MULTI-TASK PSEUDO-LABEL LEARNING FOR NON-INTRUSIVE SPEECH QUALITY ASSESSMENT MODEL

Ryandhimas E. Zezario[1 2], *Bo-Ren Brian Bai*[3], *Chiou-Shann Fuh*[1], *Hsin-Min Wang*[2], *Yu Tsao*[2]

[1]National Taiwan University [2]Academia Sinica [3]Fortemedia

ABSTRACT

This study proposes a multi-task pseudo-label learning (MPL)-based non-intrusive speech quality assessment model called MTQ-Net. MPL consists of two stages: obtaining pseudo-label scores from a pretrained model and performing multi-task learning. The 3QUEST metrics, namely Speech-MOS (S-MOS), Noise-MOS (N-MOS), and General-MOS (G-MOS), are the assessment targets. The pretrained MOSA-Net model is utilized to estimate three pseudo labels: perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and speech distortion index (SDI). Multi-task learning is then employed to train MTQ-Net by combining a supervised loss (derived from the difference between the estimated score and the ground-truth label) and a semi-supervised loss (derived from the difference between the estimated score and the pseudo label), where the Huber loss is employed as the loss function. Experimental results first demonstrate the advantages of MPL compared to training a model from scratch and using a direct knowledge transfer mechanism. Second, the benefit of the Huber loss for improving the predictive ability of MTQ-Net is verified. Finally, the MTQ-Net with the MPL approach exhibits higher overall predictive power compared to other SSL-based speech assessment models.

Index Terms: 3QUEST, PESQ, STOI, SDI, speech quality prediction, speech intelligibility prediction, self-supervised learning

1. INTRODUCTION

Speech assessment metrics are important quantitative evaluation indicators for speech-related applications, such as speech synthesis [1], speech enhancement (SE) [2], hearing aids [3], and telecommunications [4]. With the emergence of deep learning and the availability of training samples, researchers have begun to employ deep learning models to deploy speech assessment metrics in different tasks, e.g., voice conversion [5, 6], speech enhancement [7, 8], hearing aids [9, 10], and telecommunications [4]. To achieve more accurate automatic assessment, several strategies have also been explored, e.g., reducing the bias per listener [5], incorporating a self-supervised learning model [6, 7], and performing ensemble learning [11, 12]. Despite significant improvements in performance, achieving satisfactory generalization with a limited number of training samples remains a challenge. The main objective of this study is to explore how information from an established deep learning-based speech assessment model trained on a larger training set can be leveraged to improve prediction performance on a target assessment task with limited training data. In our previous work [7], we proposed MOSA-Net, a multi-objective speech assessment model that uses cross-domain features (spectral and temporal features) and latent representations from an SSL model [13] to simultaneously predict objective quality, intelligibility, and distortion scores. We also found that, through knowledge transferring, MOSA-Net trained on objective assessment metrics can be successfully adapted to predict subjective quality and intelligibility scores.

This study aims to further expand the applicability of MOSA-Net as a teacher model by introducing the multi-task pseudo-label learning (MPL) approach. MPL consists of two stages: obtaining pseudo-label scores and performing multi-task learning. MPL uses both supervised loss and semi-supervised loss to train the target model, where the supervised loss estimates the difference between the predicted score and the ground-truth label score, and the semi-supervised loss estimates the difference between the predicted score and the pseudo-label score. Additionally, the Huber loss [14], which combines mean absolute error (MAE) and mean square error (MSE), is employed as the loss function. The pseudo-label scores used in MPL include perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and speech distortion index (SDI) [15] obtained from the pretrained MOSA-Net model. In this study, we investigate the application of MPL in transferring knowledge of MOSA-Net to deploy a multi-target speech quality assessment network called MTQ-Net. The primary objective of MTQ-Net is to simultaneously predict three 3QUEST metrics [16], namely Speech-MOS (S-MOS), Noise-MOS (N-MOS), and General-MOS (G-MOS) scores, which are widely used in telecommunications. There are paid costs for using the 3QUEST tool. Additionally, it is impossible to estimate the 3QUEST scores of any speech utterance without a corresponding clean speech. To overcome these limitations, it is highly desirable to train a neural network that can estimate the 3QUEST

metrics from a single utterance. Our experimental results demonstrate the advantages of MPL over knowledge transfer and training from scratch approaches, allowing the deployed MTQ-Net model to achieve better prediction capabilities. Furthermore, utilizing the Huber loss can yield higher prediction performance compared to MAE and MSE alone. Finally, MTQ-Net with the MPL approach demonstrates improved overall prediction performance compared to other SSL-based speech assessment models [6].

The remainder of this paper is organized as follows. Section II presents the proposed MPL mechanism and its use in MTQ-Net. Section III describes the experimental setup and results. Finally, Section IV presents the conclusions and future work.

2. MULTI-TASK PSEUDO-LABEL LEARNING

In this section, we introduce the overall framework of the MPL approach. As shown in Fig. 1, MPL consists of two distinct stages: obtaining pseudo-label scores and performing multi-task learning. The primary ground-truth labels are obtained using the 3QUEST tool, where the calculation of the S-MOS, N-MOS, and G-MOS scores of a speech utterance requires its corresponding clean speech utterance. The pretrained MOSA-Net model is adopted to obtain the pseudo labels, including PESQ, STOI, and SDI scores.

In the second stage, multi-task learning is performed to train the MTQ-Net model to predict 3QUEST scores. As shown in Fig. 1, during the training phase, speech waveforms are processed by a cross-domain feature extraction module. This module generates three types of acoustic features: power spectral features from the short-time Fourier transform (STFT), learnable filter banks (LFB) from the Sinc convolution layer [17], and SSL embeddings from a self-supervised learning (SSL) model [18]. These three acoustic features are then processed by the CNN-BLSTM module as in MOSA-Net [7]. In the final step, the output of the CNN-BLSTM module is processed by six different task-specific layers aimed at estimating S-MOS, N-MOS, G-MOS, PESQ, STOI, and SDI scores, respectively. Specifically, each task-specific layer consists of an attention layer, a fully connected layer, and a global average pooling layer. It is worth noting that the three task-specific layers for estimating pseudo-label scores will be detached during inference. The purpose of adding these three additional task-specific layers is to improve the generalization ability of the encoder layer during the training phase. The objective function for training the MTQ-Net model is defined as follows:

$$\begin{aligned} O &= \mathcal{L}_{superv} + \mathcal{L}_{semi} \\ \mathcal{L}_{superv} &= \mathcal{L}_{SMOS} + \mathcal{L}_{NMOS} + \mathcal{L}_{GMOS} \\ \mathcal{L}_{semi} &= \mathcal{L}_{PESQ} + \mathcal{L}_{STOI} + \mathcal{L}_{SDI} \end{aligned} \tag{1}$$

The overall loss O consists of a supervised loss L_{superv} calculated based on the ground-truth scores of the primary (i.e., target) task and a semi-supervised loss L_{semi} calculated based on the pseudo-label scores of the auxiliary task. For each assessment metric (e.g., S-MOS), the loss (e.g., $\mathcal{L}_{SMOS}$) is calculated by adding utterance-level loss $\mathcal{L}_{utt}$ and frame-level loss $\mathcal{L}_{fr}$, following [7]. $\mathcal{L}_{utt}$ and $\mathcal{L}_{fr}$ are calculated as follows:

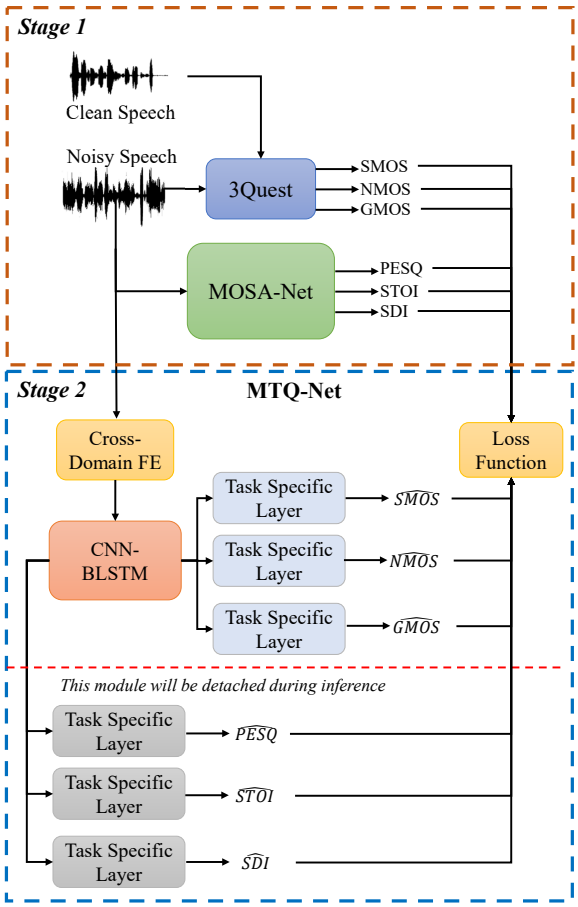

Fig. 1. Overall framework of MPL for training the MTQ-Net model.

$$\mathcal{L}_{utt} = \begin{cases} \frac{1}{U}\sum_{u=1}^{U} \frac{1}{2}(S_u - \hat{S}_u)^2 & |S_u - \hat{S}_u| \le \delta \\ \frac{1}{U}\sum_{u=1}^{U} \delta(|S_u - \hat{S}_u| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{2}$$

$$\mathcal{L}_{fr} = \begin{cases} \frac{1}{U}\sum_{u=1}^{U} \frac{\alpha_S}{F_u}\sum_{f=1}^{F_u} \frac{1}{2}(S_u - \hat{s}_f)^2 & |S_u - \hat{s}_f| \le \delta \\ \frac{1}{U}\sum_{u=1}^{U} \frac{\alpha_S}{F_u}\sum_{f=1}^{F_u} \delta(|S_u - \hat{s}_f)| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{3}$$

$S_u, \hat{S}_u, \hat{s}_f$ are the ground-truth (or pseudo-label) score, predicted utterance-level score, and predicted frame-level score, respectively. The parameter δ is a hyperparameter that determines whether the Huber loss uses MAE or MSE. U denotes

the total number of training utterances; F_u denotes the number of frames in the u-th training utterance; α_S is the weight between utterance-level and frame-level losses.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluated the proposed MTQ-Net model on the Taiwan Mandarin Hearing In Noise test - Quality & Intelligibility (TMHINT-QI) dataset [8]. The dataset includes clean, noisy, and enhanced speech utterances from five different SE systems, including Karhunen-Loeve transform (KLT) [19], minimum-mean squared error (MMSE) [20], fully convolutional network (FCN) [21], deep denoising autoencoder (DDAE) [22], and transformer-based SE [23]. TMHINT-QI provided a diverse set of metrics, including both objective and subjective measures. In this study, we specifically concentrated on additional metrics, not originally part of TMHINT-QI, which are the 3QUEST scores, consisting of S-MOS, N-MOS, and G-MOS scores. We prepared training labels for MTQ-Net by calculating 3QUEST scores based on the noisy-clean or enhanced-clean paired utterances from TMHINT-QI. It's important to emphasize that we operated under the assumption that the other metrics offered by TMHINT-QI were inaccessible for our analysis.

The training set contained 11,000 utterances with corresponding S-MOS, N-MOS, and G-MOS scores as ground-truth labels. Specifically, from 11,000 speech samples, we allocated 90% for training and 10% for validation. The test set contained 2,500 utterances with corresponding ground-truth labels. It is worth noting that there is no overlap in training and test utterances. We used three evaluation metrics, namely MSE, linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) [24] to evaluate the prediction output of all compared models. The smaller the difference between the predicted score and the ground-truth score, the smaller the MSE value; therefore, a lower MSE value indicates better performance. The LCC and SRCC values respectively represent the numerical or ranked correlation between the predicted scores and the ground-truth scores; so the higher the value, the higher the correlation and the better the performance.

3.2. MTQ-Net with Different Training Mechanisms

In the first experiment, we compared three different training mechanisms for building MTQ-Net. First, we trained MTQ-Net from scratch using three types of ground-truth labels (denoted as "From Scratch"). Second, we performed simple knowledge transfer, i.e., initialized MTQ-Net with the weights of the MOSA-Net model trained to predict PESQ, STOI, and SDI, and then trained the model using three types of ground-truth labels (denoted as "KT"). Third, we adopted the MPL approach to construct MTQ-Net (denoted as "MPL"). In addition, we also compared the embeddings of two different SSL models, namely HuBERT [13] and WavLM [18]. All models were trained using the MSE loss. From Table 1, we first notice that performing knowledge transfer helps MTQ-Net achieve better prediction performance across all evaluation metrics (MTQ-Net (HuBERT)-KT vs MTQ-Net (HuBERT)-From Scratch). Second, WavLM embeddings are slightly more effective than HuBERT embeddings in most cases, although not always (MTQ-Net (WavLM)-KT vs MTQ-Net (HuBERT)-KT). Third, MPL, which considers three additional pseudo labels, is indeed more effective than simple knowledge transfer (MTQ-Net (WavLM)-MPL vs MTQ-Net (WavLM)-KT). Fourth, fine-tuning the SSL model during MTQ-Net training can yield better performance in most cases (MTQ-Net (FT-WavLM)-MPL vs MTQ-Net (WavLM)-MPL).

Table 1. LCC, SRCC, and MSE results of MTQ-Net using different training mechanisms.

Systems	Method	LCC	SRCC	MSE
S-MOS Score Prediction				
MTQ-Net (HuBERT)	From Scratch	0.891	0.884	0.055
MTQ-Net (HuBERT)	KT	0.894	0.892	0.055
MTQ-Net (WavLM)	KT	0.899	0.887	0.048
MTQ-Net (WavLM)	MPL	0.902	0.895	0.046
MTQ-Net (FT-WavLM)	MPL	**0.913**	**0.908**	**0.045**
N-MOS Score Prediction				
MTQ-Net (HuBERT)	From Scratch	0.734	0.775	0.107
MTQ-Net (HuBERT)	KT	**0.745**	0.787	**0.102**
MTQ-Net (WavLM)	KT	0.739	0.786	0.115
MTQ-Net (WavLM)	MPL	0.737	**0.788**	0.123
MTQ-Net (FT-WavLM)	MPL	0.714	0.775	0.115
G-MOS Score Prediction				
MTQ-Net (HuBERT)	From Scratch	0.851	0.850	0.047
MTQ-Net (HuBERT)	KT	0.860	0.863	0.045
MTQ-Net (WavLM)	KT	0.868	0.865	0.044
MTQ-Net (WavLM)	MPL	0.876	**0.880**	0.043
MTQ-Net (FT-WavLM)	MPL	**0.877**	0.876	**0.042**

3.3. MTQ-Net with Different Loss Functions

In the second experiment, we compared three different loss functions used in MPL for training MTQ-Net, namely MAE, MSE, and Huber loss. MAE is known to be robust against outlier data, while MSE performs appropriate calculations by accommodating small errors equally important as large errors. On the other hand, Huber loss combines the advantages of MAE and MSE and selects the most appropriate loss through the parameter δ. In this experiment, δ was set to 1.0. As shown in Table 2, using MSE loss can achieve better overall performance than using MAE loss. Compared with MAE and MSE alone, Huber loss combining MAE and MSE can enable the MTQ-Net model to achieve better overall performance.

Table 2. LCC, SRCC, and MSE results of MTQ-Net using different loss functions.

Method	Target	LCC	SRCC	MSE
S-MOS Score Prediction				
MPL	MSE	**0.913**	**0.908**	0.045
MPL	MAE	0.904	0.897	0.074
MPL	Huber	0.912	0.903	**0.043**
N-MOS Score Prediction				
MPL	MSE	0.714	0.775	0.115
MPL	MAE	0.712	0.783	0.117
MPL	Huber	**0.739**	**0.789**	**0.108**
G-MOS Score Prediction				
MPL	MSE	0.877	0.876	0.042
MPL	MAE	0.864	0.864	0.047
MPL	Huber	**0.881**	**0.882**	**0.039**

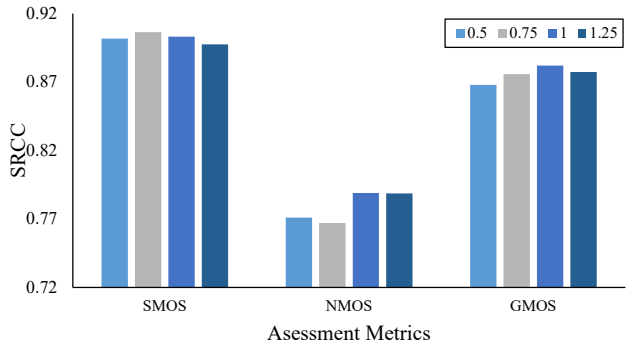

Fig. 2. SRCC results of MTQ-Net trained using Huber loss with different δ values.

3.4. Huber Loss with Different δ Values

One of the main important mechanisms of Huber Loss is the flexibility to switch between MAE and MSE through the parameter δ. Specifically, the Huber loss behaves like the MAE if the absolute difference between the label and the predicted score is larger than δ. Otherwise, the Huber loos behaves like the MSE. To study the optimal value of δ, we deployed MTQ-Net using four different δ values (0.5, 0.75, 1.00, and 1.25). As shown in Fig. 2, MTQ-Net trained with different δ values obtains different SRCC values in predicting 3QUEST scores. For S-MOS prediction, the best performance is obtained when δ is set to 0.75, while for N-MOS and G-MOS prediction, the best performance is obtained when δ is set to 1.0. This result shows that setting δ to 1.0 yields the best overall performance.

3.5. Comparing MTQ-Net with Other Models

For a more comprehensive study, we compared MTQ-Net with another SSL-based model (i.e., MOS-SSL [6]). MOS-SSL was built on the pretrained wav2vec 2.0 model by mean-pooling its output embeddings and adding a linear output layer to predict MOS scores. The wav2vec 2.0 model was fine-tuned during MOS-SSL training. We trained three versions of MOS-SSL to predict S-MOS, N-MOS, and G-MOS, respectively. We chose MOS-SSL as the baseline due to its remarkable performance as a baseline in the VoiceMOS Challenge 2022. Since most recent speech assessment models employ stacking training mechanisms and ensemble learning, they may not be directly comparable with our approach. The results in Table 3 show that MTQ-Net performs better than MOS-SSL in S-MOS and G-MOS prediction, but worse than MOS-SSL in N-MOS prediction. Overall, MTQ-Net outperforms MOS-SSL, which confirms the benefits of using the MPL approach to deploy MTQ-Net. It is worth mentioning that unlike MOS-SSL, which requires training a separate model to predict each assessment score, a single MTQ-Net model can simultaneously predict S-MOS, N-MOS, and G-MOS scores given an audio waveform as input.

Table 3. LCC, SRCC, and MSE results of MTQ-Net and MOS-SSL.

Systems	LCC	SRCC	MSE
S-MOS Score Prediction			
MOS-SSL	0.904	0.903	0.056
MTQ-Net	**0.912**	**0.903**	**0.043**
N-MOS Score Prediction			
MOS-SSL	**0.770**	**0.811**	**0.093**
MTQ-Net	0.739	0.789	0.108
G-MOS Score Prediction			
MOS-SSL	0.849	0.852	0.052
MTQ-Net	**0.881**	**0.882**	**0.039**

4. CONCLUSIONS

In this paper, we have proposed the MPL approach to achieve robust prediction capabilities of speech assessment models. MPL consists of obtaining pseudo-label scores from a well-trained speech assessment model and performing multi-task learning to deploy the target speech assessment model. In this study, the well-trained model is the MOSA-Net model for predicting PESQ, STOI, and SDI scores, while the target model is the MTQ-Net model for predicting S-MOS, N-MOS, and G-MOS scores. Our experiments lead to the following conclusions. First, we confirm the advantages of MPL over knowledge transfer and training from scratch approaches, enabling the deployed MTQ-Net model to achieve better prediction capabilities. Second, we validate the benefits of using Huber loss, which combines the strengths of mean absolute error (MAE) and mean squared error (MSE), to achieve improved prediction performance. Third, MTQ-Net trained with the MPL approach can achieve overall higher prediction performance compared to the strong SSL-based model MOS-SSL. In the future, we will investigate potential integration of MTQ-Net with speech processing applications.

5. REFERENCES

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

[2] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[3] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech*, 2022, pp. 3508–3512.

[4] G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, W. Wardah, G. Mittag, R. Culter, Z. Zhang, D. S. Williamson, F. Chen, F. Yang, and S. Shang, "ConferencingSpeech 2022 Challenge: Non-intrusive objective speech quality assessment (NISQA) challenge for online conferencing applications," in *Proc. Interspeech*, 2022, pp. 3308–3312.

[5] 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 *Proc. ICASSP*, 2021, pp. 391–395.

[6] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022.

[7] 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.

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

[9] J. Barker, M. Akeroyd, J. Trevor, J. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech*, 2022, pp. 3508–3512.

[10] R. E. Zezario, F. Chen, C. S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids," in *Proc. Interspeech*, 2022, pp. 3944–3948.

[11] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino, and Y. Zhao, "Fusion of Self-supervised Learned Models for MOS Prediction," in *Proc. Interspeech*, 2022, pp. 5443–5447.

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

[13] W.-N. Hsu, B. Bolte, Y.-Hung Hubert 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.

[14] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[15] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[16] Head Acoustics Application Note, "3QUEST: 3-fol quality evaluation of speech in telecommunications systems," 2008.

[17] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. SLT*, 2018.

[18] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[19] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[21] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA ASC*, 2017.

[22] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[23] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6649–6653.

[24] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.