QMOS: Qwen-based MOS Prediction for TTS

Ravi Sastry Kolluru, Sri Charan Devarakonda,S Radhe Shyam Salopanthula, Anil Kumar Vuppala
LTRC, International Institute of Information Technology Hyderabad
{kolluru.s, sricharan.d, radheshyam.s}@research.iiit.ac.in, anil.vuppala@iiit.ac.in

***Abstract*—Text-to-speech systems are improving fast, but measuring how natural they sound still requires expensive human listening tests. Existing automatic methods struggle to generalise well across different datasets. We present QMOS, a MOS prediction framework that extracts hierarchical speech quality features from a frozen Qwen2-Audio large audio language model and combines them with layer-weighted WavLM SSL representations through a learned cross-attention fusion. This design captures both high-level semantic naturalness and low-level acoustic distortions in a unified model. On two standard benchmarks, SOMOS and BVCC, QMOS achieves competitive system-level SRCC of 0.923 and 0.921 on SOMOS and BVCC, respectively, while using no system-ID conditioning or listener embeddings. Cross-domain evaluation yields a competitive SRCC of 0.702, suggesting the learned representations generalise across acoustic domains.**

***Index Terms*—Text-to-Speech, Mean Opinion Score, Audio Language Models, Self-Supervised Learning, Speech Quality Assessment**

I. Introduction

Text-to-speech synthesis has advanced rapidly from concatenative and HMM-based systems to deep learning approaches: Tacotron [1] introduced end-to-end mel-spectrogram synthesis, FastSpeech [2] and FastSpeech 2 [3] achieved parallel non-autoregressive generation with explicit duration and pitch control, VITS [4] combined a variational autoencoder with normalising flows for end-to-end waveform synthesis, and diffusion-based models such as NaturalSpeech2 [5] and Voicebox [6] now produce speech virtually indistinguishable from human recordings. As synthesised speech quality approaches this ceiling, the challenge of reliably evaluating and comparing TTS systems has become as important as building them.

The gold standard for TTS evaluation remains the Mean Opinion Score test, defined in ITU-T P.800 [7], where human listeners rate speech naturalness on a five-point scale and scores are averaged across raters. While MOS captures true human perceptual judgement, conducting large-scale listening tests is expensive, time-consuming, and subject to significant inter-rater variability. Raters differ in hearing sensitivity, attention, and subjective quality standards, making the resulting labels inherently noisy aggregates rather than ground truth values.These limitations have driven a long line of research into automatic MOS prediction systems that can approximate human judgements at scale.

Early automatic metrics such as PESQ [8], POLQA [9], and STOI [10] are intrusive, requiring a clean reference signal for comparison, and correlate poorly with human MOS on modern neural TTS [11]. PESQ and POLQA were designed for telephony speech assessment and cannot evaluate synthesised speech where no reference exists, while STOI measures intelligibility but does not capture naturalness. Non-intrusive prediction began with MOSNet [12], which demonstrated that convolutional-recurrent neural networks trained directly on human MOS labels could predict speech quality from magnitude spectrograms without a reference signal, establishing the basic framework of direct regression from audio features to MOS which subsequent work extends. Subsequent efforts addressed listener bias [13], scoring offsets [14], multi-objective learning [15], [16], and full score distributions [17]. SSL representations from WavLM [18] and wav2vec 2.0 [19], pre-trained on thousands of hours of speech, brought large gains: SSL-MOS [20] showed that fine-tuning WavLM on MOS labels produces dramatically stronger predictions than spectral features alone, and UTMOS [21] achieved state-of-the-art performance on the BVCC benchmark [22] via an ensemble of strong SSL-based learners and weak classical regressors. SHEET [23], DNSMOS [24] and its probabilistic successor DNSMOS Pro [25], and FUSE-MOS [26] further improved generalisation across conditions, while SAMOS [27] combined semantic and acoustic features for multi-task MOS prediction.

Despite this progress, two fundamental limitations persist in current approaches. First, most methods rely on SSL features from a single pre-trained model, treating all layers uniformly or using only the final layer. Studies on SSL layer behaviour [28] show that lower layers encode low-level acoustic characteristics such as spectral envelope and pitch, while deeper layers encode phonetic categories and semantic content, yet this stratification is rarely exploited for MOS prediction. Second, SSL models are corpus-sensitive: features learned from one TTS evaluation corpus encode corpus-specific acoustic style alongside quality-relevant content, leading to significant performance degradation when models are evaluated across datasets.

In this work, we present QMOS, a MOS prediction framework that addresses these limitations. Our key contribution is Grouped Acoustic Pooling (GAP), which exploits the hierarchical layer structure of a frozen Qwen2-Audio large audio language model [29], a 32-layer generative transformer pre-trained on diverse audio understanding tasks. GAP stratifies the encoder into early, mid, and deep layer groups and combines them via an instance-adaptive soft gate, capturing phonetic, prosodic, and semantic quality signals simultaneously. These representations are fused with learnable WavLM layer-weighted SSL features through cross-attention, while a rater

disagreement encoder provides supplementary annotation context to a heteroscedastic prediction head. Training optimises a composite objective of heteroscedastic negative log-likelihood and linear concordance correlation.

II. Methodology

An overview of the proposed QMOS framework is illustrated in Figure 1. Given a speech utterance, QMOS extracts complementary quality representations from two frozen pretrained models and combines them with annotation context features to produce a final MOS prediction.

A. Feature Extraction

1) Grouped Acoustic Pooling: We use Qwen2-Audio [29] as a frozen feature extractor, a 32-layer transformer with 4096-dimensional hidden states pre-trained on diverse audio understanding tasks. Rather than using only the final layer, we partition the encoder into three functional depth groups and extract time-averaged representations from each:

$$G_k = \frac{1}{T \cdot |\mathcal{L}_k|} \sum_{t=1}^{T} \sum_{l \in \mathcal{L}_k} H_t^{(l)} \in \mathbb{R}^{4096} \tag{1}$$

where $\mathcal{L}_e = \{1, .., 10\}$, $\mathcal{L}_m = \{11, .., 22\}$, and $\mathcal{L}_d = \{23, .., 32\}$ denote the early, mid, and deep layer groups respectively, and $H_t^{(l)}$ is the hidden state at time step t and layer l. Early layers capture phonetic and spectral characteristics, mid layers encode prosody and rhythm, and deep layers represent high-level semantic naturalness and intelligibility.

2) SSL Features: We extract frame-level representations from all 24 layers of a frozen WavLM-Large [18] model and apply frame-level mean pooling to obtain layer-wise vectors:

$$\mathbf{S} = [S_1, \ldots, S_{24}] \in \mathbb{R}^{24 \times 1024} \tag{2}$$

$$S_k = \frac{1}{T} \sum_{t=1}^{T} H_t^{(k)} \tag{3}$$

Unlike Qwen2-Audio, whose generative pretraining induces a clear functional progression across layers motivating explicit group boundaries, WavLM-Large benefits from a continuous learned weighting that adapts to corpus-specific quality-relevant layer distributions, reflected in the dataset-dependent temperature τ.

3) Rater Disagreement Features: Where annotation metadata is available, we compute a 5-dimensional descriptor per utterance capturing the spread of rater scores: inter-rater standard deviation σ_s, normalised rater count $\tilde{N}$, normalised minimum and maximum scores, and Pearson's second skewness coefficient capturing asymmetry in the rating distribution.

$$\mathbf{r} = \left[\sigma_s,\ \tilde{N},\ \frac{s_{\min}}{5},\ \frac{s_{\max}}{5},\ \frac{\bar{s} - \text{med}(s)}{\sigma_s + \epsilon}\right] \in \mathbb{R}^5 \tag{4}$$

This descriptor is motivated by the substantial annotation variability present in MOS datasets: SOMOS exhibits a mean inter-rater standard deviation of σ=1.16 (range 0–2.05), with high-disagreement systems averaging σ=1.39 compared to 0.57 for consistent ones. Providing the prediction head with an explicit disagreement signal allows it to modulate confidence as a function of annotation reliability, at negligible parameter cost.

B. GAP Encoder

Each group representation G_k is independently normalised via LayerNorm. An instance-adaptive soft gate then assigns per-sample mixing weights across the three groups:

$$\mathbf{w} = \text{Softmax}\Big(W_2 \tanh\Big(W_1[\tilde{G}_e \| \tilde{G}_m \| \tilde{G}_d]\Big)\Big) \in \mathbb{R}^3 \tag{5}$$

$$\mathbf{G}_w = \text{stack}\left[w_e \tilde{G}_e,\ w_m \tilde{G}_m,\ w_d \tilde{G}_d\right] \tag{6}$$

This allows the model to dynamically emphasise semantic layers for high-quality utterances and acoustic layers for distorted ones on a per-sample basis. The gated representations form a matrix of shape (3×4096); mean and standard deviation are computed across the three group vectors to capture the average and spread of quality signals, while max pooling surfaces the most dominant group response. These are projected and fused into a compact 256-dimensional GAP representation $\mathbf{z}_{\text{gap}}$.

C. SSL Branch and Cross-Attention Fusion

Not all WavLM layers contribute equally to MOS prediction. We learn a temperature-scaled softmax distribution over the 24 layers to softly select the most quality-informative combination:

$$w_k = \frac{\exp(s_k/\tau)}{\sum_{j=1}^{24} \exp(s_j/\tau)} \tag{7}$$

$$\mathbf{z}_{\text{ssl}} = \text{Proj}\left(\sum_{k=1}^{24} w_k S_k\right) \in \mathbb{R}^{256} \tag{8}$$

where $\mathbf{s} \in \mathbb{R}^{24}$ are learned scalar weights and τ is set to 1.0 for BVCC and 2.0 for SOMOS, reflecting their different quality representation strategies. The GAP and SSL representations are each treated as single-token sequences and fused via multi-head cross-attention, where $\mathbf{z}_{\text{gap}}$ acts as the query and $\mathbf{z}_{\text{ssl}}$ provides the key and value. This implements a query-conditioned projection in which the GAP context modulates the retrieval of SSL acoustic detail:

$$\mathbf{z}_{\text{attn}} = \text{LN}(\mathbf{z}_{\text{gap}} + \text{MHA}(Q{=}\mathbf{z}_{\text{gap}},\ K{=}V{=}\mathbf{z}_{\text{ssl}})) \tag{9}$$

This asymmetric design allows the global semantic quality context from GAP to selectively retrieve relevant acoustic detail from SSL through the residual connection.

D. Prediction Head

The rater disagreement vector is batch-normalised and projected via a linear layer with GELU activation to a 32-dimensional encoding $\mathbf{z}_{\text{rd}}$. The attended GAP output, SSL projection, and disagreement encoding are concatenated and passed through a three-layer MLP (544→256→128→64) with GELU activations and dropout. The final MOS prediction is produced by a sigmoid-scaled output head:

$$\hat{y} = 1 + 4 \cdot \sigma(\hat{\mu}) \in (1, 5) \tag{10}$$

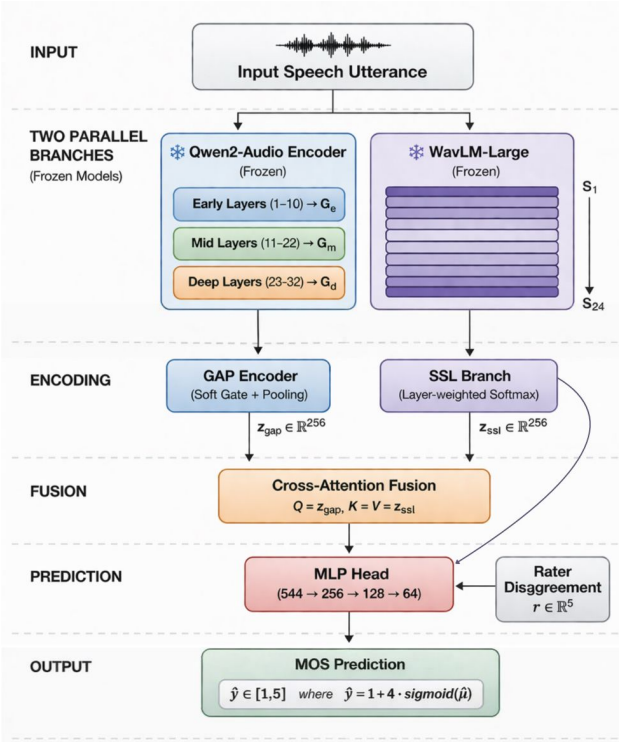

Fig. 1. Overview of the proposed QMOS methodology.

which constrains predictions to the valid MOS range. The model also produces a secondary variance output $\hat{\sigma}^2 = e^{\text{clamp}(\hat{\ell}, -6, 4)}$, used by the NLL loss during training to down-weight uncertain samples.

E. Training Objective

QMOS is trained with a composite loss combining two complementary terms:

$$\mathcal{L} = 0.3 \cdot \mathcal{L}_{\text{NLL}} + \lambda_{\text{lcc}} \cdot \mathcal{L}_{\text{LCC}} \tag{11}$$

where $\lambda_{\text{lcc}} = 0.1$ for Variant A and 0.5 for Variant B.

Heteroscedastic NLL. The model output is treated as a Gaussian $\hat{y} \sim \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2)$, giving:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \left[\hat{\ell}_i + \frac{(y_i - \hat{\mu}_i)^2}{e^{\hat{\ell}_i} + \epsilon} \right] \tag{12}$$

This formulation down-weights samples where the model predicts high variance, providing a form of implicit uncertainty weighting during training.

Linear Concordance Correlation. We minimise Pearson correlation between predictions and labels, aligning the training objective with the primary evaluation metric:

$$\mathcal{L}_{\text{LCC}} = 1 - \text{LCC}(\hat{y}, y) \tag{13}$$

III. Experimental Setup

A. Datasets and Evaluation Protocol

We evaluate QMOS on two standard benchmarks. SOMOS contains 19,504 utterances from 200 neural TTS systems with crowdsourced MOS labels on a 1–5 scale, using the 14,100 / 2,300 / 2,300 train / validation / test split. BVCC, the benchmark of VoiceMOS Challenge 2022 [22], covers both TTS and voice conversion systems using the 4,974 / 1,066 / 1,066 split. We report MSE, Pearson LCC, and Spearman SRCC at utterance and system level following the VoiceMOS evaluation protocol, with system-level SRCC as the primary metric.

B. Training Configuration

All models are trained on a single NVIDIA A100 40 GB GPU with AdamW (weight decay 5×10^{-3}) and batch size 32 for up to 100 epochs with early stopping (patience 15). We use a 5-epoch linear warmup followed by cosine annealing to zero. The learning rate is 1×10^{-4}; both Qwen2-Audio and WavLM-Large are used as frozen feature extractors throughout. C-Mixup [31] (β=0.5, $|\Delta\text{MOS}|$<0.5) is applied on BVCC for generalisation. The WavLM layer temperature τ is set to 1.0 for BVCC and 2.0 for SOMOS. Loss weights are α_{NLL}=0.3 and λ_{LCC}=0.5.

IV. Results

Baseline scores for all compared systems are taken directly from the FUSE-MOS paper [26] for both datasets, ensuring a consistent and fair comparison under identical evaluation protocols.

A. SOMOS Results

Table I presents utterance- and system-level results on SOMOS. QMOS-B achieves a system-level SRCC of 0.923 ($\pm$0.032, 95% bootstrap CI, n=3000) and utterance-level SRCC of 0.732, outperforming all baselines including FUSE-MOS (sys-SRCC 0.910, utt-SRCC 0.676) and Content-Aware SSL (sys-SRCC 0.917). QMOS-A, using only the GAP branch without SSL fusion, already reaches sys-SRCC 0.894 and utterance-level MSE of 0.190, competitive with FUSE-MOS despite using no WavLM features. The improvement from Variant A to Variant B (+0.029 sys-SRCC, $-$0.023 utt-MSE) directly reflects the contribution of WavLM cross-attention fusion: layer-weighted SSL features, selectively attended by GAP queries, recover fine-grained acoustic quality distinctions that Qwen2-Audio representations alone do not capture, consistent with the observation that deep layers of a generative audio LLM are optimised for semantic understanding rather than perceptual quality discrimination.

B. BVCC Results

Table II presents results on BVCC. QMOS-B achieves a system-level SRCC of 0.921 ($\pm$0.026, 95% bootstrap CI, n=3000) and utterance-level SRCC of 0.868, competitive with UTMOS-strong (sys-SRCC 0.936) and FUSE-MOS (sys-SRCC 0.947) while using no system-ID conditioning or listener embeddings. Notably, UTMOS-strong fine-tunes the full SSL encoder whereas QMOS keeps both Qwen2-Audio and WavLM-Large frozen throughout, making the gap smaller than the architectural difference might suggest. QMOS-A reaches sys-SRCC 0.861, trailing LDNet (0.886) and comparable to

TABLE I
RESULTS ON SOMOS. †95% BOOTSTRAP CI (n=3000) FOR QMOS-B SYS-SRCC: 0.923 ± 0.032.

Model	Utterance-level			System-level		
	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑
DeePMOS [30]	0.267	0.504	0.475	0.060	0.805	0.791
DNSMOS Pro [25]	0.429	0.484	0.462	–	–	–
Content-Aware SSL [26]	0.203	0.687	0.681	0.052	0.911	0.917
Content-Aware SSL+BERT [26]	0.179	0.668	0.659	0.021	0.906	0.913
FUSE-MOS [26]	0.175	0.686	0.676	0.020	0.907	0.910
QMOS-A (ours)	0.190	0.651	0.639	0.019	0.892	0.894
QMOS-B (ours)†	**0.167**	**0.703**	**0.732**	**0.017**	**0.924**	**0.923**

TABLE II
RESULTS ON BVCC. †95% BOOTSTRAP CI (n=3000) FOR QMOS-B SYS-SRCC: 0.921 ± 0.026.

Model	Utterance-level			System-level		
	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑
LDNet [13]	0.333	0.795	0.794	0.169	0.885	0.886
DeePMOS [30]	0.361	0.782	0.774	0.187	0.872	0.866
UTMOS-strong [21]	**0.165**	**0.899**	**0.896**	0.090	0.936	0.936
SAMOS [27]	–	–	–	0.097	0.944	0.942
FUSE-MOS [26]	0.191	0.887	0.887	0.086	**0.946**	**0.947**
QMOS-A (ours)	0.365	0.777	0.776	0.096	0.866	0.861
QMOS-B (ours)†	0.216	0.869	0.868	**0.051**	0.921	0.921

DeePMOS (0.866) without any SSL cross-attention. The improvement from Variant A to Variant B on BVCC (+0.060 sys-SRCC) is notably larger than on SOMOS (+0.029), consistent with BVCC spanning a more acoustically diverse set of systems including voice conversion, where low-level WavLM layers 5–7 provide discriminative spectral quality cues that complement the higher-level representations of Qwen2-Audio. FUSE-MOS achieves the highest system-level SRCC of 0.947, followed by SAMOS at 0.942 (LCC 0.944), while QMOS-B is competitive at 0.921 and operates as a single model without ensembling. The remaining gap relative to FUSE-MOS, SAMOS, and UTMOS-strong is attributable in part to the frozen encoder constraint, as neither Qwen2-Audio nor WavLM-Large is fine-tuned on MOS data. This gap is further amplified by the nature of BVCC, which contains a substantial proportion of voice conversion systems whose artefacts (e.g., spectral mismatch, timbre leakage, and prosodic inconsistency) are more complex and less aligned with semantic representations than typical TTS errors, making them harder to capture without task-specific adaptation.

C. Cross-Domain Generalisation

We evaluate each model on the held-out test set of the opposite dataset without any fine-tuning to assess how well quality representations transfer across corpora. The two datasets differ not only in size but in acoustic scope: SOMOS is a purely neural TTS benchmark where all 200 systems synthesise from the same source, producing a relatively homogeneous acoustic domain; BVCC, by contrast, spans both TTS and voice conversion systems, introducing a substantially wider range of spectral distortions and prosodic degradations.

Training on SOMOS and evaluating on BVCC, QMOS-B achieves sys-SRCC 0.695 against QMOS-A at 0.620, with the fusion branch again providing a consistent lift. In the reverse direction (BVCC $\rightarrow$ SOMOS), QMOS-B reaches 0.702 against QMOS-A at 0.597. The stronger transfer in the BVCC $\rightarrow$ SOMOS direction is expected: a model trained on the broader acoustic diversity of BVCC generalises well to the narrower homogeneous TTS domain of SOMOS, whereas a model trained only on clean neural TTS has not encountered the spectral artefacts characteristic of voice conversion systems, limiting transfer in the opposite direction.

D. Ablation Study

Table III reports the contribution of each training objective component and the rater disagreement encoder on the in-domain test sets of both SOMOS and BVCC.

Removing $\mathcal{L}_{\text{LCC}}$ produces the largest degradation across both benchmarks, with system-level SRCC dropping from 0.923 to 0.862 on SOMOS and from 0.921 to 0.826 on BVCC, confirming that explicit correlation alignment is the most critical training signal for both in-domain accuracy and cross-dataset consistency. Removing $\mathcal{L}_{\text{NLL}}$ has a comparatively smaller effect: system-level SRCC remains competitive at 0.921 on SOMOS and 0.918 on BVCC, though utterance-level SRCC on SOMOS rises marginally to 0.727, suggesting the heteroscedastic weighting slightly suppresses utterance-level discrimination in exchange for better-calibrated system-level ordering. Removing the rater disagreement encoder causes a modest but consistent drop in system-level SRCC on both datasets (0.920 vs. 0.923 on SOMOS; 0.913 vs. 0.921 on BVCC), demonstrating that annotation spread context provides

TABLE III
ABLATION STUDY ON SOMOS AND BVCC (QMOS-B). SYS-SRCC IS THE PRIMARY METRIC.

Variant	SOMOS		BVCC	
	Utt-SRCC↑	Sys-SRCC↑	Utt-SRCC↑	Sys-SRCC↑
QMOS-B (full)	**0.732**	**0.923**	**0.868**	**0.921**
w/o $\mathcal{L}_{\text{LCC}}$	0.693	0.862	0.810	0.826
w/o $\mathcal{L}_{\text{NLL}}$	<u>0.727</u>	0.921	0.532	0.612
w/o Rater Disagr.	0.698	0.920	0.875	0.913

a useful supplementary signal despite its negligible parameter cost.

V. CONCLUSION

We presented QMOS, a MOS prediction framework that combines Grouped Acoustic Pooling over a frozen Qwen2-Audio encoder with layer-weighted WavLM cross-attention fusion, trained using a composite objective of heteroscedastic NLL and linear concordance correlation. Experiments on SOMOS and BVCC demonstrate that fusing the high-level semantic representations of a generative audio language model with low-level SSL acoustic detail is an effective strategy for MOS prediction, with consistent gains from the cross-attention branch across both benchmarks. Ablation results confirm that the LCC objective is the single most critical training component, while the rater disagreement encoder provides a consistent supplementary benefit at negligible cost. Cross-domain evaluation further suggests that training on acoustically diverse data improves generalisation. Future work will explore fine-tuning the Qwen2-Audio encoder for quality-specific tasks, extending QMOS to non-English TTS evaluation, and moving beyond single-score prediction toward dimension-wise quality estimation that better reflects the multi-faceted nature of human perceptual judgement.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019, pp. 3165–3174.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[5] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. ICML*, 2023.

[6] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Proc. NeurIPS*, 2023.

[7] ITU-T, "Methods for subjective determination of transmission quality," ITU-T Recommendation P.800, 1996.

[8] 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*, 2001, pp. 749–752.

[9] ITU-T, "Perceptual objective listening quality analysis," ITU-T Recommendation P.863, 2011.

[10] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[11] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv:1611.06471*, 2016.

[12] C. Lo, S. Fu, W. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H. Wang, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[13] W. Yu, E. Cooper, Y. Tsao, H. Wang, T. Toda, and J. Yamagishi, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. ICASSP*, 2022, pp. 7087–7091.

[14] Y. Leng, X. Chen, A. Wu, S. Liu, Z. Zhao, X. Tan, L. He, T. Qin, W. Wang, and T. Liu, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *Proc. ICASSP*, 2021, pp. 391–395.

[15] W. Tseng, Y. Lin, H. Lee, and L. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. Interspeech*, 2021, pp. 2781–2785.

[16] S. Bott, Z. Tüske, and S. Renals, "Non-intrusive speech quality assessment with multi-task learning," in *Proc. Interspeech*, 2023.

[17] Y. Leng, X. Tan, S. Liu, H. Duan, Y. Qin, and T. Liu, "MBNET: Distribution-based MOS prediction for synthesized speech," *arXiv preprint*, 2022.

[18] 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, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.

[20] W. Tseng, Y. Lin, H. Lee, and L. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. ICASSP*, 2022, pp. 7087–7091.

[21] 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.

[22] W. Huang, E. Cooper, J. Yamagishi, and T. Toda, "The VoiceMOS challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[23] E. Cooper, W. Huang, T. Toda, and J. Yamagishi, "SHEET: Single-headed attention with efficient training for non-intrusive speech quality evaluation," in *Proc. Interspeech*, 2023.

[24] 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*, 2022, pp. 261–265.

[25] T. Cumlin, E. Cooper, and J. Yamagishi, "DNSMOS Pro: A reduced-size DNN for probabilistic MOS prediction," in *Proc. Interspeech*, 2024.

[26] M. Hoq, T. Kinnunen, and P. Alku, "FUSE-MOS: Fusion of speech embeddings for MOS prediction," in *Proc. Interspeech*, 2025.

[27] Y.-F. Shi, Y. Ai, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "SAMOS: A neural MOS prediction model leveraging semantic representations and acoustic features," *arXiv:2411.11232*, 2024.

[28] A. Pasad, J. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, 2021, pp. 914–921.

[29] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, and L. Zhou, "Qwen2-audio technical report," *arXiv:2407.10759*, 2024.

[30] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, "DeePMOS: Deep posterior mean-opinion-score of speech," in *Proc. Interspeech*, 2023, pp. 526–530.

[31] H. Yao, C. Wang, S. Li, L. Nie, M. Dong, J. Li, W. Xu, and C. Finn, "C-Mixup: Improving generalization in regression," in *Proc. NeurIPS*, 2022.