Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023

Olivier Perrotin, Brooke Stephenson, Silvain Gerber, Gérard Bailly, Simon King

Computer Speech and Language 90 (2025) 101747

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023☆

Olivier Perrotin [a,*], Brooke Stephenson [a,b], Silvain Gerber [a], Gérard Bailly [a], Simon King [c,d]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, F-38000, France
[b] CNRS, LIRIS (UMR 5205) and ICAR (UMR 5191), Lyon, France
[c] Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[d] Papercup Technologies Ltd., London, United Kingdom

ARTICLE INFO

Keywords:
Blizzard Challenge
Speech synthesis
Evaluation
Listening test

ABSTRACT

The Blizzard Challenge has benchmarked progress in Text-to-Speech (TTS) since 2005. The Challenge has seen important milestones passed, with results suggesting that synthetic speech was indistinguishable from natural speech in terms of intelligibility in 2021 and that by that same year it was perhaps even indistinguishable in naturalness. The high quality of synthetic speech generated by the latest TTS systems has thus revealed limitations with ITU-T P.800.1 Mean Opinion Score (MOS) in detecting the remaining differences between synthetic and natural speech. Yet, it was the only method used in previous Challenges and is still the most popular method in the field for speech synthesis evaluation. In the 2023 Challenge, we addressed observed limitations of past Challenges by incorporating state-of-the-art speech synthesis evaluation techniques to refine the evaluation of speech quality, speaker similarity and intelligibility. For speech quality, a relative comparison of the systems receiving the best MOS was able to discover a greater number of significant differences between systems. Regarding speaker similarity, we demonstrated that there is a strong bias depending on whether the listeners are familiar with the target voice or not. As for intelligibility, the evaluation of language-specific phenomena, such as the pronunciation of homographs, better highlighted system limits compared to global transcription tasks of synthesised utterances. In addition to reporting results for the 18 entries to the 2023 Challenge, we extend the results analysis to type of TTS module to provide some insights on the most recent advances in model design. Overall, this year's results demonstrate the need for a shift towards new methods for refining TTS evaluation to shed light on increasingly smaller and localised differences between synthesised and natural speech.

1. Introduction

Evaluation of speech synthesis is, on the one hand, a constantly evolving research field, where metrics must be adapted to contend with improved speech quality, diverse data types (e.g., under-resourced languages) and multiplication of communication contexts (Wagner et al., 2019). On the other hand, the Blizzard Challenge is a reference in benchmarking progress in TTS by

☆ This article is an extension of the Blizzard Challenge 2023 technical report (Perrotin et al., 2023).

* Corresponding author.
E-mail address: olivier.perrotin@grenoble-inp.fr (O. Perrotin).

https://doi.org/10.1016/j.csl.2024.101747
Received 17 June 2024; Received in revised form 24 September 2024; Accepted 30 October 2024
Available online 8 November 2024

providing a standardised evaluation protocol which has changed little since 2005. While this has facilitated comparisons between years (Le Maguer et al., 2022, 2024), the 2021 Challenge has revealed a saturation of subjective scores suggesting that synthetic speech was indistinguishable from natural speech in terms of both naturalness and intelligibility. Therefore, to accurately evaluate the synthesis produced by the latest neural systems, a refinement of the Blizzard Challenge evaluation paradigms is needed. In the 2023 Challenge, we strove to find the right balance between incorporating state-of-the-art evaluation techniques while maintaining a consistency of protocol with past editions. While keeping MOS-based global evaluation of synthetic utterances, we further provided fine-grained evaluations of speech quality, speaker similarity, and intelligibility. This article is an extension of the 2023 Blizzard Challenge summary report (Perrotin et al., 2023). In addition to reporting the tasks (Section 2) and evaluation methods (Section 3), we provide in this article a more thorough review of the literature and deeper discussion of the evaluation of speech synthesis (Section 1); we include a concise version of the Challenge results, with an additional analysis of the effect of the type of architecture on speech quality scores (Section 4); and we conclude with recommendations for future Challenges (Section 5).

1.1. A brief history of the Blizzard Challenge

Blizzard Challenges (e.g., Black and Tokuda, 2005; King and Karaiskos, 2013) have been held since 2005. Each Challenge has required participating teams to complete broadly similar tasks: use the released training speech data to build synthetic voices, then submit a set of synthesised test sentences which the organisers evaluate in a large-scale listening test. The Challenges have tackled widely spoken languages (Eberhard et al., 2023): English (2005–2013; 2015–2018); Asian languages such as Mandarin (2008–2010; 2019–2020) and Shanghainese (2020); Indian languages (2013–2015); and two non-English European languages (European Spanish (2021) and Metropolitan French (2023)). The data used in all Challenges is studio-recorded, mostly from a speaker reading novels and/or news articles, but also dialogue transcriptions (Ling et al., 2021). Only one Challenge provided spontaneous (but probably rehearsed) speech in the form of talk-show excerpts (Wu et al., 2019).

Whilst the language and the speech data have varied, the methods of evaluation have remained rather static over time, using four main subjective evaluation paradigms. First, speech naturalness is evaluated using an Absolute Category Rating (ACR) evaluation to obtain listeners' Mean Opinion Scores (MOS). Second, similarity to the reference speaker's voice has also been evaluated using the ACR paradigm. Third, speech intelligibility is measured with a transcription task and reported as Word Error Rate (WER). Finally, following ITU-T P.85 (1994) and Viswanathan and Viswanathan (2005), Hinterleitner et al. (2011) introduced a finer-grained evaluation of naturalness for long-form synthesis, comprising continuous scales for multiple descriptors (e.g., pleasantness, listening effort, speech pauses, intonation); this was employed in six of the Blizzard Challenges (2012, 2013, 2016–2018, 2020).

The Blizzard Challenge has now measured progress for almost 20 years, from unit-selection (with Festival as the benchmark Clark et al., 2006), through statistical parametric models (with HTS as the benchmark Tokuda et al., 2013) to the most recent deep end-to-end methods (with Tacotron2 Shen et al., 2018 and FastSpeech2 Ren et al., 2021 as benchmarks). In the 2021 edition, for the first time in a Blizzard Challenge, one system was rated as indistinguishable from natural speech in terms of MOS naturalness on a 5-point scale; in the same year, practically all systems were not significantly different in intelligibility from natural speech, culminating a progression that started in the early years of the Challenge. With ever-improving synthetic speech naturalness, evaluation methods that measure average naturalness (or intelligibility) over a set of test utterances have reached their limit in being able to pinpoint differences between natural and synthetic speech. Therefore, in the Blizzard Challenge 2023, we devised an evaluation that is able to "zoom-in" on specific contrasts between natural and synthetic speech, by building on the findings and recommendations of the speech synthesis evaluation literature, described in the next section.

1.2. The need for more refined speech synthesis evaluation paradigms

The Mean Opinion Score paradigm described by ITU-T P.800 (1996) and ITU-T P.800.1 (2016) has become a standard in the evaluation of speech generation. It is the mainstay of benchmark competitions – the Blizzard Challenges and various Voice Conversion Challenges (Toda et al., 2016; Lorenzo-Trueba et al., 2018; Yi et al., 2020; Huang et al., 2023) – mainly because it is simple and quick to implement, scales well to a large number of experimental conditions, and the results appear easy to analyse. Nevertheless, two main limitations have been raised in the literature, going back nearly twenty years. First, the apparent simplicity of the protocol hides a significant number of biases (Zielinski et al., 2008), and a series of surveys has pointed out that these biases are more often ignored than mitigated, as detailed in Section 1.2.1. Second, global MOS evaluation focuses on evaluating the implicit closeness of a single synthetic stimulus to natural speech, without reference to phonetics, language, or communicative situation (Haeb-Umbach and Wagner, 2023; Malisz, 2023). A growing number of voices are advocating for more specific evaluations that target the functional aspects of speech (Wagner et al., 2019), as detailed in Section 1.2.2.

1.2.1. Global evaluation: the need for careful calibration of Mean Opinion Score

The MOS paradigm is commonly implemented by: (1) selecting a set of **stimuli** to be evaluated, covering all experimental conditions (which usually are systems in TTS evaluation); (2) submitting these stimuli one at a time to a pool of **listeners** ; (3) asking for an ACR against some **criterion** (e.g., naturalness or speaker similarity). All three aspects have been shown to be sources of bias if not controlled and correctly specified.

When selecting a set of stimuli to evaluate, the range equalising bias (Zielinski et al., 2008) leads listeners to use the full ACR scale to rate them. As a consequence, Le Maguer et al. (2022, 2024) and Cooper and Yamagishi (2023) have shown that the same system's MOS could drop by a full point from one test to another, depending on the stimuli it is rated against. These studies therefore

demonstrate the need to provide adequate high and low anchors in MOS tests to better control the range of notation and comparison between tests.

The second critical aspect is the recruitment of listeners. Wester et al. (2015) demonstrated that there is a strong effect of the number of listeners on the MOS when less than 30 are recruited, which was the case in the majority of the Interspeech papers that they reviewed. Moreover, the *type* of listener also has a potentially large effect on the MOS, but their demographics, expertise, and language competency are rarely reported by authors, presumably because they were not controlled (Wester et al., 2015; Streijl et al., 2016; Chiang et al., 2023; Kirkland et al., 2023). The *conditions* under which listeners perform the test also matter. There has been a shift to recruiting listeners via crowdsourcing platforms, some of which provide little information about their workers. Thus Chiang et al. (2023) showed that not all platforms are equal: listeners provided by Prolific produced more discriminative results than those of Amazon Mechanical Turk. Overall, this strong variability of listener behaviours calls for the need to at least collect listener information and report it so this can be factored into analysis. Controlling the distribution of listener type and recruitment condition would provide results that are more representative of the general population.

The third determining factor is the wording of the instructions to listeners that describes the criterion to rate. Whilst "naturalness" is perhaps the most common wording, it is a poorly-defined concept that is therefore subject to uncontrolled interpretation by listeners (Dall et al., 2014; Shirali-Shahreza and Penn, 2018) and is highly multi-factorial (Seebauer et al., 2023; Shirali-Shahreza and Penn, 2023). Systems rated in terms of "naturalness" have been ordered in the same ranking but with lower absolute MOS than when rated in terms of "quality" (Kirkland et al., 2023) or "appropriateness" (O'Mahony et al., 2021). Systems have been ranked differently in terms of "naturalness" than in "distortion" (Chiang et al., 2023). Providing the communication context of the utterance to evaluate (e.g., read or in a conversation) also has an influence on the naturalness MOS (Dall et al., 2014). These studies underline the importance of providing listeners with a clear definition of the wording of instructions, so that the latter can perform the evaluation on an equal footing, and that the exact wording should also be reported in papers.

Most of the papers cited above have diagnosed a systematic misuse of MOS tests in speech generation research, which is apparent first in the lack of detail reported in publications, and more importantly in the lack of control over the parameters of an experiment.

1.2.2. Specific evaluation: targeting language-specific local events and context

Several directions have been proposed in the literature for a finer evaluation of speech synthesis. A simple approach that is commonly seen is to use a small number of systems which can be rated side-by-side in a multiple comparison setting, such as paradigms based on MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor), originally defined in ITU-R BS.1534-3 (2015) for the evaluation of broadcast audio. Adapted MUSHRA paradigms for synthetic speech evaluation include a natural speech reference hidden amongst the stimuli, which is also provided explicitly. Listeners who give a low score to the reference may be excluded from analysis. The addition of mid-range and lower anchors, recommended by ITU-R BS.1534-3 (2015), is not systematically followed, as there is no standard for mid-quality and low-quality speech synthesis. MUSHRA's explicit relative ranking paradigm focuses listener attention on the differences between stimuli, but this imposes a limit on the number of stimuli presented side-by-side on each MUSHRA 'screen' (12 at most, but preferably no more than 7).

Although TTS systems have generally improved in global quality over time, this does not guarantee better handling of the large number of rare events (i.e., rare events whose diversity is "so large that the probability of encountering at least one of these events in a particular sample approaches certainty", see Möbius (2003, p. 68), such as mispronunciation of under-represented phonemes in the training corpus or realisation of forbidden liaisons), which has led some authors to advocate for evaluations that target such local events, based on phonetics and speech production knowledge (Haeb-Umbach and Wagner, 2023; Malisz, 2023) or speaker idiosyncrasies (Bailly et al., 2023). Pairwise comparison tests are the simplest implementation of this (Shirali-Shahreza and Penn, 2018; Camp et al., 2023; Yasuda and Toda, 2023) and have been applied, for example, to the evaluation of obstruents generated by neural end-to-end systems (Pandey et al., 2023). Alternatively, listeners may be asked to annotate places in a sentence which display local prosodic mismatch (Gutierrez et al., 2021), or an artefact, etc. The histogram of listeners' responses over each utterance then reveals their locations (de Kok, 2013). The main drawbacks of pairwise evaluation and local annotation is scaling to a large number of systems.

Another direction is the evaluation of speech signals as a whole, where listeners are provided with the associated communicative intent (Wagner et al., 2019). Clark et al. (2019) showed that stimuli are given higher MOS when listened to with preceding context (although the wording of instruction differed between context conditions). Listeners are also able to detect prosodic mismatch between a synthesised sentence and its preceding context (O'Mahony et al., 2021). One step further is the evaluation of comprehensibility, i.e., whether the *information* (rather than the words) has been understood by the listener. Pisoni et al. (1987) proposed a reaction time comprehension test, where listeners were asked to decide if a sentence was true or false; they observed longer reaction times for synthetic speech than natural speech, i.e., showing longer processing for synthetic speech. In contrast to this very specific experimental setting, Wester et al. (2016) experimented with the use of questionnaires on the content of long-form synthesised excerpts, but this did not shed light on differences between synthetic and natural speech. Other researchers have investigated the online processing of speech through behavioural experiments: the gating paradigm quantifies the presence of lexical or supra-lexical information in different parts of an utterance by masking part of it (Morlec et al., 2001); close-shadowing (Bailly, 2003) highlights the sensitivity of the listener to the fluency of the utterance; bio-signals including EEG (Parmonangan et al., 2019) or pupil dilatation (Govender et al., 2019) also shed light on listener's cognitive processing. Since evaluation methods for comprehensibility are very diverse, no consensus has emerged, at least for speech synthesis (Wagner et al., 2019).

1.3. Innovations for the Blizzard Challenge 2023

The Blizzard Challenge 2023 was organised by the Université Grenoble Alpes and used read sentences in Metropolitan French (French from France).

In past Blizzard Challenges, the organisers all followed the same protocol. Generally, benchmark systems such as Festival (Clark et al., 2006) or HTS (Tokuda et al., 2013) were included; these were intended as low anchors, although they were not necessarily the lowest-rated systems. This consistency of protocol has led to a reproducible system ranking over the years (Cooper and Yamagishi, 2021; Le Maguer et al., 2022, 2024). To maintain this for 2023, *we choose to retain a global MOS-based evaluation* of speech quality and of speaker similarity, in addition to making innovations and improvements. The 2021 challenge revealed a saturation in results, which motivates our introduction of fine-grained evaluations of speech quality, speaker similarity, and intelligibility.

Innovation in global evaluation.

Careful choice of wording and reporting of instructions: to avoid the issue of the under-defined concept "naturalness", we evaluated *quality*. To ensure transparency and reproducibility of our evaluations, the exact instructions given to participants are provided in Appendix.

Objective pre-selection of stimuli: to focus the subjective evaluations on stimuli that best discriminate between the submitted systems, we sub-selected test utterances according to the systems dispersion given an objective measure (described in Section 3); this was done for every listening test.

Statistical analysis: previous challenges adopted the non-parametric pair-wise statistical analysis proposed by Clark et al. (2007). Such repeated measures have limitations, so we instead fitted a single statistical model to explain the variance of the listeners' responses in each specific test. So, for each Test, we performed a dedicated statistical analysis tailored to the listener response type (e.g., ordinal, continuous, binary) and which supports multiple comparisons across all experimental conditions at once.

Innovation in fine-grained evaluation.

Fine-grained quality test — MOS then MUSHRA: the evaluation of a large number of systems on a 5-point scale makes it hard to differentiate between similar-quality systems. As supported by Cooper and Yamagishi (2023), we added a supplementary test to refine the initial Mean Opinion Scores of quality with a MUSHRA design. The supplementary test included only the highest-rated systems, to reduce the number of experimental conditions.

Fine-grained similarity test — effect of listener: we demonstrated in the Blizzard Challenge 2023 that speaker similarity evaluation is an ill-posed problem as it is difficult to judge the identity of a voice that has never been heard before. So, we further tested the effect of familiarity of the listeners with the speaker.

Fine-grained intelligibility test — language-specific task: to address the synthesis of local events, we chose to evaluate the intelligibility of French-specific heterophonic homographs.

From all the aspects discussed previously in Section 1.2, only the evaluation of the speech signal in the context of its associated communicative intent was not addressed in the 2023 Blizzard Challenge, simply because we were not able to identify a relevant evaluation method. Nevertheless participating teams were asked to synthesise suitable materials, described in Section 2, which we have released as a resource for others to use.

Open data and reproducibility. 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. Training datasets for the Challenge tasks were also made freely-available. All links to these resources are reported in Table 1. To encourage reproducibility and provide exhaustive examples of speech evaluation methods to the community, we took great care to provide every detail of each experimental protocol in this single document, as well as the detailed instructions given to listeners in Appendix.

2. Challenge tasks and participating teams

2.1. Tasks

There were two tasks in the Blizzard Challenge 2023, each using a dedicated dataset made freely-available for the Challenge (Table 1). The complete task specification given to participating teams is on the Challenge website (Table 1). 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 was not trained by the participating team (e.g., pre-trained wav2vec, BERT, etc.).

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 proceedings	https://www.isca-archive.org/blizzard_2023/index.html
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

2.1.1. Hub task FH1: French TTS

The Hub dataset is a subset of the M-AILABS French dataset (Solak, 2019), comprising 51 h of speech from five audiobooks read by the female speaker Nadine Eckert-Boulet (NEB) in Metropolitan French; the full contents of the Hub dataset are online (Table 1). All recordings come from the free public domain audiobook LibriVox project (Kearns, 2014), and the texts were taken from the Gutenberg Project (Project Gutenberg Literary Archive Foundation, 1971). Each audio file corresponds to one chapter, originally at 44.1 kHz sample rate, but downsampled to 22.05 kHz for the challenge. The text processing described by Lenglet et al. (2021) and performed on part of the data was extended to the full Hub dataset for the Blizzard Challenge 2023. We first restored the original chapter structure by aligning the text from the Gutenberg Project with the recordings from LibriVox. We normalised all text, including spelled out abbreviations and numbers, and manually corrected misspellings and omissions. Paragraph ends were annotated with the punctuation mark "§", placed after the final punctuation mark preceding each new line. Text transcriptions were then segmented into utterances at all pauses of at least 400 ms in the corresponding audio. The audio recording of each chapter was thus paired with the text: participating teams were thus free to also use any of the room tone recording in-between utterances. In total, the Hub dataset includes 67 329 utterances from 289 original chapter recordings, each with an orthographic transcription. 40 715 of these utterances were semi-automatically aligned with phonetic transcriptions,[1] originally gathered from the Robert dictionary (edition 1995) and enriched with full form variations (case and gender for nouns and adjectives, conjugated forms for verbs). All transcriptions have been hand-corrected over the years by the fourth author. The phonetic alphabet is described on the dataset website (Table 1).

The objective of the Hub task was to build a voice from the provided French Hub dataset. Two repeatability requirements were imposed on participating teams: (1) the use of external models was allowed only if they are publicly-available pre-trained models, and a reference is cited; (2) all audio data used for training models (including for fine-tuning pre-trained models) must be publicly available and reported. Participating teams were required to synthesise 3 test sets, provided as normalised orthographic text only:

$\mathsf{MOS}_{\mathsf{FH1}}$ 1000 distinct sentences read by NEB but from an audiobook not in the Hub dataset (*"Le Vingtième Siècle : La Vie électrique" – Albert Robida)*, which were 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" (Merriam-Webster, 2024) (such as 'fils' which is either 'son', pronounced /fis/, or the plural of 'fil', a thread or wire, pronounced /fil/). From Hajj et al. (2022), we identified 36 homograph pairs and created 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 by Benoît et al. (1996).

$\mathsf{EXP}_{\mathsf{FH1}}$ sentences for evaluating expressivity and comprehensibility. $\mathsf{EXP}_{\mathsf{FH1}}$ 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 small paragraphs (40 ± 6 words), extracted from the same book as $\mathsf{MOS}_{\mathsf{FH1}}$.

2.1.2. 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 apart from the release for the Challenge. Sentences are taken from the SIWIS database (Honnet et al., 2017), which is composed of isolated sentences from French novels and French parliamentary debates. Similar audio processing (i.e., downsampled to 22.05 kHz), text processing and segmentation as that for the Hub dataset was performed, and we selected 2 h of this corpus to release as the Spoke dataset. In total, it includes the audio of 2515 utterances from 13 audiovisual recording sessions, all with aligned phonetic transcriptions. Again, participating teams were free to

[1] Some text inputs already incorporate words that are phonetically transcribed, such as acronyms or unusual pronunciations.

also use the room tone in-between utterances. The Spoke task was to build a voice from the provided dataset. None of the Hub task reproducibility requirements were imposed for this task, but they were highly encouraged. Participating teams were required to synthesise only:

MOS_{FS1} 400 distinct sentences from French parliamentary debates, to be used for quality and speaker similarity evaluation, also from the SIWIS dataset, but not part of the Spoke training set.

2.2. Systems

2.2.1. Systems submitted by participating teams

18 teams participated to the Challenge, all submitting to the Hub task FH1 and 14 to the Spoke task FS1, (summarised in Table 2). Systems are identified using letters in all published results. Letter 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 used encoder–decoder architectures: 11 had either a FastSpeech 2-like or non-attentive Tacotron-like architecture, and the remaining 7 used stochastic models, mostly with variational auto-encoders conditioned on text (see Table 2). 15 systems employed a GAN-based vocoder for waveform generation, of which 6 were trained end-to-end with the acoustic model, (i.e., the representation between the acoustic model and the vocoder is also learned instead of using a spectrogram). 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 3 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 3 teams provided links to their full system implementation, which was an optional reproducibility criterion. Full descriptions of all submitted systems are available in the Blizzard proceedings (link in Table 1).

2.2.2. Benchmark systems

BT: Tacotron 2 benchmark. Our first benchmark model combines Tacotron 2 (Shen et al., 2018) with HiFi-GAN (Kong et al., 2020). 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 BT for FH1 from scratch on the Hub dataset for 158 500 steps. We then fine-tuned the 100 000-step checkpoint of the model on the Spoke dataset for an additional 57 500 steps for the FS1 task. We used the hyperparameter values recommended in the NVIDIA repository (batch size: 32; learning rate: 1×10^{-3}; weight decay: 1×10^{-6}; see link in Table 1 for the complete settings). For waveform generation, we used the HiFi-GAN implementation from Kong et al. (2020) with the provided pre-trained universal model `UNIVERSAL_V1/g_02500000` (link provided in Table 1).

BF: FastSpeech 2 benchmark. The second benchmark model is FastSpeech 2 (Ren et al., 2021), used with the same vocoder as 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 2.1.1). 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 phone set used for the training data. The BF model for the FH1 task was trained from scratch on the Hub dataset for 333 935 steps with a batch size of 32, a constant learning rate of 5×10^{-4}, and clip-norm 5.0 (Fairseq recommended values). The BF model is then fine-tuned for an additional 7254 steps on the Spoke dataset for the FS1 task.

3. Evaluation method

3.1. 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 (2000).

3.2. Objective measures

In order to select the most informative samples for subsequent subjective evaluation, we identified those test utterances which exhibited the most variation between systems, as in Raidt et al. (2004). Two objective distance measures were employed: the *spectral distance* is the root mean square error (RMSE) between the Dynamic Time Warping (DTW)-aligned mel spectrograms of a pair of speech samples (same text, synthesised by differing systems). The *duration distance* is the ratio of the DTW path length over the average mel-spectrogram duration of the 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 samples, respectively. For the MOS_{FH1} and MOS_{FS1} test sets separately: we computed the spectral and duration distances for every utterance ($N = 996$ for MOS_{FH1}, due to missing submissions from a few systems; $N = 400$ for MOS_{FS1}), for all possible pairings of the n participating systems, resulting in a $N \times n \times n$ matrix for each of the two distance measures. Each matrix was normalised by its average value across all dimensions, since the two distances have very

Table 2
Evaluated systems in the Blizzard Challenge 2023. The first two are the benchmarks. The remaining rows are the systems submitted by the participating teams, ordered in an attempt to cluster systems which share similar properties. The method descriptions are summarised based on questionnaires and on the workshop paper from each participating team (see Saget et al., 2023; Lenglet et al., 2023; Lu et al., 2023b; Xu et al., 2023; Ma et al., 2023; Lux et al., 2023; Zalkow et al., 2023; Boros et al., 2023; Zaïdi et al., 2023; Xie et al., 2023b; Veaux et al., 2023; Bu and Zhao, 2023; Lu et al., 2023a; Qi et al., 2023; Xie et al., 2023a; Jiang et al., 2023; Shang et al., 2023; Chen et al., 2023). When the vocoder is between parentheses, it has been trained end-to-end with the acoustic model. FS1 indicates whether the participating team submitted an entry for the FS1 task. L2S and LLM stand for Letter-to-Sound module and Large Language Model, respectively.

Team name	FS1	L2S	Prosody control (during inference)	Acoustic model	Vocoder	LLM input
			Benchmark systems			
BF: FastSpeech 2	✓	■ eSpeak	■ Variance predictors from text	■ FastSpeech2	◆ HiFi-GAN	/
BT: Tacotron 2	✓	/	/	◆ Tacotron2	◆ HiFi-GAN	/
			Participating systems			
Saget et al. (2023)	✓	◆ Own L2S	■ Variance predictors from text	■ FastSpeech2 (TTS) + WavLM-Tacotron2 (VC)	■ WaveGlow	/
Lenglet et al. (2023)	✓	◆ Own L2S in encoder	■ Variance predictors from text	■ FastSpeech2-based	■ WaveGlow	/
Lu et al. (2023b)	✓	◆ Own L2S + CamemBERT + ChatGPT	◆ Prosody predictor (GST/VAE) from text + CamemBERT + Speech type	■ FastSpeech2-based with conformers	◆ HiFi-GAN	● Text + Prosody
Xu et al. (2023)	✓	◆ Own L2S + BERT	● Prosody predictor (GST) from text + BERT Variance predictors from text	■ FastSpeech2-based with conformers	◆ HiFi-GAN	● Text + Prosody
Ma et al. (2023)		■ eSpeak	● Prosody predictor (VQ-VAE) from FlauBERT Variance predictors from text	■ FastSpeech2-based	◆ HiFi-GAN	◆ Prosody
Lux et al. (2023)	✓	■ eSpeak + CamemBERT	● Prosody predictor (GST) from reference audio Variance predictors from text + GST	■ FastSpeech2-based with conformers	◆ BigVGAN	■ Text
Zalkow et al. (2023)	✓	■ eSpeak	■ Variance predictors from text	◆ Forward / Fast Tacotron	◆ StyleMelGAN	/
Boros et al. (2023)	✓	◆ Own L2S	■ Variance predictors from text + CamemBERT	◆ RNN-based	◆ (HiFi-GAN)	◆ Prosody
Zaïdi et al. (2023)	✓	■ eSpeak + CamemBERT	◆ Prosody predictor (VAE) from text	◆ VAE-Tacotron	◆ HiFi-GAN	■ Text
Xie et al. (2023b)	✓	◆ Own L2S + CamemBERT	◆ Prosody predictor (RNN) from text	◆ Non-attentive Tacotron	◆ HiFi++	■ Text
Veaux et al. (2023)	✓	◆ Own L2S + FlauBERT	◆ Prosody predictor (GST) from FlauBERT	◆ Non-attentive Tacotron	◆ HiFi-GAN-based	● Text + Prosody
Bu and Zhao (2023)		/	◆ Prosody predictor (Flow) from text	● Flow-VAE	◆ BigVGAN	/
Lu et al. (2023a)	✓	■ eSpeak	◆ Prosody predictor (Flow) from text	● VITS	◆ (BigVGAN)	/
Qi et al. (2023)	✓	■ eSpeak	◆ Prosody predictor (Flow) from text	● VITS	◆ (HiFi-GAN)	/
Xie et al. (2023a)	✓	■ eSpeak + pBART	◆ Prosody predictor (Flow) from text	● VITS	◆ (HiFi-GAN)	/
Jiang et al. (2023)		■ eSpeak	◆ Prosody predictor (Flow) from text + GPT-3	● VITS	◆ (HiFi-GAN)	◆ Prosody
Shang et al. (2023)		◆ Own L2S + BERT	◆ Prosody predictor (H-VAE) from text	● Hierarchical VAE	/	■ Text
Chen et al. (2023)	✓	■ eSpeak + CamemBERT	■ Variance predictors from text	● Diffusion Transformer	● FastDiff	■ Text
			Factor levels colour legend for statistical analysis per type of module			
		■ eSpeak L2S ◆ Own L2S module	■ Variance predictor ◆ Prosody predictor ● Variance + prosody predictors	■ FastSpeech2-based ◆ Tacotron2-based ● Stochastic	■ Flow-based ◆ GAN-based ● Diffusion-based	■ Text ◆ Prosody ● Text + Prosody

Table 3
The five listening tests described in Section 3.3.1, their designs as described in Section 3.3.2, 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 every sentence-system combination was rated in each test. Additional information about the test listeners is provided in Table 4.

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

different magnitudes. These two normalised matrices were summed elementwise to obtain a single $N \times n \times n$ distance matrix, then summed across systems 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 believe that these would be the least informative samples to present to listeners. The most informative utterances will be selected as stimuli in Section 3.3.3.

3.3. Subjective evaluation

3.3.1. Tests

Table 3 describes how the dimensions of quality, similarity and intelligibility were evaluated using five independent listening tests. Tests 1 and 4 measured Mean Opinion Scores of speech quality (1.a & 4.a) and speaker similarity (1.b & 4.b) for tasks FH1 and FS1 and using the MOS_{FH1} and MOS_{FS1} test sets, respectively. 21 systems were evaluated in Test 1: natural speech A, benchmarks BF and BT, and the systems from 18 participating teams (C to T). 17 systems were evaluated in Test 4: natural speech A, benchmarks BF and BT, and the systems from 14 participating teams.

Tests 2 and 5 measured quality, again using the MOS_{FH1} and MOS_{FS1} test sets for tasks FH1 and FS1 respectively, but this time using a MUSHRA design and only for the systems with highest quality found by Tests 1.a & 4.a respectively. The procedure for identifying those systems is described in Section 3.4.3 and they were compared against natural speech and the benchmark receiving the highest MOS in Tests 1.a & 4.a (which is BF). 5 systems were selected for Test 2: natural speech A, benchmark BF, and the 3 best rated systems found by Test 1.a. 6 systems were selected for Test 5: natural speech A, benchmark BF, and the 4 highest quality systems found by Test 4.a.

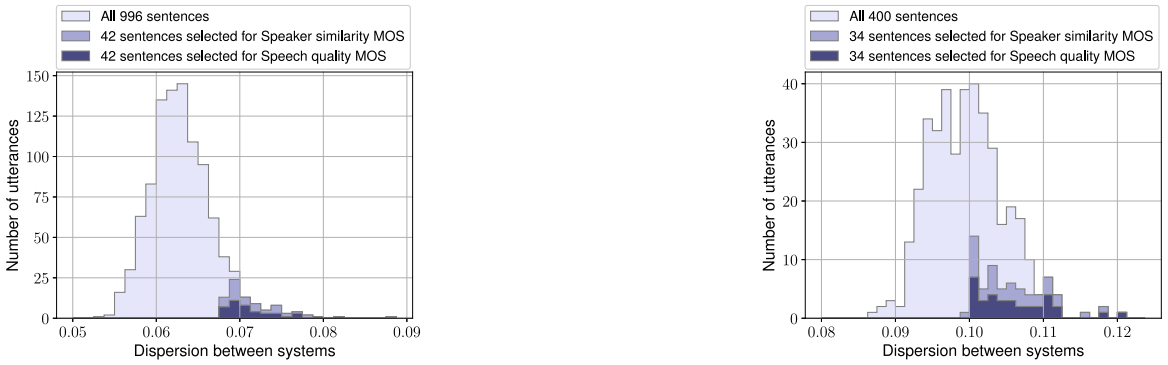
Test 3 measured intelligibility for task FH1 using the INT_{FH1} test set. (There was no evaluation of intelligibility for FS1.) In Test 3.a, listeners performed a transcription task on SUS stimuli from which we measured Word Error Rate. In Test 3.b, listeners performed and ABX task on homographs from which we measured Pronunciation Accuracy. 20 systems were evaluated in Test 3: benchmarks BF and BT, and systems from 18 participating teams. (We do not have natural speech recordings of the INT_{FH1} test set.)

3.3.2. Design

Following Fraser and King (2007), Tests 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 m in the following: $m = 21$ for Tests 1.a and 1.b ; $m = 20$ for Tests 3.a and 3.b ; and $m = 17$ for Tests 4.a and 4.b. System ordering within blocks was systematically varied by using a Latin Square design. For Tests 1.a, 1.b, 4.a, and 4.b (resp. 3.a), one experimental block consisted of presenting two different sentences (resp. one sentence) per system for a total of $2m$ (resp. m) sentences, so that all the sentences and all the systems under evaluation were heard within each block. m experimental blocks with a circular permutation of systems ensured that all sentence and system combinations were eventually rated by the m groups of listeners.

Test 3.b employed 36 pairs of homographs, each included in a context utterance, for a total of 72 stimuli. The combinations of stimuli and systems are again divided into m experimental blocks following a Latin square design ($m = 20$): each experimental block presented the 72 sentences, with a rotation of the m systems for each sentence. m experimental blocks with a circular permutation of systems ensured that all sentences and systems combinations were eventually rated by the m groups of listeners. To remove the effect of the context sentence, we designed a second version of Test 3.b with another set of 72 stimuli comprising the same homographs but in different context utterances. In practice, for Test 3, we ran a first round of tests with m groups of listeners performing Test 3.a and the first version of Test 3.b. Then, in a second round, m groups of different listeners performed Test 3.a and the second version of Test 3.b.

Tests 2 and 5 evaluate all systems side-by-side, sentence by sentence. 20 sentences were selected for each test and all listeners performed the same test. The complete set of test designs is summarised in Table 3.



(a) MOS_{FH1} utterances (b) MOS_{FS1} utterances

Fig. 1. Distributions of the MOS_{FH1} and MOS_{FS1} utterances' dispersion between systems according to the objective distance defined in Section 3.2.

Table 4
Number of listeners per recruitment type (Paid or Volunteer) for each test, given as: number of listeners retained after screening/total number of listeners who completed that test. Test 3 required only native speakers of French and therefore no volunteer listeners were used.

Test	Paid listeners	Volunteer listeners	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%)

3.3.3. Materials

Tests 1 and 4 used sentences from the MOS_{FH1} and MOS_{FS1} test sets for tasks FH1 and FS1, respectively. Fig. 1 displays the distributions of the MOS_{FH1} (left) and MOS_{FS1} (right) utterances' dispersion between systems according to the objective distance defined in Section 3.2. The MOS_{FH1} test set distribution has a long tail towards high dispersion, and for Test 1 we selected the $4m = 84$ sentences that maximised the dispersion between the systems. The MOS_{FS1} test set distribution is more narrow, and for Test 4 we randomly sampled $4m = 68$ sentences that display a dispersion above the average up to the maximum value. For both tests, half of the sentences were employed in the speech quality evaluations (1.a & 4.a) and half in the speaker similarity evaluations (1.b & 4.b). Selected sentences are displayed with darker colours in Fig. 1. After the systems that obtained the best speech quality MOS were selected (as described in Section 3.4.3), the 20 sentences from Test 1.a (resp. Section 4.a) that maximised dispersion between the selected systems were retained for Test 2 (resp. 5). Therefore, sentences for the MUSHRA quality tests are subsets of those from the MOS quality tests. For Test 3.a, the subset of the 20 SUS sentences from the INT_{FH1} test set that maximised dispersion between all systems were used for evaluation. For each of the 72 homographs in INT_{FH1}, we randomly selected two of its three contextualising sentences for the two versions of Test 3.b, respectively. As is standard practice in Blizzard Challenges, only a relatively small subset of the material synthesised by participating teams using their systems was actually used in the listening tests, leaving a large amount of synthetic speech material available to use in future listening tests. The complete listening test results, including every individual listener response to every stimulus presented, are distributed via the Blizzard Challenge archive in a package that also includes all submitted synthetic speech (Table 1).

3.3.4. Implementation

All tests were implemented with the Web Audio Evaluation Tool (Jillings N. and D., 2015) and full details of the tasks and instructions given to listeners are provided in Appendix. 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 Perrotin et al. (2023).

3.3.5. Listeners

All listeners were recruited online via the two following methods:

Paid listeners. recruited via the Prolific crowdsourcing service (see Table 1). Inclusion criteria: self-certified French native speakers from any country of origin; no self-reported hearing problems. Listeners were instructed to wear earphones or headphones for the test. All instructions for this group of listeners were given only in French (see full instruction details in Appendix). Each of the five tests was independently posted on Prolific, so a listener could participate in several tests but not in more than one experimental block of any given 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; 4 listeners for Test 3.b. The overall number of completed tests is given in Table 4. Listeners were compensated at a rate equivalent to £10 per hour,[2] based on estimated completion times 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 3.

Online volunteers. recruited via mailing lists. Inclusion criterion: no self-reported hearing problems. Listeners were required to wear earphones or headphones for the test. All the test instructions for this group of listeners were given in English (see full instruction details in Appendix). Since Test 3 required French proficiency which was not an inclusion criterion for this group of listeners, only Tests 1, 2, 4 and 5 were submitted to volunteers as four independent URLs. Because some listeners dropped the test, and they chose freely among the four URLs, we did not control the number of online volunteers per experimental block, for each test. The overall number of completed tests is given in Table 4.

We discarded the responses of a small number (1–3) of listeners for Tests 1 and 4 (MOS) who used only one or two levels from the 5-point scale across the whole test; we did not replace these listeners. For Tests 2 and 5 (MUSHRA), we discarded the responses of listeners who rated the hidden natural speech reference at less than 80 on average across the whole test; this applied to about one quarter of paid listeners, so we recruited additional listeners to achieve 30 listeners per test, after screening. The final numbers of listeners are reported in Table 4.

3.4. Analysis methodology

3.4.1. Score computation

For Tests 1 & 4 (MOS) and 2 & 5 (MUSHRA), we analysed the scores given by listeners. For Test 3.a (intelligibility using SUS), Word Error Rate (WER) was calculated, allowing for certain spelling variations including homophones (available in the results package from the Challenge archive (Table 1). Compared to MOS and MUSHRA, which are subjective judgements, Test 3.b has an objective answer: whether the pronunciation of the homograph by a system is correct. So, in this test, listeners can be seen as annotators. A minimum of 4 listeners were recruited per experimental block and we used Fleiss' kappa to obtain an inter-listener agreement value per block (Landis and Koch, 1977). 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 was selected by the majority of listeners, and compared it to the expected pronunciation. In this manner we obtained a binary correct vs. incorrect score for each sentence-system combination.

3.4.2. Statistical analysis

Previous challenges have used the statistical analysis specified by Clark et al. (2007). 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 listeners that are both speech experts and native). 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, but this correction is too strong in the sense that it conversely decreases the chance of discovering statistically significant results.
- A Wilcoxon test compares pairs of distributions based on the ranking of the samples from both distributions. In the case of ACR 5-point ratings, where there are only 5 unique values, there will be a large number of ties in any ranking, thus limiting the power of this statistical test to find differences between systems.

For the above reasons, we introduced a new statistical method for the 2023 Blizzard Challenge, comprising the following 5 steps:

Step (1) Selection of the factors under investigation. Our two main factors of interest are the *speech_expert* and *is_native* factors. The *speech_expert* factor has 2 levels: speech experts, SE, who self-identified as such among both paid listeners and volunteers ; and non-speech experts, NSE = SP + SR, with SP the subset of paid listeners (all of whom are native speakers of French) who did not self-report as a speech expert; and SR the subset of (unpaid) volunteers who also did not self-report as a speech expert. The *is_native* factor has two levels, native and non-native: listeners who self-identified as native (resp. non-native) speakers of French.

Step (2) Descriptive statistics. As in previous challenges, for each identified combination of factors, we calculated a set of descriptive statistics that comprises: 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 NONE of the scores are normally distributed. For instance, most tests are carried out on an ordinal scale. In this case, the mean and standard deviation values are not meaningful: we only use the mean to decide a system ordering to be used consistently across all plots. The descriptive statistics for all tests are available in the results package.

[2] Prolific is a UK-based platform, so payment rates were set in GBP.

Table 5
Summary of statistical tests performed on the listener responses obtained in each of the listening tests.

Test	1, 4	2, 5	3.a	3.b
Score	MOS	MUSHRA	WER	Correct score
Data type	Ordinal	Proportion	Proportion	Binary
Statistical model	Ordinal-	Beta-	Beta-	Logistic-
		regression with random effects		
R function	`clmm`	`glmmTMB`	`glmmTMB`	`glmer`
R package	`ordinal`	`glmmTMB`	`glmmTMB`	`lme4`
Post-hoc analysis	Estimated marginal means	Method from Hothorn et al. (2008)		
R function	`emmeans`	`glht`	`glht`	`glht`
R package	`emmeans`	`mutlcomp`	`mutlcomp`	`mutlcomp`

Table 6
Significance of the different factors and their interactions involved in Tests 1, 2, 4, and 5, according to the statistical models listed in Table 5, ($p < 0.01$). 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)	✓	✓	✓	✓	✓	✓
speech_expert (SE, NSE)	✓	✓	✓	✓	✓	✓
speech_expert × *system*			✓	✓		
is_native (native, non-native)	✓	✓		✓	✓	✓
is_native × *system*	✓			✓	✓	✓
speech_expert × *is_native*		✓				
speech_expert × *is_native* × *system*						

Step (3) Statistical models. For each test and identified combination of factors, we fitted a statistical model suitable for the type of data, as specified in Table 5. All statistical models included the *sentence* and *listener_ID* as random factors.

Step (4) Assessing the significance of factors. For each statistical model, the effect of individual factors and their interactions are tested by removing them one at a time from the full statistical model, and assessing if the removal of each factor has a significant impact on the model. We started with the random factors *sentence* and *listener_ID*, then proceed to interactions between factors and, only if the latter were non-significant, we removed the factors involved in those interactions. A likelihood ratio test (`anova` function in R) was used to assess the significance of each factor or interaction removal ($p < 0.01$).

Step (5) Multiple comparisons. Once the statistical model has been thus simplified, we performed multiple comparisons between levels of the remaining significant factors. The appropriate post-hoc analysis method depends on the data type (see Table 5). Statistics and p-values for each pairwise comparison are provided in the results package and statistical significance is set for $p < 0.01$.

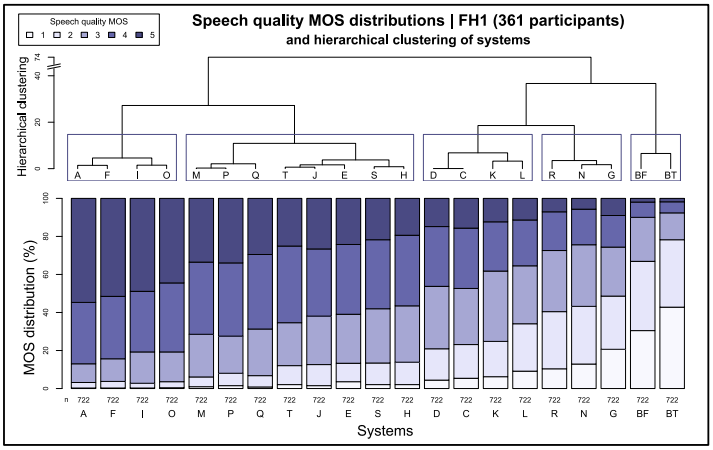
For the sake of brevity, we do not report the results of Wilcoxon's signed rank tests in this article, but we showed in the 2023 Blizzard Challenge summary paper that when the number of observations is small, Wilcoxon's signed rank tests display less statistical differences between systems than the ad hoc regressions (Perrotin et al., 2023, Fig. 7 and 8).

3.4.3. Identification of the highest-quality systems, for inclusion in the MUSHRA tests

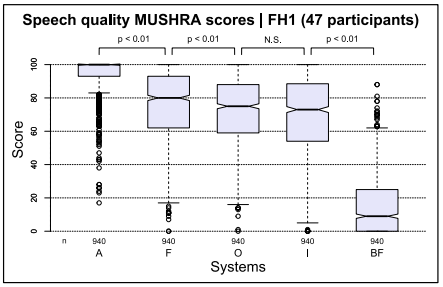
To refine the speech quality evaluation obtained with MOS in Tests 1 & 3, we submitted the systems with the highest quality to a MUSHRA test, in Tests 2 & 5. To select which systems, we first fitted a statistical model to the MOS data, with the effect of *system* only, and computed multiple comparisons between systems, leading to a matrix of statistic values with an element for each pair of systems. Then, we used the matrix of statistic values as a distance matrix to perform a hierarchical clustering of the systems (R function `agnes` from the `cluster` package). The clustering is visualised as a tree in Fig. 2, which reveals which systems have similar quality. To decide on the number of clusters, our aim was to select 3 to 5 systems with the highest quality, since this is a suitable number for a MUSHRA test. We clustered all systems into 5 clusters, for both FH1 and FS1. All systems in the cluster with the highest MOS quality scores were then selected: this gave 3 systems for FH1 and 4 for FS1. These systems, along with natural speech and the BF benchmark, were then re-evaluated using MUSHRA in Tests 2 & 5.

4. Results

The results of the speech quality and speaker similarity evaluations in Tests 1, 2, 4 and 5 were analysed with respect to the factors listed in Section 3.4.2, with the significance of each factor and their interactions being calculated using an appropriate statistical model (Table 5) then summarised in Table 6. The significant impact of the *system* factor trivially shows that the submitted systems



(a) Speech quality MOS distributions per *system* (bottom) and hierarchical clustering of systems based on the multiple comparison (top). Ordinates of the hierarchical clustering are distances based on the statistic values obtained from the multiple comparisons of systems (see Section 3.4.3 and Table 5).



(b) Speech quality MUSHRA scores per *system*.

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

generate synthetic speech that listeners perceive as significantly different; this will be discussed in Section 4.1.1. The significance of the random *sentence* and *listener_ID* factors shows that, for all tests, these factors explain a significant amount of the variance in the results. The significance of the *speech_expert* factor in all tests indicates that speech experts SE evaluated speech synthesis differently than non-experts NSE. Moreover, this difference in behaviour also affects the relative difference between systems for the MUSHRA Tests 2 & 5, given the significant interaction between the *speech_expert* and *system* factors for these tests. Similarly, the *is_native* factor has a significant effect in most tests: native listeners judged synthetic speech differently than non-native listeners. These results are further discussed in Section 4.1.3, but it is already clear that a listener's profile has a significant effect on evaluation scores, especially in the fine-grained tests MUSHRA tests.

The remainder of this summary is organised as follows, Section 4.1 presents the results for speech quality evaluation (Tests 1.a, 2, 4.a, 5), first per *system*, as in previous challenges, then Sections 4.1.2 and 4.1.3 present a deeper analysis of the effects of system architecture and listener. Section 4.2 discusses the results for speaker similarity evaluation (Tests 1.b & 4.b). Section 4.3 reports intelligibility results for both SUS and homograph syntheses (Tests 3.a & 3.b). For each test, systems are presented in a consistent order: descending average score for that test calculated from the responses of all listeners combined. As already noted, 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 indicate which systems are significantly better than others. For that, pairwise significance between systems was calculated and will be reported in the discussion. We consider two conditions as significantly different if $p < 0.01$.

4.1. Evaluation of speech quality

4.1.1. Effect of system only

Hub task (FH1). Fig. 2 summarises all the results of the speech quality evaluation for task FH1. MOS distributions are displayed in Fig. 2(a); the hierarchical clustering places the systems into five groups. The highest quality group of systems comprises A, F, I, and O, of which F and I are not statistically different from natural speech A, and the only significantly different pair is O and A. The second group of systems comprises M, P, Q, T, J, E, S, and H, all with a median MOS of 4. The benchmarks BF and BT are in the lowest quality group and each have a median MOS of 2. So, all systems entered into the challenge by participating teams generated significantly higher quality speech than the benchmarks.

The systems in the first group (A, F, I, O) plus the best benchmark system (BF) were subsequently re-evaluated using the MUSHRA paradigm in order to make finer distinctions, with the result shown in Fig. 2(b). Now *all* systems were judged as significantly different from natural speech. 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. I and O remain not significantly different from one another.

In summary, the results from this MUSHRA test highlight the limits of the preceding MOS test, which was unable to find all distinctions between a large number of systems. On the other hand, MUSHRA uncovered those finer distinctions between, with the obvious limitation of only comparing a modest number of systems at once.

Spoke task (FS1). Fig. 3 summarises the differences in median speech quality MOS per system between tasks FH1 and FS1. Recall that only 2 h of speech from the target speaker was provided for FS1. It is striking from Fig. 3 that the global range of MOS for FS1 is comparable to FH1. This is further demonstration that MOS ratings are relative to the systems within that particular test (Cooper

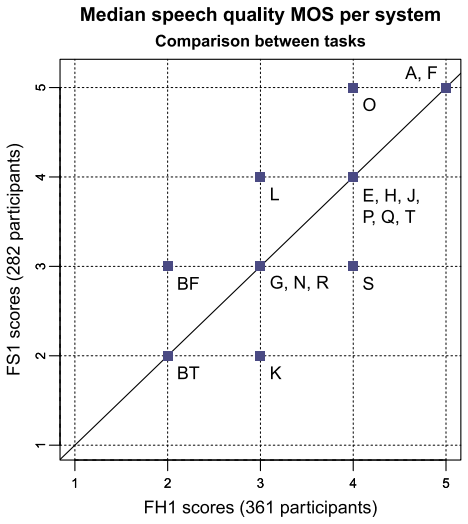


Fig. 3. Comparison of median speech quality MOS per *system* between the two tasks, FH1 and FS1.

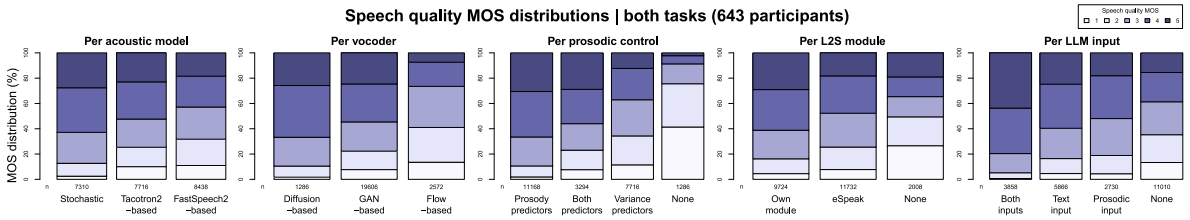


Fig. 4. Speech quality MOS distributions from both tasks per TTS system modules.

and Yamagishi, 2023; Le Maguer et al., 2022, 2024). Most systems were given approximately the same ranking and a similar score as in FH1.

Systems F, L, O, Q, BF were submitted to a follow-on MUSHRA test (Test 5), just as was done in FH1. While F and O were not statistically significant from natural speech (A) when evaluated with the MOS paradigm, all systems were significantly different from natural speech in that MUSHRA test. This is another demonstration that MOS tests containing a large number of systems are not sufficient for fine-grained evaluation of speech synthesis, which was made possible with the MUSHRA test performed on fewer selected systems.

4.1.2. Effect of system architecture

Fig. 4 breaks down the speech quality MOS of both FH1 and FS1 tasks combined per type of module used in each system architecture, as categorised in the bottom of Table 2. Before commenting on these results, it is important to note that our empirical selection of these five types of modules does not cover the full design choices of one system, and only 20 out of 432 combinations of these modules are represented by the 18 participating systems and the 2 benchmarks. Therefore, the interactions between the use of these modules cannot be modelled here, and the trends that we highlight for some modules are highly intertwined with the rest of the architecture of the participating systems. For each module taken separately, we fit an ordinal regression with the *sentence* and *listener_ID* factors as random effects. For each module, a post-hoc analysis with estimated marginal means showed that all pairs of MOS distributions are significantly different.

We first observe that systems implementing a stochastic model (VITS Kim et al., 2021, Diffusion-based, VAE-based) were given higher speech quality MOS than Tacotron2-based and FastSpeech2-based models. Regarding neural vocoders, flow-based models (all using WaveGlow Prenger et al., 2019) were judged significantly worse than Diffusion-based or GAN-based (HiFiGAN Kong et al., 2020, BigVGAN Lee et al., 2023) models. The third panel highlights that using a prosodic control significantly improves the quality of output speech, with the use of prosody predictors (GST-based Wang et al., 2018, VAE-based, or Flow-based) being more performant than FastSpeech-based variance predictors (Ren et al., 2021). As for letter-to-sound (L2S) mapping, eSpeak, which was used by 10 of the 20 systems, is better than using none at all, but participating teams who used their own L2S mapping were given higher quality scores. Finally, 11 of the 20 systems used a French large language model (LLM) such as BERT (Devlin et al., 2019),

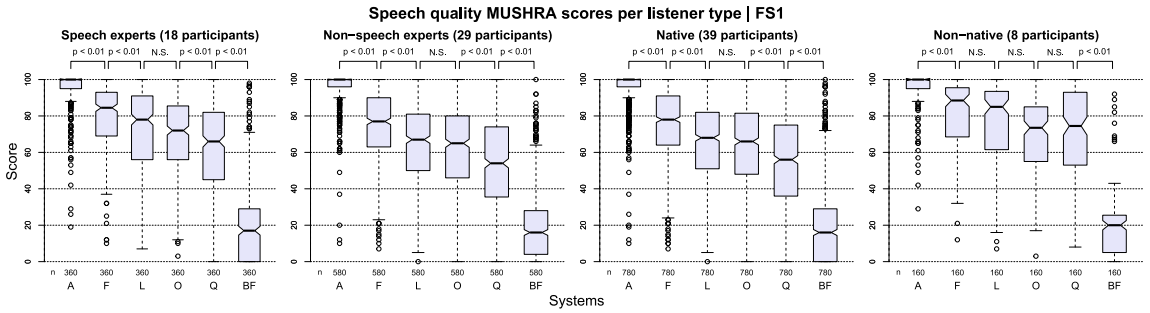


Fig. 5. Speech quality MUSHRA scores for task FS1 (Test 5) per *system*, *speech_expert* and *is_native*.

FlauBERT (Le et al., 2020), or CamemBERT (Martin et al., 2020) to complement either the text or the prosodic control input. All were beneficial to the output speech quality, with a greater effect when complementing the text input than the prosodic control input. The combination of both displays increased speech quality MOS by a significant margin.

Overall, this breakdown highlights that the most recent modules (stochastic models, prosody predictors, use of LLM) globally display the greatest perceived speech quality. However, note that the best rated system F in terms of quality in all Tests 1.a, 2, 4.a, 5, is FastSpeech2-based. Therefore there is a high variation of speech quality in the use of a given system module, that is due to interactions between modules, but also to the selection and preparation of the training data, to the design of the training routine, etc. Thus, we can *not* generalise from this study that one module is superior to another without context, but only provide general trends.

4.1.3. Effect of the type of listener

Table 6 shows that there are significant interactions between the *speech_expert* and *system* factors as well as between the *is_native* and *system* factors. Fig. 5 displays the MUSHRA scores per *speech_expert* and *is_native* factors for task FS1. Synthetic systems were given better scores by SE than NSE, relatively to natural speech (A) and the benchmark (BF). Note that this does not affect the significance of the differences between systems. Regarding the interaction between the *is_native* and *system* factors, we observe that non-native listeners gave higher scores to the synthetic systems, relatively to natural speech (A) and the benchmark (BF). It suggests that understanding the language leads to a more severe judgement of the global quality. Moreover, pairwise differences between systems indicate that non-native listeners perceived less significant differences between systems than native listeners. The results per *speech_expert* and *is_native* factors for speech quality MOS and speaker similarity MOS on tasks FH1 and FS1 display 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 NSE.
- The *is_native* factor has a significant and important effect on the results, with lower scores given by native listeners, and much less pairwise differences perceived by non-native listeners compared to native listeners.

Overall, systems were judged differently according to the *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.

4.1.4. Automatic prediction of speech quality MOS

In parallel to running our quality MOS evaluation, we provided the organisers of the VoiceMOS challenge with the same synthesised speech stimuli as a test set for automatic MOS prediction (Cooper et al., 2023). Participants of the VoiceMOS challenge were not given the true subjective MOS, that were only used to evaluate their automatic prediction. As reported in their summary paper, Cooper et al. (2023) found inconsistencies between the accuracy of automatic prediction on FH1 and FS1, as they observed that three (resp. four) out of nine systems make better MOS predictions on FS1 (resp. FH1). Only two systems perform equally well on both tasks, with one having correlations above 0.75 for both tasks. They also note a global higher difficulty to predict the MOS of systems G, D, E, H and R. Note that these five systems include all the levels of acoustic model, prosodic control, L2S modules and LLM input identified in Table 2. If the automatic prediction of subjective scores is not yet fully mature, the perspectives laid by Cooper et al. (2024) are promising to imagine on the one hand complementing subjective evaluations with such automatic predictors, not only for global MOS but for various tasks, and on the other hand use these systems as a proxy to better understand the human perception of speech.

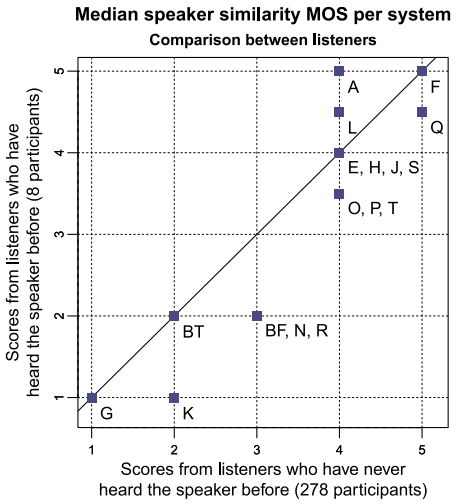


Fig. 6. Comparison of median speaker similarity MOS per *system* between listeners who have heard and not heard the voice to be synthesised before, on task FS1.

4.2. Evaluation of speaker similarity

For the first time in the Blizzard Challenge 2021 (Ling et al., 2021), some synthesised speech was judged statistically closer to the reference speaker than the natural speech itself. In this 2023 edition, this trend has expanded to more than one system. 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 rated closer to the reference speaker than the natural speech, which got a median score of 4 ("probably the same person"). A probable reason for these results, which was reported by some listeners as well as some participating teams to the challenge, is that the four reference stimuli 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 of judging 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 came from his/her voice?

The low scores given to natural speech in both tasks suggest that listeners could not 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. 6 against speaker similarity scores obtained from listeners who have never heard the speaker's voice before. 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 median score 5 ("exactly the same person"). Only System F was given similar scores, that correlate 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 natural speech stimuli 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.

4.3. Evaluation of intelligibility

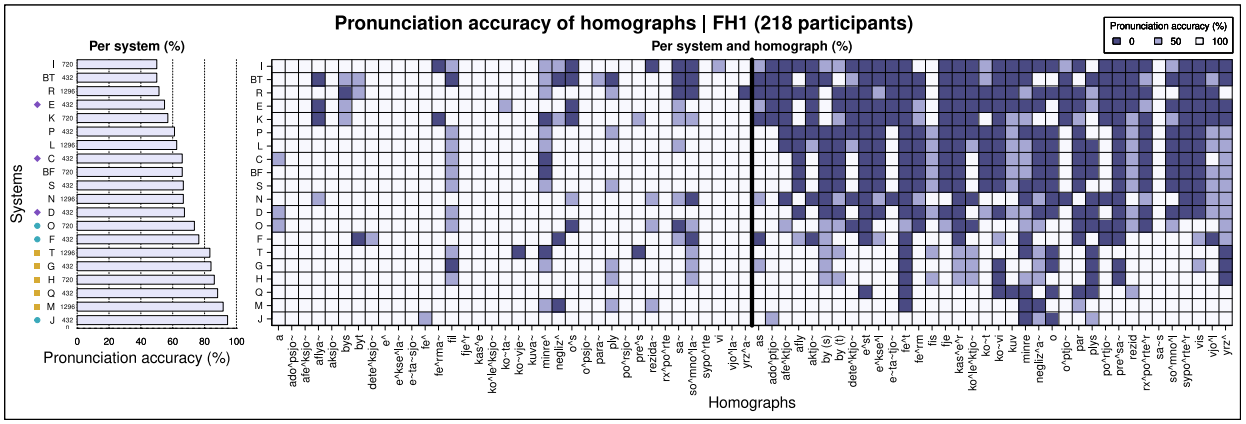


Fig. 7. Intelligibility results for task FH1, with pronunciation accuracy of homographs (Test 3.b), per *system* (left) and homograph (right). Square, diamond, and circle symbols on the left panel display the usage of LLM models as a complementary text input, prosodic control or both, respectively, for each system.

4.3.1. Synthesis of semantically unpredictable sentences

Ten years ago, the Blizzard Challenge 2013 (King and Karaiskos, 2013) reported WER on transcription of synthesised SUS between 20% for the best performing systems, and up to 50%. In the last 2021 edition (Ling et al., 2021), WER was between 12% and 25%. In this 2023 edition, 10 systems (J, F, P, O, Q, H, I, M, S, E) obtained a median WER 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. Only G got a median WER of 33%. Overall, 6 systems were significantly different from the two benchmarks: J, F, P, O, Q were significantly better, and G was significantly lower. While none of the systems have specifically been trained on SUS generation, the excellent score for almost all systems demonstrates that SUS synthesis has dramatically improved across the past years and is globally well handled by most systems. Thus, this test might have reached its limit to make a distinction between the global intelligibility of speech synthesis across systems.

4.3.2. Synthesis of homographs

Although speech synthesis performs well in global evaluations, more targeted testing can reveal remaining problem areas: the evaluation of homographs aims to identify local pronunciation errors. The left part of Fig. 7 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. 7 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 the similar behaviour of the systems with the lowest pronunciation accuracy (I, BT, R, E, K, P, L, C, BF, S, N) which 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 significantly different from each other (J, M, Q, H, G, T) manage to handle both pronunciations. Interestingly, we observe a step in performance between systems D and O which correlates well with the use of a Large Language Model (Devlin et al., 2019) to complement the text input. As indicated by the square and circle symbols in the left of Fig. 7 as well as in Table 2, Systems O, F, T, G, H, Q, M and J used one LLM for letter-to-sound mapping while the other systems did not.

5. Conclusions and perspectives

The Blizzard Challenge 2023 evaluated the synthesis of isolated sentences in French generated from models trained on read speech (audiobooks or extracts of parliament) on two tasks. The French 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). On these three criteria, this challenge has demonstrated that state-of-the-art architectures are now becoming extremely competitive for the synthesis of high-quality isolated sentences, leading us to draw five main perspectives for future Blizzard Challenges, and more generally for the evaluation of speech synthesis:

Speech quality. In this Challenge, two systems generated speech that is not significantly distinguishable from natural speech in a global MOS evaluation, confirming the saturation effect in speech quality observed in the previous Blizzard Challenge edition. Nevertheless, one should *not* conclude that the synthetic speech is 'as good as' or 'indistinguishable from' natural speech in general. With the introduction of a fine-grained comparison between the systems such as MUSHRA, we were able to highlight differences that a global MOS evaluation cannot. To, on the one hand, maintain consistency with previous challenges that allow longitudinal studies such as Le Maguer et al. (2022, 2024), Cooper and Yamagishi (2023), and, on the other hand, allow us to discriminate between the best performing systems, we advise keeping both global MOS and fine-grained MUSHRA evaluations for speech quality in future challenges, as long as not all systems have saturated the listeners MOS. Moreover, full implementation details should keep being provided for full reproducibility of the results.

Speaker similarity. By selecting reference samples that were diverse and representative of the training set, we fortuitously demonstrated that average listeners can only rate two speech samples from the same speaker as "probably the same person" instead of "exactly the same person" in comparison with listeners who were familiar with the speaker and could perfectly recognise speech samples from her voice. Therefore, it seriously calls into question the validity of such a protocol, especially with the growing interest for *expressive* speech synthesis whose purpose is to increase the diversity of speech samples from a single speaker. One key challenge in the evaluation of speech synthesis is to refine the definition of speaker similarity and derive more robust protocols for such subjective test. Moreover, we obtained very similar speech quality and speaker similarity scores for the 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 dataset containing only a few minutes of speech or even zero-shot synthesis.

Intelligibility. Evaluation of intelligibility on SUS has demonstrated excellent results with median error rates of 0 for half of the systems. Therefore, synthesising speech signals that can be transcribed without errors is an issue that is almost solved. In contrast, finer evaluation of homograph pronunciation accuracy displayed less successful results, although the use of large language models is promising as it allowed some systems to reach more than 80% accuracy on the task. In that direction, we suggest that global intelligibility test such as SUS transcription could be dropped for future challenges, to focus instead on the evaluation of local events that still display errors in the synthesis. Since local events are language- or speaker- dependent (e.g., speaker idiosyncrasies in pronunciation), extremely varied (e.g., isolated phone quality, liaisons, tones, prosodic patterns, etc.), and that relevant evaluation protocol are difficult to scale to a large number of stimuli (e.g., assessing pairwise differences, pointing at local errors), there are too many evaluations to conduct for one or even the few challenges to come to cover all of them. For this sake, we suggest that participating teams, in addition to submitting their synthesis samples for the main speech quality evaluation, could also design a test of their own on the local event of their choice, to which all participating teams would be evaluated. This would allow the evaluation of as many local events as participating teams in future Challenge editions.

Evaluation in context. Due to lack of time and resources, this Challenge edition did not address the evaluation of speech synthesis in context (cf. example of applications in Wagner et al. (2019, Table 1)) which has already been discussed and sometimes performed in the most recent literature. Instead, only isolated sentences were evaluated. With the drastic improvement of speech synthesis quality and its growing presence in everyday life as a human-machine interface, the evaluation of the communication success of speech synthesis will certainly receive more focused attention in the years to come. Criteria for communication success are numerous (e.g., putting focus on the right word accordingly to the context, maintaining the listener attention in long form synthesis, synthesising spontaneous speech with appropriate backchannels, indicating turn taking with the right prosody, generating appropriate social attitudes depending on the context and interlocutor, etc.) and each requires dedicated datasets and evaluation protocols that are not yet standardised in the community. We therefore invite people working on these specific communicative intents to disseminate their expertise in their evaluation by getting involved in the organisation of future Blizzard Challenges.

Diversity of speech data. Last but not least, most Blizzard Challenges including 2023 have tackled well-resourced languages as well as clean read speech. A shift towards the synthesis of under-resourced languages, as well as more variable forms of speech (e.g., spontaneous, noisy, child speech, etc.) is very desirable not only to increase the difficulty of the challenge but also to bring the text-to-speech state-of-the-art closer to generating speech that is representative of human communication.

CRediT authorship contribution statement

Olivier Perrotin: Writing – original draft, Validation, Supervision, Software, Project administration, Methodology, Conceptualization. **Brooke Stephenson:** Writing – review & editing, Software, Investigation. **Silvain Gerber:** Formal analysis. **Gérard Bailly:** Resources, Funding acquisition, Data curation, Conceptualization. **Simon King:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has received funding from the BPI project THERADIA (France) 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. Finally, we thank the numerous listeners who participated in the evaluation, and the 18 participating teams without whom the challenge would not exist.

Appendix. Instructions to the participants of the listening tests

A.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.

A.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.

A.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 benchmark, and three or four participating teams' 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.

A.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.

A.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.

Data availability

Links to download related data are given in the manuscript.

References

Bailly, G., 2003. Close shadowing natural versus synthetic speech. Int. J. Speech Technol. 6 (1), 11–19. http://dx.doi.org/10.1023/A:1021091720511.

Bailly, G., Lenglet, M., Perrotin, O., Klabbers, E., 2023. Advocating for text input in multi-speaker text-to-speech systems. In: Proc. ISCA Speech Synthesis Workshop. Grenoble, France, pp. 1–7. http://dx.doi.org/10.21437/SSW.2023-1.

Benoît, C., Grice, M., Hazan, V., 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Commun. 18 (4), 381–392. http://dx.doi.org/10.1016/0167-6393(96)00026-X.

Black, A.W., Tokuda, K., 2005. The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In: Proc. Interspeech. Lisbon, Portugal, pp. 77–80. http://dx.doi.org/10.21437/Interspeech.2005-72.

Boros, T., Dumitrescu, S.D., Mironica, I., Chivereanu, R., 2023. Generative adversarial training for text-to-speech synthesis based on raw phonetic input and explicit prosody modelling. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 69–74. http://dx.doi.org/10.21437/Blizzard.2023-9.

Bu, Y., Zhao, Y., 2023. Xpress: The 10AI speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 119–123. http://dx.doi.org/10.21437/Blizzard.2023-18.

Camp, J., Kenter, T., Finkelstein, L., Clark, R., 2023. MOS vs. AB: Evaluating text-to-speech systems reliably using clustered standard errors. In: Proc. Interspeech. Dublin, Ireland, pp. 1090–1094. http://dx.doi.org/10.21437/Interspeech.2023-2014.

Chen, H., He, M., de Gibson, L.C., Garner, P.N., 2023. The Idiap speech synthesis system for the Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 93–97. http://dx.doi.org/10.21437/Blizzard.2023-13.

Chiang, C.-H., Huang, W.-P., Lee, H.-y., 2023. Why we should report the details in subjective evaluation of TTS more rigorously. In: Proc. Interspeech. Dublin, Ireland, pp. 5551–5555. http://dx.doi.org/10.21437/Interspeech.2023-416.

Clark, R.A.J., Podsiadlo, M., Fraser, M., Mayo, C., King, S., 2007. Statistical analysis of the Blizzard Challenge 2007 listening test results. In: Proc. Blizzard Challenge Workshop. Bonn, Germany, URL: https://www.isca-speech.org/archive/blizzard_2007/clark07_blizzard.html.

Clark, R., Richmond, K., Strom, V., King, S., 2006. Multisyn voice for the Blizzard Challenge 2006. In: Proc. Blizzard Challenge Workshop. Pittsburgh, PA, United States, URL: http://festvox.org/blizzard/bc2006/cstr_blizzard2006.pdf.

Clark, R., Silen, H., Kenter, T., Leith, R., 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. In: Proc. ISCA Speech Synthesis Workshop. Vienna, Austria, pp. 99–104. http://dx.doi.org/10.21437/SSW.2019-18.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., Yamagishi, J., 2023. The VoiceMOS Challenge 2023: Zero-shot subjective speech quality prediction for multiple domains. In: IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, Taipei, Taiwan, pp. 1–7. http://dx.doi.org/10.1109/ASRU57964.2023.10389763.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., Yamagishi, J., 2024. A review on subjective and objective evaluation of synthetic speech. Acoust. Sci. Technol. advpub (e24.12), http://dx.doi.org/10.1250/ast.e24.12.

Cooper, E., Yamagishi, J., 2021. How do voices from past speech synthesis challenges compare today? In: Proc. ISCA Speech Synthesis Workshop. Budapest, Hungary, pp. 183–188. http://dx.doi.org/10.21437/SSW.2021-32.

Cooper, E., Yamagishi, J., 2023. Investigating range-equalizing bias in mean opinion score ratings of synthesized speech. In: Proc. Interspeech. Dublin, Ireland, pp. 1104–1108. http://dx.doi.org/10.21437/Interspeech.2023-1076.

Dall, R., Tomalin, M., Wester, M., Byrne, W., King, S., 2014. Investigating automatic & human filled pause insertion for speech synthesis. In: Proc. Interspeech. Singapore, pp. 51–55. http://dx.doi.org/10.21437/Interspeech.2014-11.

de Kok, I.A., 2013. Listening Heads (Ph.D. thesis). University of Twente, Netherlands.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. 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, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Eberhard, D.M., Simons, G.F., Fennig, C.D. (Eds.), 2023. Ethnologue: Languages of the World, twenty-sixth ed. Dallas, Texas: SIL International, URL: http://www.ethnologue.com.

Fraser, M., King, S., 2007. The Blizzard Challenge 2007. In: Proc. Blizzard Challenge Workshop. Bonn, Germany, URL: http://festvox.org/blizzard/bc2007/blizzard_2007/blz3_001.html.

Govender, A., Wagner, A.E., King, S., 2019. Using pupil dilation to measure cognitive load when listening to text-to-speech in quiet and in noise. In: Proc. Interspeech. Graz, Austria, pp. 1551–1555. http://dx.doi.org/10.21437/Interspeech.2019-1783.

Gutierrez, E., Oplustil-Gallegos, P., Lai, C., 2021. Location, location: Enhancing the evaluation of text-to-speech synthesis using the rapid prosody transcription paradigm. In: Proc. ISCA Speech Synthesis Workshop. Budapest, Hungary, pp. 25–30. http://dx.doi.org/10.21437/SSW.2021-5.

Haeb-Umbach, R., Wagner, P., 2023. How neural network architectures can inform basic research in phonetics - and vice versa. In: Proc. Interspeech. Dublin, Ireland.

Hajj, M.-L., Lenglet, M., Perrotin, O., Bailly, G., 2022. Comparing NLP solutions for the disambiguation of French heterophonic homographs for end-to-end TTS systems. In: Prasanna, S.R.M., Karpov, A., Samudravijaya, K., Agrawal, S.S. (Eds.), Speech and Computer. Springer International Publishing, pp. 265–278. http://dx.doi.org/10.1007/978-3-031-20980-2_23.

Hinterleitner, F., Neitzel, G., Möller, S., Norrenbrock, C., 2011. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In: Proc. Blizzard Challenge Workshop. Turin, Italy, URL: http://festvox.org/blizzard/bc2011/DeutcheTelekom_Blizzard2011.pdf.

Honnet, P.-E., Lazaridis, A., Garner, P.N., Yamagishi, J., 2017. The SIWIS French speech synthesis database – design and recording of that the title has the same capitalisation rules than the other references. quality French database for speech synthesis. Technical Report, Idiap-Internal-RR-06-2017, http://dx.doi.org/10.7488/ds/1705.

Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. Biom. J. 50 (3), 346–363. http://dx.doi.org/10.1002/bimj.200810425.

Huang, W.-C., Violeta, L.P., Liu, S., Shi, J., Toda, T., 2023. The Singing Voice Conversion Challenge 2023. http://dx.doi.org/10.48550/arXiv.2306.14422.

ITU, 1994. A method for subjective performance assessment of the quality of speech voice output devices. Technical Report ITU-T P.85, International Telecommunication Union, URL: https://www.itu.int/rec/T-REC-P.800-199608-I/en.

ITU, 1996. Methods for objective and subjective assessment of quality. Technical Report ITU-T P.800, International Telecommunication Union, URL: https://www.itu.int/rec/T-REC-P.800-199608-I/en.

ITU, 2000. Software tools and audio coding standardization. Technical Report, International Telecommunication Union, URL: https://www.itu.int/rec/T-REC-P.56-201112-I/en.

ITU, 2015. Method for the subjective assessment of intermediate quality level of audio systems. Technical Report ITU-R BS.1534-3, International Telecommunication Union, URL: https://www.itu.int/rec/R-REC-BS.1534.

ITU, 2016. Methods for objective and subjective assessment of speech and video quality. Technical Report ITU-T P.800.1, International Telecommunication Union, URL: https://handle.itu.int/11.1002/1000/12972.

Jiang, Y., Song, K., Yang, F., Xie, L., Meng, M., Ji, Y., Wang, Y., 2023. The Xiaomi-ASLP text-to-speech system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 109–113. http://dx.doi.org/10.21437/Blizzard.2023-16.

Jillings N., M.D., D., R.J., 2015. Web audio evaluation tool: A browser-based listening test environment. In: Sound and Music Computing Conference. SMC, URL: https://github.com/BrechtDeMan/WebAudioEvaluationTool.

Kearns, J., 2014. LibriVox: Free public domain audiobooks. Reference Rev. 28 (1), 7–8. http://dx.doi.org/10.1108/RR-08-2013-0197.

Kim, J., Kong, J., Son, J., 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Meila, M., Zhang, T. (Eds.), In: Proceedings of International Conference on Machine Learning, vol. 139, Virtual, pp. 5530–5540, URL: https://proceedings.mlr.press/v139/kim21f.html.

King, S., Crumlish, J., Martin, A., Wihlborg, L., 2018. The Blizzard Challenge 2018. In: Proc. Blizzard Challenge Workshop. Hyderabad, India, URL: http://festvox.org/blizzard/bc2018/blizzard2018_overview_paper.pdf.

King, S., Karaiskos, V., 2010. The Blizzard Challenge 2010. In: Proc. Blizzard Challenge Workshop. Kansai Science City, Japan, URL: http://festvox.org/blizzard/bc2010/summary_Blizzard2010.pdf.

King, S., Karaiskos, V., 2012. The Blizzard Challenge 2012. In: Proc. Blizzard Challenge Workshop. Portland, OR, USA, URL: http://festvox.org/blizzard/bc2012/summary_Blizzard2012.pdf.

King, S., Karaiskos, V., 2013. The Blizzard Challenge 2013. In: Proc. Blizzard Challenge Workshop. Barcelona, Spain, URL: http://www.festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.

King, S., Karaiskos, V., 2016. The Blizzard Challenge 2016. In: Proc. Blizzard Challenge Workshop. Cupertino, CA, USA, URL: http://festvox.org/blizzard/bc2016/blizzard2016_overview_paper.pdf.

Kirkland, A., Mehta, S., Lameris, H., Henter, G.E., Szekely, E., Gustafson, J., 2023. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In: Proc. ISCA Speech Synthesis Workshop. Grenoble, France, pp. 41–47. http://dx.doi.org/10.21437/SSW.2023-7.

Kong, J., Kim, J., Bae, J., 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In: Neural Information Processing Systems. NIPS, Vancouver, Canada, URL: https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174. http://dx.doi.org/10.2307/2529310.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D., 2020. FlauBERT: Unsupervised language model pre-training for French. In: Proceedings of the Language Resources and Evaluation Conference. LREC, Marseille, France, pp. 2479–2490, URL: https://aclanthology.org/2020.lrec-1.302.pdf.

Le Maguer, S., King, S., Harte, N., 2022. Back to the future: Extending the Blizzard Challenge 2013. In: Proc. Interspeech. Incheon, Korea, pp. 2378–2382. http://dx.doi.org/10.21437/Interspeech.2022-10633.

Le Maguer, S., King, S., Harte, N., 2024. The limits of the mean opinion score for speech synthesis evaluation. Comput. Speech Lang. 84, 101577. http://dx.doi.org/10.1016/j.csl.2023.101577.

Lee, S.-g., Ping, W., Ginsburg, B., Catanzaro, B., Yoon, S., 2023. BigVGAN a universal neural vocoder with large-scale training. In: International Conference on Learning Representations. ICLR, Kigali, Rwanda, URL: https://openreview.net/forum?id=iTtGCMDEzS_.

Lenglet, M., Perrotin, O., Bailly, G., 2021. Impact of segmentation and annotation in French end-to-end synthesis. In: Proc. ISCA Speech Synthesis Workshop. Budapest, Hungary, pp. 13–18. http://dx.doi.org/10.21437/SSW.2021-3.

Lenglet, M., Perrotin, O., Bailly, G., 2023. The GIPSA-lab text-to-speech system for the Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 34–39. http://dx.doi.org/10.21437/Blizzard.2023-3.

Ling, Z.-H., Zhou, X., King, S., 2021. The Blizzard Challenge 2021. In: Proc. Blizzard Challenge Workshop. Online, URL: http://festvox.org/blizzard/bc2021/BC21_ling_zhou_king.pdf.

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z., 2018. The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods. In: Proceedings of Odyssey – the Speaker and Language Recognition Workshop. Les Sables d'Olonne, France, pp. 195–202. http://dx.doi.org/10.21437/Odyssey.2018-28.

Lu, Y., Fu, R., Qi, X., Wen, Z., Tao, J., Yi, J., Wang, T., Ren, Y., Zhang, C., Yang, C., Shi, W., 2023a. The VIBVG speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 103–108. http://dx.doi.org/10.21437/Blizzard.2023-15.

Lu, C., Lee, J., Wen, X., Lou, X., Oh, J., 2023b. The Samsung speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 52–57. http://dx.doi.org/10.21437/Blizzard.2023-6.

Lux, F., Koch, J., Meyer, S., Bott, T., Schauffler, N., Denisov, P., Schweitzer, A., Vu, N.T., 2023. The IMS Toucan system for the Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 40–45. http://dx.doi.org/10.21437/Blizzard.2023-4.

Ma, Q., Liu, W., Yang, Y., Xu, C., Ling, H., Zhong, J., 2023. The SCUT text-to-speech system for the Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 58–62. http://dx.doi.org/10.21437/Blizzard.2023-7.

Malisz, Z., 2023. Realising the potential of modern speech synthesis for prosodic research. In: Proc. Interspeech. Dublin, Ireland.

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B., 2020. CamemBERT: a tasty French language model. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. Online, pp. 7203–7219. http://dx.doi.org/10.18653/v1/2020.acl-main.645.

Merriam-Webster, Last checked: 11/2024. Merriam-Webster.com dictionary, URL: https://www.merriam-webster.com/dictionary/homograph, 2024.

Möbius, B., 2003. Rare events and closed domains: Two delicate concepts in speech synthesis. Int. J. Speech Technol. 6 (1), 57–71. http://dx.doi.org/10.1023/A:1021052023237.

Morlec, Y., Bailly, G., Aubergé, V., 2001. Generating prosodic attitudes in French: Data, model and evaluation. Speech Commun. 33 (4), 357–371. http://dx.doi.org/10.1016/S0167-6393(00)00065-0.

O'Mahony, J., Oplustil-Gallegos, P., Lai, C., King, S., 2021. Factors affecting the evaluation of synthetic speech in context. In: Proc. ISCA Speech Synthesis Workshop. Budapest, Hungary, pp. 148–153. http://dx.doi.org/10.21437/SSW.2021-26.

Pandey, A., Edlund, J., Le Maguer, S., Harte, N., 2023. Listener sensitivity to deviating obstruents in WaveNet. In: Proc. Interspeech. Dublin, Ireland, pp. 1080–1084. http://dx.doi.org/10.21437/Interspeech.2023-1843.

Parmonangan, I.H., Tanaka, H., Sakti, S., Takamichi, S., Nakamura, S., 2019. Speech quality evaluation of synthesized Japanese speech using EEG. In: Proc. Interspeech. Graz, Austria, pp. 1228–1232. http://dx.doi.org/10.21437/Interspeech.2019-2059.

Perrotin, O., Stephenson, B., Gerber, S., Bailly, G., 2023. The Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 1–27. http://dx.doi.org/10.21437/Blizzard.2023-1.

Pisoni, D.B., Manous, L.M., Dedina, M.J., 1987. Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. Comput. Speech Lang. 2 (3), 303–320. http://dx.doi.org/10.1016/0885-2308(87)90014-3.

Prahallad, K., Vadapalli, A., Elluru, N., Mantena, G., Pulugundla, B., Bhaskararao, P., Murthy, H.A., King, S., Karaiskos, V., Black, A.W., 2013. The Blizzard Challenge 2013 – Indian language tasks. In: Proc. Blizzard Challenge Workshop. Barcelona, Spain, URL: http://festvox.org/blizzard/bc2013/blizzard_2013_summary_indian.pdf.

Prahallad, K., Vadapalli, A., Rallabandi, S.K., Kesiraju, S., Murthy, H., Nagarajan, T., Singh, B., T., S., Rao, K.S., Gangashetty, S.V., King, S., Tokuda, K., Black, A.W., 2015. The Blizzard Challenge 2015. In: Proc. Blizzard Challenge Workshop. Berlin, Germany, URL: http://festvox.org/blizzard/bc2015/overview_bc2015.pdf.

Prenger, R., Valle, R., Catanzaro, B., 2019. Waveglow: A flow-based generative network for speech synthesis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, Brighton, UK, pp. 3617–3621. http://dx.doi.org/10.1109/ICASSP.2019.8683143.

Project Gutenberg Literary Archive Foundation, 1971. The Gutemberg project. URL: https://www.gutenberg.org.

Qi, X., Wang, X., Wang, Z., Liu, W., Ding, M., ShuchenShi, 2023. The FruitShell French synthesis system at the Blizzard 2023 Challenge. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 114–118. http://dx.doi.org/10.21437/Blizzard.2023-17.

Raidt, S., Bailly, G., Holm, B., Mixdorff, H., 2004. Automatic generation of prosody: comparing two superpositional systems. In: Proc. Speech Prosody 2004. Nara, Japan, pp. 417–420. http://dx.doi.org/10.21437/SpeechProsody.2004-95.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y., 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In: International Conference on Learning Representations. ICLR, Virtual, URL: https://openreview.net/forum?id=Afb6Nwd6LNEz.

Saget, F., Gaudier, T., Shamsi, M., Tahon, M., 2023. LIUM-TTS entry for Blizzard 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 28–33. http://dx.doi.org/10.21437/Blizzard.2023-2.

Seebauer, F., Kuhlmann, M., Haeb-Umbach, R., Wagner, P., 2023. Re-examining the quality dimensions of synthetic speech. In: Proc. ISCA Speech Synthesis Workshop. Grenoble, France, pp. 34–40. http://dx.doi.org/10.21437/SSW.2023-6.

Shang, Z., Li, X., Shi, P., Hua, H., Zhang, P., 2023. The IOA-ThinkIT system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 124–129. http://dx.doi.org/10.21437/Blizzard.2023-19.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y., Wu, Y., 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP, Calgary, AB, Canada, pp. 4779–4783. http://dx.doi.org/10.1109/ICASSP.2018.8461368.

Shirali-Shahreza, S., Penn, G., 2018. MOS naturalness and the quest for human-like speech. In: IEEE Spoken Language Technology Workshop. SLT, Athens, Greece, pp. 346–352. http://dx.doi.org/10.1109/SLT.2018.8639599.

Shirali-Shahreza, S., Penn, G., 2023. Better replacement for TTS naturalness evaluation. In: Proc. ISCA Speech Synthesis Workshop. Grenoble, France, pp. 197–203. http://dx.doi.org/10.21437/SSW.2023-31.

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

Streijl, R.C., Winkler, S., Hands, D.S., 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Syst. 22 (2), 213–227. http://dx.doi.org/10.1007/s00530-014-0446-1.

Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., Yamagishi, J., 2016. The Voice Conversion Challenge 2016. In: Proc. Interspeech. San Francisco, CA, USA, pp. 1632–1636. http://dx.doi.org/10.21437/Interspeech.2016-1066.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K., 2013. Speech synthesis based on hidden Markov models. In: Proceedings of the IEEE, vol. 101, (no. 5), pp. 1234–1252. http://dx.doi.org/10.1109/JPROC.2013.2251852.

Vasilis, K., King, S., Clark, R.A.J., Mayo, C., 2008. The Blizzard Challenge 2008. In: Proc. Blizzard Challenge Workshop. Brisbane, Australia, URL: http://festvox.org/blizzard/bc2008/summary_Blizzard2008.pdf.

Veaux, C., Maia, R., Papendreou, S., 2023. The DeepZen speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 81–86. http://dx.doi.org/10.21437/Blizzard.2023-11.

Viswanathan, M., Viswanathan, M., 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Comput. Speech Lang. 19 (1), 55–83. http://dx.doi.org/10.1016/j.csl.2003.12.001.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., Voß e, J., 2019. Speech synthesis evaluation — State-of-the-art assessment and suggestion for a novel research program. In: Proc. ISCA Speech Synthesis Workshop. Vienna, Austria, pp. 105–110. http://dx.doi.org/10.21437/SSW.2019-19.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R.J., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: Dy, J., Krause, A. (Eds.), In: Proceedings of the International Conference on Machine Learning, vol. 80, Stockholmsmässan, Stockholm Sweden, pp. 5180–5189, URL: http://proceedings.mlr.press/v80/wang18h.html.

Wester, M., Valentini-Botinhao, C., Henter, G.E., 2015. Are we using enough listeners? no! — an empirically-supported critique of interspeech 2014 TTS evaluations. In: Proc. Interspeech. Dresden, Germany, pp. 3476–3480. http://dx.doi.org/10.21437/Interspeech.2015-689.

Wester, M., Wu, Z., Yamagishi, J., 2016. Analysis of the Voice Conversion Challenge 2016 evaluation results. In: Proc. Interspeech. San Francisco, CA, USA, pp. 1637–1641. http://dx.doi.org/10.21437/Interspeech.2016-1331.

Wu, Z., Xie, Z., King, S., 2019. The Blizzard Challenge 2019. In: Proc. Blizzard Challenge Workshop. Vienna, Austria, URL: http://festvox.org/blizzard/bc2019/blizzard2019_overview_paper.pdf.

Xie, Z., Fang, R., Zhao, M., 2023a. The BIGAI text-to-speech systems for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 98–102. http://dx.doi.org/10.21437/Blizzard.2023-14.

Xie, K., Wu, Y.-C., Xie, F.-L., 2023b. FireRedTTS: The Xiaohongshu speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 87–92. http://dx.doi.org/10.21437/Blizzard.2023-12.

Xu, Z., Zhang, S., Wang, X., Zhang, J., Wei, W., He, L., Zhao, S., 2023. MuLanTTS: The Microsoft speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 46–51. http://dx.doi.org/10.21437/Blizzard.2023-5.

Yasuda, Y., Toda, T., 2023. Analysis of mean opinion scores in subjective evaluation of synthetic speech based on tail probabilities. In: Proc. Interspeech. Dublin, Ireland, pp. 5491–5495. http://dx.doi.org/10.21437/Interspeech.2023-1285.

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z.-H., Toda, T., 2020. Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion—. In: Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge. Shanghai, China, pp. 80–98. http://dx.doi.org/10.21437/VCCBC.2020-14.

Zaïdi, J., Duchêne, C., Seuté, H., Carbonneau, M.-A., 2023. The La Forge speech synthesis system for Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 75–80. http://dx.doi.org/10.21437/Blizzard.2023-10.

Zalkow, F., Sani, P., Fast, M., Bauer, J., Joshaghani, M., Lakshminarayana, K.K., Habets, E.A.P., Dittmar, C., 2023. The AudioLabs system for the Blizzard Challenge 2023. In: Proc. Blizzard Challenge Workshop. Grenoble, France, pp. 63–68. http://dx.doi.org/10.21437/Blizzard.2023-8.

Zhou, X., Ling, Z.-H., King, S., 2020. The Blizzard Challenge 2020. In: Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge. Shanghai, China, pp. 1–18. http://dx.doi.org/10.21437/VCC_BC.2020-1.

Zielinski, S., Rumsey, F., Bech, S.r., 2008. On some biases encountered in modern audio quality listening tests – a review. J. Audio Eng. Soc. 56 (6), 427–451, URL: http://www.aes.org/e-lib/browse.cfm?elib=14393.