Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Towards Frame-level Quality Predictions of Synthetic Speech

Michael Kuhlmann[1], *Fritz Seebauer*[2], *Petra Wagner*[2], *Reinhold Haeb-Umbach*[1]

[1]Paderborn University, Germany
[2]Bielefeld University, Germany
{kuhlmann,haeb}@nt.upb.de, {fritz.seebauer,petra.wagner}@uni-bielefeld.de

Abstract

While automatic subjective speech quality assessment has witnessed much progress, an open question is whether an automatic quality assessment at frame resolution is possible. This would be highly desirable, as it adds explainability to the assessment of speech synthesis systems. Here, we take first steps towards this goal by identifying issues of existing quality predictors that prevent sensible frame-level prediction. Further, we define criteria that a frame-level predictor should fulfill. We also suggest a chunk-based processing that avoids the impact of a localized distortion on the score of neighboring frames. Finally, we measure in experiments with localized artificial distortions the localization performance of a set of frame-level quality predictors and show that they can outperform detection performance of human annotations obtained from a crowd-sourced perception experiment.

Index Terms: speech synthesis evaluation, SSQA, frame-level quality scores

1. Introduction

The automatic subjective speech quality assessment (SSQA) is an important task to track progress in the field of speech synthesis and to cut costs and time by being independent of crowd-sourced evaluation. With the introduction of self-supervised learned (SSL) encoders [1], much progress has been made with the results well correlated with human perception [2–4], and the focus now shifts to out-of-domain generalization [3,5]. Both automatic and perceptual assessments summarize the audio quality of a query signal with a single scalar value, typically in the interval $[1, 5]$, where 1 denotes poor and 5 excellent quality [6].

To train such systems, utterances produced by speech synthesis systems are annotated with quality scores gathered from crowd-sourced perception experiments. Although systems trained this way can successfully rank the overall quality of the query signal [1, 3, 4], current approaches do not look into explaining which parts of the signal contribute how much to the overall score. For the further improvement of synthesis systems, we believe that it becomes increasingly important to assess synthesis quality at a more fine-grained level. A hurdle in doing so is that there are no large-scale data sets for speech quality assessment with strong, i.e., frame-level, annotated quality scores.

So far, this problem has only been investigated by QualityNet [7]. They used a BLSTM network with negative forget gate bias initialization to discourage the use of global context information across frames and a frame-level objective to regularize frame-level scores. Although their analysis shares some similarities with ours, the forget gate bias initialization is not applicable to Transformer-based SSL encoders which we are going to study.

In this work, we approach the problem of inferring strong from weak predictions as follows. We start by identifying design choices of state of the art (SOTA) SSQA models that prevent us from extracting frame-level scores and show that there currently exists only one model from which frame-level scores are extractable. This brings us to the definition of criteria that *well-behaved* frame-level scores have to fulfill. We then introduce ChunkMOS, a new SSQA model to predict such well-behaved frame-level scores. To test whether the above-identified and the proposed frame-level score prediction network are able to localize distortions, we inject synthetic, but realistic, distortions into clean utterances and measure detection performance with an established metric from the sound event detection literature. The source code is publicly available[1].

2. Predicting well-behaved frame-level scores from utterance-level targets

Neural SSQA models follow an encoder/decoder paradigm [5, 8], where the encoder takes a raw speech signal as input and computes features with frame granularity, typically $50\,\mathrm{frames/s}$, which are further processed by the usually non-linear decoder to yield an utterance-level scalar prediction on the MOS scale. Common to all SOTA predictors is the use of SSL encoders such as wav2vec2 [9] or WavLM [10] for feature extraction. We denote the sequence of hidden embeddings in the l-th layer, $l = 1, \ldots, L$, of such encoders by $\mathcal{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \ldots, \mathbf{h}_T^{(l)}\}$. We start by analyzing the frame-level scores of SOTA SSQA models and show why they are unsuitable for judging frame-level quality.

2.1. Analyzing frame-level scores of SOTA SSQA models

The authors of [1] were the first to introduce the use of SSL encoders for SSQA. Their system performed equally well as more complicated systems in the VoiceMOS challenge 2022 [2]. All successful follow-up works relied on this SSL approach [3, 4] and recently, MOS-Bench [5], a new benchmark to compare SSQA systems, with the accompanying toolkit SHEET[2] was proposed, which offers pretrained models that build upon SSL-MOS [1]. Due to the success and popularity of SSL-MOS, we take it as the starting point for our analysis.

2.1.1. Issue 1a: Unbounded frame-level scores

SSL-MOS applies time average pooling (TAP) and a linear projection to the outputs $\mathcal{H}^{(L)}$ of the last layer of a finetuned

[1]https://github.com/michael-kuhlmann/frame-level-mos
[2]https://github.com/unilight/sheet

10.21437/Interspeech.2025-2190

wav2vec2 encoder. By rearranging the linear terms, we can write the predicted utterance-level score o as

$$o = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{\mathrm{T}}\mathbf{h}_t^{(L)} + b = \frac{1}{T}\sum_{t=1}^{T} s_t, \tag{1}$$

where we introduced the frame-level score s_t. Here, $\mathbf{w} \in \mathbb{R}^D$ is the learnable projection vector, D is the dimension of the hidden embedding $\mathbf{h}_t^{(L)} \in \mathbb{R}^D$ at time frame t, and b is a learnable scalar bias. Neither utterance- nor frame-level scores are bounded and, in theory, can take values in the interval $[-\infty, \infty]$. This is problematic as it will give different minimum and maximum scores across different utterances, preventing meaningful interpretation of frame-level scores.

The SHEET implementation of SSL-MOS [5] fixes this issue by (i) restricting the range of the scalar frame-level outputs with a $\tanh$ activation and applying proper scaling and shifting (range clipping) and (ii) performing TAP at the very last step of the model pipeline. This gives the following modified notion for the frame-level scores of SHEET SSL-MOS:

$$s_t = \gamma \tanh\left(\mathbf{w}^{\mathrm{T}} g\left(\mathbf{h}_t^{(L)}; \theta_{\mathrm{dec}}\right) + b\right) + \beta, \tag{2}$$

where $g(\cdot; \theta_{\mathrm{dec}}) : \mathbb{R}^D \to \mathbb{R}^{D_g}$ is a nonlinear decoder network and γ and β are proper scaling and shifting scalars (e.g., $\gamma = 2$ and $\beta = 3$). Note that the idea to perform range clipping was already proposed in the implementation of UTMOS [11]. However, they proposed to apply the loss to each frame instead of the time-pooled utterance-level score. We observed that this produced almost constant frame-level scores, which proved to be useless for judging frame-level quality.

2.1.2. Issue 1b: Inaccessible frame-level scores

Several works [4, 8, 12] followed the original SSL-MOS [1] by performing first TAP, but then processing the resulting embedding with a nonlinear decoder, giving the utterance-level score

$$o = \mathbf{w}^{\mathrm{T}} g\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_t^{(L)}; \theta_{\mathrm{dec}}\right) + b. \tag{3}$$

Frame-level scores cannot be derived from such a model because $g(\cdot)$ and the TAP operation cannot be swapped since the decoder is nonlinear here. Hence, frame-level scores become *inaccessible*.

2.1.3. Issue 2: Local-global coupling behavior

While the above analysis focused on the decoder side, we also observed a crucial problem on the encoder side that has been overlooked so far. Consider a model with accessible frame-level scores such as the original SSL-MOS and a query signal containing a local distortion such as additive noise. Figure 1 (top) shows the frame-level output of SSL-MOS before and after applying the local distortion to the query signal. Due to the infinite receptive field of the Transformer in the SSL encoder, the local distortion causes not only a frame-level score drop in the degraded area but also a mean shift of the frame-level scores. While this is not relevant from an utterance-level perspective, as long as the results correlate well with human perception, it becomes problematic when one is interested in attributing quality scores to frames: A single local event can have unpredictable effects on independent frames. We refer to this behavior of SSL-based SSQA models as *local-global coupling*.

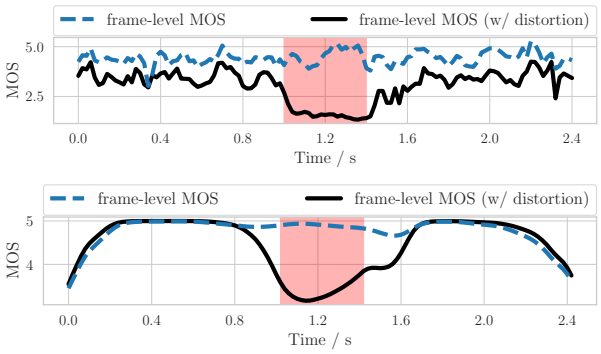

Figure 1: *Effect of adding a local distortion (pink noise) on frame-level MOS (blue: before, black: after adding the distortion). Because of the infinite receptive field of the Transformer, a local distortion additionally causes a mean shift (top), while it remains localized for ChunkMOS (bottom).*

2.2. On well-behaved frame-level scores

Based on the above analysis, we now state conditions to ensure that frame-level scores are accessible and better interpretable when training with an utterance-level loss. We define frame-level scores $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$ obtained under weak target training as *well-behaved* if they satisfy three conditions:

1. *Range clipping*: The range of frame-level scores is restricted to the MOS scale.
2. *Accessibility*: Utterance-level scores are obtained by applying affine transformations to the pooled frame-level scores.
3. *Local-global decoupling*: Given two disjoint intervals A and B over $\mathcal{S}$, A is invariant to changes in B.

Conditions 1 and 2 can be satisfied by following the design of SHEET SSL-MOS. Since MOS annotated data are scarce, there is currently no other way than relying on SSL encoders to obtain SOTA performance. Therefore, to satisfy condition 3, we propose to handle the input to the SSL encoder differently.

2.2.1. Decoupling the frame and utterance level

We propose to chunk a query signal $\mathcal{X} = \{x_1, \ldots, x_N\}$ into fixed-size blocks $\mathcal{B} = \{\mathcal{X}_1, \ldots, \mathcal{X}_{\lceil \frac{N-B}{M}+1 \rceil}\}$, each of length B with a block shift of $0 < M \leq B$ samples. If a query signal is too short ($N < B$), we zero-pad it to length B. Each chunk is processed independently by the SSL encoder which can be efficiently done in a batch-wise manner. If $M < B$, i.e., the blocks overlap, we perform overlap-add at the output of the SSL encoder before forwarding the result to the decoder.

Additionally, instead of using a single block length, we propose to chunk the query signal at different resolutions $B_1, \ldots, B_K$. Denote the output of the SSL encoder that receives inputs of block length B_k by $\mathcal{H}_k = \{\mathbf{h}_{k1}, \ldots, \mathbf{h}_{kT}\}$ after applying overlap-add. Then, the input to the decoder is a weighted sum of all outputs:

$$\mathbf{h}_t = \sum_{k=1}^{K} w_k \mathbf{h}_{kt}, \quad \sum_{k=1}^{K} w_k = 1, \tag{4}$$

where $w_1, \ldots, w_K$ are learnable parameters of the model. Note that this is conceptually similar to the spectrogram feature extractor in UTMOSv2 [12], where a weighted sum over the outputs of multiple CNNs that received spectrograms at different time-frequency resolutions was taken.

3. Experiments

3.1. Tested frame-level predictors

We experiment with the following systems: (a) as baselines, we take the pretrained checkpoints from the original SSL-MOS[3] and its SHEET implementation[4], from which we extract frame-level scores; (b) a reimplementation of SHEET SSL-MOS referred to as SHEET SSL-MOS: Re, and (c) the reimplementation with chunked inputs, called ChunkMOS. We train ChunkMOS with three block lengths of $1\,\mathrm{s}$, $600\,\mathrm{ms}$ and $400\,\mathrm{ms}$ with a shift of half a block length. SHEET SSL-MOS: Re differs from the SHEET recipe in the following points: (i) we do not finetune the CNN feature extractor of the wav2vec2 encoder; (ii) we take a weighted sum of all hidden embeddings of the encoder [4, 12], and (iii) the utterance-level loss function is a combination of a clipped MSE loss [13] with clipping $\tau = 0.1$ and a contrastive loss [11] with margin $\alpha = 0.1$. We found that these adaptations could improve the results. As decoder, we experiment with CNN and BLSTM architectures. For the CNN, we use three layers with kernel size three, hidden size 512 and leaky ReLU activation. For the BLSTM, we use three layers with hidden size 256 per direction. We normalize the loudness of each utterance to $-18\,\mathrm{dBFS}$ before forwarding it to the SSL encoder. We train models (b) and (c) for 60 epochs with an initial learning rate of 1×10^{-5}, which linearly decays to 1×10^{-6} at the end of the training. All models, including the baselines, were trained on the BVCC corpus [14]. For evaluation, we take the checkpoint with the highest system-level spearman rank correlation coefficient (SRCC) on the validation data.

3.2. Utterance- and system-level performance

We start by comparing the systems on the utterance and system levels, following the evaluation protocol of the recent VoiceMOS challenges [2, 3]. Since out-of-domain performance is an important aspect of SSQA models, we also evaluate on the test clean split of SOMOS [15]. Table 1 shows the results. For the in-domain scenario, all models perform equally well, with ChunkMOS performing a bit worse than SSL-MOS. This gap becomes more extreme for the out-of-domain scenario, where ChunkMOS clearly underperforms.

3.3. Local-global coupling behaviors

In Section 2.1.3, we have reasoned the local-global coupling behavior to be problematic for interpreting frame-level scores. Here, we are performing a large-scale quantitative evaluation to show that this problem indeed occurs. We take utterances from the test-clean split of LibriTTS-R [16], add a $1\,\mathrm{s}$ long pink noise distortion to each utterance and compute the frame-level scores. Around the distortion, we apply a collar of $200\,\mathrm{ms}$ and take the left and right subsequences of the frame-level scores. We extract these subsequences before and after applying the distortion. To compare the similarity of the subsequences before and after distortion, we calculate their Pearson correlation coefficient (PCC) and their distance given by the minimum alignment cost from dynamic time warping (DTW) (Table 2).

Clearly, the chunked processing effectively resolves the local-global coupling, in particular when using CNN or BLSTM decoder layers. Not only is the PCC the highest, the low DTW costs indicate that frame-level score trajectories are very similar without and with a distorted segment in the neighborhood. This is also exemplified for a single utterance in Figure 1 (bottom). For SHEET SSL-MOS, we made observations similar to those for UTMOS [11], i.e., the frame-level scores are mostly constant, leading to a more prominent mean shift. This has extreme detrimental effects on the detection performance (see below).

[3] https://github.com/nii-yamagishilab/mos-finetune-ssl/blob/main/run_inference.py

[4] https://huggingface.co/unilight/sheet-models/tree/main/bvcc/sslmos/2337

Table 1: *Utterance- and system-level results of various MOS predictors for in-domain (top) and out-of-domain (bottom) data. Best results are given in* **bold**. *Best results for each model family (resp. baselines) are underlined. MSE: mean squared error; LCC: linear correlation coefficient.*

	Utterance-level			System-level		
Model	MSE	LCC	SRCC	MSE	LCC	SRCC
SSL-MOS [1]	.277	.869	.869	.145	.924	.922
SHEET SSL-MOS [5]	.271	.867	.870	.123	.930	.931
SHEET SSL-MOS: Re	**.230**	.870	.872	**.088**	**.933**	**.930**
+ CNN	.248	**.879**	**.878**	.116	.922	.921
+ BLSTM	.303	.869	.867	.147	.918	.911
ChunkMOS	.537	.851	.856	.320	.916	.914
+ CNN	.550	.858	.859	.322	.922	.916
+ BLSTM	.490	.855	.855	.275	.915	.911

	Utterance-level			System-level		
Model	MSE	LCC	SRCC	MSE	LCC	SRCC
SSL-MOS [1]	.640	**.481**	**.490**	.453	.662	**.708**
SHEET SSL-MOS [5]	.849	.451	.467	.614	.643	.691
SHEET SSL-MOS: Re	**.546**	.476	.457	**.356**	**.683**	.693
+ CNN	.601	.467	.450	.416	.681	.693
+ BLSTM	.716	.458	.445	.537	.641	.667
ChunkMOS	1.296	.348	.346	1.125	.512	.582
+ CNN	1.345	.374	.375	1.174	.537	.612
+ BLSTM	1.272	.367	.363	1.088	.541	.601

3.4. Frame-level performance

There exist no dataset with frame-level quality predictions which can serve as ground truth for frame-level prediction. For this reason, we added synthetic localized distortions to a clean speech signal and evaluated how well the frame-level predictor can localize the corrupted frames. We acknowledge that this approach has deficiencies because the introduced distortions are artificial. Nevertheless, this serves as an indication whether frame-level predictions from utterance-level annotated training data are, in principle, possible.

3.4.1. Performance measure

To measure the localization accuracy of the detected events, we employ metrics from the evaluation of sound event detection systems. Specifically, we use the intersection-based criterion [17] to classify a prediction as a hit, miss, or false alarm, depending on how large the overlap of the detected segment containing a distortion is with the ground truth: If the overlap is larger than ρ_{DTC}, where $\rho_{\mathrm{DTC}} \in [0, 1]$, it is considered a hit, else a miss, and a false alarm is indicated if there was no distorted frame in the ground truth annotation of the detected region. To be independent of the detection threshold, we report the area under the receiver operating characteristic curve (AUC).

Table 2: *Change of frame-level scores after introducing a distortion. We measure Pearson correlation coefficient and dynamic time warping alignment cost of the unaffected subsequences (i.e., to the left and to the right of the distortion) before and after applying the distortion. Best results are given in* **bold**.

Model	lPCC	rPCC	lDTW	rDTW
SSL-MOS [1]	.792	.789	32.99	28.48
SHEET SSL-MOS [5]	.377	.386	144.61	114.70
SHEET SSL-MOS: Re	.804	.772	66.81	80.63
+ CNN	.880	.842	38.43	43.58
+ BLSTM	.681	.659	67.84	46.04
ChunkMOS	.867	.822	1.82	2.88
+ CNN	**.924**	**.905**	2.42	**3.31**
+ BLSTM	.913	.809	**1.73**	3.89

3.4.2. Artificial distortions

Additive noise. We use additive noise to simulate synthesis artifacts in the generation of fricatives. As such, distortion areas are sampled where the forced alignment indicates a fricative and we add **pink noise** with standard deviation $\sigma = 0.1$.
Phase artifacts. We compute the STFT of a single distortion area and replace the phase vectors with **random phase** vectors drawn from the uniform distribution $\mathcal{U}(-\pi, \pi)$. For phase randomziation, we constrain distortion areas to voiced segments.
Prosody artifacts. We simulate two types of prosody artifacts. First, we **shift the pitch** by either a quarter or half octave. We constrain distortion areas for pitch shifting to segments where a pitch was detected. Second, we **stretch the duration of vowels** by a random factor from the uniform distribution $\mathcal{U}(2, 3)$.

The injection is done as follows: For each utterance to be evaluated, we randomly sample a distortion class. Next, we sample $N = 3$ distortion areas such that the onset of an area aligns with the constraints of the sampled distortion and apply this distortion to all areas. We sample the duration for each area from the uniform distribution $\mathcal{U}(400\,\mathrm{ms}, 700\,\mathrm{ms})$ which corresponds to the average duration of an English word stem.

3.4.3. Detection performance

To measure detection performance, we use the intersection-based criterion with $\rho_{\mathrm{DTC}} = 0.5$ and $\rho_{\mathrm{DTC}} = 0.7$, where the latter penalizes less accurate detections more severely. Following post-processing in sound event detection, we apply median filters of varying duration ($0\,\mathrm{ms}$ to $500\,\mathrm{ms}$ in steps of $50\,\mathrm{ms}$) and choose the filter length with the best AUC on the dev-clean split. We use `sed_scores_eval`[5] [18] for hyperparameter tuning and AUC metric computation. Table 3 shows the results.

The results indicate that detection of distortions is possible using weakly obtained frame-level scores. For $\rho_{\mathrm{DTC}} = 0.5$, we achieve competitive performance with chunked inputs and a BLSTM decoder compared to an encoder with a global attention window. When detection requirements become stricter ($\rho_{\mathrm{DTC}} = 0.7$), ChunkMOS falls behind by a large margin. We made two observations that may explain this inexact localization. First, frame-level scores of ChunkMOS are inert, especially in combination with a BLSTM decoder, tending to predict too long activity, as can be seen in the comparison in Figure 1. Second, drops in the frame-level score curve tend to lag such that a local minimum falls together with the offset of a distortion, decreasing the possible overlap to a maximum of 50%.

[5] `https://github.com/fgnt/sed_scores_eval`

Table 3: *Intersection-based detection results of artificially distorted clean speech utterances from LibriTTS-R [16] test-clean, averaged over all distortion classes. Best results are highlighted in* **bold**. *Best results for each model family (resp. baselines) are underlined.*

Model	$\rho_{\mathrm{DTC}} = 0.5$		$\rho_{\mathrm{DTC}} = 0.7$	
	medfilt	AUC	medfilt	AUC
SSL-MOS [1]	400 ms	.650	350 ms	.513
SHEET SSL-MOS [5]	450 ms	.028	200 ms	.007
SHEET SSL-MOS: Re	300 ms	.603	250 ms	.451
+ CNN	450 ms	**.683**	300 ms	**.551**
+ BLSTM	450 ms	.595	350 ms	.451
ChunkMOS	500 ms	.544	500 ms	.266
+ CNN	500 ms	.622	500 ms	.317
+ BLSTM	500 ms	.682	500 ms	.341

Table 4: *Detection performance (AUC) of human annotations versus frame-level scores of SSQA models on 18 distorted utterances from the EARS [20] corpus.*

Method	$\rho_{\mathrm{DTC}} = 0.5$	$\rho_{\mathrm{DTC}} = 0.7$
Human Annotated	.361	.064
SSL-MOS [1]	.454	.234
SHEET SSL-MOS: Re+CNN	.543	.332
ChunkMOS+BLSTM	.408	.193

3.5. Comparison to human detection performance

In a final experiment, we compare the distortion localization to that of human annotations. In a small-scale listening experiment similar to [19], we asked 40 laypeople to mark areas they thought were detrimental to the overall quality of the speech signal. In total, 18 distorted utterances (6 per distortion class) from the EARS [20] corpus were queried. We counted how many participants marked a frame as detrimental and normalized this count by the number of ratings for that utterance to obtain soft scores. To tune the hyperparameters, we performed a 5-fold cross validation using the `SEBBs` toolbox[6] [21]. Table 4 shows the results. We find that weakly obtained frame-level scores from all models outperform the localization performance of laypeople, especially for strict detection requirements ($\rho_{\mathrm{DTC}} = 0.7$). This outcome does not encourage us to gather crowd-sourced frame-level targets but instead to further pursue weak learning of frame-level scores.

4. Discussion and Outlook

We have shown that weak learning of frame-level scores for SSL-based SSQA is possible and might be the best way to obtain these given the discouraging results of the crowd-sourced perception experiment. Although ChunkMOS is the only model that predicts *well-behaved* frame scores, it cannot yet achieve the detection performance of models that produce ill-behaved frame-level scores. Improving the performance of ChunkMOS is of ongoing interest. Ultimately, our goal is to apply the idea of frame-level scores to actual synthetic speech to discover system-specific error patterns.

[6] `https://github.com/merlresearch/sebbs`

5. Acknowledgements

This research was funded by Deutsche Forschungsgemeinschaft (DFG), project 446378607. Computational resources were provided by the Paderborn Center for Parallel Computing.

6. References

[1] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2022.

[2] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Interspeech*, 2022, pp. 4536–4540.

[3] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The Voicemos Challenge 2024: Beyond Speech Quality Prediction," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 803–810, 2024.

[4] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-n. Hsu, "Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound," *ArXiv*, vol. abs/2502.05139v1, 2025.

[5] W.-C. Huang, E. Cooper, and T. Toda, "MOS-Bench: Benchmarking Generalization Abilities of Subjective Speech Quality Assessment Models," *ArXiv*, vol. abs/2411.03715, 2024.

[6] I. T. Union, "ITU-T Rec. P.85, A method for subjective performance assessment of the quality of speech voice output devices," 1994.

[7] S. wei 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," in *Interspeech*, 2018, pp. 1873–1877.

[8] H. Wang, S. Zhao, X. Zheng, and Y. Qin, "RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in *Interspeech*, 2023, pp. 1095–1099.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.

[11] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Interspeech*, 2022, pp. 4521–4525.

[12] 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," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024.

[13] Y. Leng, X. Tan, S. Zhao, F. K. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391–395, 2021.

[14] E. Cooper and J. Yamagishi, "How do Voices from Past Speech Synthesis Challenges Compare Today?" in *Speech Synthesis Workshop*, 2021.

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

[16] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus," in *Interspeech*, 2023, pp. 5496–5500.

[17] Çağdaş Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2020.

[18] J. Ebbers, R. Serizel, and R. Haeb-Umbach, "Threshold Independent Evaluation of Sound Event Detection Scores," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1021–1025, 2022.

[19] F. Seebauer, M. Kuhlmann, R. Haeb-Umbach, and P. Wagner, "Re-examining the quality dimensions of synthetic speech," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[20] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Interspeech*, 2024, pp. 4873–4877.

[21] J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Sound Event Bounding Boxes," in *Interspeech*, 2024, pp. 562–566.