Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features

Ryandhimas E. Zezario , *Member, IEEE*, Szu-Wei Fu , Fei Chen , *Senior Member, IEEE*,
Chiou-Shann Fuh , *Member, IEEE*, Hsin-Min Wang , *Senior Member, IEEE*, and Yu Tsao , *Senior Member, IEEE*

Abstract—This study proposes a cross-domain multi-objective speech assessment model, called MOSA-Net, which can simultaneously estimate the speech quality, intelligibility, and distortion assessment scores of an input speech signal. MOSA-Net comprises a convolutional neural network and bidirectional long short-term memory architecture for representation extraction, and a multiplicative attention layer and a fully connected layer for each assessment metric prediction. Additionally, cross-domain features (spectral and time-domain features) and latent representations from self-supervised learned (SSL) models are used as inputs to combine rich acoustic information to obtain more accurate assessments. Experimental results show that in both seen and unseen noise environments, MOSA-Net can improve the linear correlation coefficient (LCC) scores in perceptual evaluation of speech quality (PESQ) prediction, compared to Quality-Net, an existing single-task model for PESQ prediction, and improve LCC scores in short-time objective intelligibility (STOI) prediction, compared to STOI-Net, an existing single-task model for STOI prediction. Moreover, MOSA-Net can be used as a pre-trained model to be effectively adapted to an assessment model for predicting subjective quality and intelligibility scores with a limited amount of training data. Experimental results show that MOSA-Net can improve LCC scores in mean opinion score (MOS) predictions, compared to MOS-SSL, a strong single-task model for MOS prediction. We further adopt the latent representations of MOSA-Net to guide the speech enhancement (SE) process and derive a quality-intelligibility (QI)-aware SE (QIA-SE) approach. Experimental results show that QIA-SE outperforms the baseline SE system with improved PESQ scores in both seen and unseen noise environments over a baseline SE model.

Index Terms—Deep learning, multi-objective learning, non-intrusive speech assessment models, speech enhancement.

I. INTRODUCTION

SPEECH assessment metrics are indicators that quantitatively measure the specific attributes of speech signals. These metrics are vital to the development of speech-related application systems. A direct assessment approach measures the difference between the distorted/processed speech and clean reference at the signal level. The speech distortion index (SDI) [1] is a well-known example that calculates the distortion of the distorted/processed speech compared with the clean speech. Meanwhile, the signal-to-noise-ratio (SNR) [2] and segmental SNR (SSNR) [3] are other well-known metrics that indicate the difference between processed and noisy speech. Scale-invariant source-to-noise ratio (SI-SNR) [4] and optimal scale-invariant signal-noise ratio (OSI-SNR) [5] are improved versions of SNR that have also been proven effective in assessing speech signals. Although these signal-level metrics can directly indicate the distortion or SNR of the distorted/processed speech compared to the clean reference, they may not fully reflect the quality and intelligibility of the distorted/processed speech. Therefore, many evaluation metrics have been proposed for measuring speech quality and intelligibility.

Existing speech quality and intelligibility evaluation metrics can be classified into two categories: subjective and objective metrics. The subjective evaluation metrics are based on test scores from human listeners. To obtain subjective scores, speech samples are played to a group of human subjects, and these subjects provide feedback regarding the quality or intelligibility levels of the played speech signals. In terms of speech quality, the mean opinion score (MOS) is a typical numerical indicator in

listening tests. In most cases, the MOS metric categorizes speech quality into five levels, ranging from one to five, with a higher score indicating better quality. By contrast, the intelligibility score refers to the ratio of the number of accurately recognized words/phonemes/sentences in the played speech samples, where in this study, we focus on word-level speech intelligibility scores. To attain an unbiased assessment of speech quality and intelligibility, it is necessary to recruit a sufficient number of human subjects, and each subject must listen to a large amount of speech utterances encompassing diverse acoustic conditions, including speakers and distortion sources. This testing strategy is prohibitive and may not always be feasible. Hence, several objective evaluations metrics have been developed as surrogates for human listening tests [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

Generally, a conventional objective quality evaluation metric comprises two stages. The first stage includes a series of signal processing units designed to convert speech waveforms into handcrafted acoustic/auditory features. The second stage derives a mapping function to predict the speech quality score based on acoustic/auditory features. The mapping function can be implemented via linear regression [8], polynomial regression [6], [7], multivariate adaptive regression spline [14], and machine learning methods, such as Gaussian mixture models [9], [17], [18], support vector regression [10], [15], and artificial neural networks [19], [20], [21]. Depending on whether clean reference speech is required, objective speech quality metrics can be further classified into two categories: intrusive metrics [6] and non-intrusive metrics [7], [8], [9], [10], [11], [12], [13], [14], [15]. Well-known intrusive metrics include perceptual evaluation of speech quality (PESQ) [6] and perceptual objective listening quality analysis (POLQA) [32]. Compared with intrusive metrics, non-intrusive metrics do not require a clean reference; therefore, they are more suitable for real-world scenarios, but generally have lower assessment capabilities.

Objective intelligibility evaluation metrics can be classified into two categories. One category first segregates the speech signal under analysis into frequency subbands, and assumes that each subband independently contributes to the intelligibility. Next, the long-term subband SNR is calculated and then normalized to a value between 0 and 1. Finally, the intelligibility score is obtained using the perceptually weighted average of the normalized subband SNRs. Notable examples of this category include the articulation index (AI) [22], speech intelligibility index (SII) [23], extended SII (ESII) [24], and coherence SII (CSII) [27]. The other category is derived based on the observation that reverberation and/or additive noise tends to reduce the modulation depth of the distorted signal, compared with the clean reference signal. Well-known approaches of this category include the speech transmission index (STI) [25], spectro-temporal modulation index (STMI), normalized-covariance measure (NCM) [26], short-time objective intelligibility (STOI) [28], extended STOI (eSTOI) [29], spectrogram orthogonal polynomial measure (SOPM) [33], neurogram orthogonal polynomial measure (NOPM) [34], neurogram similarity index measure (NSIM) [35] and weighted

spectro-temporal modulation index (wSTMI) [36]. To avoid the necessity for clean reference speech, several non-intrusive approaches have been proposed. Most of them adopt statistical models of clean speech signals or psychoacoustic features for speech understanding [37]. Notable non-intrusive speech intelligibility metrics include modulation-spectrum area (ModA) [30], speech-to-reverberation modulation energy ratio (SRMR) [31], and the non-intrusive STOI [28].

Recently, the emergence of deep learning algorithms has resulted in the development of many deep learning-based speech assessment models. These models are trained to predict human subjective ratings [38], [39], [40], [41], [42], [43] or objective evaluation scores, in terms of speech quality [44], [45], [46] and intelligibility [46], [47]. To attain a higher assessment accuracy, the MBNet adopts the BiasNet architecture to compensate for the biased scores of a certain judge [48]. In addition, the multi-task learning criterion that simultaneously optimizes multiple metrics is used to train the assessment model [46], [49]. Meanwhile, different acoustic features are used as input to the assessment model to consider information from different acoustic domains [50], [51].

In contrast to current deep learning-based speech assessment models that use a single acoustic feature type and one particular objective to build the model, the proposed MOSA-Net aims to exploit rich acoustic information from multiple domains and utilize multiple objectives for model training. More specifically, MOSA-Net uses three different types of features: traditional spectral features, waveforms processed by learnable filters (based on the Sinc convolutional network [52]), and latent representations from self-supervised learned (SSL) models (wav2vec 2.0 [53] and HuBERT [54]). Note that wav2vec2.0 learns context information by identifying representations to the true quantized latent speech representation [53], and HuBERT predicts (hidden) cluster assignments [54]. Instead of directly using the outputs, we use the embeddings of these SSL models as the SSL features. For more details, please refer to [51] and [55]. Additionally, MOSA-Net adopts a multi-task learning criterion that simultaneously predicts multiple objective assessment metrics, including speech quality, intelligibility, and distortion scores. MOSA-Net is composed of a convolutional neural network (CNN) and a bidirectional long short-term memory (BLSTM) with an attention mechanism. We systematically evaluated the performance of MOSA-Net based on various model architectures, training targets, acoustic features, and datasets. Experimental results (in terms of mean square error (MSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC)) demonstrate the advantages of cross-domain features, multi-tasking learning, and attention mechanism. Furthermore, experimental results also show that MOSA-Net, originally trained to predict objective speech assessment scores (i.e., PESQ, STOI, and SDI), can serve as a pre-trained model to be adapted to an assessment model with limited training data to predict subjective quality and intelligibility scores. To the best of our knowledge, this is the first work that adapts a pre-trained objective metric assessment model to a new model that can predict human subjective ratings with a small amount of adaptation data.

In the literature, there have been several studies incorporating speech assessment models to improve SE performance [56], [57], [58], [59], such as MetricGAN [56] and MetricGAN+ [57]. In addition, some SE methods prepare multiple SE systems and use speech assessment models to select the SE system that is most suitable for the test utterance, such as SSEMS [60] and ZMOS [61]. In contrast to existing approaches, we propose integrating the latent representation of MOSA-Net into the SE system, and derive a novel quality-intelligibility-aware (QIA)-SE system. Experimental results show that QIA-SE achieves notable improvements over the baseline SE systems and several existing SE systems, which confirms the advantage of combining the knowledge in the speech assessment model to improve the enhancement capability.

The remainder of this paper is organized as follows. We first review related work in Section II. Subsequently, we elaborate the proposed methods in Section III. In Section IV, we describe the experimental setup, report the experimental results, and discuss our findings. Finally, we conclude our work in Section V.

II. RELATED WORK

A. Deep Learning-Based Assessment Metrics

To date, deep learning models have been widely used to build speech assessment systems. In this section, we review several deep learning-based assessment metrics based on different targets and model architectures.

As mentioned earlier, the assessment targets can be classified into two types. The first type is human subjective ratings, and the second is objective assessment scores. When the target is the human subjective ratings, the learned assessment metric through appropriate modeling can directly predict the human subjective ratings [43], [48], [51], [62], [63], [64]. However, a significant number of subjective listening tests encompassing many listeners and acoustic conditions must be conducted in advance to prepare ground-truth labels for an unbiased training set. In addition, it is difficult to extend the dataset to new domains, because additional subjective tests must be conducted. According to the training target criterion, the human subjective ratings can be classified into two categories: quality [43], [48], [62] and intelligibility scores [33], [65], [66]. Notable systems associated with subjective quality metrics include the following: (1) The MOSNet [62], which combines the utterance-level and frame-level scores to estimate the MOS of an utterance; (2) the DNSMOS [43], which uses the teacher-student architecture to eliminate subjective bias; and (3) the MBNet [48], which compensates for individual judgement biases using a BiasNet architecture. Compared with speech quality assessment, there is less work on predicting subjective intelligibility scores using deep learning models. For example, (1) Andersen et al. [63] used a CNN model to accommodate the entire signal composed of multiple sentences to estimate the scalar value of the intelligibility score, and (2) Pedersen et al. [64] used a CNN architecture to calculate scores locally in a short time to achieve more efficient learning with limited listening test data.

The second group adopts objective speech assessment metrics as the ground-truth labels for model training. Similar to the first group, the objective speech assessment metrics can also be divided into two categories: quality and intelligibility. For objective speech quality assessment, the PESQ [44], [45], [46], POLQA [67], and HASQI [46] scores obtained by comparing the test speech with the reference speech are often used as the ground-truth scores for training the deep-learning-based assessment metrics. For speech intelligibility assessment, the STOI [45], [46], [47] and hearing-aid speech perception index (HASPI) [68] scores are used as the training targets.

Many model architectures have been used to construct the deep learning-based assessment metrics, e.g., BLSTM [44], pyramid BLSTM [69], CNN [41], [43], and CNN-BLSTM [48], [62]. In addition, attention mechanism [46], [47], multi-task learning [46], [49], and additional network that compensate for score biases [48] have been used to improve assessment capabilities. In terms of input, different acoustic features have been explored, which can be classified into three categories. The first category includes traditional spectral features such as log Mel features [43] and power spectral (PS) features [44], [47], [62]. The second category uses learnable filters to extract features from raw waveform [45], [70]. The third category is based on the end-to-end features of the self-supervised learned network [51].

B. Incorporating Speech Assessment Metrics to SE

The idea of incorporating informative latent representations from pre-trained models to guide target speech processing tasks has been extensively studied. For example, SE systems using speaker embedding [71], [72], [73], [74] and noise embedding [75] have been shown to provide improved SE performance. Since the goal of speech assessment metrics is to estimate speech quality/intelligibility attributes given a distorted/processed speech signal, it is feasible to use the information from these assessment metrics to guide the SE process to achieve better speech quality and intelligibility. These approaches can be classified into two categories. The first category directly uses speech assessment metrics as training targets to train the SE system [56], [57], [76]. The second category uses assessment metrics to determine the best model architecture or select the most appropriate output [60], [61], [77].

III. PROPOSED METHODS

In this section, we first present the proposed MOSA-Net model. Subsequently, we will explain how to use latent representations to guide SE systems for obtaining better speech quality or intelligibility.

A. Multi-Target Speech Assessment Model With Cross-Domain Features (MOSA-Net)

Fig. 1 shows the overall architecture of the MOSA-Net model. As shown in the figure, MOSA-Net adopts cross-domain acoustic features and predicts multiple assessment scores. Given a speech waveform $\mathbf{X} = [x_1, \ldots, x_n \ldots, x_N]$, the model takes two branches of the input. In the first branch, the speech waveform $\mathbf{X}$ is processed by STFT and learnable filter banks (LFB)
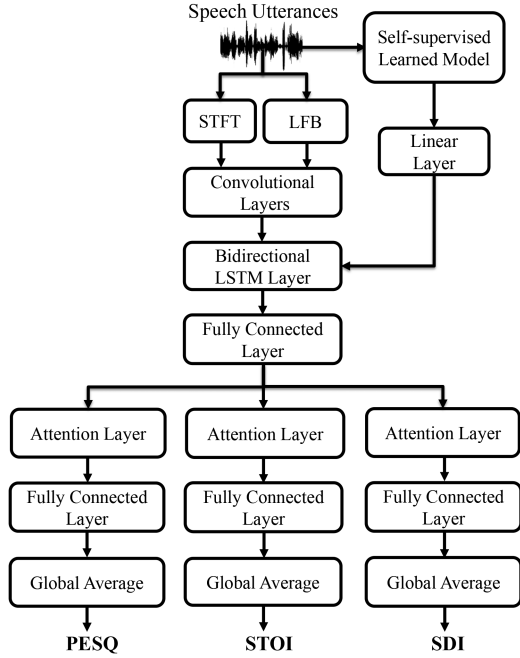
Fig. 1. Architecture of the MOSA-Net model.

$$L_{SDI} = \frac{1}{N} \sum_{n=1}^{N} \left[(S_n - \hat{S}_n)^2 + \frac{\alpha_S}{L(U_n)} \sum_{l=1}^{L(U_n)} (S_n - \hat{s}_{nl})^2 \right] \tag{1}$$

where Q_n, I_n, and S_n are the true utterance-level scores of the PESQ, STOI, and SDI, respectively, of the n-th training utterance; $\hat{Q}_n$, $\hat{I}_n$, and $\hat{S}_n$ are the predicted utterance-level scores of the PESQ, STOI, and SDI, respectively, of the n-th training utterance; N denotes the total number of training utterances; $L(U_n) = L(X_n) + L(F_n) + L(C_n)$ denotes the total number of frames in the n-th training utterance; $L(X_n)$, $L(F_n)$, and $L(C_n)$ are the number of frames of the STFT, LFB, and SSL features, respectively; $\hat{q}_{nl}$, $\hat{i}_{nl}$, and $\hat{s}_{nl}$ are the predicted frame-level PESQ, STOI, and SDI scores of the l-th frame of the n-th training utterance, respectively; as shown in (1), for the n-th training utterance, there are $L(U_n)$ predicted frame-level PESQ, STOI, and SDI scores; α_Q, α_I, and α_S determine the weights between utterance-level and frame-level losses; and γ_1, γ_2, and γ_3 determine the weights between PESQ, STOI, and SDI losses. In (1), for each metric, the first and second terms estimate the accuracy of the utterance-level score and the frame-level score, respectively. The frame-level STOI, PESQ, and SDI ground-truth scores are the same as the utterance-level STOI, PESQ, and SDI ground-truth scores, respectively. Although γ_1, γ_2, γ_3, α_Q, α_I, and α_S can be adjusted, we set them all to 1 in our experiments. A linear layer was applied to reduce the high-dimensional SSL features to the same dimension as the STFT and LFB features.

B. QIA-SE Model

The QIA-SE model is designed to incorporate the latent representation from MOSA-Net to guide the SE process. The overall QIA-SE architecture is illustrated in Fig. 2. As shown in the figure, the noisy speech waveform is first converted to spectral features, $\mathbf{Y} = [y_1, \ldots, y_n \ldots, y_L]$, where L is the total number of frames. QIA-SE aims to convert $\mathbf{Y}$ to enhanced spectral features $\hat{\mathbf{X}}$, referring to the latent representation features $\mathbf{A} = [a_1, \ldots, a_n \ldots, a_L]$ extracted by the MOSA-Net from the input noisy speech waveform. In our current implementation, $\mathbf{A}$ is the output of the fully-connected layer after attention in MOSA-Net, and is incorporated into the middle layer of QIA-SE via vector concatenation to guide the enhancement process,

$$\mathbf{H_1} = F_\theta^{(1)} (\mathbf{Y})$$
$$..$$
$$\mathbf{H}_k = F_\theta^{(k)} (\mathbf{H}_{k-1})$$
$$\mathbf{H}_{k+1} = F_\theta^{(k+1)} ([\mathbf{H}_k^T, \mathbf{A}^T]^T)$$
$$..$$
$$\mathbf{H}_K = F_\theta^{(K)} (\mathbf{H}_{K-1})$$
$$\hat{\mathbf{X}} = F_\theta^{(K+1)} (\mathbf{H}_K) \tag{2}$$

separately, and the estimated spectral and filtered signal features are fed into a convolutional layer. In the second branch, the speech waveform $\mathbf{X}$ is processed by a self-supervised learned model (HuBERT [54] or wav2vec 2.0 [53]). It is noteworthy that different types of features are concatenated by temporal dimension. That is, the total number of frames of an utterance is the summation of the frame numbers of the PS, LFB, and SSL features. Besides, linear transformations are used to make these features have the same feature dimension. The combined features are further processed by a bidirectional layer and a fully connected layer. Subsequently, a set of attention layers is used for the corresponding objective assessment metrics. In our implementation, multiplicative attention is used in the attention layers because of its high efficiency and decent performance. Next, for each metric, a fully connected layer is used to generate the frame-wise scores. Finally, based on the frame-level scores, a global average operation is applied to calculate the final predicted PESQ, STOI, and SDI scores.

Considering that speech utterances may contain stationary and/or non-stationary noise in different segments of frames, directly estimating the utterance level score may result in less accurate estimation. Therefore, the MOSA-Net aims to combine utterance-level and frame-level score estimations. Accordingly, the objective function of MOSA-Net is defined as follows:

$$L_{All} = \gamma_1 L_{PESQ} + \gamma_2 L_{STOI} + \gamma_3 L_{SDI}$$

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

$$L_{STOI} = \frac{1}{N} \sum_{n=1}^{N} \left[(I_n - \hat{I}_n)^2 + \frac{\alpha_I}{L(U_n)} \sum_{l=1}^{L(U_n)} (I_n - \hat{i}_{nl})^2 \right]$$
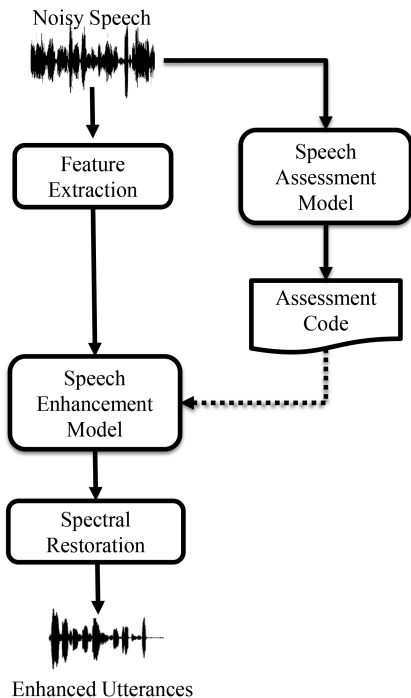
Fig. 2.   Architecture of the QIA-SE model.

where k denotes the layer index, and $F_\theta^{(k)}$ indicates the transformation model of the k-th hidden layer. Parameter θ is optimized by minimizing the following loss function based on the MSE:

$$\theta = \arg\min_\theta L(\mathbf{X}, \hat{\mathbf{X}}), \qquad (3)$$

where $\mathbf{X}$ denotes the clean speech reference. In this work, QIA-SE was built with a CNN model [61], which comprised 12 convolutional layers, followed by a fully connected layer consisting of 128 neurons. Each convolutional layer contained four channels {16, 32, 64, 128} with three types of strides {1, 1, 3} in each channel. Note that during training, MOSA-Net is frozen, and the parameters in the SE model are optimized. We can take $\mathbf{A}$ from the PESQ, STOI, or SDI branch, or from multiple branches by concatenating the representations by feature dimension. In the testing stage, noisy speech is first input into MOSA-Net to generate the latent representation $\tilde{\mathbf{A}}$, and then input into the QIA-SE model to obtain the enhanced spectral features, as defined in (2). The enhanced speech waveform is generated by performing ISTFT on the enhanced spectral features along with phase information from the noisy speech.

IV. EXPERIMENTS

A. Analysis of MOSA-Net

In this section, we systematically investigate the correlations between the performance of MOSA-Net and different input features, model architectures, and output labels. A fair comparison of MOSA-Net with related neural evaluation metrics is presented.

1) Experimental Setup: We used the Wall Street Journal (WSJ) dataset [78], which comprises 37,416 training utterances

and 330 test utterances. The training and test utterances were recorded at a sampling rate of 16 kHz.

We artificially contaminated the clean training utterances with 100 types of noises [79] at 31 different SNR levels, ranging from -10 to 20 dB with an interval of 1 dB, to prepare noisy utterances. We used 37,416 noisy-clean utterance pairs (randomly sampled a corresponding noisy utterance for each clean training utterance) to train an SE system, which was constructed by a BLSTM model with two bidirectional hidden layers, each containing 300 neurons. Then, we used the SE model to prepare enhanced utterances. Finally, we randomly sampled 1,500 clean utterances and corresponding 15,000 noisy utterances and 15,000 enhanced utterances and computed their PESQ, STOI, and SDI scores to form the training set for the MOSA-Net model.

We prepared two test sets: a seen test set and an unseen test set. For the seen test set, we randomly selected 300 clean utterances from the utterances other than the 1500 utterances in the training set and their corresponding 2,350 noisy utterances and 2,350 enhanced utterances. For the unseen test set, we selected 300 utterances from the test set of the WSJ dataset and artificially contaminated them with four unseen noise types (i.e., car, pink, street, and babble) at six SNR levels (i.e., -10, -5, 0, 5, 10, and 15 dB), amounting to 7,200 noisy utterances. The same SE model was applied to generate enhanced utterances. Note that the speakers in the unseen test set were not involved in the training set. We randomly selected 2,350 noisy utterances and 2,350 enhanced utterances together with the 300 clean utterances to form the unseen test set.

To evaluate the proposed MOSA-Net model, we adopted three evaluation metrics, namely the MSE, LCC, and SRCC [80]. Lower MSE scores indicate that the predicted scores are closer to the ground-truth assessment scores (the lower the better), whereas higher LCC and SRCC scores indicate that the predicted scores are of higher correlations to the ground-truth assessment scores (the higher the better).

2) MOSA-Net With Different Model Architectures: First, we compared the MOSA-Net with different model architectures, including the CNN [56], BLSTM [44], CNN-BLSTM [62], and CNN-BLSTM with attention [47]. In the following, the MOSA-Net implemented with CNN-BLSTM will be denoted as CRNN, while the MOSA-Net implemented with CNN-BLSTM with attention will be denoted as CRNN+AT. In our implementation of the CNN, CRNN, and CRNN+AT models, we used a batch size of one and the Adam optimizer [81] with a learning rate of 0.0001. BLSTM was trained with the RMSprop optimizer [82] with a learning rate of 0.001. For a fair comparison, we adopted the same acoustic features PS (power spectrogram) and a single-metric (either the PESQ or STOI score) learning criterion to train the model. To extract the PS features, each speech waveform was converted into a 257-dimensional spectrogram by applying a 512-point STFT with a Hamming window of 32 ms and a hop of 16 ms. The results of the MOSA-Net using the CNN, BLSTM, CRNN, and CRNN+AT are shown in Table I, where the results of both the seen and unseen tests are reported. As shown in Fig. 1, the CRNN+AT model included 12 convolutional layers, each comprising four channels 16, 32, 64, and 128, a one-layered BLSTM (with 128 nodes), and a fully connected layer (with 128

TABLE I
LCC, SRCC, AND MSE RESULTS OF MOSA-NET USING CNN, BLSTM, CNN-BLSTM (CRNN) AND CRNN WITH ATTENTION (CRNN+AT) MODEL ARCHITECTURES. THE PS FEATURES ARE USED AS THE INPUT, AND A SINGLE METRIC (EITHER THE PESQ OR STOI SCORE) IS USED TO TRAIN MOSA-NET

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
BLSTM	0.964	0.945	0.074	0.957	0.932	**0.075**
CNN	0.975	0.959	0.055	0.947	0.931	0.117
CRNN	0.981	0.965	0.042	**0.966**	0.949	0.078
CRNN+AT	**0.982**	**0.967**	**0.040**	0.965	**0.954**	0.092
STOI score prediction						
BLSTM	0.923	0.929	0.005	0.764	0.784	0.029
CNN	0.936	0.939	0.004	0.698	0.694	**0.012**
CRNN	0.964	0.962	0.002	0.789	0.797	0.016
CRNN+AT	**0.970**	**0.968**	**0.001**	**0.827**	**0.815**	0.015

The bold entities represent the best results.

neurons). An attention layer was used to estimate the assigned objective assessment metric. Finally, the output of the attention layer was forwarded to a fully connected layer (with one output node), and a global average operation was applied to generate the prediction score. The CRNN model architecture resembled CRNN+AT, where no attention layer was involved. For the CNN, we used the same model architecture as that reported in [56]. The model comprised of four two-dimensional convolutional layers with the following filters and kernels configurations: [15, (5, 5)], [25, (7, 7)], [40, (9, 9)], and [50, (11, 11)]. In addition, the two-D global average pooling was added to fix feature dimension into 50, and the feature was mapped into three fully connected layers with the following configurations: 50 and 10 LeakyReLU nodes, and one output node. For BLSTM, we used the same model architecture as that reported in [44]. The model comprised of one bidirectional LSTM layer with 100 nodes, followed by two fully connected layers with 50 exponential linear unit (ELU) nodes and one output node.

As shown in Table I, CRNN slightly outperformed the CNN and BLSTM, in terms of both the PESQ and STOI predictions for the seen and unseen test sets. The results suggest that combining the abilities of CNN in extracting local invariant features and BLSTM in characterizing temporal characteristics can yield better performance than using individual CNN and BLSTM in this task. Additionally, CRNN+AT outperformed CRNN. This indicates that by incorporating the attention mechanism, the model can focus on the more important regions and hence allow MOSA-Net to achieve better prediction performance.

Next, we compared the CRNN+AT model with two existing systems, Quality-Net [44] (based on BLSTM) and STOI-Net (based on CRNN) [47]. Note that the CRNN+AT model is actually a single-task MOSA-Net model, and the single-task MOSA-Net model with a single type of features (PS) is the same as the CRNN+AT based STOI-Net model in [47]. To qualitatively analyze the advantages of CRNN+AT, we used scatter plots[1] to compare the prediction results of these models.

[1] We drew the scatter plots using the open-source tool: https://seaborn.pydata.org/generated/seaborn.regplot.html. The regression lines were estimated with a confidence interval of 95%.
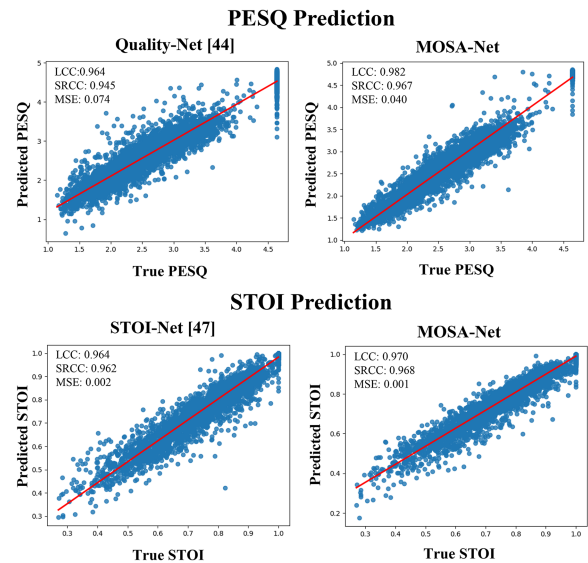

Fig. 3. Scatter plots of speech assessment predictions of MOSA-Net, Quality-Net [44], and STOI-Net [47].

TABLE II
LCC, SRCC, AND MSE RESULTS OF MOSA-NET, QUALITY-NET, AND STOI-NET

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
Quality-Net [44]	0.964	0.945	0.074	0.957	0.932	**0.075***
MOSA-Net	**0.982***	**0.967***	**0.040***	**0.965***	**0.954***	0.092
STOI score prediction						
STOI-Net [47]	0.964	0.962	0.002	0.789	0.797	0.016
MOSA-Net	**0.970***	**0.968***	**0.001***	**0.827***	**0.815***	**0.015***

*Denotes that the performance deference of the two models in each condition is statistically significant.

Scatter plots can show the correlation between predicted and true scores. More specifically, if an assessment model can accurately predict the metric scores, the points will be densely distributed on the diagonal; otherwise, the points will be spread out from the diagonal. As shown in Fig. 3, we can note that the predicted PESQ and STOI scores by MOSA-Net are more correlated with the true scores than Quality-Net and STOI-Net as the points of MOSA-Net are more densely distributed on the diagonal. More comparisons of MOSA-Net, Quality-Net, and STOI-Net are shown in Table II. From the table, we can again confirm that MOSA-Net outperforms these two previous works in almost all LCC, SRCC, and MSE metrics (except for the MSE metric for PESQ prediction under the unseen noise condition). We further compared two methods (MOSA-Net vs Quality-Net and MOSA-Net vs STOI-Net) by performing t-test on the individual average MSE/LCC/SRCC scores of 20 matched pairs (where each individual score was computed from 5 utterances), and found that all the performance differences in Table II are statistically significant, with a p-value less than 0.05.

3) MOSA-Net With Single and Multi-Task Training: Next, we aimed to compare the performance of the MOSA-Net with single- and multi-task training criteria. In the previous section, we used a single-task training criterion. Specifically, when the

TABLE III
LCC, SRCC, AND MSE RESULTS OF MOSA-NET TRAINED WITH SINGLE (Q/I/D), DOUBLE (Q+I/Q+D/I+D), AND TRIPLE (Q+I+D) METRICS FOR PREDICTING PESQ, STOI, SDI SCORES UNDER SEEN AND UNSEEN CONDITIONS. Q, I, AND D DENOTE PESQ, STOI, AND SDI SCORES, RESPECTIVELY

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
Q	0.982	0.965	0.043	0.965	**0.954**	0.092
Q+I	0.987	0.974	0.028	**0.966**	0.952	**0.068***
Q+D	0.986	0.975	0.028	0.956	0.939	0.100
Q+I+D	**0.988***	**0.977***	**0.026***	0.965	0.950	0.075
STOI score prediction						
I	0.970	0.968	**0.001**	**0.827***	0.815	0.015
Q+I	0.971	0.968	0.002	0.802	0.815	**0.014**
I+D	0.973	0.968	**0.001**	0.785	0.792	0.015
Q+I+D	**0.977***	**0.974***	**0.001**	0.790	**0.816**	0.016
SDI score prediction						
D	0.883	0.904	0.045	0.826	0.822	0.050
Q+D	0.939	0.947	0.024	**0.866***	**0.866***	0.038
I+D	**0.952***	**0.955***	**0.019***	0.848	0.826	0.042
Q+I+D	0.947	0.954	0.022	0.850	0.859	**0.036***

The bold entities represent the best results.
*Denotes that the performance deference of the best multi-task model and the single-task model in each condition is statistically significant.

Fig. 4.   Scatter plots of speech assessment predictions of the single-task and multi-task MOSA-Net models.

prediction task was the PESQ/STOI, the MOSA-Net was trained using PESQ/STOI labels. In this section, we used multiple assessment targets to train the MOSA-Net, and the model architecture is shown in Fig. 1. The results of single-, double-, and triple-task learning are shown in Table III, where the prediction targets are the PESQ, STOI, and SDI scores, respectively.

As shown in Table III, in PESQ score prediction, under the seen noise condition, the MOSA-Net models trained with double-task criteria (Q+I and Q+D) yielded better results in terms of all LCC, SRCC, and MSE metrics than MOSA-Net trained with a single-task criterion (Q), and the triple-task criterion (Q+I+D) yielded the best performance. However, for the unseen noise condition, neither double-task nor triple-task criteria led to performance gains. For STOI score prediction, similar trends can be observed in Table III. For SDI score prediction, MOSA-Net (Q+D) and MOSA-Net (I+D) achieved better performance than MOSA-Net (D) in both seen and unseen noise conditions. Additional consideration of PESQ or STOI in model training really helps with SDI score prediction. While additionally considering both PESQ and STOI in model training yielded good results, these results were not the best. We performed t-test on the results in Table III. For each testing condition, the results of the best multi-task model and the corresponding single-task model were compared. For example, for the SRCC metric in PESQ score prediction under the seen noise condition, MOSA-Net (Q+I+D) and MOSA-Net (Q) were compared. From Table III, we can see that when a multi-task model outperformed the corresponding single-task model, the performance improvement (e.g., MOSA-Net (Q+I+D) outperformed MOSA-Net (Q) in PESQ prediction in terms of SRCC: 0.977 vs 0.965) was almost always statistically significant, with a p-value less than 0.05. In contrast, when the best multi-task model performed worse than the corresponding single-task model, there was only
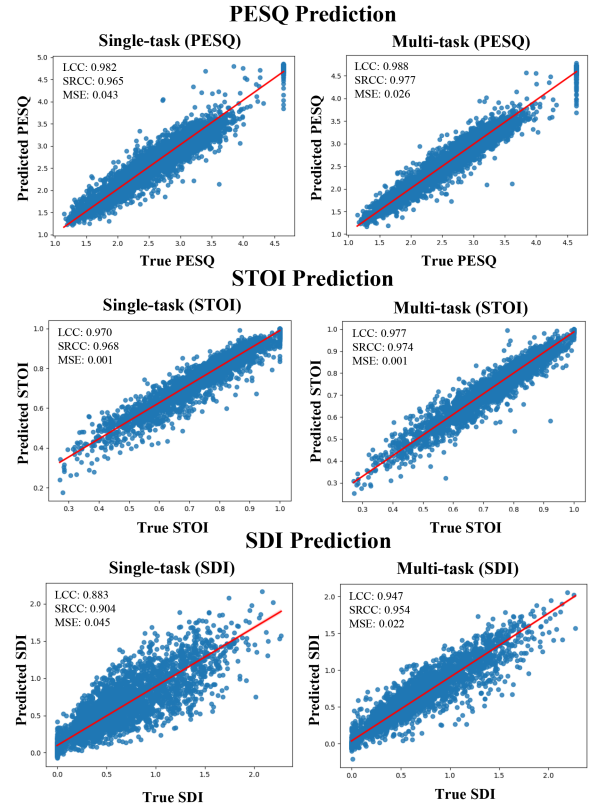
one case where the difference in performance was statistically significant, i.e., MOSA-Net (Q+I) vs MOSA-Net (I) in STOI prediction in terms of LCC under the unseen noise condition. Overall, the results in Table III suggest that PESQ, STOI, and SDI predictions are correlated to some extent and that it is beneficial to employ a multi-task learning criterion when training speech assessment models.

In addition to quantitative analyses, we conducted qualitative analyses on the MOSA-Net trained with single- and multi-task training criteria. As shown in Fig. 4, the multi-task MOSA-Net could estimate the assessment scores more accurately than the single-task MOSA-Net models, as evidenced by the denser distribution of points along the diagonal.

To develop a comprehensive analysis, we visualized the hidden layer representation of the MOSA-Net with single-task and triple-task learning. We extracted the output of the attention layer from MOSA-Net trained with single- and multi-task criteria in Figs. 5 and 6. From Fig. 5, the representations of the single-task MOSA-Net trained with individual PESQ, STOI, and SDI values yielded different patterns when predicting the individual metrics (the PESQ, STOI, and SDI). This shows that the MOSA-Net trained with a distinct metric is learned to focus on particular regions. By contrast, as shown in Fig. 6, the multi-task MOSA-Net that was trained simultaneously on three assessment metrics yielded different visualization results. Unlike the single-task models, the multi-task MOSA-Net model yielded a similar pattern in each of the branches. Therefore, it
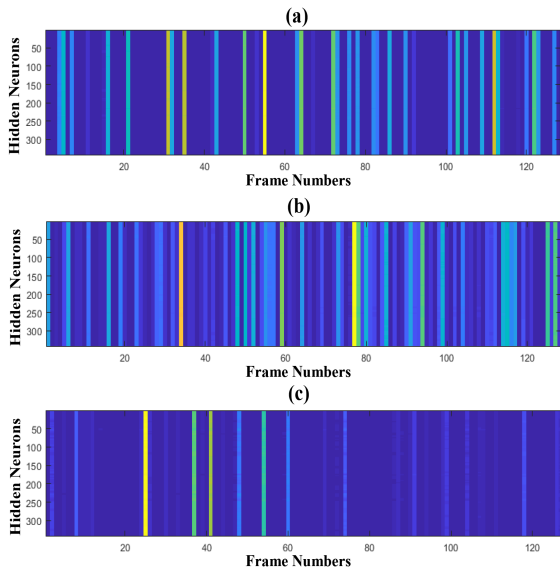
Fig. 5.　Latent representations of a speech utterance at the attention layer of the single-task MOSA-Net (a) PESQ, (b) STOI, and (c) SDI. The horizontal and vertical axes denote the frame index and attention weight, respectively.
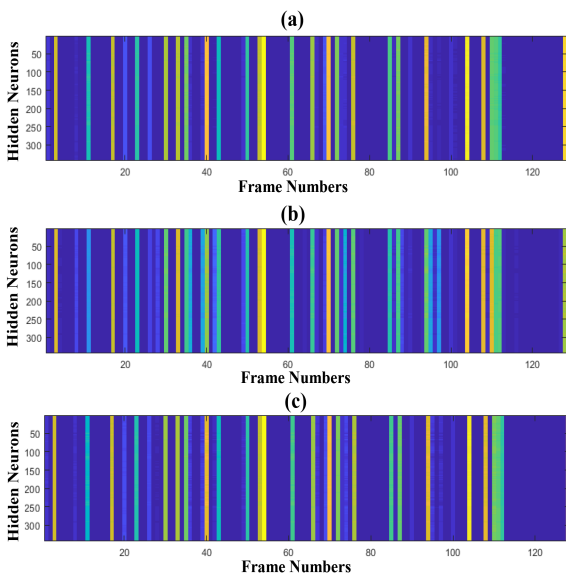

Fig. 6.　Latent representations of a speech utterance at the attention layer of the multi-task MOSA-Net (a) PESQ, (b) STOI, and (c) SDI. The horizontal and vertical axes denote the frame index and attention weight, respectively.

may further suggest that the MOSA-Net aims to share useful representations and achieve more general weights by optimally considering all metrics.

4) Comparison With Another Multi-Task Method: In this section, we compare the performance of the MOSA-Net with that of another multi-task speech assessment model, namely, attention enhanced multi-task speech assessment (AMSA) [46]. Specifically, we compared two different strategies for constructing the objective function. In our proposed work, we combined the estimated loss from the utterance- and frame-level scores to define the objective function. By contrast, AMSA uses the

TABLE IV
LCC, SRCC, AND MSE RESULTS OF MOSA-NET AND AMSA

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
AMSA [46]	0.985	0.973	0.031	0.962	0.946	0.080
MOSA-Net	**0.988**	**0.977**	**0.026**	**0.965**	**0.950**	**0.075**
STOI score prediction						
AMSA [46]	0.975	0.973	**0.001**	0.783	0.794	0.018
MOSA-Net	**0.977**	**0.974**	**0.001**	**0.790**	**0.816**	**0.016**
SDI score prediction						
AMSA [46]	0.929	0.942	0.029	0.835	0.847	0.039
MOSA-Net	**0.947**	**0.954**	**0.022**	**0.850**	**0.859**	**0.036**

The bold entities represent the best results.

regression loss based on the utterance-level score and the classification loss based on the classification-aided model to define the objective function. For a fair comparison, the same model architecture with the same number of assessment targets was used in both systems. When training the AMSA system, we followed the same parameter as defined in [46] to adjust the classification-aided model.

As shown in Table IV, MOSA-Net consistently outperformed AMSA in PESQ, STOI, and SDI predictions across all LCC, SRCC, and MSE metrics. Therefore, these evaluation results demonstrate the benefit of combining the utterance-level score and the frame-level score to form the objective function.

5) MOSA-Net With Cross-Domain Features: In this section, we investigate the effectiveness of different acoustic features on the performance of MOSA-Net [2] and whether the combination of multiple acoustic features can lead to more accurate predictions for MOSA-Net. In addition to PS features, which have been used in the previous experiments, MOSA-Net employed complex features (termed complex), learnable filter banks (termed LFB features), and the output of a self-supervised learned model (termed SSL features). The goals of using these three types of features are as follows: (1) Complex features can reserve the phase information; (2) LFB features can retain the raw-waveform information more completely; (3) SSL features can exploit the context-information of phones. For (1), we used real and imaginary (RI) spectrograms. For (2), we used SincNet [52] as the learnable feature extraction model. For (3), we used two self-supervised learned models, namely wav2vec 2.0 [53] and HuBERT [54], to generate the SSL features. The corresponding features are termed SSL(W2V) and SSL(Hub), respectively. The results of MOSA-Net using PS, Complex, LFB, SSL(W2V), and SSL(Hub) features are shown in Table V.

As shown in Table V, the PS features tended to achieve slightly better performance than the other features when estimating the PESQ. By contrast, in assessing the STOI score, the SSL(Hub) features achieved better performance in both the seen and unseen environments. Meanwhile, in assessing the SDI score, the PS features achieved better performances in the seen and unseen environments, respectively. Hence, it is indicated that these acoustic features have different and complementary information

[2][Online]. Available: https://github.com/dhimasryan/MOSA-Net-Cross-Domain

TABLE V
LCC, SRCC, AND MSE RESULTS OF MOSA-NET USING DIFFERENT INPUT FEATURES

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
PS	**0.988**	**0.977**	**0.026**	**0.965**	0.950	**0.075**
Complex	0.985	0.975	0.031	**0.965**	**0.951**	0.081
LFB	0.981	0.971	0.040	0.957	0.942	0.091
SSL(W2V)	0.984	0.972	0.033	0.961	0.947	0.084
SSL(Hub)	0.981	0.967	0.041	0.954	0.933	0.088
STOI score prediction						
PS	0.977	0.974	**0.001**	0.790	0.816	0.016
Complex	0.976	0.974	**0.001**	0.765	0.794	**0.014**
LFB	0.972	0.970	**0.001**	0.778	0.787	0.016
SSL(W2V)	0.970	0.968	0.002	0.804	0.820	0.017
SSL(Hub)	**0.980**	**0.978**	**0.001**	**0.807**	**0.821**	0.015
SDI score prediction						
PS	**0.947**	**0.954**	**0.022**	**0.850**	**0.859**	**0.036**
Complex	0.945	0.953	0.023	0.818	0.839	0.047
LFB	0.936	0.944	0.025	0.827	0.834	0.056
SSL(W2V)	0.890	0.918	0.043	0.822	0.836	0.074
SSL(Hub)	0.935	0.952	0.026	0.842	0.830	0.068

The bold entities represent the best results.

TABLE VI
LCC, SRCC, AND MSE RESULTS OF MOSA-NET USING CROSS-DOMAIN FEATURES

Model	Seen Noises			Unseen Noises		
	LCC	SRCC	MSE	LCC	SRCC	MSE
PESQ score prediction						
PS+SSL(Hub)	**0.991**	**0.981**	**0.020**	0.968	**0.957**	**0.066**
Complex+SSL(Hub)	0.990	0.979	0.023	0.968	0.956	0.084
LFB+SSL(Hub)	0.989	0.978	0.024	0.963	0.951	0.085
PS+LFB+SSL(Hub)	0.990	0.980	0.021	**0.969**	**0.957**	0.070
Complex+LFB+SSL(Hub)	0.990	0.980	0.022	0.967	0.956	0.081
STOI score prediction						
PS+SSL(Hub)	**0.989**	0.985	**0.001**	0.814	0.820	**0.016**
Complex+SSL(Hub)	**0.989**	**0.986**	**0.001**	0.826	0.828	0.015
LFB+SSL(Hub)	0.986	0.984	**0.001**	0.834	0.834	0.022
PS+LFB+SSL(Hub)	0.985	0.984	**0.001**	**0.836**	**0.839**	0.017
Complex+LFB+SSL(Hub)	**0.989**	0.985	**0.001**	0.831	0.826	0.016
SDI score prediction						
PS+SSL(Hub)	0.961	0.966	0.016	0.878	0.866	0.044
Complex+LFB	0.942	0.950	0.023	0.839	0.857	0.048
Complex+SSL(Hub)	**0.971**	**0.973**	**0.012**	0.890	0.888	0.037
LFB+SSL(Hub)	0.964	0.970	0.014	0.851	0.836	0.060
PS+LFB+SSL(Hub)	0.964	0.967	0.015	0.878	0.872	0.045
Complex+LFB+SSL(Hub)	0.969	0.971	0.012	**0.895**	**0.899**	**0.033**

The bold entities represent the best results.

for speech assessment. In addition, by considering phase information, the Complex features can reserve useful information that is particularly more useful when conducting assessment evaluations in seen environments. Because SSL(Hub) generally provides better performance than SSL(W2V), SSL(Hub) is used as the representative SSL features in the following discussion.

Next, we further investigated the MOSA-Net that combines cross-domain features as input. As shown in Fig. 1, the STFT and learnable neural network (SincNet in this study) were applied to the speech waveform to obtain the PS/Complex and LFB features, which were then used as the input to the MOSA-Net. For the SSL(W2V) and SSL(Hub) features, the speech waveform was processed by the wav2vec 2.0 and HuBERT models, respectively, and the latent representations were input to the middle layer of the MOSA-Net model. The results of the MOSA-Net with different combinations of acoustic features are shown in Table VI.

Comparing the results in Tables V and VI, the benefits of incorporating cross-domain features to train the MOSA-Net model are evident. For example, the combination of Complex and SSL(Hub), denoted as Complex+SSL(Hub) in Table VI, consistently outperformed the individual Complex and SSL(Hub) in PESQ, STOI, and SDI predictions in both the seen and unseen environments. Furthermore, Table VI shows that the Complex+SSL(Hub) features achieved the best performance among all combinations for STOI and SDI predictions in the seen environments. Finally, the combination of three acoustic features, namely Complex+LFB+SSL(Hub)/PS+LFB+SSL(Hub), consistently achieved better performance in the unseen environments as compared with Complex+SSL(Hub) / PS+SSL(Hub). We also present the scatter plots of predictions of the MOSA-Net with the single-domain (PS) and cross-domain (PS+LFB+SSL(Hub)) features in Fig. 7. From the figure, the MOSA-Net with the cross-domain features can achieve a more accurate estimation than the MOSA-Net with the single-domain
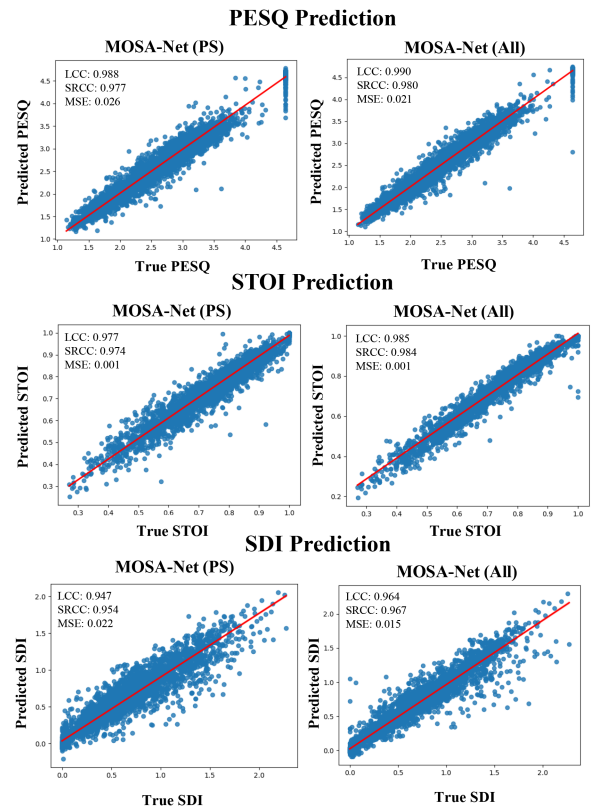

Fig. 7. Scatter plots of speech assessment predictions of the MOSA-Net with the single-domain (PS) and cross-domain (PS+LFB+SSL(Hub)) features.

features. In addition, the t-test shows that the improvements of the MOSA-Net with cross-domain (PS+LFB+SSL(Hub)) features over the MOSA-Net with single-domain (PS) features in PESQ, STOI, and SDI score predictions for both seen and unseen environments in terms of LCC, SRCC, and MSE are all statistically significant, with a p-value less than 0.05. The results

TABLE VII
LCC, SRCC, AND MSE RESULTS OF MOSA-NET TESTED ON THE TIMIT DATASET

Model	LCC	SRCC	MSE
PESQ Score Prediction			
AMSA [46]	0.728	0.673	0.765
MOSA-Net	0.754	0.710	0.654
MOSA-Net(Cross-Domain)	**0.960**	**0.948**	**0.111**
STOI Score Prediction			
AMSA [46]	0.701	0.584	0.015
MOSA-Net	0.746	0.608	0.011
MOSA-Net(Cross-Domain)	**0.920**	**0.936**	**0.004**
SDI Score Prediction			
AMSA [46]	0.873	0.678	**0.042**
MOSA-Net	0.859	0.643	0.047
MOSA-Net(Cross-Domain)	**0.895**	**0.901**	0.051

The bold entities represent the best results.

confirm the benefit of cross-domain features, which provide more complete information for the speech assessment model.

6) MOSA-Net Tested on the TIMIT Dataset: In this section, we aim to analyze the generalization ability of MOSA-Net by testing the models trained on the WSJ dataset on another standard English dataset, TIMIT [83]. The testing noisy set was obtained by injecting the same set of noises used in the WSJ task at eight SNR levels (from -10 dB to 25 dB with a step of 5 dB) to the original clean utterances, and the enhanced set was obtained by applying the same BLSTM-based SE model used in the WSJ task to the noisy utterances. Therefore, the evaluation experiment was performed under the seen noise condition. In total, we randomly selected 750 noisy utterances, 750 enhanced utterances, and 500 clean utterances to test the assessment models, including AMSA [46], MOSA-Net using PS features (denoted as MOSA-Net), and MOSA-Net using cross-domain features (denoted as MOSA-Net(Cross-Domain)). As shown in Table VII, MOSA-Net consistently outperforms AMSA in PESQ and STOI predictions while slightly underperforms AMSA in SDI prediction, but MOSA-Net(Cross-Domain) achieves the best performance in almost all testing conditions except for the MSE metric for SDI prediction. The results show that by using cross-domain features, MOSA-Net(Cross-Domain) is obviously more robust across datasets, with clear advantages over AMSA and MOSA-Net using single features. The t-test confirmed that the improvements of MOSA-Net(Cross-Domain) over AMSA in 8 out of 9 testing conditions are all statistically significant, with a p-value less than 0.05. Overall, the results confirm the effectiveness of using cross-domain features to improve generalization to new datasets.

7) Adapting MOSA-Net to Predict Human Subjective Ratings: Collecting subjective scores is time consuming and expensive because multiple subjects are required for listening tests. However, compared with models that predict objective evaluation scores, it is much more challenging to train a model to predict subjective evaluation scores that are highly correlated with human subjective ratings because human subjective ratings vary greatly, as shown in some previous work [48], [51], [62]. Therefore, there is an urgent need for a method that can effectively and efficiently train a model that replaces subjective evaluation.
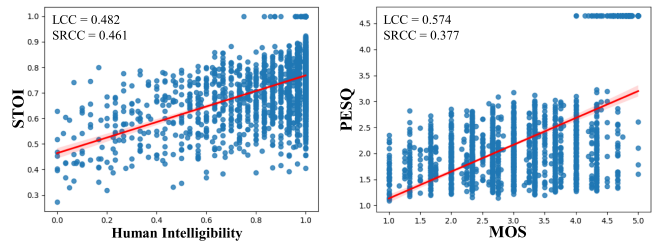

Fig. 8. Scatter plots between subjective intelligibility and STOI (left), and subjective quality and PESQ (right).

In the experiments, we used the utterances and corresponding subjective quality and intelligibility scores in the TMHINT-QI dataset [85].[3] The dataset includes clean, noisy, and enhanced utterances of five SE methods (namely Karhunen-Loeve transform (KLT) [86], minimum-mean square error (MMSE) [87], fully convolutional network (FCN) [88], deep denoising autoencoder (DDAE) [89], and transformer-based SE [90]). 226 subjects participated in the listening test (subjective quality and intelligibility).[4] Most utterances were evaluated by one subject, but some utterances were evaluated by more than one subject. In [84], it is presented that there are two setups for speech metric predictions: the corresponding prediction targets can be either average testing-condition-specific rating scores from multiple subjects or rating scores from an individual or single subject. In this study, we followed the later setup. The quality score ranges from 1 to 5, where a higher score indicates higher perceived quality. The intelligibility score ranges from 0 to 1, indicating the percentage of correctly recognized characters. We used 1,900 utterances evaluated by multiple subjects for testing and selected 15,000 utterances, each evaluated by one subject for training. For each test utterance, the average score was used as the ground-truth score. It is noteworthy that the training and test utterances do not overlap. Besides, the number of clean, noisy, and enhanced utterances of five SE models is roughly balanced.

First, we investigate whether MOSA-Net can be adapted to a new model for estimating human subjective ratings with a limited amount of training data. We used scatter plots to show the correlations between human-listening test scores (intelligibility and MOS) and objective test scores (STOI and PESQ). We drew two scatter plots in Fig. 8 (left panel: subjective intelligibility vs STOI; right panel: subjective MOS vs PESQ). We used all the test utterances to conduct the evaluation. The average MOS scores for individual speech utterances appear to be distributed only over a limited number of values because each subject rated an utterance as 1, 2, 3, 4, or 5. Considering each utterance consists of 10 Chinese characters, the human-assessed recognition rate is either 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 or 1.0. Therefore, the average speech intelligibility scores for individual speech utterances also appear to be distributed

---

[3]TMHINT-QI dataset: http://gofile.me/6PGhz/4U6GWaOtY; TMHINT-QI dataset description: https://github.com/yuwchen/InQSS

[4]Written informed consent approved by the Academia Sinica Institutional Review Board for this study was obtained from each participant before conducting the experiment.

only over a limited number of values. We also observe that the PESQ (or STOI) scores of utterances with the same average MOS (or intelligibility) score are diverse. From the two scatter plots, we note that human-listening test results (intelligibility and MOS) have moderate correlations with the corresponding objective scores (STOI and PESQ). The correlation between speech intelligibility and STOI is evaluated as 0.482 (LCC),and 0.461 (SRCC), while the correlation between MOS and PESQ is evaluated as 0.574 (LCC), and 0.377 (SRCC). These findings motivated us to use a model adaptation strategy to adapt the MOSA-Net model pre-trained on objective scores to a new model that can predict human subjective ratings.

In the next experiment, we used two state-of-the-art speech assessment models as our baselines: (1) MOSNet: a model that is based on a CNN-BLSTM architecture for predicting MOS scores [62]; (2) MOS-SSL: a model that uses features from fine-tuned wav2vec 2.0 to predict MOS scores [55]. Both models were trained on the TMHINT-QI dataset with a single-task criterion to predict the quality or intelligibility score separately. MOSNet was trained from scratch. For MOS-SSL, its linear output layer was trained from scratch while the wav2vec 2.0 module was fine-tuned. In addition, several intrusive speech quality prediction approaches such as CSIG: MOS predictor of signal distortion, CBAK: MOS predictor of background-noise, and COVL: MOS predictor of overall signal quality [84] were selected for speech quality prediction. Along with that the intrusive speech intelligibility prediction approach called ESTOI [29] was used for speech intelligibility prediction. We compared five MOSA-Net models: (1) MOSA-Net(WSJ): the best model trained on WSJ (i.e., PS+LFB+SSL(Hub) in Table VI); (2) MOSA-Net(Scratch): a model that is trained from scratch on the TMHINT-QI dataset with the same configuration as the best MOSA-Net; (3) MOSA-Net(Adapt): a model that is adapted from MOSA-Net(WSJ) using the TMHINT-QI dataset where the SSL model was not fine-tuned; (4) MOSA-Net(Scratch)$_{FT-SSL}$: same as MOSA-Net(Scratch), except that the SSL model was the one that was fine-tuned in MOS-SSL; (5) MOSA-Net(Adapt)$_{FT-SSL}$: same as MOSA-Net(Adapt), except that the SSL model was the one that was fine-tuned in MOS-SSL. The learning rate for adaptation was set to 0.00005 (half the learning rate used in MOSA-Net training).

The results are listed in Table VIII. Obviously, due to data mismatch and the gap between the PESQ/STOI metrics and the subjective quality/intelligibility scores, MOSA-Net(WSJ) could not yield satisfactory performance (especially intelligibility score prediction). In contrast, MOSA-Net(Scratch) and MOSA-Net(Adapt) performed notably better than MOSA-Net(WSJ), and several well-known intrusive evaluation metrics (CSIG, CBAK, COVL, and ESTOI). Meanwhile, MOSA-Net(Adapt) was superior to MOSA-Net(Scratch) in both speech quality and intelligibility predictions. It is also noted that with the fine-tuned SSL features, MOSA-Net(Scratch)$_{FT-SSL}$ outperformed MOSA-Net(Scratch) and MOSA-Net(Adapt), and that by utilizing the prior weight information, MOSA-Net(Adapt)$_{FT-SSL}$ achieved the best performance among all MOSA-Net models. Furthermore, MOSA-Net(Adapt)$_{FT-SSL}$ outperformed the two

TABLE VIII
LCC, SRCC, AND MSE RESULTS OF MOSA-NET FOR HUMAN LISTENING TEST PREDICTION

Model	LCC	SRCC	MSE
MOS Prediction			
CSIG [84]	0.555	0.453	-
CBAK [84]	0.545	0.343	-
COVL [84]	0.556	0.450	-
MOSNet [47]	0.724	0.656	0.489
MOS-SSL [55]	0.787	0.746	0.440
MOSA-Net(WSJ)	0.535	0.371	1.636
MOSA-Net(Scratch)	0.777	0.724	0.411
MOSA-Net(Adapt)	0.795	0.742	0.389
MOSA-Net(Scratch)$_{FT-SSL}$	0.804	0.758	0.360
MOSA-Net(Adapt)$_{FT-SSL}$	**0.805**	**0.763**	**0.356**
Speech Intelligibility Prediction			
ESTOI [29]	0.461	0.465	0.162
MOSNet [47]	0.658	0.607	0.027
MOS-SSL [55]	0.760	0.655	0.024
MOSA-Net(WSJ)	0.385	0.378	0.056
MOSA-Net(Scratch)	0.740	0.698	0.023
MOSA-Net(Adapt)	0.756	0.702	0.021
MOSA-Net(Scratch)$_{FT-SSL}$	0.796	0.712	0.018
MOSA-Net(Adapt)$_{FT-SSL}$	**0.807**	**0.730**	**0.017**

The bold entities represent the best results.

well-known baseline models, MOSNet and MOS-SSL. The t-test shows that the improvements of MOSA-Net(Adapt)$_{FT-SSL}$ over MOS-SSL in both MOS and subjective intelligibility predictions in all metrics are statistically significant, with a p-value less than 0.05. The results confirm that MOSA-Net(WSJ) can serve as a decent pre-trained model to be adapted to a new model to predict human subjective ratings. Furthermore, fine-tuning the SSL model can provide more meaningful feature representations, leading to improved prediction performance.

B. Experiments of SE With Assessment Information

In this section, we evaluate the QIA-SE system that incorporates the knowledge from the MOSA-Net model for improving the SE performance. To date, several methods have been proposed to incorporate the knowledge from the speech assessment models into an SE system, e.g., [60], [61]. We intend to compare the proposed QIA-SE system with the comparative SE systems. We tested the proposed QIA-SE system on two SE datasets, namely the WSJ corpus and the Taiwan Mandarin version of the Hearing in Noise Test (TMHINT) dataset [91]. PESQ, STOI, signal distortion (CSIG) [84], and segmental signal to noise ratio improvement (SSNRI) scores were used to evaluate the SE performance.

1) Experiments on the WSJ Dataset: We used the same 37,416 noisy-clean pairs in Section IV.A.1 to form the training set. From the test set of WSJ, we used four seen noise types (i.e., white, engine, bell, and traffic) and four unseen noise types (i.e., car, pink, street, and babble) to prepare 330 seen noisy test utterances and 330 unseen noisy test utterance at six SNR levels (i.e., -10, -5, 0, 5, 10, and 15 dB). All training and test utterances were converted to 257-dimensional log-power-spectra (LPS)

TABLE IX
THE AVERAGE SE PERFORMANCE (PESQ, STOI, CSIG, SSNRI) IN SEEN AND
UNSEEN ENVIRONMENTS

See	Noisy	CNN	SSEMS	ZMOS	QIA-SE
PESQ evaluation					
Seen	2.211	2.652	2.675	2.678	**2.953**
Unseen	2.118	2.478	2.484	2.507	**2.658**
STOI evaluation					
Seen	0.830	0.850	0.851	0.851	**0.868**
Unseen	0.799	0.820	0.820	0.821	**0.828**
CSIG evaluation					
Seen	1.790	2.091	2.089	2.060	**2.169**
Unseen	1.989	2.238	2.227	2.207	**2.312**
SSNRI evaluation					
Seen	-	4.135	4.541	3.640	**5.516**
Unseen	-	3.178	3.506	2.683	**5.731**

The bold entities represent the best results.

Fig. 9. SE performance (PESQ) of Noisy, CNN, QIA-SE (S-PESQ), QIA-SE(S-STOI), and QIA-SE(M).

features with a Hamming window of 32 ms, a hop of 16 ms, and a 512-point STFT.

The baseline SE system was built with a CNN model [61], which comprised 12 convolutional layers, followed by a fully connected layer consisting of 128 neurons. Each convolutional layer contained four channels {16, 32, 64, 128} with three types of strides {1, 1, 3} in each channel. Two comparative systems, namely specialized speech enhancement model selection (SSEMS) [60] and zero-shot model selection ZMOS [61], were constructed to evaluate the effectiveness of the proposed QIA-SE system. In SSEMS, multiple component SE models were prepared, with each model characterizing a particular noisy-clean mapping. Subsequently, a speech assessment model was incorporated to select the most suitable component model based on the estimated PESQ score. In ZMOS, the latent representation of a speech assessment model was used to prepare multiple component models in the offline stage. In the online process, the noisy speech was input into the speech assessment model to obtain the latent representation, which was then used to select the most suitable component model to perform SE. By contrast, the proposed QIA-SE system directly incorporates the latent representation into the hidden layer of the SE model, and is thus a speech-assessment-aware SE system. The enhancement results in terms of PESQ, STOI, CSIG, and SSNRI for the SSEMS, ZMOS, and QIA-SE models are shown in Table IX. Note that the baseline CNN-based SE model and the SSEMS, ZMOS, and QIA-SE models were all implemented on the same CNN architecture.

From Table IX, we first note that both SSEMS and ZMOS outperformed the baseline CNN model. Next, the proposed QIA-SE model significantly outperformed SSEMS and ZMOS in both the seen and unseen environments. In terms of STOI, similar trends were observed, i.e., QIA-SE outperformed the other SE models. In terms of CSIG and SSNRI, QIA-SE also achieved the best performance among all SE systems in both seen and unseen environments. The results confirmed the benefits of applying the speech assessment model as a supportive tool for the main SE task. It is noteworthy that SSEMS and ZMOS adopted deep learning-based speech assessment to prepare multiple component models offline and selected the best
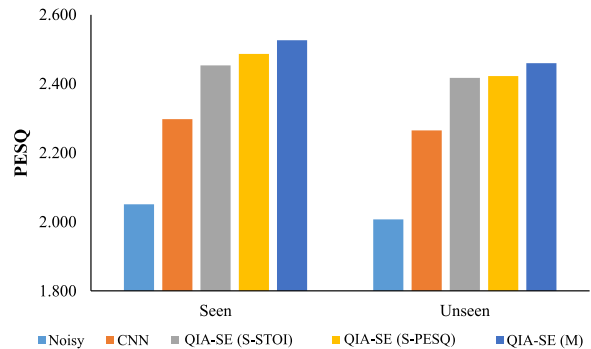
one online, where additional models and selection computations are required. By contrast, QIA-SE directly incorporates the latent representation of the assessment model. The results suggest that directly combining the assessment model into the SE system may be a more feasible and hardware-friendly approach.

2) Experiments on the TMHINT Dataset: In this experiment, we used the TMHINT dataset to evaluate the proposed QIA-SE model to fulfill three objectives: (1) to further verify the effectiveness of the QIA-SE model on a different SE task (from an English corpus to a Mandarin corpus); (2) to confirm the effectiveness of speech assessment codes across corpora in different languages; and (3) to verify the correlation of the SE performance with the MOSA-Net trained with different training criteria (single-task and multi-task learning). The training set comprised of 1,200 utterances recorded by three male and three female speakers (each speaker provided 200 utterances). We used 100 types of noises [79] to generate 36,000 noisy training utterances at 31 SNR levels (from -10 dB to 20 dB, with a step of 1 dB). The test set comprised 120 utterances recorded by another two speakers (one male and one female). We used four seen noise types (i.e., white, engine, bell, and traffic) and four unseen noise types (i.e., car, pink, street, and babble) to generate 120 seen noisy test utterances and 120 unseen noisy test utterances at six SNR levels (i.e., -10, -5, 0, 5, 10, and 15 dB).

Similar to the previous experiments, we used the same CNN architecture to develop all SE systems. We denote the QIA-SE with the MOSA-Net trained with single-task and multi-task criteria as QIA-SE(S) and QIA-SE(M), respectively. The MOSA-Net was constructed based on the best model configuration. The PESQ and STOI results under the seen and unseen noise conditions are shown in Figs. 9 and 10, respectively. QIA-SE(S-PESQ) and QIA-SE(S-STOI) indicate that the PESQ and STOI scores were used to train the single-task MOSA-Net, respectively.

As shown in Figs. 9 and 10, we note that QIA-SE(S-PESQ), QIA-SE(S-STOI), and QIA-SE(M) outperformed the baseline CNN model, whereas QIA-SE(M) achieved better performance than the other two QIA-SE models. The results again confirmed the effectiveness of QIA-SE, which leverages the speech assessment model to attain better SE capability. Meanwhile, as
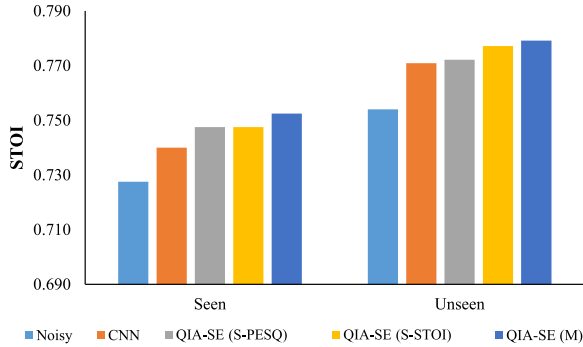
Fig. 10. SE performance (STOI) of Noisy, CNN, QIA-SE (S-PESQ), QIA-SE(S-STOI), and QIA-SE(M).
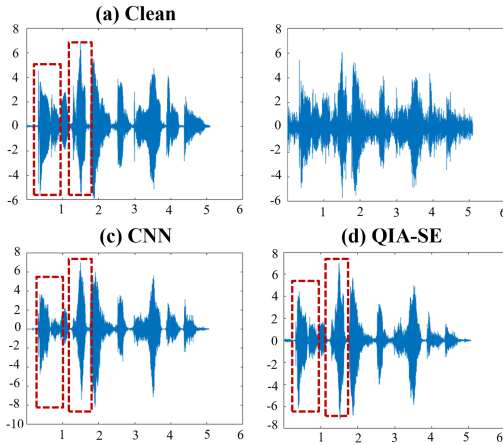

Fig. 12. Spectrograms of a clean utterance (Clean), its noisy version (car noise at 5 dB SNR) (Noisy), and the CNN and QIA-SE enhanced ones.

Fig. 11. Waveforms of a clean utterance (Clean), its noisy version (car noise at 5 dB SNR) (Noisy), and the CNN and QIA-SE enhanced ones.

Fig. 13. Amplitude envelopes from the second-channel frequency band of a clean utterance (Clean), its noisy version (car noise at 5 dB SNR) (Noisy), and the CNN and QIA-SE enhanced ones.

shown in Table III, the multi-task learning criterion allows the MOSA-Net to more accurately predict the speech assessment scores. The results in Figs. 9 and 10 show that the SE system combined with a better speech assessment model can achieve better enhancement performance.

3) Qualitative Analysis: In addition to objective evaluation, Figs. 11, 12, and 13 show the waveform, spectrogram, and amplitude envelope plots of a clean utterance, its noisy version (car noise), and the enhanced utterances (by the CNN and QIA-SE models). As shown in Fig. 11, both CNN and QIA-SE effectively removed the noise components from noisy speech. Compared with the CNN baseline, the QIA-SE model preserved more detailed structures (cf. the red rectangles in Figs. 11(a), 11(c), and 11(d)). From Fig. 12, we also note that both CNN and QIA-SE effectively reduced noise components, and the QIA-SE model preserved more details in the spectrogram than the CNN baseline (cf. the red rectangles in Figs. 12(a), 12(c), and 12(d)).

Several previous studies have shown that the amplitude envelope of the middle-frequency bands has a significant impact on speech intelligibility [92], [93]. In this study, we adopted the four-channel tone-vocoder used in [93] to extract the amplitude envelope containing 457–1202 Hz components from the speech waveform. Fig. 13 shows the amplitude envelopes of the clean, noisy, and enhanced utterances processed by the CNN and QIA-SE models, where the x- and y-axes denote the
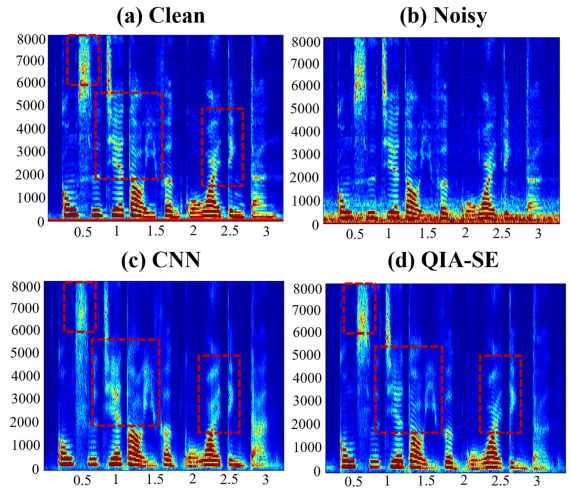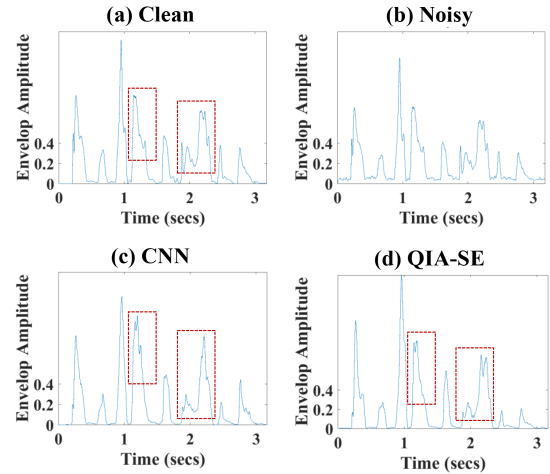
time index and amplitude magnitude, respectively. The results in Figs. 13(a), 13(c), and 13(d) show that compared with CNN enhanced speech, the amplitude envelope of QIA-SE enhanced speech is more similar to that of the original clean waveform. The results further confirmed the benefits of the QIA-SE approach.

V. CONCLUSION

In this article, we proposed a cross-domain speech assessment metric called MOSA-Net. We first systematically investigated the performance of the MOSA-Net with different model architectures and compared the prediction capability based on different training criteria (single-task vs multi-task training). Experimental results showed that the CRNN with the attention mechanism achieved the best performance as compared with the other models in terms of the LCC, SRCC, and MSE metrics. Next, the MOSA-Net with multi-task training consistently and

significantly outperformed that with single-task training. Subsequently, we tested the MOSA-Net based on different acoustic features, including spectral features, waveform processed by learnable filter banks, and representations from SSL models. The results showed that the MOSA-Net that used cross-domain features (combining information from spectral features, complex features, raw-waveform, and SSL features) achieved the best performance. Finally, we confirmed that MOSA-Net can be used as a pre-trained model to be adapted to an assessment model for predicting subjective quality and intelligibility scores.

In the second part, we proposed QIA-SE, an SE system that incorporates the information from the MOSA-Net. Experimental results showed that the QIA-SE, which directly combined the latent representations from the MOSA-Net, yielded better performance than previous SSEMS and ZMOS models, which utilized speech assessment models for offline ensemble model preparation and online model selection. In addition to better performance, the QIA-SE model required less model storage requirements and online computation. Finally, we observed that when combined with a better speech assessment model, the SE system yielded better performance.

Recently, we have used the same model architecture as MOSA-Net to predict speech intelligibility for hearing loss [94] and word error rate for ASR [95]. In the future, we will explore applying the model to predict other speech assessment metrics. Moreover, how to automatically optimize the scaling factors of different losses according to specific speech processing tasks will be further investigated. Meanwhile, we will also investigate improving the robustness of MOSA-Net in real-world scenarios.

REFERENCES

[1] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[2] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 629–632.

[3] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 2819–2822.

[4] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[5] C. Ma, D. Li, and X. Jia, "Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 711–715.

[6] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation*, 2001, Art. no. 862.

[7] T. Murphy, D. Picovici, and A. E. Mahdi, "A new single-ended measure for assessment of speech quality," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 177–180.

[8] D. Sharma, L. Meredith, J. Lainez, D. Barreda, and P. A. Naylor, "A non-intrusive PESQ measure," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2014, pp. 975–978.

[9] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, "Low-complexity, non-intrusive speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

[10] Q. Li, Y. Fang, W. Lin, and D. Thalmann, "Non-intrusive quality assessment for enhanced speech signals based on spectro temporal features," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2014, pp. 1–6.

[11] Q. Li, W. Lin, Y. Fang, and D. Thalmann, "Bag-of-words representation for non-intrusive speech quality assessment," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 616–619.

[12] L. Ding, Z. Lin, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Commun.*, vol. 49, pp. 477–489, 2007.

[13] F. Rahdari, R. Mousavi, and M. Eftekhari, "An ensemble learning model for single-ended speech quality assessment using multiple-level signal decomposition method," in *Proc. IEEE 4th Int. Conf. Comput. Knowl. Eng.*, 2014, pp. 189–193.

[14] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[15] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and L.-T. Chia, "Non-intrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1217–1232, May 2012.

[16] M. Narwaria, W. Lin, I. V. McLoughlin, S. Emmanuel, and C. L. Tien, "Non-intrusive speech quality assessment with support vector regression," in *Proc. Int. Conf. Multimedia Model.*, 2010, pp. 325–335.

[17] T. H. Falk, H. Yua, and W.-Y. Chan, "Single-ended quality measurement of noise suppressed speech based on kullback leibler distances," *J. Multimedia*, vol. 2, pp. 19–26, 2007.

[18] R. K. Dubey and A. Kumar, "Non-intrusive speech quality assessment using several combinations of auditory features," *Int. J. Speech Technol.*, vol. 16, pp. 88–101, 2013.

[19] T.-Y. Yan, M. Wei, W. Wei, and Z.-M. Xu, "A new neural network measure for objective speech quality evaluation," in *Proc. IEEE 6th Int. Conf. Wireless Commun. Netw. Mobile Comput.*, 2010, pp. 1–4.

[20] M. Hakami and W. B. Kleijn, "Machine learning based non-intrusive quality estimation with an augmented feature set," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5105–5109.

[21] M. H. Soni and H. A. Patil, "Effectiveness of ideal ratio mask for non-intrusive quality assessment of noise suppressed speech," in *Proc. IEEE 25th Eur. Signal Process. Conf.*, 2017, pp. 573–577.

[22] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[23] "Methods for calculation of the speech intelligibility index," ANSI Standard S3.5-1997, Acoustical Society of America, New York, NY, USA, 1997.

[24] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[25] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, 1980.

[26] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3679–3689, 2004.

[27] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[29] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[30] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 311–314, 2012.

[31] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[32] J. G Beerends 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," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, Jun. 2013.

[33] N. Mamun, M. S. A. Zilany, J. H. L. Hansen, and E. E. Davies-Venn, "An intrusive method for estimating speech intelligibility from noisy and distorted signals," *J. Acoust. Soc. Amer.*, vol. 150, no. 3, pp. 1762–1778, 2021.

[34] N. Mamun, W. A. Jassim, and M. S. A. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 760–773, Apr. 2015.

[35] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320, Feb. 2012.

[36] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speech intelligibility prediction using spectro-temporal modulation analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 210–225, 2020.

[37] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomed. Signal Process. Control*, vol. 71, 2022, Art. no. 103204.

[38] J. Ooster, R. Huber, and B. Meyer, "Prediction of perceived speech quality using deep machine listening," in *Proc. Interspeech*, 2018, pp. 976–980.

[39] P. Seetharaman, G. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 591–595.

[40] J. Ooster and B. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy based measures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 636–640.

[41] H. Gamper, C. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and nonintrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 85–89.

[42] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 631–635.

[43] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6493–6497.

[44] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-W. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech*, 2018, pp. 1873–1877.

[45] X. Jia and D. Li, "A deep learning-based time-domain approach for non-intrusive speech quality assessment," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 477–481.

[46] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 911–915.

[47] 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 *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 482–486.

[48] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 391–395.

[49] Y. Choi, Y. Jung, and H. Kim, "Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2020, pp. 462–469.

[50] C.-H. Hu, Y.-H. Peng, J. Yamagishi, Y. Tsao, and H.-M. Wang, "SVSNet: An end-to-end speaker voice similarity assessment model," *IEEE Signal Process. Lett.*, vol. 29, pp. 767–771, 2021.

[51] W.-C. Tseng, C. Y. Huang, W.-T. Kao, Y. Lin, and H. Y. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. Interspeech*, 2021, pp. 2781–2785.

[52] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 1021–1028.

[53] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[54] 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 Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[55] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8442–8446.

[56] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, vol. 97, pp. 2031–2041.

[57] S.-W. Fu et al., "MetricGAN : An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[58] Z. Xu, M. Strake, and T. Fingscheidt, "Deep noise suppression maximizing non-differentiable PESQ mediated by a non-intrusive PESQNet," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1572–1585, 2022.

[59] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7412–7416.

[60] R. E. Zezario, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Specialized speech enhancement model selection based on learned non-intrusive quality assessment metric," in *Proc. Interspeech*, 2019, pp. 3168–3172.

[61] R. E. Zezario, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Speech enhancement with zero-shot model selection," in *Proc. IEEE Eur. Signal Process. Conf.*, 2021, pp. 491–495.

[62] C.-C. Lo et al., "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[63] A. H. Andersen, J. M. De Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1925–1939, Oct. 2018.

[64] M. B. Pedersen, A. Heidemann Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 336–340.

[65] G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel, and C. A. Gwaltney, "Monosyllabic word recognition at higher-than-normal speech and noise levels," *J. Acoust. Soc. Amer.*, vol. 105, no. 4, pp. 2431–2444, 1999.

[66] J. R. Dubno, A. R. Horwitz, and J. B. Ahlstrom, "Word recognition in noise at higher-than-normal levels: Decreases in scores and increases in masking," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 914–922, 2005.

[67] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 7125–7129.

[68] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Commun.*, vol. 131, pp. 35–46, 2021.

[69] X. Dong and D. S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," in *Proc. Interspeech*, 2020, pp. 4631–4635.

[70] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 316–320.

[71] F.-K. Chuang, S.-S. Wang, J. w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," in *Proc. Interspeech*, 2019, pp. 3173–3177.

[72] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 181–185.

[73] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5554–5558.

[74] K. Zmolikova et al., "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, 2019.

[75] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," 2020, *arXiv:2008.11920*.

[76] K. M. Nayem and D. S. Williamson, "Incorporating embedding vectors from a human mean-opinion score prediction model for monaural speech enhancement," in *Proc. Interspeech*, 2021, pp. 216–220.

[77] Y.-T. Chang et al., "MoEVC: A mixture of experts voice conversion system with sparse gating mechanism for online computation acceleration," in *Proc. IEEE Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[78] D. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, pp. 899–902.

[79] D. Hu and D. Wang, "PNL 100 nonspeech sounds[online]," 2010. [Online]. Available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html

[80] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 15, no. 1, pp. 72–101, 1904.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[82] T. Tieleman et al., "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[83] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus CD-ROM TIMIT," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. 4930, 1993.

[84] Y. Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2007.

[85] Y.-W. Chen and Y. Tsao, "InQSS: A speech intelligibility assessment model using a multi-task learning network," in *Proc. INTERSPEECH*, to be published.

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

[87] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[88] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.

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

[90] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6649–6653.

[91] M. Huang, "Development of Taiwan mandarin hearing in noise test" Department of speech language pathology and audiology, National Taipei University of Nursing and Health Science, 2005.

[92] R. van Hoesel, M. Böhm, R. D. Battmer, J. Beckschebe, and T. Lenarz, "Amplitude-mapping effects on speech intelligibility with unilateral and bilateral cochlear implants," *Ear Hear.*, vol. 26, no. 4, pp. 381–382, 2005.

[93] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2584–2594, Nov. 2017.

[94] 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*, to be published.

[95] R. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MTI-Net: A multi-target speech intelligibility prediction model," in *Proc. INTERSPEECH*, to be published.

Szu-Wei Fu received the M.S. and Ph.D. degrees from the Graduate Institute of Communication Engineering and the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in 2014 and 2020, respectively. He is currently a Senior Applied Scientist with Microsoft, Vancouver, WA, Canada. His research interests include speech processing, speech enhancement, machine learning, and deep learning speech/speaker recognition and deep learning.

Fei Chen (Senior Member, IEEE) received the B.Sc. and M.Phil. degrees from the Department of Electronic Science and Engineering, Nanjing University, Nanjing, China, in 1998 and 2001, respectively, and the Ph.D. degree from Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, in 2005. He continued his research as Postdoctor and Senior Research Fellow with the University of Texas at Dallas, supervised by Prof. Philipos Loizou, and The University of Hong Kong, Hong Kong. He is currently a Full Professor with the Department of Electrical and Electronic Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China. Dr. Chen is leading the speech and physiological signal processing Research Group in SUSTech. He has authored or coauthored more than 100 journal papers and more than 100 conference papers in IEEE journals/conferences, Interspeech, Journal of Acoustical Society of America. His research interests include speech communication and assistive hearing technologies, brain-computer interface, and biomedical signal processing. He was tutorial speaker of Interspeech2022, Interspeech2020, EUSIPCO2022, APSIPA2021, and APSIPA2019, and organized special session Signal processing for assistive hearing devices at ICASSP2015. Dr. Chen is an APSIPA distinguished Lecturer (2022-2023), and is currently an Associate Editor for *Biomedical Signal Processing and Control* and *Frontiers in Human Neuroscience*.

Ryandhimas E. Zezario (Member, IEEE) received the M.S. degree in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2017. He is currently working toward the Ph.D. degree in computer science and information engineering with National Taiwan University, Taipei, Taiwan. He is also currently a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include speech enhancement, non-intrusive quality assessment, speech processing, speech/speaker recognition and deep learning. He was the recipient of the Gold Prize for the best non-intrusive systems and 1st place for the Hearing Industry Research Consortium student prizes at the Clarity Prediction Challenge 2022.

Chiou-Shann Fuh (Member, IEEE) received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1983, the M.S. degree in computer science from Pennsylvania State University, University Park, PA, USA, in 1987, and the Ph.D. degree in computer science from Harvard University, Cambridge, MA, USA, in 1992. From 1992 to 1993, he was with AT&T Bell Laboratories and engaged in performance monitoring of switching networks. From 1993 to 2000, he was an Associate Professor with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan and then promoted to a Full Professor. His research interests include digital image processing, computer vision, pattern recognition, mathematical morphology, artificial intelligence, deep learning, chatbot, Big Data analysis, cloud computing, and their applications to defect inspection, automatic optical inspection, industrial automation, digital still camera, digital video camcorder, surveillance camera, and camera module, such as color interpolation, auto exposure, auto focus, auto white balance, color calibration, and color management.

Hsin-Min Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently a Research Fellow. He also holds a joint appointment as a Professor with the Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan. His main research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition. From 2016 to 2020, he was an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He is currently a Senior Editor of APSIPA Transactions on Signal and Information Processing. He was the General Co-Chair of ISCSLP2016 and ISCSLP2018 and the Technical Co-Chair of ISCSLP2010, O-COCOSDA2011, APSIPAASC2013, ISMIR2014, and ASRU2019. He was the recipient of the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA distinguished Lecturer for 2014–2015. He is also a Member of the International Speech Communication Association and ACM.

Yu Tsao (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. He is also a Jointly Appointed Professor with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. His research interests include assistive oral communication technologies, audio coding, and bio-signal processing. He is currently an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS. He was the recipient of the Academia Sinica Career Development Award in 2017, national innovation awards in 2018–2021, Future Tech Breakthrough Award 2019, and Outstanding Elite Award, Chung Hwa Rotary Educational Foundation 2019–2020. He is the corresponding author of a paper that received the 2021 IEEE Signal Processing Society (SPS), Young Author, Best Paper Award.