Acoust. Sci. & Tech. **45**, 4 (2024)

INVITED REVIEW

A review on subjective and objective evaluation of synthetic speech

Erica Cooper[1,*], Wen-Chin Huang[2], Yu Tsao[3], Hsin-Min Wang[3], Tomoki Toda[2] and Junichi Yamagishi[1]

[1]*National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan*
[2]*Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464–8601 Japan*
[3]*Academia Sinica, No. 128, Sec. 2, Academia Rd., Nangang Dist., Taipei 115, Taiwan*

(*Received 7 February 2024, Accepted for publication 29 March 2024, J-STAGE Advance published date: 4 April 2024*)

Abstract: Evaluating synthetic speech generated by machines is a complicated process, as it involves judging along multiple dimensions including naturalness, intelligibility, and whether the intended purpose is fulfilled. While subjective listening tests conducted with human participants have been the gold standard for synthetic speech evaluation, its costly process design has also motivated the development of automated objective evaluation protocols. In this review, we first provide a historical view of listening test methodologies, from early in-lab comprehension tests to recent large-scale crowdsourcing mean opinion score (MOS) tests. We then recap the development of automatic measures, ranging from signal-based metrics to model-based approaches that utilize deep neural networks or even the latest self-supervised learning techniques. We also describe the VoiceMOS Challenge series, a scientific event we founded that aims to promote the development of data-driven synthetic speech evaluation. Finally, we provide insights into unsolved issues in this field as well as future prospects. This review is expected to serve as an entry point for early academic researchers to enrich their knowledge in this field, as well as speech synthesis practitioners to catch up on the latest developments.

Keywords: Synthetic speech evaluation, Mean opinion score, Automatic speech quality prediction, VoiceMOS Challenge

1. INTRODUCTION

Synthesized speech, which is artificial speech generated by a computer, requires evaluation in order to judge whether it is understandable to listeners, natural-sounding, well-matched to the target speaker or speaking style, and generally acceptable for its intended purpose. Evaluation is also required to judge whether some new synthesis method is better than a previous one, or whether some new proposed modification gives an improvement. For as long as researchers have been developing synthesized speech, they have also been considering how to evaluate it. Historically, such evaluation has mainly relied on listening tests conducted with human listeners. Human opinions are the gold standard for evaluating synthesized speech because, after all, it will be humans who will listen to it. However, such evaluations are very costly and time-consuming, and researchers have also considered more automated evaluation methods to streamline the experimental iteration process. From acoustic correlates of human opinions to signal-processing-based methods developed for telephony, to machine learning-based approaches trained on listening test data, researchers have been considering and testing out these automated evaluation methods to make their experiments more efficient. This review will outline a history of evaluation for speech synthesis, including different types of subjective listening tests and efforts to find suitable objective metrics. We will also describe our two years of experience running the VoiceMOS Challenge, a shared task for data-driven opinion prediction for the quality of synthesized speech. Finally, we will describe ongoing work in this area as well as unsolved issues and future prospects.

2. LISTENING TESTS

This section will overview the listening test methodologies that have historically been used for synthesized speech, as well as current popular methodologies. We will also make note of some critiques of these methodologies that have arisen. In the case of modular synthesizers which have a front-end for linguistic processing, each of the

*e-mail: ecooper@nii.ac.jp
[doi:10.1250/ast.e24.12]

linguistic processing components (such as text normalization and grapheme-to-phoneme conversion) may have their own evaluation methodologies, which can be automated in the presence of sufficient labeled data. However, evaluations of the linguistic components of speech synthesizers are outside of the scope of the paper and we mainly focus on evaluation of the final synthesized speech.

2.1. 1980s to Early 1990s: Intelligibility and Comprehension

During the 1980s and early 1990s, popular approaches for computerized speech synthesis were rule-based formant synthesizers and concatenative unit-selection-based synthesizers based on small units such as diphones or groups of phonemes. The evaluation mainly focused on aspects related to intelligibility at the phoneme, word, or sentence level, as well as comprehension at the multi-sentence or paragraph level.

A 1990 survey paper [1] describes several evaluation methodologies for speech synthesis focusing on popular ones in the decade prior, and their advantages and disadvantages. The authors point out that most evaluations for synthesized speech at that point focused on intelligibility. They outline three types of evaluations: *comparative*, which will reveal which synthesis system is best, *diagnostic*, which will assist in identifying problems with a synthesizer, and *applied*, which will demonstrate how well-suited the synthesizer is for a particular application. An example of an applied evaluation is a **controlled field test**, in which a synthesizer is evaluated in the real world in its intended application with real users as the listeners — it is stated that this kind of test is relatively uncommon, and how to design ecologically-valid listening tests still remains an open question.

The **Modified Rhyme Test** (MRT) [2] is a test developed for telephony for evaluating intelligibility at the phoneme level, and it is cited as the most frequently-used listening test at the time for comparative evaluations of speech synthesizers in controlled conditions. In an MRT, listeners hear a monosyllabic word, often in the context of a carrier sentence, and then they must choose the word that they heard from a list of six choices which vary by either the initial or final consonant. The MRT is a variation of the **Diagnostic Rhyme Test** (DRT) [3], which only presents listeners with two possible choices. Positive aspects of the modified rhyme test are that it is reliable as well as easy to conduct. Its drawbacks are that the test scenario is unrealistic compared to actual use cases for synthesized speech, and this limited scenario results in artificially high scores. The multiple-choice paradigm also does not reveal confusions that listeners may have made that are not presented among the six choices. Furthermore, consonant clusters are not evaluated in a standard MRT. Therefore, variations of the MRT were sometimes deployed, such as ones making use of open responses [4] and ones including consonant clusters in the testing material [5].

A 1993 paper [6] describes several perceptual tests used by Bell Labs that have a more diagnostic focus. Their in-house speech synthesis system was a concatenative one using a database of units ranging in length from one to five phonemes, and they evaluated two versions of this along with the commercial DECtalk formant synthesizer and natural speech. Their motivations were to evaluate the coverage of their dataset and to identify bad units that may require re-recording. This paper describes how listening tests were typically conducted at the time — listeners were recruited from the local area near the office and in most cases they were not only unfamiliar with speech synthesis but also with computers in general and they did not have keyboard typing skills. So, listeners either used very simple user interfaces that only required pushing one or two buttons to make a choice, or else they were asked to write down what they heard using pencil and paper in the case of transcription tasks, and their responses would be entered into a computer later on. They conducted a **word pointing test** for detecting bad units, in which listeners point at words in a sentence where they hear problems and rate their severity; a **minimal pairs intelligibility test**, similar to the DRT except using word lists that cover more types of sounds such as vowels, consonant clusters, and multi-syllabic words; an **orthographic name transcription task** in which listeners write down proper names that they hear; **quality ratings scores with problem categorization**, similar to quality MOS but with a follow-up question asking listeners to choose a category of the problem they identify with the audio if their rating is low; and **paired comparisons with certainty ratings**, a modification of a simple pairwise comparison test where listeners also indicate the strength of their preference on a 1-6 scale. The authors reiterate that ratings of synthesized speech are inherently context-sensitive and therefore cannot be meaningfully compared across tests.

Beyond the phoneme level, word-level and sentence-level intelligibility of synthesized speech can be assessed by way of a **transcription task** that asks listeners to write down the word or sentence that they hear. Results are reported in terms of the percentage of words that were correctly identified. The testing material can consist of meaningful sentences that are chosen to have good coverage of the sounds of the language or to be representative of the target use case for the synthesizer, or *semantically-unpredictable sentences* (SUS) [7], which are grammatically-correct nonsense sentences. Although meaningful sentences better reflect a real-life use case, testing with SUS can give a more realistic picture of the intelligibility

of the systems under comparison because the effect of contextual clues is removed. Intelligibility tests using SUS are more rigorous and provide a kind of "lower bound" [8] for how well the system can be expected to be understood.

A 1980 paper [8] on the evaluation of the MITalk rule-based formant synthesizer critiques past (uncited) synthesis studies for focusing too much on intelligibility and not enough on *comprehension*, a listener's ability to understand and retain information from what they heard. In addition to conducting MRT tests and tests for word recognition in a sentence using both meaningful sentences and SUS, they also conducted a **listening comprehension test** in which listeners heard narrative passages and several multiple-choice questions about their contents, based on standardized reading comprehension tests. They compared listening to synthesized speech to both listening to natural speech and reading the passage to measure the comprehension differences in each case. However, this same study shows that this kind of comprehension test is already *saturated*, that is, it is not sensitive enough to reveal differences between natural and synthesized speech. A decade later, the 1990 survey paper notes that paragraph-level comprehension evaluations still remain scarce despite this being an important and realistic use case for speech synthesizers. Sentence-level intelligibility and comprehension can also be measured using a **sentence verification task**, in which listeners have to quickly decide whether a sentence is factually correct or not, and results can be measured in terms of listener latency and accuracy.

The 1990 survey paper gives a short overview of listening tests that evaluate intonation, which were starting to receive some attention at the time but were not yet as widely used as intelligibility tests. These include **pairwise comparison tests** in which listeners hear the same sentence realized by two different synthesizers and choose which one they prefer, **Mean Opinion Score tests** (MOS) [9] in which listeners rate an audio sample on an Absolute Category Rating (ACR) scale for some characteristic of the audio such as its naturalness, and **magnitude estimation tests**, in which listeners assign numbers of their own choosing to describe their perception of the magnitude of some aspect of the audio and their answers are normalized later [10]. Pairwise comparison tests can reveal fine-grained differences between systems as listeners are forced to make direct comparisons, and the human auditory system has a better ability to make comparisons rather than absolute judgments [11]; however, pairwise comparison tests are not well-suited for longer passages and they don't scale well to a large number of synthesizers as generally all pairs must be compared. MOS tests have become very popular, which will be discussed further in later sections. The magnitude estimation test was found to give unreliable results [10] and has not caught on as a major evaluation paradigm for speech synthesis.

As the MOS testing paradigm was gathering some interest for speech synthesis by this point, researchers started to consider the best ways to use it. A 1992 study [10] investigated two aspects of listening test design in the evaluation of four unnamed Swedish synthesizers. First, the granularity of the rating scale was evaluated by comparing a standard 5-point scale to an 11-point scale with half-point increments. Next, the question of which which systems to include in the listening test "context" was also addressed — they considered a "narrow" context where listeners only hear the four synthesizers, a middle context where natural speech is added, and a wide context where a low-quality reference is also included in the form of natural speech distorted with noise. Their results showed that increasing the listening test context reduced the scores that listeners gave to the synthesis systems in both the "middle" and "wide" context cases, and that scores also became reduced and more compressed in the case of the 11-point rating scale compared to the 5-point scale. These early results demonstrate that MOS is relative, not absolute, and depends on the testing materials and interface provided to the listeners.

2.2. Mid-1990s and 2000s: Naturalness, Intelligibility, and Efforts to Standardize

As the storage capacity of computers improved, concatenative synthesizers were developed to make use of larger databases from which longer units could be selected. With the improved quality of speech synthesis, evaluation during this era shifted from focusing mainly on intelligibility to including a more comprehensive evaluation of naturalness and prosodic factors. Efforts to develop standards for evaluating synthesized speech also increased, as did some introspection by researchers in the field as to whether evaluations are being conducted in a valid way or if improvements could be made.

The 1994 ITU-T Recommendation P.85 [12] represents an early effort to develop a recommended evaluation protocol specifically for synthesized speech. Based on the ITU-T Rec. P.80 [13] developed for telephony, this specification recommends using a set of ACR-based questions for different aspects of listener opinion such as overall impression, self-reported listening effort, articulation, and a final binary question about whether the voice is acceptable. Listeners conduct the ACR part of the test after an initial round of testing for comprehension of the same audio material, but P.85 gives no recommendation about how to use the results of this part of the test, indicating that comprehension is considered necessary but is still not a main focus. P.85 also outlines some standardization details, such as that at least 5 different synthesis systems should

be included in the test, audio samples should range from 10–30 seconds in duration, a training session should be provided for listeners, and one listening session should range from 40–60 minutes. A 1996 paper [14] provides similarly detailed recommendations for running SUS-based transcription tests, recommending the use of several different sentence types that generalize well across European languages, using short and common words. In 2000, guidelines were published [15] for evaluating Japanese synthesizers, focusing on intelligibility at the syllable, word, and sentence level, and overall quality, with recommendations to ask listeners about rhythm, intonation, and overall suitability.

A 2002 study of the reliability of P.85 [16] noted that, eight years later, P.85 had not seen much adoption, possibly because it seems complicated to run and is not embedded in a real task. They evaluated six different commercial English concatenative synthesizers using both the P.85 paradigm and a simple pairwise test. They aimed to test the effect of genre (domain) by including material from four different domains, and of listening session, by bringing back the same group of listeners one week later to repeat the same test. The authors found very strong correlations across several of the different P.85 rating scales, indicating that they may not really be testing different factors and therefore the complexity of P.85 tests may not in fact be worthwhile. They found a significant effect of genre, although system rankings did not change. Results were found to be very consistent across sessions. The only scales that differed significantly across sessions were "listening effort" and "comprehension problems," which had differences indicating that listeners had an easier time listening and comprehending during the second session and that there was a learning effect. The pairwise test gave almost the same rankings of systems, with less variability, and with more significant differences being revealed between systems with the same number of listeners. Another study from four years later [17] comparing P.85 to more commonly-used SUS tests for intelligibility and MOS tests for naturalness found that P.85 was not suitable for measuring intelligibility, with SUS tests being more rigorous and producing more useful results. However, contrary to the previous study, they found that the P.85 scales were *not* all correlated, and P.85 could provide a much more nuanced and informative picture of the naturalness and overall quality of a system. They attribute this different result to the fact that synthesized speech had improved substantially in the intervening years and that listeners were able to identify more subtle differences between the systems.

In 1997, a questionnaire was sent out to researchers in the field of speech synthesis, and the results were reported [18]. 16 researchers from around the world responded to questions about which listening test methods they knew about, and which ones they had actually used, revealing that pairwise comparison tests had the most users, and DRT/MRT tests, comprehension tests, and MOS were also well-known and used. They also collected free-text opinions and found that researchers were aware that there are plenty of choices for listening tests, but it would be useful to have some kind of guidelines for which test(s) to choose or adapt for a given case.

In 2005, the first Blizzard Challenge was held [19] to compare corpus-based text-to-speech synthesis techniques using standardized datasets and evaluations. The Blizzard Challenge has run almost every year since, and it has been an important initiative for documenting the progress of speech synthesis research. Inspired by the benefits of standardized datasets and evaluation metrics in the speech recognition community, the challenge was developed to provide a common ground for comparing different synthesis techniques. As there is no obvious "best" evaluation for synthesized speech, the challenge organizers chose to run a variety of listening tests, namely MOS and sentence transcriptions of both SUS and carrier sentences containing words from MRT/DRT word lists. This first edition of the challenge drew six participating teams from three continents, and evaluations were conducted online via the Blizzard homepage, with participants being recruited through word of mouth from communities of speech experts, volunteers, and US undergraduate students. MOS tests and SUS transcription tasks continued to be used in every subsequent Blizzard challenge, with tests for speaker similarity being added in later editions on the recommendation of the 2005 organizers. The evaluation methodologies set forth by the Blizzard organizers have set a strong precedent for speech synthesis evaluation.

Although evaluations for naturalness and intelligibility had become standard by this point, a 2007 book chapter [20] predicts that tone of voice, manner of speaking, emotional expressiveness, and, generally speaking, "interpersonal skills" will become more important for speech synthesis in the future, and so we will need to find ways to evaluate these aspects as well.

Listening tests were very common by this point, and a 2008 review [21] considers the various forms of bias that should be considered when designing and reporting them. Three categories of bias are described: biases arising from affective judgments (e.g., appearance of the testing equipment, expectations, personal preferences, emotions and mood), response mapping bias arising from the test design (e.g. stimulus spacing and frequency, perceptually nonlinear scales, and range-equalizing bias, the inclination of listeners to try to use the entire range of choices available to them), and interface bias (e.g., the layout of the assessment scale and the words chosen for the labels).

2.3. 2010s to the Present: Crowdsourcing, MOS, and Critiques

Unit selection-based synthesizers were still widely used at the start of this time period, and the more flexible and smaller-footprint statistical parametric speech synthesizers such as hidden Markov model (HMM) based ones [22] had also emerged. Currently, neural network-based synthesizers are dominant, with many commercial systems achieving very natural-sounding synthesis that cannot always be identified as computer-generated by listeners [23].

Although listening tests were already being conducted online in cases such as the Blizzard Challenge, laboratory tests remained commonplace until crowdsourcing platforms such as Amazon Mechanical Turk*, which launched in 2005, became popular, and especially after the CrowdMOS [24] open-source toolkit for running MOS listening tests on Mechanical Turk was published in 2011. Crowdsourcing platforms remove geographic constraints and expand the pool of potential participants, and they also allow listeners to participate in experiments in their own homes at a time of their choosing. Crowdsourced tests also remove much of the control that researchers have over their experiments, so thorough quality control must be performed. A 2013 book chapter on crowdsourcing for speech synthesis evaluation [25] notes that the number of papers using crowdsourcing for listening tests had increased dramatically by that point. The authors describe the most popular listening test paradigms at this time (SUS transcription, MOS for naturalness, and pairwise comparisons; we also begin to see references to listening tests measuring speaker similarity, expressivity, and speaking style) and best practices for crowdsourcing them, as well as lessons learned from their own experiences (for example, they recommend discarding a listener's first three answers since these fall into the listener's learning or calibration phase) and ways to filter out inattentive listeners (e.g., by including some "gold" samples which should always receive a high rating). The authors also described listening tests that did not work well with crowdsourcing, such as asking listeners to write a free-text description of their impression of the audio and a more diagnostic test including categorical choices about potential types of synthesis errors, which had low listener agreement and was found to be too complicated.

The trend of introspection into how the speech synthesis field conducts evaluations was continued in 2015 [26] in a study revisiting papers presented at Interspeech 2014. The authors list up the current most popular evaluation methodologies: MOS, differential MOS (DMOS; similar to MOS tests but with a reference sample, and listeners rate how different the test sample is; frequently used to rate speaker similarity), preference tests with and without references, transcription tasks, and **MUSHRA tests** [27] (MUltiple Stimuli with Hidden Reference and Anchor), which had emerged as a popular evaluation methodology by this time. MUSHRA is a test originally developed for broadcast audio in which listeners are presented with several systems' samples at the same time and they rate them on a sliding scale from 0–100. A reference sample of natural speech is included, representing the upper bound, and the MUSHRA specification also requires the inclusion of a middle and lower anchor, typically the reference sample low-pass-filtered at 7 kHz and a 3.5 kHz in the case of broadcast audio, although these are typically excluded in evaluations for synthesized speech.

In the 2015 study [26], it was observed that although there are many published guidelines for conducting the various types of listening tests, in many cases these guidelines are not followed. Analysis of the 2013 Blizzard Challenge revealed that at least 30 listeners should participate in listening tests in order to obtain reliable results, and the authors list some best practices for conducting listening tests for synthesized speech and reporting on their design. Another paper from the following year [28] observes the overwhelming popularity of MOS tests for evaluating various types of media, and points out many shortcomings, such as that MOS is not suitable for longer clips or for distinguishing fine-grained differences, the typical labels used for scoring are not perceptually linear, the typical procedure of removing outliers may remove completely valid differing opinions, and that the same final MOS can be obtained even after averaging some very different distributions of scores, so two systems may appear to be equivalent when they are actually quite different. Furthermore, they reiterate the warning that it is not valid to compare MOS across different studies.

With crowdsourced MOS tests becoming the most widely used evaluation paradigm for synthesized speech, a trend of critiques of MOS arose, along with a call to design more thoughtful, less saturated, and more ecologically valid evaluations. A 2019 position paper [29] points out that typical MOS tests evaluate isolated sentences, which is neither realistic nor especially meaningful, and that the community should consider contextual appropriateness and more task-driven evaluations as well as revisiting comprehension. A 2016 study [30] reexamined the multiple-choice comprehension test paradigm, testing natural speech and statistical parametric speech synthesis (SPSS) in a conversational domain. Although listeners reported that the comprehension task with synthesized speech was more difficult, the measured comprehension results were not significantly different between systems, reaffirming that comprehension tests are saturated and that more sensitive

*https://www.mturk.com

methodologies need to be developed. A 2017 study [31] designed a novel evaluation approach for interactions with a virtual avatar with different voice synthesis conditions and compared the outcome to the usual audio-only evaluation, observing smaller effects in the results of the interactive study, highlighting the difficulty in designing ecologically-valid evaluation paradigms. Attention also turned to the choice of listening test material, with a 2019 study [32] comparing MOS evaluations of synthesized paragraphs and the same sentences in isolation. They found that synthesized paragraphs were rated lower than the same sentences in isolation, even though natural speech paragraphs were rated higher than their isolated sentences, and that variations in the context provided to listeners changed their responses. A follow-up study from 2021 [33] evaluating sentences in isolation and in context found that the instructions presented to listeners had a strong effect, with different results obtained by asking them about "naturalness" vs. "appropriateness."

One common critique of MOS tests is that "naturalness" is not well-defined and that listeners may be considering different aspects of the audio when they decide their ratings [29,34,35]. Although it has been observed that the reliability of MOS tests shows that listeners still somehow know what to do despite the apparent vagueness of the task [36], it is a valid point that we may want to know which facets of naturalness are affecting listeners' judgments, for more diagnostic purposes. A 2023 study [35] asked listeners to provide a short free-text response at the end of the listening test to describe what criteria they used to assess naturalness, finding that listeners generally interpreted this as how "human-like" the audio sounded. Another study from around the same time [37] conducted a very fine-grained listening test, asking listeners to rate synthesized samples on over 40 properties in eight broader categories such as human-likeness and audio quality. Listeners were also asked to mark audio samples in the time domain for where points of unnaturalness especially occur. Listener agreement was found to be lower on these more fine-grained and well-defined categories compared to prior studies using basic naturalness MOS; however, these kinds of in-depth listener studies are an important step towards better understanding listeners' behavior and designing more thoughtful, comprehensive, and diagnostically-useful evaluations.

A 2018 study [38] advocated for the use of pairwise comparisons instead of MOS, showing evidence that MOS may have become saturated around 2013, and emphasizing the ability of pairwise comparison tests to make finer-grained distinctions. Two later studies from Interspeech 2023 both independently came to the conclusion that standard confidence intervals computed from MOS tests tend to be overly optimistic and that pairwise comparison tests are preferable based on empirical studies of real listening test data. The first of these [39] looked at replicated MOS and comparison tests and found that the fact that the same listener rates many stimuli breaks the independence assumptions that are made when computing confidence intervals, and that using cluster-based methods to compute them mitigated this. They also found that the results of pairwise tests were less influenced by the number of listeners compared to MOS tests, indicating that they are preferable to use especially if few listeners are available or if the synthesis systems under comparison are similar in quality. The second study [40] demonstrated empirically on a large-scale MOS dataset that huge amounts of samples, more than it is realistic to collect, are required in order to obtain small enough confidence intervals, computed using various different tail probability methods, to result in meaningful system rankings using MOS. They also advocated for the use of pairwise comparisons instead. Several other studies [34,41–43] revisit the context dependency of MOS, showing how changing the systems included in the listening test, scale instruments, and instructions can affect final MOS results, and that notably [43], MOS tests as they are typically run have become saturated and lost their ability to make meaningful distinctions between current systems, indicating a need for better evaluation methodologies going forward.

3. AUTOMATIC EVALUATION FOR SYNTHETIC SPEECH

Despite the relative ease of conducting crowdsourced listening tests online, especially compared to the days of scheduling local listeners to come to the lab in person and record their answers on paper, evaluation is still a bottleneck for experimental iteration and development of speech synthesizers. Many speech and language tasks come with automatic objective evaluation metrics, such as word error rate for automatic speech recognition and the BLEU score [44] for machine translation. In contrast, speech synthesis still lacks strong and agreed-upon objective evaluation metrics. Researchers have made efforts to address this gap by using metrics developed for telephony, finding acoustic correlates of human evaluations, measuring degradations compared with a ground-truth audio sample, and using data-driven machine-learning-based approaches.

Although claims have been made that automatic evaluation of speech synthesis should be difficult or impossible, this has not discouraged researchers from directing efforts toward more objective evaluation methodologies. There are several ways to categorize these methods. First, a *model-based* method learns a model from data to make the prediction, while a *signal-based* metric does not require learning such a model. Second, an *intrusive* (or *double-ended*) method requires a reference

signal for comparison, and a *non-intrusive*, (or reference-free, *single-ended*) method does not require a reference.

3.1. Difficulties in Automatic Evaluation of Synthetic Speech

Objective evaluation of synthesized speech is expected to be difficult for several reasons. First of all, there is the so-called "one-to-many" problem — for any given condition (text, style, environment, etc.), there may be countless ways to realize it that would be considered correct and natural. While we may evaluate a synthesized utterance by comparing it to a ground-truth one, we may be unfairly penalizing perfectly valid differences in prosody, timing, and pronunciation. How humans produce and perceive prosodic variations is also still not well understood [20].

Second, the types of artifacts encountered in synthesized speech, and the types of unnaturalness, are varied and also fundamentally different from those encountered in telephony. While noise can be a major cause of degradation in communication networks, this is typically not an issue for speech synthesis models that are developed using clean data, whereas discontinuities arising from concatenation points in unit selection synthesis, and issues like unnatural prosody are sources of unnaturalness that are specific to synthesized speech. It is also unknown whether certain subjective traits of speech correspond to objectively measurable components of a signal [29].

Last but not least, listening tests fundamentally collect information about subjective preferences, which can be expected to vary based on individual differences or contextual elements of the test [21]. Calibration to different listening test contexts would be necessary, but the best practice is still unknown [28].

3.2. Speech Quality Assessment Metrics from Telephony

Although there have been many signal-based metrics for objectively measuring the signal quality of speech that is transmitted over noisy telecommunication networks, in this section we will limit the scope to those that have been adopted for evaluating speech synthesis in particular.

The *Mel-cepstral distance (MCD)* measure [45] computes the difference between the Mel cepstra of a reference and test speech sample. The perceptually motivated Mel cepstrum was hypothesized to be a better match for subjective ratings than the previously used standard cepstrum, which was validated by experiments showing that MCD had better correlations with subjective ratings of low-bitrate coded speech with simulated channel conditions than cepstral distance. 15 years later, MCD was tested for evaluating synthesized speech [46] in the context of facilitating the development of speech synthesizers in new languages by non-experts. While small differences in delay have to be accounted for when using MCD for telephony, the alignment between a ground-truth speech sample and the corresponding synthesized sample may be completely different. The authors propose both the use of dynamic time warping (DTW) to address this, as well as the idea of using "gold" durations for synthesizing samples to be evaluated with MCD. While small differences in phoneme durations are unlikely to affect naturalness ratings, this approach cannot identify problems with duration modeling. The authors also point out that MCD is not suitable for finding problems like discontinuities in the f0 contour. Nevertheless, they found it to be a reasonable proxy for subjective opinions during development.

The *Perceptual Evaluation of Speech Quality (PESQ)* was developed for the objective evaluation of speech over narrow-band telephone networks and codecs, and standardized as ITU-T Recommendation P.862 [47]. This metric was designed to model human perception by estimating MOS. The PESQ algorithm aligns test and reference signals taking into account the possibly variable time delays that can occur in VoIP. Then, measures of absolute and additive disturbances are computed which are converted into a final score [48]. A third-order polynomial is fitted to a real MOS dataset to convert the final PESQ score into a final MOS-like value. Although learning is involved in the development, we discuss PESQ here because (1) a third-order polynomial is too simple for PESQ to be categorized as a model-based approach, and (2) most researchers use it off-the-shelf without re-training it on new datasets. It is noted in the recommendation that PESQ specifically measures the effects of one-way speech distortion and noise, and that it was not designed to measure loudness loss, delay, sidetone, echo, or other impairments.

Despite not being designed for the assessment of TTS, several works have employed PESQ for this purpose anyway, to determine its usefulness as a potential objective measure, with widely varying results. A 2005 study of single-word synthesis using three diphone synthesizers [49] found very high Pearson correlations of 0.99 between PESQ and MOS ratings, and the authors concluded that it would be possible to use PESQ instead of listening tests going forward. A later study in 2011 [50] did similar experiments using data from Blizzard 2008–2010 listening tests and found very low correlations of 0.17. They hypothesized that time alignment was the issue and that the previous study was affected less by that since they had been using short samples of individual words instead of full sentences. Nevertheless, they also tried using individual words cut from the Blizzard samples, but correlations between PESQ and MOS remained low, indicating that PESQ is not well-suited for evaluating the larger variety of synthesis methods represented in the Blizzard data. One

more 2015 study [51] evaluating PESQ correlations with Blizzard datasets from 2008–2013 found that the correlations with MOS were close to 0.

The subsequent *P.563 recommendation* [52,53] was the first reference-free measurement developed by the ITU, and the ANIQUE+ [54] reference-free model for narrow-band telephony was adopted as an ANSI standard shortly thereafter. Following PESQ, P.563 was designed to model human perception and predict MOS for narrow-band telecommunications and is not recommended for other purposes. P.563 considers three main categories of distortion: unnaturalness of speech (with separate analysis for male voices, female voices, and voices that sound strongly robotic due to distortion), strong additional noise (including low static signal-to-noise ratio (SNR) and low segmental SNR), and other distortions such as interruptions, mutes, and time clipping, in which the algorithm distinguishes between normal word endings and signal interruptions. A dominant distortion class is identified and the distortion measures are combined with distortion-dependent weightings. These final scores can be converted into a MOS-type value using a third-order polynomial calibrated against real MOS data, similar to PESQ. While P.563 correlates well with MOS for the intended conditions, it was shown to not correlate well when used for TTS [55]. Nonetheless, there still exist several works using P.563 as a baseline for comparison [56,57].

The *root mean squared error (RMSE) and correlation of f0* are measures that have been used for evaluating intonation of synthesized speech, with RMSE f0 measuring the distance between two f0 sequences, and the correlation measuring how well changes in direction of the f0 contour in the test sample match a reference sample. As these measures grew in popularity, it became necessary to verify their validity in terms of matching well with human perception of differences in intonation contours. This was done in a 1999 study [58] that collected listeners' ratings of audible similarity or difference in the intonation of pairs of speech samples and then evaluated their correlations with RMSE f0, f0 correlation, and other measures. While none of the correlations were very high, RMSE f0 matched best with human ratings.

A 2013 study [59] investigated popular objective and spectrum-based measures such as MCD, frequency-weighted signal-to-noise ratio (FWS), cepstral distance, log-likelihood ratio (LLR) based on linear prediction models, and weighted spectral slope, and their correlations with MOS ratings on three scales: speaker similarity, naturalness, and how much background noise was audible. The synthesizers under investigation were HMM-based TTS systems that had been speaker-adapted using either clean or noisy and enhanced target speaker data. Linear correlations between the measures under study and the MOS ratings showed that FWS correlated best with speaker similarity, MCD was the best correlate of naturalness in the clean adaptation data condition, and LLR was the best correlate for all three measures in the noisy adaptation data scenario.

3.3. Models for Evaluation of Synthetic Speech

Early model-based approaches used methods such as linear regression and support vector machines (SVMs), with neural network-based approaches growing in popularity as more and larger MOS-labeled datasets became available. In more recent years, self-supervised learning (SSL) based speech models have been proven to be useful for a huge variety of downstream tasks, including MOS prediction for synthesized speech.

It is worth mentioning that by 2015, commercial API-based automatic speech recognition (ASR) models had been shown to have good correlations with human transcriptions of SUS [60], and this remains a popular method for evaluating intelligibility of synthesized speech today [61,62]. Attention errors such as skips and repeats can also be counted in the case of attention-based neural synthesizers [63]. However, this section will mainly focus on prediction of more subjective aspects of synthesized speech such as naturalness and quality.

3.3.1. Early attempts at machine learning based synthetic speech quality prediction

A study in 2008 [55] was one of the first works that investigated model-based methods. They first showed that although the P.563 measure [52,53], which was designed for narrow-band telephony, had poor correlations with subjective quality ratings of synthesized speech, several internal features in fact had higher correlations (with some dataset dependency), indicating that useful information was being extracted. They thus proposed an approach using a regression tree and several of the internal P.563 features that were determined to be informative.

A follow-up study in 2010 [64] investigated different combinations of three sets of features for predicting subjective opinions: internal P.563 features, log-likelihoods from a reference HMM trained on natural speech, and a large set of over 1,500 general acoustic features such as ones related to signal duration, formants, intensity, pitch, and spectrum. Experiments using a linear regression model revealed that the best correlations (in the range of 0.7–0.8) with listening test results from Blizzard 2008 [65] and 2009 [66] were obtained when all three types of features were used.

The same group further investigated whether prosodic and MFCC-based acoustic features correlated with MOS on a more challenging Blizzard 2012 dataset [67], which had an evaluation of synthesized paragraphs in the audiobook domain [68]. They investigated a set of prosodic and

micro-prosodic features such as f0 mean, standard deviation, dynamics, rhythm parameters, jitter, and shimmer, as well as a set of MFCC-based features. Using feature selection methods and SVM classifiers, they found that the MFCC-based features were more informative than the prosodic ones, but that once again their combination produced the best results. In 2015, they continued their investigations predicting aspects of voice naturalness, prosody, and intelligibility using large acoustic feature sets and SVMs, this time introducing a nonlinearity in the form of a "regular perception range" which is derived for each quality aspect — this is the range in which correlation between the acoustic features and the quality rating is maximized, with values outside of this range hypothesized to be less perceptually salient and therefore ignored. Incorporating this nonlinear perceptual regularization into their SVM prediction pipeline produced correlation coefficients upwards of 0.9, leading the authors to conclude that nonlinear modeling is necessary for this prediction task.

While the above-mentioned works focused on single-ended methods, one study [51] developed a double-ended naturalness prediction method for past Blizzard data based on demiphone (a cluster of HMM states) level degradations relative to a reference signal using spectral and f0 features, with warping required to align the durations. They found that this metric had the added complication that sometimes there is not a perfect phonemic correspondence between synthesized and natural speech due to valid pronunciation variations and optional silences; nevertheless, they were able to obtain system-level correlations in the 0.8 range with this approach.

3.3.2. Neural network-based synthetic speech quality prediction

Deep neural networks (DNNs) have emerged as the most popular approach for modern classification and regression tasks in the past decade as the computational resources and large-scale datasets needed to train them have become more available. The task of evaluating synthesized speech has been no exception, with synthesis challenges providing ample MOS-labeled data.

Scientific challenges focusing on speech synthesis are naturally suitable sources for training DNN-based speech quality prediction models, owing to their scale. In addition to the Blizzard challenge series which focuses on TTS, in recent years the Voice Conversion Challenge (VCC) series has also become a popular data source. Founded in 2016 [69] and subsequently run in 2018 [70], 2020 [71], and 2023 [72], the VCC provides a platform for teams to compete in the task of voice conversion (VC) using shared datasets and evaluations. The VCC organizers also make the submitted audio samples and their ratings from listening tests available, which has been a valuable resource of large-scale listening test data that has been widely adopted by researchers building MOS predictors, with the 2016 and 2018 datasets being especially popular.

Most of the papers described in this section use some or all of the following evaluation metrics, which have become standard, at both the utterance level and the system level (averaging all the ratings for one synthesis system into a system-level score) to measure how well a predictor can predict human MOS ratings:

- **Root mean squared error (RMSE):** The average difference between actual and predicted MOS values.
- **Linear correlation coefficient (LCC):** The basic correlation between actual and predicted MOS values.
- **Spearman rank correlation coefficient (SRCC):** Correlations of the rankings of the actual and predicted MOS values — it may be more useful for MOS predictors to predict the ranks of systems correctly than to predict exact MOS values.
- **Kendall Tau correlation coefficient (KTAU):** Proposed for evaluating MOS predictors [73] because it measures rank correlations in a manner that is more robust to outliers.

Perhaps the first effort to attempt using neural networks for MOS prediction was in 2016 [57]. Several types of models were applied to the prediction of MOS ratings from six past years of Blizzard Challenges, with per-year mean normalization and per-system variance normalization applied to enable combining the datasets into one large-scale dataset. They compared linear regression models to neural networks and a hierarchical approach was used, with a system-level score being predicted first and then that prediction being used as a feature for stimulus-level prediction, based on the observation that system-level scores were more predictable and the intuition that knowing the system-level score should be informative for predicting the score for a single sample generated by that system. They found that using neural networks for both prediction stages worked better than linear regression, and that features extracted from a convolutional neural network (CNN) improved correlations over using features from P.563, MFCCs, and cepstral features.

AutoMOS [74] appeared shortly thereafter for the evaluation of production-grade unit selection synthesizers. As in prior studies optimizing the cost functions for concatenative synthesizers, the authors aimed to develop a metric they could use to better tune their cost functions. They investigate LSTM-based architectures with some automated hyperparameter tuning for predicting MOS of multiple years' worth of internal listening test datasets of ratings of several iterations of their US English TTS system. They found that there was inadvertently some predictive power of the text in their dataset, with more common utterance types having higher MOS. They also began using AutoMOS to tune the development of their

TTS, and they also suggest that it can be used in the future to automatically select utterances to include in human listening tests to make those tests more efficient.

MOSNet [75] was the first attempt to automatically predict the subjective quality of converted speech. The methodology was largely based on Quality-Net [76], a model for subjective assessment of enhanced speech. In the MOSNet paper, they used the VCC 2018 data to train CNN, BLSTM, and a combination CNN-BLSTM architecture with raw magnitude spectrogram input for MOS prediction, finding that the last one worked the best, with a system-level SRCC of around 0.9. The VCC evaluations also contain DMOS ratings for speaker similarity in addition to MOS ratings for naturalness, so the MOSNet authors also modified their system to predict speaker similarity by accepting two input audio samples, a test sample, and a reference, making MOSNet the first deep learning based model for both quality and similarity prediction of voice-converted samples. Their system was published as open-source code and it was trained on freely-available data, so it became a popular benchmark system in subsequent work.

A later study [73] trained the MOSNet architecture on ASVSpoof 2019 Logical Access data [23] which contains samples from both TTS and VC, comparing eight different input representations in addition to the original spectrogram input to determine the best one. These included image-based embeddings of the spectrogram as well as several x-vector [77] variants designed to extract different types of information, and it was found that the embeddings of spectrograms worked well for evaluating TTS systems. Crucially, they observed that pretrained MOSNets did not generalize well to new datasets and synthesis systems, and they recommend retraining MOSNet when switching datasets.

Some improvements to MOSNet were proposed [78] including the use of learned global quality tokens (GQT), inspired by global style tokens for TTS [79] and intended to reflect the criteria used by listeners in making their judgments, and an encoding layer that better aggregates frame-level scores into utterance-level ones by incorporating information about their distributions. The combination of the two proposed improvements was shown to improve MOS prediction on the in-domain VCC 2018 test set, but the original MOSNet had the best system-level correlations when testing on VCC 2016 in a cross-dataset condition, once again revealing the difficulty of cross-domain MOS prediction.

3.3.3. Listener modeling in synthetic speech evaluation

A technique that has been gaining attention is listener modeling. In MOS tests, ratings from multiple listeners are averaged together to get one value representing the quality of each utterance. This results in datasets that are a fraction of the size of the actual number of collected labels. With the intuition that modeling individual listeners' ratings could effectively increase the amount of available training data and explain variations in the scores, listener-dependent modeling has emerged as an effective approach to MOS prediction.

MBNet (Mean-Bias Network) [80] was the first work to consider the use of per-listener scores for MOS prediction. They incorporate a mean subnet and a bias subnet that allows the network to learn the personal preferences of individual listeners, which can vary widely, in addition to the averaged scores. The mean subnet predicts the averaged score similar to previous works, and the bias subnet predicts the difference between the mean score and an individual listener's score, given the listener ID. During inference, a specific listener ID can be input to predict his or her rating, or the bias net can be discarded and a prediction may be generated by the mean net only. Trained on VCC 2018 and evaluated on both its test set and the VCC 2016 data, MBNet was shown to improve over the MOSNet baseline.

LDNet (Listener-Dependent Network) [81] further proposed several improvements over MBNet. The authors first hypothesized that the speaker bias can be modeled using few parameters and made the bias net lightweight. Instead of discarding the bias net during inference, they further proposed two inference modes: (1) an "all listeners" mode, which averages over the predicted decisions of all of the listeners seen in training, and (2) a "mean listener" mode, where a "virtual" listener is created during training whose rating is always the mean score of a given audio sample. LDNet was shown to outperform MBNet, with the "mean listener" mode giving the best results.

DeePMOS (Deep Posterior Mean-Opinion-Score) [82] estimated a posterior Gaussian distribution of MOS ratings. This was accomplished by extending MBNet to output a predicted variance in addition to the mean. This approach improved MSE and system-level correlations over MBNet, and also provides more interpretable predictions in the form of distributions as opposed to point estimates.

3.3.4. SSL-based approaches

In parallel to listener modeling, another technique that has been gaining more attention is the adaptation of SSL-based speech models. The application of SSL to speech was first shown to produce excellent results in speech recognition [83], and has since become the dominant approach in almost all speech processing tasks [84]. Using an SSL-based speech model requires two stages: (1) self-supervised pre-training on large quantities of unlabeled speech audio for some pretext task, such as contrastive learning, as in the case of Wav2vec 2.0 [83], or prediction of masked regions, as in the case of HuBERT (Hidden-Unit

Bidirectional Encoder Representations from Transformers) [85], and (2) appending a task-specific prediction head on the SSL model and fine-tuning with a downstream labeled dataset. The representations learned in the first stage have been shown to have excellent capabilities for a large variety of speech downstream tasks at many levels, from phoneme recognition and speaker identification to emotion recognition and intent classification [86].

In 2021, the first effort to use SSL models for the MOS prediction task was made [87]. They compared the use of several different pretrained SSL models as encoders, followed by attention-based pooling of the output frame-level vectors, to the use of classical features like MFCCs. The SSL model parameters get updated during training with MOS-labeled data along with the rest of the model parameters. They also incorporated listener modeling in a similar manner to MBNet. As in previous studies, they trained on VCC2018 and tested on both VCC2018 and VCC2016, showing improvements in both cases but especially in the in-domain scenario.

The SSL-MOS model [88] used an even simpler SSL-based architecture for MOS prediction without any listener modeling, trained and evaluated on more diverse datasets including both voice conversion and TTS samples in several languages. Compared to MOSNet pretrained on VCC 2018 and finetuned to a mixed dataset of voice conversion and TTS samples, the SSL-based MOS predictors showed superior generalization ability to out-of-domain datasets with unseen systems, even showing reasonable correlations in the very challenging zero-shot condition.

A later work [89] considered that prosody and content are important to consider when evaluating naturalness. After collecting a large-scale dataset called SOMOS (Samsung Open MOS) [90] of MOS ratings for samples from 200 neural synthesizers with an emphasis on producing prosodic variations, they proposed a content-aware approach to MOS prediction. They included prosodic features (phoneme-level f0 and duration) and linguistic features extracted from the known text of the synthesized utterances (Tacotron [91] encoder outputs, part-of-speech tags, and BERT [92] embeddings), and added encoders for these features to several different MOS predictors trained on SOMOS. Their results demonstrated that the linguistic features gave improved results for SSL-MOS, especially at the utterance level, and also that training converged more quickly when these features were used.

The authors of a system called RAMP (Retrieval-Augmented MOS Prediction) [93] considered adding a non-parametric component to the decoder of an SSL-based MOS predictor based on k nearest neighbors, consisting of a datastore of the SSL representations of the training audio samples and their corresponding MOS scores. A score can be predicted by retrieving the k nearest neighbors of an input audio sample and obtaining a distance-weighted sum of their scores, which can then be fused with the prediction from the standard decoder. This approach was shown to improve predictions compared to basic SSL-MOS, especially for out-of-domain prediction.

The SQuId project [94], while also based on SSL, was the first massively multilingual research effort towards MOS prediction for synthesized speech. The basis of their model is a multi-modal language model pre-trained on a combination of unlabeled speech, text, and paired text and speech data. Audio is input as a spectrogram, and the model is fine-tuned for the MOS prediction task as regression. They trained it on MOS ratings from over 2000 internal projects for 52 language locales, and the evaluation data included several "zero-shot" locales that had been unseen during training. They found some benefits of transfer learning by including data from many languages, and they also found evidence that these benefits did not in fact depend on language similarity, indicating that there are aspects of naturalness that are not language-dependent.

3.3.5. Unsupervised approaches for synthetic speech quality prediction

The above-mentioned models are trained in a *supervised* manner on datasets with score labels, which can be difficult to collect. It is therefore of interest to develop *unsupervised* approaches that require no MOS-labeled data. A popular idea is to develop a *reference model*. This idea is opposed to the use of a *reference sample* in double-ended quality prediction methods, which suffers from two problems: (1) a reference sample may not always be available, and (2) the consequence of the "one-to-many" problem is that perfectly acceptable utterances may be unfairly penalized if they differ from a reference. In contrast, a reference *model* can be viewed as prior knowledge of natural speech, and it can be trained using a large amount of natural speech samples. Quality prediction could be achieved by measuring the distance between the input speech and natural speech defined by the reference model. This line of work shares the same concept with unsupervised anomaly detection [95]. Although this approach to synthesized speech quality prediction has remained largely under-explored, there have been several studies on this topic.

In 2008 [56], gender-dependent HMMs were trained on natural speech and used as reference models. A log-likelihood measure with respect to these reference models was then computed from features extracted from synthesized speech. These log-likelihoods were shown to be useful for predicting several quality dimensions of a P.85-type listening test, with MOS scores for overall impression, naturalness, and fluency having the best correlations.

Correlations on all eight of their rating scales were also higher with this log-likelihood measure than with P.563 scores. Because this study used "black box" commercial API-based synthesizers, they did not have access to the original training data for those systems, so they built their reference models using speech from different speakers.

A 2013 study [96] built Gaussian mixture models (GMMs) of the distributions of acoustic feature vectors extracted from the natural speech of their TTS training data, vocoded using the same vocoder as their HMM-based synthesis pipeline, resulting in reference models that are very well-matched to their target conditions. They measured the likelihoods of the reference GMMs with respect to their model's generated acoustic space and proposed this as an objective measure of synthesis quality. A later study [97], inspired by methods in automatic speech recognition, measured the Kullback-Leibler divergence between GMMs of natural and synthesized speech from three different HMM synthesizer variants. They found very high (over 0.9) correlations with MOS ratings.

Most recently, SpeechLMScore [98] measures the likelihood of a synthesized audio sample with respect to a generative speech unit language model which has been pretrained on natural speech. They evaluated their score's correlations with MOS ratings for voice conversion, TTS, and speech enhancement datasets. In particular, they found that their approach had better correlations than supervised systems that were trained on data from mismatched domains, indicating the superior generalization ability of this approach.

A paper analyzing the behavior of pretrained SSL models [99] demonstrated that uncertainty measures derived from these models correlated well with MOS ratings, without any finetuning for the MOS prediction task. The authors investigated a variety of different pretrained SSL models as well as languages of the synthesized speech and found that in particular, contrastively-pretrained wav2vec [100] models had the best correlations in all settings. These experiments demonstrate that even without being exposed to any synthesized speech data during training, SSL models still encode some information corresponding to human judgments of naturalness.

3.3.6. Beyond predicting quality of synthetic speech

Objective scores for synthesized speech may not only be used for evaluation, but also for other purposes such as system development. Introducing models of human perception into the development of speech synthesis engines can serve to produce synthesized speech that better matches human preferences. There were many such efforts during the age of concatenative speech synthesis. Concatenative text-to-speech synthesizers typically select speech units from a database of recordings by optimizing for a weighted combination of *target cost*, which measures how well a selected unit matches the desired phonetic sequence and the target prosodic aspects such as f0 and stress, and *join cost*, which penalizes unit concatenation boundaries that are audibly jarring or discontinuous. The main function of a concatenative synthesizer is to perform a search over the available speech units using dynamic programming to find the best sequence of units that optimizes both of these costs. One of the primary research tasks during the era of concatenative synthesis was to find the best features to express these costs, and to find the best way to weigh them. Much of this research used human listening test data to find expressions of these costs, especially the join cost, that best matched human perception.

Many of these works investigated different distance measures, such as Euclidean distance, the Kullback-Leibler measure, and Mahalanobis distance, of a variety of acoustic features that can be extracted from a speech audio signal, such as Mel-frequency cepstral coefficients (MFCCs), power spectra, and line spectrum frequencies [101–103], measuring correlations of the final costs of synthesized utterances with their human ratings. Cost weights could be optimized towards the perceptual results using methods such as downhill simplex [104], and the best ways to compute and combine costs from individual units were also considered, with one study [105] finding that the average cost of a unit sequence correlated better with human perception than maximum cost, and that the root mean squared cost, which combines both the average and the maximum, having the best correlation of all. Later work [106] considered how to extract common useful knowledge from differing listener opinions. Naturalness predictions were also applied in a dialogue system setting [107], where predictions of naturalness based on the costs reported by the synthesizer as well as textual features using knowledge of how often words appeared in the recording script for their unit selection database were used to choose the paraphrasing of the generated dialogue that was most likely to sound the best when realized by the synthesizer.

More recently [108], a Transformer-based TTS model [109] was trained using a loss function that included a predicted MOS term. While they did not observe significant improvements in the synthesized speech under normal training conditions when they included this term, they found that including it gave improvements under low-resource data scenarios and when using knowledge distillation to compress the model. Another recent paper [110] found that using MOS prediction in a loop of selecting multi-speaker audio data found online and training synthesizers on that data could help to quickly evaluate whether a given speaker's data can improve the synthesis model or not.

The works mentioned thus far mostly focused on speech quality assessment. However, certain dimensions in

synthetic speech can be important, depending on the target task. For instance, evaluating speaker similarity is essential in tasks such as VC, voice cloning, or speaker-adaptive TTS. While models like MOSNet can be modified to model speaker similarity, designing models tailored to similarity prediction is an underexplored field.

SVSNet [111] was based on MOSNet and focused on speaker similarity evaluation. The authors considered that symmetry needs to be enforced — speaker similarity ratings should remain the same regardless of the order in which the test and reference samples are presented. They achieved this using a co-attention mechanism which can also handle content and length mismatches between the two samples, and they also introduced raw waveform input using a SincNet [112] learnable filterbank in the encoder. They experimented with both regression- and classification-based learning objectives, finding that regression produced better results and that their proposed system overall made better speaker similarity predictions than MOSNet.

Although machine learning models of human judgments of speaker similarity can be trained from listening test data, it has instead become common practice to use cosine similarity between speaker embeddings extracted from an original audio sample from the target speaker and a synthesized one [113,114] using speaker recognition models that have been trained on data from thousands of speakers. An analysis of the results of VCC 2020 [115] showed high Pearson correlations above 0.8 between x-vector cosine similarities and listener ratings for speaker similarity.

Another task related to MOS prediction for synthesized speech is the task of spoofing detection. Multi-task learning was considered for MOS prediction for the first time in 2021 [116], where two auxiliary classification tasks were added to MOSNet: spoofing detection (a binary decision about whether the speech is synthesized or real), and spoofing type classification (a multi-class task that identifies which synthesis system was used). They found that both auxiliary tasks improved prediction on VCC data, with their combination producing the best results of system-level SRCCs of around 0.96. Conversely, MOS prediction was also shown to aid in the task of fake audio detection [117].

3.3.7. Predicting rank order and pairwise preferences

Critiques of MOS point out that MOS is not absolute, but highly affected by biases inherent in the entire context of the listening test. Therefore, MOS values cannot be meaningfully compared across different studies, and more importantly, MOS datasets should not be naively combined. This is a crucial limitation as modern neural networks require very large quantities of labeled training data. Some prior work [57] combined MOS datasets by using score normalization, while a later study [118] allowed the network to learn how to do the normalization by incorporating a bias-aware loss function that approximates the specific bias of each dataset with a first-order polynomial function. In effect, absolute errors caused by the biases are not penalized, and the model learns the correct *ranking order* within each dataset. A later study [119] introduced a loss function that measures the differences between the predicted MOS values between all pairs of samples in a mini-batch compared to the differences in their actual values, effectively measuring the correctness of the ranking of samples in a mini-batch. This approach was shown to have benefits over a traditional L1 loss, especially in zero-shot and semi-supervised scenarios.

A system called PrefNet [120] outputs the probability that one waveform would be preferred over another, given a pair of audio samples that are expected to contain the same lexical content but may vary in duration. Similar to SVSNet, it is important that the results are the same regardless of the order in which the two samples are input; this requirement is enforced by the use of anti-symmetric twin neural networks, and the durational alignment problem is solved using attention and RNNs. They derived large-scale pairwise preference data from several MUSHRA tests by labeling pairs of audio with how often one sample was rated more highly than the other. They evaluate their model using percent accuracy rather than correlations; depending on the testing conditions, accuracies ranged from about 50–100%. A later work aiming to predict pairwise preferences [121] derived a dataset of preference scores from pairs of MOS ratings from the same listener, intending to reduce listener bias — they found that this "same-listener" constraint did in fact result in predictions with better correlations. Their proposed model predicts utterance scores for two samples, from which it can produce a preference score and also system-level scores aggregated over all utterances from a given system. Experimental results showed significant improvements over a model that only predicts MOS for one input audio sample.

3.3.8. Learning from speech quality prediction in other domains

In parallel with research on opinion predictors for synthesized speech, researchers in speech quality estimation for telephony and speech enhancement have also been considering the use of deep neural networks for predicting MOS ratings [122,123] and relative comparisons [124] of degraded and processed natural speech. Although the types of degradations that appear in synthesized speech are different from degradations to natural speech due to noise and processing, these tasks are nevertheless very related and it may be beneficial to consider approaches such as transfer learning.

In one such effort [125], the authors use a CNN-LSTM architecture that they had previously developed and trained for speech quality estimation of degraded natural speech [126], and then they continued training the model on MOS datasets for synthesized speech, including Blizzard and VCC. They considered that it was important to evaluate on unseen synthesis systems and so they held out some of the datasets for this purpose, unlike many past works which use a randomly selected test set. Despite evaluating on challenging unseen conditions, they nevertheless had best correlations upwards of 0.9, although interestingly they also found that ablating the pretraining by replacing it with randomly-initialized values only degraded the correlations by a small amount.

4. THE VOICEMOS CHALLENGE 2022

In 2022, we launched the first VoiceMOS Challenge [127], a shared task on the topic of MOS prediction for synthesized speech, with the goals of encouraging research on this topic and of unifying datasets and evaluation to make direct comparisons between different approaches. The challenge attracted 22 participating teams from academia and industry, accelerated research and generated discussion on this topic.

4.1. Data and Tracks

There were two tracks in the VoiceMOS Challenge 2022, namely the main track and the out-of-domain track. Table 1 summarizes the datasets.

In 2021, we collected a large-scale dataset of MOS ratings for synthesized audio samples as well as reference natural speech samples. We gathered English-language synthesized audio samples from several past Blizzard Challenges from 2008–2016 [65,66,128–131], as well as from all previous Voice Conversion Challenges [69–71]. We also wanted to include samples from more recent systems, so we added publicly-shared samples from ESPnet-TTS [132]. Altogether, we collected samples from 187 different systems (where natural reference speech for each challenge is considered a "system" as well) and selected 38 samples per system. We conducted a large-scale listening test in which we obtained 8 ratings per sample from 304 unique listeners. This data is described in detail in our prior work [133] and has been publicly released[†] as the BVCC dataset. This dataset provided the material for the main track of the challenge. We created standard training, development, and testing splits of the data containing 70%, 15%, and 15% of the data respectively, holding out some unseen speakers, synthesis systems, and listeners in the development and test sets.

Table 1 Summary of the main track and out-of-domain (OOD) track datasets in the VoiceMOS Challenge 2022.

Track	Lang	# Samples			# ratings per sample
		Train	Dev	Test	
Main	Eng	4,974	1,066	1,066	8
OOD	Chi	Label: 136 Unlabel: 540	136	540	10–17

Additionally, we wanted to encourage participants to investigate semi-supervised and unsupervised approaches to MOS prediction, so we also ran an out-of-domain (OOD) track in the challenge. We used listening test data from the Blizzard Challenge 2019 including the original MOS ratings from the challenge evaluation, and provided participants with 10% labeled training data and 40% unlabeled (audio samples only) training data. The remaining data was divided into a 10% development set and a 40% test set, again including unseen systems and listeners (but not speakers, as all of the samples were from models trained on one speaker's data). Blizzard 2019 focused on Mandarin-language text-to-speech synthesis, so this track was challenging both in terms of the smaller amount of labeled training data as well as the language mismatch with respect to the main track.

The evaluation metrics used in the challenge were system-level and utterance-level MSE, LCC, SRCC, and KTAU, and system-level SRCC was chosen as the primary metric.

4.2. Baselines

Baseline systems were a simple SSL-based MOS predictor (SSL-MOS) fine-tuned on the BVCC data [88], LDNet also trained on BVCC [81], and MOSA-Net [123] which is also trained on BVCC. All three baselines are publicly available, where participants were given access to the pretrained models as well as the recipes for training, finetuning, and making predictions on the challenge datasets. These baselines represent a range of approaches, with MOSA-Net using features extracted from pretrained SSL models and other sources, SSL-MOS conducting finetuning of SSL models, and LDNet conducting listener-dependent modeling.

4.3. Team Approaches

In the following, we reference and briefly summarize papers released by participants, which either described their submitted system or provided an analysis.

- The UTMOS (University of Tokyo MOS) system (T17) [134] was one of the best-performing systems, scoring the highest on several metrics. It ensembles strong and weak learners. The strong learners were

[†]https://doi.org/10.5281/zenodo.6572573

modified SSL-MOS systems with additional techniques, including contrastive learning, listener-dependent modeling, and phoneme encodings. The weak learners were regression models including linear regression, decision tree, and kernel methods. The inputs to these weak learners were SSL features. Finally, the team conducted a listening test with the unlabeled set in the OOD track and added a listening test ID to combine multiple datasets from different listening tests. They were in fact one of the few teams that made use of the unlabeled set.

- The DDOS (Domain adaptive pre-training and Distribution of Opinion Scores) system (T19) [135] ranked second in three out of the four system-level metrics in the main track. In their system, in addition to regressing to the MOS score, they also modeled the distribution of the MOS ratings. They also applied data augmentation by changing the voice pitch. On the OOD track, they used a domain-adaptive pre-training technique which reduced MSE. They also provided zero-shot results on the OOD track.
- The ZevoMOS system (T01) [136] was based on SSL-MOS, but they used two SSL inputs and an ASR confidence score. Moreover, the SSL models were first fine-tuned on the FoR dataset [137] to classify natural and synthetic speech, then fine-tuned on the BVCC dataset.
- The system by JAIST (T08) [138] was based on MOSA-Net with two key concepts: an auditory filterbank and temporal modulation. A temporal modulation feature on the gammatone filterbank (TMGF) was concatenated with the HuBERT features. They showed that this method could improve prediction on utterances with a low MOS.
- The system from NICT (T11) [139] ranked first in LCC, SRCC, and KTAU in both the main track system level metrics and the OOD track utterance level metrics. They proposed a fusion framework exploiting seven SSL models. For the OOD track, they applied semi-supervised learning to the unlabeled set, which was shown to be very effective.
- The system from ByteDance AI-LAB (T20) [140] ranked 4th in terms of both system- and utterance-level SRCC. It was based on LDNet, and they combined the main and OOD track datasets with a shared encoder and separate decoders. The encoder was essentially a wav2vec 2.0 model fine-tuned for phoneme recognition.
- MooseNet [141] was based on SSL-MOS, and their main idea was to apply PLDA. It transformed frame-by-frame acoustic features into time-invariant features by global pooling, an operation similar to that used to compute speaker vectors for speaker recognition tasks. Applying PLDA showed improvements compared to the vanilla SSL-MOS.
- A comparison of SSL features and raw acoustic features like spectrograms was made in [142]. Starting from LDNet, they showed that combining wav2vec 2.0 and Mel spectrograms or F0 values can improve the performance, implying that there is complementary information found in raw acoustic features.
- An analysis focused on various factors when fine-tuning SSL models [143]. Starting from SSL-MOS based on wav2vec 2.0, they experimented with not only synthetic speech but also natural speech in noisy environments and transmitted over communication networks, and showed that fine-tuning with mixed-lingual datasets and larger dataset sizes could improve generalization performance.
- Another analysis focused on the metadata of the BVCC dataset [144]. They used the SSL-MOS model and added metadata information. They showed the amount of error and correlation that can be explained by metadata predictors such as system and rater identifiers. They also showed that since there were often only very few utterances per system in the development and test sets, utterance-level metrics were more informative than the system-level ones.

4.4. Lessons Learned

Overall, we observed that finetuning SSL models for the MOS prediction task is a powerful approach that can produce predictions with very high correlations with real listener ratings. However, we observed that predictions for unseen systems in the OOD track were substantially more difficult. This is important because this case corresponds the most with a real-life use case for MOS predictors — predicting MOS for a system which has not been evaluated in a listening test before, and therefore for which no MOS labels already exist. Furthermore, we asked participating teams to fill out a survey including questions about what types of tasks they would like to see in future challenges, and many responses were about including a larger variety of audio to evaluate, including synthesis in more different languages, singing synthesis, and noisy and enhanced speech.

5. THE VOICEMOS CHALLENGE 2023

The outcomes of the first challenge motivated our design of the 2023 edition of the challenge [145]. We focused on real-life MOS prediction in a variety of speech domains. In the 2023 challenge, we did not provide any MOS-labeled audio samples in two of the target domains, and listening tests were ongoing at the same time as the challenge, meaning that team predictions were made before the actual ground-truth MOS values were known to

Table 2 Summary of the test phase data for each track in the VoiceMOS Challenge 2023.

Track	Type	Lang	Systems	Samples per system	# ratings per sample
Track 1a Track 1b	TTS	Fre	Hub: 21 Spoke: 17	42 34	15
Track 2	Singing VC	Eng	In-dom: 25 Cross-dom: 24	80	6
Track 3	Noisy & enhanced	Chi	97	20	5.3

anyone. In total, 10 teams participated in this year's challenge.

5.1. Data and Tracks

There were three tracks in the VoiceMOS Challenge 2023. Table 2 summarizes the datasets for each track.

We collaborated with the organizers of the Blizzard Challenge 2023 [146] as well as the Singing Voice Conversion Challenge 2023 (SVCC) [72] to acquire synthesized samples from their teams for conducting MOS prediction by our teams, while the Blizzard and SVCC listening tests were still ongoing.

The Blizzard Challenge 2023 focused on French text-to-speech synthesis, with a Hub track of their challenge providing 51 hours of training data from a single speaker, and a Spoke track providing 2 hours of data from a different speaker, intended for speaker adaptation. Therefore, Track 1 of the VoiceMOS Challenge 2023 was French TTS, and since the Blizzard listening tests were conducted separately for their Hub and Spoke tasks, we likewise divided this track into corresponding Tracks 1a and 1b.

Since spoken voice conversion has reached near-human levels of naturalness [71], in 2023, the VCC organizers decided to focus on the task of singing voice conversion — that is, converting a sung audio sample to a different speaker identity, using either sung (matched) or spoken (mismatched) reference audio from the target speaker. Track 2 of the VoiceMOS Challenge 2023 was therefore singing voice conversion.

There was substantial interest from the participating VoiceMOS teams in 2022 to expand to noisy and enhanced speech, and we also noticed many parallel efforts towards more automatic evaluation methodologies in the speech synthesis and speech enhancement communities. Considering that these are similar tasks and that there could be benefits from more communication and collaboration between these communities, Track 3 was noisy and enhanced speech. Unlike the other two tracks, where no MOS-labeled training data was provided to participants, we provided the TMHINT-QI [147] dataset as training data, with a held-out development set providing evaluation material for displaying team scores on a leaderboard on our challenge website during the initial few weeks of the challenge to encourage early participation and friendly competition. During the evaluation phase, a separate test set called the TMHINT-QI2 [148] was curated, with the same noise generation process, partially different speech enhancement systems, and completely different raters.

5.2. Baselines

The baseline systems we included were SSL-MOS [88], which had been the best-performing baseline in the previous challenge, and UTMOS [134], which was one of the top team systems from the 2022 challenge that also had an open-source implementation. We used models that were pretrained on BVCC as baselines without any additional development.

5.3. Team Approaches

At the time of writing, only two teams have released papers describing their systems, so we will briefly summarize them below.

- The LE-SSL-MOS (Listener-Enhanced Self Supervised Learning Mean Opinion Score) system (T06) [149] showed promising results on all tracks. This was considered impressive since most teams did well on one track and performed badly on the other tracks. There were several key ideas. First, they employed model ensembling, combining scores for multiple models. These models include supervised learners, including a vanilla SSL-MOS model and an SSL-MOS model enhanced with listener-dependent modeling. They also included unsupervised learners, where "unsupervised" was defined as not using any MOS labels during training. These unsupervised approaches include a fine-tuned SpeechLMScore [98] model, as well as ASR confidence scores.
- The SQAT-LD system (Speech Quality Assessment Transformer) (T03) ranked 4th in Track 1a, 2nd in Track 1b, and 1st in Track 2 [150]. They also combined SSL-MOS with listener-dependent modeling, where their SSL model was SSAST [151]. They also proposed to combine the weighted scores of each frame to better predict the overall score. Their model was trained on the main and OOD datasets from the VoiceMOS Challenge 2022, and they also included a bias-aware loss [118] to enable training on multiple datasets.

5.4. Lessons Learned

From the system descriptions that the teams submitted, we found that listener-dependent modeling was more popular this year, and teams that used a mix of different

training datasets (BVCC, SOMOS [90], past Blizzard original data, etc.) also tended to do better. Although we shared pointers to the training datasets of the Blizzard Challenge 2023 and Singing Voice Conversion Challenge 2023, as well as other relevant datasets without MOS labels, no teams made use of those datasets.

We were surprised that many teams had good prediction results for the singing track, especially since none of the teams reported using any singing data to develop their systems. We suspect that the domain mismatch between synthesized singing and speech is not as large as we had assumed. Furthermore, we also observed that many teams had large gaps between their results for Tracks 1a and 1b, although not in any consistent direction across teams. Upon investigating the training data for the Blizzard Hub and Spoke tasks, we observed that the Spoke data contained audible reverberation whereas the Hub data did not, which may have been one of the reasons for this result.

For Track 3, we observed generally higher scores, where some training data was made available, compared to the other tracks, where there was not. We also observed that most teams' scores for the different tracks are very different, and no team had high scores on all tracks using the same model trained on the same data, indicating that general-purpose MOS prediction can still be considered an open research problem.

6. FUTURE PROSPECTS AND CHALLENGES

Researchers in speech synthesis have long considered the best ways to compare and evaluate speech synthesis methods, and reliable objective evaluation metrics have been a long sought-after goal. Several decades ago, listeners visited research laboratories in person to listen to synthesized samples and transcribe them by hand. Now, crowdsourced MOS tests can be conducted quickly and conveniently, and powerful self-supervised speech representations and large-scale MOS datasets have brought us closer to the goal of objective metrics. Predicted MOS values are already being reported in some TTS and voice conversion papers as an objective evaluation metric alongside subjective listening test results [72,152] as well as other now commonly accepted automatic measures such as ASR word error rate and cosine similarities of speaker embeddings. MOS predictors are also being used in research applications such as in loss functions for training TTS systems [108], data selection for building TTS models from found data [110], and to aid in fake audio detection systems [117].

There have been several studies demonstrating that MOS tests may have become saturated and lost their ability to meaningfully differentiate between modern-day synthesizers [38,40,43]. Pairwise comparison tests have been shown to mitigate this. The testing material can also be chosen to better highlight differences between systems, thereby making listening tests more efficient [153], a task that some have suggested can be facilitated by automatic quality predictors as well [74].

Zero-shot general-purpose quality prediction of synthesized speech still remains an open research problem, with calibration to different domains and listening test contexts remaining a challenge. MOS predictors can be used for applications where the audio data is from a similar domain to that with which the predictor was trained; however, what constitutes "similar enough" still remains an open question. Care must be taken when reporting and understanding objective evaluation results given by MOS predictors, and we still need to accumulate more knowledge on MOS predictors and their behavior on different out-of-domain datasets before we can fully accept them as a replacement for human listening tests.

With the increasing attention being paid to the problems with MOS tests and "naturalness" as a target, there is a growing interest in other evaluation methodologies and their automation. Pairwise preference predictors are one step in this direction, and, more generally, objective evaluation methods that can directly output a ranking of multiple systems as opposed to MOS values would be interesting future work. There is also substantial evidence that MOS as a listening test methodology is no longer sufficient. It is important to consider more comprehensive listening test methodologies that consider factors such as context appropriateness and other aspects of listener opinions, as well as how we can incorporate these factors into automated evaluations. Data scarcity will always be an issue in terms of the availability of MOS-labeled data for every possible domain, context, or question that we ask listeners, so unsupervised and semi-supervised methods are an important future research direction. Methods that enable the combination of smaller or heterogeneous datasets, such as models that learn pairwise predictions or rankings, will be useful for addressing this as well. Furthermore, *interpretable* opinion prediction for synthesized speech remains under-explored — there are many possible reasons why a listener might assign a sample a low score, and knowing the reason why a sample's predicted quality is low would be very useful from a diagnostic point of view. This line of research would first require a better understanding of how listeners assign their ratings, and studies asking listeners for reasons or explanations for their ratings have been an important step in this direction. Conversely, if MOS predictors become very accurate and interpretable, we can consider using them as psychoacoustic tools to better understand human perception of speech. In the long term, we aim to be able to comprehensively model human preferences about synthesized

speech, including arbitrary aspects of human opinions, and to be able to use those predictions during model development to produce the next generation of more realistic, diverse, and context-adaptable synthesized speech.

ACKNOWLEDGMENTS

We would like to thank the organizers of the Blizzard Challenges and Voice Conversion Challenges as well as the authors of ESPnet-TTS for making their audio samples and listening test data available. We would also like to thank the participants of the VoiceMOS Challenge for engaging with this topic and making the challenges a success. We would like to thank Sébastien Le Maguer, Yusuke Yasuda, and Gustav Eje Henter for many valuable discussions on this topic. This work was supported by JST CREST Grant Numbers JPMJCR18A6 and JPMJCR19A3, MEXT KAKENHI grant 21K11951, and MOST 110-2221-E-001-015-MY3.

REFERENCES

[1] R. Van Bezooijen and L. C. Pols, "Evaluating text-to-speech systems: Some methodological aspects," *Speech Commun.*, **9**, 263–270 (1990).

[2] A. S. House, C. Williams, M. H. Hecker and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *J. Acoust. Soc. Am.*, **35**, 1899 (1963).

[3] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technol.*, **1**, 30–39 (1983).

[4] M. Spiegel, M. J. Altom, M. Macchi and K. Wallace, "A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech," *Proc. Speech Input/Output Assessment and Speech Databases*, Vol. 2, pp. 5–10 (1989).

[5] U. Jekosch, "The cluster-based rhyme test: A segmental synthesis test for open vocabulary," *Proc. Speech Input/Output Assessment and Speech Databases*, Vol. 2, pp. 15–18 (1989).

[6] J. P. van Santen, "Perceptual experiments for diagnostic testing of text-to-speech systems," *Comput. Speech Lang.*, **7**, 49–100 (1993). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230883710041

[7] M. Grice, "Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages," *Proc. Speech Input/Output Assessment and Speech Databases*, Vol. 2, pp. 19–22 (1989).

[8] D. Pisoni and S. Hunnicutt, "Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.,* (*ICASSP*) *'80*, Vol. 5, pp. 572–575 (1980).

[9] "Methods for subjective determination of transmission quality," *ITU-T Rec. P.800*, International Telecommunication Union (ITU-R) (1996).

[10] M. Goldstein, B. Lindström and O. Till, "Some aspects on context and response range effects when assessing naturalness of Swedish sentences generated by 4 synthesiser systems," *Proc. 2nd Int. Conf. Spoken Lang. Process.* (*ICSLP 1992*), pp. 1339–1342 (1992).

[11] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Commun.*, **16**, 225–244 (1995). [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939400047E

[12] "A method for subjective performance assessment of the quality of speech voice output devices," *ITU-T Rec. P.85*, International Telecommunication Union (ITU-R) (1994).

[13] "Methods for subjective determination of transmission quality," *ITU-T Rec. P.80*, International Telecommunication Union (ITU-R) (1993).

[14] C. Benoît, M. Grice and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Commun.*, **18**, 381–392 (1996). [Online]. Available: https://www.sciencedirect.com/science/article/pii/016763939600026X

[15] S. Itahashi, "Guidelines for Japanese speech synthesizer evaluation," *Proc. 2nd Int. Conf. Lang. Resour. Eval.* (*LREC'00*), M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer, Eds. (European Language Resources Association (ELRA), Athens, Greece, 2000). [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/pdf/77.pdf

[16] Y. V. Alvarez and M. Huckvale, "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems," *Proc. 7th Int. Conf. Spoken Lang. Process.* (*ICSLP*) *2002*, pp. 329–332 (2002).

[17] D. Sityaev, K. Knill and T. Burrows, "Comparison of the ITU-T P.85 standard to other methods for the evaluation of text-to-speech systems," *Proc. Interspeech 2006*, paper 1233–Tue2WeO.3 (2006).

[18] L. C. W. Pols and U. Jekosch, *A Structured Way of Looking at the Performance of Text-to-Speech Systems* (Springer New York, New York, 1997), pp. 519–527. [Online]. Available: https://doi.org/10.1007/978-1-4612-1894-4_41

[19] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets," *Proc. Interspeech 2005*, pp. 77–80 (2005).

[20] N. Campbell, *Evaluation of Speech Synthesis* (Springer Netherlands, Dordrecht, 2007), pp. 29–64. [Online]. Available: https://doi.org/10.1007/978-1-4020-5817-2_2

[21] S. Zielinski, F. Rumsey and S. Bech, "On some biases encountered in modern audio quality listening tests-A review," *J. Audio Eng. Soc.*, **56**, 427–451 (2008).

[22] K. Tokuda, H. Zen and A. W. Black, "An HMM-based speech synthesis system applied to English," *Proc. IEEE Speech Synthesis Workshop*, IEEE Santa Monica, pp. 227–230 (2002).

[23] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang and Z.-H. Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Comput. Speech Lang.*, **64**, p. 101114 (2020). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230820300474

[24] F. Ribeiro, D. Florêncio, C. Zhang and M. Seltzer, "CrowdMOS: An approach for crowdsourcing mean opinion score studies," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2011*, pp. 2416–2419 (2011).

[25] S. Buchholz, J. Latorre and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," in *Crowdsourcing for Speech Processing: Applications to Data, Collection, Transcription and Assessment*, M. Eskénazi, G.-A. Levow, H. Meng, G. Parent and D. Suendermann, Eds. (John Wiley &

Sons, Chichester, 2013), Chap. 7, pp. 173–214.

[26] M. Wester, C. Valentini-Botinhao and G. E. Henter, "Are we using enough listeners? No! — An empirically-supported critique of interspeech 2014 TTS evaluations," *Proc. Interspeech 2015*, pp. 3476–3480 (2015).

[27] "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *Recommendation ITU-R BS.1534-3*, International Telecommunication Union (ITU-R) (2015).

[28] R. C. Streijl, S. Winkler and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, **22**, 213–227 (2016).

[29] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander and J. Voße, "Speech synthesis evaluation: State-of-the-art assessment and suggestion for a novel research program," *Proc. 10th ISCA Workshop on Speech Synthesis* (*SSW 10*), pp. 105–110 (2019).

[30] M. Wester, O. Watts and G. E. Henter, "Evaluating comprehension of natural and synthetic conversational speech," *Proc. Int. Conf. Speech Prosody 2016*, pp. 766–770 (2016).

[31] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to TTS evaluation," *Proc. Interspeech 2017*, pp. 249–253 (2017).

[32] R. Clark, H. Silen, T. Kenter and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *Proc. 10th ISCA Workshop on Speech Synthesis* (*SSW 10*), pp. 99–104 (2019).

[33] J. O'Mahony, P. O. Gallegos, C. Lai and S. King, "Factors affecting the evaluation of synthetic speech in context," *Proc. 11th ISCA Speech Synthesis Workshop* (*SSW11*), International Speech Communication Association, pp. 148–153 (2021).

[34] R. Dall, J. Yamagishi and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," *Proc. 7th Int. Conf. Speech Prosody 2014*, pp. 1012–1016 (2014). [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2014-191

[35] S. Shirali-Shahreza and G. Penn, "Better replacement for TTS naturalness evaluation," *Proc. 12th ISCA Speech Synthesis Workshop* (*SSW2023*), pp. 197–203 (2023).

[36] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, **1**, p. e006 (2014). [Online]. Available: https://loquens.revistas.csic.es/index.php/loquens/article/view/6

[37] F. Seebauer, M. Kuhlmann, R. Haeb-Umbach and P. Wagner, "Re-examining the quality dimensions of synthetic speech," *Proc. 12th ISCA Speech Synthesis Workshop* (*SSW2023*), pp. 34–40 (2023).

[38] S. Shirali-Shahreza and G. Penn, "MOS naturalness and the quest for human-like speech," *Proc. 2018 IEEE Spoken Language Technology Workshop* (*SLT*), pp. 346–352 (2018).

[39] J. Camp, T. Kenter, L. Finkelstein and R. Clark, "MOS vs. AB: Evaluating text-to-speech systems reliably using clustered standard errors," *Proc. Interspeech 2023*, pp. 1090–1094 (2023).

[40] Y. Yasuda and T. Toda, "Analysis of mean opinion scores in subjective evaluation of synthetic speech based on tail probabilities," *Proc. Interspeech 2023*, pp. 5491–5495 (2023).

[41] E. Cooper and J. Yamagishi, "Investigating range-equalizing bias in mean opinion score ratings of synthesized speech," *Proc. Interspeech 2023*, pp. 1104–1108 (2023).

[42] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely and J. Gustafson, "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," *Proc. 12th ISCA Speech Synthesis Workshop* (*SSW2023*), pp. 41–47 (2023).

[43] S. Le Maguer, S. King and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Comput. Speech Lang.*, **84**, p. 101577 (2024). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000967

[44] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. 40th Annu. Meet. Assoc. Computational Linguistics*, pp. 311–318 (2002).

[45] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Proc. IEEE Pacific Rim Conf. Communications Computers and Signal Processing*, Vol. 1, pp. 125–128 (1993).

[46] J. Kominek, T. Schultz and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean Mel cepstral distortion," *Proc. Speech Technology for Under-Resourced Languages* (*SLTU-2008*), pp. 63–68 (2008).

[47] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862* (2001).

[48] S. Ipswich, "PESQ: An Introduction White Paper" (2001).

[49] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," *Proc. Eur. Congr. Acoustics*, pp. 2725–2728 (2005).

[50] F. Hinterleitner, S. Zabel, S. Möller, L. Leutelt and C. Norrenbrock, "Predicting the quality of synthesized speech using reference-based prediction measures," in *Konferenz Elektronische Sprachsignalverarbeitung* (TUDpress, Dresden, 2011), pp. 99–106.

[51] L. Latacz and W. Verhelst, "Double-ended prediction of the naturalness ratings of the Blizzard Challenge 2008–2013," *Proc. Interspeech 2015*, pp. 3486–3490 (2015).

[52] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T Rec. P.563* (2004).

[53] L. Malfait, J. Berger and M. Kastner, "P. 563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio Speech Lang. Process.*, **14**, 1924–1934 (2006).

[54] D.-S. Kim and A. Tarraf, "Anique+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs. Tech. J.*, **12**, 221–236 (2007).

[55] T. H. Falk, S. Möller, V. Karaiskos and S. King, "Improving instrumental quality prediction performance for the Blizzard Challenge," *Proc. Blizzard Challenge Workshop* (2008).

[56] T. H. Falk and S. Moller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Process. Lett.*, **15**, 781–784 (2008).

[57] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks," *Proc. Interspeech 2016*, pp. 342–346 (2016).

[58] R. Clark and K. Dusterhoff, "Objective methods for evaluating synthetic intonation," *Proc. 6th Euro. Conf. Speech Communication and Technology* (*Eurospeech*) *'99*, Budapest, Hungary, pp. 1623–1626 (1999).

[59] U. Remes, R. Karhila and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," *Proc. 8th ISCA Workshop on Speech Synthesis* (2013).

[60] F. Hinterleitner, S. Zander, K.-P. Engelbrecht and S. Möller, "On the use of automatic speech recognizers for the quality and intelligibility prediction of synthetic speech," *Proc. Konferenz Elektronische Sprachsignalverarbeitung* (TUDpress, Dresden, 2015), pp. 105–111.

[61] O. Sharoni, R. Shenberg and E. Cooper, "SASPEECH: A Hebrew single speaker dataset for text to speech and voice conversion," *Proc. Interspeech 2023* (2023).
[62] S. Mehta, R. Tu, J. Beskow, É. Székely and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2024* (2024) (to appear).
[63] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," *Proc. Int. Conf. Learning Representations* (2018).
[64] F. Hinterleitner, S. Möller, T. H. Falk and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009," *Proc. Blizzard Challenge Workshop*, Vol. 2010, pp. 48–60 (2010).
[65] V. Karaiskos, S. King, R. A. Clark and C. Mayo, "The Blizzard Challenge 2008," *Proc. Blizzard Challenge Workshop*, Citeseer (2008).
[66] A. W. Black, S. King and K. Tokuda, "The Blizzard Challenge 2009," *Proc. Blizzard Challenge Workshop*, pp. 1–24 (2009).
[67] S. King and V. Karaiskos, "The Blizzard Challenge 2011," *Proc. Blizzard Challenge Workshop* (2011).
[68] C. R. Norrenbrock, F. Hinterleitner, U. Heute and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks: Blizzard Challenge 2012," *Proc. Blizzard Challenge Workshop* (2012).
[69] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu and J. Yamagishi, "The Voice Conversion Challenge 2016," *Proc. Interspeech 2016*, pp. 1632–1636 (2016).
[70] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," *Proc. Speaker and Language Recognition Workshop* (*Odyssey 2018*), pp. 195–202 (2018).
[71] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling and T. Toda, "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *Proc. Jt. Workshop BC and VCC 2020*, pp. 80–98 (2020).
[72] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi and T. Toda, "The Singing Voice Conversion Challenge 2023," *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*) (2023).
[73] J. Williams, J. Rownicka, P. Oplustil and S. King, "Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis," *Proc. Speaker and Language Recognition Workshop* (*Odyssey 2020*), pp. 222–229 (2020).
[74] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv: 1611.09207* (2016).
[75] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," *Proc. Interspeech 2019*, pp. 1541–1545 (2019).
[76] S.-W. Fu, Y. Tsao, H.-T. Hwang and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," *Proc. Interspeech 2018*, pp. 1873–1877 (2018).
[77] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2018*, pp. 5329–5333 (2018).
[78] Y. Choi, Y. Jung and H. Kim, "Deep MOS predictor for synthetic speech using cluster-based modeling," *Proc. Interspeech 2020*, pp. 1743–1747 (2020).
[79] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *Proc. Int. Conf. Machine Learning* (*PMLR*), pp. 5180–5189 (2018).
[80] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2021*, pp. 391–395 (2021).
[81] W.-C. Huang, E. Cooper, J. Yamagishi and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2022*, pp. 896–900 (2022).
[82] X. Liang, F. Cumlin, C. Schüldt and S. Chatterjee, "DeePMOS: Deep posterior mean-opinion-score of speech," *Proc. Interspeech 2023*, pp. 526–530 (2023).
[83] A. Baevski, Y. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, **33**, 12 449–12 460 (2020).
[84] A. Mohamed, H.-Y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE J. Sel. Top. Signal Process.*, **16**, 1179–1210 (2022).
[85] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29**, 3451–3460 (2021).
[86] S. W. Yang, P. H. Chi, Y. S. Chuang, C. I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed and H.-y. Lee, "SUPERB: Speech processing Universal PERformance Benchmark," *Proc. Interspeech 2021*, pp. 3161–3165 (2021).
[87] W.-C. Tseng, C.-Y. Huang, W.-T. Kao, Y. Y. Lin and H.-Y. Lee, "Utilizing self-supervised representations for MOS prediction," *Proc. Interspeech 2021*, pp. 2781–2785 (2021).
[88] E. Cooper, W.-C. Huang, T. Toda and J. Yamagishi, "Generalization ability of MOS prediction networks," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2022*, pp. 8442–8446 (2022).
[89] A. Vioni, G. Maniati, N. Ellinas, J. S. Sung, I. Hwang, A. Chalamandaris and P. Tsiakoulis, "Investigating content-aware neural text-to-speech MOS prediction using prosodic and linguistic features," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2023*, pp. 1–5 (2023).
[90] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris and P. Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the evaluation of neural text-to-speech synthesis," *Proc. Interspeech 2022*, pp. 2388–2392 (2022).
[91] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010 (2017).
[92] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language

understanding," *Proc. NAACL-HLT 2019*, pp. 4171–4186 (2019).

[93] H. Wang, S. Zhao, X. Zheng and Y. Qin, "RAMP: Retrieval-augmented MOS prediction via confidence-based dynamic weighting," *Proc. Interspeech 2023*, pp. 1095–1099 (2023).

[94] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh and J. Riesa, "SQuId: Measuring speech naturalness in many languages," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2023*, pp. 1–5 (2023).

[95] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, **22**, 85–126 (2004).

[96] S. Le Maguer, N. Barbot and O. Boeffard, "Evaluation of contextual descriptors for HMM-based speech synthesis in French," *Proc. 8th ISCA Speech Synthesis Workshop* (*SSW 8*) (2013).

[97] C. T. Do, M. Evrard, A. Leman, C. d'Alessandro, A. Rilliard and J. L. Crebouw, "Objective evaluation of HMM-based speech synthesis system using Kullback-Leibler divergence," *Proc. Interspeech 2014*, pp. 2952–2956 (2014).

[98] S. Maiti, Y. Peng, T. Saeki and S. Watanabe, "Speech-LMScore: Evaluating speech generation using speech language model," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2023*, pp. 1–5 (2023).

[99] A. Ravuri, E. Cooper and J. Yamagishi, "Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot MOS prediction," *Proc. IEEE ICASSP 2024 Workshop Self-supervision in Audio, Speech and Beyond* (2024) (to appear).

[100] S. Schneider, A. Baevski, R. Collobert and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, pp. 3465–3469 (2019).

[101] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *Proc. 5th Int. Conf. Spoken Lang. Process.* (*ICSLP*) *1998*, paper 0115 (1998).

[102] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) (*Cat. No. 01CH37221*), Vol. 2, pp. 837–840 (2001).

[103] J. Vepa, S. King and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," *Proc. 7th Int. Conf. Spoken Lang. Process.* (*ICSLP*) *2002*, pp. 2605–2608 (2002).

[104] M. Lee, "Perceptual cost functions for unit searching in large corpus-based concatenative text-to-speech," *Proc. EUROSPEECH*, Aalborg, Denmark, pp. 2227–2230 (2001).

[105] T. Toda, H. Kawai, M. Tsuzaki and K. Shikano, "Perceptual evaluation of cost for segment selection in concatenative speech synthesis," *Proc. 2002 IEEE Speech Synthesis Workshop* (*SSW 2002*), pp. 183–186 (2002).

[106] L. Formiga and F. Alías, "Extracting user preferences by GTM for AiGA weight tuning in unit selection text-to-speech synthesis," *Proc. Int. Work-Conf. Artificial Neural Networks*, pp. 654–661 (2007).

[107] C. Nakatsu and M. White, "Learning to say it well: Reranking realizations by predicted synthesis quality," *Proc. 21st Int. Conf. Computational Linguistics and 44th Annu. Meet. Assoc. Computational Linguistics*, pp. 1113–1120 (2006).

[108] Y. Choi, Y. Jung, Y. Suh and H. Kim, "Learning to maximize speech quality directly using MOS prediction for neural text-to-speech," *IEEE Access*, **10**, 52 621–52 629 (2022).

[109] N. Li, S. Liu, Y. Liu, S. Zhao and M. Liu, "Neural speech synthesis with transformer network," *Proc. AAAI Conf. Artificial Intelligence*, **33**, pp. 6706–6713 (2019).

[110] K. Seki, S. Takamichi, T. Saeki and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2023* (2023).

[111] C.-H. Hu, Y.-H. Peng, J. Yamagishi, Y. Tsao and H.-M. Wang, "SVSNet: An end-to-end speaker voice similarity assessment model," *IEEE Signal Process. Lett.*, **29**, 767–771 (2022).

[112] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," *Proc. 2018 IEEE Spoken Language Technology Workshop* (*SLT*), pp. 1021–1028 (2018).

[113] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Adv. Neural Inf. Process. Syst.*, **31** (2018).

[114] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. de Oliveira, A. Candido Jr., A. da Silva Soares, S. M. Aluisio and M. A. Ponti, "SC-GlowTTS: An efficient zero-shot multi-speaker text-to-speech model," *Proc. Interspeech 2021*, pp. 3645–3649 (2021).

[115] R. K. Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian and T. Toda, "Predictions of subjective ratings and spoofing assessments of Voice Conversion Challenge 2020 submissions," *Proc. Jt. Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 99–120 (2020).

[116] Y. Choi, Y. Jung and H. Kim, "Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," *Proc. 2021 IEEE Spoken Language Technology Workshop* (*SLT*), pp. 462–469 (2021).

[117] W. Zhou, Z. Yang, C. Chu, S. Li, R. Dabre, Y. Zhao and T. Kawahara, "MOS-FAD: Improving fake audio detection via automatic mean opinion score prediction," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2024* (to appear), (2024).

[118] G. Mittag, S. Zadtootaghaj, T. Michael, B. Naderi and S. Möller, "Bias-aware loss for training image and speech quality prediction models from multiple datasets," *Proc. 13th Int. Conf. Quality of Multimedia Experience* (*QoMEX*) *2021*, pp. 97–102 (2021).

[119] H. Yadav, E. Cooper, J. Yamagishi, S. Sitaram and R. R. Shah, "Partial rank similarity minimization method for quality MOS prediction of unseen speech synthesis systems in zero-shot and semi-supervised setting," *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*), pp. 1–7 (2023).

[120] C. Valentini-Botinhao, M. S. Ribeiro, O. Watts, K. Richmond and G. E. Henter, "Predicting pairwise preferences between TTS audio stimuli using parallel ratings data and anti-symmetric twin neural networks," *Proc. Interspeech 2022*, pp. 471–475 (2022).

[121] C.-H. Hu, Y. Yasuda and T. Toda, "Preference-based training framework for automatic speech quality assessment using deep neural network," *Proc. Interspeech 2023*, pp. 546–550 (2023).

[122] C. K. Reddy, V. Gopal and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2021*, pp. 6493–6497 (2021).

[123] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **31**, 54–70 (2022).

[124] P. Manocha, B. Xu and A. Kumar, "NORESQA: A frame-

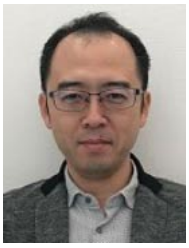
work for speech quality assessment using non-matching references," *Adv. Neural Inf. Process. Syst.*, **34**, 22 363–22 378 (2021).
[125] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *Proc. Interspeech 2020*, pp. 1748–1752 (2020).
[126] G. Mittag and S. Möller, "Full-reference speech quality estimation with attentional Siamese neural networks," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2020*, pp. 346–350 (2020).
[127] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda and J. Yamagishi, "The VoiceMOS Challenge 2022," *Proc. Interspeech 2022*, pp. 4536–4540 (2022).
[128] S. King and V. Karaiskos, "The Blizzard Challenge 2010" (2010).
[129] S. King and V. Karaiskos, "The Blizzard Challenge 2011" (2011).
[130] S. King and V. Karaiskos, "The Blizzard Challenge 2013" (2013).
[131] S. King and V. Karaiskos, "The Blizzard Challenge 2016" (2016).
[132] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*) *2020*, pp. 7654–7658 (2020).
[133] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" *Proc. 11th ISCA Speech Synthesis Workshop* (*SSW 11*), pp. 183–188 (2021).
[134] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," *Proc. Interspeech 2022*, pp. 4521–4525 (2022).
[135] W.-C. Tseng, W.-T. Kao and H.-Y. Lee, "DDOS: A MOS prediction framework utilizing domain adaptive pre-training and distribution of opinion scores," *Proc. Interspeech 2022*, pp. 4541–4545 (2022).
[136] A. Stan, "The ZevoMOS entry to VoiceMOS Challenge 2022," *Proc. Interspeech 2022*, pp. 4516–4520 (2022).
[137] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," *Proc. Int. Conf. Speech Technology and Human-Computer Dialogue* (*SpeD*), pp. 1–10 (2019).
[138] H. Nguyen, K. Li and M. Unoki, "Automatic mean opinion score estimation with temporal modulation features on gammatone filterbank for speech assessment," *Proc. Interspeech 2022*, pp. 4526–4530 (2022).
[139] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino and Y. Zhao, "Fusion of self-supervised learned models for MOS prediction," *Proc. Interspeech 2022*, pp. 5443–5447 (2022).
[140] X. Tian, K. Fu, S. Gao, Y. Gu, K. Wang, W. Li and Z. Ma, "A transfer and multi-task learning based approach for MOS prediction," *Proc. Interspeech 2022*, pp. 5438–5442 (2022).
[141] O. Plátek and O. Dusek, "MooseNet: A trainable metric for synthesized speech with a PLDA module," *Proc. 12th ISCA Speech Synthesis Workshop* (*SSW2023*), pp. 48–54 (2023).
[142] A. Kunikoshi, J. Kim, W. Jun and K. Sjölander, "Comparison of speech representations for the MOS prediction system," *arXiv preprint arXiv:2206.13817* (2022).
[143] H. Becerra, A. Ragano and A. Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," *Proc. Interspeech 2022*, pp. 4088–4092 (2022).
[144] M. Chinen, J. Skoglund, C. K. A. Reddy, A. Ragano and A. Hines, "Using rater and system metadata to explain variance in the VoiceMOS Challenge 2022 dataset," *Proc. Interspeech 2022*, pp. 4531–4535 (2022).
[145] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda and J. Yamagishi, "The VoiceMOS Challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*) (2023).
[146] O. Perrotin, B. Stephenson, S. Gerber and G. Bailly, "The Blizzard Challenge 2023," *Proc. 18th Blizzard Challenge Workshop*, pp. 1–27 (2023).
[147] Y.-W. Chen and Y. Tsao, "InQSS: A speech intelligibility and quality assessment model using a multi-task learning network," *Proc. Interspeech 2022*, pp. 3088–3092 (2022).
[148] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," *arXiv preprint arXiv:2309.12766* (2023).
[149] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu and X. Xu, "LE-SSL-MOS: Self-supervised learning MOS prediction with listener enhancement," *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*) (2023).
[150] K. Shen, D. Yan, L. Dong, Y. Ren, X. Wu and J. Hu, "SQAT-LD: SPeech Quality Assessment Transformer utilizing listener dependent modeling for zero-shot out-of-domain MOS prediction," *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*) (2023).
[151] Y. Gong, C.-I. Lai, Y.-A. Chung and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," *Proc. AAAI*, **36**, pp. 10 699–10 709 (2022).
[152] T. Saeki, S. Maiti, X. Li, S. Watanabe, S. Takamichi and H. Saruwatari, "Learning to speak from text: Zero-shot multilingual text-to-speech with unsupervised text pretraining," *Proc. 32nd Int. Jt. Conf. Artificial Intelligence, IJCAI-23*, E. Elkind, Ed., International Joint Conferences on Artificial Intelligence Organization, pp. 5179–5187 (2023), main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2023/575
[153] J. Chevelu, D. Lolive, S. Le Maguer and D. Guennec, "How to compare TTS systems: A new subjective evaluation methodology focused on differences," *Proc. Interspeech 2015*, pp. 3481–3485 (2015).

Erica Cooper received the B.Sc. and M.Eng. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2009 and 2010, respectively, and the Ph.D. degree in computer science from Columbia University, New York, NY, USA, in 2019. She is currently a Project Associate Professor with the National Institute of Informatics, Tokyo, Japan. Her research interests include statistical machine learning and speech synthesis. She was a co-organizer of the VoiceMOS Challenge in 2022 and 2023. Dr. Cooper's awards include the 3rd Prize in the CSAW Voice Biometrics and Speech Synthesis Competition, the Computer Science Service Award from Columbia University, and the Best Poster Award in the Speech Processing Courses in Crete.

Wen-Chin Huang received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 2018 and the M.S. degree in 2021 from Nagoya University, Nagoya, Japan, where he is currently working toward the Ph.D. degree. From 2017 to 2019, he was a Research Assistant with the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He was the Co-Organizer of the Voice Conversion Challenge 2020 and VoiceMOS Challenge 2022. His research interests include deep learning applications to speech processing, with a main focus on voice conversion and speech quality assessment. He was the recipient of the Best Student Paper Award in ISCSLP2018, the Best Paper Award in APSIPA ASC 2021, and the Research Fellowship for Young Scientists (DC1) from the Japan Society for the Promotion of Science in 2021.

Yu Tsao received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. He is also a Jointly Appointed Professor with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. His research interests include assistive oral communication technologies, audio coding, and biosignal processing. Dr. Tsao was a recipient of the Academia Sinica Career Development Award in 2017, National Innovation Awards from 2018 to 2021, Future Tech Breakthrough Award 2019, Outstanding Elite Award, Chung Hwa Rotary Educational Foundation from 2019 to 2020, NSTC Future Tech Award 2022, and Young Author, Best Paper Award. He is the corresponding author of a paper that received the 2021 IEEE SIGNAL PROCESSING SOCIETY (SPS). He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS.

Hsin-Min Wang received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow. His main research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition. From 2016 to 2020, he was an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He is currently a Senior Editor of APSIPA Transactions on Signal and Information Processing. He is the General Co-Chair of ISCSLP2016, ISCSLP2018, and ASRU2023 and a Technical Co-Chair of ISCSLP2010, O-COCOSDA2011, APSIPAASC2013, ISMIR2014, ASRU2019, and APSIPAASC2023. He was the recipient of the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA distinguished Lecturer for 2014–2015. He is a Senior Member of IEEE and a Member of the International Speech Communication Association and ACM.



Tomoki Toda received the B.E. degree from Nagoya University, Nagoya, Japan, in 1999, and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), Ikoma, Japan, in 2001 and 2003, respectively. From 2003 to 2005, he was a Research Fellow of the Japan Society for the Promotion of Science. He was an Assistant Professor from 2005 to 2011 and an Associate Professor from 2011 to 2015 with NAIST, respectively. Since 2015, he has been a Professor at the Information Technology Center, Nagoya University. His research interests include statistical approaches to sound media information processing. He was the recipient of several awards, including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (*Speech Communication Journal*).

Junichi Yamagishi received the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006. From 2007 to 2013, he was a Research Fellow with the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, Edinburgh, U.K. He was appointed as an Associate Professor with the National Institute of Informatics (NII), Tokyo, Japan, in 2013. He is currently a Professor with NII. His research interests include speech processing, machine learning, signal processing, biometrics, digital media cloning, and media forensics. He was a co-organizer for the bi-annual ASVspoof Challenge and the bi-annual Voice Conversion Challenge. He was also a member of the IEEE Speech and Language Technical Committee from 2013–2019, an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING from 2014–2017, a Chairperson of ISCA SynSIG from 2017–2021, and the Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2019–2023. He is currently a PI of JST-CREST and ANR supported VoicePersonae Project.