HighRateMOS: Sampling-Rate Aware Modeling for Speech Quality Assessment

Wenze Ren[1†], Yi-Cheng Lin[1†], Wen-Chin Huang[2], Ryandhimas E. Zezario[3], Szu-Wei Fu[4], Sung-Feng Huang[4], Erica Cooper[5], Haibin Wu[6], Hung-Yu Wei[1], Hsin-Min Wang[3], Hung-yi Lee[1], Yu Tsao[3]
[1]*National Taiwan University*, [2]*Nagoya University*, [3]*Academia Sinica*, [4]*NVIDIA*
[5]*National Institute of Information and Communications Technology*, [6]*Independent Researcher*
{r11942166, f12942075, hungyilee}@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw

arXiv:2506.21951v1 [eess.AS] 27 Jun 2025

***Abstract*—Modern speech quality prediction models are trained on audio data resampled to a specific sampling rate. When faced with higher-rate audio at test time, these models can produce biased scores. We introduce HighRateMOS, the first non-intrusive mean opinion score (MOS) model that explicitly considers sampling rate. HighRateMOS ensembles three model variants that exploit the following information: (i) a learnable embedding of speech sampling rate, (ii) Wav2vec 2.0 self-supervised embeddings, (iii) multi-scale CNN spectral features, and (iv) MFCC features. In AudioMOS 2025 Track3, HighRateMOS ranked first in five out of eight metrics. Our experiments confirm that modeling the sampling rate directly leads to more robust and sampling-rate-agnostic speech quality predictions.**

***Index Terms*—Speech quality assessment, Mean opinion score (MOS)**

I. Introduction

Speech quality is critical for modern communication and multimedia systems, as degradation of audio quality leads to loss of intelligibility and poor user experience. To improve audio quality, recent Text-to-Speech (TTS) systems produce speech at sampling rates higher than 16 kHz. For example, Parallel WaveGAN [1] generates 24 kHz audio in a Transformer-based TTS framework, while HiFi-GAN [2] synthesizes 22.05 kHz speech. As TTS moves to higher sampling rates, evaluations must handle diverse sampling rates to account for differences in frequency range and listening performance.

Most non-intrusive neural predictors of speech quality are trained on data with a single sampling rate. For example, Quality-Net [3], SSL-MOS [4], MOSNet [5], and MOSA-Net [6] were trained on speech uniformly downsampled to 16 kHz. This leaves the question of how well these models perform on audio recorded or played back at other sampling rates.

In this work, we propose **HighRateMOS**, an ensemble of three models to predict speech quality at different sampling rates. Each model takes sampling rate embeddings, self-supervised learning (SSL) model representations, and Mel-spectrograms as input features. The complete architecture is depicted in Figure 1. HighRateMOS achieved first place in five of the eight evaluation metrics in Track 3 of the AudioMOS Challenge 2025. Compared to the baseline system, the system-level Spearman Rank Correlation Coefficient (SRCC) is improved by 27.5% and Kendall's τ (KTAU) is improved by 53.9%.

II. Methodology

A. Feature Extraction

We encode the sampling rate (16 kHz, 24 kHz, or 48 kHz) of each utterance into a learned embedding vector, which allows the model to adjust to the sampling rate of the input signal. To capture a broad range of acoustic information, we use SSL embeddings along with Mel-spectrograms and MFCCs. Pretrained SSL models such as wav2vec 2.0 [7], HuBERT [8], and WavLM [9] are trained on raw waveforms sampled at 16 kHz, and their convolutional front ends assume this sampling rate. In previous studies, audio recorded at 24 kHz or 48 kHz is first converted to 16 kHz using downsampling. We posit that an SSL encoder pretrained at 16 kHz can treat audio at a higher sampling rate as the same content played at a slower speed. For example, a 48 kHz waveform resembles 16 kHz speech stretched three times, with the same temporal pattern as the encoder. Because SSL models are trained on large and diverse corpora covering a wide range of speaking rates and acoustic conditions, they can characterize content even when the speech rate is slower.

For the Mel-spectrogram, we use a multi-scale CNN to capture both fine-grained and broad spectral–temporal patterns that affect perceived speech quality. Convolutional filters at different time–frequency scales enable the model to detect narrowband artifacts alongside wider distortions within the same architecture.

B. Feature Aggregation

In this challenge, we trained three models. As shown in Figure 1, Model 1 takes an intuitive approach by concatenating the frame-level outputs of the SSL encoder, the sampling rate vector, and the multi-scale Mel-spectrograms along the channel dimension to obtain a unified representation; Model 2 introduces a cross-attention mechanism on top of Model 1, which selectively focuses on relevant acoustic patterns from the self-supervised features and the spectral representations; Model 3 integrates all components. After combining multiple features, temporal aggregation is performed via BLSTM, and the final MOS prediction is obtained by a fully connected layer. This design ensures that each model variant can capture different aspects of speech quality.

C. Loss Function

We divide the eight losses into four categories: point-wise regression loss, ranking-based loss, correlation-based loss, and hybrid loss.

1) Point-wise regression loss: We explore the Mean Squared Error (MSE) [3] and Mean Absolute Error (MAE) [10] for minimizing the average squared/absolute difference.

2) Ranking-based loss: Ranking-based loss functions encourage the model to preserve the relative ordering of samples by MOS. We employ the contrastive loss [11], which encourages the model not only to predict correct absolute MOS values but also to match the relative difference between any two utterances. We also employ the Relative Ranking Loss (RelRank), which is the triplet loss among extreme values in a batch [12].

3) Correlation-based loss: Correlation-based loss functions directly optimize the statistical agreement between predictions and human ratings. We explore Linear Correlation Coefficient (LCC) loss

[†] Equal Contribution to this work.

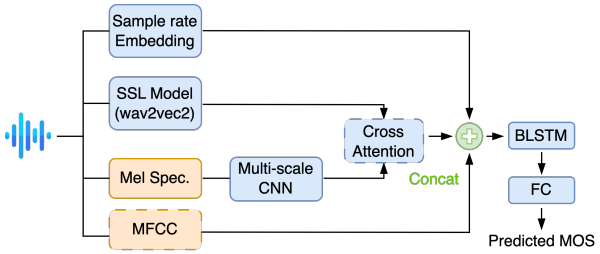


Fig. 1. Proposed model architecture. Three models are trained. Model 1 does not include the cross-attention module and MFCC, Model 2 adds the cross-attention module to Model 1, and Model 3 additionally incorporates MFCC training with all blocks.

(defined as one minus the Pearson linear correlation between predicted and reference MOS) and Concordance Correlation Coefficient (CCC) loss (defined as one minus the CCC), which jointly maximizes both precision and accuracy of the predictions.

4) Hybrid loss: Hybrid losses aim to leverage the complementary strengths of different objectives. We also incorporate widely used losses such as Dual Criterion Quality (DCQ) loss [13], which jointly enforces both precise numerical estimation and correct ordinal ranking, and UTMOS loss [14], which is the sum of clipped MSE loss and $0.5\times$ contrastive loss. We use the same clipping parameter of 0.25 as the original paper.

III. Experimental Setup

A. Dataset

We use the AudioMOS 2025 Track 3 dataset, which contains 1,200 speech samples produced by TTS and neural vocoder systems, each rated for naturalness on a 1-5 Likert scale by 10 annotators. The dataset covers three sampling rates: 16 kHz, 24 kHz, and 48 kHz. The training and development sets share the same 400 utterances but use different listening protocols. In the training set, speech samples at each sampling rate were evaluated in separate sessions, so listeners only compared stimuli with the same sampling rate. Unlike the training set, in the development set, speech samples of the three sampling rates were interleaved in a listening session. The evaluation set follows the same mixed-rate protocol as the development set, requiring listeners to rate stimuli of 16 kHz, 24 kHz, and 48 kHz in one session.

We report two sets of results: development performance during training and final evaluation performance. During the training phase, since labels for the development set are not available, we can only obtain the performance of the model on it by uploading predictions to the challenge platform. We split the training set into five folds and perform five-fold cross-validation, training on four folds and validating on the held-out fold. We then aggregate the predictions from the five held-out folds to evaluate the development performance. In the evaluation phase, we follow the standard protocol of training on the full training set, selecting the best model on the development set (the labels are available in the evaluation phase), and reporting results on the evaluation set.

B. Models

We use the SHEET[1] [15] toolkit and SSL feature extractors from the S3PRL[2] [16] toolkit to implement our models. The SSL models explored include wav2vec 2.0 Base (W2V2), Large (W2V2 L) [7]; WavLM Base+ (WavLM B), Large (WavLM L) [9]; Data2vec Large (D2V) [17], and XLSR-1B (XLSR) [18]. We use AdamW [19] optimizer with a learning rate of 1e-3, and train the model until reaching system-level SRCC convergence (stops increasing for 2000 steps). The batch size is 8.

[1] https://github.com/unilight/sheet
[2] https://github.com/s3prl/s3prl

C. Evaluation Metrics

We employ a set of metrics at the utterance and system levels, including MSE for measuring prediction accuracy, LCC for evaluating linear relationships, SRCC for capturing monotonic associations, and KTAU for evaluating the ordinal agreement between predicted scores and ground truth scores [20].

TABLE I
Official evaluation results of all submitted systems for AudioMOS Track 3. **Bold**: the best results.

System	Utterance-level MSE	LCC	SRCC	KTAU	System-level MSE	LCC	SRCC	KTAU
B03	0.273	0.821	0.695	0.508	0.119	0.941	0.749	0.547
T01	0.303	0.804	0.643	0.459	0.104	0.957	0.866	0.705
T08	0.277	0.811	0.716	0.529	**0.056**	0.978	0.913	0.758
T11	0.282	0.813	0.714	0.536	0.085	0.968	0.917	0.789
T13	0.298	0.796	0.671	0.487	0.090	0.972	0.926	0.779
T16	0.287	0.830	0.723	**0.589**	0.071	0.952	0.891	0.750
T19	**0.238**	0.846	0.694	0.513	0.080	0.955	0.914	0.758
Ours	0.303	**0.847**	**0.742**	0.556	0.116	**0.982**	**0.955**	**0.842**

TABLE II
Comparison between losses. **Bold**: best; underline: second-best.

Method	Development MSE	LCC	SRCC	KTAU	Evaluation MSE	LCC	SRCC	KTAU
MAE	0.095	0.942	0.768	0.611	0.130	0.958	0.926	0.768
MSE	0.081	0.955	0.779	0.632	0.136	0.956	0.905	0.758
RelRank	0.049	0.954	0.824	0.653	0.088	0.958	0.929	0.800
LCC	0.545	0.938	**0.887**	**0.811**	0.424	0.937	**0.962**	**0.863**
CCC	0.069	0.969	0.836	0.674	**0.063**	0.973	0.956	0.832
DCQ	**0.041**	0.967	0.863	0.663	0.067	0.973	0.946	0.811
UTMOS	0.092	0.947	0.773	0.621	0.117	0.962	0.920	0.779
contrastive	0.049	**0.972**	0.880	0.716	0.124	**0.982**	0.949	0.821

TABLE III
System-level performance of different SSL upstreams on development and evaluation sets. **Bold**: best; underline: second-best.

Upstream	Development MSE	LCC	SRCC	KTAU	Evaluation MSE	LCC	SRCC	KTAU
W2V2 L	0.055	0.969	0.832	0.663	0.047	0.984	**0.980**	**0.905**
D2V L	0.070	0.949	0.814	0.684	0.239	0.986	0.976	0.884
XLSR	**0.020**	**0.980**	0.872	**0.726**	0.199	**0.987**	0.976	0.874
WavLM L	0.036	0.964	0.812	0.663	0.100	0.975	0.956	0.832
WavLM B	0.033	0.968	0.862	0.716	**0.040**	0.978	0.941	0.800
W2V2	0.049	0.972	**0.880**	0.716	0.124	0.982	0.949	0.821

IV. AudioMOS 2025 Challenge Results

Table I shows the official evaluation results of our system on the evaluation set compared with the baseline and other systems. Our system achieved first place in 5 out of 8 metrics and second place in one metric. It is worth noting that our system showed a substantial performance advantage over all other systems in the system-level SRCC and KTAU metrics.

V. Experimental Results

A. Preliminary Study on Losses

We conducted a preliminary study on SSL-MOS using wav2vec 2.0 to probe the effect of different losses before selecting the loss

TABLE IV
ABLATION STUDY ON MODEL COMPONENTS ON THE EVALUATION SET. **BOLD**: THE BEST RESULTS, ✓ AND ✗ : RETAINED AND REMOVED COMPONENTS.

SSL	SR Emb	Mel	Multi-CNN	MFCC	Cross-Attn	BLSTM	Utterance-level MSE	LCC	SRCC	KTAU	System-level MSE	LCC	SRCC	KTAU
✓	✓	✓	✓	✓	✓	✓	0.537	0.761	0.666	0.477	0.247	0.982	0.956	0.842
✓	✓	✓	✓	✗	✗	✓	0.360	0.821	0.720	0.533	0.187	0.975	0.919	0.779
✗	✓	✓	✓	✓	✓	✓	0.632	0.749	0.648	0.458	0.420	0.967	0.892	0.726
✓	✗	✓	✓	✓	✓	✓	0.450	0.806	0.717	0.524	0.223	**0.985**	0.955	0.842
✓	✓	✗	✗	✓	✓	✓	0.377	0.762	0.675	0.484	0.107	0.975	0.938	0.800
✓	✓	✓	✗	✓	✓	✓	0.541	0.772	0.662	0.470	0.322	0.967	0.902	0.737
✓	✓	✓	✓	✗	✓	✓	**0.276**	**0.837**	**0.724**	**0.540**	**0.063**	0.973	0.943	0.811
✓	✓	✓	✓	✓	✗	✓	0.601	0.798	0.704	0.513	0.352	0.975	0.949	0.832
✓	✓	✓	✓	✓	✓	✗	0.321	0.801	0.689	0.502	0.084	**0.985**	**0.981**	**0.905**

TABLE V
COMPARISON ACROSS DIFFERENT ENSEMBLE CONFIGURATIONS, **BOLD**: THE BEST RESULTS.

Ensemble	Development MSE	LCC	SRCC	KTAU	Evaluation MSE	LCC	SRCC	KTAU
Setting 1	0.343	0.832	0.717	0.531	0.129	0.978	0.950	0.832
Setting 2	0.279	0.835	0.712	0.521	0.133	0.967	0.922	0.768
Setting 3	0.333	0.840	0.726	0.541	0.144	0.980	0.949	0.821
Setting 4	0.372	0.822	0.704	0.518	0.204	0.957	0.893	0.737
HighRateMOS	**0.303**	**0.847**	**0.742**	**0.556**	**0.116**	**0.982**	**0.955**	**0.842**

in our final system. Table II shows that different loss functions yield varied performance trade-offs across metrics. On the development set, **contrastive loss** delivers the best LCC, and the second-best SRCC, KTAU, and MSE. Given that the primary challenge objective is to maximize system-level SRCC, the **contrastive loss** emerges as the most suitable choice. By comparison, training with an LCC loss yields the highest SRCC but suffers from a very large MSE, indicating poor absolute calibration.

When development labels are available, LCC loss achieves the highest system SRCC and KTAU on the evaluation set, making it the best choice for pure ranking. Contrastive loss remains a robust middle ground, achieving the highest LCC (0.982) and competitive SRCC (0.949), and thus is a strong alternative when both ranking and calibration matter.

B. Preliminary Study on SSL Upstream

To identify the best SSL encoder for robust MOS prediction across varied sampling rates, we compared several SSL encoders using contrastive loss. In Table III, the W2V2 achieves an exceptional development-set ranking capability, with the highest SRCC and the second-highest LCC. This led us to select W2V2 as our upstream model. However, our contrastive-trained W2V2 on the evaluation set does not outperform all other models. W2V2 L achieves the highest SRCC, while XLSR attains the highest LCC. This may reflect the distribution shift between training and development/evaluation data.

C. Ablation Study

We retrained the proposed model and conducted a comprehensive ablation by removing each of the seven core components (SSL features, sampling rate embedding, Mel spectrogram features, multi-scale CNN, MFCC features, cross-attention, and BLSTM) one by one. Table IV shows the results on the evaluation set. SSL features were proven to be the most critical component; removing them drops the system-level SRCC significantly from 0.956 to 0.892, indicating that the pretrained SSL representations provide core acoustic information for cross-sampling rate speech quality prediction. The joint removal of Mel spectrum maps and multi-scale CNNs also led to a significant decrease in system-level SRCC, proving that spectral features are very important for predicting the quality of audio with different sampling rates. In contrast, the removal of MFCC features, cross-attention mechanisms, and sampling rate embeddings had a relatively small impact on performance.

Removing BLSTM actually improved system performance, a counterintuitive result that suggests that simple FC layers may be more effective than complex sequence modeling in speech quality assessment tasks. We believe this happens because the BLSTM tended to overfit on our relatively small training set, whereas the FC layers remained more robust.

D. Ensemble Strategy

In Table V, we explored four ensemble configurations combining our three models. The "5-fold" refers to the evaluation scores of the five cross-fold validation models trained during the training phase. **Setting 1**: Model 1 (5-fold average) + Models 2 and 3 (standard training). **Setting 2**: All models use 5-fold averaging. **Setting 3**: All models use standard training. **Setting 4**: All models use the best among 5-fold. **HighRateMOS**: Model 1 (best among 5-fold) + Models 2 and 3 (standard training). HighRateMOS significantly outperforms other configurations and achieves the highest evaluation performance. Since Model 1 shows significant performance variation across its cross-validation folds, choosing its single best fold captures the configuration most aware of sampling-rate differences. This validates that our hybrid integration method is very effective for speech quality prediction across different sampling rates.

VI. CONCLUSION

HighRateMOS tackles the distribution shift that plagues non-intrusive speech-quality prediction when signals arrive at different sampling rates by unifying three key ideas: explicit sampling-rate embeddings, multi-scale feature fusion, and ensemble averaging. It encodes 16 kHz, 24 kHz, and 48 kHz as learnable vectors, combines wav2vec 2.0's self-supervised acoustic representations with Mel and MFCC spectral details through cross-attention and BLSTM aggregation, and then averages the outputs of complementary model variants. This design preserves the synergy between high-level acoustic and fine-grained spectral cues while sharply reducing cross-bandwidth prediction drift. On AudioMOS 2025 Track 3, HighRateMOS delivers a system-level SRCC of 0.955 and KTAU of 0.842, significantly exceeding the baseline and other systems. Although its performance is still limited by the size and linguistic diversity of the training data, this work demonstrates that explicit sampling-rate modeling is a practical approach to bandwidth-agnostic, robust speech-quality assessment, setting the stage for future advances such as continuous-rate encoding and multilingual expansion.

REFERENCES

[1] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," 2020. [Online]. Available: https://arxiv.org/abs/1910.11480

[2] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[3] 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*, 2018, pp. 1873–1877.

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

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

[6] 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 Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[7] 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*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] 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 Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8442–8446.

[11] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "CDPAM: Contrastive learning for perceptual audio similarity," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 196–200.

[12] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 1220–1230.

[13] D. Yuan and L. Wang, "Dual-criterion quality loss for blind image quality assessment," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 7823–7832. [Online]. Available: https://doi.org/10.1145/3664647.3681250

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

[15] W.-C. Huang, E. Cooper, and T. Toda, "Mos-bench: Benchmarking generalization abilities of subjective speech quality assessment models," 2024. [Online]. Available: https://arxiv.org/abs/2411.03715

[16] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[17] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 1298–1312. [Online]. Available: https://proceedings.mlr.press/v162/baevski22a.html

[18] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[20] 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," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 803–810.