

Advancing Speech Quality Assessment Through Scientific Challenges and Open-source Activities

Wen-Chin Huang
Nagoya University, Japan

Abstract—Speech quality assessment (SQA) refers to the evaluation of speech quality, and developing an accurate automatic SQA method that reflects human perception has become increasingly important, in order to keep up with the generative AI boom. In recent years, SQA has progressed to a point that researchers started to faithfully use automatic SQA in research papers as a rigorous measurement of goodness for speech generation systems. We believe that the scientific challenges and open-source activities of late have stimulated the growth in this field. In this paper, we review recent challenges as well as open-source implementations and toolkits for SQA, and highlight the importance of maintaining such activities to facilitate the development of not only SQA itself but also generative AI for speech.

I. INTRODUCTION

Evaluating the quality of a speech sample, which is also known as speech quality assessment (SQA) [1]–[3], can be a complicated process, and human is often considered the gold standard. Not only because the end user is human, but the assessment involves considering multiple dimensions simultaneously, including naturalness, intelligibility, and other intended purposes, which is a highly difficult task. While the main purpose of SQA was to monitor the quality of telecommunication services, its application to evaluate speech generation systems has been gaining attention in recent years due to the generative AI boom. However, having human listeners to judge the speech quality in the development pipeline can be costly, and has thus motivated the development of automated evaluation protocols.

Compared to other attributes like intelligibility, the highly-subjective nature of speech quality [2] emphasizes how important it is for a metric to be well-correlated with human perception. It has therefore become increasingly popular to develop SQA methods that are directly optimized using human preference data. Since such approaches are data-driven and thus based on machine learning models, the field of SSQA greatly benefits from the rapid development of deep neural networks (DNNs) in the past decade [4]–[6]. As a result, SQA methods nowadays have been shown to correlate well with human ratings, leading to the adaptation of such methods to evaluate speech generation models in scientific papers.

The success of a research field often arises from the collective efforts of the community as a whole. In the era of deep learning, this can be realized in two key ways. First, scientific challenges can increase visibility and attract interest to a field. These challenges are not about competition or ranking; rather, the main goal is to advance on specific problems. By providing a standardized framework – including shared

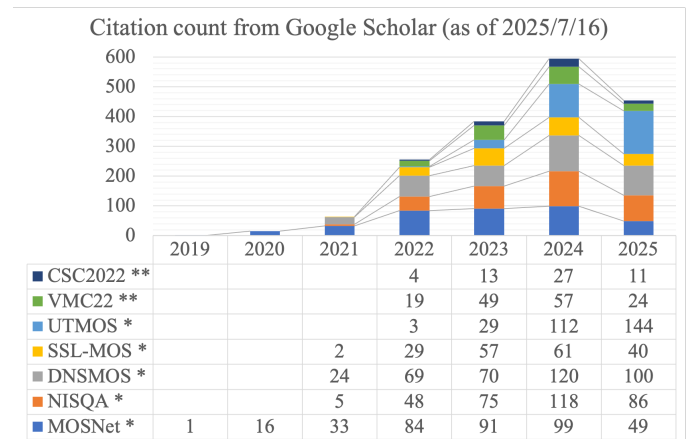


Fig. 1. Google Scholar Citations count of recent SQA papers.*: papers on SQA with open-source implementations. **: summary papers of scientific challenges

datasets and evaluation protocols – they enable systematic analysis and comparison of different approaches. The insights and findings can further be disseminated to benefit researchers across related fields. As evidence, in 2022, two scientific challenges were organized: the ConferencingSpeech Challenge (CSC) [7], and the VoiceMOS Challenge (VMC) [8]. Since then, the citation count of recent SQA papers increased rapidly, as shown in Figure 1.

Open-source activities have also contributed to the rapid advancement of SQA. The most direct application of any SQA method is its integration into real-world scenarios. Therefore, ease of use becomes a crucial factor. Conventional non-DNN-based SQA methods such as PESQ [9] can be computed analytically. In contrast, DNN-based methods require model training, posing a barrier for users who simply want to assess the quality of, for instance, a speech generation system. Converting human-annotated quality labels into trained models and making them openly accessible enables a broader adoption of modern SQA methods. This, in turn, has played a key role in accelerating progress in the field.

In this perspective paper, we first share our four years of experience in running the VoiceMOS Challenge series, whose scope was further expanded from SQA to quality assessment of general audio, thus rebranded to the AudioMOS Challenge. We share the task design, key insights, and feedback from participants. Then, we review recent open-source activities in SQA, which greatly benefit the development of this field. Finally, we discuss future directions.

TABLE I
SUMMARY OF THE TRACKS IN VOICEMOS CHALLENGE 2022-2024 AND AUDIOMOS CHALLENGE 2025.

| Challenge | Track | Dataset | Audio type | Evaluation axis | Setting |
|-------------------------|---------|----------------|---|---|---------------------------------------|
| VoiceMOS Challenge 2022 | Main | BVCC | English TTS & VC | Naturalness | in-domain |
| | OOD | BC19 | Chinese TTS | Naturalness | in-domain |
| VoiceMOS Challenge 2023 | Track 1 | BC23 | French TTS | Naturalness | out-of-domain |
| | Track 2 | SVCC23 | English singing voice conversion | Naturalness | out-of-domain |
| | Track 3 | TMHINT-QI(S) | Chinese noisy & enhanced speech | Naturalness | out-of-domain |
| VoiceMOS Challenge 2024 | Track 1 | Zoomed-in BVCC | English TTS & VC | Naturalness | Same audio, zoomed-in labels |
| | Track 2 | SingMOS | Chinese & Japanese singing voice synthesis & conversion | Naturalness | in-domain |
| | Track 3 | - | English noisy & enhanced speech | Signal distortion & background noise & overall quality | out-of-domain, semi-supervised |
| AudioMOS Challenge 2025 | Track1 | MusicEval | Text-to-music | Overall quality & Textual alignment | in-domain |
| | Track 2 | - | Natural speech/audio/music, text-to-speech/audio/music | Product quality & Product complexity & Content enjoyment & Content usefulness | out-of-domain |
| | Track 3 | - | English TTS at different sampling rate | Naturalness | same audio, different listening tests |

II. THE VOICEMOS AND AUDIOMOS CHALLENGE SERIES

The VMC series was initiated in 2022, and has been held annually since then¹. The term ‘‘MOS’’ stands for ‘‘mean opinion score’’, which is a common listening test type [10]. As the name suggests, the task involves predicting the MOS for a given voice sample. From 2022 to 2024, the challenge focused primarily on the evaluation of speech. In 2025, however, the scope was broadened to include other audio modalities such as music and general environmental sounds. To reflect this expansion, the challenge was rebranded as the AudioMOS Challenge. Table I summarizes the tracks in each year.

In SQA, out-of-domain generalization is a crucial aspect, due to the nature of how listening tests are conducted: each listening test represents a unique context, with different contents (text, speakers, etc.), recruited listeners, ranges of systems being evaluated, and even instructions. Thus, with respect to an SQA model trained on a specific dataset, the testing scenario can be either *in-domain* or *out-of-domain*, where the former means the test samples and the ratings come from the same listening test as that of the training set, and the latter means they come from different listening tests.

The challenge was hosted on CodaLab [11] (which was further renamed to CodaBench²). In each year, the challenge was divided into the training phase, evaluation phase, and post-challenge. The training phase was usually two to three months long, and participants could use the training set (if provided) to develop their system. In the evaluation phase, which was always one week long, participants made their predictions of the quality of the test set samples. After the submission deadline, we released the test set ground truth

labels for participants to perform analysis and wrap up their paper. We also asked each team to submit a system description form based on the template we distributed, including surveys on their opinions on the challenge and future directions.

To assess SQA models, several evaluation criteria were used, but for most of the time, we used system-level Spearman rank correlation coefficient (SRCC) as the primary metric for determining the ranking.

Each year, we provided one or more open-sourced baseline systems, giving participants access to pretrained models along with recipes for training, fine-tuning, and making predictions on the challenge datasets. This component is essential to a scientific challenge for several reasons. First, it significantly lowers the barrier to participation. Second, the baseline system typically represents the state-of-the-art or the most representative approach available at the time. As a result, it becomes feasible to assess whether the challenge drives meaningful progress in the field.

A. The VoiceMOS Challenge 2022

1) *Track and setting*: There were two tracks in VMC 2022, namely the main track and the out-of-domain track. The main track was based on a large-scale dataset of MOS ratings for synthesized audio samples as well as reference natural speech samples, which we collected in 2021 [12]. We mainly collected English-language synthesized audio samples from several past Blizzard Challenges (BCs) from 2008-2016 and from all previous Voice Conversion Challenges, as well as publicly-shared samples from ESPnet-TTS, one of the most commonly used TTS toolkits at that time [13]. Altogether, we collected samples from 187 different systems, and we conducted a large-scale listening test where each sample received eight ratings. Although we carefully designed a training,

¹<https://sites.google.com/view/voicemos-challenge>

²<https://www.codabench.org/>

development, and testing split of the data while holding out some unseen speakers, synthesis systems, and listeners in the development and test sets, we still consider this track to be an “in-domain” setting.

Additionally, we also ran an out-of-domain (OOD) track by making use of the BC 2019 data, which focused on Chinese text-to-speech (TTS) samples, and provided participants with a very small amount of labeled training data (136 samples). Although we also designed the training, development, and testing split to include unseen systems and listeners, since labeled training data was provided, we considered the OOD track to be somewhat “in-domain”. Still, 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.

2) *Results and insights:* The challenge attracted **22** participating teams from academia and industry, which we consider to be very successful compared to popular scientific challenges like the ImageNet Large Scale Visual Recognition Challenge [14], which had on average 20-30 participants per year. In addition, we provided three baselines [15]–[17]. For the main track, the best baseline had a system-level MSE and SRCC of 0.148 and 0.921, ranking 18th and 12th in the two metrics, respectively. The top prediction systems in system-level MSE and SRCC scored **0.090** and **0.939**, respectively. As for the OOD track, the top prediction systems had an MSE of **0.030** and a SRCC of **0.979**, respectively.

Overall, we observed that fine-tuning SSL models for the MOS prediction task was a powerful approach that can produce predictions with **very high correlations with real listener ratings**, even in the case of the OOD track where a very small amount of label was available. Feedback from participants was about including a larger variety of audio to evaluate, including synthesis in more different languages, singing voice synthesis, and noisy and enhanced speech.

B. The VoiceMOS Challenge 2023

The outcomes of the first challenge motivated our design of the 2023 edition of the challenge [18]. If, even with a very small amount of labeled training data, the correlation coefficient between the machine’s prediction and that of humans can be larger than 0.95, how about an entirely **out-of-domain, zero-shot** setting? Therefore, we focused on real-life MOS prediction in a variety of speech domains, as the labels in the test sets are completely new listening tests: they were collected at the same time as the challenge, meaning that team predictions were made before the actual ground-truth MOS values were known to anyone.

1) *Data and tracks:* There were three tracks in VMC 2023. The first track was based on the Blizzard Challenge 2023 [19], which focused on French text-to-speech synthesis. The second track was based on the Singing Voice Conversion Challenge 2023 (SVCC) [20], where the input singing voice sample was converted to a different speaker identity, using either sung (matched) or spoken (mismatched) reference audio from the target speaker. Unlike VMC 2022, where the listening tests were completed in advance and divided into training,

development, and test sets, when the VMC 2023 training phase took place, the Blizzard and SVCC listening tests were still ongoing. Thus, no official training data was provided for these two tracks.

For the third track, we considered the quality assessment of noisy and enhanced speech for the first time. Unlike the previous two tracks, we provided the TMHINT-QI [21] dataset as training data, and for the test set, a separate listening test was conducted, with the same noise generation process, partially different speech enhancement systems, and completely different raters. We named the test set TMHINT-QI(S) [22]. The design of this track was to answer the substantial interest from the participating teams in VMC 2022, as 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, we believed there could be benefits from more communication and collaboration between these communities.

2) *Results and insights:* In total, **ten** teams participated in the 2023 challenge. This year we also provided two baseline systems: one was SSL-MOS, the best baseline system in VMC 2022 [15], and the other one was UTMOS, the best performing system in VMC 2022 [23]. In each track, at least one team outperformed the baseline, showing that progress was indeed made. The most important result was 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**.

As for other findings, we were surprised that many teams performed well on the second track, whose focus was singing voice conversion samples. At the time being, quality assessment for singing voices was still an underexplored field, so we suspected that the domain mismatch between synthesized singing and speech was not as large as we had assumed. Looking at each team’s approach, we found that listener-dependent modeling [16] was more popular this year, and teams that used a mix of different training datasets also tended to do better.

C. The VoiceMOS Challenge 2024

1) *Data and tracks:* There were three tracks in VMC 2024 [24]. The first track was MOS prediction for “zoomed-in” systems, motivated by a real-world scenario where the researchers wish to evaluate a speech generation model under development, whose quality is expected to be better than any previous system. Based on the BVCC dataset, we collected a set of “zoomed-in” subjective ratings [25], where new listening tests were conducted using approximately 50%, 25%, 12%, and 6% of the highest-rated systems. No new training data was provided, and the test set consists of 1000 samples from the 25% subset, 500 of which are also included in the 12% subset. Although the samples in the test sets are all from BVCC, since new listening tests were conducted, this track was considered out-of-domain.

The second track was based on SingMOS [26], a newly collected dataset consisting of natural singing voice samples, vocoder analysis-synthesis samples, and singing voice synthesis/conversion samples in Chinese and Japanese. The official split was used, with partially unseen samples in the training set. Thus, this track was considered in-domain.

The third track focused on MOS prediction for noisy and enhanced speech. While this may appear similar to the VMC 2023 track 3, we introduced two key novelties. First, instead of relying on a single MOS value, the evaluation employed three perceptual dimensions defined in ITU-T P.835 [27]: signal quality (SIG), background intrusiveness (BAK), and overall quality (OVRL). Second, the training data was further limited in size. Specifically, the training and validation sets were derived from the UDASE task of the 7th CHiME Challenge [28], [29], containing only 60 and 40 samples, respectively. The evaluation set was constructed using data from a separate listening test involving samples from the VoiceBank-DEMAND dataset [30].

2) *Results and insights*: This year, we received submissions from five teams from academia and three teams from industry, for **eight** in total from six different countries. We also had a baseline system for each track. For track 1, although multiple teams outperformed the baseline, the overall performance was significantly lower than the original labels, highlighting the difficulty in ranking high-quality speech synthesis systems. For track 2, although no team outperformed the baseline in terms of system-level SRCC, all systems had a system-level SRCC higher than 0.8. While participants questioned whether the baseline was too strong, we suspect that the overly-easy in-domain setting was the main cause. Finally, in track 3, among the three axes, SIG was the most difficult dimension to predict, and participating teams had diverse behaviors.

D. The AudioMOS Challenge 2025

The rapid progress of music and general audio generation led to an urgent need for an automatic evaluation method for text-to-music (TTM) and text-to-audio (TTA) systems that reflect human perception, as demanded by participants in the 2024 challenge. In light of this, we expanded the VMC series and rebranded to the AudioMOS Challenge (AMC).

1) *Data and tracks*: There were three tracks in AMC 2025. The first track focused on MOS prediction of TTM systems, where we used MusicEval [31], a dataset containing music clips generated by 31 modern TTM systems. Music experts were recruited to rate each clip in terms of overall musical quality and alignment with the text prompt, which respectively emphasizes the importance of both the quality of the generated music and its consistency with the given prompt. The second track was based on Meta Audiobox Aesthetics [32], a suite of unified assessment methods for speech, music, and sound. Instead of a single MOS, the evaluation protocol consists of four new evaluation dimensions: production quality, production complexity, content enjoyment, and content usefulness. The task was to assess synthetic samples from TTS, TTA, and TTM along the four axes. The third track focused on MOS prediction

for synthesized speech in different sampling frequencies. During the training phase, participants were provided with samples in 16, 24, and 48 kHz, along with their ratings obtained from listening tests that only contained samples in the same sampling frequencies. For the test set, the participants were asked to make predictions of synthetic samples that reflect their scores in a listening test that contains samples from all sampling frequencies.

2) *Results and insights*: We received submissions from **24** unique teams, which was the most among the previous VMCs, demonstrating the increasing interest in audio quality assessment. We also prepared a baseline for each track. For track 1, not only did all teams outperform the baseline, but the best-performing team achieved system-level SRCCs over 0.95 on both axes, again reflecting the overly-easy in-domain setting. For the second track, as it was difficult for a single system to excel in the prediction of all four axes, more than half of the teams outperformed the baseline. For track 3, although all teams outperformed the baseline, the limited number of synthesis systems in the dataset made many participants question whether the track was properly designed.

E. Key factors to the success of a challenge

The most important aspect of organizing a scientific challenge is attracting a sufficient number of participants. Across the four editions, we observed fluctuating levels of participation, with 22, 10, 8, and 24 teams joining in 2022 through 2025, respectively. Upon reflection, we identified several key factors that contributed to higher engagement:

- **Well-defined task**: Although tasks such as the “zoomed-in” setting and the “different sampling rate” condition were frequently requested by past participants, it proved challenging for the organizers to design task setups that were both fair and easy to understand. Striking a balance between research novelty and clarity of formulation remains a key difficulty in challenge design.
- **User-friendly baseline**: In VMC 2022 and AMC 2025, we dedicated more effort to developing comprehensive and easy-to-use baseline implementations, including pre-trained models, clear documentation, and ready-to-run scripts. These efforts greatly lowered the entry barrier for new participants, especially those unfamiliar with the task.
- **Marketing and advertisement**: This is arguably the most critical yet often underestimated factor in organizing a successful challenge. In 2023 and 2024, we did not actively promote the challenge through mailing lists and social media, and this lack of visibility likely contributed to the lower participation in those years.

III. OPEN-SOURCE TOOLKITS

In this section, we are particularly interested in SQA methods based on DNNs. This means they cannot be expressed in analytical forms, and are too costly for users to train the models from scratch. As a result, open-sourcing these models and making them easy to use becomes very important.

TABLE II
COMPARISON OF EXISTING OPEN-SOURCED SPEECH QUALITY
ASSESSMENT METHODS.

| Name | Inference | Model training | Multi-model | Multi-dataset |
|-----------------------|-----------|----------------|-------------|---------------|
| MOSNet [5] | ✓ | ✓ | | |
| DNSMOS [6] | ✓ | | ✓ | |
| NISQA [33] | ✓ | ✓ | | |
| SSL-MOS [15] | ✓ | ✓ | | |
| UTMOS [23] | ✓ | ✓ | | |
| TorchAudio-Squim [34] | ✓ | | ✓ | |
| SHEET [35] | ✓ | ✓ | ✓ | ✓ |
| VERSA [36] | ✓ | | ✓ | |

Table II shows a comparison between representative open-source SQA methods. In the early stages of SQA development, many publicly available codebases were primarily released as supplementary materials to their corresponding scientific papers, including MOSNet, DNSMOS, NISQA, SSL-MOS, and UTMOS [5], [6], [15], [23], [33]. While open-sourcing code has become a standard expectation in modern machine learning research, it was relatively uncommon in a smaller field like SQA at the time. This scarcity contributed to the widespread adoption of these early methods. However, these toolkits typically supported only the specific method proposed in the original paper, and were often trained on a single dataset, limiting their generalizability and practical applicability.

To support the *research* development of SQA systems, the SHEET toolkit [35] was developed with a particular focus on providing complete *training* and evaluation scripts, supporting a large collection of datasets and several representative SQA models. The goal was to provide a benchmark playground for both experienced researchers and newcomers to easily start working on this area.

Another line of work aims to provide a user-friendly interface to multiple off-the-shelf metrics and pre-trained SQA models, with the goal to enhance the accessibility of existing SQA metrics. For instance, TorchAudio-Squim [34] was designed to provide non-intrusive, reference-free quality measures for speech. Although only providing four metrics, its tight integration with TorchAudio [37], the official audio domain library of PyTorch, made it an easy-to-use building block for speech signal processing system development. Recently, the VERSA toolkit [36] was introduced as a unified, lightweight evaluation toolkit designed for not only speech but music and general audio. VERSA was already supporting 65 metrics with 729 variations based on different configurations by the time the paper was published, and it is planned to be continuously developed.

IV. CONCLUSION AND DISCUSSIONS

In this perspective paper, we first reflected on our experiences organizing the VoiceMOS and AudioMOS Challenge series over the past four years, highlighting key lessons on task design, baseline development, and community engagement. We also reviewed recent progress in open-source SQA toolkits and their role in accelerating research. These collective efforts have definitely inspired innovation and collaboration.

As challenge organizers, we have been constantly listening to the voice of the community. Although the field continues to expand toward general audio quality assessment, we would like to point out that despite the progress, automatic quality assessment of speech itself is still not solved, as we still constantly receive requests on evaluating more complex speech types like multi-lingual speech, expressive TTS, and prompt-based TTS. In addition, we are also seeing benchmarks that are publicly available and continuously evolving [38], [39]. We hope this paper not only serves as a reflection on past efforts but also as a call to continue building a shared foundation for the future of SQA and beyond.

Acknowledgments: This work was partly supported by JSPS KAKENHI Grant Number 25K00143.

REFERENCES

- [1] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech Quality Estimation: Models and Trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.
- [2] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*, W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 623–654.
- [3] 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*, vol. 45, no. 4, pp. 161–183, 2024.
- [4] 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.
- [5] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.
- [6] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.
- [7] G. Yi, W. Xiao, Y. Xiao, *et al.*, "ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications," in *Proc. Interspeech*, 2022, pp. 3308–3312.
- [8] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.
- [9] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.
- [10] "Methods for subjective determination of transmission quality," in *ITU-T Rec. P.800*, International Telecommunication Union (ITU-R), 1996.

- [11] A. Pavao, I. Guyon, A.-C. Letournel, *et al.*, “CodaLab Competitions: An Open Source Platform to Organize Scientific Challenges,” *Journal of Machine Learning Research*, vol. 24, no. 198, pp. 1–6, 2023.
- [12] E. Cooper and J. Yamagishi, “How do voices from past speech synthesis challenges compare today?” In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.
- [13] S. Watanabe, T. Hori, S. Karita, *et al.*, “ESPnet: End-to-End Speech Processing Toolkit,” in *Proc. Interspeech*, 2018, pp. 2207–2211.
- [14] O. Russakovsky, J. Deng, H. Su, *et al.*, “Imagenet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [15] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of MOS prediction networks,” in *Proc. ICASSP*, 2022, pp. 8442–8446.
- [16] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, “LDNet: unified listener dependent modeling in MOS prediction for synthetic speech,” in *Proc. ICASSP*, 2022, pp. 896–900.
- [17] 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 TASLP*, vol. 31, pp. 54–70, 2023.
- [18] 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 *Proc. ASRU*, 2023, pp. 1–7.
- [19] 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 & Language*, vol. 90, p. 101747, 2025.
- [20] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, “The Singing Voice Conversion Challenge 2023,” in *Proc. ASRU*, 2023, pp. 1–8.
- [21] Y.-W. Chen and Y. Tsao, “InQSS: a speech intelligibility and quality assessment model using a multi-task learning network,” in *Proc. Interspeech*, 2022, pp. 3088–3092.
- [22] 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,” in *Proc. ICME*, 2024, pp. 1–6.
- [23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4521–4525.
- [24] W.-C. Huang, S.-W. Fu, E. Cooper, *et al.*, “The Voicemos Challenge 2024: Beyond Speech Quality Prediction,” in *Proc. SLT*, 2024.
- [25] E. Cooper and J. Yamagishi, “Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech,” in *Proc. Interspeech*, 2023, pp. 1104–1108.
- [26] Y. Tang, J. Shi, Y. Wu, and Q. Jin, “Singmos: An extensive open-source singing voice dataset for mos prediction,” *arXiv preprint arXiv:2406.10911*, 2024.
- [27] I. Recommendation, “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm,” *ITU-T recommendation*, p. 835, 2003.
- [28] S. Leglaive, L. Borne, E. Tzinis, *et al.*, “The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement,” in *7th International Workshop on Speech Processing in Everyday Environments (CHiME)*, Aug. 2023.
- [29] S. Leglaive, M. Fraticelli, H. ElGhazaly, *et al.*, “Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge,” *arXiv preprint arXiv:2402.01413*, 2024.
- [30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating rnn-based speech enhancement methods for noise-robust text-to-speech,” in *SSW*, 2016, pp. 146–152.
- [31] C. Liu, H. Wang, J. Zhao, *et al.*, “MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation,” in *Proc. ICASSP*, 2025.
- [32] A. Tjandra, Y.-C. Wu, B. Guo, *et al.*, “Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound,” *arXiv preprint arXiv:2502.05139*, 2025.
- [33] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets,” in *Proc. Interspeech*, 2021, pp. 2127–2131.
- [34] A. Kumar, K. Tan, Z. Ni, *et al.*, “Torchaudio-Squim: Reference-Less Speech Quality and Intelligibility Measures in Torchaudio,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [35] W.-C. Huang, E. Cooper, and T. Toda, “SHEET: A Multi-purpose Open-source Speech Human Evaluation Estimation Toolkit,” in *Proc. Interspeech*, 2025.
- [36] J. Shi, H.-J. Shim, J. Tian, *et al.*, “VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music,” in *Proc. NAACL-HLT (System Demonstrations)*, 2025, pp. 191–209.
- [37] Y.-Y. Yang, M. Hira, Z. Ni, *et al.*, “Torchaudio: Building Blocks for Audio and Speech Processing,” in *Proc. ICASSP*, 2022, pp. 6982–6986.
- [38] W.-C. Huang, E. Cooper, and T. Toda, “MOS-Bench: Benchmarking Generalization Abilities of Subjective Speech Quality Assessment Models,” *arXiv preprint arXiv:2411.03715*, 2024.
- [39] Christoph Minixhofer and Ondřej Klejch and Peter Bell, “TTSDS2: Robust Objective Evaluation for Human-Quality Synthetic Speech,” in *Proc. SSW*, 2025, 68–75.