A Non-Intrusive Speech Quality Assessment Model using Whisper and Multi-Head Attention

Guojian Lin*, Yu Tsao† and Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 12432635@mail.sustech.edu.cn, fchen@sustech.edu.cn Tel/Fax: +86-755-88018554
† Academia Sinica, Research Center for Information Technology Innovation, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw Tel/Fax: +886-2-2787-2390

Abstract—**Speech quality assessment serves as an important tool for speech related applications. In this study, we propose a non-intrusive model QUAL-Net, which is able to estimate subjective quality scores of the target speech. QUAL-Net combines acoustic features extracted by a large-scale model Whisper with spectral features and time-domain waveform features. Furthermore, QUAL-Net employs a CNN-BiLSTM-Attention architecture and introduces multi-head attention mechanism into attention layer to enhance model's performance. Experimental results demonstrate that Whisper embedding features have more powerful speech quality characterization ability than other self-supervised learning (SSL) embedding features. Additionally, the feature combination utilizing all three types of acoustic features obtains optimal improvement in model performance. Furthermore, the results prove that multi-head attention has potential to capture more key information from acoustic features than multiplicative self-attention. We tested QUAL-Net's performance on the noisy and enhanced track of VoiceMOS Challenge 2023. Compared with MOSA-Net and other speech quality assessment models, QUAL-Net achieves significant improvement when it is trained to estimate subjective quality scores. QUAL-Net outperforms the top-ranked MOSA-Net+ in all evaluation metrics. QUAL-Net uses a simpler CNN architecture compared to the MOSA-Net+, contributing to reduction of the model complexity.**

I. Introduction

In real life, accurate speech quality assessment is of great significance to the development of speech-related applications, such as speech enhancement, speech synthesis and hearing aids. Listening test based on listeners is acknowledged as the most direct and accurate method to evaluate speech quality. The mean opinion score (MOS) is the most widely used evaluation metric of speech quality in subjective listening test, ranging from one to five. However, listening test is time-consuming and costly, and requires specific listening environments. Due to the limitation of subjective listening test, objective metrics for speech quality assessment have been proposed, such as perceptual evaluation of speech quality (PESQ) [1], perceptual objective listening quality analysis (POLQA) [2], signal-to-distortion ratio (SDR) [3] and hearing aid speech quality index (HASQI) [4]. But these metrics based on signal processing algorithm need clean speech with the same frequency and time length as reference.

In recent years, researchers have developed non-intrusive speech quality assessment models based on deep learning. By learning features from a large amount of speech data, these models are able to accurately predict speech quality scores without clean reference. Deep learning-based methods can be divided into two types on the target evaluation metrics. The first type is to predict objective evaluation metrics. Quality-Net [5] uses the network BiLSTM to predict scores of PESQ. STOI-Net [6] employs a CNN-BiLSTM architecture with attention to predict scores of objective speech intelligibility. AMSA [7] utilizes reference-less multi-task learning (MTL) framework to predict multiple objective speech quality and intelligibility scores. The second type focus on predicting scores from subjective listening test. MOSNet [8], a CNN-BiLSTM based model, is proposed to estimate quality of the converted speech. MBNet [9] uses two networks to seprately predict the mean quality score of an utterance and the difference between the mean score and listener score, respectively. LDNet [10] employs raw speech and listener-dependency information of listeners as input to the model for MOS prediction. In VoiceMOS Challenge 2022 [11], which aims to encourage the research on development of MOS predictor for synthesized speech, numerous innovative systems [12], [13], [14], [15] based on self-supervised learning (SSL) models have been proposed, yielding great improvement in the performance of MOS prediction.

Whisper [16], a large-scale pre-trained model, has demonstrated its advanced performance and powerful generalization ability in various speech processing tasks. Zezario et al. [17] utilizes Whisper to extract phonetic embedding representations to assess HASPI scores for hearing aids, contributing to an improvement of approximately 30% in the ranking correlation between predicted scores and actual scores compared to the baseline model using WavLM [18] embedding features. It is more difficult to obtain MOS from human listening test than from objective evaluation metrics, which results in the limited amount of labeled speech data for training. Therefore, it is necessary to consider the importance of each speech frame's information to ensure accurate evaluation. For example, there are quiet segments in speech frames that contain redundant information and should receive less attention. More attention should be paid to the frames that contain more useful information. However, self-attention mechanism overly focuses on its own location information when encoding features, ignoring the importance of other location information. In contrast, the multi-head attention mechanism maps feature information to multiple subspaces to compute location weights in parallel, enabling the model to focus on different subspace information at different locations, thus capturing more effective information. Liang et al. [19] utilizes multi-head attention mechanism to design a non-intrusive speech quality evaluation model for hearing aids. The value of the Pearson correlation coefficient (PCC), which describes the correlation between the predicted quality scores and the actual quality scores is improved from 0.943 to 0.985.

Based on CNN-BiLSTM with multiplicative self-attention, MOSA-Net [20] combines cross-domain features with embedding representations from SSL model to evaluate quality and intelligibility of the noisy speech. In VoiceMOS Challenge 2022, MOSA-Net has achieved performance close to that of the baseline system [11]. We tested the performance of MOSA-Net using noisy and enhanced speech dataset from track 3 of the recent VoiceMOS Challenge 2023 [21] and found its prediction accuracy can be further improved. Based on the architecture of MOSA-Net, this work proposes a novel model QUAL-Net for speech quality evaluation. SSL pre-trained models are replaced with Whisper to extract acoustic embedding features. QUAL-Net employs multi-head attention mechanism instead of multiplicative self-attention. QUAL-Net also utilizes a simpler CNN architecture with less convolution layers than MOSA-Net. In the experiments, we first compare the performance of Whisper-based features with other four SSL-based features. Then, we investigate the effect of different audio features combinations on the accuracy of quality prediction. Subsequently, we analyze the impact of multi-attention and compare it with other attention mechanism. Finally, We compare QUAL-Net's performance on noisy and enhanced speech dataset with other systems.

II. QUAL-Net

The overall architecture of the proposed QUAL-Net is presented in Fig. 1. QUAL-Net is composed of the feature extraction module and the quality prediction module. The specific parameters of the model are described in Table I. In the feature extraction module, given a speech waveform $X = [x_1, x_2, \ldots, x_n, \ldots, x_N]$, the model takes three input branches. In the first branch, X is converted by 512-point STFT(Short Time Fourier Transformation) with a Hamming window of 32 ms and a hop of 16 ms to obtain a 257-dimensional spectrogram. In the second branch, X is fed into SincNet [22], a convolution network based on "Sinc" function, with filter dimension of 257. The output of the SincNet is a 257-dimensional filtered time-domain waveform, namely LFB feature. Subsequently, spectral features and LFB features are fed into CNN. CNN has five convolution networks, with a two-dimensional convolution layer, a batch normalization layer, a ReLU activation function and power average pooling layer in each network. The pooling layers calculate the p-th root of the p-th power sum of X in the moving window. The calculation process is described as follows, where the value of p is 4:

$$f_{(x)} = \sqrt[p]{\sum_{x \in X} x^p} \tag{1}$$

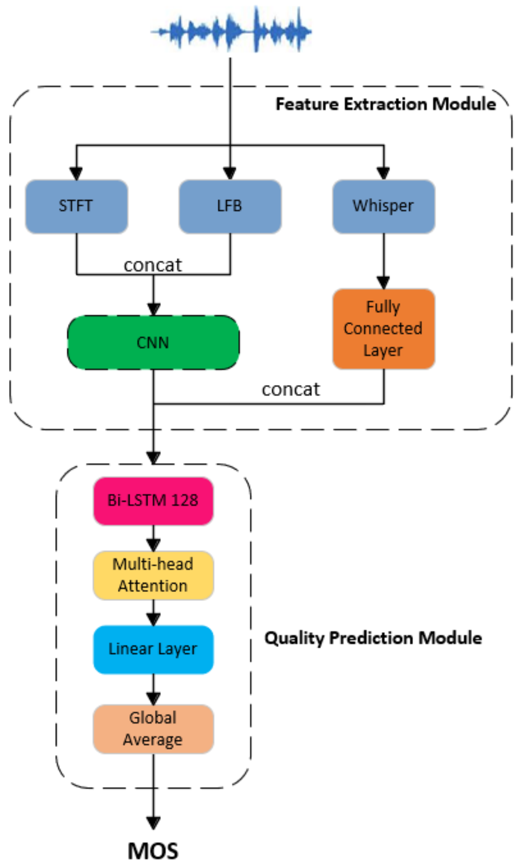

Fig. 1. Architecture of the proposed QUAL-Net

In the third branch, X is first transformed into a spectral signal by STFT, the frequency axis is divided into a series of Mel frequency bands, and then the energy within each Mel frequency band is summed and logarithmized to obtain the log mel-spectrogram M. Then, M is fed into Whisper pre-trained model to obtain 512-dimensional Whisper embedding features WE. A fully connected layer is utilized to ensure that WE have the same feature dimension as deep frame-level features and WE are concatenated with deep frame-level features extracted by CNN. The combined features are mapped into quality prediction module to predict quality scores.

Quality prediction module consists of a BiLSTM, a multi-head attention layer, a linear layer and an adaptive average pooling function. The input features are first processed by BiLSTM with 128 nodes. BiLSTM comprises a forward LSTM and a backward LSTM, which models temporal information of feature sequences using backward and forward propagation, contributing to effective leveraging context information of long time sequences. BiLSTM is utilized to process combined features frame-by-frame and capture contextual dependencies in acoustic features. Multi-head attention layer with 8 attention heads, is applied to learn different attention weights based on correlations within combined features, enabling the model to focus on more effective feature information to ensure accurate quality prediction. Then, a linear layer with one node is utilized

TABLE I
PARAMETERS OF THE QUAL-NET

Name	Layer	Parameter	Size of output
CNN	Conv1	32,3×3	32×N×257
	Batch Normaliztion	32	32×N×257
	ReLU	/	32×N×257
	LPPool2d	4,1×4	32×N×64
	Conv2	32,3×3	32×N×64
	Conv3	64,3×3	64×N×64
	Batch Normaliztion	64	64×N×64
	ReLU	/	64×N×64
	LPPool2d	4,1×4	64×N×16
	Conv4	64,3×3	64×N×16
	Conv5	128,3×3	128×N×16
	Batch Normaliztion	128	128×N×16
	ReLU	/	128×N×16
	LPPool2d	4,1×4	128×N×4
	Reshape	/	N×512
Score Prediction Module	BiLSTM	128	N×256
	Dense	128	N×128
	Multi-Head Attention	128,heads=8	N×128
	Dense	1	1×128
	Global Average	1	1×1

to generate frame-level scores. The number of frames of a speech utterance is the sum of frame numbers of the deep frame-level features and WE. The output of linear layer is processed through global average pooling to calculate the final quality score.

Moreover, the model integrates both frame-level quality scores and utterance-level quality scores into the loss function for training. The loss function is described as follows:

$$\mathcal{L}_{quality} = \frac{1}{N}\sum_{n=1}^{N}\left[\left(Q_n - \hat{Q}_n\right)^2 + \frac{\alpha_Q}{F_u}\sum_{l=1}^{F_u}\left(Q_n - \hat{q}_{nl}\right)^2\right] \tag{2}$$

where Q_n represents actual Quality scores of the n-th training utterance. $\hat{Q}_n$ represents the predicted Quality scores of the n-th training utterance. The total number of training utterances is denoted by N. F_u represents the total number of frames in the n-th training utterance, which is the frame number of combined feature. $\hat{q}_{nl}$ is the predicted frame-level scores of the l-th frame of the n-th training utterance. For the n-th training utterance, there are F_u predicted frame-level scores. The weights between utterance-level and frame-level losses are determined by α_Q.

III. EXPERIMENTS

A. Dataset

The dataset in this experiment is from the noisy and enhanced track of VoiceMOS Challenge 2023 [21]. Training data is based on TMHINT-QI [23] dataset, a Mandarin corpus containing 24,408 ten-word utterances. It contains totally 8201 samples, including 360 clean speech samples, 1874 noisy speech samples with four types of noises (babble, street, pink, and white) at four signal-to-noise ratio (SNR) levels (-2, 0, 2, and 5) and 5967 enhanced speech samples derived from five speech enhancement systems: KLT, MMSE, FCN, DDAE, and Transformer. 226 listeners were recruited to take the listening test and predicted speech quality scores on a range from 1 to 5. The mean of the subjective quality scores for each utterance was used as the actual score of the speech.

A seperate dataset TMHINT-QI (S) was created as test set, containing 360 noisy samples with the same noise type as those in training data and 1600 enhanced samples processed by five speech enhancement models: MMSE, FCN, Trans, DEMUCS, and CMGAN, including two unseen enhancement systems. A total of 110 listeners were recruited for listening test of the test set.

B. Evaluation Metrics

To evaluate the performance of the model, system-level and utterance-level mean squared error (MSE), Linear Correlation Coefficient (LCC) and Spearman Rank Correlation Coefficient (SRCC) are used. MSE indicates the difference between predicted scores and actual quality scores. LCC describes the linear correlation between predicted scores and actual scores. SRCC represents the rank correlation between predicted scores and actual scores. Since it is more useful for quality assessment models to predict the ranks of systems accurately than to predict actual quality scores, we use system-level SRCC as the primary evaluation metric for model performance.

C. Model Training

In the training phase of the model, we use Adam optimizer with initial learning rate of 0.001 and adopt a dynamic learning rate adjustment strategy. If validation loss does not decrease after 10 training iteration rounds, the learning rate of Adam optimizer decreased by a factor of 10 with the minimum learning rate of 0.000001 to help the model find the optimal parameters better as well as to avoid overfitting. Totally 50 epochs are trained with batch size of 1. 90% of training data are used for training and 10% for validation. All speech samples are in 16 kHz.

D. Comparison of Different Embedding Features

In the first experiment, we aim to compare different speech pre-trained models and select the optimal pre-trained model as feature extractor. Five pre-trained models, BEATs [24], Hubert [25], XLSR [26], WavLM, and Whisper are employed to extract speech embedding features, where the feature dimensions extracted by the Hubert, XLSR, WavLM, and Whisper models are 1024, while the feature dimension of the BEATs model is 768.

As presented in Table II, Hubert, WavLM, and Whisper embedding features have respectively high correlation with speech quality. The models using BEATs and XLSR embedding features achieve low prediction accuracy, particularly

TABLE II
PERFORMANCE OF QUAL-NET WITH DIFFERENT EMBEDDING FEATURES

Embedding Feature	System-level			Utterance-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
BEATs [24]	0.584	0.582	0.792	0.537	0.505	1.090
Hubert [25]	0.961	0.955	0.067	0.793	0.753	0.352
XLSR [26]	0.888	0.883	0.257	0.721	0.693	0.586
WavLM [18]	**0.964**	0.962	0.063	0.794	0.757	0.348
Whisper [16]	0.961	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

TABLE III
PERFORMANCE OF QUAL-NET WITH DIFFERENT FEATURES COMBINATIONS

Combination	Feature	System-level			Utterance-level		
		LCC	SRCC	MSE	LCC	SRCC	MSE
1	STFT+LFB	0.808	0.796	0.227	0.638	0.572	0.599
2	WE	0.926	0.921	0.106	0.786	0.752	0.374
3	STFT+WE	0.954	0.954	0.058	**0.808**	0.777	**0.323**
4	LFB+WE	0.950	0.949	0.066	0.803	0.773	0.327
5	STFT+LFB+WE	**0.961**	**0.966**	**0.053**	0.807	**0.780**	**0.323**

BEATs embedding feature shows significantly weaker correlation with speech quality than other features. The model with Whisper embedding features achieves the best performance in all evaluation metrics except for the system-level LCC, which is slightly lower than that of the WavLM. Based on training data amount of labeled speech of 680,000 hours and training parameters of large scale, Whisper reveals its robustness in audio feature extraction. Experimental results demonstrate the benefits of utilizing Whisper to deploy quality prediction model.

E. Comparison of Different Feature Combinations

In the second experiment, we investigate the effect of different acoustic features on the prediction performance through ablation experiment. As shown in Table III, five different feature combinations are employed, where STFT represents spectral features, LFB represents waveform features after SincNet filter processing and WE represents the embedding representations extracted by Whisper. Comparing combination 1, combination 2 and combination 5, the Whisper embedding features significantly outperform STFT+LFB features in both system-level and utterance-level evaluations, indicating that the Whisper embedding features play a major role in quality prediction. Comparing combination 3 and combination 4, with the same embedding features, model using spectral features has slightly better performance than that of using LFB features. It indicates that although LFB features retain the raw waveform more completely, spectral features which retain the speech phase information and the short-time transform features can obtain more useful phonetics information. The performances of combination 3 and combination 4 have been very close to that of the combination 5. It further demonstrates that compared with WE, spectral features and LFB features do not have a large magnitude of enhancement on prediction performance of the model.

Combination 5 achieves the most accurate quality prediction, demonstrating the effectiveness of combining different acoustic features for speech quality estimation.

F. Impact of Multi-Head Attention

In the third experiment, to investigate the impact of multi-head attention in the model, we compare QUAL-Net (*No-ATT*), which does not have attention layer with QUAL-Net (*Mul-ATT*) and QUAL-Net (*Multi-Head-ATT*). It needs to be clarified that Mul-ATT denotes that the model uses the self-attention layer which employs multiplicative attention and Multi-Head-ATT denotes that the model uses multi-head attention layer.

As present in Table IV, the models with attention layer achieve more accurate quality prediction than the model without that. Compared with QUAL-Net (*No-ATT*), QUAL-Net (*Multi-Head-ATT*) improves on system-level SRCC by 0.051. Additionally, QUAL-Net (*Multi-Head-ATT*) outperforms QUAL-Net (*Mul-ATT*) in all evaluation metrics, improving on system-level SRCC by 0.022. It confirms that multi-head attention enables the model to focus on more useful information in time frames and effectively capture contextual association of speech, thus enhancing model's performance. Multi-head attention also exhibits superior ability to process frame-level feature of speech over multiplicative self-attention.

G. Comparison with Other Systems

In this experiment, we compare QUAL-Net with other speech quality evaluation models on the noisy and enhanced speech dataset of VoiceMOS Challenge 2023. Table V exhibits the performance of different systems. MOS-SSL [12] uses finetuned SSL models to predict MOS and LE-SSL-MOS [27] constructs a MOS predictor based on SSL models and the scores of individual listener augmentation branches, introducing new unsupervised metrics to improve prediction accuracy. UTMOS [13] builds strong and weak learners based on SSL models and classical machine learning algorithms to predict MOS. It is noted that MOSA-Net [20] serves as the baseline model of our system, and its enhanced version MOSA-Net+ [28] achieves top-ranked performance on the noisy and enhanced track of VoiceMOS Challenge 2023.

TABLE IV
PERFORMANCE OF QUAL-NET WITH DIFFERENT ATTENTION LAYER

Model	System-level			Utterance-level		
	LCC	**SRCC**	**MSE**	**LCC**	**SRCC**	**MSE**
QUAL-Net (*No-ATT*)	0.920	0.915	0.243	0.771	0.742	0.521
QUAL-Net (*Mul-ATT*)	0.947	0.944	0.179	0.786	0.758	0.454
QUAL-Net (*Multi-Head-ATT*)	**0.961**	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

TABLE V
PERFORMANCE OF ALL SYSTEMS ON ENHANCED AND NOISY SPEECH TRACK OF VOICEMOS CHALLENGE 2023

System	System-level			Utterance-level		
	LCC	**SRCC**	**MSE**	**LCC**	**SRCC**	**MSE**
MOS-SSL [12]	0.637	0.487	2.986	0.518	0.403	3.356
UTMOS [13]	0.769	0.621	1.763	0.611	0.477	2.216
LE-SSL-MOS [27]	0.769	0.749	0.635	0.684	0.636	0.688
MOSA-Net [20]	0.954	0.941	0.067	0.781	0.749	0.358
MOSA-Net+ [28]	0.952	0.956	0.082	0.803	**0.780**	0.343
QUAL-Net	**0.961**	**0.966**	**0.053**	**0.807**	**0.780**	**0.323**

Our system QUAL-Net yields notable improvement in the performance of MOS prediction, outperforming all the compared systems including the top-ranked model MOSA-Net+. Compared with MOSA-Net, our system improves system-level SRCC from 0.941 to 0.966. In comparison to MOSA-Net+, our system improves system-level SRCC from 0.956 to 0.966. It demonstrates the advantages of creating multi-domain acoustic features using Whisper model and adopting multi-head attention for enhancing model's speech quality evaluation performance. It is worth noting that although UTMOS achieves top-ranked performance in VoiceMOS Challenge 2022, its performance on the noisy and enhanced dataset is much lower than that of QUAL-Net and MOSA-Net, which to some extent reflects the difficulty in realizing the quality assessment of multi-domain speech dataset.

IV. CONCLUSIONS

This study proposes a non-intrusive speech quality assessment model QUAL-Net. QUAL-Net incorporates pre-trained model Whisper into feature extraction and introduces multi-head attention mechanism to quality prediction. Experimental results demonstrate that Whisper embedding features can capture more effective acoustic information compared to other SSL-based features. Additionally, ablation experiment confirms that Whisper embedding features play a major role in speech quality evaluation. The model using spectral features, time-domain waveform features and Whisper features achieves the optimal performance. Furthermore, multi-head attention has superior performance over multiplicative self-attention and further improves prediction accuracy of the model. QUAL-Net achieves a notable improvement over MOSA-Net and outperforms MOSA-Net+ and other speech quality assessment models. Since QUAL-Net uses CNN architecture with less number of convolution layers than MOSA-Net+, it achieves better performance with simpler model structure. In future work, we plan to test the performance of QUAL-Net on different types of datasets and explore QUAL-Net's potential for semi-supervised speech quality assessment.

REFERENCES

[1] 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)*, IEEE, vol. 2, 2001, pp. 749–752.
[2] J. G. Beerends, C. Schmidmer, J. Berger, *et al.*, "Perceptual objective listening quality assessment (POLQA), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.
[3] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[4] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.
[5] 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.
[6] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2020, pp. 482–486.
[7] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 911–915.
[8] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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*, 2022, pp. 4521–4525.

[14] 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," in *Proc. Interspeech 2022*, 2022, pp. 4541–4545.

[15] Z. Yang, W. Zhou, C. Chu, *et al.*, "Fusion of self-supervised learned models for MOS prediction," in *Proc. Interspeech 2022*, 2022, pp. 5443–5447.

[16] 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.

[17] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Utilizing Whisper to enhance multi-branched speech intelligibility prediction model for hearing aids," *arXiv preprint arXiv:2309.09548*, 2023.

[18] S. Chen, C. Wang, Z. Chen, *et al.*, "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.

[19] R. Liang, Y. Xie, J. Cheng, C. Pang, and B. Schuller, "A non-invasive speech quality evaluation algorithm for hearing aids with multi-head self-attention and audiogram-based features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2166–2176, 2024.

[20] 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.

[21] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–7.

[22] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 1021–1028.

[23] Y.-W. Chen and Y. Tsao, "InQSS: A speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. Interspeech 2022*, 2022, pp. 3088–3092.

[24] S. Chen, Y. Wu, C. Wang, *et al.*, "BEATs: Audio pre-Training with acoustic tokenizers," in *International Conference on Machine Learning*, PMLR, 2023, pp. 5178–5193.

[25] 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.

[26] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[27] Z. Qi, X. Hu, W. Zhou, *et al.*, "LE-SSL-MOS: Self-supervised learning MOS prediction with listener enhancement," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–6.

[28] 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," *arXiv preprint arXiv:2309.12766*, 2023.