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

FUSE-MOS: Fusion of Speech Embeddings for MOS Prediction with Uncertainty Quantification

Enjamamul Hoq, Nikhil Gupta, Danielle Omondi, Ifeoma Nwogu

State University of New York at Buffalo, USA

ehoq@buffalo.edu, ngupta22@buffalo.edu, drakosua@buffalo.edu,inwogu@buffalo.edu

Abstract

The rapid advancements in text-to-speech (TTS) and voice conversion (VC) technologies necessitate evaluating the quality of synthesized speech. In this paper, we propose a novel network, FUSE-MOS, which combines the learned latent representations from raw audio waveforms and their corresponding Log-Mel spectrograms, to estimate the posterior distribution of Mean Opinion Score (MOS). Our method thus learns a broader and more nuanced representation of the speech signal. At inference, it predicts MOS value (point estimate) and also provides a measure of uncertainty of that prediction. By leveraging the combined latent representation, FUSE-MOS achieves significant improvements in performance metrics when compared to other existing approaches on benchmark datasets. We also explore an intelligent form of uncertainty filtering strategy to filter out low-confidence (high-uncertainty) samples. It shows FUSE-MOS's capability to maintain strong performance even with reduced data.

Index Terms: speech quality assessment, MOS prediction, uncertainty estimation, whisper, mean opinion score.

1. Introduction

Evaluating the quality of synthesized speech is a major task in the text-to-speech (TTS) synthesis and Voice Conversion (VC) field. The most often used and conventional approach to evaluate synthetic speech quality is the Mean Opinion Score (MOS). Human listeners evaluate voice sample quality by using a five-point rating system where 1 and 5 represent "bad" and "excellent," respectively. This evaluation considers several factors like vocal coherence, tone, pronunciation, naturalness, and noise. This evaluation is time-consuming and resource-intensive. Thus, there is a need for a robust deep learning-based MOS evaluation model.

1.1. Related Work

With the advances in deep neural network(DNN) techniques, many MOS prediction models have been developed to learn the mapping between speech and its quality score. Early works have moved away from techniques like Gaussian mixture models [1], support vector regression [2] to more advanced techniques developed by AutoMOS[3] and QualityNet[4]. AutoMOS uses a long-short-time-memory (LSTM) network and QualityNet uses a bi-directional LSTM(Bi-LSTM) network for end-to-end training to predict Perceptual Evaluation of Speech Quality (PESQ) [5]. MOSA-Net combined cross-domain features like Mel-spectrogram using CNN/Bi-LSTM model and self-supervised learning (SSL) features to improve MOS prediction [6]. Recent studies include MOSNet uses a CNN-BLSTM architecture and combines frame-level and utterance-level loss functions to improve accuracy [7]. MOSNet overlooks individual biases and this drawback is further improved by MBNet[8] and LDNet[9] by conditioning the model on human listeners during training.

In recent years, the self-supervised learning (SSL) model has gained attention as a MOS prediction model and has proven very effective in generalizing to out-of-domain data[10]. DDOS implements a domain-adaptive pre-training strategy to further pre-train wav2vec 2.0 features to minimize domain mismatch in SSL models [11]. The VoiceMOS Challenge 2022 [12] paved the way to promote research on automatic MOS prediction models. This challenge showed the effectiveness of fine-tuning SSL models for the MOS prediction task. Winning teams proposed ensemble techniques [13, 14] and multi-task learning [15] strategy to enhance the MOS prediction performance. Other works like [16] have demonstrated the success of ensembling of seven different models including wav2vec 2.0 [17], LDNet[9], CNN-RNN based architecture and QuartzNet [18]. Ravuri *et al.*[19] shown that uncertainty measures derived from pre-trained SSL models like wav2vec [20] correlate with MOS scores and have been shown effective for zero-shot settings. In VoiceMOS Challenge 2023, MOSA-Net+ uses a multi-objective speech assessment system that fuses cross-domain features to capture both subjective quality and intelligibility scores[21]. DeePMOS addresses the inherent noise and limited data in MOS training datasets and predicts the posterior distribution of MOS [22]. DNSMOS Pro [23] introduces a lightweight model built on a Gaussian likelihood framework and predicts the posterior distribution of MOS.

Although previous studies have greatly improved automatic MOS prediction, there are limited approaches that integrate time-domain and frequency-domain features. Furthermore, few methods provide uncertainty estimation strategies to enhance model performance.

1.2. Contribution

To address existing gaps and leverage the strengths of current research, we propose a novel fusion-based architecture that learns from various acoustic representations and quantifies the confidence of its predictions. We call our work FUSE-MOS. It combines both raw waveform and frame-level spectrogram features, to learn a broader and more nuanced representation of speech. Unlike other approaches, our proposed model aims to estimate the posterior distribution of MOS thus resulting in the ability to estimate the confidence level of each prediction made by the model. We also present an intelligent form of "cleaning our data", by eliminating low-confidence predictions from the training set. In our experiment, as much as 10% of the

10.21437/Interspeech.2025-2532

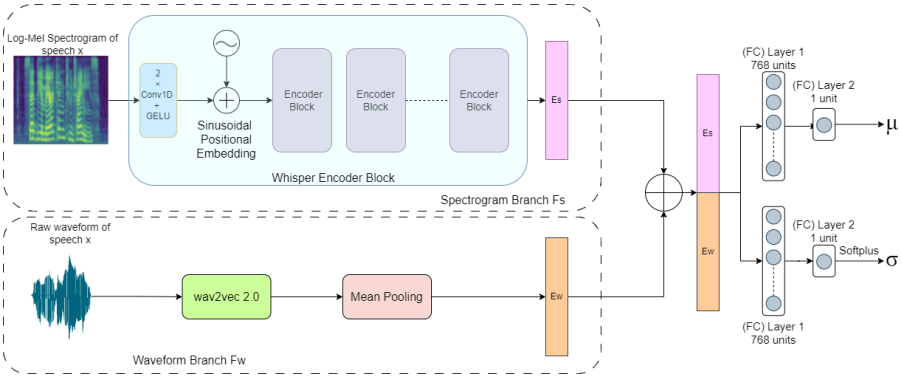

Figure 1: *FUSE-MOS architecture.*

data was removed to investigate whether this reduced dataset improves model performance. When the model was retrained on the reduced dataset, we found that FUSE-MOS maintain robust performance even after discarding these data points. We also introduce novel training features for MOS by utilizing the pre-trained, large-scale, weakly supervised model (for speech recognition) called Whisper [24].

2. Methodology

2.1. Framing the Problem

Building on the dataset creation framework from [8], consider a MOS dataset D consisting of N speech samples. Each sample x_i is evaluated by a group of m random listeners $\{l_{i1}, \ldots, l_{im}\}$, providing a set of judge scores $\{j_i^1, \ldots, j_i^m\}$. The average score for each sample, its MOS value, is denoted as $\bar{j}_i$. It is important to recognize that the same listener may rate multiple samples within the dataset. The dataset D comprises M total listeners, with M typically being much larger than m due to budget constraints during data collection.

MOS prediction can be treated as a regression task, but in this work, we introduce the use of a full regression architecture, also capable of providing the measure of uncertainty for each prediction. Such an architecture is composed of three components: (i) **the feature extraction network**, (ii) **the regressor network** and (iii) a **prediction layer** which consists of two branches, one predicting the mean output value and the other, the standard deviation, which when passed through the `Softplus` function (given by $f(x) = \ln(1 + e^x)$) converts it to a probability distribution.

2.2. Network Architecture

2.2.1. The Feature Extraction Network

Our proposed feature extraction network is presented in the leftmost part of the FUSE-MOS[1] architecture, as shown in Fig. 1. It comprises a spectrogram branch(F_s) which uses the Whisper encoder [24] and a waveform branch(F_w) which relies on raw waveform. Whisper encoder takes Log-Mel spectrogram feature as an input which captures harmonic structure and overall frequency content of speech signals. We use the small-size Whisper model to extract high-quality 768-dimensional audio embeddings for MOS prediction. To use Whisper, the speech signal is preprocessed by zero-padding it to the 30s duration to match the input dimension of the Whisper encoder E_s. During training, the whisper encoder module is kept frozen. In the waveform branch(F_w), we use the Fairseq[2] wav2vec 2.0 base model which is pre-trained on LibriSpeech to extract latent representations from the raw audio waveform that preserve fine-grained temporal information. Mean-pooling is used on wav2vec 2.0 to get 768-dimensional output embeddings. All audio files were downsampled to 16kHz before passing through the waveform branch.

2.2.2. The Regression Network

To explore the complementary nature of features extracted from Log-Mel spectrograms and raw waveforms, as described in Section 2.1, we employ the use of a deep neural network (DNN) model as the regression function f_θ. This model takes both Log-Mel spectrogram features x_s and raw waveform features x_w as inputs. The output of the DNN regression function $f_\theta(x_s, x_w)$ is a posterior distribution $p_\phi(y \mid x)$, where y represents the MOS, x is the set of extracted speech features, and ϕ are the parameters of this distribution.

Assuming that the posterior follows a Gaussian distribution, it can be fully described by its mean $\mu(x)$ and variance $\sigma^2(x)$. Therefore, the regression function f_θ outputs the parameters of this distribution as follows: $p_\phi(y \mid x) = \mathcal{N}(y; \mu(x), \sigma^2(x))$, where $\phi(x) = \{\mu(x), \sigma^2(x)\} = f_\theta(x_s, x_w)$.

Our goal is to learn a mapping from the extracted features (x_s, x_w) to the mean and variance of the MOS distribution. By utilizing this mapping, we can determine the uncertainty associated with our model's predictions. Since our target distributions are Gaussian, estimating the distribution for each target variable enables us to generate confidence values for the MOS. Achieving these mapping results ensures that the model provides not only accurate predictions but also reliable measures of uncertainty.

We then perform channel-wise concatenation of the outputs from the spectrogram branch and waveform branch each producing 768- dimensional vector (mean-pooled over time) and feed the concatenated output through two fully connected layers to predict the mean μ_x. To get the standard deviation σ_x, the concatenated output is passed through another set of two fully connected layers and a Softplus function to generate a valid probability distribution.

[1] `https://github.com/enjamamulhoq/FUSE-MOS`

[2] `https://github.com/pytorch/fairseq`

Table 1: *Comparison of FUSE-MOS with other methods using VCC2018 and BVCC datasets. The best values are in boldface.*

Model	VCC2018						BVCC					
	Utterance level			System level			Utterance level			System level		
	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑
Other's Models:												
MOSNet[25][26]	0.538	0.642	0.589	0.084	0.957	0.888	-	-	-	-	-	-
MBNet[8]	0.426	0.680	0.647	0.029	0.977	0.949	-	-	-	-	-	-
LDNet[9]	0.479	0.648	0.613	0.021	0.983	0.979	0.333	0.795	0.794	0.169	0.885	0.886
SSL-MOS[10]	0.453	0.724	0.695	0.065	0.992	0.965	0.212	0.884	0.882	0.139	0.941	0.941
UTMOS [13]	-	-	-	-	-	-	**0.165**	**0.899**	**0.896**	0.090	0.936	0.936
DeePMOS[22]	0.497	0.662	0.628	0.055	0.981	0.963	0.361	0.782	0.774	0.187	0.872	0.866
DNSMOS Pro[23]	0.441	0.677	0.641	-	-	-	0.338	0.787	0.783	-	-	-
Our Models:												
SSL-MOS+NLL	0.389	0.723	0.693	**0.017**	0.986	0.930	0.212	0.879	0.875	0.122	0.945	0.941
LDNet+SSL-MOS	0.450	0.661	0.621	0.028	0.989	0.969	0.328	0.788	0.785	0.188	0.863	0.862
LDNet+SSL-MOS+NLL	0.440	0.673	0.636	0.022	0.988	**0.983**	0.429	0.764	0.770	0.273	0.850	0.856
FUSE-MOS (with NLL)	**0.386**	**0.729**	**0.698**	0.019	**0.993**	0.963	0.191	0.887	0.887	**0.086**	**0.946**	**0.947**

2.3. Model Training and Inference

This DNN is trained to output the parameters of the Gaussian distribution that best describes the MOS for each speech sample, using a negative log-likelihood(NLL) loss function based on the principles of maximum likelihood estimation (MLE). We optimize this loss function to observe MOS values under the predicted distribution. Specifically, for each speech sample x_i with its corresponding MOS score y_i, the DNN predicts the mean $\mu(x_i)$ and standard deviation $\sigma(x_i)$. Given the assumption that the MOS scores follow a Gaussian distribution, the likelihood of observing y_i is:

$$\mathcal{L} = -\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_{x_i}^2}} \exp\left(-\frac{(y_i-\mu_{x_i})^2}{2\sigma_{x_i}^2}\right)\right) \quad (1)$$

Simplifying, this loss function can be rewritten as:

$$\mathcal{L} = \arg\min_{\theta} \frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i-\mu_{x_i})^2}{\sigma_{x_i}^2} + \log(2\pi) + 2\log(\sigma_{x_i})\right] \quad (2)$$

By training the DNN using this loss function, the model effectively learns a mapping from the input speech features x to the MOS distribution parameters $\phi(x) = \{\mu(x), \sigma^2(x)\}$.

At inference, for a given input speech signal's waveform feature x_{w_i} and its corresponding Log-Mel spectrogram feature x_{s_i}, our model predicts the Gaussian parameters of MOS y_i which are the mean $\mu(x_i)$ and standard deviation $\sigma(x_i)$; $\mu(x_i)$ represents the best single point estimate for the MOS and $\sigma(x_i)$ is the uncertainty associated with predicting $\mu(x_i)$.

3. Experimental Settings

3.1. Datasets

We used the VCC2018 [27], BVCC [28], and SOMOS-clean [29] datasets in our experiments. They contain audio samples generated by various text-to-speech (TTS) systems, where each sample was rated by different listeners on a 5-point scoring scale. We followed predefined training/validation/testing splits of 13580/3000/4000 for VCC2018, 4974/1066/1066 for BVCC, and 14100/3000/3000 for SOMOS.

Table 2: *Comparison on the SOMOS dataset*

Model	Utterance level				System level			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
DeePMOS[22]	0.267	0.504	0.475	–	0.060	0.805	0.791	–
DNSMOS Pro[23]	0.429	0.484	0.462	–	–	–	–	–
Content-Aware SSL[33]	0.203	**0.687**	**0.681**	**0.493**	0.052	**0.911**	**0.917**	**0.741**
Content-Aware SSL (BERT)[33]	0.179	0.668	0.659	0.475	0.021	0.906	0.913	0.736
FUSE-MOS (ours)	**0.175**	0.686	0.676	0.490	**0.020**	0.907	0.910	0.732

3.2. Training Details and Evaluation Metrics

We trained FUSE-MOS and baseline models across VCC2018, BVCC, and SOMOS datasets. The training was carried out on an NVIDIA GeForce RTX 3090 24GB RAM GPU. The models were trained for 1000 epochs with early stopping criteria. Each training roughly takes 6 hours. The batch size was set to 4 for training and 2 for validation, and the optimizer used was SGD with a learning rate of 0.0001 and a momentum of 0.9.

We used standard performance evaluation metrics, mean-squared-error (MSE), linear-correlation-coefficient (LCC)[30], and Spearman's rank-correlation-coefficient (SRCC)[31], and Kendall Tau Rank Correlation Coefficient (KTAU) [32] to assess our models at the utterance and system levels.

4. Tests, Results and Discussion

4.1. Quantitative Analysis

We present the results of comparing our proposed model with various state-of-the-art (SOTA) models in Table 1, using the VCC2018 and BVCC datasets at both the utterance and system levels. We reproduced the baseline models SSL-MOS (VCC2018, BVCC), DNSMOS Pro(SOMOS) and DeePMOS (BVCC, SOMOS) in our environment and obtained the LDNet, MBNet,MOSNet, UTMOS and DNSMOS Pro(BVCC,VCC2018) results as reported in their respective papers. We used the official implementations for SSL-

Table 3: *Ablation study results showing the impact of individual components on model performance for the BVCC dataset at utterance and system levels.*

Model	Utterance			System		
	MSE↓	LCC↑	SRCC↑	MSE↓	LCC↑	SRCC↑
FUSE-MOS	**0.199**	**0.883**	**0.882**	**0.088**	**0.940**	**0.942**
FUSE-MOS (uncertainty filtering)	0.214	0.871	0.871	0.108	0.934	0.936
only wav2vec2-branch	0.241	0.875	0.874	0.163	0.936	0.934
only wav2vec2-branch+NLL	0.266	0.872	0.872	0.200	0.935	0.936
only Whisper-branch	0.297	0.814	0.815	0.123	0.892	0.891
only Whisper-branch+NLL	0.298	0.814	0.815	0.128	0.892	0.890

MOS[3], DeePMOS[4], DNSMOS Pro [5] and LDNet[6] mean listener setup for our experiments. We provide the best score of our FUSE-MOS (with NLL) model and observe that it achieves stronger or comparable performance across key metrics. On the VCC2018 dataset, When compared against a similar fusion strategy (LDNet+SSL-MOS+NLL) and other baselines, FUSE-MOS outperforms them at the utterance level while retaining competitive correlations at the system level. On BVCC dataset, FUSE-MOS shows highly competitive performance when compared with UTMOS. While, UTMOS does a slightly better job at tracking individual utterance ratings, FUSE-MOS is more consistent at ranking entire systems. Table 2 presents the results of evaluating our proposed model using the most recent TTS-based dataset, SOMOS. We show that our proposed FUSE-MOS architecture consistently outperforms DeepMOS and DNSMOS Pro by a significant margin while matching or exceeding Content-Aware SSL model in key aspects. Content-Aware SSL shows strong results, particularly in system-level LCC and SRCC. It indicates that both models capture MOS trends effectively. These results show great generalizing capability and robustness of our FUSE-MOS model across several datasets.

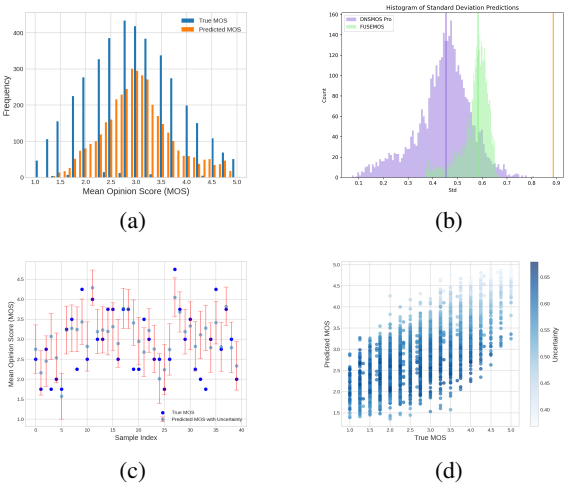

Figure 2: *Visualizations on the VCC2018 test dataset: (a) Utterance-Level Histogram Plot; (b)Histogram showing the standard deviation predictions from the posteriors against the prior standard deviation 0.888; (c) Predicted MOS with Uncertainty; (d) Uncertainty Scatter Plot.*

[3]https://github.com/nii-yamagishilab/mos-finetune-ssl

[4]https://github.com/Hope-Liang/DeePMOS

[5]https://github.com/fcumlin/DNSMOSPro

[6]https://github.com/unilight/LDNet

4.2. Qualitative Analysis

From Fig. 2a, we observe that the shape of the distribution of predicted MOS values very closely follows that of the true (or human-annotated) MOS values, thus validating the predicted scores to some degree. Because the true MOS values were integers $1, 2 \ldots, 5$ and the predicted MOS, from the regression model, were continuous values between $[1, 5]$, the resulting distributions could not perfectly align, but their shapes were similar. Fig. 2b shows that FUSE-MOS produces a wider standard deviation range yet outperforms DNSMOS Pro in MSE and correlation metrics. It suggests that FUSE-MOS's posterior distribution is more accurately calibrated, providing reliable MOS predictions. Fig. 2c shows the true and predicted MOS values, along with the estimated uncertainty bars, for 40 randomly selected samples. The plot suggests that when the true MOS score is greater than about 3.25, the true and predicted values are likely to be closer to each other and the uncertainty bars are smaller. In some instances, there is complete overlap (e.g. Samples #17, 18, 39). Lastly, Fig. 2d shows how uncertainties behave with increasing true and predicted values. We observe again, as before, that as the true MOS values go above 3.5, the volume of uncertainty reduces.

4.3. Ablation Studies

Table 3 reports a BVCC ablation study in which we systematically removed or added parts of FUSE-MOS. We trained the full FUSE-MOS, then wav2vec 2.0-only and Whisper-only variants, and finally added an NLL uncertainty loss to each single-branch model. In all cases, training was performed five times with different random seeds and we averaged the scores for each metric to account for performance variations. While each branch alone provides reasonable performance, Wav2Vec 2.0 branch contribute most to the FUSE-MOS model's overall performance. These experiments reinforce the effectiveness of FUSE-MOS as a fusion-based approach. Lastly, across five training runs, we filtered out around 10% of audio samples with high σ and retrained FUSE-MOS with only the 'good' samples. Although, performance didn't directly improve but it demonstrates our posterior-based system can handle some data reduction without significant loss.

5. Conclusion

In summary, we introduce a novel MOS prediction model that uses wav2vec 2.0 and Whisper encoder to extract latent representations from the raw waveform and log Mel-spectrogram inputs. By fusing the learned representations, it estimates both the mean and standard deviation of MOS scores. It sets SOTA results on VCC2018 and BVCC at utterance and system levels. It also gives a competitive performance on the newer SOMOS dataset, showing its potential for evaluating any neural TTS system. Our results show the potential of fusing cross-domain features (frequency and time-domain representations) alongside statistical uncertainty measures to improve automated MOS prediction.

6. Acknowledgements

This material is based upon work supported under the AI Research Institutes program by National Science Foundation Award # 2223507 and the Institute of Education Sciences, U.S. Department of Education, Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and

conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

7. References

[1] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, nonintrusive speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1948–1956, 2006.

[2] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1217–1232, 2011.

[3] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.

[4] 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," *arXiv preprint arXiv:1808.05344*, 2018.

[5] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[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, 2022.

[7] 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," *arXiv preprint arXiv:1904.08352*, 2019.

[8] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[9] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 896–900.

[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)*. IEEE, 2022, pp. 8442–8446.

[11] W.-C. Tseng, W.-T. Kao, and H.-y. Lee, "Ddos: A mos prediction framework utilizing domain adaptive pre-training and distribution of opinion scores," *arXiv preprint arXiv:2204.03219*, 2022.

[12] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2022," *arXiv preprint arXiv:2203.11389*, 2022.

[13] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[14] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino, and Y. Zhao, "Fusion of self-supervised learned models for mos prediction," *arXiv preprint arXiv:2204.04855*, 2022.

[15] X. Tian, K. Fu, S. Gao, Y. Gu, K. Wang, W. Li, and Z. Ma, "A transfer and multi-task learning based approach for mos prediction," *Interspeech 2022*, 2022.

[16] M. Kunešová, J. Matoušek, J. Lehečka, J. Švec, J. Michálek, D. Tihelka, M. Bulín, Z. Hanzlíček, and M. Řezáčková, "Ensemble of deep neural network models for mos prediction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[19] A. Ravuri, E. Cooper, and J. Yamagishi, "Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 580–584.

[20] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[21] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating whisper for robust speech assessment," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[22] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, "Deepmos: deep posterior mean-opinion-score of speech," in *Proceedings of INTERSPEECH*, 2023, pp. 526–530.

[23] F. Cumlin, X. Liang, V. Ungureanu, C. K. A. Reddy, C. Schüldt, and S. Chatterjee, "Dnsmos pro: A reduced-size dnn for probabilistic mos of speech," in *Interspeech 2024*, 2024, pp. 4818–4822.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[25] Y. Choi, Y. Jung, and H. Kim, "Deep mos predictor for synthetic speech using cluster-based modeling," *arXiv preprint arXiv:2008.03710*, 2020.

[26] ——, "Neural mos prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 462–469.

[27] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[28] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" *arXiv preprint arXiv:2105.02373*, 2021.

[29] 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," *arXiv preprint arXiv:2204.03040*, 2022.

[30] K. Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.

[31] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[32] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of measurement and statistics*, vol. 2, pp. 508–510, 2007.

[33] A. Vioni, G. Maniati, N. Ellinas, J. S. Sung, I. Hwang, A. Chalamandaris, and P. Tsiakoulis, "Investigating content-aware neural text-to-speech mos prediction using prosodic and linguistic features," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.