The limits of the Mean Opinion Score for speech synthesis evaluation

The limits of the Mean Opinion Score for speech synthesis evaluation

Sébastien Le Maguer[a], Simon King[b], Naomi Harte[a]

[a]*Sigmedia Lab, ADAPT Centre, Department of Electronic and Electrical Engineering, School of Engineering, Trinity College Dublin, Ireland*
[b]*University of Edinburgh, Centre for Speech Technology Research, United Kingdom*

Abstract

The release of WaveNet and Tacotron has forever transformed the speech synthesis landscape. Thanks to these game-changing innovations, the quality of synthetic speech has reached unprecedented levels. However, to measure this leap in quality, an overwhelming majority of studies still rely on the Absolute Category Rating (ACR) protocol and compare systems using its output; the Mean Opinion Score (MOS). This protocol is not without controversy, and as the current state-of-the-art synthesis systems now produce outputs remarkably close to human speech, it is now vital to determine how reliable this score is.

To do so, we conducted a series of four experiments replicating and following the 2013 edition of the Blizzard Challenge. With these experiments, we asked four questions about the MOS: How stable is the MOS of a system across time? How do the scores of lower quality systems influence the MOS of higher quality systems? How does the introduction of modern technologies influence the scores of past systems? How does the MOS of modern technologies evolve in isolation?

The results of our experiments are manyfold. Firstly, we verify the superiority of modern technologies in comparison to historical synthesis. Then, we show that despite its origin as an absolute category rating, MOS is a relative score. While minimal variations are observed during the replication of the 2013-EH2 task, these variations can still lead to different conclusions for the intermediate systems. Our experiments also illustrate the sensitivity of MOS to the presence/absence of lower and higher anchors. Overall, our experiments suggest that we may have reached the end of a cul-de-sac by only evaluating the overall quality with MOS. We must embark on a new road and develop different evaluation protocols better suited to the analysis of modern speech synthesis technologies.

Keywords: Speech Synthesis Evaluation, Absolute Category Rating, Mean Opinion Score, Blizzard Challenge
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

The advent of Deep Learning and the subsequent significant progress in Machine Learning not only led to game-changing innovations in speech synthesis [1, 2], but also democratised the access to this technology. With numerous systems proposed every year, we are now in critical need of adequate methodologies to evaluate and analyse the output of modern speech synthesis systems. A broad census of the submissions dedicated to speech synthesis of INTERSPEECH 2021, INTERSPEECH 2022 and SSW 2021 [3] shows that two-thirds rely on one evaluation paradigm, the Absolute Category Rating (ACR), and its output: the Mean Opinion Score (MOS).

An ACR (or MOS-test) is a listening test protocol proposed in the ITU P.800 recommendation [4]. This protocol implies that, for each step of the evaluation, the subject (or listener) has to rate one given sample using a predetermined Likert scale. The scale defined in the P.800 is composed of 5 labelled points from 1-bad to 5-excellent. We present the ACR in more detail in Section 2, but a key factor of its popularity already stands out: it is easy to implement and is scalable for large subjective evaluation campaigns with several systems to evaluate.

While the popularity of the ACR is unquestionable, using this protocol is not without controversy. [5] presents a comprehensive analysis of the different limitations and problems related to the use of MOS. The authors covered various topics, from audio to video processing. The first issue they point out is the compression of the results due to the use of a Likert scale. This well-known phenomenon has also been observed in speech synthesis evaluation [6,

7, 8]. A more fundamental issue lies in the relative nature of the MOS. Several studies [9, 10] in image and video processing have emphasised the relative nature of MOS. Numerous studies [11, 12, 13], reviews [6, 7, 8] and even recommendations [14] already pointed out this issue for speech quality assessment. The P.910 [15] issued in 1996 was originally proposed to address this issue by the introduction of the Absolute Category Rating with Hidden Reference (ACR-HR) protocol. The ACR-HR differs from the ACR by imposing the presence of a hidden reference and by postprocessing the ratings relatively to the score of this reference. The outcome, the Differential Mean Opinion Score (DMOS)[1], still aims to capture the divergence of magnitude between the evaluated systems. By contrast, Clark et al. [11] recommend considering the MOS as an ordinal score and propose the use of ranking statistical testing to compare the evaluated systems. Rosenberg et al. [12] extend the study from Clark et al. and show an important source of biases is the choice of listeners and utterances. The authors point out that, due to these biases, the significance of observed differences is difficult to estimate. Finally, in [13], the authors consider that the ACR paradigm is fundamentally flawed and suggest discarding the MOS altogether.

Despite all of these controversies, the ACR protocol remains stubbornly predominant in the face of more robust alternatives such as the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [16]. The most likely reason behind the popularity of the ACR is convenience. Unfortunately, this apparent simplicity also leads researchers to neglect adequate design of their evaluation, as demonstrated by Wester et al. [17] in a meta-analysis of the INTERSPEECH 2014 submissions or by Kirkland et al. [3] in a meta-analysis of the submissions from INTERSPEECH 2021, INTERSPEECH 2022 and SSW 2021.

Another consequence of a "simple" score such as MOS is the emergence of automatic MOS predictors using Deep Learning. The results of the 2022 VoiceMOS challenge [18] show these models are now able to predict a MOS at a system level with an impressive accuracy; the Spearman's Ranking Correlation Coefficient (SRCC) most of the submitted systems reached is above 0.9. While this accuracy is impressive, these systems rely on a fundamental hypothesis: MOS is an absolute score and Deep Learning can predict this score.

As ACR remains the most popular evaluation paradigm of synthetic speech, and as the use of automatic MOS predictors will inescapably increase, it is now vital to determine the validity of the conclusions we draw when using an ACR evaluation.

To do so, we propose a series of four experiments extending the 2013 edition of the Blizzard Challenge. Naturalness is poorly defined, but has been admitted but is widely accepted to equate to "human-likeness" [19]. Most evaluations conducted nowadays only consider Naturalness, and we will mainly focus our experiments on this dimension. Each experiment is dedicated to challenging one potential limit of the ACR. The first experiment (Section 4) is an exact repeat of the Blizzard Challenge 2013 evaluation taking place almost a decade after the original run. It aims to challenge the robustness of MOS across a long period of time. The second experiment (Section 5) focuses on only evaluating the top-tier synthesis systems from 2013. The goal is to determine how lower-quality systems influence the MOS of systems producing a better synthesis. The third experiment (Section 6) introduces modern synthesis into the equation. By doing so, our goal is to answer two questions. First, we will determine how competitive the best systems developed in 2013 are with modern speech synthesis technologies. Then, we want to ascertain how stable a MOS given to a system is when potentially better systems are presented. The fourth experiment (Section 7) is the counterpart of the second experiment but for modern technologies. With this last experiment, we investigate how the MOS of modern systems evolves without the presence of 2013 top-tier systems in the evaluation. Before this, we need to contextualise the ACR protocol (Section 2) as well as present the Blizzard Challenge, the 2013 edition, and to then describe the dataset used to conduct our experiments (Section 3). Finally, the last section (Section 8) provides an overall discussion about the use of ACR in the present and future speech synthesis landscape.

2. From MOS to MOS-test for speech synthesis evaluation

In order to study the implications of using the ACR to evaluate synthetic speech, we first need to provide an overview of how this protocol came to be, and its inherent biases. In recent years, automatic MOS prediction has grown rapidly as an application within the speech synthesis evaluation field. As this task strongly relies on the assumed

[1] not to be mistaken with the Degradation Mean Opinion Score (DMOS) result of the Degradation Category Rating (DCR) also described in the P.800 [4]

robustness of the ACR and the associated source of reliable MOS as training data, we also provide an overview of automatic MOS predictors.

2.1. From MOS to the ACR listening test protocol

The use of MOS to evaluate the output quality of a speech technology system is not recent and can be traced to the 1960s [20]. Early papers that reference the use of MOS in speech technology involve applications in speech coding and transmission over the telephone [21, 22]. In the 1980s, early publications in computer-based Text-To-Speech (TTS) relied on this score for the evaluation of the proposed methodologies [23]. As MOS became more widely used in speech technology, several recommendations were published to establish a more rigorous evaluation protocol using MOS as the metric. Two key recommendations are the P.82 [24], published by the CCITT in 1984, and the P.800 [4] published by ITU-T (formerly known as the CCITT) in 1996.

The publication of the recommendation P.82 provided the telecommunication industry with a standard to conduct telephone surveys to evaluate speech transmission quality. The survey is composed of multiple questions that are designed to reveal the user's general impressions of transmission performance. One of these questions (Question 9.0 in Annexe A from [24]) is dedicated to determining the quality of the speech signal during the conversation. For this question, the user is asked to define the quality of the connection using one of four words: poor; fair; good; or excellent. The interviewer collects the answer as a numerical value between 1 and 4. The MOS is then computed as the arithmetic mean of these answers to obtain a quantitative result in order to be able to compare different tests.

The recommendation P.800 [4] was released in 1996 to expand the scope of P.82. The P.800 introduced a new type of evaluation: the listening-opinion test. It differs from the survey evaluation, proposed in the P.82 and reconducted in the P.800, by imposing a common set of speech signals to be evaluated by multiple listeners. The recommended listening-opinion test method is the Absolute Category Rating (ACR) which uses the MOS as its quantitative results. An ACR imposes that the listener must rate each speech sample independently; at each step of the test, only one sample is presented to the listeners. To ensure the validity of the test, the P.800 imposes a set of conditions that must be fulfilled. First, it is stipulated (see section B.2.3 of the P.800) that a degraded reference has to be introduced during the test to ensure the reproducibility of the evaluation in another environment. The duration of the test (between 20 min and 45 min) as well as the duration of the stimuli (between 2 s and 3 s) are also prescribed. Finally, the recommendation recognises the uses of various scales depending on the purpose of the evaluation. To evaluate the listening quality, the proposed rating scale is now composed of 5 levels, with the addition of a lower level (bad) to the quality rating scale introduced in the P.82.

The P.800 was proposed for speech quality evaluation, but it was quickly adopted by the speech synthesis community [25]. Since then, the ACR has become the standard evaluation methodology to compare speech synthesis systems. Now, when researchers report that they conducted a listener test or measured a MOS for a speech synthesis system, they in essence (should) have run an ACR in compliance with P.800. The main departure from the original recommendation relates to the reference signal. While the P.800 explicitly states that a degraded signal should be part of the evaluation, the natural speech signal is generally used when evaluating the quality of synthetic speech. Furthermore, the ACR is not restricted to speech processing evaluation. Image and video processing also relies on this protocol for multimedia application evaluation (the P.910 [15]) or, in the case of broadcasting (the BT.500 [26]). Closer to speech, audio engineering also heavily relies on this protocol [7, 8]. The widespread use of this protocol demonstrates its versatility, and means that experiences from diverse fields can inform our thinking about the limits of the ACR in the context of speech synthesis.

Over the years, limitations and issues of the ACR have become apparent, resulting in additional recommendations being proposed. The most prominent is the BS.1534 [16] or MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA). MUSHRA deviates from ACR on several points. Firstly, instead of a discrete scale, MUSHRA relies on a continuous scale from 0 to 100 divided into 5 equal intervals. The labels of these intervals are the labels used in the ACR. Secondly, in contrast to the ACR which is monadic, MUSHRA requires all the systems to be evaluated at each step. By doing so, it allows the listeners to fine-tune their comparison. Thirdly, a fundamental addition from MUSHRA is the introduction of a mandatory hidden reference and mandatory anchors (low-level and mid-level). The reference sample is not only given to the listener, but is also introduced into the set of evaluated signals. As the evaluated systems are randomised, the position of this reference is therefore hidden to the listener. The goal is to ensure the ability of the listener to detect the signal corresponding to the reference and, therefore, control the validity of the listener's ratings. The anchors are artificially constructed, degraded speech signals which have a crucial role

in the protocol as they aim to stabilise the rating scale. The third revision of MUSHRA recommendation, BS.1534-3 [27] emphasises the importance of these anchors by introducing a dedicated appendix about the selection of optimum anchors.

Finally, in 2016, the P.800.2 [14] was introduced to complete the P.800. The P.800.2 provides a guideline on how to report the results of the Mean Opinion Score. Not only does this recommendation emphasise the relative nature of the MOS (especially in the sections 7 and 8 of the P.800.2), but it also provides the information that needs to be communicated when reporting the results of a MOS test. However, it is important to note that the statistical analysis to be conducted when reporting the results of a MOS test is considered outside of the scope of the P.800.2.

2.2. Biases in listening tests

Bias	**Description**
Scale contraction	Compression of the score distribution **regardless** of the actual range of the evaluated stimuli
Range equalising	Scores span the entire scale **regardless** of the actual range of the evaluated stimuli
Centering	Systematic shift of all scores
Stimulus spacing	Listeners' self-calibrating scale **regardless** of the actual quality difference between the stimuli
Scale non-linearity	Semantics of the labels do not convey the same level of differences

Table 1: Summary of the listening test biases presented in [7] and analyzed in depth in [8].

As the review article of Zieliński et al. [7] demonstrates, listening tests are subject to numerous biases (or systematic errors). Following the classification proposed by Poulton [28] for biases of subjective tests in psychology, [8] completes [7] by exploring five biases widespread in speech quality evaluation using empirical data. These five biases are presented in Table 1. Both [7] and [8] provide an extensive analysis of these biases and their implications. However, as demonstrated by the analyses presented in [8], detecting the influence of one specific effect is a complex task. In the present study, we only consider their manifestation (as presented in Table 1) to raise awareness, as these biases are part of every subjective evaluation. We strongly encourage the reader to refer to these reviews to gain more insights about them. The analyses presented in [8] demonstrate the prevalence of these biases in MUSHRA listening tests, despite the protocol imposing anchors to reduce the impact of such biases. In the same review, Zieliński also indicates that these conclusions are applicable to the ACR. He further notes that the impact of some biases, such as the contraction bias, are amplified due to the nature of the ACR. While these analyses were conducted for speech coding evaluation, their conclusions also apply to speech synthesis evaluation. This is confirmed by Cooper et al. [29] who provides an observation of the range equalising bias in the context of an ACR evaluation of speech synthesis systems.

From these biases, it becomes apparent that MOS is a relative score. This is also clearly stated in the P.800.2 [14], which clarifies the term "absolute" to refer to "the fact that subjects are asked to independently rate each sample". Zieliński et al. [7] notes that for all the previously mentioned biases, the order is preserved. In other words, the MOS should be considered as ordinal values. While both [7] and [8] provide directions to reduce the impact of these biases on the results, none of the solutions fully cancel the impact of these biases. Another source of biases, whose impact is not clear, concerns the expertise of the recruited listeners. The outcome of the experiments conducted by Schinkel et al. [30] demonstrated that, counter-intuitively, expert listeners shift their scale more than experienced or even naïve listeners. Work in [8] completes this analysis by emphasising that, despite the fact that expert listeners are likely to provide more robust ratings, too little research is conducted to determine how listeners are influenced by these biases. The best solution to avoid a misinterpretation remains to strictly follow the recommendations as well as the reporting guidelines such as the P.800.2 [14] or as provided by Wester et al.[17]. In addition, as pointed out in [7], the scale non-linearity makes it impossible for the assessor to determine the exact nature of the differences between the two systems. As we can only consider the ordinal nature of the results to be preserved, it strengthens the argument of Clark et al.[11] to use a Wilcoxon ranking test to conduct the statistical analysis.

Despite the aforementioned literature providing guidelines on conducting listening tests, analysing their outcomes and reporting their results, Wester et al. [17] demonstrate that researchers overlook these recommendations. Any conclusions provided using listening tests which do not conform to these guidelines can lead to misinterpreted results.

Considering the prevalence of studies in speech synthesis improperly conducting their evaluation, again demonstrated by [17] and [3], any claim of advances can effectively be put into question.

Speech synthesis also faces additional biases. Contrary to speech coding, speech synthesis does not aim to reconstruct a given speech signal. In other words, the output of a speech synthesis system can differ from the reference speech signal while still being considered good. This introduces another bias: the reference should, in fact, be considered a high-quality anchor. To our knowledge, no studies have been conducted to determine this anchor's influence on the outcome of the test. A final fundamental bias is introduced due to the orthogonal nature of the synthesis artefacts. A classic example is the difference in nature between unit-selection concatenation artefacts and HMM-synthesis buzziness. The homogenisation of the speech synthesis landscape due to the ubiquitous use of deep learning may reduce this bias. This has yet to be investigated.

2.3. Objective/Instrumental MOS

While a MOS-test might be straightforward to implement, the cost of conducting a subjective evaluation campaign can be prohibitive. Hence researchers have investigated ways to predict the rating of a given sample automatically. We can consider two generations of MOS predictors. The first generation, initiated by Perceptual Evaluation of Speech Quality (PESQ) [31], relies on signal processing techniques to model the listener. A more detailed presentation of these models is available in a dedicated tutorial [32]. However, as these systems were developed for speech coding, they did not lead to convincing results when applied in speech synthesis and voice conversion [33, 34].

The second generation of MOS predictors, initiated by AutoMOS [35], now rely on Deep Learning architectures. Since then, there has been a growing research interest due to the impressive results obtained by these models. In 2022, 22 teams participated in the first MOS prediction challenge - the VoiceMOS Challenge [18]. Most of the submitted predictors obtained a Spearman's Ranking Correlation Coefficient (SRCC) above 0.9 at a system level, showing an impressive correlation between the predicted MOS and the ground-truth score. However, despite using the SRCC as the main metric, the task of MOS prediction is still treated as a regression task. Consequently, biases such as the scale non-linearity are simply ignored during the evaluation of the accuracy of these systems. This is even more problematic for models such as the MOS-SSL baseline [36], which also discard any information related to the listeners during the modelling.

As we have seen previously, biases are prevalent in listening tests. Predicting MOS using models introduces additional biases. Again, in his review [8], Zielínski provides a list of the three biases MOS predictors face:

- **offset effect** the predicted scores are consistently higher/lower than the ground truth;
- **gradient effect** the distribution of the predicted scores has a steeper/shallower slope than the one of the ground truth;
- **non-linear warping effect** the distribution of the predicted scores has a non-linear relationship to the one of the ground truth.

In addition, any combination of these biases should also be kept in mind when using an automatic MOS predictor. These biases were reported for signal-processing based MOS predictors, before the break-through of Deep Learning based MOS predictors. The analysis of the influences of these biases on these new models would be a worthwhile task.

The rapid growth of research on MOS predictors and the appeal of using these models to replace subjective evaluation has already led to their extensive adoption. In addition to this use case, researchers have also been investigating the use of MOS predictors as a loss measure in the training of "perceptually-guided" speech synthesis models [37]. Yet these predictors are trained using data encompassing biases that are not fully understood and are known to not generalise well [18]. While the prediction accuracy of these models is quantitatively impressive, it is fundamental to analyze the relevance of their predictions. In other words, we urgently need to determine the limitations of the conclusions obtained using an ACR test. Otherwise automatic MOS predictors may unwittingly exacerbate our problems by reliably predicting an unreliable score. But before we dive into our experiments, we first need to introduce the dataset we used.

3. The Blizzard Challenge and the 2013 Edition

In order to conduct our analysis, we need a dataset with evaluated samples that are obtained in a strictly controlled manner. This implies that the dataset should be composed of only one speaker recorded in a professional environment and utterances composed of only one dedicated linguistic domain. This is necessary to ensure that the variations present in the corpus do not bias our experiments' conclusions. In addition, as we aim to determine the consistency of a MOS across time, we need to rely on a dataset covering a sufficiently long time span to reach meaningful conclusions.

The only dataset covering such a time span was proposed by Cooper et al. [38]. This dataset was composed by pooling utterances from different editions of the Voice Conversion Challenge and the Blizzard Challenge that have been taking place during the previous 15 years. The authors then conducted a large-scale evaluation campaign to rate the different stimuli. While this data is ideal as a corpus for MOS prediction systems, it is too heterogeneous to conduct our analysis: this dataset consists of utterances of different linguistic content (e.g. audiobook for adult or children, news) and samples of different target speakers (e.g. male or female, US English or British English).

As a result, we decided to design our own dataset by taking advantage of the Blizzard Challenge [39, 40] and, in particular, the resources provided for the 2013 edition [41]. We consider this is the ideal dataset to investigate the reliability of the MOS and, to support this point, we first need to present a brief history of the Blizzard Challenge. We then describe the dataset produced by the organisers for Blizzard 2013.

3.1. A brief history of the Blizzard Challenge

The Blizzard Challenge [39, 40] is an annual speech synthesis challenge, initiated in 2005. The challenge aims to compare different speech synthesis systems in a uniform setup. Participants are first all provided with the same training dataset. Then, after a couple of months, the text of a set of utterances to be synthesised is released to the participants who have one week to submit their samples. Participants then submit the synthesis samples, which are evaluated in a large-scale subjective evaluation campaign. Since 2005, the main evaluated aspects are the similarity to the speaker of the training dataset, the naturalness and the intelligibility. The similarity and the naturalness have been investigated using an ACR test; the evaluation of the intelligibility has been conducted using a Semantically Unpredictable Sentences (SUS) protocol [42]. To evaluate the similarity, listeners were provided a set of 4 speech samples from the original speaker to be considered as reference. The task for the listener was to determine how the tested sample compare to these references. For the naturalness, each sample was evaluated in isolation without reference. This is similar to the ACR defined in the P.800, but the question (presented in Table 4 in the row "Naturalness") and the labels of the rating scale were modified to evaluate the naturalness of the speech sample instead of its overall quality. While more dimensions, such as the acceptance or the emotion appropriateness, have been investigated for alternative horizons [43] (e.g. paragraph), participants still mainly refer to the MOS at a sentence level.

Since 2005, the Blizzard Challenge has provided a diverse set of speech corpora for different languages, accents and even different domains (e.g. audiobooks, broadcasts). Over nearly 20 years, various teams have submitted samples of numerous systems from state of the art implementations of the synthesis families of the day. As a result, the organisers have built up a set of ideal corpora to investigate the evolution of speech synthesis technologies.

Among these corpora, the one provided in 2013 [41] stands out. The corpus used in 2013 is in English, composed of audiobooks spoken by an American professional voice female actor, and the recording was also conducted in a professional setup. These audiobooks have a more controlled prosody than the corpora provided for the editions between 2016 and 2018. The Blizzard Challenge 2013 is the last edition in "controlled" English, but it took place before the publication of WaveNet [1] (2016) and Tacotron [2] (2017). These publications radically changed the way speech synthesisers are designed. The overwhelming majority of speech synthesisers are now derived from these technologies. In the spirit of studies such as [44], extending the 2013 edition to include these systems can provide a valuable dataset to investigate how modern technologies manage to capture key aspects of the speech signal that earlier methods (e.g. HMM-Based synthesis) failed to.

Another important argument in favour of the 2013 edition is that its training corpus is the most downloaded by researchers (2011 and 2013 each have around 550 licenses issued in 2022). Compared to the 2011 corpus [45], whose duration was 16.6h, the 2013 edition proposed a corpus of at least 19h of uncompressed WAV files and up to 300h of MP3 compressed speech signals.

3.2. The 2013 edition and the task 2013-EH2

The 2013 edition comprises two tasks for English [41]: 2013-EH1 and 2013-EH2. The organisers provided an unsegmented dataset of 300h of speech for 2013-EH1, whereas for 2013-EH2, there was a sentence-level segmented subset of 19 h of speech with corresponding transcriptions. For 2013-EH1, most audio files are only available in MP3 compressed form. For 2013-EH2, the audio files are uncompressed WAV files sampled at 44.1 kHz. We chose to use 2013-EH2 because uncompressed audio is available, and segmentation into sentences is provided. The resulting corpus comprises 9733 utterances whose durations vary between 0.5 s and 35.3 s. [41] already presents an extensive analysis of the results for 2013-EH1 as it was the main task of the challenge. As we choose to use 2013-EH2, we need to present a brief analysis of the results for this specific task. We restrict our analysis to the naturalness evaluation as it is the focus of our paper.

In 2013, 14 systems were submitted to the 2013-EH2 task from three synthesis families: Unit-Selection, Parametrical Hidden Markov Model (HMM) and Hybrid Synthesis. While the natural voice is sampled at 44.1 kHz, the majority of participants submitted signals sampled at 16 kHz. The only exception is the signals submitted by team L which are sampled 44.1 kHz. The evaluation is composed of 9 sections, including 4 dedicated to naturalness (2 about novel sentences, 2 about news sentences). During the evaluation campaign, each listener rated only one sample per system for each section. Therefore, for the two sections dedicated to the novel sentences, listeners rated 15 sentences including one sentence of the natural voice. As there are no samples for the natural voice for the news sentences, only 14 sentences were evaluated for these two sections.

Three categories of listeners took part in the evaluation campaign: onsite recruited listeners (labelled EE in [41]), online recruited listeners (labelled ER) and speech technology experts (labelled ES). By taking into account only the listeners who completed the sections and are native English speakers, we obtain the following distribution: 61 EE listeners, 18 ER listeners and only 4 ES listeners. In addition, the listeners labelled EE were recruited by the challenge organisers, and the evaluation was run in person at the University of Edinburgh's perceptual testing laboratory. Therefore, we compare the results of our experiments to the ratings of the listeners labelled EE as they are less impacted by uncontrollable external biases.

The overall results for the evaluation of naturalness are presented in a standard boxplot (Figure 1) and in a summary table (Table 2a). As recommended in [11], a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests were conducted to determine the significance of the difference between systems. Table 2b shows the significant differences between the submitted systems with $\alpha = 0.01$.

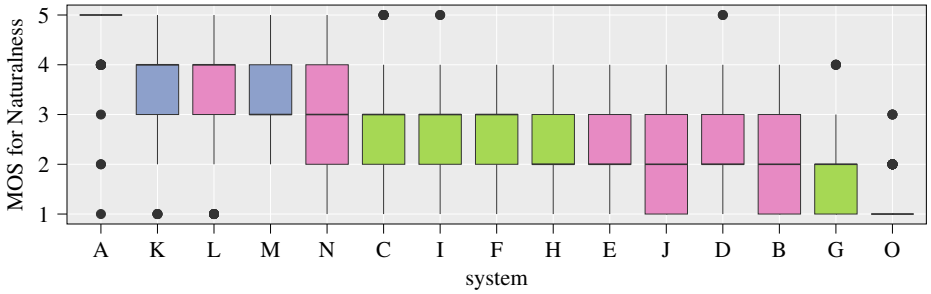


Figure 1: Listening test results related to the naturalness of the task 2013-EH2 for the group of listeners EE. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS. Each color represents a family of synthesis: ■ = Hybrid synthesis, ■ = Unit Selection synthesis, ■ = HMM-based synthesis. **System A** corresponds to the natural voice and **system O** is part of the HMM-based synthesis family.

From these results, we can detect five main clusters. First, the natural voice is perceived as more natural than any synthetic voice. The most natural synthetic voices are produced by systems K, M (both Hybrid synthesis) and system L (Unit-Selection). The second block comprises system N (Unit-Selection) and most of HMM voices (C - the HMM baseline [46], I, F, H). The next block is represented by unit selection voices (E, J, D, B - The Festival baseline [47]) followed by system G (HMM synthesis). Finally, system O (HMM Synthesis) produces the synthetic voice perceived as the least natural.

These results for 2013-EH2 are almost identical to those obtained for 2013-EH1. The only difference of note

system	mean	std	median	mad	min	max	count
A	4.80	0.61	5.00	0.00	1.00	5.00	124
K	3.60	1.00	4.00	1.00	1.00	5.00	124
L	3.35	1.03	4.00	1.00	1.00	5.00	124
M	3.33	0.93	3.00	1.00	1.00	5.00	124
N	2.90	0.99	3.00	1.00	1.00	5.00	124
C	2.75	1.04	3.00	1.00	1.00	5.00	124
I	2.68	0.96	3.00	1.00	1.00	5.00	124
F	2.56	1.05	3.00	1.00	1.00	5.00	124
H	2.35	0.97	2.00	1.00	1.00	4.00	124
E	2.27	0.96	2.00	1.00	1.00	4.00	124
J	2.25	0.98	2.00	1.00	1.00	4.00	124
D	2.24	0.97	2.00	1.00	1.00	5.00	124
B	2.13	0.95	2.00	1.00	1.00	4.00	124
G	1.78	0.83	2.00	1.00	1.00	4.00	124
O	1.20	0.46	1.00	0.00	1.00	3.00	124

(a) Overall statistics

	A	K	L	M	N	C	I	F	H	E	J	D	B	G	O
A	NA														
K	■	NA													
L	■		NA												
M	■			NA											
N	■	■			NA										
C	■	■	■	■		NA									
I	■	■	■	■			NA								
F	■	■	■	■				NA							
H	■	■	■	■					NA						
E	■	■	■	■	■					NA					
J	■	■	■	■	■						NA				
D	■	■	■	■	■							NA			
B	■	■	■	■	■	■	■						NA		
G	■	■	■	■	■	■	■	■	■	■				NA	
O	■	■	■	■	■	■	■	■	■	■	■	■	■	■	NA

(b) Significance test results

Table 2: Statistical summary of the results related to the naturalness of the task 2013-EH2 for the group of listeners EE. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. The significance test consists of a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests. Each cell marked with ■ indicates that the two systems are considered different with a p-value < 0.01; The systems are ordered using the mean MOS.

relates to system M. In 2013-EH1, system M (median MOS=4, average MOS=3.9) was considered significantly more natural than system K (median MOS=3, average MOS=2.8). In 2013-EH2, system M is at the same level as system K. This can be explained by the difference of sampling rates used by team M. While the submission of this team is sampled at 48 kHz for 2013-EH1, the sampling rate they used for 2013-EH2 is 16 kHz.

Nonetheless, the outcome of 2013-EH2 leads to identical conclusions to the outcome of 2013-EH1: the natural voice is the most natural; hybrid synthesis produces synthetic speech perceived as the most natural; there is an important variability between the systems of the other synthesis paradigms. In the context of our paper, the 2013-EH2 task provides an ideally controlled dataset to determine how reliable the conclusions of an ACR evaluation are. As a first experiment, we determine how the scores obtained in 2013 hold a decade later. In [48], the authors analyzed the difference of influence between the novel and the news sentences, and the results show it was marginal. Therefore, we only consider the novel sentences for our analysis.

4. Experiment 1 - Repeating 2013-EH2

By definition, the scores obtained using an ACR should remain steady. In other words, even if time has passed during two evaluation campaigns, the MOS obtained during these two campaigns should remain identical. In order to test this assumption, we conducted an experiment which consists of a repeat of the 2013-EH2 with limited adaptations.

4.1. Evaluation protocol

Our goal is to remain as close as possible to the protocol used in the original run of the challenge [41]. This protocol was comprised of 9 sections which focused on assessing the speaker similarity (Section 1 of the test), the naturalness at a sentence level (Sections 2 to 5), multiple dimensions - defined in [43] - at a paragraph level (Sections 6 and 7), and the intelligibility of SUS (Sections 8 and 9). It is important to note that Section 3 and Section 5 are dedicated to evaluating news sentences for which no natural samples are available.

In this experiment, we focus on the naturalness at the sentence level. As a result, we only included the first five sections for our repeat experiment. During the original campaign, when listeners evaluated the naturalness (Sections 2, 3, 4 and 5) they had already become familiar with the system variability as they heard one sentence of each system in Section 1. As a result, while Section 1 is not the core of our study, we included it to avoid creating biases during our evaluation run. For the evaluation conducted in 2013, listeners took the test in a controlled environment. In our study, we rely on a CrowdMOS [49] setup where listeners take the test online.

In addition to the actual evaluation, each listener had to fill pre- and post-test questionnaires. For the pre-test questionnaire, the listeners were asked about their age, accent (identified as “dialect” in the questionnaire), and listening conditions. This last part is even more important in our case as it helps us to post-analyze the submissions. This

ensures that the listening conditions of our experiment are comparable to those of the reference group (i.e., listeners labelled EE) from the original run. For the post-test questionnaire, the listeners were asked about their familiarity with speech technologies and the Blizzard Challenge. The questions were identical to those from the original edition. The only difference was that the questionnaire was split in pre-test and post-test ones to avoid biasing the listener about their task.

We recruited listeners via Prolific [50]. We also pre-screened participants to balance their birth sex and English accent (UK/US). A location filter was also applied to ensure the self-declared English accent balancing. Each participant was paid £4; a rate suggested as good remuneration by Prolific.

4.2. Summary of the listeners

Among the 60 recruited native English listeners, 30 indicated being American speakers, 29 being British speakers, and one native Indian speaker. We discarded this last listener during our analysis to avoid introducing biases. Nevertheless, the final set of listeners remains gender balanced with 30 males and 29 females.

The majority of listeners reported a limited familiarity with TTS technologies. From the pool of 59 listeners, only 16 report using TTS technologies more than once a week. 10 reported having never used TTS technologies at the time of the evaluation campaign. The majority of listeners (45 out of 59) reported being under 40 years old with 19 being under 30 years old. As a result, we can consider our pool of listeners as naïve and, therefore, equivalent to the ones who participated in the original campaign.

4.3. Results

The overall results are presented in a standard boxplot (Figure 2) and in a summary table (Table 3a). Following the Blizzard Challenge guidelines, a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests were conducted to determine how significant the differences between the systems are. The results of these tests are presented in Table 3b. In addition, to follow the guidelines provided by [8], we also describe the deviation by reporting two additional metrics. The first metric is the deviation as a percentage of the 5-point scale. We identify this metric using R_d. We also report Cohen's d, identified as d, as [8] points out that other disciplines (e.g. psychology) require a report of the effects size normalised to a pooled variance.

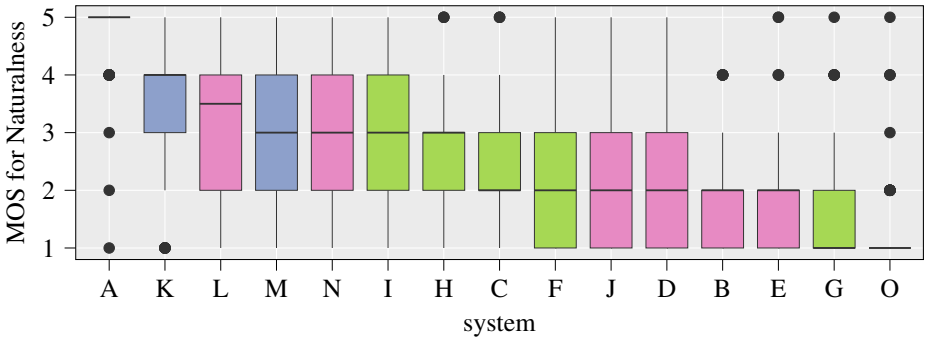


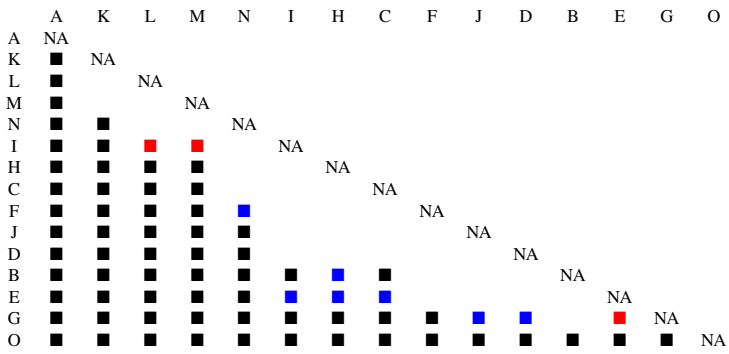
Figure 2: Listening test results for the repeat experiment. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS. Each color represents a family of synthesis: ■ = Hybrid synthesis, ■ = Unit Selection synthesis, ■ = HMM-based synthesis. **System A** corresponds to the natural voice and **system O** is part of the HMM-based synthesis family.

Overall, our results are equivalent to those obtained during the original run and presented in Figure 2. The natural voice is perceived as the most natural. The hybrid systems K and M, as well as system L (Unit Selection) are still producing synthetic the voices as the most natural. There is an important variability in terms of perceived naturalness between synthesis produced by HMM and Unit-Selection systems. System O produces the synthesis which is perceived as the least natural. In addition, the average MOS and the median MOS for most of the systems remain equivalent.

Nonetheless, when comparing the results of the present experiment to the ones of the original run, we can note some subtle differences that affect the interpretation of the results. First, in 2013, systems I and C were considered

system	mean	std	median	mad	min	max	count
A	4.81	0.57	5.00	0.00	1.00	5.00	118
K	3.70	1.14	4.00	1.00	1.00	5.00	118
L	3.28	1.08	3.50	0.50	1.00	5.00	118
M	3.27	1.06	3.00	1.00	1.00	5.00	118
N	3.03	1.12	3.00	1.00	1.00	5.00	118
I	2.81	1.03	3.00	1.00	1.00	5.00	118
H	2.57	1.07	3.00	1.00	1.00	5.00	118
C	2.47	1.04	2.00	1.00	1.00	5.00	118
F	2.43	1.17	2.00	1.00	1.00	5.00	118
J	2.37	1.18	2.00	1.00	1.00	5.00	118
D	2.26	1.02	2.00	1.00	1.00	5.00	118
B	1.92	0.86	2.00	1.00	1.00	4.00	118
E	1.86	0.88	2.00	1.00	1.00	5.00	118
G	1.64	0.90	1.00	0.00	1.00	5.00	118
O	1.22	0.67	1.00	0.00	1.00	5.00	118

(a) Overall statistics)



(b) Significance test results)

Table 3: Statistical summary of the results for the repeat experiment. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. The significance test consists of a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests with $\alpha = 0.05$. ■ indicates a significant difference only for the repeat; ■ indicates a significant difference only for the original campaign; ■ indicates a significant difference both for the original campaign and repeat. The systems are ordered using the mean MOS.

equivalent. In the repeat, we see a trend of these systems going in opposite directions. The average MOS of system I is now 2.81 in comparison to 2.68 in 2013; the median MOS remains at 3. The relative difference R_d for system I is $-2.5\,\%$ and the corresponding Cohen's d is -0.12 The average MOS of system C is now 2.47 in comparison to 2.75 in 2013; the median MOS here drops to 2. The corresponding relative difference R_d is $5.5\,\%$ and the Cohen's d is 0.26. Despite these minor differences, the outcome of the Wilcoxon tests shows that system I is now considered only different from the natural voice and system K. Yet, systems I and C are still not considered significantly different. Another difference is related to system E. Its average MOS drops from 2.27 to 1.86 ($R_d = 8.2\,\%$, $d = 0.43$) but its median remains constant. As a result, System E is perceived as less natural than systems C, I and H for Experiment 1 while, for the original edition, these systems were not considered significantly different. Conversely, the average MOS of system G drops from 1.78 to 1.64 ($R_d = 2.93\,\%$, $d = 0.16$). Again, despite a limited difference, systems E and G are not considered significantly different anymore. Furthermore, system G is now significantly perceived as less natural than systems J and D.

These subtle variations can be due to a less controlled environment during our evaluation. Our campaign was conducted online, whereas our reference listener group took the test in a sound booth. As we have seen in Section 3.2, the group ER contains the listeners recruited online for the original campaign. By selecting the native listeners from this group, we can obtain a cohort of listeners equivalent to the one of our campaign. The number of listeners is too low to obtain statistically valid results, but reproducing the analysis with this cohort can provide an idea of the influence of the environment on our results. The results obtained for the group of listeners ER are equivalent to the ones obtained for the group EE in 2013. This suggests that these variations are not due to the differences between the testing environments.

The overall trend seems reassuring - the broad conclusions obtained in 2013 still hold, with the ranking between systems remaining stable. Nonetheless, among the systems of intermediate quality, we observed variations which are likely the manifestation of the biases presented in Section 2.2. With an overall protocol such as the ACR, the exact cause of these variations cannot be explained. Consequently, it is imperative to precisely describe the evaluation protocol as emphasised in [17].

Experiment 1 - Summary

- Overall ranking results hold after 10 years
- Variations in the statistical significance for intermediate systems which can impact the conclusions

5. Experiment 2 - Focus on the top-tier systems of 2013-EH2

From the results of Experiment 1, the synthesis produced by system O is considered significantly less natural than other submissions. If we consider the stimulus spacing bias, such a low quality (median MOS=1, average MOS=1.22) is likely to influence the scores of the other systems by leading them to have, artificially, a higher MOS. This second experiment aims to determine how the scores of top-tier systems will evolve by removing lower tier systems. Focusing the analysis on the top-tier systems also allows us to investigate the use of different scores for different purposes. By doing so, we aim to determine how other dimensions relate to naturalness.

5.1. Selected systems

As we have seen in the previous section, the evaluation covered all synthesis families available in 2013: HMM-based, Unit-Selection and Hybrid synthesis. For each family, several systems of different qualities were evaluated leading to diverse MOS per family. To focus our analysis and explore the robustness of the ACR in more detail, we select a reduced number of systems satisfying the following criteria: only keep top-tier systems; one representative system per family; avoid additional variables in our analysis.

Following these criteria, system K is selected as the representative of the hybrid synthesis family. While system K is not considered significantly different from system M, the submissions for the latter used different sampling rates for 2013-EH1 (48 kHz) and 2013-EH2 (16 kHz). Similarly, submission L is sampled at 44.1 kHz. As a result, we select system N as the Unit Selection synthesis representative. Systems N and L are not considered significantly different despite their average MOS difference. For HMM synthesis, the conclusions of the previous experiment point to system I being the ideal candidate. Yet, we propose to use system C as the representative for HMM synthesis. Systems I and C are not considered significantly different, and system C is the HTS baseline [46]. The implementation of system C is freely available[2] and was widely used by the community at the time. In order to identify which family the system represents, we have added the family to the system identifiers. System K is now identified as K-Hybrid, system N as N-US and system C as C-HMM. Finally, to avoid introducing an unnecessary variable during the analysis, we downsampled the natural speech to 16 kHz.

5.2. Evaluation protocol

For this experiment, we evaluate far less systems than for Experiment 1. We propose to take advantage of this opportunity to conduct a more in-depth analysis of ACR. To do so, each listener rated all sentences synthesised by all systems in the reduced set. Furthermore, at each step, the listener was instructed to rate the sample across five dimensions: the Overall Quality defined in the recommendation P.800 [4]; the Naturalness as defined in the Blizzard Challenge evaluation protocol [41]; the Listening Effort, the Pronunciation as defined in the questionnaire I from the P.85 [51]; and the Voice Pleasantness extracted from questionnaire Q from the P.85 as well. The exact questions submitted to the listener are presented in Table 4.

Dimension	Question	Reference
Overall Quality	How do you rate the quality of the sound of what you have just heard?	P.800 [4]
Naturalness	Choose a score for how **natural** or **unnatural** the sentence ***sounded***.	Blizzard [41]
Listening Effort	How would describe the effort your were required to make in order to understand the message?	P.85 (I) [51]
Pronunciation	Did you notice any anomalies in pronunciation?	P.85 (Q) [51]
Voice Pleasantness	How would you describe the voice?	P.85 (Q) [51]

Table 4: Questions proposed for each dimension to the listener during the evaluation

While the Overall Quality and the Naturalness are fairly standard evaluated dimensions, the others are less frequently used. The Listening Effort and the Pronunciation dimensions are targeting the intelligibility or the comprehensibility of the synthetic signal. As Pommée et al. [52] demonstrates, the comprehensibility relates to the reconstruction

[2] http://hts.sp.nitech.ac.jp

of the meaning of the message conveyed by the signal. We focus our analysis on isolated sentences, so we consider that the comprehensibility remains too high-level and that only an indication of the intelligibility of the message is measured by these tow dimensions. The Listening Effort and the Pronunciation dimensions also have the advantage of providing outcomes which are relatively easy to interpret. More subjective than the previous dimensions, the Voice Pleasantness is of particular importance as it provides information about the pleasure of the listening experience of the evaluated voice [53]. While some studies [54] are available, this dimension remains under-investigated.

To satisfy the duration constraints prescribed in recommendation P.800 [4], we selected 8 utterances for each section to be evaluated. The core of the test therefore comprises two sections of 32 samples. For each section, the samples are randomized to ensure that there are no biases due to the presentation order.

As for Experiment 1, listeners were recruited via Prolific [50] using the same pre-screen filters (i.e., balance on birth-sex and English accent). We also collected demographic information from the same pre- and post-test questionnaires used in Experiment 1.

5.3. Summary of the listeners

A total of 76 listeners were recruited. Among these listeners, and despite being explicitly instructed to do so, 13 didn't use headphones. These users are discarded from the present analysis. Of the 63 remaining listeners, 30 are native US English speakers, 31 are native British English speakers, one is a native Irish English speaker and one of an unknown accent. The two last listeners are excluded from the analysis. Therefore, our analysis is based on 61 listeners' responses, of which 30 are from females and 31 are from males. As for Experiment 1, the strong majority of the listeners are under 40 years old (43 out of 61) and only 18 listeners reported using speech technologies more than once per week.

5.4. Results of Experiment 2

As for Experiment 1, the overall results are presented in a standard boxplot (Figure 3) and in a summary table (Table 5a). Similarly, a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests were conducted to determine how significant the differences between the systems are. The results of these tests are presented in Table 5b.

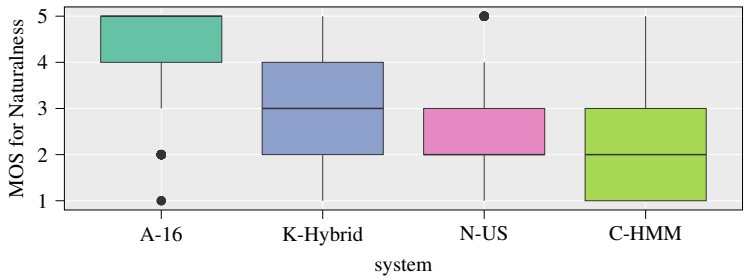


Figure 3: Listening test results for the experiment focusing on the top-tier systems of 2013. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS. Each color represents a family of synthesis: ■ = natural voice, ■ = Hybrid synthesis, ■ = Unit Selection synthesis, ■ = HMM-based synthesis.

system	mean	std	median	mad	min	max	count
A-16	4.43	0.70	5.00	0.00	1.00	5.00	945
K-Hybrid	2.99	1.00	3.00	1.00	1.00	5.00	943
N-US	2.38	0.96	2.00	1.00	1.00	5.00	944
C-HMM	2.21	0.97	2.00	1.00	1.00	5.00	944

(a) Overall statistics

	A-16	K-Hybrid	N-US	C-HMM
A-16	NA			
K-Hybrid	■	NA		
N-US	■	■	NA	
C-HMM	■	■	■	NA

(b) Wilxocon results

Table 5: Statistical summary of the results for the experiment focusing on the top-tier systems of 2013. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. The significance test consists of a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests with $\alpha = 0.05$. The systems are ordered using the mean MOS.

As we first focus on the extension results, we can see that the natural voice (A-16) is considered the best, with a median MOS of 5 and an average MOS of 4.45. The significance tests confirm that A-16 is perceived as more natural. The synthesis systems are rated substantially below as system K-Hybrid obtained a median MOS of 3, while the median MOS for both systems N-US and C-HMM is 2. Again, the significance tests confirm this order with the following additional information: N-US now significantly differs from C-HMM.

From Table 5a, we see that downsampling the natural recorded speech led to a decrease of the average MOS from 4.8 to 4.45. However, the median and the order remain identical. This is reassuring: the natural recorded speech is still considered the most natural.

From the same table, the most striking point is the drop in MOS (both the average and the median) for all the synthesis systems when comparing their scores to the 2013 listening test with the present focused repeat. Specifically, system K-Hybrid average MOS drops from 3.60 to 3.03 ($R_d = 7.7\,\%$, $d = 0.37$); system N-US average MOS from 2.9 to 2.39 ($R_d = 4.6\,\%$, $d = 0.21$) and system C-HMM average MOS from 2.75 to 2.24 ($R_d = -6.8\,\%$, $d = -0.3$). The same stimuli were presented to listeners in both listening tests, yet the MOS scores given by the listeners to each and every one of these systems dropped up to 0.6. Despite this drop, the values of the Cohen's d coefficients indicate a moderate effect. Yet, systems N-US and C-HMM are now considered significantly different. When we refined the analysis, we observed that only the native American English Male speakers caused the significance of the distinction between the systems N-US and C-HMM. The series of Bonferroni-corrected Wilcoxon tests show no significant differences between systems for any other combination of gender and accent. When investigating the other scores (Overall Quality, Voice Pleasantness), we observed the same pattern: the difference between systems N-US and C-HMM is only statistically significant for native American English Male speakers. As MOS is an overall score, it is difficult to determine the underlying reasons for this difference, and [55] pointed out that the interpretation of small variations between gender should be taken with caution. Determining the underlying reasons for this difference is beyond the scope of this article, but it remains important to note that the demographic of the listeners also introduces a relative element to the ACR.

These results demonstrate the relative nature of the MOS obtained ACR. In opposition to other protocols such as MUSHRA [16], ACR does not impose the use of anchors. However, the listeners are likely to base their decisions on samples previously rated during the test. In other words, the listeners are likely to select their own anchors. By comparing the results of Experiment 1 and Experiment 2, we can hypothesize that system O was used as the lower-quality anchor by the listeners.

	Overall		List. Effort		Pronunciation		Voice Pleas.		Naturalness	
system	M	Mdn	M	Mdn	M	Mdn	M	Mdn	M	Mdn
A-16	4.41	5.00	4.50	5.00	4.67	5.00	4.16	4.00	4.43	0.70
K-Hybrid	3.44	4.00	3.63	4.00	3.52	4.00	3.04	3.00	2.99	1.00
N-US	2.96	3.00	3.14	3.00	2.97	3.00	2.59	3.00	2.38	0.96
C-HMM	2.92	3.00	3.41	4.00	3.11	3.00	2.46	2.00	2.21	0.97

Table 6: Statistics for all the dimensions for Experiment 2. "Mdn" stands for *Median*. "M" stands for *mean*.

Results for all the dimensions presented in Section 5.2 (Overall Quality, Listening Effort, Pronunciation and Voice Pleasantness) are presented in Table 6. As for the Naturalness, we also tested the significance of the difference using a series of Bonferroni-corrected pairwise Wilcoxon signed-rank tests. As expected, for a majority of these dimensions, the trends are identical to the one of Naturalness: The natural voice is rated higher than the synthesis system K-Hybrid, followed by N-US, and, finally, the system C-HMM. The outcome of the significance testing validates this outcome, but the testing also shows that systems N-US and C-HMM are not significantly different for the Overall Quality and the Pronunciation. The synthesis family of these two systems differs: system N-US implements Unit Selection synthesis, and system C-HMM relies on the HMM-based synthesis paradigm. As a result, the synthesis of these systems has issues of a different nature (buzziness vs concatenation artefacts). Due to the general nature of MOS, we cannot speculate as to the exact impact of these artefacts on the listeners' perception. Nonetheless, the listeners considered that these artefacts have a similarly negative impact on the synthesis quality.

The Listening Effort does not follow this pattern. For this dimension, system C-HMM is considered significantly

above system N-US. System C-HMM is actually considered equivalent to system K-Hybrid. The Listening Effort and the Pronunciation are dimensions related to the Intelligibility aspect of a speech signal. Yet they aim to analyze different aspects of this speech signal component. The Listening Effort targets the - self-assessed - cognitive load required by the listener to understand the content of the sample. The Pronunciation focuses on the anomalies that the listener defines as an annoyance. Therefore, the result of our analysis suggests that system N-US's pronunciation anomalies are more perturbing than the ones of system C-HMM. As system N-US is a unit selection system, this is likely the result of concatenation artefacts that can be quite prominent.

When focusing on the natural voice, the Voice Pleasantness is noticeable. It is the only dimension with a median MOS different from 5. For the same voice, this dimension also obtains lower average MOS (4.08) than the other dimensions (between 4.43 for the Overall Quality and the Naturalness, and 4.68 for the Pronunciation). The Voice Pleasantness is the dimension which can be considered the most subjective and listener specific. This is confirmed by a brief analysis of this specific dimension which shows that the average of MOS per user varies from 2.4 to 4.9; the median from 2 to 5. Yet the average MOS and the median for the Voice Pleasantness per user are either close to the ones for Naturalness or can even be above them. Consequently, these dimensions are the most sensitive to the scale contraction and the scale non-linearity biases. When conducting a Spearman correlation analysis, the strongest observed correlation is between these two dimensions with a correlation coefficient of $\rho = 0.62$. This correlation remains moderate, confirming that listeners distinguish these two dimensions.

Experiment 2 - Summary

- Strong influence of lower quality system on the MOS due to absence of an anchor
- Listening Effort and Pronunciation show different trends ⇒ concatenation artefacts more likely to cause the pronunciation trend
- Voice Pleasantness is the closest dimension to Naturalness

6. Experiment 3 - Expanding 2013-EH2 with modern systems

Experiment 1 demonstrates that MOS is relatively stable across time. Experiment 2 demonstrates that, without well defined anchors, the MOS (result of an ACR) is relative to the presence of lower quality systems. In the following experiment, we aim to investigate how the score given to a historical system evolves when more modern technologies are present in the test. In other word, as better technologies have been developed, we want to determine if the quality category of top-tier systems from 2013 will remain or if the classification of these systems will change. To do so, we conducted an ACR evaluation which includes the top-tier systems of 2013-EH2 and four modern synthesis systems. This experiment was previously presented in [48].

6.1. Adding new submissions

We built four additional systems to cover the combinations of two acoustic modelling architectures with two neural vocoders to represent modern technologies. For the acoustic models, we used Tacotron [2] as implemented by [56] and FastPitch [57, 58]. Tacotron is the original Sequence-To-Sequence architecture. It uses an attention mechanism both to learn phone-spectrogram alignment during training and to predict duration during inference. FastPitch, on the other hand, is fully-supervised regarding phone-spectrogram alignment during training and performs explicit duration prediction during inference. Externally-provided F0 (referred to as "pitch" in the name of the model) provides additional supervision and is likewise explicitly predicted during inference. As a result, FastPitch is significantly faster to train, taking one day compared to Tacotron's three-and-a-half weeks (on a single NVIDIA GeForce RTX 2080 TI). As the corpus contains utterances with long pauses, which can derail the training of attention, we used guided-attention for Tacotron [59]. This, and FastPitch, both require a phone-spectrogram alignment, which we obtained using Montreal Forced Aligner (MFA) [60].

The neural vocoders we selected are WaveNet [1, 61] and Parallel WaveGAN [62, 63]. WaveNet is the first neural vocoder and has been the initial step in the paradigm shift from source/filter modelling to modern speech synthesis.

Therefore, it is an important milestone in the evolution of speech synthesis. While the quality produced by WaveNet is known to be outstanding, training and inference are slow due to its auto-regressive architecture. Parallel WaveGAN uses the Generative Adversarial Network (GAN) paradigm, with a generator based on components developed for WaveNet. The main difference with WaveNet is that Parallel WaveGAN is not auto-regressive, and the loss function is a multi-resolution STFT. This allows Parallel WaveGAN to be trained faster; it is also fast at inference time. We use the following identifiers for each system: FastPitch/WaveNet (F-N), FastPitch/Parallel WaveGAN (F-G), Tacotron/WaveNet (T-N) and Tacotron/Parallel WaveGAN (T-G).

The input to these models is a phonetic sequence enriched by appending elementary prosodic information (consistent with what was standard in 2013). For each vowel, we provide the lexical stress of the containing syllable. We associate the punctuation type with each pause label, if available. To obtain this input, we used MaryTTS [64].

When training these systems, operating at 16 kHz is a necessary compromise to allow comparison with the selected historical systems. As part of the implementation of these systems, developers also provided training recipes. For our experiments, we used the training recipes for either the Blizzard Challenge 2013 dataset, when available, or *cmu-slt-artic*, which is another standard corpus with waveforms sampled at 16 kHz. We then experimented with multiple acoustic parameter extraction and normalization strategies proposed for training neural vocoders [61, 65, 63, 56]. The outcome was that using the acoustic parameters proposed for Parallel WaveGAN with a z-score normalization of the mel-spectrograms led to more stable results for both neural vocoders. The mel-spectrogram is extracted using 80 filters with cutoff frequencies from 50 Hz for the lower bound and 7600 Hz for the upper bound.

6.2. Evaluation Protocol

The evaluation protocol for this experiment is equivalent to the one described in Experiment 1 (Section 4.1). The main difference comes from the systems which are evaluated. The present experiment includes seven systems (3 top-tier systems from 2013 and 4 modern synthesis systems). The established Blizzard Challenge protocol requires that listeners rate only one sample per system for each section. Given that we are evaluating fewer systems than in 2013, we had to select, per section, a subset of 8 sentences. These sentences are the one evaluated in Experiment 2 (see Section 5.2). We only analyze the ratings for the same subset of utterances from the original challenge in the analysis section.

As for Experiment 1, we recruited listeners via Prolific [50], and they performed the evaluation online. However, for this experiment, each participant was paid £3.75[3].

6.3. Summary of the listeners

68 native speakers of English participated in the listening test, but 8 listeners failed to use headphones as per instructions. We excluded these listeners from our analysis. Of the 60 remaining listeners, 30 are native British English speakers, 29 are native US English speakers and one of an unknown accent. This last listener is excluded from the analysis. Therefore, our analysis is based on 59 listeners' responses, of which 28 are from females and 31 from males. From the pool of 59 listeners, 35 are under 40 years old with 23 being under 30 years old. For this experiment, the majority of listeners reported not using speech technologies (19 out of 59); only 6 listeners reported using speech technologies more than once per week.

6.4. Results of the expansion

Similarly to the previous experiments, the overall results are presented in (Figure 4) and (Table 7a). Table 7b presents the results of the series of Bonferroni-corrected pairwise Wilcoxon signed rank tests between systems' MOS.

Firstly, the natural voice is still considered more natural than the synthetic voices. As expected, modern systems produce more natural-sounding synthetic speech than all 2013 state-of-the-art technologies. The modern systems are all perceived as equally natural according to the results of the significance tests. Finally, for the historical systems K-Hybrid, N-US, C-HMM, we obtain identical results as for Experiment 1: the order remains K-Hybrid, N-US and C-HMM, with only system K-Hybrid being perceived significantly different from the other historical systems..

[3]The reason is that this evaluation took place earlier in the year 2022. Between this experiment and the other ones, Prolific increased the suggested rate to 4£ to account for inflation.

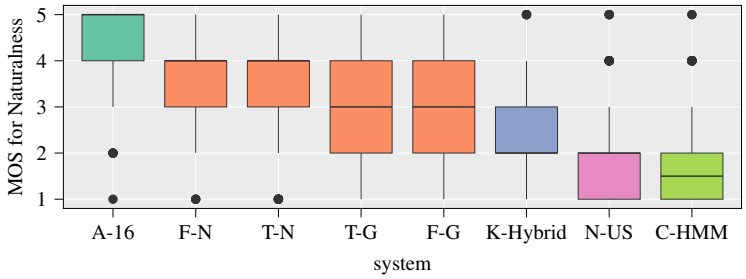


Figure 4: Listening test results for the experiment comparing 2013 top-tier systems to modern TTS technologies. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS. Each color represents a family of synthesis: ■ = natural voice, ■ = Hybrid synthesis, ■ = Unit Selection synthesis, ■ = HMM-based synthesis and ■ = DNN-based synthesis.

system	mean	std	median	mad	min	max	count
A-16	4.40	0.89	5.00	0.00	1.00	5.00	118
F-N	3.54	1.06	4.00	1.00	1.00	5.00	118
T-N	3.41	1.13	4.00	1.00	1.00	5.00	118
T-G	3.36	1.22	3.00	1.00	1.00	5.00	118
F-G	3.16	1.15	3.00	1.00	1.00	5.00	118
K-Hybrid	2.49	1.10	2.00	1.00	1.00	5.00	118
N-US	1.95	1.10	2.00	1.00	1.00	5.00	118
C-HMM	1.81	1.04	1.50	0.50	1.00	5.00	118

(a) Overall statistics

	A-16	F-N	T-N	T-G	F-G	K-Hybrid	N-US	C-HMM
A-16	NA							
F-N	■	NA						
T-N	■		NA					
T-G	■			NA				
F-G	■				NA			
K-Hybrid	■	■	■	■	■	NA		
N-US	■	■	■	■	■	■	NA	
C-HMM	■	■	■	■	■	■		NA

(b) Wilxocon results

Table 7: Statistical summary of the results for the experiment comparing 2013 top-tier systems to modern TTS technologies. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. The significance test consists of a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests with $\alpha = 0.05$. The systems are ordered using the mean MOS.

By comparing the results to the ones of Experiment 2, we can see the results for the natural voices are equivalent. The natural voice obtains a median MOS of 5 and an average MOS of 4.39. However, we can see that the MOS of historical systems drop. The median MOS of system K-Hybrid drops now from 3 to 2 and its average to 2.49 (R_d = 16.7 %, d = 0.79). While the median MOS for systems N-US and C-HMM remains at 2, their averages drop, respectively, to 1.95 (R_d = 18.1 %, d = 0.79) and 1.81 (R_d = 20.2 %, d = 0.89). This is a large drop which confirms the outcome of previous studies, such as [11, 12], that MOS is also relative to the systems presented during the evaluation.

When comparing these results to the ones obtained for Experiment 1, the relative nature of MOS is even more striking. The median MOS of system K-Hybrid drops by 2 full levels (from 4 in Experiment 1 to a score of 2 in Experiment 3). The average MOS of system K-Hybrid also falls by 1 full level (from 3.70 to 2.49, R_d = 24.2 %, d = 0.95). While the drop is less pronounced for systems N-US and C-HMM than for system K-Hybrid, it is still a significant one considering the lower scores obtained by these systems in Experiment 1. The median MOS of systems N-US and C-HMM drops by 1 level overall (from 3 to 2), their average MOS also drops by just over one level for system N-US (from 3.03 to 1.95, R_d = 21.7 %, d = 0.88) and just under one level for system C-HMM (from 2.47 to 1.81, R_d = 13.4 %, d = 0.61).

From the results of Experiments 1 and 2, it is evident that using an ACR does not lead to an absolute score. One fundamental flaw of the ACR is the absence of anchors. As outlined in Section 2, anchors have been introduced in more recent recommendations to stabilize the scale (i.e, reduce the stimulus spacing bias); quoting the recommendation BS-1534 [16], "the inclusion of appropriate and relevant anchors in testing enables **stable** use of the subjective rating scale". In the case of speech synthesis, designing such anchors is a challenging task considering the versatility of the production of such systems. Nonetheless, our results suggest that if speech synthesis researchers want to rely on the ACR for their evaluation, they will have to determine standardized anchors. The Annexe 1 of the third revision of MUSHRA [16] provides a starting point on how to design these anchors.

Experiment 3 - Summary

- Modern speech synthesis is perceived as more natural than the top-tier systems from 2013
- **Large drop** of the MOS of historical system compared to Experiments 1 and 2 but **order maintained**
- Reporting magnitude differences between MOS is meaningless

7. Experiment 4 - Exploring the perception of modern systems

The previous experiments demonstrate the relative nature of the ACR. Experiment 2 proves the influence of the lower anchor on the ratings; Experiment 3 shows the influence of more recent technologies on the ratings of older systems. One question remains to be answered: if we remove the older systems from the test, how will the scores of modern technologies be impacted? In other words, if we remove the lower anchors from the listener standpoint, would it allow systems to be better differentiated?

To answer this question, we conducted an experiment equivalent to Experiment 2. The only difference is that the three older systems (K-Hybrid, N-US and C-HMM) are substituted by the modern technologies introduced in the previous section (F-G, F-N, T-G, T-N). The protocol is identical to the one described in Section 5.2.

7.1. Summary of the listeners

Of the 59 recruited listener, 57 used headphones. One listener indicated being a speech synthesis expert and was therefore discarded. Of the 56 remaining listeners, 28 are native US English speakers, 27 native British English speakers and one native Indian English speaker. We also discarded the Indian speaker to ensure better consistency during the analysis. Therefore, our analysis is based on 56 listeners' reponses of which 28 are from females and 27 from males. Finally, for this experiment, the majority of listeners reported not using speech technologies (16 out of 55); only 7 listeners reported using speech technologies more than once per week. The majority of listeners are again under 40 years old (43 out of 55) with 20 of them being under 30 years old.

7.2. Results

As in the previous experiments, the overall results are presented in a boxplot (Figure 5) and a summary table (Table 8a). Table 8b presents the results of the series Bonferroni-corrected pairwise Wilcoxon signed rank tests between systems' MOS.

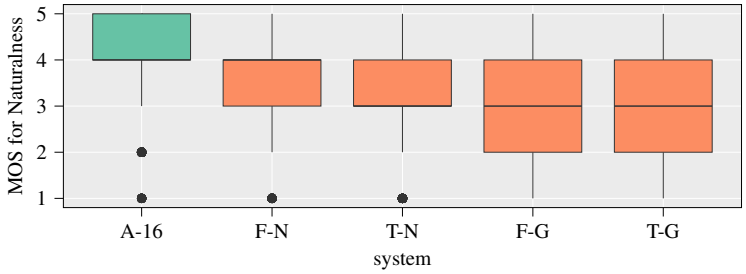


Figure 5: Listening test results for the experiment focusing on the modern speech synthesis technologies. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS. Each color represents a family of synthesis: ■ = natural voice and ■ = DNN-based synthesis.

By comparing the results of this experiment to those from Experiment 3, we can see little or no variation in the average MOS value associated with each system. However, when comparing the MOS medians, we can see two differences: the median MOS of the natural voice drops to 4, and that of Tacotron/WaveNet drops to 3. However, when analyzing the significance results, the only difference with Experiment 3 is that FastPitch/WaveNet is now considered more natural than any other modern methodologies. This pattern is similar to the one we observed when comparing

system	mean	std	median	mad	min	max	count
A-16	4.14	0.95	4.00	1.00	1.00	5.00	864
F-N	3.53	1.13	4.00	1.00	1.00	5.00	864
T-N	3.36	1.12	3.00	1.00	1.00	5.00	865
F-G	3.32	1.16	3.00	1.00	1.00	5.00	863
T-G	3.25	1.13	3.00	1.00	1.00	5.00	864

(a) Overall statistics

	A-16	F-N	T-N	F-G	T-G
A-16	NA				
F-N	■	NA			
T-N	■	■	NA		
F-G	■	■		NA	
T-G	■	■			NA

(b) Wilxocon results

Table 8: Statistical summary of the results for the experiment focusing on the modern speech synthesis technologies. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. The significance test consists of a series of Bonferroni-corrected pairwise Wilcoxon signed rank tests with $\alpha = 0.05$. The systems are ordered using the mean MOS.

the outcomes of Experiments 1 and 2: by removing systems considered less natural, we allow the listeners to quantify more subtitle differences between higher-quality systems.

Nevertheless, if we consider the drop of the average MOS of top-tier systems between Experiment 1 and Experiment 2, it is surprising to see how little the average MOS changed in comparison to Experiment 3. The only noticeable deviation is observed for Fastpitch/WaveGAN (F-G) with a medium effect (Cohens $d = 0.52$). Yet this deviation is not statistically significant in regard to the Wilcoxon test. This stabilisation of the scores could be explained by the saturation of the scale: listeners clearly identify the natural speech but acknowledge that modern synthesis systems are efficient. To explore differences between modern speech synthesis technologies further, using MOS in short sentences is too limited. Studies such as [66] already demonstrated that changing the length of the stimuli changes the ratings as listeners have more material to distinguish different systems. Furthermore, playing one sample per step, in addition to using a randomisation algorithm, prevents the listeners from spotting more subtle differences which could be critical to qualifying different systems.

	Overall		List. Effort		Pronunciation		Voice Pleas.		Naturalness	
system	M	Mdn	M	Mdn	M	Mdn	M	Mdn	M	Mdn
A-16	4.39	5.00	4.31	5.00	4.64	5.00	3.99	4.00	4.14	0.95
F-N	4.02	4.00	3.93	4.00	4.36	5.00	3.61	4.00	3.53	1.13
F-G	3.82	4.00	3.85	4.00	4.25	5.00	3.44	3.00	3.36	1.12
T-N	3.59	4.00	3.61	4.00	4.13	5.00	3.45	3.00	3.32	1.16
T-G	3.44	4.00	3.60	4.00	4.08	4.00	3.40	3.00	3.25	1.13

Table 9: Statistics for all the dimensions for Experiment 4. "Mdn" stands for *Median*. "M" stands for *mean*.

The results for all dimensions, presented in Table 9, follow a pattern similar to the ones for Naturalness. The Natural Voice is rated superior to all synthesis systems; the combination FastPitch/WaveNet (F-N) is the best synthesis system. This is confirmed by the significance testing. However, some differences are noticeable for the other systems. Systems relying on FastPitch produce speech which seems to be more intelligible and easier to listen to than Tacotron-based systems. The significance testing confirms this result as FastPitch systems are significantly different from Tacotron systems for both the Listening Effort and the Pronunciation.

Focusing on the Natural Voice, we observe similar results to those obtained for Experiment 2 (see Section 5.4): all dimensions have a median MOS of 5 except the Voice Pleasantness whose median MOS is 4. As expected, listeners clearly identified the natural voice as a reference. Focusing on the synthesis systems, we observe a shift toward the upper part of the scale for all dimensions compared to the results obtained for Experiment 2 (see Section 5.4). A dimension still stands apart. For all systems except for Tacotron-WaveGan (T-G), the Pronunciation has reached the peak median of 5. This proves the robustness of modern technologies. It is also important to note that our corpus is ideal for such systems: it contains well curated audio signals whose speech relies on a well-resourced and studied language.

Finally, the only dimensions which obtained a MOS median below 4 are the Naturalness and the Voice Pleasant-

ness. As we have seen previously, these dimensions are the most subjective. Considering the results of Experiment 2, this outcome suggests that we could be observing the influence of the scale contraction bias. As a result, a dedicated listener-centric approach is required to analyze these dimensions.

Experiment 4 - Summary

- Stability of the MOS in comparison to Experiment 3
- Pronunciation scale is saturated ⇒ modern synthesis is robust **for ideal corpora**
- Voice Pleasantness and Naturalness rated below the 5 ⇒ manifestation of the scale contraction bias?

8. Discussion and conclusions

In this paper, we first presented an overview of the ACR in Section 2. We then presented a series of four experiments replicating and following the 2013 edition of the Blizzard Challenge. To conduct these experiments, we trained additional models and conducted additional subjective evaluation campaigns. The subjective evaluation results and the resources to train the models are available in an open-source repository[4]. The models are available on the Blizzard resources website for non-commercial use only[5]. From our experiments, we reached the following conclusions.

Firstly, and as expected, modern speech synthesis technologies outperform the top-tier systems in 2013. This outcome is clear from the results of Experiment 3. While cross-evaluation comparisons should be made with extreme caution, we still can see some trends. Namely, the distribution of MOS is shifting upward for all dimensions when using Deep-Learning compared to 2013 state-of-the-art technologies. Minor differences have also been observed when comparing the MOS of modern speech synthesis technologies. These differences remain challenging to explain and suggest that we need more precise evaluation protocols.

Focusing on the MOS itself, these four experiments confirm its relative nature despite its origin as an Absolute Category Rating. This is especially apparent when comparing the evolution of the scores of top-tier systems from 2013 across Experiments 1, 2 and 3. If we consider system K-Hybrid, its median MOS for Experiment 1 is 4 while for Experiment 3 it is 2; leading to a drop of 2 levels. Experiment 2 demonstrates the influence of the absence of lower anchors, while Experiment 3 demonstrates the impact of the presence of new higher anchors. Furthermore, while the results of Experiment 1 suggest an overall stability of the scores across time, variations were still observed between the original run and the repeat experiment. As suggested by [7] and [11], only the ranking remains constant across the different experiments. It is also important to note that the models are as good as the quality of the dataset used to trained the models. As we have seen in Section 3.2, the dataset we used is of high-quality. With a less meticulously curated corpus, not only the scores will be affected, but it is also possible that this ranking could fluctuate. This is because various systems may exhibit varying degrees of sensitivity to different types of artifacts.

During the original campaign, listeners were asked if they faced problems with the evaluation task of Naturalness. The main reported problem (19 out of 38 reports) is about the scale (too big, too small or confusing). As detailed in [5], a comparison of scales has been conducted to evaluate video quality. Huynh-Thu et al. [67] compared 5-, 9-, 11-point discrete and continuous scales. However, they found no statistically significant differences between MOS. We did not find an equivalent analysis for speech processing, but this study suggests that individual listeners may prefer different scales, but a 5-point scale is an ideal tradeoff. A promising alternative to ACR is seen in MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [16]. The introduction of anchors and the hidden reference aimed to stabilise the scores. Yet numerous studies listed in [68, 7, 8] show that this protocol remains subjective to the same biases even if their impact is reduced. Nonetheless, the nature of MUSHRA as an hybrid rating/ranking task removes the temptation to quote absolute scores.

In parallel, the automatic prediction of MOS has now achieved impressive accuracy due to the eruption of Deep-Learning based MOS predictors. If we consider the results of 2022 VoiceMOS Challenge [18], we can see that the

[4] https://github.com/sigmedia/bc_2013_extension/
[5] https://www.cstr.ed.ac.uk/projects/blizzard/2013/lessac_blizzard2013/

Self-Supervised Learning (SSL) baseline obtains a SRCC of 0.93. This is an impressive result but as a community, we need to reflect on this. The VoiceMOS Challenge corpus (presented in [38]) contains samples of various quality produced by different methodologies. Confronted with a more homogeneous dataset with systems of closer quality and producing artefacts of the same nature, it is likely that these predictors will not be able to rank the samples of this dataset with such high accuracy. The work in this paper clearly shows the problems with relying on MOS when conducting evaluations with consistent data and carefully controlled conditions. In other words, our work demonstrates that we may be increasing our ability to reliably predict an unreliable measure of the absolute rating of synthetic speech in predicting MOS. As a result and considering the biases that MOS faces, these predictors should be used with extreme caution.

From our review of the existing literature on the evaluation of synthetic speech and the results of our experiments, it has become apparent that we are now at a crossroads. Deep-Learning has enabled synthesis systems to produce artificial speech of outstanding quality. Standard protocols such as ACR, and maybe even MUSHRA, are not enough to help researchers to get a meaningful diagnosis of the quality of a speech synthesis system. The goal of speech synthesis evaluation is to provide methodologies to determine how good synthetic speech is. In the spirit of [69], we consider that a paradigm shift is in order as we need to answer this fundamental question: what constitutes "good" synthetic speech? To answer this question, it will be necessary to consider multiple point of views about speech synthesis as multiple research communities (e.g. Human-Computer Interaction (HCI) [70], avatars [71] or even trust [72]) already assess speech synthesis for their own use-cases, hence providing other perspectives and ways to evaluate speech synthesis.

9. Acknowledgements

The authors would like to thank Rob Clark, Rasmus Dall, Cassia Valentini-Botinhao, Erica Cooper, Xing Wang and Junichi Yamagishi for the fruitful discussions on the topic presented in this study. This research was conducted with the financial support of Irish Research Council (IRC) under Grant Agreement No. 208222/15425 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme under `Grant No. 13/RC/2106_P2`.

References

[1] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A generative model for raw audio, CoRR abs/1609.03499 (2016). `arXiv:1609.03499`.
URL http://arxiv.org/abs/1609.03499

[2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, Tacotron: A fully end-to-end text-to-speech synthesis model, CoRR (2017). `arXiv:1703.10135`.
URL http://arxiv.org/abs/1703.10135

[3] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, J. Gustafson, Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation, in: Proceedings of ISCA Speech Synthesis Workshop (SSW), 2023, pp. 41–47. `doi:10.21437/SSW.2023-7`.

[4] ITU, Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, International Telecommunication Union (ITU-P), Geneva (1996).

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

[6] M. Viswanathan, M. Viswanathan, Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale, Computer Speech & Language 19 (1) (2005) 55–83. `doi:10.1016/j.csl.2003.12.001`.

[7] S. Zieliński, F. Rumsey, S. Bech, On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review, Journal of the Audio Engineering Society (JAES) 56 (6) (2008) 427–451.

[8] S. Zielinski, On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion, Journal of the Audio Engineering Society 64 (1/2) (2016) 55–74.

[9] A. M. Van Dijk, J.-B. Martens, A. B. Watson, Quality asessment of coded images using numerical category scaling, in: Advanced Image and Video Communications and Storage Technologies, Vol. 2451, SPIE, 1995, pp. 90–101.

[10] R. K. Mantiuk, A. Tomaszewska, R. Mantiuk, Comparison of four subjective methods for image quality assessment, in: Computer graphics forum, Vol. 31, Wiley Online Library, 2012, pp. 2478–2491.

[11] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, Statistical analysis of the blizzard challenge 2007 listening test results, in: The Blizzard Challenge Workshop, 2007, http://festvox.org/blizzard/bc2007/blizzard_2007/full_papers/blz3_003.pdf.

[12] A. Rosenberg, B. Ramabhadran, Bias and statistical significance in evaluating speech synthesis with mean opinion scores, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2017, pp. 3976–3980. doi:10.21437/Interspeech.2017-479.
URL http://dx.doi.org/10.21437/Interspeech.2017-479
[13] S. Shirali-Shahreza, G. Penn, Mos naturalness and the quest for human-like speech, IEEE Spoken Language Technology Workshop (SLT) (Dec 2018). doi:10.1109/slt.2018.8639599.
[14] ITU, Mean opinion score interpretation and reporting, ITU-T Recommendation P.800.2, International Telecommunication Union (ITU-P), Geneva (2016).
[15] ITU, Subjective video quality assessment methods for multimedia applications, ITU-T Recommendation P.910, International Telecommunication Union (ITU-P), Geneva (1996).
[16] ITU-T, Method for the subjective assessment of intermediate sound quality (MUSHRA), Tech. Rep. BS.1534-1, International Telecommunication Union (ITU-R) (2001).
[17] M. Wester, C. Valentini-Botinhao, G. E. Henter, Are we using enough listeners? no! - an empirically-supported critique of interspeech 2014 TTS evaluations, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2015, pp. 3476–3480. doi:10.21437/Interspeech.2015-689.
[18] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, J. Yamagishi, The VoiceMOS Challenge 2022, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2022, pp. 4536–4540. doi:10.21437/Interspeech.2022-970.
[19] H. C. Nusbaum, A. L. Francis, A. S. Henly, Measuring the naturalness of synthetic speech, International Journal of Speech Technology 1 (1) (1995) 7–19. doi:10.1007/bf02277176.
URL http://dx.doi.org/10.1007/bf02277176
[20] Ieee recommended practice for speech quality measurements, IEEE Transactions on Audio and Electroacoustics 17 (3) (1969) 225–246. doi:10.1109/TAU.1969.1162058.
[21] G. Williams, L. Moye, Subjective evaluation of unsuppressed echo in simulated long-delay telephone communications, Proceedings of the Institution of Electrical Engineers 118 (3–4) (1971) 401. doi:10.1049/piee.1971.0074.
URL http://dx.doi.org/10.1049/piee.1971.0074
[22] W. R. Daumer, J. R. Cavanaugh, A subjective comparison of selected digital codecs for speech, Bell System Technical Journal 57 (9) (1978) 3119–3165. doi:10.1002/j.1538-7305.1978.tb02197.x.
URL http://dx.doi.org/10.1002/j.1538-7305.1978.tb02197.x
[23] M. C. Hall, Objective quality evaluation of parallel-formant synthesised speech, in: First European Conference on Speech Communication and Technology (Eurospeech 1989), ISCA, 1989. doi:10.21437/eurospeech.1989-315.
URL http://dx.doi.org/10.21437/eurospeech.1989-315
[24] CCITT, Method for evaluation of service from the standpoint of speech transmission quality, CCITT Recommendation P.82, International Telegraph and Telephone Consultative Committee (CCITT) (1984).
[25] A. Kain, M. Macon, Spectral voice conversion for text-to-speech synthesis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1998. doi:10.1109/icassp.1998.674423.
URL http://dx.doi.org/10.1109/ICASSP.1998.674423
[26] ITU, Methodology for the subjective assessment of the quality of television pictures, ITU-R Recommendation BT.500-13, International Telecommunication Union (ITU-R), Geneva (1996).
[27] ITU-T, Method for the subjective assessment of intermediate sound quality (MUSHRA), Tech. Rep. BS.1534-3, International Telecommunication Union (ITU-R) (2015).
[28] E. C. Poulton, S. Poulton, Bias in quantifying judgements, Taylor & Francis, 1989.
[29] E. Cooper, J. Yamagishi, Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2023, pp. 1104–1108. doi:10.21437/Interspeech.2023-1076.
[30] N. Schinkel-Bielefeld, A. K. Leschanowsky, How much is the use of a rating scale by a listener influenced by anchors and by the listener's experience?, in: Audio Engineering Society Convention 138, Audio Engineering Society, 2015.
[31] ITU-T, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Tech. Rep. P.862, International Telecommunication Union (ITU-R) (2001).
[32] S. Möller, W.-Y. Chan, N. Cote, T. Falk, A. Raake, M. Waltermann, Speech quality estimation: Models and trends, IEEE Signal Processing Magazine 28 (6) (2011) 18–28. doi:10.1109/msp.2011.942469.
URL http://dx.doi.org/10.1109/msp.2011.942469
[33] M. Cernak, M. Rusko, An evaluation of synthetic speech using the PESQ measure, in: European Congress on Acoustics, 2005, pp. 2725–2728.
[34] F. Hinterleitner, S. Zabel, S. Möller, L. Leutelt, C. Norrenbrock, Predicting the quality of synthesized speech using reference-based prediction measures, in: B. J. Kröger, P. Birkholz (Eds.), Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011., TUDpress, Dresden, 2011, pp. 99–106.
[35] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, D. Sculley, AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech, in: NIPS - End-to-end Learning for Speech and Audio Processing Workshop, 2016.
[36] E. Cooper, W.-C. Huang, T. Toda, J. Yamagishi, Generalization ability of mos prediction networks, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8442–8446. doi:10.1109/ICASSP43922.2022.9746395.
[37] Y. Choi, Y. Jung, Y. Suh, H. Kim, Learning to maximize speech quality directly using mos prediction for neural text-to-speech, IEEE Access 10 (2022) 52621–52629. doi:10.1109/ACCESS.2022.3175810.
[38] E. Cooper, J. Yamagishi, How do Voices from Past Speech Synthesis Challenges Compare Today?, in: 11th ISCA Speech Synthesis Workshop

(SSW 11), ISCA, 2021. doi:10.21437/ssw.2021-32.
URL http://dx.doi.org/10.21437/ssw.2021-32
[39] A. W. Black, K. Tokuda, The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets, in: INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005, 2005, pp. 77–80.
[40] S. King, Measuring a decade of progress in text-to-speech, Loquens 1 (1) (2014) 006. doi:10.3989/loquens.2014.006.
[41] S. King, V. Karaiskos, The blizzard challenge 2013, in: The Blizzard Challenge Workshop, 2013, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.
[42] C. Benoît, M. Grice, V. Hazan, The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, Speech Communication 18 (4) (1996) 381–392. doi:https://doi.org/10.1016/0167-6393(96)00026-X.
URL http://www.sciencedirect.com/science/article/pii/016763939600026X
[43] F. Hinterleitner, G. Neitzel, S. Möller, C. Norrenbrock, An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks, in: The Blizzard Challenge Workshop, 2011.
URL http://festvox.org/blizzard/bc2011/DeutcheTelekom_Blizzard2011.pdf
[44] O. Watts, G. Eje Henter, J. Fong, C. Valentini-Botinhao, Where do the improvements come from in sequence-to-sequence neural TTS?, in: Proc. 10th ISCA Speech Synthesis Workshop, 2019, pp. 217–222. doi:10.21437/SSW.2019-39.
URL http://dx.doi.org/10.21437/SSW.2019-39
[45] S. King, V. Karaiskos, The blizzard challenge 2011, in: The Blizzard Challenge Workshop, 2011, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.
[46] H. Zen, T. Toda, An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005, in: European Conference on Speech Communication and Technology (Eurospeech), Lisbon, 2005.
[47] P. Taylor, A. W. Black, R. Caley, The architecture of the festival speech synthesis system, in: The third ESCA/COCOSDA workshop (ETRW) on speech synthesis, 1998, pp. 147–152.
[48] S. Le Maguer, S. King, N. Harte, Back to the Future: Extending the Blizzard Challenge 2013, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2022, pp. 2378–2382. doi:10.21437/Interspeech.2022-10633.
[49] F. Ribeiro, D. Florencio, C. Zhang, M. Seltzer, CROWDMOS: An approach for crowdsourcing mean opinion score studies, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011. doi:10.1109/icassp.2011.5946971.
URL http://dx.doi.org/10.1109/icassp.2011.5946971
[50] Prolific · Quickly find research participants you can trust.
URL https://www.prolific.co
[51] ITU-T, A method for subjective performance assessment of the quality of speech voice output devices, Tech. Rep. P.85, International Telecommunication Union (ITU-R) (1994).
[52] T. Pommée, M. Balaguer, J. Mauclair, J. Pinquier, V. Woisard, Intelligibility and comprehensibility: A Delphi consensus study, International Journal of Language & Communication Disorders 57 (1) (2021) 21–41. doi:10.1111/1460-6984.12672.
URL http://dx.doi.org/10.1111/1460-6984.12672
[53] F. Hinterleitner, C. Norrenbrock, S. Möller, Is intelligibility still the main problem? a review of perceptual quality dimensions of synthetic speech, in: Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8), 2013, pp. 147–151.
[54] L. Pinto-Coelho, D. Braga, M. Sales-Dias, C. Garcia-Mateo, On the development of an automatic voice pleasantness classification and intensity estimation system, Computer Speech & Language 27 (1) (2013) 75–88, special issue on Paralinguistics in Naturalistic Speech and Language. doi:https://doi.org/10.1016/j.csl.2012.01.006.
URL https://www.sciencedirect.com/science/article/pii/S0885230812000083
[55] E. Gaudrain, J. Undurraga, N. Grimault, D. Başkent, Comments on "differences in common psychoacoustical tasks by sex, menstrual cycle, and race" by d. mcfadden et al., 2018, and methodological pitfalls in human population research (2020). doi:10.31234/osf.io/ghfpv.
[56] C. Valentini-Botinahao, A. Govender, O. McCarthy, Tacotron + WaveRNN (2020).
URL https://github.com/cassiavb/Tacotron/commit/946408f8cd7b5fe9c53931c631267ba2a723910d
[57] A. Łańcucki, Fastpitch: Parallel Text-to-Speech with Pitch Prediction, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 6588–6592. doi:10.1109/ICASSP39728.2021.9413889.
[58] NVIDIA Group, FastPitch (2021).
URL https://github.com/NVIDIA/DeepLearningExamples/commit/6a642837c471c596aab7edf204384f66e9483ab2
[59] X. Zhu, Y. Zhang, S. Yang, L. Xue, L. Xie, Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis, IEEE Access 7 (2019) 65955–65964. doi:10.1109/access.2019.2914149.
URL http://dx.doi.org/10.1109/ACCESS.2019.2914149
[60] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal forced aligner: Trainable text-speech alignment using kaldi., in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2017, pp. 498–502.
[61] R. Yamamoto, WaveNET (2020).
URL https://github.com/r9y9/wavenet_vocoder/commit/a35fff76ea3687b05e1a10023cad3f7f64fa25a3
[62] R. Yamamoto, E. Song, J.-M. Kim, Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6199–6203. doi:10.1109/ICASSP40776.2020.9053795.
[63] T. Hayashi, Parallel WaveGAN (2021).
URL https://github.com/kan-bayashi/ParallelWaveGAN/commit/6d4411b65f9487de5ec49dabf029dc107f23192d
[64] I. Steiner, S. L. Maguer, Creating New Language and Voice Components for the Updated MaryTTS Text-to-Speech Synthesis Platform, in: International Conference on Language Resources and Evaluation (LREC), 2018.
[65] NVIDIA Group, waveglow (2020).
URL https://github.com/NVIDIA/waveglow/commit/8afb643df59265016af6bd255c7516309d675168

[66] R. Clark, H. Silen, T. Kenter, R. Leith, Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs, in: Proceedings of the Speech Synthesis Workshop (SSW), 2019, pp. 99–104. doi:10.21437/SSW.2019-18.

[67] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, A. Raake, Study of rating scales for subjective quality assessment of high-definition video, IEEE Transactions on Broadcasting 57 (1) (2011) 1–14. doi:10.1109/tbc.2010.2086750.
URL http://dx.doi.org/10.1109/tbc.2010.2086750

[68] S. Zieliński, P. Hardisty, C. Hummersone, F. Rumsey, Potential biases in mushra listening tests, in: Audio Engineering Society Convention 123, Audio Engineering Society, 2007.

[69] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. L. Maguer, Z. Malisz, É. Székely, C. Tånnander, et al., Speech Synthesis Evaluation—State-of-the-Art Assessment and Suggestion for a Novel Research Program, in: Speech Synthesis Workshop (SSW), 2019, pp. 105–110. doi:10.21437/SSW.2019-19.

[70] M. Cohn, E. Raveh, K. Predeck, I. Gessinger, B. Möbius, G. Zellou, Differences in Gradient Emotion Perception: Human vs. Alexa Voices, in: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), 2020, pp. 1818–1822. doi:10.21437/Interspeech.2020-1938.

[71] D. Higgins, K. Zibrek, J. Cabral, D. Egan, R. McDonnell, Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans, Computers & Graphics 104 (2022) 116–128. doi:https://doi.org/10.1016/j.cag.2022.03.009.
URL https://www.sciencedirect.com/science/article/pii/S0097849322000474

[72] I. Torre, J. Goslin, L. White, D. Zanatto, Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience, in: Proceedings of the Technology, Mind, and Society, TechMindSociety '18, Association for Computing Machinery, New York, NY, USA, 2018. doi:10.1145/3183654.3183691.
URL https://doi.org/10.1145/3183654.3183691