Progress and Challenges in DNN-based Objective Quality Assessment of Synthesized Speech

Erica Cooper
National Institute of Information and Communications Technology, Japan
E-mail: ecooper@nict.go.jp

***Abstract*—The field of speech synthesis has advanced rapidly in recent years, and evaluation methodologies for synthesized speech have evolved as well. While listening tests are the gold standard for evaluating synthesized speech, they are costly and time-consuming, leading researchers to consider more automatic and objective metrics for evaluation. In this paper, we give an overview of machine learning based approaches to the prediction of quality of synthesized speech, with a focus on modern deep neural network (DNN) based approaches for MOS prediction, including supervised task-specific training, approaches making use of pretrained self-supervised speech models, unsupervised approaches, and more recent approaches making use of large language models. We will also discuss the current state of objective evaluation of synthesized speech including open research challenges and future directions.**

I. Introduction

The field of speech synthesis, including text-to-speech synthesis (TTS) and voice conversion (VC), has advanced rapidly in recent years, and evaluation methodologies have evolved as well. We have generally moved past the days of inviting local participants to come to the laboratory and transcribe the synthesized words that they hear using pencil and paper, and with the convenience of crowdsourcing platforms, large-scale listening tests can be completed rapidly by participants around the world. While listening tests are the gold standard for evaluating synthesized speech, they are costly and time-consuming, leading researchers to consider objective evaluation metrics.

The Mean Opinion Score (MOS) test [1], in which listeners rate audio samples on a scale for some aspect such as quality or naturalness, has become the most popular evaluation methodology due to its simplicity and scalability. MOS tests also result in paired data of audio samples and ratings, enabling this data to be used for supervised training of regression models. As an abridged summary of our previous survey paper [2], this paper will give an overview of machine learning based approaches to quality prediction for synthesized speech, with a focus on deep neural network (DNN) based approaches including supervised task-specific training, use of pretrained self-supervised speech models, and unsupervised approaches, also including newer works which have since been published. We will additionally describe more recent approaches making use of large language models and discuss the current state of the field, including open research challenges and future directions.

II. Background

Early approaches to quality prediction for synthesized speech took inspiration from intelligibility metrics from telephony. These are categorized broadly into two approaches: *intrusive* or *double-ended*, requiring a ground-truth matched reference sample, and *non-intrusive* (also known as *reference-free* or *single-ended*), in which no reference sample is required. For speech synthesis evaluation, non-intrusive methods are usually preferable due to the so-called *one-to-many problem,* meaning that there are many valid but different ways that a sentence can be synthesized while still sounding correct.

The *Mel-cepstral distance (MCD)* measure [3], developed for telephony, is the difference between the Mel cepstra of a reference and a test speech sample. It was adapted for synthesized speech [4] by applying dynamic time warping (DTW) for alignment. *Root mean squared error (RMSE)* and *correlation of f0* have also been used for evaluating intonation.

The *Perceptual Evaluation of Speech Quality (PESQ)* metric [5] was developed for evaluation of speech over telephone networks. Disturbances are measured and aggregated into a final score [6], and then a third-order polynomial is fitted to real MOS data to convert that score into a final value. The *P.563 recommendation* [7], [8] was the first reference-free measurement developed by the ITU. P.563 considers several different distortion categories, identifies the dominant one, and computes a weighted combination of the different distortion measures. Similar to PESQ, a third-order polynomial is then fitted to real MOS data to scale the scores.

As more MOS datasets have become available, machine learning based approaches started to be applied. A 2008 study [9] found that although P.563 had poor correlations for synthesized speech, several internal features of that model had higher correlations and thus a regression tree was used to find the informative ones. Follow-up studies investigated additional features [10] and SVMs with non-linear regularization [11].

III. Datasets

Datasets of human ratings of synthesized speech are few and relatively small in size compared to typical datasets for other machine learning tasks, but they have been invaluable for advancing the field. The Blizzard Challenge (BC), which began in 2005 [12], was started for the purpose of comparing corpus-based TTS. At the end of each challenge, a crowdsourced listening test is conducted. The Blizzard Challenge has set a strong precedent for TTS evaluation, and they have released their listening test data. The Voice Conversion Challenge (VCC), which began in 2016 [13], is a similar challenge for voice conversion that also shares listening test data.

979-8-3315-7206-8/25/$31.00

While listening tests for research papers typically include samples from 5-10 systems, usually ablations or other variants of a proposed system, BC and VCC draw submissions representing a wide variety of synthesis methods and typically evaluate about 20-30 systems at a time. Crucially, datasets from separate MOS tests cannot be combined to create a larger dataset, given the various biases in listening tests [14]. This motivated the creation of the BVCC dataset [15], in which samples from various BCs, VCCs, and ESPnet-TTS [16] were evaluated together in one large-scale listening test containing samples from 187 systems in total. A parallel effort was SOMOS [17], in which samples from 200 Tacotron variants were evaluated together. These larger-scale datasets only contain naturalness ratings – to address the lack of a large-scale dataset for speaker similarity, VoxSim [18], a dataset of speaker similarity ratings of natural speech utterances from 1251 VoxCeleb speakers, was collected. The VCC2020-Pref dataset [19] extends the MOS ratings in VCC 2020 by collecting additional pairwise preference ratings for the same samples. Towards the aim of collecting data for LLM-based evaluation, the QualiSpeech dataset [20] combines samples from BVCC with additional samples from ten more recent TTS models, with annotations for eleven specific problem categories as well as natural language descriptions of the audio quality.

IV. Evaluation metrics

MOS predictors are evaluated with respect to their match to ground-truth labels collected in a listening test. Typically, both utterance level and system-level metrics are reported. **Root mean squared error** (RMSE) measures the average difference between ground-truth and predicted MOS values. Several correlation metrics are also typically reported:

- **Linear correlation coefficient (LCC):** a simple measure of correlation between ground-truth and predicted ratings.
- **Spearman rank correlation coefficient (SRCC):** Correlations of the ground-truth and predicted *rankings*.
- **Kendall Tau correlation coefficient (KTAU):** Proposed in [21] for evaluating MOS predictors; measures rank-order correlations with more robustness to outliers.

V. DNN-based approaches

A 2016 study [22] used a hierarchical approach to predict a system-level score and then use it as a feature for sample-level prediction. Comparing DNNs to linear regression models, they found that DNNs worked better for both prediction stages. AutoMOS [23] was proposed shortly thereafter to use LSTMs trained on MOS ratings from internal listening test data to evaluate unit selection TTS. The authors found that the resulting predictor could be used to tune their cost function.

In 2019, MOSNet [24] was the first DNN model to predict quality of voice-converted speech. They used VCC 2018 data to train CNN, BLSTM, and combination architectures, finding that the combination worked best with a system-level SRCC of around 0.9. Since the authors open-sourced their code, MOSNet became a popular objective evaluation metric.

NISQA-TTS [25] is a CNN-LSTM pretrained for quality estimation of degraded natural speech [26] that was further trained on TTS MOS data in several languages. Correlations above 0.9 were obtained even with unseen conditions.

Listening test datasets often contain listener ID labels. MBNet (Mean-Bias Network) [27] incorporated mean and bias subnets that learn preferences of individual listeners in parallel with learning the averaged scores, resulting in improvements over a MOSNet baseline. LDNet (Listener-Dependent Network) [28] incorporated several further improvements, including a lighter-weight bias subnet as well as two inference modes: "all listeners," which outputs an averaged predicted decision of all the listeners seen during training, and "mean listener," wherein a "virtual" listener, whose rating is the mean score of a given audio sample, is learned during training. LDNet outperformed the MBNet baseline, with the best results produced in "mean listener" mode. DeePMOS (Deep Posterior Mean-Opinion-Score) [29] extended MBNet to predict variance in addition to the mean, which improved MSE and system-level correlations over the MBNet baseline while also providing more interpretable predictions.

VI. SSL-based approaches

Large-scale speech models trained using self-supervised learning (SSL) based objectives have demonstrated excellent results on speech-related machine learning tasks [30], and speech quality prediction has been no exception. In 2021, a first effort to fine-tune SSL models for MOS prediction was made [31]. They compared the use of several different pretrained SSL models as encoders to the use of classical features like MFCCs, finding improvements when SSL encoders were used. The SSL-MOS model [32] used an even simpler SSL-based architecture for MOS prediction by simply average-pooling the frame-level output of an SSL model and adding a linear output layer. Compared to MOSNet, the SSL-based MOS predictors showed better generalization ability to out-of-domain datasets, even showing reasonable correlations in zero-shot scenarios.

Several later works improved upon SSL-MOS. Encoders for prosodic features and linguistic features extracted from text [33] were added. RAMP (Retrieval-Augmented MOS Prediction) [34] added a non-parametric component, generating predictions based on a distance-weighted sum of the k nearest neighbors of a test audio sample, which improved out-of-domain prediction. A follow-up work, RAMP+ [35], improved the distance calculation and retrieval by using synthesis system ID. DDOS [36] modeled the distribution of MOS ratings in addition to the mean, and also added domain-adaptive pretraining with synthesized speech, which was shown to reduce MSE. The UTMOS system [37] obtained good results by combining strong learners (SSL models) and weak learners (regression models), and due to their open-source code, UTMOS has recently become a popular objective metric for TTS. A follow-up work, UTMOSv2 [38], used a pretrained model for image features to capture information from speech spectrograms, and fusing this with an SSL-based predictor gave improvements in differentiating between high-quality speech synthesizers.

SQuId [39] was the first massively multilingual work on TTS MOS prediction. They started with a multi-modal language model and fine-tuned it using data from over 2000 internal listening tests for 52 different language locales. Their evaluation included several unseen locales and demonstrated the benefits of transfer learning.

VII. Unsupervised approaches

Supervised training of MOS predictors requires datasets which are relatively few in number and expensive to collect. Furthermore, there is extensive evidence [32], [40], [41] that supervisedly-trained MOS predictors have worse predictive ability on new and unseen domains such as different languages. This has been the motivation to explore *unsupervised* strategies, which typically rely on building a *reference model*, which represents prior knowledge about the distribution of natural speech, and then measuring the distance between the distributions of natural speech and synthesized speech.

This distribution modeling approach has precedent in the era of hidden Markov model (HMM) based TTS – in 2008 [42], gender-dependent HMMs were trained on natural speech, and log-likelihood with respect to these reference models was then computed from features of synthesized speech. A 2013 study [43] built Gaussian mixture models (GMMs) of the acoustic feature vectors of their TTS training data, copy-synthesized using the same vocoder as their synthesis pipeline, resulting in well-matched reference models.

More recently, SpeechLMScore [44] extracted token sequences from a pretrained SSL, followed by k-means clustering, and perplexity of these token sequences is measured with respect to a speech language model trained on natural speech. UNIQUE [45] used a distribution model of k-means tokens instead of a sequence model. VQScore [46] trained a VQ-VAE on natural clean speech and then reconstruction error was used as a measure of quality, demonstrating improved prediction over SpeechLMScore with lower data requirements. SpeechBERTScore [47] is an unsupervised but intrusive approach – lexically-matched (but not duration-matched) reference samples are required, but not any MOS-labeled data. Considering that researchers typically have ground-truth natural test samples, embeddings of the ground-truth and synthesized speech samples are extracted using an SSL model and their cosine similarity is computed. This method outperformed traditional measures like MCD, as well as pretrained UTMOS and SpeechLMScore, in terms of SRCC.

We also note the related literature on Fréchet Inception Distance (FID) and Fréchet Audio Distance (FAD). FID was proposed in 2017 as an evaluation metric for image generation models [48] – it compares the distribution of two datasets in an embedding space using the 2-Wasserstein distance. FAD [49] similarly evaluates audio. FAD was adapted for TTS specifically [50] by using an ASR model to obtain the embeddings. Inspired by these approaches, TTSDS [51] used an ensemble of reference models to measure aspects of speech such as prosody, speaker identity, and intelligibility. TTSDS2 [52] extended the approach to a multilingual setting by replacing the English reference models with multilingual ones.

VIII. LLM-based approaches

Large language models (LLMs) have emerged as powerful tools for natural language tasks, and in particular, auditory large language models have demonstrated exceptional capabilities in many speech and audio perception and understanding tasks. In 2024, a study [53] investigated the use of auditory LLMs for text-to-audio, text-to-music, text-to-speech, and deep noise suppression evaluation tasks, finding strong correlations with human listeners and good generalization ability to out-of-domain conditions. A 2025 study [54] investigated fine-tuning auditory LLMs on MOS-labeled datasets, not only demonstrating good MOS prediction abilities but also providing A/B comparisons and natural-language descriptions of the audio quality. A similar approach [55] applied audio LLMs to predict and describe four quality dimensions of noisy and degraded speech, and another study [56] used Whisper in combination with multimodal GPT-4o to obtain text representations of noisy and enhanced speech to evaluate their naturalness.

IX. Beyond MOS

While MOS is a popular evaluation methodology for synthesized speech, there are also many drawbacks such as bias [14]. There is also abundant evidence that MOS tests have become saturated [57], [58] and that other testing methodologies such as pairwise A/B comparisons [59] and MUSHRA-style tests [60] have better discriminative power. We may also want to evaluate aspects besides naturalness, such as intelligibility, speaker similarity, expressivity, context appropriateness, diagnostic identification of specific synthesis issues, and so on.

Pairwise preference data can be derived from MOS or MUSHRA data. PrefNet [61] takes a pair of audio samples with the same lexical content but different durations and outputs the probability that one waveform would be preferred over another. Since it is important for the results to be the same regardless of the input order, anti-symmetric twin neural networks are used. The authors derived pairwise preference data from MUSHRA tests. A later work [62] derived preference scores from pairs of MOS ratings constrained to be from the same listener, and in a follow-up study called E2EPref [19], an investigation of different approaches to pair generation and preference aggregation was additionally conducted and cross-dataset and out-of-domain evaluations were performed. Modeling MOS data as a pairwise preference problem was shown to give improvements over an UTMOS baseline as well as better robustness to range-equalizing bias, enabling effective combination of multiple datasets during training.

Speaker similarity is typically evaluated for voice conversion and speaker-adaptive TTS, and automatic metrics are popular – one approach is to use a pretrained speaker recognition model to extract a speaker embedding of the synthesized speech and one for the natural speech of the target speaker, and to measure cosine similarity. Correlations of this approach with human ratings were observed at around 0.85 when using x-vectors

[63]. The later VoxSim study [18] found lower correlations of about 0.75 using ECAPA embeddings, so there is likely some effect of dataset and embedding. Data-driven approaches have also been applied – SVSNet [64] modified MOSNet to take two audio inputs and have a symmetry constraint similar to PrefNet, resulting in predictions with correlations to human ratings above 0.9. Improved cross-dataset generalization ability was also observed when training using the VoxSim data.

Intelligibility is no longer commonly evaluated for synthesized speech as state-of-the-art synthesizers generally produce intelligible speech, but it can still be important to evaluate in low-resource data conditions. Word error rate (WER) obtained from automatic speech recognizers (ASR) is commonly used, and by 2015, commercial ASR APIs were shown to have correlations of 0.94 with WERs of human transcriptions [65].

Machine learning approaches and objective measures can also help researchers run more effective listening tests. To combat the saturation of MOS tests, audio samples can be selected that are maximally different, measured by acoustic differences such as DTW-aligned MFCCs or Mel spectrograms [66], [67]. Furthermore, while pairwise comparison tests have better discriminative power than MOS tests, researchers often avoid running them because these tests scale quadratically – however, an online learning approach can be used to decide which systems need additional comparisons to differentiate them [68]. Lifelong learning has also been applied to the training of MOS predictors [69] to mitigate both the issue of cross-corpus incomparability as well as the need to continually adapt predictors to newer and more advanced synthesis methods.

Speech synthesis researchers may also want to know the reasons *why* their system has low quality – diagnostic information about specific issues can be informative for making improvements. LLM-based approaches are taking steps to address this, but there have also been efforts towards dedicated modeling for this task. ADTMOS [70] considered that the perceptual salience of several manually-chosen distortion-related features such as zero crossing rate, energy, and spectral flux could be learned by learning weights for those features during training. ChunkMOS [71] predicts localized, frame-level scores trained on data with synthetic distortions at known locations, resulting in a fine-grained, interpretable model.

X. Open challenges and future directions

Generalization ability to out-of-domain conditions, such as unseen syntheszers, languages, and synthesis types is perhaps the most important use case for speech quality assessment because researchers will always want to evaluate new systems. In-domain, supervised MOS prediction has become quite accurate, but the evidence shows that out-of-domain quality prediction still remains more challenging. Unsupervised reference modeling approaches are a promising direction; however, this approach makes an assumption that appropriate reference data exists. As the field expands to more use cases, more different and creative types of synthesis, and various low-resource data conditions, this assumption might not always hold. The application of LLMs to the task of speech quality prediction, especially for synthesized speech, is a promising but still relatively very new area. The generalist abilities of LLMs show promise for evaluating out-of-domain audio and for use in novel evaluation paradigms – further research will be necessary to determine whether LLMs will ultimately be useful for this task, and for the development of new approaches.

Currently, the most popular aspect to evaluate is quality or naturalness, but as state-of-the-art speech synthesizers approach a high level of quality (and as MOS tests become saturated), it is becoming important to evaluate different and more comprehensive aspects of speech. Given the ability of LLMs to handle very long context, they are expected to be useful for evaluation of long-form synthesis as well as aspects of context-appropriateness and ecological validity of synthesized speech (for example, in a conversational setting); however, this is still yet to be determined. Evaluation of synthesized nonverbal vocalizations will also become important for the development of realistic conversational agents.

All datasets of human ratings of synthesized speech show large differences between listeners' ratings of the same audio sample. The reasons for these differences are currently not well-understood, but it is important to understand them in order to create quality predictors that truly represent human opinions and that are more accurate and more interpretable. Accurate, interpretable, and fine-grained quality predictors can also help us to better understand human perception of speech. In the long term, more accurate, comprehensive, ecologically-valid, and diagnostic evaluation tools will help to push forward the state of the art and help researchers to develop the next generation of high-quality, multilingual, and adaptable generated speech.

References

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

[2] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, 2024.

[3] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993.

[4] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean Mel cepstral distortion," in *SLTU*, 2008.

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

[6] S. Ipswich, "PESQ: An Introduction White Paper," 2001.

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

[8] L. Malfait, J. Berger, and M. Kastner, "P. 563—The ITU-T standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

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

[10] 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," in *Blizzard Challenge Workshop*, 2010.

[11] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Quality prediction of synthesized speech based on perceptual quality dimensions," *Speech Communication*, 2015.

[12] A. W. Black and K. Tokuda, "The Blizzard Challenge - 2005: evaluating corpus-based speech synthesis on common datasets," in *Interspeech*, 2005.

[13] T. Toda, L.-H. Chen, D. Saito, *et al.*, "The Voice Conversion Challenge 2016," in *Interspeech*, 2016.

[14] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, 2008.

[15] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" In *ISCA SSW*, 2021.

[16] T. Hayashi, R. Yamamoto, K. Inoue, *et al.*, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP*, 2020.

[17] G. Maniati, A. Vioni, N. Ellinas, *et al.*, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Interspeech*, 2022.

[18] J. Ahn, Y. Kim, Y. Choi, *et al.*, "VoxSim: a perceptual voice similarity dataset," 2024.

[19] C.-H. Hu, Y. Yasuda, and T. Toda, "E2EPref: an end-to-end preference-based framework for speech quality assessment to alleviate bias in direct assessment scores," *Computer Speech & Language*, vol. 93, no. C, 2025.

[20] S. Wang, W. Yu, X. Chen, *et al.*, "QualiSpeech: A speech quality assessment dataset with natural language reasoning and descriptions," in *ACL*, 2025.

[21] J. Williams, J. Rownicka, P. Oplustil, and S. King, "Comparison of Speech Representations for Automatic Quality Estimation in Multi-Speaker Text-to-Speech Synthesis," in *Odyssey*, 2020.

[22] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks," in *Interspeech*, 2016.

[23] 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.

[24] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Interspeech*, 2019.

[25] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Interspeech*, 2020.

[26] G. Mittag and S. Möller, "Full-reference speech quality estimation with attentional Siamese neural networks," in *ICASSP*, 2020.

[27] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *ICASSP*, 2021.

[28] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *ICASSP*, 2022.

[29] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, "DeePMOS: Deep Posterior Mean-Opinion-Score of Speech," in *Interspeech*, 2023.

[30] S. W. Yang, P. H. Chi, Y. S. Chuang, *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," in *Interspeech*, International Speech Communication Association, 2021.

[31] W.-C. Tseng, C.-Y. Huang, W.-T. Kao, Y. Y. Lin, and H.-Y. Lee, "Utilizing Self-Supervised Representations for MOS Prediction," in *Interspeech*, 2021.

[32] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *ICASSP*, 2022.

[33] A. Vioni, G. Maniati, N. Ellinas, *et al.*, "Investigating Content-Aware Neural Text-to-Speech MOS Prediction Using Prosodic and Linguistic Features," in *ICASSP*, 2023.

[34] H. Wang, S. Zhao, X. Zheng, and Y. Qin, "RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in *Interspeech*, 2023.

[35] H. Wang, S. Zhao, X. Zheng, J. Zhou, X. Wang, and Y. Qin, "RAMP+: Retrieval-augmented MOS prediction with prior knowledge integration," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[36] 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," in *Interspeech*, 2022.

[37] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Interspeech*, 2022.

[38] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The T05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *SLT*.

[39] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, "SQuId: Measuring speech naturalness in many languages," in *ICASSP*, IEEE, 2023.

[40] 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," in *ASRU*, 2023.
[41] S. Udupa, S. Maiti, and P. K. Ghosh, "IndicMOS: multilingual mos prediction for 7 indian languages," in *Interspeech*, 2024.
[42] T. H. Falk and S. Moller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Processing Letters*, 2008.
[43] S. L. Maguer, N. Barbot, and O. Boeffard, "Evaluation of contextual descriptors for HMM-based speech synthesis in French," in *SSW*, 2013.
[44] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "SpeechLMScore: Evaluating speech generation using speech language model," in *ICASSP*, IEEE, 2023.
[45] J. Yoon, W. Ko, S. Um, *et al.*, "UNIQUE: Unsupervised network for integrated speech quality evaluation," in *Interspeech*, 2024.
[46] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, "Self-supervised speech quality estimation and enhancement using only clean speech," in *ICLR*, 2024.
[47] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "SpeechBERTScore: reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics," in *Interspeech*, 2024.
[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, 2017.
[49] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," 2019.
[50] M. Bińkowski, J. Donahue, S. Dieleman, *et al.*, "High fidelity speech synthesis with adversarial networks," in *ICLR*, 2020.
[51] C. Minixhofer, O. Klejch, and P. Bell, "TTSDS - text-to-speech distribution score," in *IEEE SLT*, 2024.
[52] C. Minixhofer, O. Klejch, and P. Bell, "TTSDS2: resources and benchmark for evaluating human-quality text to speech systems," *ISCA SSW*, 2025.
[53] S. Deshmukh, D. Alharthi, B. Elizalde, *et al.*, "Pam: Prompting audio-language models for audio quality assessment," *Interspeech*, 2024.
[54] S. Wang, W. Yu, Y. Yang, *et al.*, "Enabling auditory large language models for automatic speech quality evaluation," in *ICASSP*, 2025.
[55] C. Chen, Y. Hu, S. Wang, *et al.*, "Audio large language models can be descriptive speech quality evaluators," in *ICLR*, 2025.
[56] R. E. Zezario, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, "A study on zero-shot non-intrusive speech assessment using large language models," in *ICASSP*, 2025.
[57] S. Shirali-Shahreza and G. Penn, "MOS Naturalness and the Quest for Human-Like Speech," in *IEEE SLT*, 2018.
[58] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*,
[59] Y. Yasuda and T. Toda, "Analysis of Mean Opinion Scores in Subjective Evaluation of Synthetic Speech Based on Tail Probabilities," in *Interspeech*, 2023.
[60] M. S. Ribeiro, J. Yamagishi, and R. A. Clark, "A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis," in *Interspeech*, 2015.
[61] 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," in *Interspeech*, 2022.
[62] C.-H. Hu, Y. Yasuda, and T. Toda, "Preference-based training framework for automatic speech quality assessment using deep neural network," in *Interspeech*, 2023.
[63] R. K. Das, T. Kinnunen, W.-C. Huang, *et al.*, "Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020.
[64] 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 Processing Letters*, vol. 29, pp. 767–771, 2022.
[65] 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," in *Konferenz Elektronische Sprachsignalverarbeitung*, 2015.
[66] J. Chevelu, D. Lolive, S. L. Maguer, and D. Guennec, "How to compare TTS systems: a new subjective evaluation methodology focused on differences," in *Interspeech*, 2015.
[67] O. Perrotin, B. Stephenson, S. Gerber, G. Bailly, and S. King, "Refining the evaluation of speech synthesis: A summary of the blizzard challenge 2023," *Computer Speech and Language*, 2024.
[68] Y. Yasuda and T. Toda, "Automatic design optimization of preference-based subjective evaluation with online learning in crowdsourcing environment," *arXiv preprint arXiv:2403.06100*, 2024.
[69] F. Saget, M. Shamsi, and M. Tahon, "Lifelong learning MOS prediction for synthetic speech quality evaluation," in *Interspeech*, 2024.
[70] Q. Liang, Y. Shen, T. Chen, L. Zhang, and S. Zhao, "ADTMOS–synthesized speech quality assessment based on audio distortion tokens," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
[71] M. Kuhlmann, F. Seebauer, P. Wagner, and R. Haeb-Umbach, "Towards frame-level quality predictions of synthetic speech," in *Interspeech 2025*, 2025.