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

***Abstract*—Modern speech quality prediction models are trained on audio data resampled to a specific sampling rate. When tested on audio with a higher sampling rate, these models can produce biased scores. We present 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 Track 3, HighRateMOS ranked first in five of eight metrics. Our experiments confirm that modeling sampling rate leads to more robust and sampling-rate-agnostic speech quality predictions.**

***Index Terms*—Speech quality assessment, MOS, Sampling rate**

I. Introduction

Speech quality is critical for modern communication and multimedia systems, as degraded quality leads to decreased 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 should 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 at a single sampling rate. For example, Quality-Net [3], SSL-MOS [4], MOSNet [5], MOSA-Net [6], UTMOS [7], and DNSMOS [8] were all trained on speech uniformly downsampled to 16 kHz. This raises 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 for predicting speech quality at different sampling rates. Each model takes a sampling rate embedding, self-supervised learning (SSL) model representations, and Mel-spectrogram 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 of each utterance (16 kHz, 24 kHz, or 48 kHz) into a learned embedding vector, which allows the model to adapt to the sampling rate of the input signal. To capture a wide range of acoustic information, we use SSL embeddings, Mel-spectrograms computed with a window size of 1024 and 64 mel-bands, and MFCCs. Pretrained SSL models, such as wav2vec 2.0 [9], HuBERT [10], and WavLM [11], are trained on raw waveforms sampled at 16 kHz, and their convolutional front-ends also assume this sampling rate. In previous studies, audio recorded at 24 kHz or 48 kHz was first downsampled to 16 kHz. 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 at slower speech rates.

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 both narrowband artifacts and 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 output 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 builds on Model 1 by introducing a cross-attention mechanism that selectively focuses on relevant acoustic patterns from the self-supervised features and spectral representations. Model 3 integrates all components. After combining multiple features, temporal aggregation is performed using a BLSTM, and the final MOS prediction is obtained through a fully connected layer. This design ensures that each model variant can capture different aspects of speech quality.

C. Loss Function

We explored 8 losses, grouped into 4 categories: point-wise regression loss, ranking-based loss, correlation-based loss, and hybrid loss.

1) Point-wise regression loss: We used Mean Squared Error (MSE) [3] and Mean Absolute Error (MAE) [12] to minimize the average squared/absolute difference.

2) Ranking-based loss: Ranking-based loss encourages the model to preserve the relative ordering of samples by MOS. We used a contrastive loss [13], which encourages the model not only to predict the correct MOS values but also to match the relative difference between two utterances. We also used the Relative Ranking Loss (RelRank), a triplet loss among extreme values in a batch [14].

3) Correlation-based loss: Correlation-based loss directly optimizes the statistical agreement between the predictions and human ratings. We explored Linear Correlation Coefficient (LCC) loss (defined as one minus the Pearson linear correlation between the

[†]Equal Contribution to this work.
EC was supported by JSPS KAKENHI Grant Number JP25K00143.

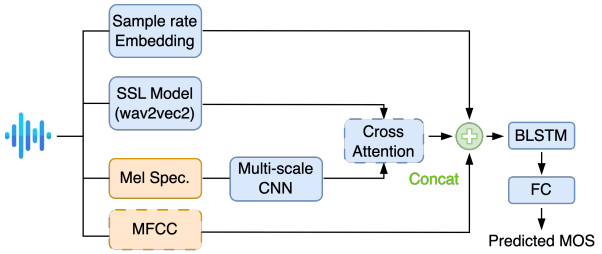


Fig. 1. Proposed model architecture. Three models are trained. Model 1 does not include the cross-attention module and MFCC. Model 2 adds the cross-attention module to Model 1. Model 3 adds MFCC features to Model 2.

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

4) Hybrid loss: Hybrid loss aims to leverage the complementary strengths of different objectives. We explored Dual Criterion Quality (DCQ) loss [15], which jointly enforces accurate numerical estimation and correct ordinal ranking, and UTMOS loss [7], which is the sum of clipped MSE loss and $0.5\times$ contrastive loss. We used the same clipping parameter of 0.25 as in [7].

III. Experimental Setup

A. Dataset

We used 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 at the three sampling rates were interleaved within a listening session. The evaluation set followed the same mixed-rate protocol as the development set, requiring listeners to rate stimuli at 16 kHz, 24 kHz, and 48 kHz in a session.

During the training phase, the labels for the development set were unavailable, and we could only obtain the model performance by uploading predictions for the development set to the challenge platform, which was inconvenient. Therefore, we decided to perform 5-fold cross-validation on the training data to determine the most appropriate loss function, SSL feature extractor, and model architecture. During the evaluation phase, we followed the standard training protocol, training on the full training set, selecting the best model on the development set (the labels were available in the evaluation phase), and submitting its predictions for the evaluation set. After the evaluation phase, the labels for the evaluation set were also released; therefore, we can report the results on both the development and evaluation sets in this paper.

B. Models

We implemented our models using the SHEET[1] [16] toolkit and SSL feature extractors from the S3PRL[2] [17], [18] toolkit. The SSL models explored include wav2vec 2.0 Base (W2V2), Large (W2V2 L) [9]; WavLM Base+ (WavLM B), Large (WavLM L) [11]; Data2vec Large (D2V) [19], and XLSR-1B (XLSR) [20]. We used AdamW [21] optimizer with a learning rate of 1e-3, and trained the models until system-level SRCC convergence (stop-growth for 2000 steps). The batch size was 8.

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

C. Evaluation Metrics

We employed a set of metrics at the utterance and system levels, including MSE to measure prediction accuracy, LCC to evaluate linear relationships, SRCC to capture monotonic associations, and KTAU to evaluate the ordinal agreement between predicted and ground-truth scores [22].

TABLE I
Official evaluation results for all submitted systems in AudioMOS Track 3. **Bold**: best result for each metric.

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
System-level performance of different losses. **Bold**: best result for each metric; underline: second-best result for each metric.

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. **Bold**: best result for each metric; underline: second-best result for each metric.

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 ranks first in 5 of 8 metrics and second in one metric. Notably, our system demonstrates substantial performance advantages in the system-level SRCC and KTAU metrics, surpassing all other systems.

V. Experimental Results

A. Preliminary Study on Losses

During the training phase, we first performed cross-validation on the training set using the SSL-MOS model with W2V2 features to explore the effects of different loss functions. We found that different loss functions exhibit varying trade-offs across metrics. The

TABLE IV
ABLATION STUDY ON MODEL COMPONENTS ON THE EVALUATION SET. **BOLD**: BEST RESULT FOR EACH METRIC. ✓ AND ✗ : RETAINED AND REMOVED.

SSL	SR Emb	Mel	Multi-CNN	MFCC	Cross-Attn	BLSTM	Utterance-level MSE	Utterance-level LCC	Utterance-level SRCC	Utterance-level KTAU	System-level MSE	System-level LCC	System-level SRCC	System-level 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
SYSTEM-LEVEL PERFORMANCE OF DIFFERENT ENSEMBLE CONFIGURATIONS. **BOLD**: BEST RESULT FOR EACH METRIC.

Ensemble	Development MSE	Development LCC	Development SRCC	Development KTAU	Evaluation MSE	Evaluation LCC	Evaluation SRCC	Evaluation 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**

contrastive loss achieved the best LCC, and the second-best SRCC, KTAU, and MSE. Given that the primary goal of this challenge is to maximize system-level SRCC, the **contrastive loss** is undoubtedly the most appropriate choice. In contrast, the LCC loss achieved the highest SRCC and KTAU but suffered from a very large MSE, indicating poor absolute calibration.

Table II shows the model's performance on the development and evaluation sets. We can see that the LCC loss achieves the highest system-level SRCC and KTAU on the evaluation set, making it the best choice for pure ranking. The contrastive loss remains a strong middle ground, achieving the highest LCC (0.982) and competitive SRCC (0.949), making it a strong alternative when both ranking and calibration are important.

B. Preliminary Study on SSL Upstreams

During the training phase, we also performed cross-validation on the training set using the SSL-MOS model trained with the contrastive loss to explore the effects of different SSL encoders. We found that W2V2 performed best and decided to use it in our final model.

Table III shows SSL-MOS's performance on the development and evaluation sets. We can see that W2V2 performed well on the development set, achieving the highest SRCC and the second-highest LCC. However, W2V2 did not outperform other models on the evaluation set. W2V2 L achieved the highest SRCC, while XLSR attained the highest LCC. This may reflect a distribution shift between the training and the development/evaluation data.

C. Ablation Study

We conducted comprehensive ablations 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 proved to be the most critical component; removing them resulted in a significant decrease in system-level SRCC from 0.956 to 0.892, indicating that the pretrained SSL representations provide core acoustic information for cross-sampling rate speech quality prediction. Jointly removing the Mel-spectrogram and multi-scale CNN also led to a significant decrease in system-level SRCC, proving the importance of spectral features for predicting audio quality at different sampling rates. In comparison, removing MFCC features, cross-attention mechanism, and sampling rate embedding had relatively little impact on performance.

Removing the BLSTM actually improved system performance, a counterintuitive result suggesting that simple FC layers may be more effective than complex sequence modeling for speech quality assessment tasks. We believe that this happens because the BLSTM tends to overfit on the relatively small training set, while the FC layers remain more robust.

D. Ensemble Strategy

In Table V, we explore four ensemble configurations combining our three models. **Setting 1**: Model 1 (5-fold average) + Models 2 and 3 (standard training). **Setting 2**: All models use the 5-fold average. **Setting 3**: All models use standard training. **Setting 4**: All models use the best of the 5-fold models. **HighRateMOS**: Model 1 (5-fold best) + Models 2 and 3 (standard training). "5-fold" refers to the five 5-fold validation models trained during the training phase (see Section III-A). HighRateMOS significantly outperforms the 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 makes the configuration more aware of sampling-rate differences. This validates the effectiveness of our hybrid integration method for speech quality prediction at 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 embedding, multi-scale feature fusion, and ensemble averaging. It encodes 16 kHz, 24 kHz, and 48 kHz into learnable vectors, combines self-supervised acoustic representations of wav2vec 2.0 with Mel and MFCC spectral details via 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 achieved a system-level SRCC of 0.955 and a KTAU of 0.842, significantly outperformed 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 extensions.

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.-w. 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 *Proceedings of Interspeech*, 2018, pp. 1873–1877.

[4] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proceedings of the 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 *Proceedings of Interspeech*, 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] 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.

[8] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 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

[10] 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.

[11] 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.

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

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

[14] 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.

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

[16] 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

[17] S.-w. 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.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proceedings of Interspeech*, 2021, pp. 1194–1198.

[18] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2884–2899, 2024.

[19] 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

[20] 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 *Proceedings of Interspeech*, 2022, pp. 2278–2282.

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

[22] 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 *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 803–810.