The Spoke task is to build a voice from the provided Spoke dataset. None of the Hub task reproducibility requirements were imposed for this task, but they were highly encouraged. Participants were required to synthesise only:

- $\mathsf{MOS}_{\mathsf{FS1}}$: 400 distinct sentences from French parliamentary debates, to be used for quality and speaker similarity evaluation, from the same corpus as the training data but not part of the training Spoke dataset.

3. Systems

3.1. Systems submitted by participating teams

18 teams submitted to Hub task FH1 and 14 to Spoke task FS1, summarised in Table 6. Systems are identified using letters in all published results. This year, A denotes natural speech; BF and BT denote two benchmark systems described below; C to T are assigned (in no particular order) to the submitted systems. Each team was free to reveal their identifier in their workshop paper, but no global mapping will be published.

All systems this year used encoder-decoder architectures: 11 systems had either a FastSpeech 2-like or non-attentive Tacotron-like architecture, and 7 were variational auto-encoders conditioned on text. 15 systems employed a GAN-based vocoder for waveform generation, of which 6 of were trained end-to-end with the acoustic model.

All teams fulfilled the two mandatory reproducibility requirements for the Hub task. 11 of 14 teams also fulfilled these optional criteria for the Spoke task; the other three teams used private internal data. Despite the fact that many of the submitted systems are likely to have been built using open-source code, only GIPSA-lab, IMS, IOA-thinkIT, and TTS-cube provided links to their full system implementation, which was an optional reproducibility criterion.

3.2. Benchmark systems

3.2.1. BT: Tacotron 2 baseline

Our first baseline model combines Tacotron 2 [13] with HiFi-GAN [14]. We used the open-source NVIDIA Tacotron 2 implementation (Table 1) with minor changes (i.e., we converted the code to be compatible with TensorFlow 2). We used orthographic characters as input, pre-processed using the transliteration cleaner function provided in the implementation (which we modified to retain case). We trained all layers of the FH1 model from scratch on the Hub dataset for a total of 158 500 training steps. We fine-tuned the 100 000-step checkpoint of the FH1 model on the Spoke dataset for an additional 57 500 steps to create the FS1 model. We used the hyperparameter values recommended in the repository (batch size: 32; learning rate: 1×10^{-3}; weight decay: 1×10^{-6}). For full list, see Table 1. For waveform generation, we used the the original HiFi-GAN implementation with the provided pre-trained universal model `UNIVERSAL_V1/g_02500000` (Table 1).

3.2.2. BF: FastSpeech 2 baseline

The second baseline model combines FastSpeech 2 [15] with the same vocoder as for benchmark system BT. We trained a model using the Fairseq FastSpeech 2 implementation (Table 1) from scratch on the 43 747 utterances in the Hub dataset which have phone alignments (as described in Section 2.1). To synthesise the test set, we used eSpeak for letter-to-sound mapping (Table 1), which provides IPA transcriptions that we then map to the phoneme set used for the training data. The FH1 model for the Hub Task was trained from scratch for 333 935 training steps with a batch size of 32, a learning rate of 5×10^{-4}, and clip-norm 5.0 (Fairseq recommended values). The FS1 model for the Spoke Task took the final FH1 model and fine-tuned it for an additional 7254 steps on the Spoke dataset.

4. Evaluation method

4.1. Innovations for 2023

Evaluations focused on speech quality, speaker similarity, and intelligibility, which is largely consistent with previous Challenges, although four innovations were introduced:

Instructions: using the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets, we evaluated *quality* instead of *naturalness* because the former is a more easily-understood concept among non-experts. A recent evaluation of speech synthesis systems demonstrated that the choice of either quality or naturalness has little impact on the relative ranking of the systems [16]. Most importantly, it is vital to report the exact wording of instructions and rating scales presented to listeners [17]; these can be found in Appendix 9.2.

Fine-grained quality test – MOS then MUSHRA: the evaluation of a large number of systems on a 5-point scale makes the differentiation of perceptively close systems difficult. As supported by [18], we added a supplementary test to refine the initial Mean Opinion Scores of quality of the best-rated system only, also using the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets but with a MUSHRA design.

Fine-grained intelligibility test – language-specific task: although TTS systems are improving in global quality over time, this does not guarantee better handling of the large number of rare events [19]. To assess this, we employed the $\mathsf{INT}_{\mathsf{FH1}}$ test set of sentences that include heterophonic homographs in the intelligibility test.

Objective pre-selection of stimuli: in order to focus the subjective evaluations on stimuli that best discriminate between the capabilities of the submitted systems, we identified the test utterances that maximised the dispersion of output syntheses according to an objective measure described below, for all tests.

Besides speech quality, speaker similarity, and speech intelligibility, the number of dimensions along which TTS can be subjectively evaluated is large [20]. Although we asked participants to synthesise the $\mathsf{EXP}_{\mathsf{FH1}}$ test set, intended for evaluation of comprehensibility [11] or speech in context [21, 22], using enumerations and paragraphs respectively, we were not able to identify a relevant evaluation method. Therefore the $\mathsf{EXP}_{\mathsf{FH1}}$ material was not evaluated but instead released as a resource for others to use.

4.2. Pre-processing

All submitted synthetic audio was at a sampling rate of 16, 22.05, 24, 44.1, or 48 kHz and therefore no re-sampling was performed. Every natural and synthetic utterance was normalised to an Active Speech Level of -26 dB as measured using the sv56demo implementation of ITU P.56 [23].

4.3. Objective measures

In order to select the most informative samples for the human subjective evaluations described later, we identified the test utterances which exhibited the most variation across systems. Two objective distances were employed:

- *spectral distance* is the Dynamic Time Warping (DTW)-aligned root mean square error (RMSE) between the mel-spectrograms of a pair of speech samples.
- *duration distance* as the ratio of the DTW path length over the average mel-spectrogram length of two speech samples.

Mel-spectrograms were computed on the synthetic signals using 80 mel-bands, and window size and hop size of 1024 and 256 bins, respectively. For each task FH1 and FS1 separately: we computed the two distances for every test utterance of the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets, respectively ($N = 996$ for FH1, due to missing submissions from some of the systems; $N = 400$ for FS1), for all possible pairs of the n participating systems, resulting in a $N \times n \times n$ matrix for each distance. Each matrix was normalised by its average value across all dimensions, since the two objective distances have very different magnitudes. These two normalised matrices were summed elementwise to obtain a single $N \times n \times n$ distance matrix, then summed for each utterance to arrive at a single N-dimension vector. This 'dispersion' value captures – for each test utterance – how much the total objective distance between all system pairs varies. Utterances with a very small dispersion value are those for which all systems generated very similar synthetic speech; we assume that these would be the least informative samples to present to listeners. The most informative utterances will be selected as stimuli in section 4.4.3

4.4. Subjective evaluation

4.4.1. Tests

In total, five independent listening tests were conducted (see Table 2). Tests 1 and 4 measured Mean Opinion Score quality (section a) and speaker similarity (section b) for tasks FH1 and FS1 respectively, using the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets. 21 systems were evaluated in Test 1: ground truth A, benchmarks BF and BT, and the systems from 18 participants. 17 systems were evaluated in Test 4: ground truth A, benchmarks BF and BT, and the systems from 14 participants.

Tests 2 and 5 measured quality, again using the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets for tasks FH1 and FS1 respectively, but following a MUSHRA design and only for the few best-performing systems obtained in Tests 1.a and 4.a respectively. The procedure to identify those systems is described in section 4.5.3. Five systems were evaluated in Test 2: ground truth A, benchmark BF, and systems from three participants. Six systems were evaluated in Test 5: ground truth A, benchmark BF, and systems from four participants.

Test 3 measured intelligibility for task FH1 using the $\mathsf{INT}_{\mathsf{FH1}}$ test set. (There was no evaluation of intelligibility for FS1.) Test 3 measured Word Error Rate in a transcription task of SUS stimuli (section a), and Pronunciation Accuracy of homographs with an ABX task (section b). 20 systems were evaluated in Test 3: benchmarks BF and BT, and systems from 18 participants. (We do not have ground truth recordings of the $\mathsf{INT}_{\mathsf{FH1}}$ test set.)

4.4.2. Design

Following [24], test sections 1.a, 1.b, 3.a, 3.b, 4.a and 4.b were divided into experimental blocks that were each assigned to a different listener group. The number of experimental blocks and sentences to be evaluated was determined by the total number of systems under evaluation denoted by n in the following: $n = 21$ for Sections 1.a and 1.b ; $n = 20$ for Section 3.a and 3.b ; and $n = 17$ for Sections 4.a and 4.b. System orderings within blocks were systematically varied by using a Latin Square design. For test sections 1.a, 1.b, 4.a and 4.b (resp. 3.a):

- One experimental block consisted in the presentation of two different sentences (resp. one sentence) per system for a total of $2n$ (resp. n) sentences, so that all the sentences and all the

Table 2: *The five listening tests introduced in Section 4.4.1, their design as described in Section 4.4.2 and their approximate median duration per subject as reported by Prolific. "sent." stands for sentence, and system orderings within blocks were systematically varied by using a Latin Square design to ensure that all sentences and systems combinations were eventually rated for each test. Complementary information on test participants is provided in Table 3.*

Test	Task	Dimension	Design	# systems	# sent.	Implementation	Duration
1.a	FH1	Quality	Mean Opinion Score	21 (A + BF + BF + 18 systems)	42	2 sent. per system × 21 blocks	20 min
1.b	FH1	Similarity	Mean Opinion Score		42	2 sent. per system × 21 blocks	
2	FH1	Quality	MUSHRA	5 (A + BF + 3 best systems in 1.a.)	20	5 systems per sent.	27 min
3.a	FH1	Intelligibility	Transcription (SUS)	20 (BF + BF + 18 systems)	20	1 sent. per system × 20 blocks	22 min
3.b	FH1	Intelligibility	ABX (Homographs)		72	36 pairs of homo. × 20 blocks	
4.a	FS1	Quality	Mean Opinion Score	17 (A + BF + BF + 14 systems)	34	2 sent. per system × 17 blocks	13 min
4.b	FS1	Similarity	Mean Opinion Score		34	2 sent. per system × 17 blocks	
5	FS1	Quality	MUSHRA	6 (A + BF + 4 best systems in 4.a.)	20	6 systems per sent.	30 min

systems under evaluation were heard within one block.

- n experimental blocks with a circular permutation of systems ensured that all sentences and systems combinations were eventually rated by the n groups of listeners

For Test section 3.b, the test set comprised 36 pairs of homographs, each included in three different context utterances, for a total of $72 \times 3 = 216$ stimuli. We designed 3 versions of Test section 3.b, one for each context utterance. In each version, each pair of homographs is therefore presented in one context, and the combinations of homographs and systems are again divided into n experimental blocks following a Latin square design ($n = 20$):

- One experimental block consisted in the presentation of 72 sentences (the 36 pairs of homographs), with a rotation of the n systems for each sentence.
- n experimental blocks with a circular permutation of systems ensured that all sentences and systems combinations were eventually rated by the n groups of listeners

For Test 3, we ran a first round of tests with n groups of listeners performing Section 3.a and the first version of Section 3.b. Then, in a second round, n groups of different listeners performed Section 3.a and the second version of Section 3.b. Due to a lack of time, only Versions 1 and 2 were evaluated in the current results.

Tests 2 and 5 adopted a simpler design since for each sentence, all systems were evaluated at once, comparatively. 20 sentences were selected for each test and all listeners performed the same test.

4.4.3. Materials

Tests 1 and 4 used sentences from the $\mathsf{MOS}_{\mathsf{FH1}}$ and $\mathsf{MOS}_{\mathsf{FS1}}$ test sets for tasks FH1 and FS1, respectively. For each test, we selected the $4n$ sentences that maximised the objective distance dispersion between the systems (see section 4.3). Half were assigned to the speech quality evaluation (Section a) and the $2n$ others to the speaker similarity evaluation (Section b).

Once the systems that obtained the best speech quality MOS were selected (see details in Section 4.5.3), the 20 sentences from Section 1.a (resp. section 4.a) that maximised the objective distance dispersion between the selected systems were kept for Test 2 (resp. 5). Therefore, sentences for the MUSHRA quality tests are subsets of those from the MOS quality tests.

For Test section 3.a, a subset of the 20 SUS sentences from the $\mathsf{INT}_{\mathsf{FH1}}$ test set that maximised the objective distance dispersion between systems were used for evaluation.

Table 3: *Number of listeners per recruitment type (Prolific and Volunteers) for each test. The number of retained listeners after screening over the total number of completed tests is indicated.*

Test	Prolific	Volunteers	Total	
1.a	322 / 324	39 / 39	361 / 363	(99%)
1.b	316 / 317	32 / 32	348 / 349	(99%)
2	30 / 43	17 / 20	47 / 63	(75%)
3.a	228 / 228	/	228 / 228	(100%)
3.b	218 / 218	/	218 / 218	(100%)
4.a	257 / 260	25 / 25	282 / 285	(99%)
4.b	255 / 258	31 / 31	286 / 289	(99%)
5	30 / 46	17 / 18	47 / 64	(73%)

For Test section 3.b, the full list of homographs to synthesise included three versions of each pair of homographs, and was split in three test versions for the evaluation, each containing one unique pair of each homograph. Only two of the test versions were evaluated.

Overall, only a relatively small subset of the test sets were actually used in the listening tests, leaving a large amount of synthetic speech material available to use in future listening tests. The detailed listening test results are distributed via the Blizzard Challenge archive in a package also including all submitted synthetic speech (Table 1).

4.4.4. Implementation

All tests were implemented with the Web Audio Evaluation Tool [25] and full details of the tasks and instructions given to participants are provided in Appendix 9.2. For all tests, the listeners were required to fill in one questionnaire at the beginning to provide information about themselves, and one questionnaire at the end to provide feedback on the test. Responses to these questionnaires are summarised in Tables 7 to 21.

4.4.5. Listeners

Similarly to previous years, listeners were recruited via the two following methods:

- *Paid listeners* via the crowdsourcing Prolific platform (see Table 1). Inclusion criteria were: self-certified French native speakers from any country of origin ; no self-reported hearing

problems. Participants were instructed to wear earphones or headphones for the test. All the test instructions for this group of listeners were given in French (see Appendix 9.2).

All five tests were independently submitted to Prolific: listeners could participate in several tests but not in several experimental blocks of the same test. For each experimental block, we recruited a minimum of 15 listeners for Tests 1 and 4 ; 30 listeners for Tests 2 and 5 ; 10 listeners for Test 3.a ; and four listeners for Test 3.b. The overall number of completed tests is given in Table 3. Listeners were compensated at a rate of 10£/ hour, with an estimated completion time of 24 min for Tests 1, 3 and 4 ; and 30 min for Tests 2 and 5. The actual median completion time is reported for each test in Table 2.

- *Online volunteers via mailing lists.* Inclusion criterion was: no self-reported hearing problems. Participants were required to wear earphones or headphones for the test. All the test instructions for this group of listeners were given in English (see Appendix 9.2).

  Since Test 3 required French proficiency which was not an inclusion criterion for this group of participants, only Tests 1, 2, 4 and 5 were submitted to volunteers as four independent URLs. Because some participants dropped the test, and they chose freely among the four URLs, we didn't control the number of online volunteers per experimental block, for each test. The overall number of completed tests is given in Table 3.

Following previous challenges, the organisers asked participating teams to help recruit volunteer listeners. Yet, the listening test completion rate by Blizzard participants is low, since participating teams reported a total of 85 team members and less than 40 online volunteers self-reported as speech experts for each test. One reason could be the shorter time frame given to online volunteers this year for listening test completion: Three weeks for Tests 1 and 4 ; One week for Tests 2 and 5.

We screened participants for Tests 1, 2, 4 and 5. For Tests 1 and 4 (MOS), we removed listeners that used only two or fewer levels from the 5-point scale across the whole test. Few listeners were in this case (between 1 and 3 for each test) and we did not run new experiments to replace the excluded listeners. For Tests 2 and 5 (MUSHRA), we removed listeners that rated the hidden natural speech reference high anchor less than 80 on average over the whole test. This removed about a quarter of the participants recruited via Prolific. Therefore we increased the number of listeners to attain 30 participants for each test after screening. The number of participants after screening over the total number of completed tests are reported in Table 3.

4.5. Analysis methodology

4.5.1. Score computation

For Tests 1, 4 (MOS), 2 and 5 (MUSHRA), we analysed the raw scores given by listeners. For Test section 3.a (intelligibility on SUS), Word Error Rate (WER) was calculated for each transcription. We allowed certain spelling variations in listener responses. In particular, homonyms were accepted, as listed in the results package provided on the Challenge archive (Table 1).

Compared to MOS and MUSHRA which are subjective judgements, Test section 3.b has an objective answer: whether the pronunciation of the homograph in a given sentence/context by a synthesiser is correct or incorrect. So in this test, listeners can be seen as annotators. A minimum of 4 raters were recruited per experimental block and we used the Fleiss' kappa test to obtain an inter-listener agreement value per block [26].

Table 4: *Summary of statistical tests performed on the outcomes of the five listening tests.*

Test	**1, 4**	**2, 5**	**3.a**	**3.b**
Score	MOS	MUSHRA	WER	Correct score
Data type	Ordinal	Proportion		Binary
Statistical model	Ordinal-	Beta- regression with random effects		Logistic-
R function	`clmm`	`glmmTMB`		`glmer`
R package	`ordinal`	`glmmTMB`		`lme4`
Post-hoc analysis	Estimated marginal means	Method from [28]		
R function	`emmeans`	`glht`		
R package	`emmeans`	`mutlcomp`		

We increased the number of raters per block until we reached at least a substantial agreement (0.6 on a [0-1] scale). Then, for each sentence and each system, we selected the homograph pronunciation that has been recognised by the majority of listeners, and compared it to the expected pronunciation. In this manner we obtained a binary correct/non-correct score for each sentence and system.

4.5.2. Statistical analysis

Previous challenges adopted the statistical analysis presented in [27]. In particular, when sufficient data was available, a Wilcoxon's signed rank test was applied between each pair of systems given the factor levels under investigation (e.g., between each pair of systems for speech experts and native listeners). There are two major drawbacks with this test:

- The high number of statistical tests that are performed artificially increases the chance of getting significant results. Of course, Bonferroni correction can be applied to compensate for this phenomenon, but this correction is too strong in a sense that it conversely decreases the chance of getting significant results.
- A Wilcoxon test compares pairs of distributions based on the ranking of the samples from both distributions. In tests like 5-point scale MOS, where the samples can take only five different values, there is a dramatic number of ties in the ranking of the samples, which limits the power of this statistical test.

For those reasons, we introduced a new statistical method composed of the following steps:

1) Selection of the factors under investigation: Our two main factors of interest are the *listener_type* (three levels, SE: speech experts (who self-identified as such), that were recruited either via Prolific or as online volunteers ; SP: paid participants (all native speakers of French), who took the test on Prolific and did not self-report as a speech expert ; and SR: volunteers, who took the test as online volunteers and did not self-report as a speech expert) and the *is_native* factor (two levels, native and non-native: listeners who self-identified as native (resp. non-native) speakers of French). The number of listeners for each factor and each test is given in Tables 7 and 8, respectively.

For each listening test, we first check the contingency table of the *listener_type* and *is_native* factors. If the table is full (i.e., there were listeners for all combinations of factor levels), we perform the statistical analysis on both factors. If the table is not full (i.e., there were missing listeners for some combinations of factor levels), we group or remove factors

until we obtain one full contingency table. One example of a grouping is the consideration of the *speech_expert* factor with 2 levels (SE: speech experts ; and N-SE = SP + SR: non-speech experts).

2) *Descriptive Statistics:* As in previous challenges, for each identified combination of factors, we output a descriptive statistics list that includes: median, median absolute deviation, mean, standard deviation, the number of data points used in the calculations, and the number of data points excluded due to missing data. Please note that all score distributions do NOT meet the normality requirements. For instance, most tests are carried out on an ordinal scale. Therefore, the mean and standard deviation values are not meaningful and should not be reported. In practice, we only used the mean value as a criterion to order the plots. The descriptive statistics for all tests are available in the results package.
3) *Statistical models:* For each test and identified combination of factors, we fit a statistical model whose type depends on the type of data (see Table 4). All statistical models also include the *sentence* and *listener_ID* as random factors.
4) *Assessing the significance of factors:* For each statistical model, the effect of individual factors and their interactions are tested by removing them one by one from the full statistical model, and assessing if the removal of each factor has a significant impact on the model. We start with random factors, then with the interactions between factors, and only if the latter were non-significant, we try to remove the factors involved in those interactions. A likelihood ratio test (`anova` function of R software) is used to assess the significance of each factor or interaction removal ($p < 0.01$).
5) *Multiple comparisons:* Once the statistical model is simplified, we perform multiple comparisons between levels of the remaining significant factors. The appropriate post-hoc analysis method depends on the data type (see Table 4).

   For the sake of comparison with previous challenge editions, we sometimes also report pairwise comparisons of systems with Wilcoxon's signed rank test using Bonferroni correction. In all cases, like in previous challenges, the multiple comparison output is presented in this paper as binary images where black squares indicate that two systems are significantly different ($p < 0.01$) given some factor levels. Statistics and p-values for each pairwise comparison are provided in the results package.

4.5.3. Identification of the systems for the MUSHRA tests

To refine the speech quality evaluation obtained with MOS, we submitted the systems with the best quality MOS to a MUSHRA test. To select the systems, we used the following method:

- On the MOS quality data, we fit a statistical model with the effect of system only, and compute multiple comparisons between systems, leading to a matrix of statistic values with an element for each pair of systems.
- We use the matrix of statistic values as a distance matrix to perform a hierarchical clustering of the systems. The latter can be represented as a tree (see Fig. 1 and 2), and allows us to cluster models that are given similar scores

To choose the number of clusters, our criteria is to get between 3 to 5 models in the cluster that includes the models with the highest MOS quality scores. We used 5 clusters for both FH1 and FS1. All the models in the cluster with the highest MOS quality scores were submitted to the MUSHRA test along with the ground truth and the BF benchmark (a total of five models for FH1 and six models for FS1).

Table 5: *Significance of the different factors and their interactions involved in Tests 1, 2, 4 and 5, according to the statistical models listed in Table 4, ($p < 0.01$). A dark background indicates when factors are not included in the model.*

Test	1.a	4.a	2	5	1.b	4.b
system	✓	✓	✓	✓	✓	✓
sentence (random)	✓	✓	✓	✓	✓	✓
listener_ID (random)	✓	✓	✓	✓	✓	✓
listener_type (SE, SP, SR)	✓				✓	
listener_type × *system*	✓				✓	
speech_expert (SE, N-SE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						

5. Results

Results of speech quality and speaker similarity evaluations (Tests 1, 2, 4 and 5) are analysed with respect to several factors listed in Section 4.5.2, and the significance of each factor and their interactions calculated with the appropriate statistical model (see Table 4) are summarised in Table 5. The significant impact of the *system* factor trivially shows that the submitted systems provide significantly different perceived outputs, which will be discussed in Section 5.1. The significance of the random *sentence* and *listener_ID* factors shows that for all tests, these factors explain a significant part of the variance in the results.

One important result is the significance of the *speech_expert* factor for all tests. This means that speech experts SE evaluated speech synthesis differently than non-speech experts N-SE. Moreover, this difference in behaviour also affects the relative ordering of systems for the MUSHRA Tests 2 and 5, given the significant interaction between the *speech_expert* and *system* factors for these tests. These results are further discussed in Section 5.2, but already demonstrate the importance of the listener's profile on the evaluation scores, especially in fine-grained tests such as MUSHRA. Inversely, the distinction between paid listeners SP and volunteers SR among non-speech experts is only significant for MOS tests on task FH1 (1.a and 1.b), given the significance of the *listener_type* factor. Similarly to the *speech_expert* factor, the *is_native* factor has a significant effect on most tests, showing that native listeners judged speech synthesis differently than non-native listeners. This had an effect on the relative ordering of the systems for Tests 1.a, 1.b, 5 and 4.b.

The remainder of this summary is organised as follows, Section 5.1 presents the listening test results per *system*, as in previous challenges. Then, Section 5.2 presents a further analysis of the results, by introducing the effects of the *listener_type*, *speech_expert*, and *is_native* factors. Section 5.3 summarises feedback we received from listeners. All results are presented using standard boxplots, except for the pronunciation accuracy (barplots). For mean opinion scores, the distribution of scores in the form of a stacked barplot is also reported, as it allows to better visualise the proportion of each score a system has been given. For each test, systems are presented in descending order of the average score calculated from the responses of all listen-

ers combined. 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 which systems are significantly better than others. By contrast, pairwise significance between systems is systematically reported as binary matrices, where a solid black box indicates a significant difference between the scores given to two systems ($p < 0.01$).

5.1. Global results, effect of the *system* only

5.1.1. FH1 - Quality

Figure 1 summarises all the results of the speech quality evaluation for task FH1. This year, two systems (F and I) were given scores that are not statistically different from natural speech (A). More generally, the hierarchical clustering (top-right of Fig. 1a) calculated from the multiple comparisons (Fig. 1b, left) highlights five groups:

1. The best rated systems include A, F, I, and O, O being not significantly different than F and I but significantly different from the natural voice A.
2. A second group includes M, P, Q, T, J, E, S and H that all received a median MOS of 4.
3. A third group includes D, C, K and L which received a median MOS of 3.
4. The fourth group includes R, N, and G, which also received a median MOS of 3 but with a higher dispersion.
5. The two benchmarks BF and BT are in the last group with a median MOS of 2. Thus all systems were judged significantly better than the benchmarks.

The systems of the first group (A, F, I, O) were subsequently submitted to a MUSHRA test for finer evaluation, along with the best benchmark system (BF), and results are shown in the bottom of Fig. 1. Interestingly, all systems were judged significantly different from natural speech. Moreover, System F which was not significantly different from I and O in the MOS test was judged significantly better than I and O in the MUSHRA test. The difference between I and O remains non-significant. To conclude, the MUSHRA test highlights the limits of the MOS test, which when including a large number of systems to rate does not finely discriminate between them. Conversely, the MUSHRA test allows a finer distinction between systems, but is limited in the number of systems that can be included in the test. Very importantly, what these results demonstrate is that if a MOS test does not show any significant differences between synthetic speech and natural speech, one should NOT conclude that the synthetic speech is 'as good as' or 'indistinguishable from' natural speech in general. Fine-grained comparisons between the systems such as MUSHRA (or other types of tests that focus on specific speech features) can highlight differences that a global MOS evaluation cannot.

5.1.2. FS1 - Quality

Figure 2 summarises all the results of the speech quality evaluation for task FS1. Recall that compared to FH1, only 2 hours of speech from the target speaker was provided in FS1. Firstly, the global range of scores got for this task is comparable if not slightly better than in FH1. We again grouped the systems that were given similar scores in clusters (top-right of Fig. 2a). The systems of the first group are again A, F, and O (I did not submit an entry for FS1), and F and O are not significantly different from natural speech. The second group with a median MOS of 4 includes L, Q, H, J, P, and T and the third group includes E and S that were given fewer scores of 5 than the systems in Group 2. The fourth group includes G, R, N, K and the benchmark system BF with median MOS between 2 and 3. BT is in the fifth group. Compared to FH1 task, System L was given much higher scores, while System K was given lower scores. Also BF became not statistically different from the systems in the fourth group. All other systems were given approximately a similar ranking and similar scores as in FH1. This shows that overall, the task of fine-tuning a TTS system on 2 hours of speech is performing well, and future challenges could reduce the size of the training data for the target speaker.

Since there were only two systems in the first group, we also included the two best systems of the second group according to the hierarchical clustering to a MUSHRA test for finer evaluation, along with the best benchmark system (BF), and the results are shown in the bottom of Fig. 2. Again, all systems were judged significantly different from natural speech, and System F which was not significantly different from O in the MOS test was judged significantly better than all the other systems in the MUSHRA test. Inversely, L and O were given significantly different MOS, but not significantly different MUSHRA scores. Q was rated significantly lower than all participants' systems. Other striking results are the lowest scores of the systems with a highest range of variation in the MUSHRA test compared to MOS. For instance, 90% of the MOS were 4 and 5 for System O (bottom-right of Fig. 2a) (last quarter of the 1-5 scale), while 50% of the MUSHRA scores where between 50% and 80% of the MUSHRA scale (Fig. 2c). This again demonstrates that MUSHRA allows participant to better discriminate between systems, and that MOS tests are not sufficient for fine-grained evaluation of speech synthesis.

5.1.3. FH1 and FS1 - Similarity

Figure 3 (resp. 4) summarises all the results of the speaker similarity evaluation for task FH1 (resp. FS1). One first surprising result is that for task FH1, Systems F, M, Q, J, and P are not statistically different from natural speech, with F and M that were given higher similarity scores than A. For task FS1, Systems Q, F, J and L were judged statistically closer to the reference speaker than the natural speech itself which got a median score of 4 (Probably the same person).

A probable reason for these results, that was reported by some listeners as well as some participants to the challenge, is that the four reference signals given in both tests sounded different from each other, although they belonged to the same speaker. This is because we selected samples that sounded the most different so the reference samples were representative of the speaker's voice range in the training data. The reported difficulty of the task to judge similarity of synthetic samples with such varied references is reflected in the low overall scores given for the task FH1: score 5 (Exactly the same person) was rarely given to both synthetic and natural speech.

Therefore, this raises the question of defining speaker similarity: do we want to assess the similarity between synthetic speech and references which are in the centre of the distribution of the speaker's voice range of variation, to which the syntheses might be close, but that is not representative of the speaker's full voice range? Or should we provide references that are representative of the speaker's full voice range, with wide timbre variations, and see how the synthesis can match the speaker's voice variability? In this evaluation, we chose the second option which, in our opinion, is more representative of an ecological speaker recognition task, but this raises a second question:

can we ask listeners who have never heard the voice of the reference speaker before if a sound sample could come from his/her voice?

The low scores given to natural speech in both tasks suggest that listeners couldn't create a mental representation of the speaker's full voice range given the few reference samples that they heard. We further tested this hypothesis on task FS1 by recruiting eight listeners who were familiar with the speaker's voice (family and friends). They reported hearing the speaker's voice either daily (one listener), weekly (two listeners), monthly (three listeners) or annually (two listeners), and their speaker similarity results are reported in Fig. 4c. Due to the low number of listeners, we did not perform any statistical analysis on this data. Nevertheless, we can observe that in this case, the natural voice was given a score 5 (Exactly the same person) more than 70% of the time. Only System F was given similar scores, that correlates well with its high score on speech quality. This innovative experiment demonstrates that:

- Only listeners that are familiar with the speaker's voice are able to correctly perform the speaker similarity task on the ground truth signal when the test references have a wide range of variation, representative of the speaker's voice.
- Listeners that are not familiar with the speaker's voice may only be able to perform a speaker similarity task where the reference given is in the centre of the distribution of the speaker's voice range of variation. Although this task is commonly performed in speech synthesis evaluations, this puts into question the validity of such a task, that is non-ecological and non-representative of the full speaker's voice range.

5.1.4. FH1 - Intelligibility (SUS)

Figure 5 summarises the results of the SUS intelligibility task. Note that 10 systems (J, F, P, O, Q, H, I, M, S, E) obtained a WER median of 0, meaning that at least half of the 20 sentences they synthesised were error-free. Results from these systems were not significantly different from each other. Systems N, C, T, BF, BT, K, R, and D got a median of approximately 15% which corresponds to approximately 1 erroneous word per sentence, since there were seven words per SUS on average. Overall, only 6 systems were significantly different from the two benchmarks: J, F, P, O, Q were significantly better, and G was significantly lower. The excellent score for almost all systems demonstrates that SUS synthesis is globally well handled by most systems, and that this test might have reached its limit to make a distinction between the global intelligibility of speech synthesis across systems.

5.1.5. FH1 - Intelligibility (Homographs)

If speech synthesis is becoming excellent when evaluated globally, the evaluation of the intelligibility of homographs is an attempt to specifically target the evaluation of synthesis on the low percentage of error that remains in the global evaluation. Figure 6 summarises the results of the homographs intelligibility task. Note that 50% accuracy corresponds to the case where both homographs of a pair are pronounced similarly (one is always right, the other is always wrong). Hence it is in practice the worst score that can be obtained globally. The right part of Fig. 6 shows the pronunciation accuracy per homograph and system, with one element of each of the 36 homograph pairs presented in alphabetical order from left to right, followed by their respective counterparts on the second side of the plot, also from left to right. This representation highlights a similar behaviour of the systems with the lowest pronunciation accuracy (I, BT, R, E, K, P, L, C, BF, S, N) and that are globally not significant from each other. They systematically tend to favour one pronunciation for each pair, reaching almost 100% accuracy on the left side of the plot (one element of each pair) vs. close to 0% accuracy on the right side of the plot (the second element of each pair). Inversely, the best rated systems that are also not significant between each other (J, M, Q, H, G, T) manage to handle both pronunciations. Interestingly, we can observe a step in performance between systems D and O which correlates well with the use of a Large Language Model [29] by the text encoder. Systems O, F, T, G, H, Q, M and J used one for letter-to-sound mapping while the other systems did not.

5.2. Results per factor

This section aims at exploring further the effects of the *listener_type*, *speech_expert* and *is_native* factors on the results. Although all plots are provided, only a few are described here, with the aim to initiate discussions.

5.2.1. Effect of listener_type

The *listener_type* factor only has an effect on MOS evaluations of task FH1 (Tests 1.a and 1.b, see Table 5), and Figure 7 (resp. 8) summarises all the results of the speech quality (resp. speaker similarity) evaluation for task FH1, per *listener_type*. First, looking at Fig. 7b and 8b, we can see that a low number of listeners for SE and SR reduces the significance between systems, with the Wilcoxon pairwise test providing less significant differences than the multiple comparison. Looking at the scores, we can observe different rating strategies by the three groups of participants (SE, SP and SR). For instance, system E was given better quality MOS by SE and SR compared to SP. For SR, only the quality MOS scores of Systems R, N, G and the two benchmarks were significantly worse than those of the other systems. Regarding speaker similarity, SE gave higher scores than SP suggesting that speech experts could be better at recognising a reference speaker with a high variance in the presented samples than non-speech experts. The system ordering by SR is quite different than SP and SE, although the small number of participants in SR could be responsible of the high dispersion in the results.

5.2.2. Effect of speech_expert and is_native

Figures 9 and 10 display the MUSHRA scores per *speech_expert* and *is_native* factors for tasks FH1 and FS1, respectively. Table 5 shows that there are significant interactions between the *speech_expert* and *system* factors for both FH1 and FS1 tasks, i.e. the speech experts SE and non-experts N-SE gave different scores to systems. Indeed, we can observe on Figures 9a and 10a that synthetic systems were given better scores by SE than N-SE, relatively to the ground truth (A) and the baseline (BF). Note that this difference does not affect the significance of the differences between systems, as shown by the identical binary images for SE and N-SE in both Figures 9c and 10c.

The *is_native* factor does not significantly impact the MUSHRA scores of task FH1 (see Table 5) and results of Fig. 9b are given indicatively. Although the latter displays large differences between native and non-native listeners, the non-significance of these differences might come from the low number of non-native listeners. Inversely, there is a significant interaction between the *is_native* and *system* factors on

the MUSHRA scores of task FS1 (see Table 5), illustrated on Fig. 10b. First, we observe that non-native listeners gave higher scores to the synthetic systems, relatively to the ground truth (A) and the baseline (BF). It suggests that understanding the language leads to a more severe judgement of the global quality. Moreover, pairwise differences between system showed in Fig. 10d indicate that non-native listeners perceived less significant differences between systems than native listeners.

The results per *speech_expert* and *is_native* factors for quality MOS on tasks FH1 and FS1, and similarity MOS on tasks FH1 and FS1 are presented in Figures 11, 12, 13 and 14, respectively. We observe similar behaviours as for the MUSHRA scores:

- The *speech_expert* factor has a significant but small effect on the results, with slightly better scores given by SE, but similar pairwise differences between systems for SE and N-SE.
- The *is_native* factor has a significant and important effect on the results, with lower scores given by non-native listeners, and much less pairwise differences perceived by non-native listeners compared to native listeners.

Overall, we demonstrated in this section that systems were judged differently according to the *listener_type*, *speech_expert* and *is_native* factors, the latter having the largest effect on the results. Therefore, this emphasises that great care must be taken in selecting listeners for perceptive tests, giving preference to native listeners of the synthesised language, even for global speech quality and speaker similarity evaluation.

5.3. Listeners feedback

On completing the evaluation, listeners were given the opportunity to tell us what they thought through an online feedback form. Feedback forms included many detailed comments and suggestions from all listener types. Listener information and feedback are summarised in Tables 7 to 21. Notably, Table 20 reports appreciations of task difficulty: as expected, similarity ratings (1.b and 4.b) and transcription of SUS (3.a) are judged rather difficult. Compilation of free answers in Table 21 confirms that similarity ratings face the problem of the comparison between varied references (several listeners even reported hearing different speakers) and more limited variations of synthetic renderings.

6. Conclusions

This year's challenge evaluates the synthesis of isolated sentences generated from read speech (audiobooks or extract of parliament) on two tasks. The Hub (resp. Spoke) task was to generate a voice from a 51-hour (resp. 2-hour) single female speaker dataset. 18 (resp. 14) text-to-speech synthesis systems were evaluated on the Hub (resp. Spoke) task. All systems used a deep neural network encoder-decoder architecture. 11 systems followed a FastSpeech-like or a non-attentive Tacotron-like design, and seven adopted a variational auto-encoder conditioned by text design. 15 systems used GANs for the training of the waveform generation process. Evaluation focused on speech quality (global and fine-grained), speaker similarity, and intelligibility (global and fine-grained).

Evaluation of speech quality demonstrated that if the best synthesis output are still perceived with lower quality than natural speech in fine-grained evaluation (MUSHRA with the systems obtaining the best MOS), some systems generate speech that is almost indistinguishable from natural speech in a global MOS evaluation. The validity of the evaluation protocol for speaker similarity have been lengthily discussed in this paper. Yet, it showed that best generated speech samples are often perceived as exactly the same person as natural speech. This has been observed from both Hub and Spoke tasks. Therefore, future Blizzard Challenges can safely introduce more challenging tasks for speaker adaptation with fewer training data, such as synthesis from a few minutes dataset or even zero-shot synthesis. Finally, evaluation of intelligibility on SUS has demonstrated excellent results with median error rates of 0 for half of the systems. Finer evaluation of homographs displayed less successful results, but the use of large language models is promising as it allowed some systems to reach more than 80% accuracy on the task.

Overall, this challenge has demonstrated that current architectures are now becoming very competitive for the synthesis of high-quality isolated sentences in terms of speech quality, speaker similarity and intelligibility. The evaluation of speech synthesis in context (cf. example of applications in Table 1 in [20]) that has already been discussed and sometimes performed in the most recent literature could be introduced in future Blizzard Challenge editions, has it will certainly receive some of the broadest attention in the years to come.

7. Acknowledgements

This research has received funding from the BPI project THERADIA and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). We wish to thank a number of additional contributors without whom running the challenge would not be possible. Sébastien Le Maguer (Trinity College Dublin) helped to normalise the submitted data. Martin Lenglet (Univ. Grenoble Alpes, CNRS, GIPSA-lab), helped with web development for the listening test. We thank Aurélie Derbier for sharing her voice for the challenge and Romain Legrand and Frederic Elisei (Univ. Grenoble Alpes, CNRS, GIPSA-lab) for the recording of her voice. Damien Lolive (Univ. Rennes, CNRS, IRISA, France) and Nicolas Obin (IRCAM, Sorbonne University, CNRS) advised in the design of the challenge tasks. Simon King (Papercup / Univ. of Edinburgh) advised on the challenge organisation and task design. Finally, we thank the numerous listeners who participated in the evaluation, and the 18 participating teams without which the challenge wouldn't exist.

8. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proceedings of Interspeech*, Lisbon, Portugal, September 4-8 2005, pp. 77–80. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2005/black05_interspeech.html

[2] S. King and K. Vasilis, "The Blizzard Challenge 2013," in *Evaluation of text-to-speech systems*, Barcelona, Spain, September 3 2013. [Online]. Available: http://www.festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf

[3] Z.-H. Ling, X. Zhou, and S. King, "The Blizzard Challenge 2021," in *The Blizzard Challenge*, Online, October 23 2021. [Online]. Available: http://festvox.org/blizzard/bc2021/BC21_ling_zhou_king.pdf

[4] I. Solak, "The M-AILABS speech dataset," https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.

[5] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, vol. 28, no. 1, pp. 7–8, 2023/05/16 2014. [Online]. Available: https://doi.org/10.1108/RR-08-2013-0197

[6] "The Gutemberg Project," https://www.gutenberg.org.

[7] M. Lenglet, O. Perrotin, and G. Bailly, "Impact of Segmentation and Annotation in French end-to-end Synthesis," in *ISCA Speech Synthesis Workshop*, Budapest, Hungary, August 26-29 2021, pp. 13–18. [Online]. Available: https://www.isca-speech.org/archive/ssw_2021/lenglet21_ssw.html

[8] Merriam-Webster. Merriam-webster.com dictionary. [Online]. Available: https://www.merriam-webster.com/dictionary/homograph

[9] M.-L. Hajj, M. Lenglet, O. Perrotin, and G. Bailly, "Comparing nlp solutions for the disambiguation of french heterophonic homographs for end-to-end tts systems," in *Speech and Computer*, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. Cham: Springer International Publishing, 2022, pp. 265–278. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-20980-2_23

[10] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939600026X

[11] D. B. Pisoni, L. M. Manous, and M. J. Dedina, "Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility," *Computer Speech & Language*, vol. 2, no. 3, pp. 303–320, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0885230887900143

[12] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The siwis french speech synthesis database – design and recording of a high quality french database for speech synthesis," Idiap-Internal-RR-06-2017, Tech. Rep., 2017. [Online]. Available: https://doi.org/10.7488/ds/1705

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, April 15-20 2018, pp. 4779–4783.

[14] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 6-12 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

[15] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations (ICLR)*, Virtual, May 3-7 2021. [Online]. Available: https://openreview.net/forum?id=Afb6Nwd6LNEz

[16] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," in *ISCA Speech Synthesis Workshop*, Grenoble, France, August 26-28 2023, pp. 41–47. [Online]. Available: https://www.isca-speech.org/archive/ssw_2023/kirkland23_ssw.html

[17] C.-H. Chiang, W.-P. Huang, and H.-y. Lee, "Why we should report the details in subjective evaluation of tts more rigorously," in *Proceedings of Interspeech*, Dublin, Ireland, August 20-24 2023, pp. 5551–5555. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2023/chiang23_interspeech.html

[18] E. Cooper and J. Yamagishi, "Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech," in *Proceedings of Interspeech*, Dublin, Ireland, August 20-24 2023, pp. 1104–1108. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2023/cooper23_interspeech.html

[19] B. Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 57–71, 2003. [Online]. Available: https://doi.org/10.1023/A:1021052023237

[20] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander, and J. Voße, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, Vienna, Austria, September 20-22 2019, pp. 105–110. [Online]. Available: https://www.isca-speech.org/archive/ssw_2019/wagner19_ssw.html

[21] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs," in *ISCA Speech Synthesis Workshop*, Vienna, Austria, September 20-22 2019, pp. 99–104. [Online]. Available: https://isca-speech.org/archive_v0/SSW_2019/abstracts/SSW10_O_3-1.html

[22] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *ISCA Speech Synthesis Workshop*, Budapest, Hungary, August 26-28 2021, pp. 148–153. [Online]. Available: https://www.isca-speech.org/archive/ssw_2021/omahony21_ssw.html

[23] International Telecommunication Union, "Software tools and audio coding standardization," International Telecommunication Union, Tech. Rep., 2000. [Online]. Available: https://www.itu.int/rec/T-REC-P.56-201112-I/en

[24] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Evaluation of text-to-speech systems*, Bonn, Germany, August 25 2007. [Online]. Available: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_001.html

[25] M. D. Jillings N., De Man B. and R. J. D., "Web Audio Evaluation Tool: A Browser-Based Listening Test Environment," in *Sound and Music Computing Conference (SMC)*, 2015. [Online]. Available: https://github.com/BrechtDeMan/WebAudioEvaluationTool

[26] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[27] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Evaluation of text-to-speech systems*, Bonn, Germany, August 2007. [Online]. Available: https://www.isca-speech.org/archive/blizzard_2007/clark07_blizzard.html

[28] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous Inference in General Parametric Models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200810425

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

9. Appendices

9.1. Summary of submitted systems

Table 6: *The participating systems and their short names. The first three rows are the benchmarks and correspond to the system identifiers A, BF and BT in that order. The remaining rows are in alphabetical order according to the systems' short names and not the systems' identifiers. The method descriptions are summarised based on the questionnaires and the workshop papers from participants. When the vocoder is between parentheses, it has been trained end-to-end with the acoustic model. L2S and LLM stand for Letter-to-Sound module and Large Language Models, respectively.*

Short name	Team	L2S	Prosody control (inference)	Acoustic model	Vocoder	LLM
A (reference)	Natural speech from the same speaker as the corpus					
BF	FastSpeech benchmark	eSpeak	Variance predictors from text	FastSpeech2	HiFi-GAN	
BT	Tacotron benchmark	/	/	Tacotron2	HiFi-GAN	
10AI (Xpress)	Beijing Yiling Intelligence Technology Co., Ltd.	/	Prosody predictor (Flow) from text	Flow-VAE	BigVGAN	
AudioLabs	International Audio Laboratories Erlangen	Lexicons + eSpeak	Variance predictors from text	ForwardTacotron / FastTacotron	StyleMelGAN	
BIGAI	Beijing Institute of General Artificial Intelligence	eSpeak + pBART	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
CASIA Speech (VIBVG)	Institute of Automation, Chinese Academy of Sciences	eSpeak	Prosody predictor (Flow) from text	VITS	(BigVGAN)	
DeepZen	DeepZen Ltd.	Lexicons + FlauBERT (POS)	Prosody predictor (GST/LST) from FlauBERT	Non-attentive Tacotron	HiFi-GAN-based	✓
FireRedTTS	Xiaohongshu Inc.	Lexicons + CamemBERT (POS, DEP)	Prosody predictor (RNN) from text Rhythmic rules predictor from POS, NER, DEP	Non-Attentive Tacotron	HiFi++	✓
Fruit shell 2023	University of Chinese Academy of Sciences	eSpeak	Prosody predictor (Flow) from text	VITS	(HiFi-GAN)	
GIPSA-lab	Univ. Grenoble Alpes, CNRS, Grenoble INP, France	Phonetic prediction task in encoder	Variance predictors from text	FastSpeech2-based	WaveGlow	
Idiap	Idiap Research Institute, Martigny, Switzerland	eSpeak + CamemBERT (POS)	Variance predictors from text	Diffusion Transformer	FastDiff	✓
IMS (Toucan)	Institute for Natural Language Processing University of Stuttgart, Germany	eSpeak + CamemBERT (POS)	Prosody predictor (GST) as input Variance predictors from text+GST	FastSpeech2-based with conformers	BigVGAN	✓
IOA-ThinkIT	Institute of Acoustics, Chinese Academy of Sciences	Own L2S + BERT (word embedding)	Prosody predictor (H-VAE) from text	Hierarchical VAE		✓
La Forge	Ubisoft	eSpeak + CamemBERT (POS)	Prosody predictor (VAE) from text	VAE-Tacotron	HiFi-GAN	✓
LIUM-TTS	Laboratoire d'Informatique Le Mans Université (LIUM)	Data-driven L2S	Variance predictors from text	FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	WaveGlow	
MuLanTTS	Microsoft	Own L2S + BERT (liaisons and homographs)	Prosody predictor (GST) from text Variance predictors from text	FastSpeech2-based with conformers	HiFi-GAN	✓
Samsung TTS	Samsung Electronics HQ and Samsung Research China, Beijing	CART + CamemBERT (breaks, liaisons, POS) + ChatGPT (some homographs)	Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	FastSpeech2-based with conformers	HiFi-GAN	✓
SCUT SCSE	South China University of Technology	eSpeak	Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	FastSpeech2-based	HiFi-GAN	✓
TTS-Cube	Adobe Systems, SCC	Data-driven L2S	Variance predictors from text + CamemBERT	RNN-based	(HiFi-GAN)	✓
Xiaomi-ASLP	Xiaomi AI Lab and Audio Speech and Language Processing Group (ASLP@NPU) Northwestern Polytechnical University	eSpeak	Prosody predictor (Flow) from text + GPT-3	VITS	(HiFi-GAN)	✓

9.2. Instructions to the participants of the listening tests

9.2.1. MOS Quality

At the beginning of the section, listeners had to listen one sentence synthesised by 10 different systems to get familiar with the range of variation of the synthesis. This is to encourage listeners to use the full rating scale. Then, for each panel, listeners listened to one audio sample at a time and were asked to choose a score from a scale, following the instruction:

Instruction (EN): Please evaluate the quality of the audio.

Instruction (FR): Veuillez évaluer la qualité de la synthèse.

Scale (EN |FR):

1.	Very Poor	Très mauvaise
2.	Poor	Mauvaise
3.	Fair	Passable
4.	Good	Bonne
5.	Excellent	Excellente

The text content of the sentence was displayed on the screen. Listeners had to listen to the audio sample entirely at least once to be able to go to the next panel.

9.2.2. MOS Similarity

At the beginning of the section, and then every seven stimuli, listeners had to listen four reference samples of the original speaker. Then, for each panel, listeners listened to one audio sample at a time and were asked to choose a score from a scale, following the instruction:

Instruction (EN): Please evaluate the similarity between the reference speaker and the voice in the present audio.

Instruction (FR): Veuillez évaluer la similarité entre la locutrice de l'extrait audio présenté, et la locutrice de référence.

Scale (EN |FR):

1.	Completely different person	Personne totalement différente
2.	Probably a different person	Personne probablement différente
3.	Similar	Proche
4.	Probably the same person	Probablement la même personne
5.	Exactly the same person	Exactement la même personne

During each stimuli evaluation, the four reference samples of the original speaker were available to listen freely. The text content of the sentence was NOT displayed on the screen. Listeners had to listen to the audio sample entirely at least once to be able to go to the next panel.

9.2.3. MUSHRA Quality

For each panel, listeners listened to one explicit reference of the original speaker, and five (Test 2) or six (Test 5) non-identified audio samples among which there were one hidden reference (the same audio file than the explicit reference), one baseline, and three or four participants' systems presented in a random order. All audio samples of one panel played the same sentence. Listeners were asked to rate the non-identified audio samples on a continuous scale from 0 to 100, with the following instructions and graduations:

Instructions (EN): Please evaluate the quality of speech synthesis:

1. Listen to the reference audio.
2. Listen to the other audio clips and rate them relative to one another using the rating scales.
3. Once you rated all [5/6] audios, click on the sort button to place your ratings in order.
4. Re-listen to the audios from worst to best (left to right) and refine your ratings.
5. You may re-order, re-listen and refine your ratings as many times as you like.

It is required to perform steps 1 to 4 to go to the next audio sample.

Instructions (FR): Veuillez évaluer la qualité de la synthèse de parole :

1. Ecoutez l'audio de référence.
2. Ecoutez les autres extraits audio et notez-les relativement aux autres en utilisant toute l'échelle de notation.
3. Une fois notés, cliquez sur "Ordonner" pour ordonner les extraits audios dans l'ordre croissant des notes que vous leurs avez attribuées.
4. Réécoutez chaque extrait dans l'ordre (de gauche à droite) et affinez votre jugement.
5. Vous pouvez réordonner les extraits, les réécouter et ajuster leurs notes autant de fois que vous le souhaitez.

Il est nécessaire de suivre les étapes 1-4 pour pouvoir passer à l'extrait suivant.

Scale:

0:	Very poor	Très mauvais
25:	Poor	Mauvais
50:	Fair	Passable
75:	Good	Bon
100:	Excellent	Excellent

As an indirect way to enforce these instructions, listeners had to listen to the reference entirely at least once and to the samples to rate entirely at least twice to be able to go to the next panel. The text content of the sentence was NOT displayed on the screen.

9.2.4. SUS Intelligibility

For each panel, listeners listened to one audio sample (one utterance) at a time and were asked to transcribe the words that they heard according to the spelling rules of French, following the instruction:

Instruction (FR): Transcrivez ci-dessous les mots entendus, selon les règles orthographique du Français.

Listeners were allowed to listen to each sentence only once.

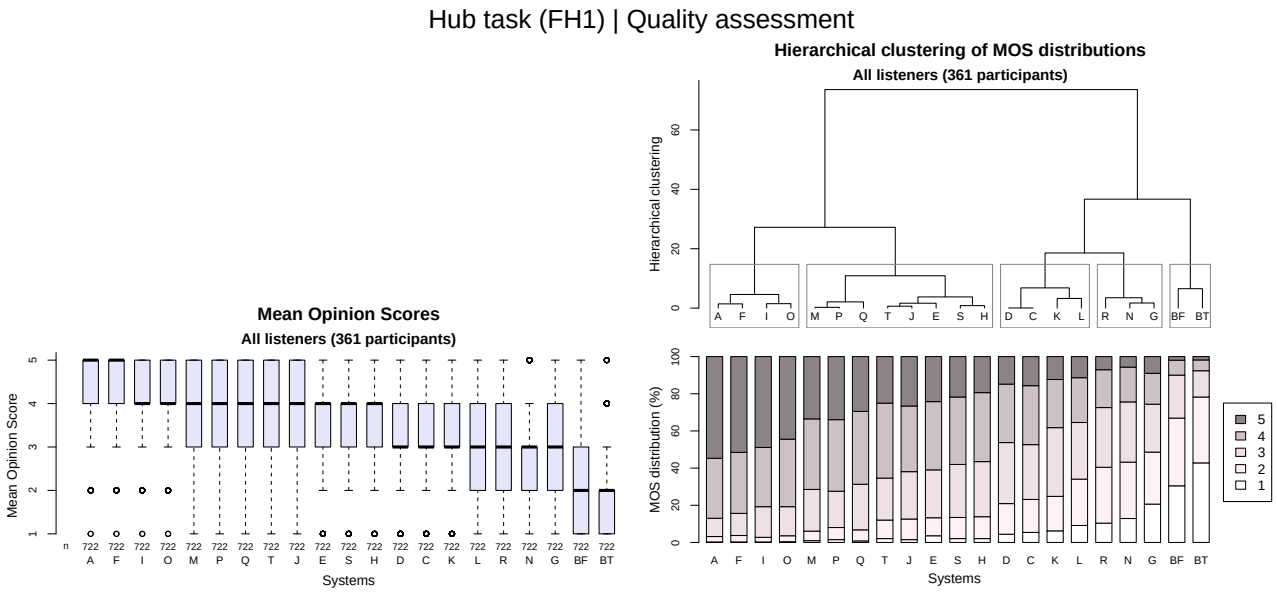
9.2.5. Homographs Intelligibility

For each panel, listeners listened to three audio samples. One audio sample was the synthesis of an utterance that contained a homograph. The text content of the sentence was displayed on the screen and the homograph was written in capital letters. The two other audio samples were the two versions of the homograph as an isolated word, uttered by a reference speaker (one of the authors of this paper, different from the voice to synthesise). Listeners were asked to select the reference audio that corresponded the best to the pronunciation of the homograph in the synthesis, regardless of the correctness of the pronunciation:

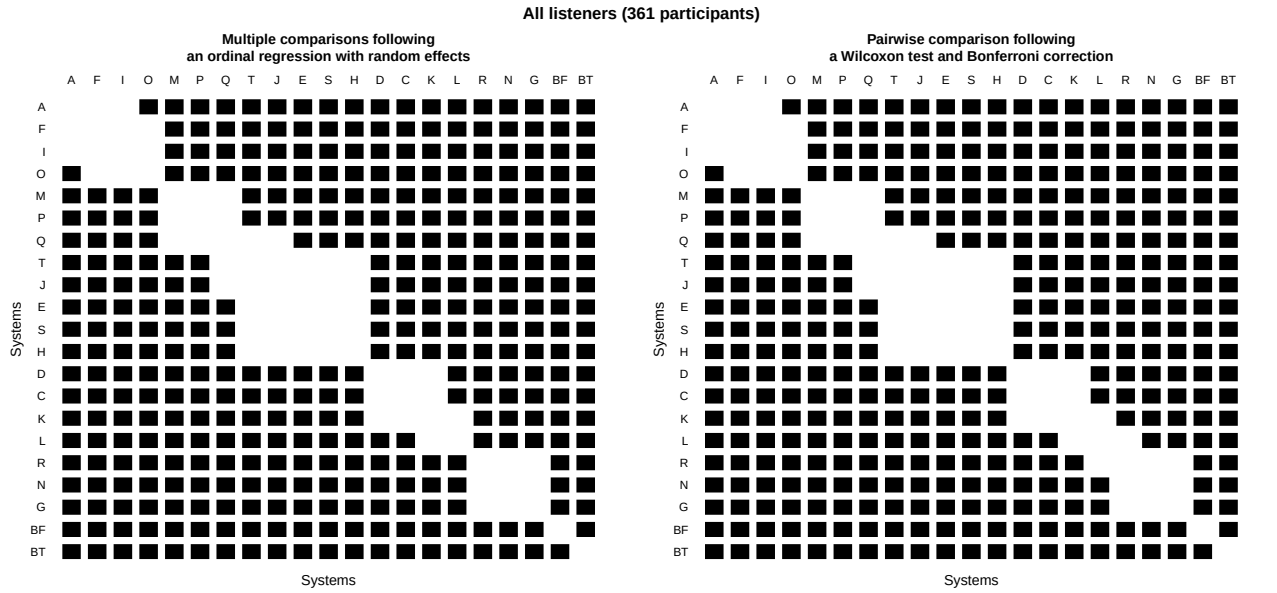
Instruction (FR): Sélectionnez l'extrait audio (en cliquant sur A ou B) dont la prononciation du mot ressemble le plus à celle du mot en majuscule dans la phrase à évaluer. Fondez votre réponse sur la prononciation du mot uniquement, et indépendamment de la grammaire de la phrase.

Listeners had to listen to the three audio samples entirely at least once to be able to go to the next panel.

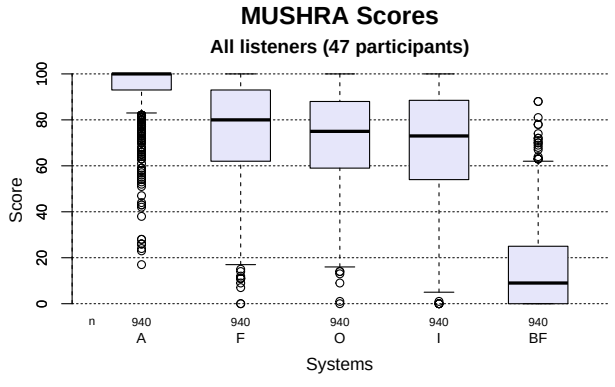
9.3. Plots and questionnaires



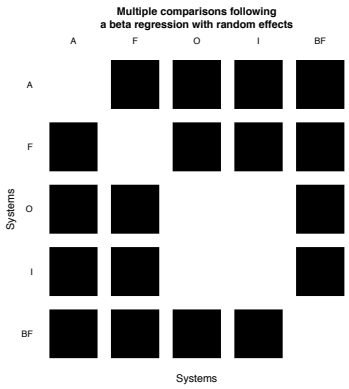
(a) *Speech quality mean opinion scores (left) and hierarchical clustering of systems (top-right) based on the MOS distributions (bottom-right).*



(b) *Significant differences in speech quality MOS between systems, indicated by solid black boxes* ($p < 0.01$). *Left: with multiple comparisons (Ordinal regression) ; Right: with pairwise comparisons (Wilcoxon)*

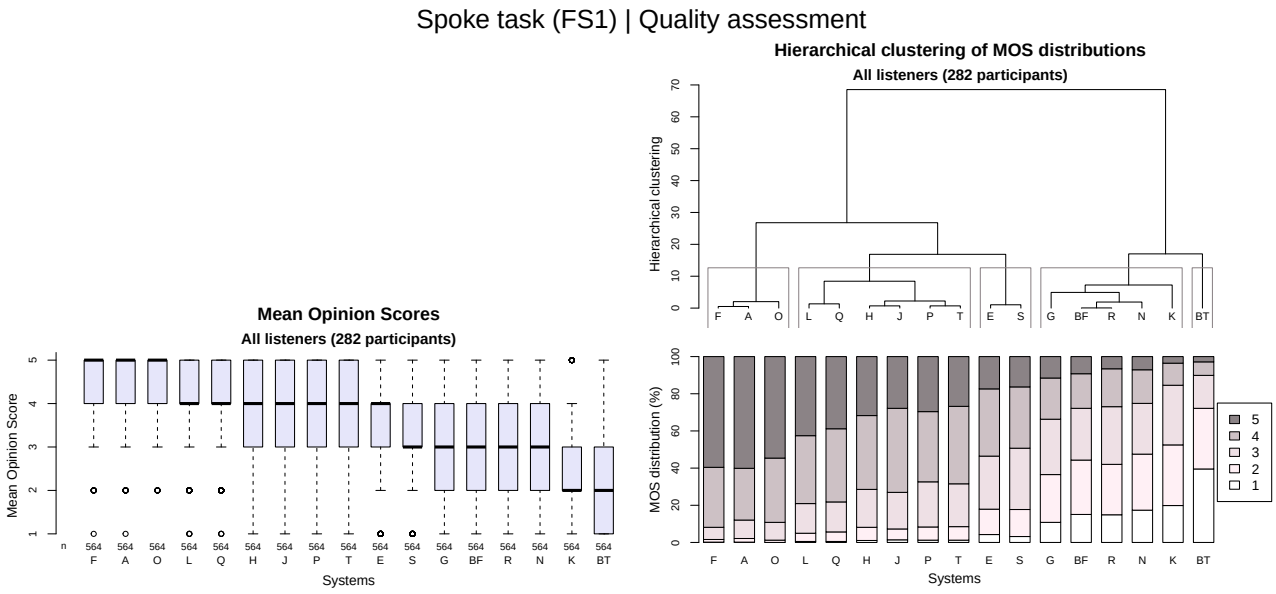


(c) *Speech quality MUSHRA scores.*

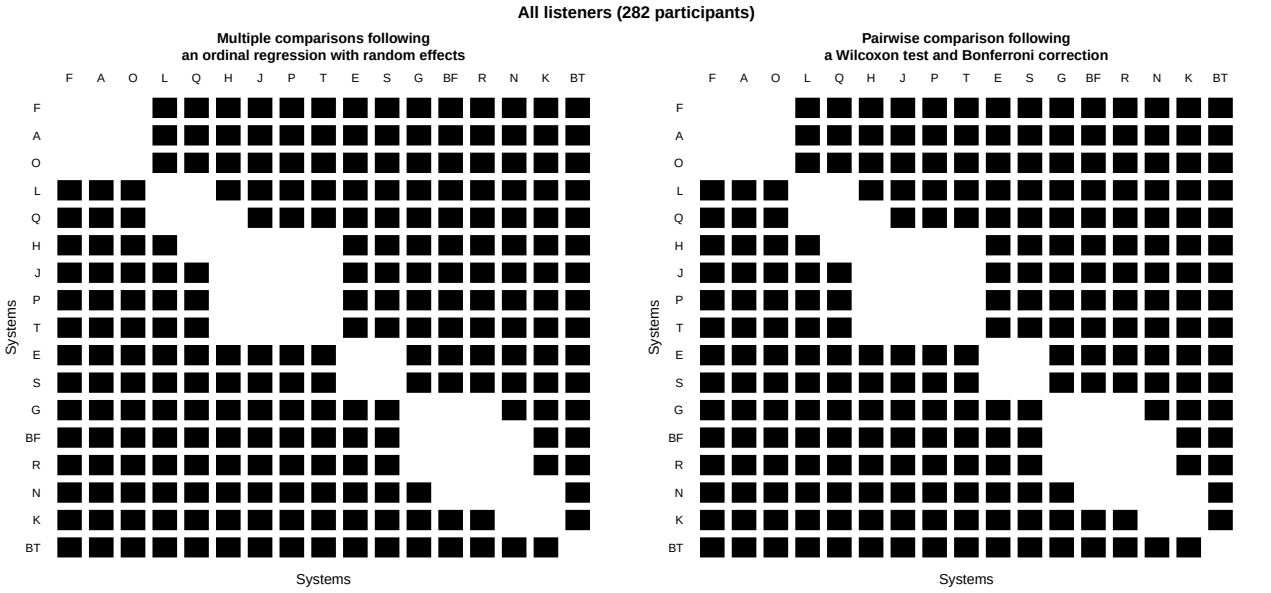


(d) *Significant differences in speech quality MUSHRA scores between systems, indicated by solid black boxes* ($p < 0.01$)

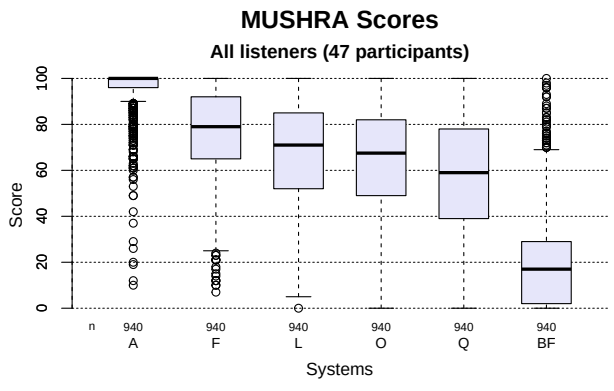
Figure 1: *Speech quality results for FH1, with MOS (Test 1.a) and MUSHRA (Test 2) evaluations, per system.*



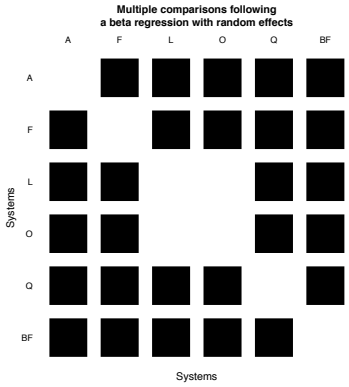
(a) *Speech quality mean opinion scores (left) and hierarchical clustering of systems (top-right) based on the MOS distributions (bottom-right).*



(b) *Significant differences in speech quality MOS between systems, indicated by solid black boxes* ($p < 0.01$). *Left: with multiple comparisons (Ordinal regression) ; Right: with pairwise comparisons (Wilcoxon)*

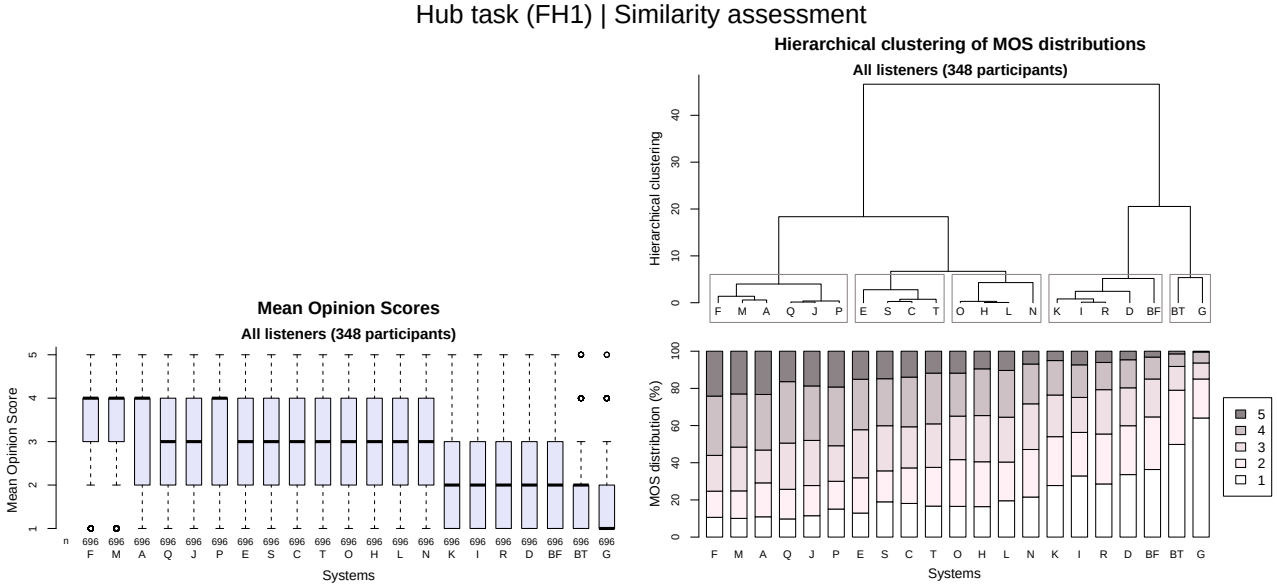


(c) *Speech quality MUSHRA scores.*

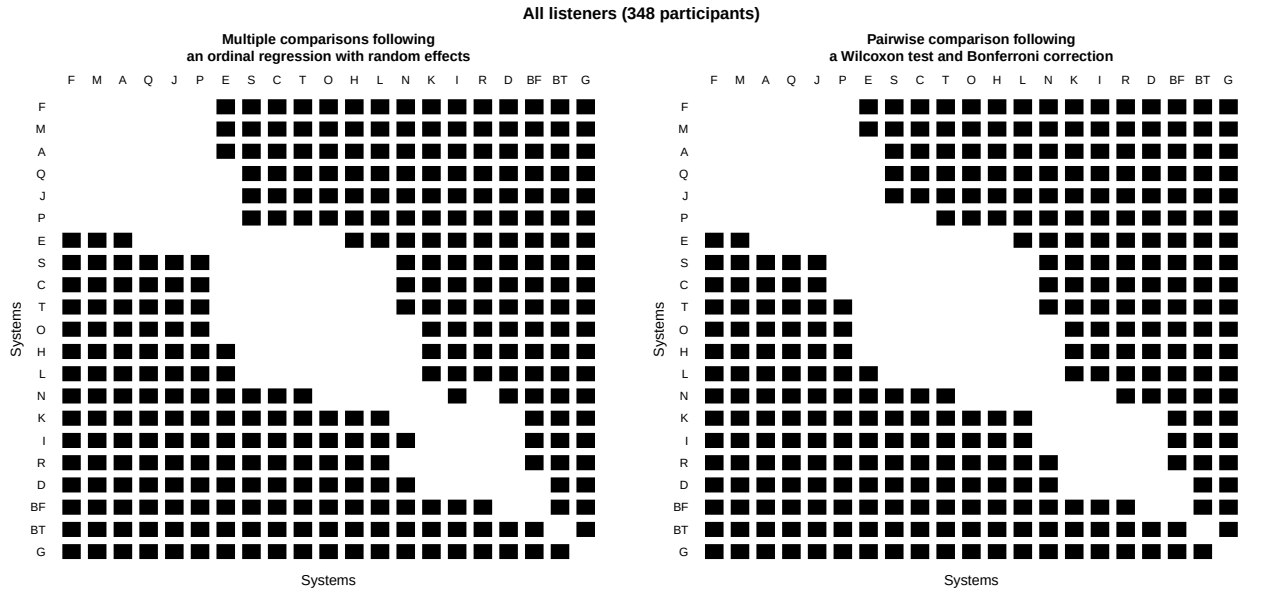


(d) *Significant differences in speech quality MUSHRA scores between systems, indicated by solid black boxes* ($p < 0.01$)

Figure 2: *Speech quality results for FS1, with MOS (Test 4.a) and MUSHRA (Test 5) evaluations, per system.*

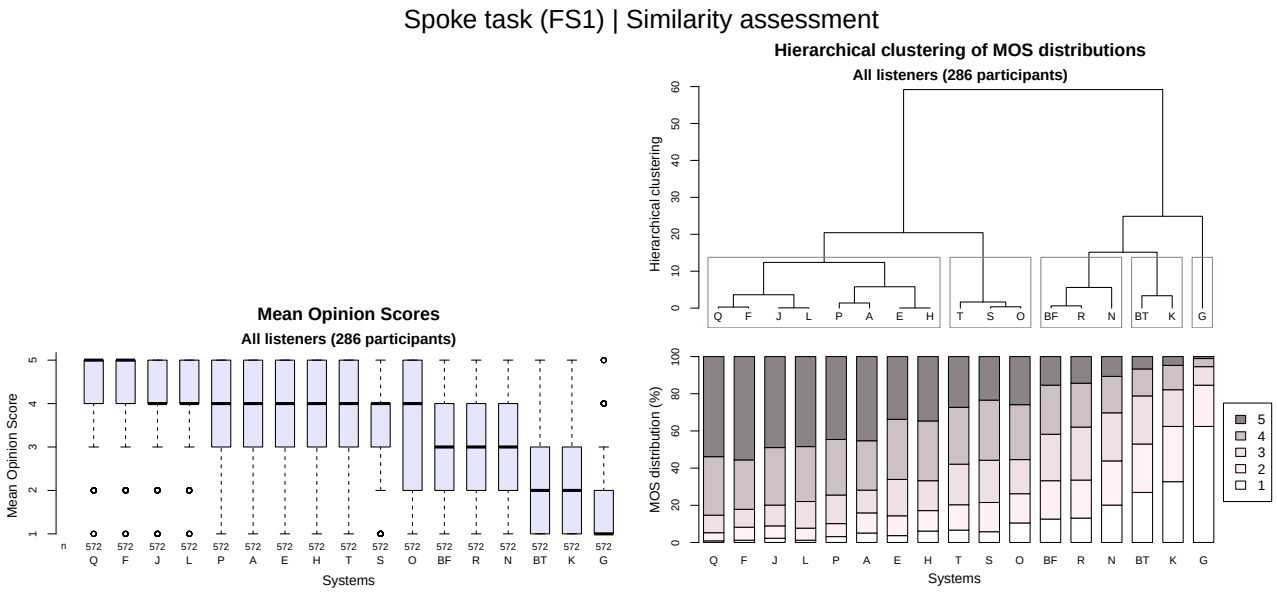


(a) *Speaker similarity mean opinion scores (left) and hierarchical clustering of systems (top-right) based on the MOS distributions (bottom-right).*

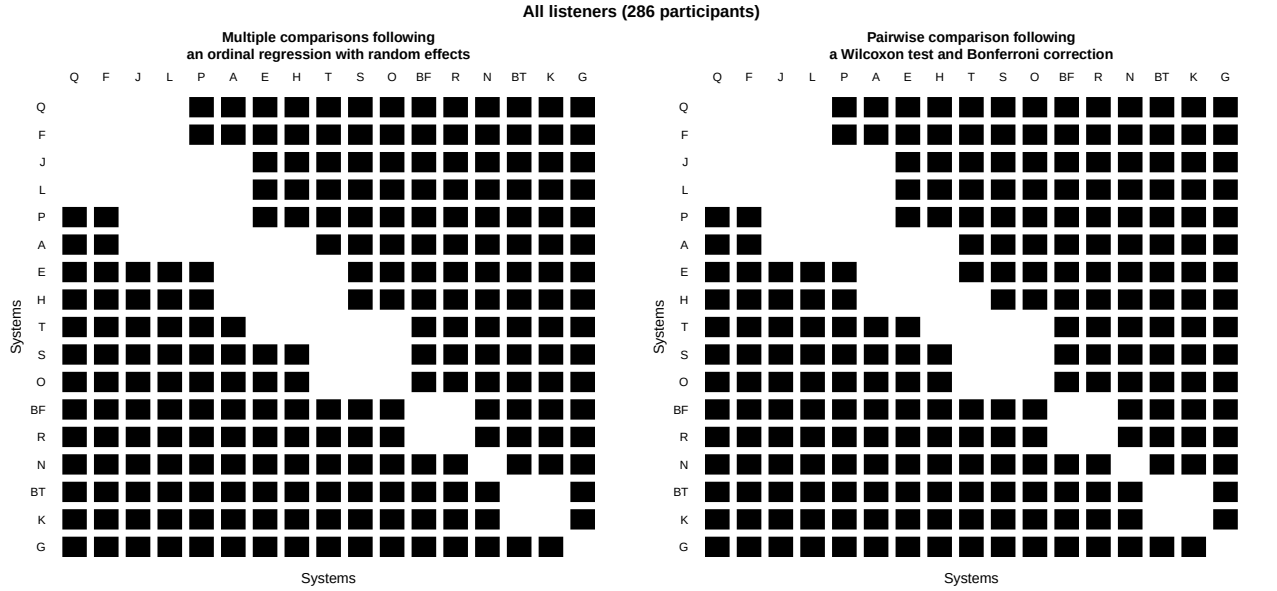


(b) *Significant differences in speaker similarity MOS between systems, indicated by solid black boxes* ($p < 0.01$). *Left: with multiple comparisons (Ordinal regression) ; Right: with pairwise comparisons (Wilcoxon)*

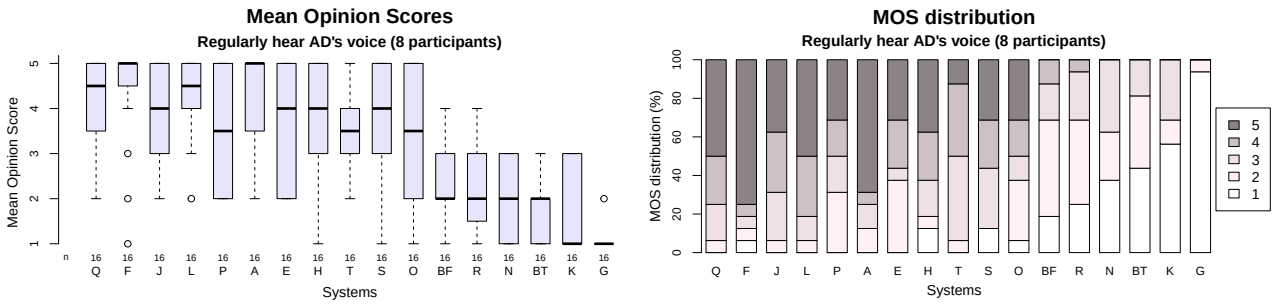
Figure 3: *Speaker similarity results for FH1, with MOS evaluation (Test 1.b), per system.*



(a) *Speaker similarity mean opinion scores (left) and hierarchical clustering of systems (top-right) based on the MOS distributions (bottom-right).*

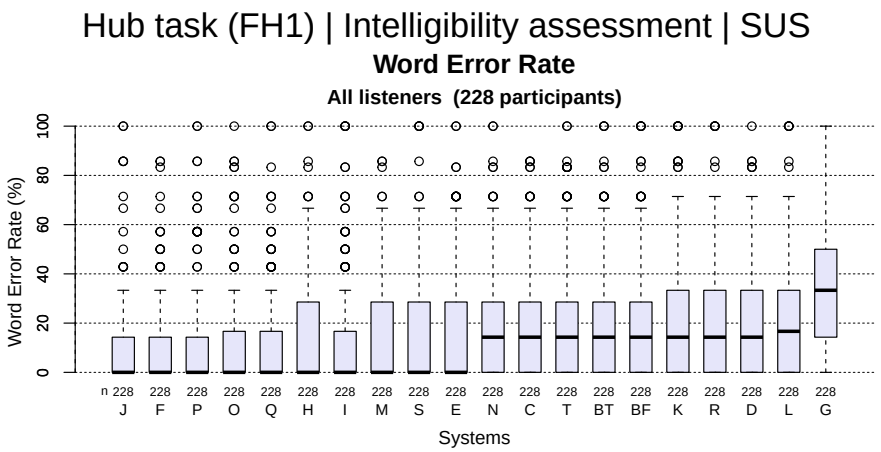


(b) *Significant differences in speaker similarity MOS between systems, indicated by solid black boxes* ($p < 0.01$). *Left: with multiple comparisons (Ordinal regression) ; Right: with pairwise comparisons (Wilcoxon)*

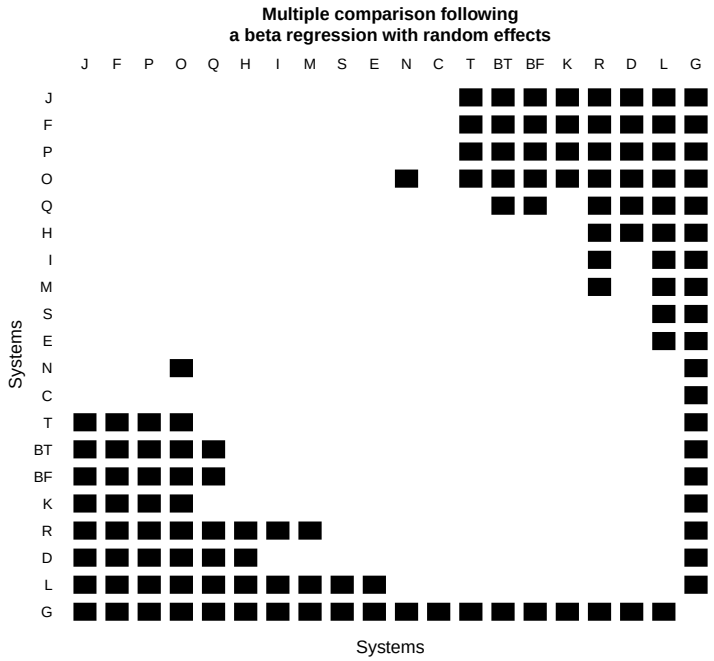


(c) *Speaker similarity MOS (left) and their distributions (right) per system, evaluated by listeners that are familiar with the speaker's voice. Results are presented with the same systems ordering than Fig. 4a.*

Figure 4: *Speaker similarity results for FS1, with MOS evaluation (Test 4.b), per system.*

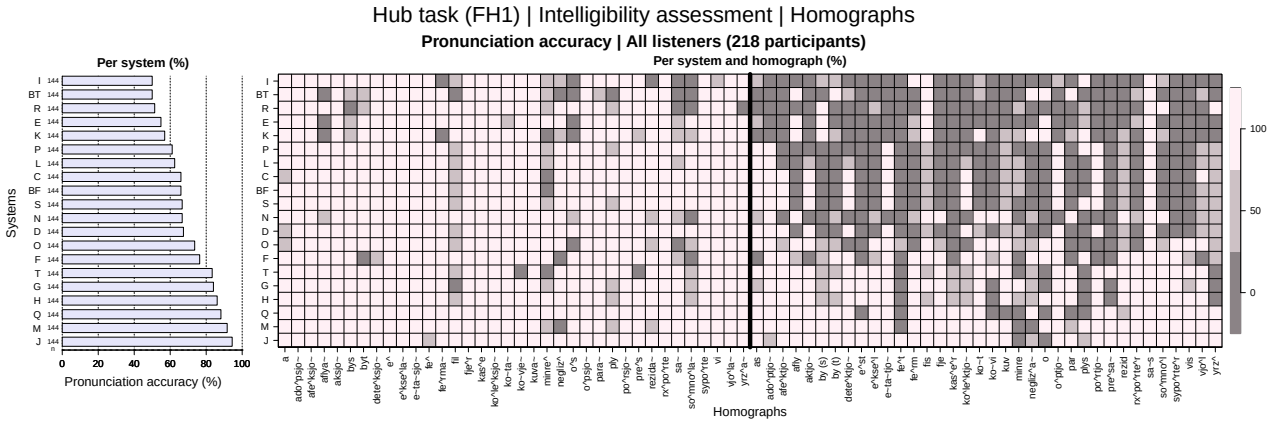


(a) *Word error rate on SUS transcription, per system.*

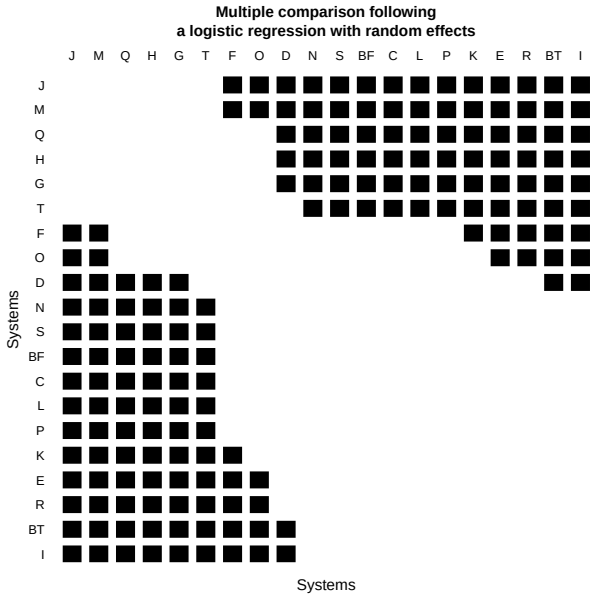


(b) *Significant differences in SUS WER between systems, indicated by solid black boxes* ($p < 0.01$).

Figure 5: *SUS intelligibility results for FH1, with WER evaluation on transcriptions (Test 3.a), per system.*

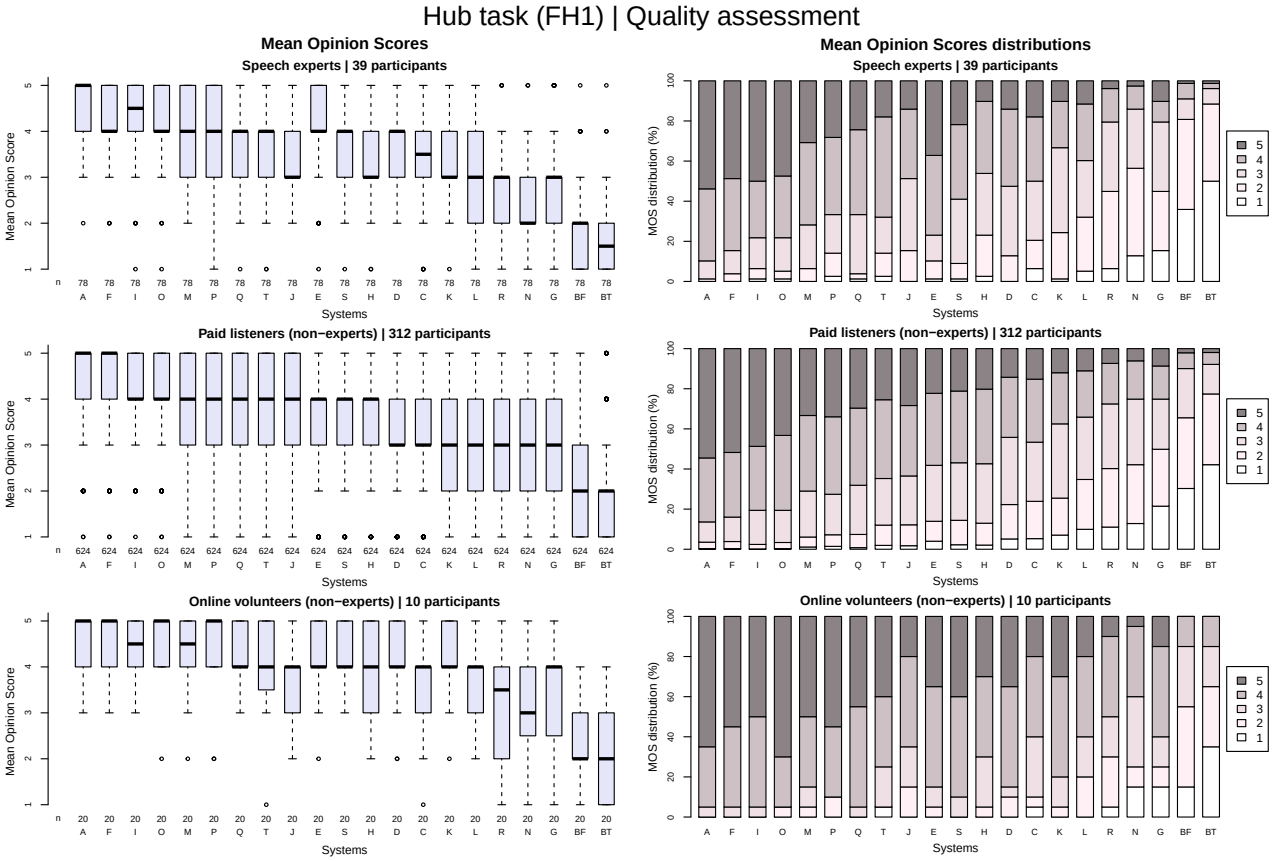


(a) *Pronunciation accuracy per system (left) and per system and homograph (right). On the right panel, one element (resp. the second element) of each pair of homograph is displayed on the left (resp. on the right) of the black line, with the same pair ordering.*

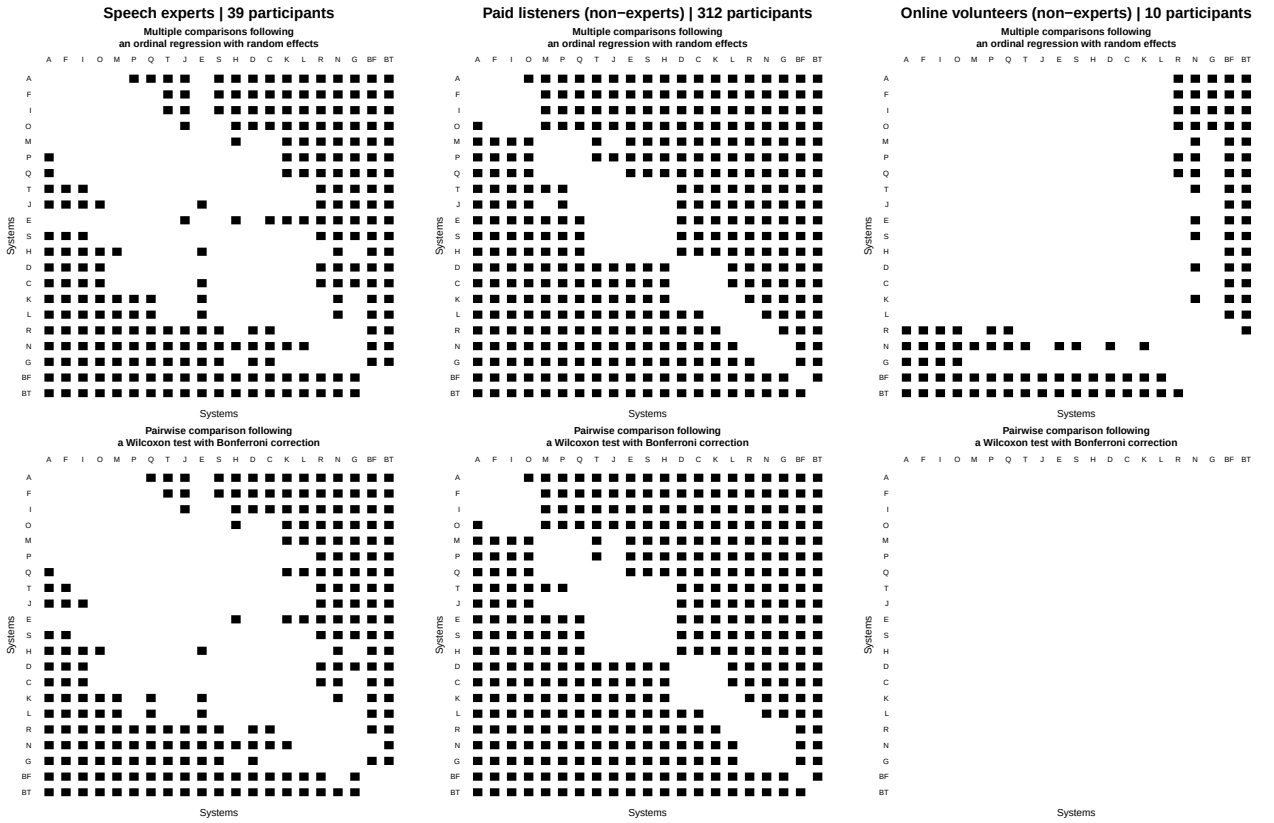


(b) *Significant differences in homograph pronunciation accuracy, indicated by solid black boxes* ($p < 0.01$).

Figure 6: *Homographs intelligibility results for FH1, with pronunciation accuracy evaluation (Test 3.b), per system and homographs.*

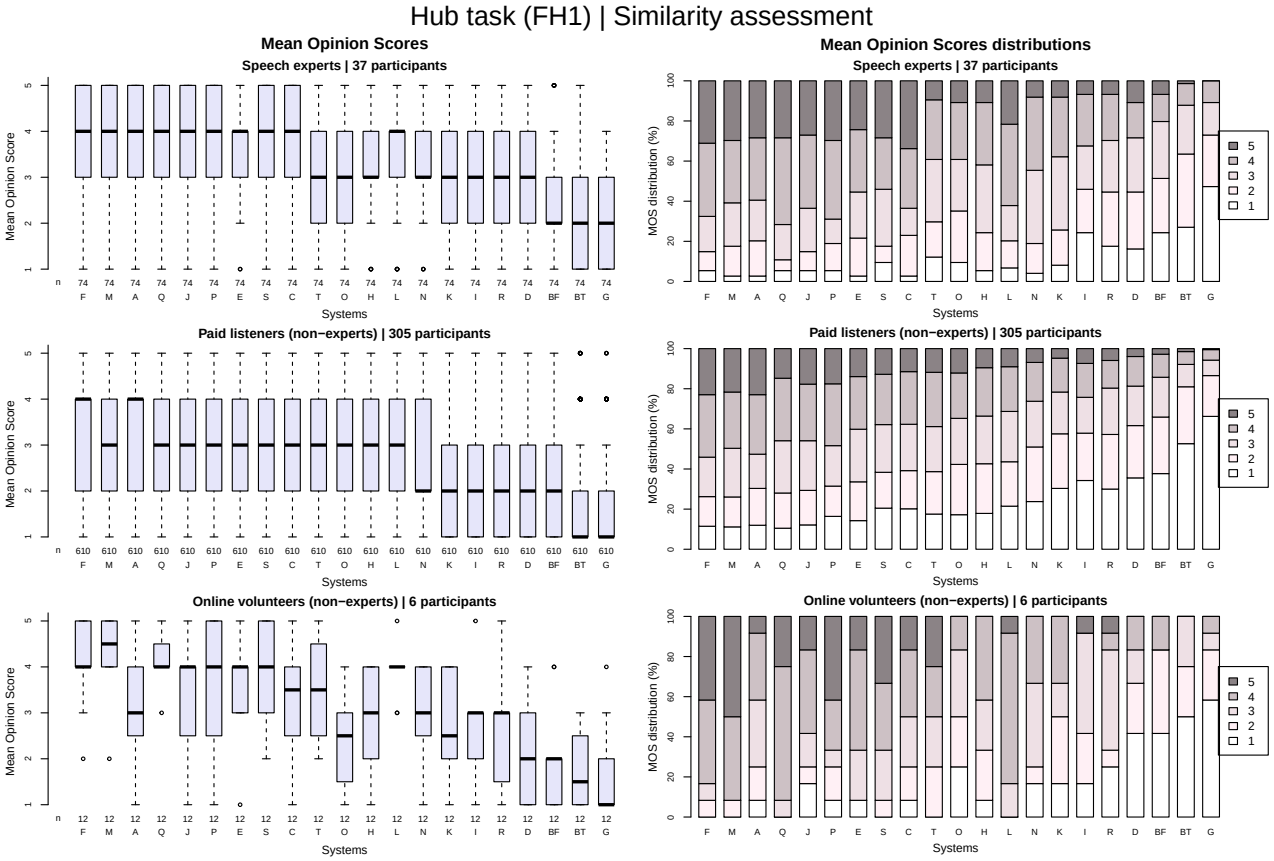


(a) *Speech quality MOS (left) and their distributions (right) per system and listener_type (top to bottom).*

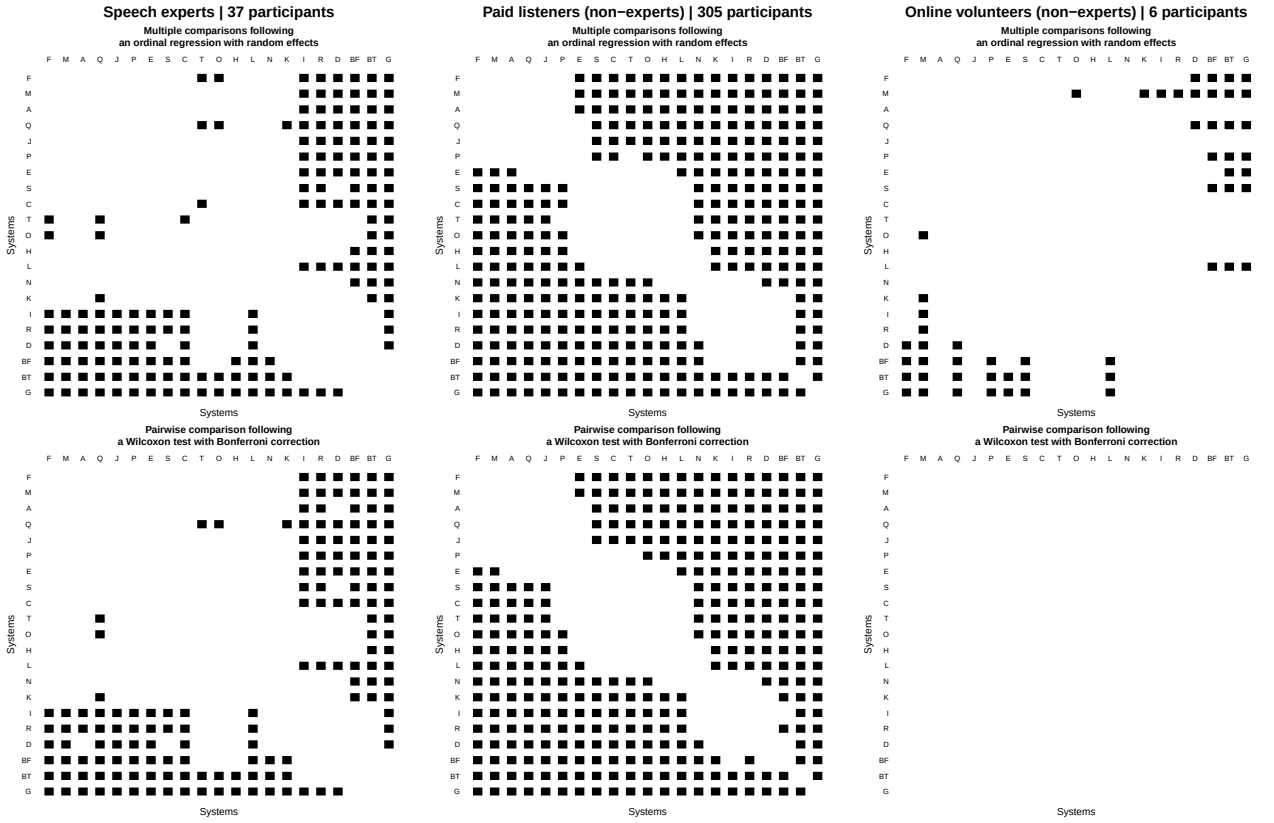


(b) *Significant differences in speech quality MOS between systems and per listener_type, indicated by solid black boxes* ($p < 0.01$). *Top: with multiple comparisons (Ordinal regression) ; Bottom: with pairwise comparisons (Wilcoxon)*

Figure 7: *Speech quality results for FH1, with MOS evaluation (Test 1.a), per system and listener_type.*

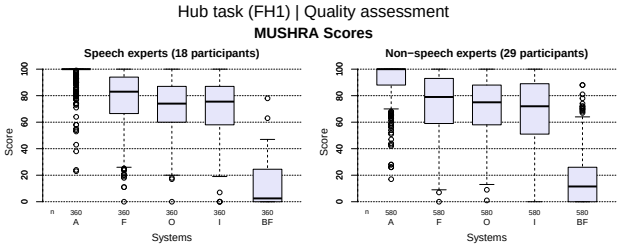


(a) *Speaker similarity MOS (left) and their distributions (right) per system and listener_type (top to bottom).*

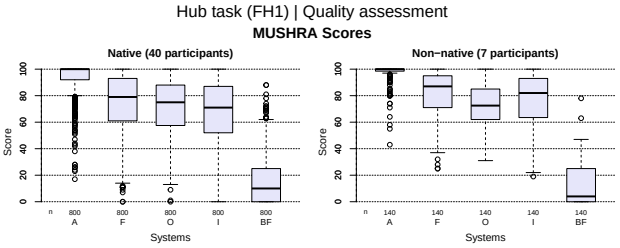


(b) *Significant differences in speaker similarity MOS between systems and per listener_type, indicated by solid black boxes* ($p < 0.01$). *Top: with multiple comparisons (Ordinal regression) ; Bottom: with pairwise comparisons (Wilcoxon)*

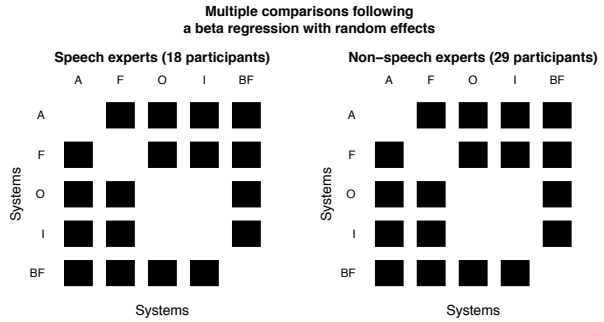
Figure 8: *Speaker similarity results for FH1, with MOS evaluation (Test 1.b), per system and listener_type.*



(a) *Speech quality MUSHRA scores, per system and speech_expert.*



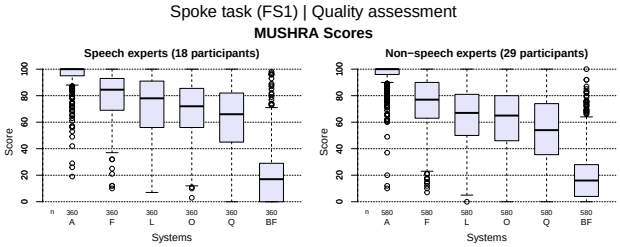
(b) *Speech quality MUSHRA scores, per system and is_native.*



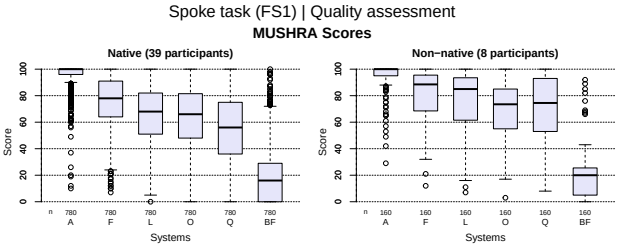
(c) *Significant differences in speech quality MUSHRA scores between systems, per speech_expert, indicated by solid black boxes* ($p < 0.01$).

(d) *Significant differences in speech quality MUSHRA scores between systems and per is_native are not reported here, since the is_native factor does not have a significant impact (see Table 5).*

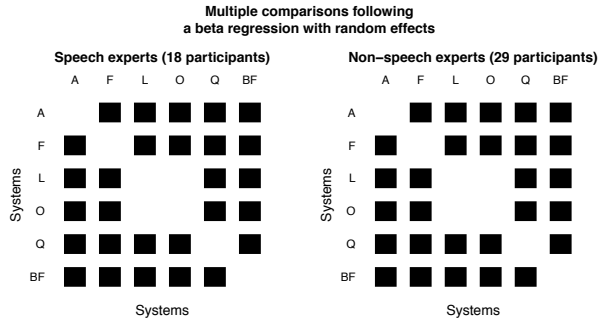
Figure 9: *Speech quality results for FH1, with MUSHRA evaluation (Test 2), per system, speech_expert and is_native.*



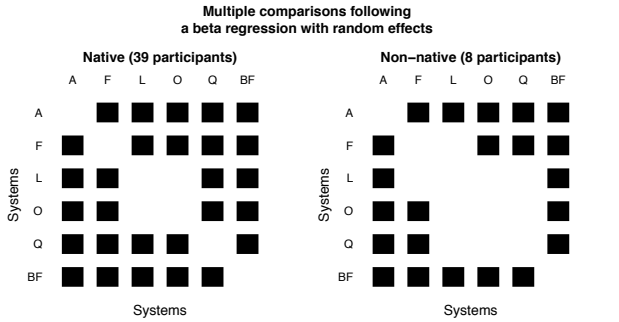
(a) *Speech quality MUSHRA scores, per system and speech_expert.*



(b) *Speech quality MUSHRA scores, per system and is_native.*

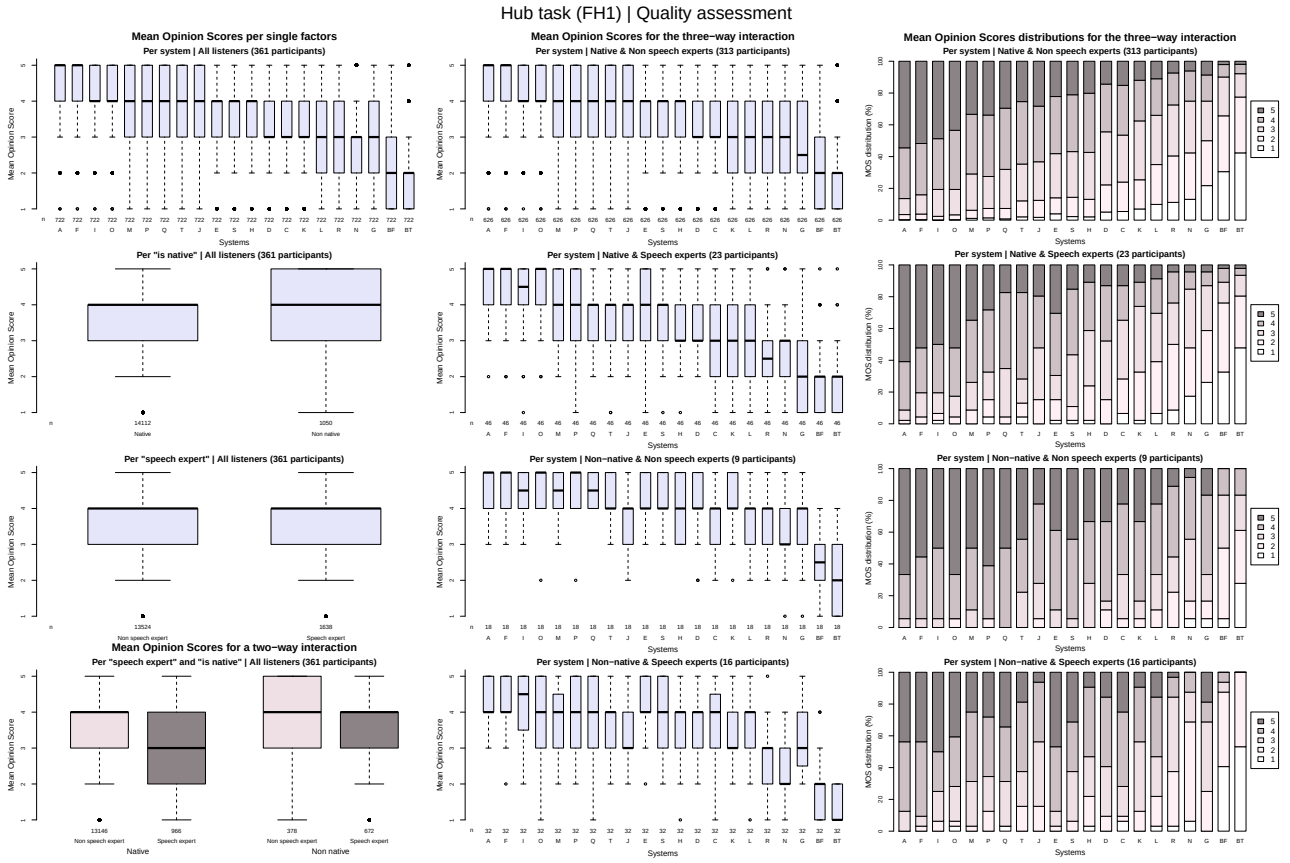


(c) *Significant differences in speech quality MUSHRA scores between systems, per speech_expert, indicated by solid black boxes* ($p < 0.01$).

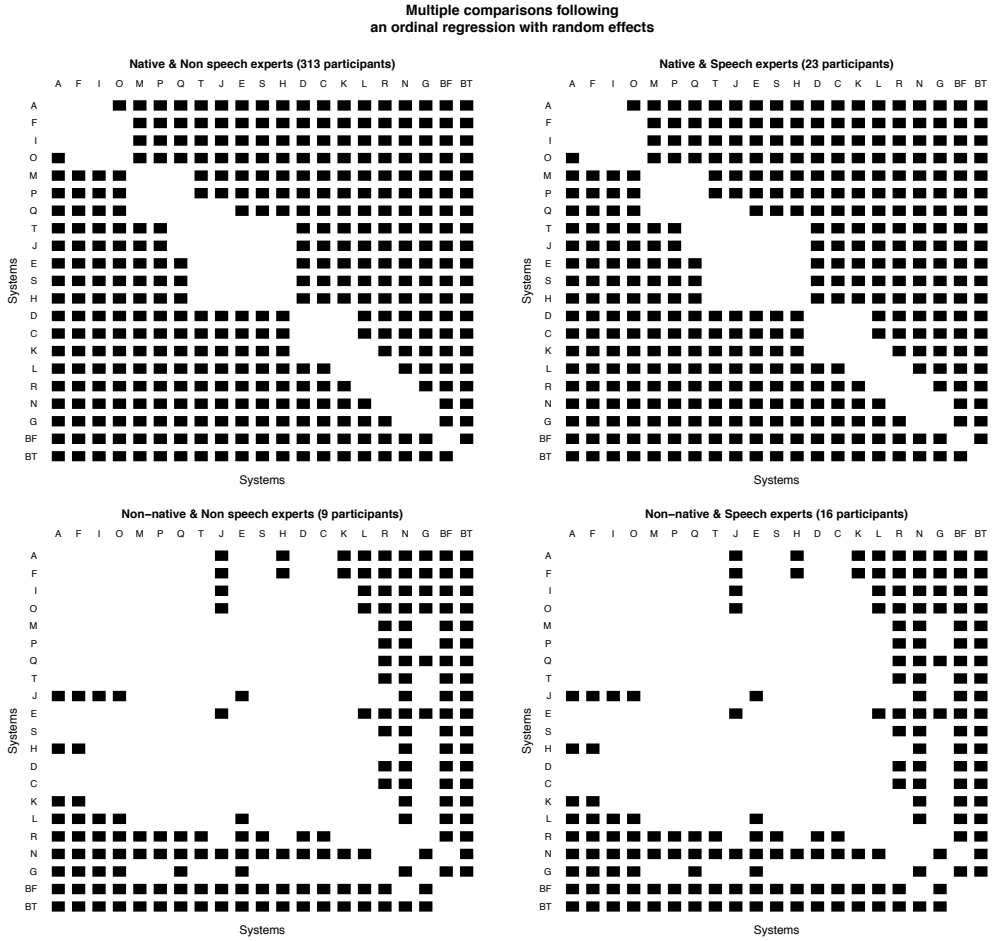


(d) *Significant differences in speech quality MUSHRA scores between systems, per is_native, indicated by solid black boxes* ($p < 0.01$).

Figure 10: *Speech quality results for FS1, with MUSHRA evaluation (Test 5), per system, speech_expert and is_native.*

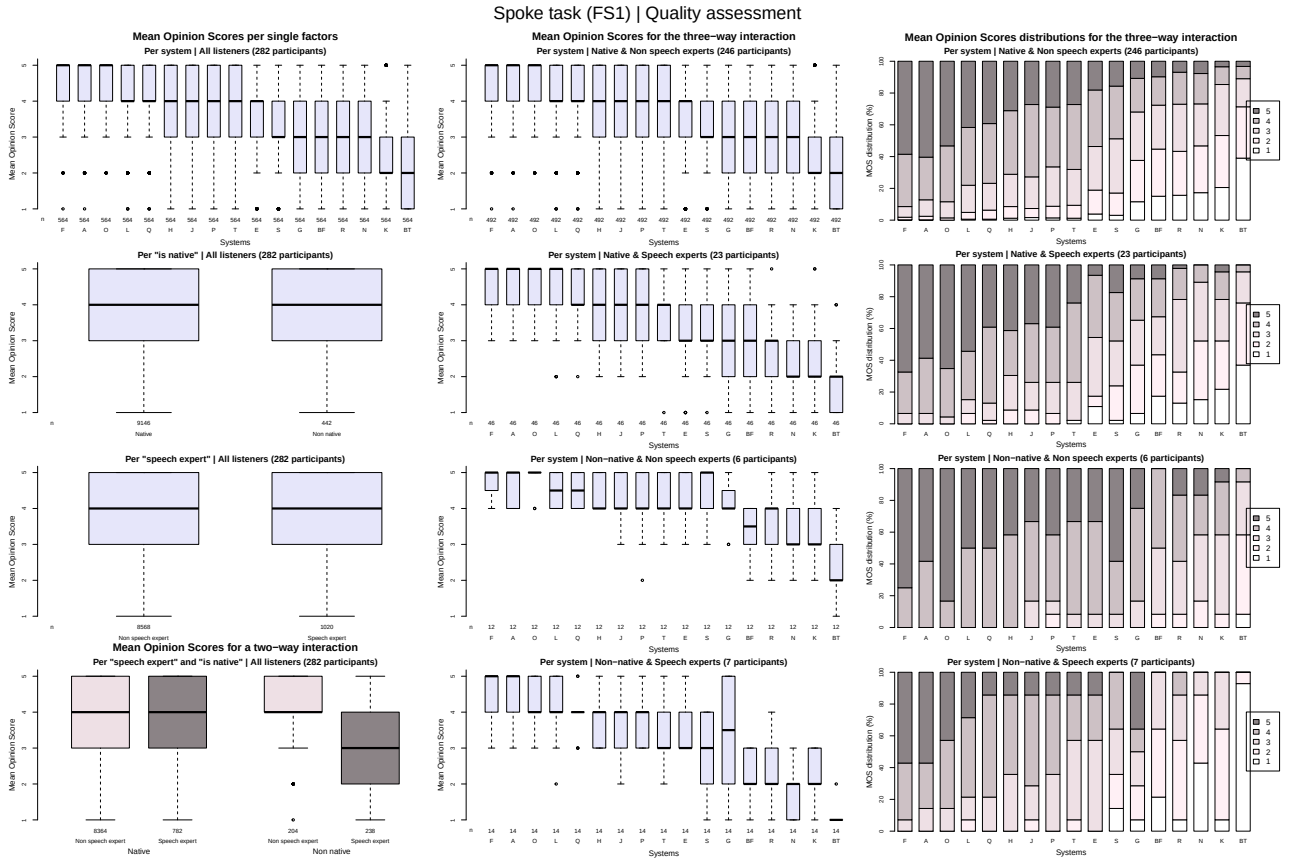


(a) *Speech quality MOS per system, per is_native, per speech_expert and per speech_expert × is_native (left column) ; and given the three factors (second column) along with their distributions (third column).*

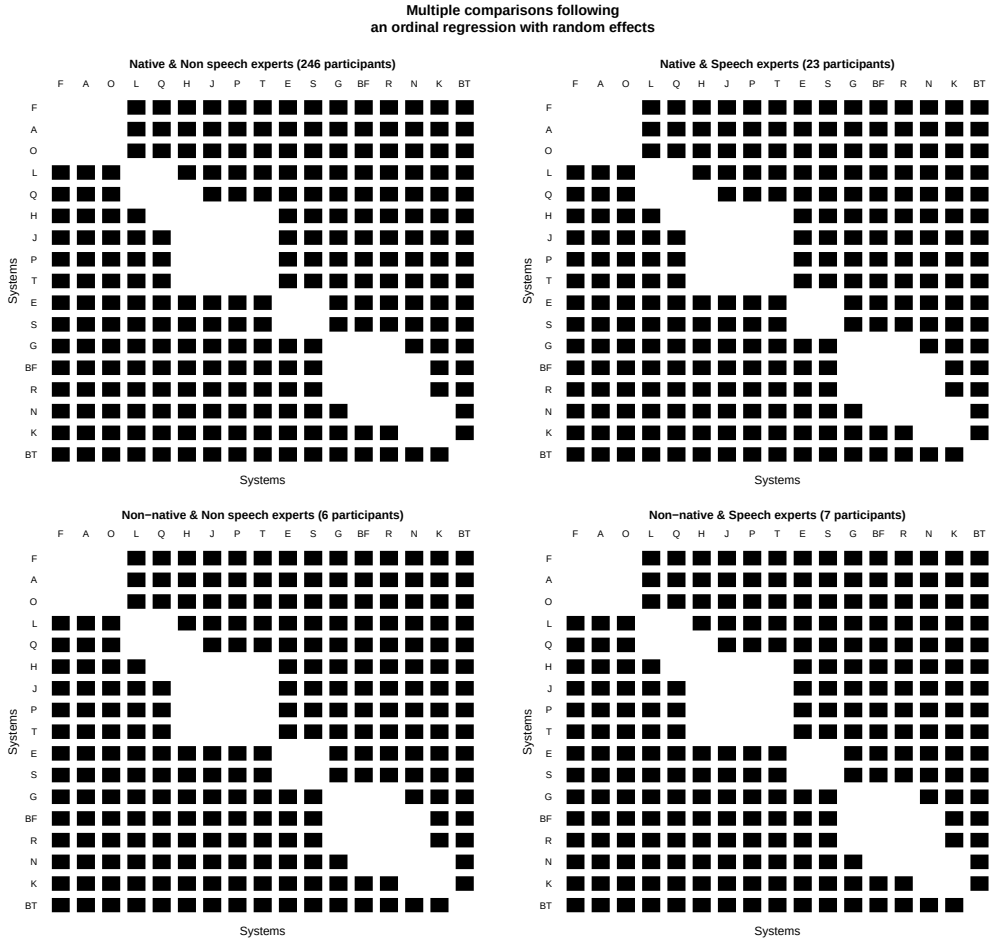


(b) *Significant differences in speech quality MOS between systems and per speech_expert × is_native, indicated by solid black boxes* ($p < 0.01$).

Figure 11: *Speech quality results for FH1, with MOS evaluation (Test 1.a), per system, speech_expert and is_native.*

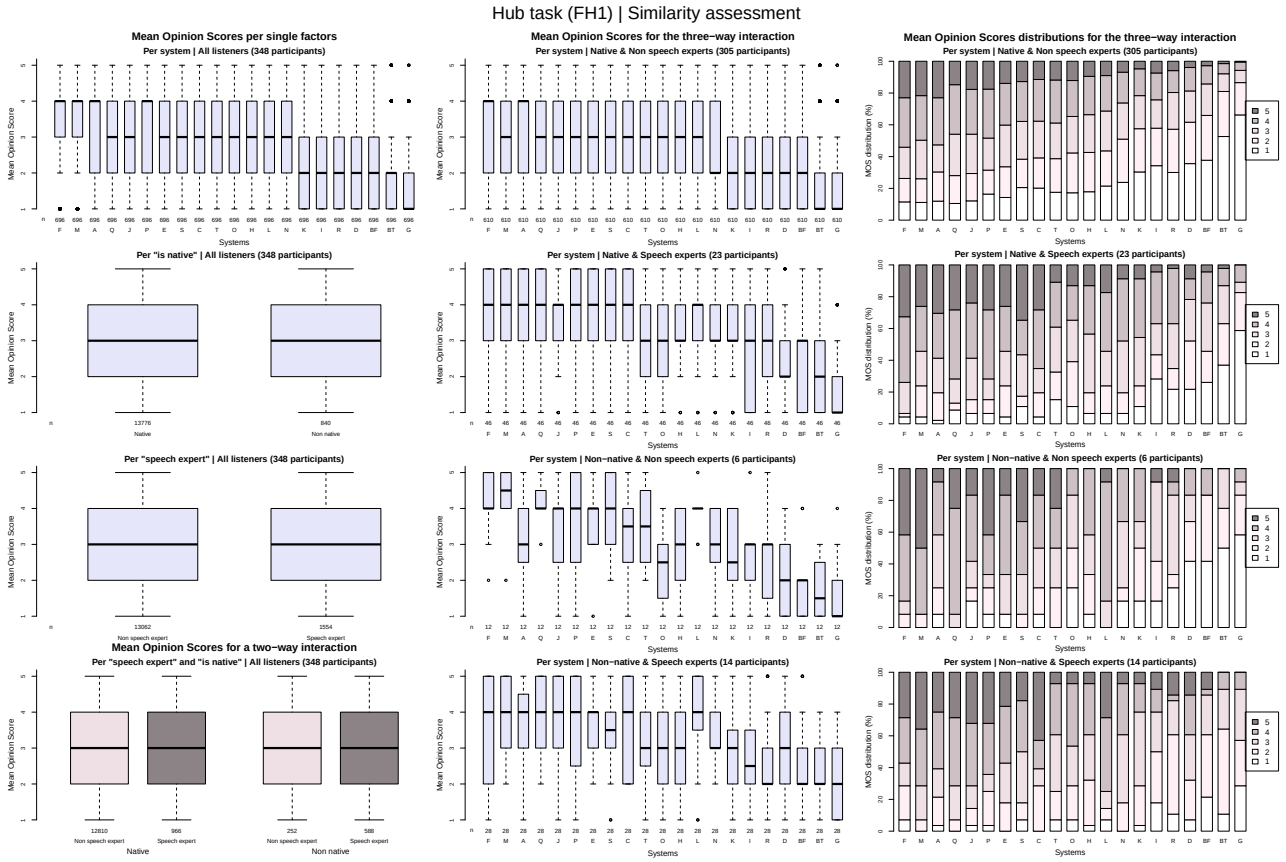


(a) *Speech quality MOS per system, per is_native, per speech_expert and per speech_expert × is_native (left column) ; and given the three factors (second column) along with their distributions (third column).*

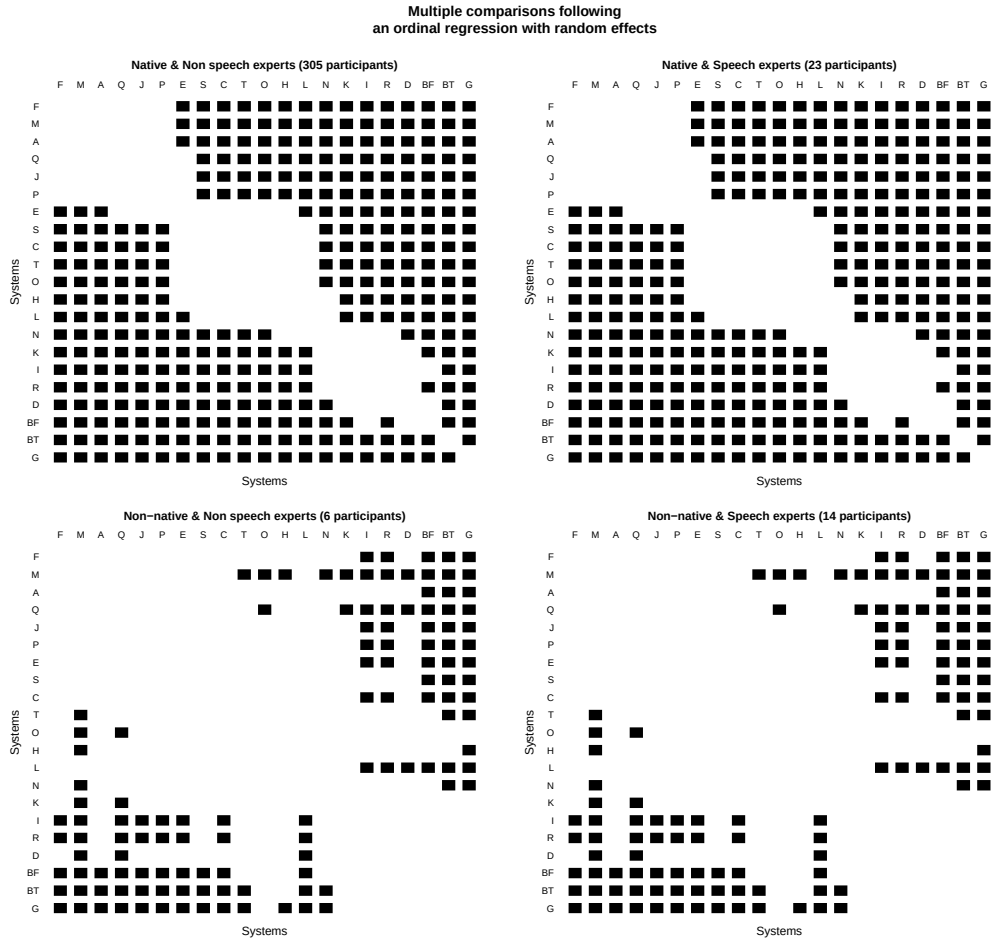


(b) *Significant differences in speech quality MOS between systems and per speech_expert × is_native, indicated by solid black boxes* ($p < 0.01$).

Figure 12: *Speech quality results for FS1, with MOS evaluation (Test 4.a), per system, speech_expert and is_native.*

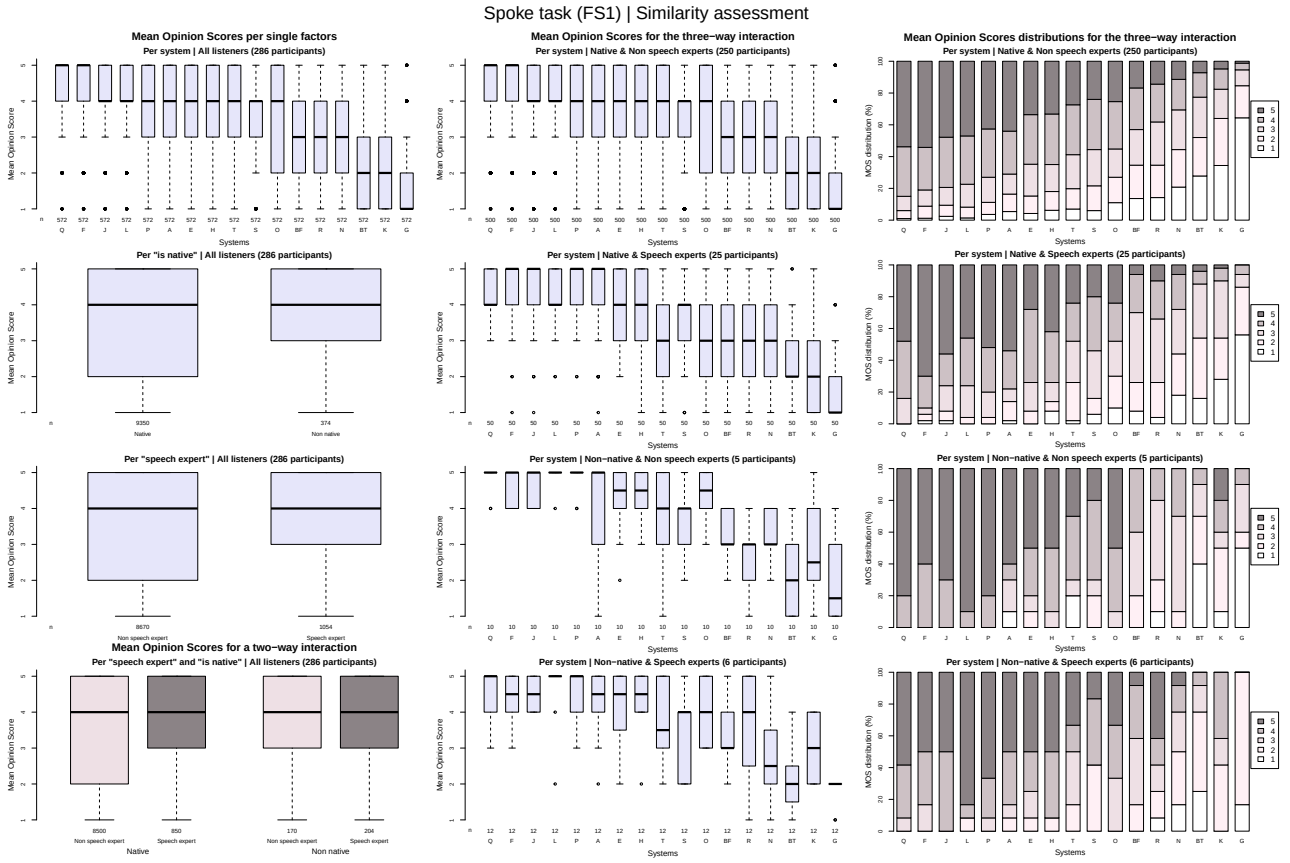


(a) *Speaker similarity MOS per system, per is_native, per speech_expert and per speech_expert × is_native (left column) ; and given the three factors (second column) along with their distributions (third column).*

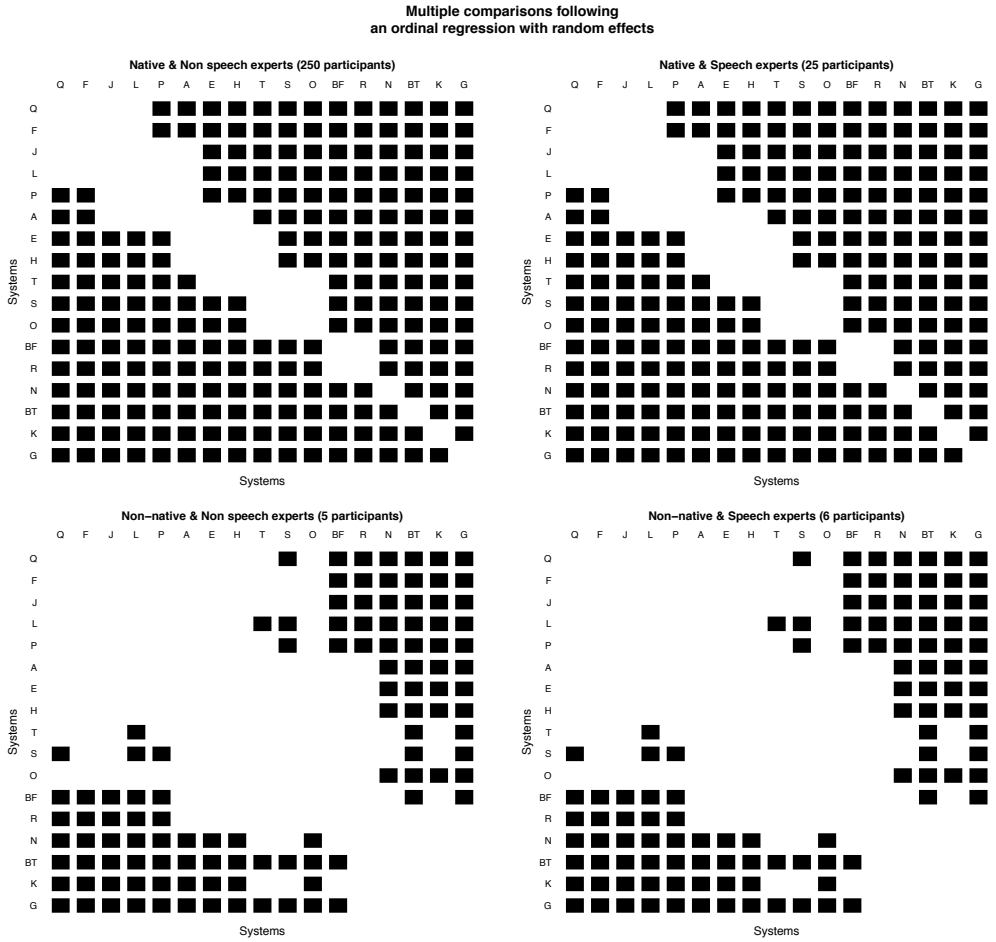


(b) *Significant differences in speech similarity MOS between systems and per speech_expert × is_native, indicated by solid black boxes* ($p < 0.01$).

Figure 13: *Speaker similarity results for FH1, with MOS evaluation (Test 1.b), per system, speech_expert and is_native.*



(a) *Speaker similarity MOS per system, per is_native, per speech_expert and per speech_expert × is_native (left column) ; and given the three factors (second column) along with their distributions (third column).*



(b) *Significant differences in speaker similarity MOS between systems and per speech_expert × is_native, indicated by solid black boxes* ($p < 0.01$).

Figure 14: *Speaker similarity results for FS1, with MOS evaluation (Test 4.b), per system, speech_expert and is_native.*

Table 7: *Number of listeners per listener type for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
SE	39	37	30	31	18	18	11	10
SP	312	305	245	243	29	28	217	208
SR	10	6	7	12	0	1	0	0

Table 8: *Number of self-reported native/non-native listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
native	336	328	269	275	40	39	228	218
non-native	25	20	13	11	7	8	0	0

Table 9: *Self-reported French (Fr.) dialect of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Metropolitan	265	259	204	207	29	36	170	162
Quebecois	31	31	33	33	6	2	21	20
Belgian	15	14	10	11	1	2	11	9
West African	10	9	4	5	1	0	3	3
Central African	7	7	7	7	2	0	5	5
Swiss	5	5	3	3	0	0	6	4
Antillean	4	3	2	2	0	0	2	2
Maghrebi	3	2	4	4	0	1	1	0
Cajun / Acadian	2	2	2	2	0	0	2	2
Indian Ocean	1	1	2	2	2	1	5	5
Other	3	3	3	3	0	0	2	6
Not a Fr. speaker	15	12	8	7	6	5	0	0

Table 10: *Self-reported French (Fr.) proficiency of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Not a Fr. speaker	9	8	3	2	3	2	0	0
Beginner	4	4	2	2	2	2	0	0
Intermediate	3	2	3	2	1	1	0	0
Advanced	5	3	1	1	1	2	0	0
Fluent	4	3	4	4	0	1	0	0
Native	336	328	269	275	40	39	228	218

Table 11: *Self-reported gender of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Female	149	144	123	123	22	19	113	104
Male	203	195	155	159	24	28	113	108
Non binary	9	9	4	4	1	0	2	2
Unanswered	0	0	0	0	0	0	0	4

Table 12: *Self-reported age of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Under 20	3	3	1	1	0	0	1	1
20-29	155	146	124	123	23	20	106	97
30-39	129	126	96	96	12	16	66	64
40-49	40	41	34	38	8	6	31	29
50-59	20	20	16	17	1	4	15	15
60-69	13	11	10	10	2	1	8	7
70-79	1	1	1	1	0	0	1	1
Unanswered	0	0	0	0	0	0	0	4

Table 13: *Self-reported highest level of education of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
High school	29	27	17	16	4	2	20	19
Some university	51	51	38	39	7	7	36	32
Bachelor's Degree	89	87	71	71	5	8	61	58
Master's Degree	169	162	137	140	22	20	99	95
Doctorate	17	15	14	15	7	10	8	6
Other	6	6	5	5	2	0	4	4
Unanswered	0	0	0	0	0	0	0	4

Table 14: *Self-reported Computer Science/Engineering experience of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Yes	130	122	94	99	24	24	65	62
No	231	226	188	187	23	23	163	152
Unanswered	0	0	0	0	0	0	0	4

Table 15: *Self-reported listening speech synthesis experience of listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Daily	50	50	38	38	13	8	25	24
Weekly	97	92	82	82	14	15	58	54
Monthly	83	78	57	60	6	10	49	45
Yearly	13	13	15	13	2	3	8	8
Rarely	105	102	82	84	9	8	78	73
Never	13	13	8	9	3	3	10	10
Unanswered	0	0	0	0	0	0	0	4

Table 16: *Self-reported devices used by listeners for each test after participant screening.*

Test ID	1.a	1.b	4.a	4.b	2	5	3.a	3.b
Headphones	201	195	159	161	32	33	110	102
Earphones	160	153	123	125	15	14	118	112
Unanswered	0	0	0	0	0	0	0	4

Table 17: *Self-reported browser used by listeners for each test after participant screening.*

Test ID	**1.a**	**1.b**	**4.a**	**4.b**	**2**	**5**	**3.a**	**3.b**
Chrome	219	215	169	169	30	21	129	124
Firefox	64	63	53	54	13	16	47	40
Safari	20	17	17	19	1	4	16	15
Edge	33	29	26	26	3	2	20	20
Opera	7	7	5	5	0	1	6	6
Other	18	17	12	13	0	3	10	9
Unanswered	0	0	0	0	0	0	0	4

Table 18: *Self-reported Aurélie Derbier listening frequency of listeners for each test after participant screening.*

Test ID	**1.a**	**1.b**	**4.a**	**4.b**	**2**	**5**	**3.a**	**3.b**
Daily	0	0	0	1	0	0	0	0
Weekly	0	0	0	2	0	0	0	0
Monthly	0	0	0	3	0	0	0	0
Yearly	0	0	0	2	0	0	0	0
Never	361	348	282	278	47	47	228	218

Table 19: *Self-reported environment of listeners for each test after participant screening.*

Test ID	**1.a**	**1.b**	**4.a**	**4.b**	**2**	**5**	**3.a**	**3.b**
Calm the whole time	316	317	255	263	43	40	194	198
Calm most of the time	29	29	21	21	3	7	16	16
Sometimes calm, sometimes noisy	1	0	1	1	1	0	3	3
Noisy most of the time	0	0	1	1	0	0	1	1
Always noisy	0	0	0	0	0	0	0	0
Unanswered	15	1	4	0	0	0	14	0

Table 20: *Listeners' impression of the difficulty of each test after participant screening.*

Test ID	**1.a**	**1.b**	**4.a**	**4.b**	**2**	**5**	**3.a**	**3.b**
Easy	320	139	260	112	37	37	89	208
Difficult	26	208	18	174	10	10	125	10
Unanswered	15	1	4	0	0	0	14	0

Table 21: *Listeners' feedback on each test after participant screening (from a compilation of free answers).*

Test ID	**1.a**	**1.b**	**4.a**	**4.b**	**2**	**5**	**3.a**	**3.b**
Positive feedback	136	77	127	94	15	16	40	100
Few differences perceived between stimuli	6	32	6	20	1	0	/	/
Several reference speakers perceived	5	49	0	1	0	0	/	/
Difficulty to find an evaluation criterion	18	6	5	3	3	3	/	/
Low audio quality	6	9	9	17	0	0	2	0
Scale too wide	3	3	2	3	0	1	/	/
Lack of sense of the utterances	0	0	0	0	0	0	39	0
Instructions unclear	0	0	0	0	0	0	0	13
Others	23	27	10	37	4	9	61	19
Unanswered	164	145	123	111	24	18	76	86