# Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks

Xuan Dong, and Donald S. Williamson

Citation: The Journal of the Acoustical Society of America **148**, 3348 (2020); doi: 10.1121/10.0002702
View online: https://doi.org/10.1121/10.0002702
View Table of Contents: https://asa.scitation.org/toc/jas/148/5
Published by the Acoustical Society of America

JASA ARTICLE

# Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks

Xuan Dong[a] and Donald S. Williamson
*Department of Computer Science, Indiana University, Bloomington, Indiana 47408, USA*

**ABSTRACT:**
Objective metrics, such as the perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and signal-to-distortion ratio (SDR), are often used for evaluating speech. These metrics are intrusive since they require a reference (clean) speech signal to complete the evaluation. The need for a reference signal reduces the practicality of these metrics, since a clean reference signal is not typically available during real-world testing. In this paper, a two-stage approach is presented that estimates the objective score of these intrusive metrics in a non-intrusive manner, which enables testing in real-world environments. More specifically, objective score estimation is treated as a machine-learning problem, and the use of speech-enhancement residuals and convolutional long short-term memory (SER-CL) networks is proposed to blindly estimate the objective scores (i.e., PESQ, STOI, and SDR) of various speech signals. The approach is evaluated in simulated and real environments that contain different combinations of noise and reverberation. The results reveal that the proposed approach is a reasonable alternative for evaluating speech, where it performs well in terms of accuracy and correlation. The proposed approach also outperforms comparison approaches in several environments. © *2020 Acoustical Society of America*.
https://doi.org/10.1121/10.0002702
(Received 19 February 2020; revised 11 October 2020; accepted 3 November 2020; published online 30 November 2020)
[Editor: John H. L. Hansen] Pages: 3348–3359

## I. INTRODUCTION

Quality and intelligibility are two important attributes of speech. Speech quality refers to the pleasantness of a speech signal, while intelligibility measures how well one can recognize what is said in a given utterance (Loizou, 2013). Quality and intelligibility assessments are important because they directly correlate to the usefulness of speech-based applications, such as hearing aids and multimedia communication services. One form of evaluation involves subjective testing using human participants. These studies are accurate and effective, however, they are also expensive and time consuming. As a result, objective metrics are often used instead.

Intrusive metrics assess speech by computing the similarity or correlation between a clean reference signal and its degraded version (noisy, reverberant, or enhanced). Popular intrusive measures include the perceptual evaluation of speech quality (PESQ) (Rix *et al.*, 2001), short-time objective intelligibility (STOI) (Taal *et al.*, 2011), extended STOI (ESTOI) (Jensen and Taal, 2016), signal-to-distortion ratio (SDR) (Vincent *et al.*, 2006), and scale-invariant SDR (SI-SDR) (Le Roux *et al.*, 2019). Speech recognition measures (e.g., word and character error rates) can evaluate speech, but they do not assess human-level quality or intelligibility. One major shortcoming of intrusive measures is that they require a clean reference signal. Unwanted interferences are always present in real environments, which renders a clean signal inaccessible. This hinders objective assessment in real environments, which ultimately restricts improvement (Falk *et al.*, 2015).

Non-intrusive objective metrics overcome the above shortcoming by directly evaluating the signal of interest without a reference signal. Many works have been proposed (Falk *et al.*, 2010; Kim and Tarraf, 2007; Sørensen *et al.*, 2017), but P.563 (Malfait *et al.*, 2006) is the first non-intrusive speech quality assessment model. It assesses the quality of speech that is transmitted over a narrow-band telephone communication channel. Its performance, however, heavily depends on the chosen set of parameters, which limits its applicability.

Data-driven approaches have recently been proposed for non-intrusive objective evaluation. More specifically, Falk and Chan (2006) use Gaussian mixture models, support vector machines and random forest classifiers to produce intermediate features from clean and degraded signals, which are then combined by a classifier to estimate a mean opinion score (MOS). A classification and regression tree is used in Sharma *et al.* (2016), where several short- and long-term features are extracted to estimate the quality and intelligibility of speech degraded by additive noise and channel distortions. Approaches involving neural networks have been gaining in popularity. In Soni and Patil (2016), sub-band autoencoders extract acoustic features that are mapped

[a]Electronic mail: xuandong@iu.edu, ORCID: 0000-0003-0630-0701.

https://doi.org/10.1121/10.0002702

to quality scores using a single-layer neural network. In Seetharaman *et al.* (2018), a full convolutional network is used to regress from a reverberant speech signal to the speech transmission index (STI), which is an objective metric used for evaluating speech that is transmitted over a communication channel. Andersen *et al.* (2018) utilize a single convolutional layer to predict a scalar in the range of 0 to 1 that corresponds to 0%–100% speech intelligibility from subjective datasets. The non-intrusive speech quality assessment (NISQA) model (Mittag and Möller, 2019) has a CNN that first estimates per-time frame quality, and then subsequently uses a long short-term memory (LSTM) network to aggregate the per-frame values, to predict overall quality. Although these approaches offer advances, they have not been extensively evaluated in unseen noisy and reverberant environments.

Current intrusive metrics provide quality and intelligibility scores that correlate with human evaluations (Hu and Loizou, 2008; Taal *et al.*, 2011), but they cannot be used in real-world environments where the reference signal is unavailable. The existing non-intrusive approaches enable reference-less evaluation, however, their performance is not satisfactory in extremely low signal-to-noise ratios (SNR) or highly reverberant environments (Falk *et al.*, 2010; Fu *et al.*, 2018a; Kinoshita *et al.*, 2016; Sharma *et al.*, 2016).

Our goal is to predict the objective quality and intelligibility scores of intrusive metrics from a degraded speech signal using a two-stage deep learning framework. The amount of environmental noise and reverberation play a dominant role in degrading the objective quality and intelligibility of clean speech, so our idea is to use the amount of noise and reverberation as a feature for objective metric prediction. In the first stage, degraded speech is provided to a speech-enhancement system that separates a mixture into enhanced speech and residuals. The second stage then uses the residuals to predict the objective scores with a stacked convolutional long short-term memory network (ConvLSTM) (Xingjian *et al.*, 2015). We view enhancement as the first stage, although the focus of this study is on the second stage that estimates quality and intelligibility scores from the residual. We focus on predicting scores from three relatively reliable intrusive metrics (i.e., PESQ, STOI, and SDR) under noisy, reverberant, and noisy-reverberant environments. The findings serve as a pilot study for future work on predicting subjective MOS. A preliminary version of this work is published in Dong and Williamson (2018).

This paper is organized as follows. A detailed description of our approach is given in Sec. II. Experimental results and comparisons are given in Sec. III. Sections IV and V conclude the discussion of the proposed system.

## II. SYSTEM DESCRIPTION

A high-level depiction of the proposed approach is provided in Fig. 1. It consists of a speech enhancement stage and an objective score prediction stage. Speech enhancement separates the target speech from background noise. Residuals are computed by subtracting the speech estimate from the input mixture. Then the residual's spectrogram is provided to the speech assessment module for estimating the objective scores. These steps are described in more detail in the following.

### A. Motivation for using enhancement residuals

Assume that a clean speech signal, $s(t)$, is corrupted by distortions. In reverberant environments, the reverberant speech, $y(t)$, is defined as the convolution of anechoic speech, $s(t)$, with a room impulse response (RIR), $h(t)$,

$$y(t) = s(t) * h(t) = d(t) + e_r(t) + l_r(t), \tag{1}$$

where * indicates convolution and $t$ represents a time index. $d(t)$ is the direct sound, and $e_r(t)$ and $l_r(t)$ are the early and late reflections, respectively. $d(t)$ is used as the reference signal when calculating the intrusive metrics of reverberant speech. In noise-only environments, the distorted signal is merely the clean signal with additive noise [e.g., $y(t) = s(t) + n(t)$]. Distorted signals in noisy-reverberant environments are generated by combining these two distortions.

A speech enhancement approach estimates the clean speech (noise) or direct sound (reverberation), $\hat{s}(t)$, from the distorted signal. The enhancement residual, $\hat{r}(t)$, is then computed as

$$\hat{r}(t) = y(t) - \hat{s}(t) \tag{2}$$

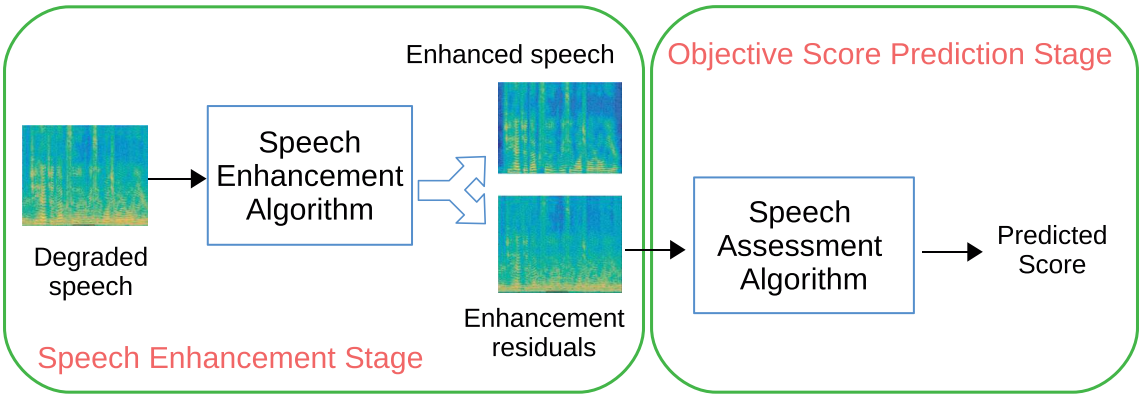



FIG. 1. (Color online) Overview of the proposed two-stage objective metric prediction approach.

and it is an estimate of the amount of additive noise or reverberation (e.g., distortion), $r(t)$, in a given signal. Subsequently, $r(t)$ equates to the noise signal ($n(t)$) in noisy environments and the reverberation $[e_r(t) + l_r(t)]$ in reverberant environments.

In general, intrusive measures generate assessment scores by computing the quantitative distance or correlation between a clean reference signal and the signal being evaluated. For instance, PESQ is a perceptually motivated measure, where the differences between the loudness spectra of the reference and degraded signals are calculated using two different approaches. One calculation results in a symmetric disturbance, $d_{SYM}$, while the other results in an asymmetric disturbance, $d_{ASYM}$ (Rix *et al.*, 2001). The final PESQ score is a direct function of these disturbances,

$$PESQ = 4.5 - 0.1 \times d_{SYM} - 0.0309 \times d_{ASYM}, \tag{3}$$

where larger disturbances lead to lower PESQ scores.

When computing STOI scores, denote the $k$th short-time Fourier transform (STFT) frequency bin of the $m$th time frame of the clean speech signal as $S(k, m)$, and of the distortion as $R(k, m)$. The STFT of the degraded speech is then, $Y(k,m) = S(k,m) + R(k,m)$. The norms of the $j$th one-third octave band are then defined as

$$S_j(m) = \sqrt{\sum_{k \in j} |S(k,m)|^2}, \tag{4}$$

$$Y_j(m) = \sqrt{\sum_{k \in j} |S(k,m) + R(k,m)|^2}. \tag{5}$$

STOI compares a sequence of $N$ one-third octave bands of the clean speech, $\boldsymbol{s}_{j,m}$, to that of the degraded speech, $\boldsymbol{y}_{j,m}$, by means of a correlation coefficient, $d_{j,m}$, as follows:

$$\boldsymbol{s}_{j,m} = \left[S_j(m-N+1), \ldots, S_j(m)\right]^T, \tag{6}$$

$$\boldsymbol{y}_{j,m} = \left[Y_j(m-N+1), \ldots, Y_j(m)\right]^T, \tag{7}$$

$$d_{j,m} = \frac{(\boldsymbol{s}_{j,m} - \mu_{\boldsymbol{s}_{j,m}})^T (\bar{\boldsymbol{y}}_{j,m} - \mu_{\bar{\boldsymbol{y}}_{j,m}})}{||\boldsymbol{s}_{j,m} - \mu_{\boldsymbol{s}_{j,m}}|| ||\boldsymbol{y}_{j,m} - \mu_{\boldsymbol{y}_{j,m}}||}, \tag{8}$$

where $\bar{\boldsymbol{y}}_{j,m}$ is a normalized and clipped version of $\boldsymbol{y}_{j,m}$, $\mu_{()}$ refers to the averaging operation, and $||\cdot||$ represents the $l_2$ norm. Upon inserting Eq. (5) into Eqs. (7) and (8), it shows that the distortion that is present within each time-frequency bin has a strong effect on $d_{j,m}$, which is expected to have a monotonic inverse relation with average intelligibility (Taal *et al.*, 2011). More specifically, as the amount of distortion increases, the subsequent objective STOI score lowers from the ideal score of 1, which occurs when distortions are not present.

For SDR, the degraded source is decomposed into the summation of the clean speech and three error terms: interference, $e_{interf}$, noise, $e_{noise}$, and artifact, $e_{artif}$, based on the principle of orthogonal projection (Vincent *et al.*, 2006),

$$s_{target} := P_{s_i} y, \tag{9}$$

$$e_{interf} := P_{\boldsymbol{s}} y - P_{s_i} y, \tag{10}$$

$$e_{noise} := P_{\boldsymbol{s},\boldsymbol{n}} y - P_{\boldsymbol{s}} y, \tag{11}$$

$$e_{artif} := y - P_{\boldsymbol{s},\boldsymbol{n}} y, \tag{12}$$

where $P_{s_i}$, $P_{\boldsymbol{s}}$, and $P_{\boldsymbol{s},\boldsymbol{n}}$ are the projection matrices for the $i$th sound source, all sound sources (including distortions), and all sound sources and sensor noises, respectively. Then the SDR is defined as the energy ratio of the true source and these errors in decibels,

$$SDR := 10 \log_{10} \frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}. \tag{13}$$

Since the latter two projection matrices (i.e., $P_{\boldsymbol{s}}$ and $P_{\boldsymbol{s},\boldsymbol{n}}$) depend on the distortions (e.g., noise or reverberation), this impacts the interference, noise and artifact errors, which impacts the final SDR.

These three intrusive metrics implicitly use a measure of distortion (or residual) in their calculations. Hence, the residual serves as a strong indicator of the resulting objective quality or intelligibility score, and could serve as an input to a score-prediction module, if it can be appropriately and accurately estimated.

### B. Speech enhancement stage

Recent work has shown that enhancement approaches that estimate a time-frequency (T-F) mask perform well (Wang *et al.*, 2014; Weninger *et al.*, 2015). Many T-F masks have been proposed (Erdogan *et al.*, 2015; Srinivasan *et al.*, 2006; Wang, 2005), but we, however, elect to use the complex ideal ratio mask (cIRM), since it has been shown to perform best in noisy, reverberant and noisy-reverberant environments (Williamson and Wang, 2017; Williamson *et al.*, 2016), where it outperformed many of the above referenced options. Later we compare the cIRM to other masks, but note that the primary focus of this work is to determine if residuals can be successfully used for score prediction, so much of our efforts are dedicated to this task.

The cIRM, which enhances the magnitude and phase of degraded speech, is constructed from the STFTs of the clean (or direct) and degraded speech,

$$\text{cIRM}(k,m) = \frac{|S(k,m)|}{|Y(k,m)|} e^{i(\phi_s(k,m) - \phi_y(k,m))}, \tag{14}$$

where $\phi_s(k,m)$ and $\phi_y(k,m)$ denote the phase response of the clean and degraded speech, respectively. Williamson *et al.* (2016) shows that the cIRM substantially improves speech quality according to objective and subjective evaluations, in many environments, so we expect it to help with estimating the residual.

The left half of Fig. 2 illustrates the speech enhancement stage. The complementary features (Wang *et al.*, 2013) from the input signal are extracted, and then provided as inputs to a deep neural network (DNN) with three dense

https://doi.org/10.1121/10.0002702

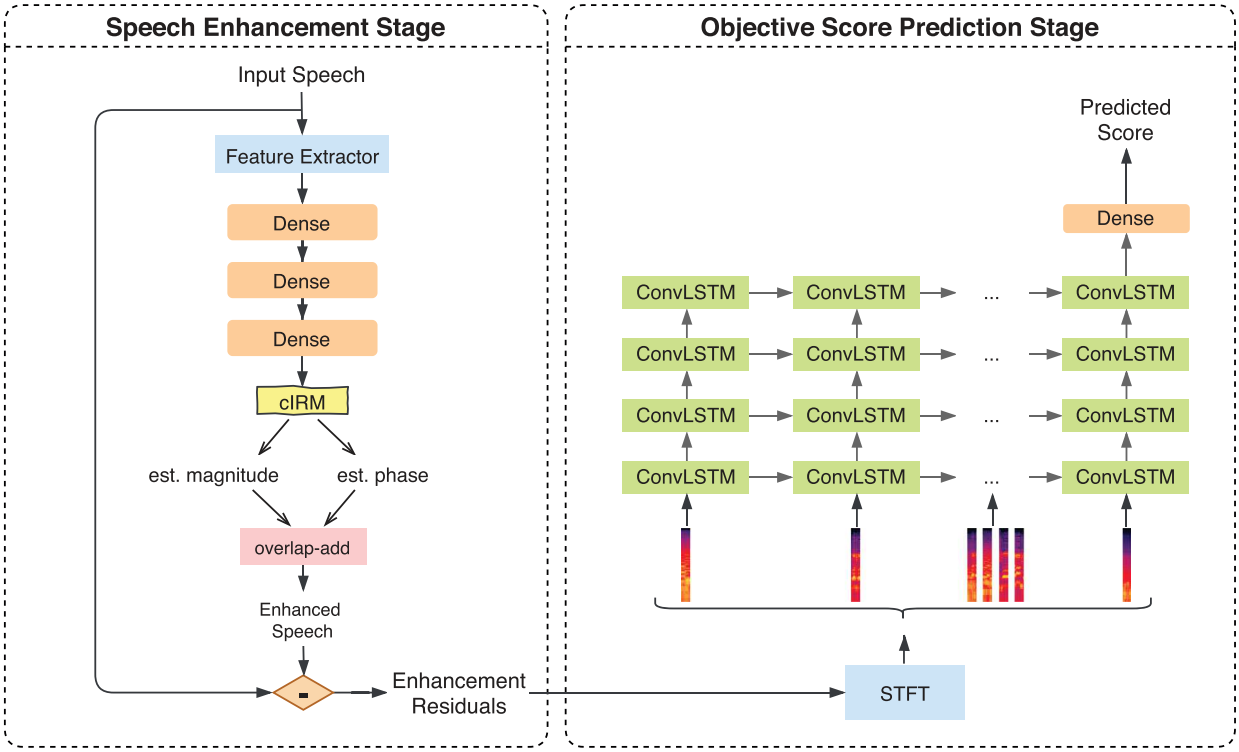



FIG. 2. (Color online) A depiction of our proposed two-stage objective metric prediction approach (cIRM-CL).

layers to estimate the cIRM. Each dense layer has 1024 units, and uses rectified linear unit (ReLU) activation functions except that a linear activation function is employed in the output layer. Adam optimization with a mean square error (MSE) loss is used to train the DNN for 200 epochs with a constant learning rate of $10^{-4}$. The remaining parameters match those as defined in the original paper. The T-F domain speech estimate is generated by applying (via Hadamard product) the estimated mask to $Y(k, m)$. Equation (2) is used to compute the enhancement residual, after performing overlap-add synthesis.

### C. Objective score prediction stage

The ConvLSTM that is proposed by Xingjian *et al.* (2015) has been successfully applied in many fields, including speech recognition (Zhang *et al.*, 2017), and speech translation (Weiss *et al.*, 2017). A LSTM effectively addresses temporal correlations, however, it is not good at maintaining local structure and it is more prone to overfitting (Salehinejad *et al.*, 2017). ConvLSTM preserves the local structure by replacing the Hadamard product ($\odot$) of the LSTM with a convolution operation (*),

$$\begin{aligned}
\boldsymbol{i}_m &= \sigma(\boldsymbol{W}_i * [\boldsymbol{x}_m, \boldsymbol{h}_{m-1}] + \boldsymbol{b}_i),\\
\boldsymbol{f}_m &= \sigma(\boldsymbol{W}_f * [\boldsymbol{x}_m, \boldsymbol{h}_{m-1}] + \boldsymbol{b}_f),\\
\boldsymbol{o}_m &= \sigma(\boldsymbol{W}_o * [\boldsymbol{x}_m, \boldsymbol{h}_{m-1}] + \boldsymbol{b}_o),\\
\boldsymbol{c}_m &= \boldsymbol{f}_m \odot \boldsymbol{c}_{m-1} + \boldsymbol{i}_m \odot \tanh(\boldsymbol{W}_c * [\boldsymbol{x}_m, \boldsymbol{h}_{m-1}] + \boldsymbol{b}_c),\\
\boldsymbol{h}_m &= \boldsymbol{o}_m \odot \tanh(\boldsymbol{c}_m),
\end{aligned} \tag{15}$$

where $\boldsymbol{x}_m$, $\boldsymbol{c}_m$, and $\boldsymbol{h}_m$ stand for the input data, cell state and hidden state at time step $m$, respectively. $\boldsymbol{i}_m$, $\boldsymbol{f}_m$, $\boldsymbol{o}_m$ denote the input, forget, and output gates, respectively. $\boldsymbol{W}$ and $\boldsymbol{b}$ are the filter matrices and bias vectors connecting different gates. $\sigma$ is the sigmoid activation function. We apply ConvLSTM units here since it not only independently captures discriminative features in frequency and time, but it also explores the co-occurrence relationship between the frequency and time domains.

The right half of Fig. 2 illustrates how enhancement residuals are utilized to predict the objective scores. Specifically, the residual's spectrogram is inputted to a stack of four ConvLSTM layers, each consisting of 16, 32, 64, and 96 filters with a $1 \times 3$ kernel (i.e., convolving only across the frequency dimension within each time step), respectively. Therefore, the characteristics of the residual can be captured by the ConvLSTM units, where local features of each frame are extracted by convolutional kernels and temporal features by recurrent LSTM networks. The hidden state for the last time step of the last ConvLSTM layer is then passed to a dense layer with 32 units, which finally outputs the predicted objective score. Training is performed for 50 epochs with an MSE loss. The learning rate is initially set to 0.01, but then it decays by 0.1 every 20 epochs. Note that the models of the first stage and the second stage are trained individually.

Generally, a two-dimensional CNN needs a fixed length feature as input, which limits the input size of the signal. The length using a ConvLSTM is not restricted in this manner. Alternatively, if using a one-dimensional CNN, fine

frequency structure is not leveraged. Additionally, the CNN-LSTM architecture learns frequency and temporal features consecutively, while ConvLSTM is able to learn temporal-frequency patterns simultaneously, which improves computational efficiency. This also may facilitate the learning of discriminative features.

## III. EVALUATIONS AND RESULTS

We evaluate the proposed approach in three different environments: noisy, reverberant, and noisy-reverberant conditions. The TIMIT speech corpus (Garofolo *et al.*, 1993) is used for the initial experiments. Each signal has been down-sampled to 16 kHz. In the speech enhancement stage, we use the same parameter configuration for calculating the complementary features as described in Wang *et al.* (2013), since they show success in modeling both noisy and reverberant speech. The STFT of the residual uses a 40 ms Hanning window, a 512 point fast Fourier transform (FFT) and 25% overlap between adjacent frames.

### A. Baseline speech enhancement approaches

We apply several T-F masking based speech enhancement approaches as baselines, and investigate how the enhancement performance of the first stage effects the prediction accuracy of the second stage. We first consider the ideal binary mask (IBM), which is a two-dimensional binary matrix that labels each T-F unit as being speech or noise dominant. The SNR of a T-F unit, $\mathrm{SNR}(k,m)$, and a local criterion (LC) are used to classify whether speech or noise dominates at each T-F point

$$\mathrm{IBM}(k,m) = \begin{cases} 1, & \text{if } \mathrm{SNR}(k,m) > LC, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The ideal ratio mask (IRM) makes soft decisions for each T-F unit, and can be computed as follows:

$$\mathrm{IRM}(k,m) = \frac{|S(k,m)|}{|S(k,m)| + |R(k,m)|}, \tag{17}$$

where $|S(k,m)|$ and $|R(k,m)|$ refer to the magnitude responses of the speech and distortion (noise or reverberation) of a distorted speech signal, respectively. The IRM returns values in-between 0 and 1 (inclusively), hence it can be thought of as the percentage of energy that is attributed to speech at each T-F unit.

Two separate DNNs are trained to estimate one of the above training targets. The DNNs each have architectures that are similar to that used for the cIRM. However, sigmoid activation functions are used in the output layer, since these labels have values between 0 and 1. The complementary features are provided as inputs as well. The same training hyper-parameters and datasets are used. In reverberant conditions, the direct sound, $D(k, m)$, is used in place of $S(k, m)$.

### B. Comparison non-intrusive methods

We initially compare our system with six non-intrusive methods: noise power spectral density (PSD) (Hendriks *et al.*, 2010), speech PSD (Erkelens *et al.*, 2007), non-intrusive speech assessment (NISA) (Sharma *et al.*, 2016), AutoMOS (Patton *et al.*, 2016), a CNN-based model (Gamper *et al.*, 2019), and Quality-Net (Fu *et al.*, 2018a), which can be used for estimating the quality and intelligibility of speech degraded by additive noise and reverberation. Noise and speech PSD are unsupervised approaches that have been used for SNR estimation (Chinaev and Haeb-Umbach, 2016; Martin, 1993; Suhadi *et al.*, 2010), which is a closely related problem. We follow this idea, but modify it for general objective metric prediction. Noise PSD is based on a minimum mean-squared error (MMSE) estimate of the noise power spectrum coefficients. Clean speech is obtained by subtracting the estimated noise PSD from the noisy speech PSD. It then is employed as the reference signal to estimate the true objective score. The speech PSD algorithm operates in a similar manner, but it directly estimates the clean speech PSD. We also select four recently developed data-driven approaches for comparison. NISA combines feature extraction with a classification and regression tree model. AutoMOS uses the log-Mel spectrogram as input to a stack of LSTMs. The outputs of the last LSTM layer are fed to two fully connected (FC) layers for estimating the final assessment score. The CNN approach utilizes Mel-frequency coefficients and pitch features as the input to a model that has four convolutional and two FC layers. Quality-Net has one bidirectional LSTM layer followed by two FC layers. For these approaches that do not release publicly available code, we followed the configurations (e.g., features and model parameters) that are used in the original papers. We verified the correctness of the implementation by obtaining nearly identical results.

### C. Performance measures

The accuracy of the prediction is evaluated by the mean absolute error (MAE) between the true objective score and the predicted score. We also use the epsilon-insensitive root MSE (RMSE$^\star$), which gives an impression of how much the prediction error exceeds the 95% confidence interval. It is recommended by ITU-T P.1401 (ITU, 2012) to compare the statistical performance of different evaluation algorithms. Last, the Pearson's correlation coefficient (PCC) is used to measure the linear relationship between a model's prediction and the ground-truth score.

### D. Noisy condition

#### 1. Experimental data

For the noisy-speech case, ten noise signals from the NOISEX-92 noise database (Varga and Steeneken, 1993) (e.g., speech-shaped noise, babble, factory floor, machine gun, cockpit, fighter jets, vehicle, radio channel, operating room, and white noise) are separately combined with 1500

https://doi.org/10.1121/10.0002702

TABLE I. Performance comparison of different speech enhancement approaches averaged across seen and unseen noisy conditions. Δ-(·) denotes performance improvement.

| | IBM | IRM | cIRM |
|---|---|---|---|
| Δ-PESQ | 0.39 | 0.42 | **0.60** |
| Δ-STOI | 0.12 | 0.14 | **0.15** |
| Δ-SDR | 5.40 | 5.60 | **7.07** |

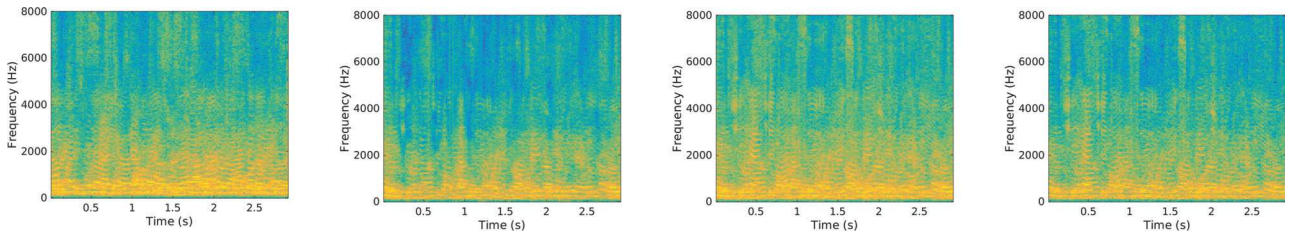
TIMIT clean speech utterances at one of ten SNR levels (from −15 to 30 dB with 5 dB increments). This results in a total of 15 000 training mixtures. Random contiguous segments from the first half of each noise are used in generating the training noisy speech mixtures. The approach is tested with 200 different TIMIT utterances that are combined with random segments from the second half of the seen noise signals at the same SNR levels as above. This results in 2000 testing mixtures for the seen noise-type dataset, even though the exact realization of the noise has not been seen during training. To further test generalization, we create a second testing set of signals that are generated from completely unseen noises and utterances. It uses 300 different clean utterances that are mixed with five unseen noise types (e.g., cafeteria, destroyer engine, live restaurant, pink, and tank noise) at one of ten SNR levels (from −15 to 30 dB with a 5 dB step). This results in 1500 testing mixtures for the unseen noise-type dataset. Note that all testing signals (seen and unseen) are composed of signals from unseen speakers. The training labels for metric prediction (e.g., PESQ, STOI, and SDR scores) are calculated from the true clean speech and generated noisy speech signals.

### 2. Results and analysis

In the first stage, we separately employ different speech-enhancement algorithms in our two-stage approach, aiming to investigate how the enhancement capabilities influence prediction accuracy. PESQ, STOI, and SDR are employed to evaluate the enhancement performance. We define Δ-(·) as the improvement of the corresponding score compared to the unprocessed noisy speech, where larger values for Δ-(·) indicate better enhancement performance. Table I shows the average improvement of each algorithm under noisy conditions. We see that every approach improves PESQ performance when compared to the unenhanced mixture. IBM offers the smallest PESQ improvement over the noisy speech, while the estimated cIRM performs best. When evaluating objective intelligibility with STOI, all approaches show noticeable improvement. Similar trends are shown when evaluating with Δ-SDR. Overall, the estimated cIRM performs better than other approaches for this task. Figure 3 shows spectrograms for the true and estimated residuals from the three enhancement approaches, where the residuals are computed from a 0 dB noisy speech signal using babble noise. In the ideal sense, the residuals would be more like the noise, since this indicates better enhancement. Notice that the estimated residual from the IBM does not capture all details of the noise. The estimated residuals from the IRM and cIRM approaches are similar.

The upper portion of Table II illustrates the prediction errors of the proposed two-stage approach using different speech-enhancement algorithms. "CL" indicates that the proposed second stage is applied. In general, we see that in noisy environments, cIRM-CL gives the best estimates compared to IBM-CL and IRM-CL. For PESQ prediction, using a cIRM in the enhancement stage significantly reduces the RMSE$^\star$ from 0.38 to 0.26 compared to IBM-CL. In terms of STOI, the MAE and RMSE$^\star$ of cIRM-CL are halved relative to the other enhancement options. For SDR, cIRM-CL clearly outperforms IBM-CL and IRM-CL in reducing the prediction error. cIRM-CL also increases the PCCs for all objective metrics. Furthermore, we observe a consistent pattern when comparing Tables I and II. That is, better enhancement performance leads to lower score prediction errors. Although the spectrograms of the IRM and cIRM are similar, the cIRM also enhances phase, which results in a better residual estimate. It reveals that improving speech-enhancement performance clearly improves prediction accuracy. Therefore, we use the cIRM for the first stage. We now denote our proposed two-stage approach that uses speech-enhancement residuals (SER) as SER-CL for the remainder of the paper.

An extreme experiment was conducted in the noisy speech condition. Assuming that a speech enhancement algorithm can ideally separate clean speech from noise in the first stage, then two input configurations to the second-stage are tested: (1) The true noise is supplied as an input to




(a) Oracle residual (b) IBM-estimated residual (c) IRM-estimated residual (d) cIRM-estimated residual

FIG. 3. (Color online) Spectrograms of the true and estimated residuals that are generated by three speech enhancement algorithms on a 0 dB seen noisy (babble) speech signal.(a) Oracle residual. (b) IBM-estimated residual. (c) IRM-estimated residual. (d) cIRM-estimated residual.

TABLE II. Comparisons between the proposed two-stage system, one-stage approaches, and different non-intrusive methods under noisy conditions. The average performance across seen and unseen testing noisy conditions is reported.

| | | PESQ | | | STOI | | | SDR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) |
| Two- stage | IBM-CL | 0.35 | 0.38 | 90.9 | 0.07 | 0.07 | 97.1 | 1.86 | 1.73 | 98.2 |
| | IRM-CL | 0.27 | 0.34 | 91.7 | 0.05 | 0.06 | 96.3 | 1.53 | 1.44 | 99.1 |
| | cIRM-CL (SER-CL) | **0.16** | **0.26** | **94.6** | **0.02** | **0.03** | **98.7** | **1.07** | **0.87** | **99.5** |
| Ideal two-stage | True noise-only | 0.23 | 0.31 | 91.2 | 0.03 | 0.03 | 95.6 | 1.59 | 1.42 | 98.1 |
| | Clean speech & noise | 0.12 | 0.18 | 96.2 | 0.02 | 0.02 | 99.2 | 0.88 | 0.79 | 99.3 |
| One- stage | IBM | 0.48 | 0.50 | 88.2 | 0.14 | 0.11 | 90.3 | 4.70 | 3.92 | 93.9 |
| | IRM | 0.42 | 0.51 | 87.9 | 0.11 | 0.09 | 90.1 | 3.97 | 3.69 | 95.1 |
| | cIRM | 0.40 | 0.48 | 89.3 | 0.08 | 0.08 | 90.8 | 3.15 | 3.02 | 95.8 |
| Other non-intrusive methods | Noise PSD | 1.20 | 1.13 | 77.4 | 0.23 | 0.19 | 80.5 | 8.22 | 7.31 | 85.3 |
| | Speech PSD | 0.80 | 0.83 | 83.6 | 0.14 | 0.13 | 85.4 | 5.37 | 4.17 | 90.0 |
| | NISA | 0.53 | 0.59 | 86.9 | 0.12 | 0.10 | 88.6 | 4.05 | 3.82 | 94.8 |
| | AutoMOS | 0.31 | 0.38 | 90.8 | 0.09 | 0.08 | 91.9 | 2.05 | 2.02 | 97.3 |
| | Quality-Net | 0.24 | 0.30 | 93.1 | 0.05 | 0.04 | 96.9 | 1.32 | 1.16 | 99.2 |
| | CNN | 0.20 | 0.29 | 94.2 | 0.04 | 0.04 | 97.8 | 1.63 | 1.54 | 98.8 |

the second score-prediction stage and (2) the true noise and the corresponding clean speech are provided as inputs to the original section stage. The results of two test cases are indicated as "ideal two-stage" in Table II. Generally, the estimation errors of the model trained with true noise are worse than the proposed two-stage model trained with the cIRM (i.e., cIRM-CL). Many factors can lead to the above results. One reason is that the network has not been designed and optimized for the pure noise, where it is possible that an alternative network structure (number of layers, number of nodes per layer, hyperparameters, etc.) may be needed. Additionally, it is quite possible that the second-stage network leverages some of the errors of the first stage, where the errors (removal of speech or retention of noise) provide distinguishing information that improves performance. It is worth noting that the model trained with both clean speech and pure noise obtains much lower error and higher correlation compared to the other models. It confirms that in the ideal case, if both original speech and noise are available, the prediction performance using the neural network is also ideal. These results can be regarded as a performance upper bound of the proposed framework.

In order to evaluate the importance of the second stage, we also directly use the enhanced speech from the first stage as an estimate of the clean reference to calculate intrusive scores (e.g., using enhanced speech as the substitute for clean speech in PESQ calculation). We denote this approach as "one-stage," since the second prediction stage is not applied. The comparison results are shown in the middle of Table II. Consistently, for each enhancement algorithm, the prediction results using one stage are severely worse than the proposed two-stage approach. The reason is that speech enhancement algorithms do not perform perfectly, so the difference between the true clean speech and the estimated clean speech leads to the inaccurate estimates of true intrusive scores, especially when the testing data include many challenging noise types and SNRs. It indicates that a speech enhancement stage followed by a metrics prediction stage is preferred. Thus, we elect to further process the enhancement residuals for score prediction.

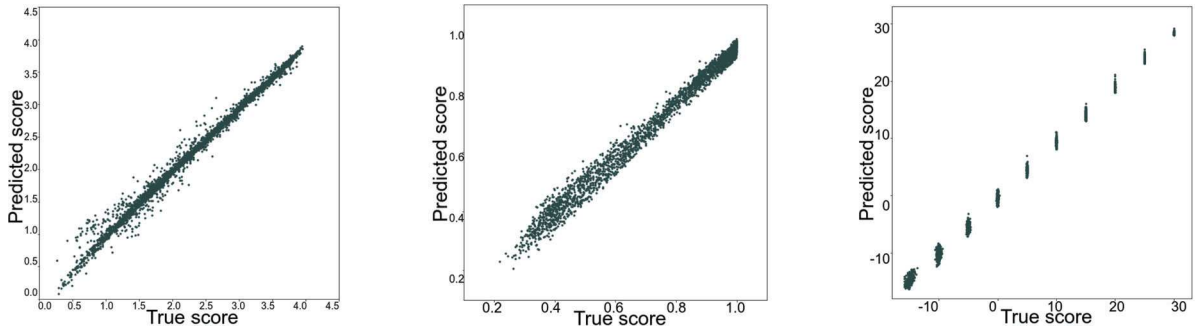
The results when the proposed system is compared to other non-intrusive approaches in seen and unseen noisy environments are shown in the bottom portion of Table II. In general, the four deep-learning approaches significantly outperform the two unsupervised (i.e., noise PSD and speech PSD) and tree based (i.e., NISA) approaches across all metrics. It demonstrates that neural network methods are beneficial for speech assessment. It is noticeable that the results of SER-CL are superior to AutoMOS, Quality-Net, and CNN approaches. For PESQ prediction, the RMSE$^\star$ of SER-CL is 0.26, which is lower than the competitive CNN approach by 0.03 and Quality-Net by 0.04. The same trends occur when predicting STOI and SDR, that is, SER-CL always produces the lowest error and highest correlation across all cases. Specifically, the STOI MAE of Quality-Net and CNN are 0.05 and 0.04, respectively, which are both higher than the 0.02 of SER-CL. Quality-Net shows better performance (1.32 MAE and 1.16 RMSE$^\star$) over CNN in terms of SDR prediction, however, it is worse than SER-CL (1.07 MAE and 0.87 RMSE$^\star$). Also, SER-CL reduces the MAE and RMSE$^\star$ of STOI and SDR by half compared to AutoMOS. In Fig. 4, the predicted scores from SER-CL are plotted against the true scores for all three objective measures. A strong and almost linear correlation between the predicted and true values are observed.

### E. Reverberant condition

#### 1. Experimental data

Reverberation adversely affects speech quality and intelligibility because the sound reflections smear speech structure across time and frequency. The reverberant speech dataset used for evaluation in reverberant conditions is

https://doi.org/10.1121/10.0002702




(a) PESQ prediction (b) STOI prediction (c) SDR prediction

FIG. 4. (Color online) Scatter plots of the true ($x$ axis) and predicted scores ($y$ axis) of SER-CL under noisy conditions. (a) PESQ prediction. (b) STOI prediction. (c) SDR prediction.

generated with the TIMIT corpus and an imaging method (Habets, 2006). Specifically, five simulated rooms are selected for both training and testing since different room dimensions will produce diverse reverberation effects. The dimensions of the five seen rooms are $4 \times 3 \times 3$ m$^3$, $6 \times 5 \times 3$ m$^3$, $8 \times 7 \times 4$ m$^3$, $9 \times 8 \times 4$ m$^3$, and $10 \times 9 \times 5$ m$^3$, respectively. Another two unseen rooms are created to further test generalization. The dimensions of these rooms are $5 \times 4 \times 3$ m$^3$ and $7 \times 6 \times 4$ m$^3$.

RIRs are generated by placing the target speaker and microphone in random positions in the simulated rooms, where the distance between the speaker and microphone is fixed at 1 m. With this configuration, 10 RIRs at 3 $T_{60}$'s (i.e., 0.2, 0.4, and 0.6 s) are generated for each of the 5 training rooms, resulting in 150 training conditions. Two additional RIRs at the same 3 $T_{60}$'s are generated for each of the 5 seen testing rooms, to serve as the seen room but unseen RIR test cases. For the unseen room case, 2 RIRs using 5 different $T_{60}$'s (i.e., 0.1, 0.2, 0.3, 0.4, and 0.5 s) are generated for each of the 2 unseen testing rooms. In total, 30 conditions from seen rooms and 20 from unseen rooms are used for testing. Each of the 1500 TIMIT utterances is convolved with 10 random training RIRs, resulting in a set of 15 000 reverberant training signals. For testing, each of the 500 different utterances from 100 speakers are convolved with 5 random testing RIRs, resulting in 2500 testing speech signals. Thus, individual models are trained and tested for each objective metric using utterances from many speakers, multiple RIRs and rooms.

#### 2. Results and analysis

Table III reports the average performance of PESQ, STOI, and SDR prediction across seen and unseen reverberant testing conditions. The average across the seen and unseen sets are shown together for brevity, but the trends are similar in each case. SER-CL is evidently superior to the PSD-based and NISA approaches in all cases. When comparing with other deep learning approaches (i.e., AutoMOS, Quality-Net, and CNN), SER-CL achieves the best result in predicting all objective metrics. For PESQ prediction, the RMSE$^\star$ of SER-CL is 0.29, which is lower than the 0.40 of AutoMOS, 0.35 of Quality-Net, and 0.33 of CNN. Similar trends occur for STOI and SDR prediction. Specifically, the STOI RMSE$^\star$ of SER-CL is 0.03, while AutoMOS and Quality-Net scores are 0.10 and 0.05, respectively. The performance of the CNN approach is comparable to that of the SER-CL in STOI prediction, where the MAE and RMSE$^\star$ are both 0.04, while PCC is 93.7%, which is close to the 93.9% of SER-CL. In terms of SDR, the MAE, RMSE$^\star$ and PCC of SER-CL are 1.36%, 1.24%, and 92.4%, which surpasses the second place Quality-Net. The experimental results show the robustness of the two-stage approach in reverberant environments.

### F. Noisy-reverberant condition

Background noise and reverberation are both present in real world environments. To evaluate performance in this challenging scenario, we generate noisy-reverberant speech by

TABLE III. Comparison of different non-intrusive methods under reverberant conditions. The average performance across seen and unseen testing data is reported.

| | PESQ | | | STOI | | | SDR | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) |
| Noise PSD | 0.67 | 0.92 | 78.4 | 0.16 | 0.18 | 82.2 | 5.54 | 4.98 | 78.9 |
| Speech PSD | 0.56 | 0.64 | 77.5 | 0.12 | 0.15 | 81.4 | 4.41 | 4.12 | 77.4 |
| NISA | 0.42 | 0.56 | 84.1 | 0.10 | 0.11 | 85.9 | 3.55 | 3.98 | 81.0 |
| AutoMOS | 0.32 | 0.40 | 88.3 | 0.09 | 0.10 | 87.9 | 2.47 | 2.33 | 83.1 |
| Quality-Net | 0.27 | 0.35 | 90.4 | 0.06 | 0.05 | 91.5 | 1.73 | 1.60 | 90.5 |
| CNN | 0.26 | 0.33 | 90.3 | 0.04 | 0.04 | 93.7 | 1.86 | 1.83 | 88.8 |
| SER-CL | **0.21** | **0.29** | **90.8** | **0.03** | **0.03** | **93.9** | **1.36** | **1.24** | **92.4** |

combining different utterances with various RIRs and noises. Specifically, 150 pairs of RIRs are generated in 5 rooms for training. 30 pairs are generated in 5 seen rooms and 20 pairs in 2 unseen rooms, resulting in 50 RIR pairs for testing. The utterances and random cuts of noise are each convolved with the corresponding RIR from the pairs. 1500 utterances are mixed with one of 150 training RIRs and 10 types of noises ($1500 \times 10 = 15\,000$ training signals). The SNR in each case is randomly set to 0, 3, 6, and 9 dB. The testing dataset is generated by combining 200 different utterances with 50 testing RIRs and 10 noises ($200 \times 10 = 2000$ testing signals). Both proposed and comparison approaches are trained and tested on the same datasets. The average prediction errors under noisy-reverberant conditions are shown in Table IV. The results are broadly consistent with those of the two previous experiments. SER-CL's performance is superior to the other approaches in most cases, except for STOI prediction, where the RMSE$^\star$ of CNN is the same as SER-CL. The average error of PSD-based and NISA approaches are substantially higher than the proposed approach in all cases. SER-CL obtains a noticeable lower error than AutoMOS and Quality-Net across all metrics.

When comparing noise PSD, speech PSD and NISA to our two-stage approach, we find that they do not give reliable and consistent predictions, especially in extreme conditions, e.g., very low SNR, or severe reverberation. The errors of other competitive deep learning approaches in each case also have noticeable growth. On the contrary, the proposed approach performs well in several testing conditions. One reason is that our two-stage approach depends on speech-enhancement residuals, which are a reasonable indicator of degradation in objective quality and intelligibility. The other reason is that we use a stack of convolutional LSTMs to map the enhancement residuals to a target metric. The ConvLSTM unit bonds the local feature extraction ability of deep convolutional neural networks with the temporal consistency of LSTMs. Thus, our model outperforms the LSTM-based AutoMOS, the BLSTM-based Quality-Net, and the CNN-based approaches.

### G. Real-world conditions

To further test generalization capabilities and the importance of inputs to the second stage, we also consider a real-world dataset, namely, the conversational speech in noisy environments (COSINE) speech corpus (Stupakov *et al.*, 2009). COSINE captures multi-party conversations in real-world environments that contain background noise and interfering speakers, using multiple microphones. The recordings were captured indoors and outdoors, and included car engine sounds, birds, wind noise, street noise, and busy cafeteria conditions, to name a few. The recordings from the close-talking microphone and the body microphones (e.g., shoulder or chest) are used as the clean reference and distorted speech, respectively, when calculating the ground-truth objective scores. Two experiments are conducted to first compare alternative inputs to our second network stage, and to compare real-world performance against other well-performing approaches. In the first, we separately develop models that use the concatenations of (1) the estimated speech and residual; (2) the degraded (unprocessed) speech and residual; (3) the estimated speech, residual, and degraded speech; and (4) only the residual (e.g., proposed approach), as the input into the second metric-prediction stage. The model parameters and configurations in both stages are as previously defined. In the second experiment, we compare our system with three of the previously used approaches (e.g., AutoMOS, Quality-Net, and CNN) and a fourth deep-learning approach, NISQA (Mittag, 2019; Mittag and Möller, 2019), since it has recently been shown to perform well. Five thousand distorted signals from the COSINE corpus are used to train each model, and 1000 signals are used for testing. Since the COSINE corpus is recorded in everyday life environments, the training and testing conditions (e.g., noise, speakers, and environments) are not the same. In addition to PESQ, two additional metrics, ESTOI (Jensen and Taal, 2016) and SI-SDR (Le Roux *et al.*, 2019), are used as the training targets, to demonstrate the flexibility of our two-stage approach in predicting emerging objective measures. An additional performance evaluation metric, Spearman's rank correlation coefficient (SRC) is included, since it describes the monotonicity between the true and predicted scores.

We summarize the results of these experiments in Table V. The upper half of the table shows the average performance when applying different inputs to the second

TABLE IV. Comparison of different non-intrusive methods under noisy-reverberant conditions. The average performance across seen and unseen testing data is reported.

| | PESQ | | | STOI | | | SDR | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) | MAE | RMSE$^\star$ | PCC(%) |
| Noise PSD | 0.59 | 0.63 | 79.7 | 0.35 | 0.28 | 73.6 | 9.02 | 9.15 | 78.1 |
| Speech PSD | 0.56 | 0.58 | 81.9 | 0.28 | 0.25 | 77.8 | 7.58 | 8.32 | 77.3 |
| NISA | 0.55 | 0.53 | 83.2 | 0.18 | 0.16 | 82.5 | 7.82 | 6.53 | 79.5 |
| AutoMOS | 0.41 | 0.49 | 85.4 | 0.13 | 0.14 | 85.8 | 3.23 | 3.06 | 85.3 |
| Quality-Net | 0.32 | 0.39 | 89.5 | 0.07 | 0.07 | 90.1 | 2.49 | 2.37 | 88.1 |
| CNN | 0.29 | 0.34 | 89.2 | 0.06 | **0.05** | 92.8 | 2.65 | 2.54 | 86.9 |
| SER-CL | **0.24** | **0.30** | **90.1** | **0.04** | **0.05** | **93.1** | **1.73** | **1.66** | **91.2** |

https://doi.org/10.1121/10.0002702

TABLE V. Performance comparison of proposed models (upper half) and comparison approaches (lower half) using the real-world COSINE corpus.

| | PESQ | | | | ESTOI | | | | SI-SDR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE* | PCC | SRC | MAE | RMSE* | PCC | SRC | MAE | RMSE* | PCC | SRC |
| (est.+res.)-CLa | 0.27 | 0.30 | 0.80 | 0.74 | 0.07 | 0.09 | 0.76 | 0.74 | 3.02 | 3.11 | 0.78 | 0.74 |
| (deg.+res.)-CLa | 0.25 | 0.30 | 0.80 | 0.78 | 0.07 | **0.07** | 0.77 | 0.75 | 2.71 | **2.74** | 0.79 | **0.76** |
| (est.+res.+deg.)-CLa | 0.28 | 0.31 | 0.80 | 0.74 | 0.08 | 0.10 | 0.75 | 0.75 | 3.21 | 3.05 | 0.76 | **0.76** |
| SER-CL (only res.) | 0.26 | 0.32 | 0.81 | 0.76 | 0.07 | 0.09 | 0.76 | 0.75 | 2.91 | 2.95 | 0.76 | 0.74 |
| AutoMOS | 0.48 | 0.44 | 0.72 | 0.65 | 0.17 | 0.17 | 0.67 | 0.65 | 3.70 | 3.88 | 0.72 | 0.72 |
| Quality-Net | 0.34 | 0.46 | 0.76 | 0.72 | 0.13 | 0.11 | 0.74 | 0.73 | 3.30 | 3.45 | 0.76 | 0.74 |
| CNN | 0.36 | 0.41 | 0.78 | 0.72 | 0.12 | 0.12 | 0.72 | 0.72 | 3.21 | 3.24 | 0.80 | 0.75 |
| NISQA | **0.24**[a] | **0.29**[a] | **0.82**[a] | **0.79**[a] | **0.06** | **0.07** | **0.78** | **0.77** | **2.70**[b] | **2.83**[b] | **0.79**[b] | **0.75**[b] |

[a]Statistically significant results when compared to SER-CL, according to a two-tailed t-test with 95% confidence interval.
[b]Statistically insignificant results when compared to the (deg.+res)-CL model.

stage. The model that uses the concatenation of the degraded speech and residual [e.g., (deg.+res.)-CL] yields the lowest error in most cases, except PCC for PESQ prediction. For each alternative proposed model, the two-tailed t-test with a 0.05 significance level is conducted against the proposed SER-CL. There are no statistically significant differences between the means of each pair of predicted scores ($p$-value $> 0.1$). Since the alternative models do not show a substantial advantage over SER-CL in these experiments, we recommend that the residual only model is used when computational efficiency is required, since the input size is smaller, resulting in fewer computations. These results also confirm that the residual can be used to predict objective scores.

The comparison with other data-driven approaches is shown in the lower half of Table V. Not surprisingly, the performance of SER-CL on real-world COSINE data declined compared to the results on the simulated TIMIT dataset, but its drop is much smaller than that of AutoMOS, Quality-Net, and the CNN-based approach. NISQA, however, performs the best overall. The performance, however, is similar for PESQ prediction (not statistically significant), but NISQA performs slightly better than SER-CL for ESTOI ($p$-value $< 0.01$) and SI-SDR predictions ($p$-value $< 0.05$). When comparing NISQA to our (deg.+res.)-CL model, (deg.+res.)-CL performs better on average for SI-SDR, while NISQA performs better for ESTOI. The PESQ results are not statistically significant. Note that the results from the proposed approaches are statistically significant against the results from AutoMos, Quality-Net, and CNN, except for the few cases where the results are the same.

## IV. DISCUSSION

Our experimental results show that improving speech enhancement performance can noticeably reduce the prediction error of objective metrics. One of the advantages of our two-stage framework is that the enhancement algorithm can be replaced by any state-of-the-art approach, e.g., deep clustering (Hershey *et al.*, 2016), Wave-U-Net (Stoller *et al.*, 2018), or speech enhancement generative adversarial network (Pascual *et al.*, 2017). From the experiments of various simulated and real-world conditions, we notice that the distortions present in speech signals have a large impact on the accuracy and robustness of metric prediction. Therefore, the choice of the speech enhancement algorithm in the first stage depends on the overall performance improvement in the intended environments. In other words, a speech enhancement approach that has been proven to be successful for an intended type of distortion or environment should be a suitable enhancement approach for our two-stage system. If across environmental performance is not satisfactory, a per-environment optimal enhancement approach may be needed.

In previous experiments, the first stage and second stage are trained with the same dataset. To further test the generalization capability of the system on mismatched training datasets, we use the already-trained speech-enhancement model on TIMIT noisy data as the first stage, and the already-trained metric-score prediction model on COSINE data as the second stage, then test this combined two-stage system with the COSINE data. This experiment will determine the effectiveness of metric prediction, assuming the enhancement stage uses a pre-trained model from a different dataset. The results are shown in Table VI. The estimation errors (MAE and RMSE*) have a moderate increase when compared to the matched-stage training case, but the correlation coefficients drop noticeably compared to the model trained on the same dataset (i.e., SER-CL in Table V). It indicates that the mismatched training data between the two stages indeed has an impact on the prediction performance since unseen distortions are included. When comparing to the average performance of CNN and Quality-Net approaches in Table V, the mismatched SER-CL model shows comparable prediction errors (e.g., did slightly better

TABLE VI. Performance of the proposed SER-CL approach on mismatched training data.

| | MAE | RMSE* | PCC | SRC |
|---|---|---|---|---|
| PESQ | 0.37 | 0.40 | 0.70 | 0.69 |
| ESTOI | 0.12 | 0.15 | 0.63 | 0.65 |
| SI-SDR | 3.30 | 3.50 | 0.61 | 0.58 |

in PESQ while worse in ESTOI and SI-SDR), but worse correlation performance. The reason is that the first stage of SER-CL is trained from noisy TIMIT dataset, which did not include the same types of distortions or speakers, as the COSINE dataset, where the COSINE dataset also includes reverberation. It is possible that performance differences are also due to differences between the simulated and real data. Nevertheless, it shows that matched training is preferred.

The poor performance of certain approaches (e.g., Quality-Net) may be attributed to making frame-level score predictions, where each frame of the signal is given an utterance-level assessment score as a label, and the predicted score of the signal is obtained by averaging the estimated values over all frames. This is a major shortcoming, as frame-level (objective or subjective) scores ($\sim$ over millisecond length windows) are not the same as utterance-level scores ($\sim$ over 4–5 s), as the degree of distortion might vary much over this time period. Our approach overcomes this drawback as a single objective metric prediction is made for the utterance.

Our approach may potentially be used to alleviate the well-known metric discrepancy problem (Fakoor *et al.*, 2018; Fu *et al.*, 2018b), where the training objective function (e.g., MSE per frame) and performance evaluation metrics (e.g., PESQ, STOI, and WER) are mismatched. A model that obtains the lowest MSE during training does not necessarily achieve an improvement in speech quality and intelligibility. On the contrary, our model can potentially facilitate a multi-task learning framework, where the speech enhancement and objective metric prediction stages are trained simultaneously to improve the quality and intelligibility of the enhanced speech and lower the error of objective metric prediction.

Successfully predicting objective quality and intelligibility scores enables real-world testing. These objective metrics, however, may not be the gold standard, since they may not always strongly correlate with user sentiment, in all environments. Predicting human-provided mean-opinion scores (MOS) is more ideal, but a large-scale publicly available speech corpus that contains subjective ratings currently does not exist to facilitate human-level score prediction. We plan to conduct large-scale listening tests using real-world corpora, and evaluate the proposed two-stage system with human-level assessment in the future. In the meantime, objective-score prediction is still viable, since it will enable researchers to assess and analyze real-world performance, which can lead to algorithm improvements.

## V. CONCLUSION

In this paper, we propose a data-driven framework that uses deep learning to perform objective speech quality and intelligibility assessment. It is clear that environmental noise and reverberation play a dominant role in degrading the objective metrics of clean speech, so our idea is to use the amount of noise and reverberation distortions as an indicator of the scores of these objective metrics. Specifically, we propose a two-stage non-intrusive prediction model. In the first stage, degraded speech is passed to a speech enhancement system that separates the mixture into enhanced speech and residuals. The residuals contain mostly noise and reverberation components and can be regarded as a reasonable distortion feature for the next stage. Then the second stage uses the residuals to predict the objective metrics of mixture speech with a stack of four convolutional LSTM layers. The results show that our two-stage approach can accurately and reliably predict many objective metrics, e.g., PESQ, STOI, and SDR, when compared to other state-of-the-art methods, in noisy, reverberant, and combined testing environments.

In the future, we will investigate training the entire system in an end-to-end manner. We expect that jointly training both stages could be beneficial to each other, that is, using the second stage to help update the first stage can improve the performance of the speech enhancement task, in return, boosting the accuracy and robustness of the assessment prediction task.

Andersen, A. H., Haan, J. M., Tan, Z., and Jensen, J. (**2018**). "Nonintrusive speech intelligibility prediction using convolutional neural networks," IEEE Trans. Audio Speech Lang. Process. **26**, 1925–1939.

Chinaev, A., and Haeb-Umbach, R. (**2016**). "A priori SNR estimation using a generalized decision directed approach," in *Proceedings of INTERSPEECH*, pp. 3758–3762.

Dong, X., and Williamson, D. S. (**2018**). "Long-term snr estimation using noise residuals and a two-stage deep-learning framework," in *Proceedings of LVA/ICA*, Springer, pp. 351–360.

Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (**2015**). "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of ICASSP*, IEEE, pp. 708–712.

Erkelens, J., Hendriks, R., Heusdens, R., and Jensen, J. (**2007**). "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," IEEE Trans. Audio Speech Lang. Process. **15**, 1741–1752.

Fakoor, R., He, X., Tashev, I., and Zarar, S. (**2018**). "Constrained convolutional-recurrent networks to improve speech quality with low impact on recognition accuracy," in *Proceedings of ICASSP*, IEEE, pp. 3011–3015.

Falk, T. H., and Chan, W.-Y. (**2006**). "Single-ended speech quality measurement using machine learning methods," IEEE Trans. Audio Speech Lang. Process. **14**, 1935–1947.

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., and Scollie, S. (**2015**). "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," IEEE Sign. Process. Mag. **32**, 114–124.

Falk, T. H., Zheng, C., and Chan, W.-Y. (**2010**). "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," IEEE Trans. Audio Speech Lang. Process. **18**, 1766–1774.

Fu, S.-W., Tsao, Y., Hwang, H.-T., and Wang, H.-M. (**2018a**). "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proceedings of INTERSPEECH*.

Fu, S.-W., Wang, T.-W., Tsao, Y., Lu, X., and Kawai, H. (**2018b**). "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," IEEE Trans. Audio Speech Lang. Process. **26**(9), 1570–1584.

Gamper, H., Reddy, C. K., Cutler, R., Tashev, I. J., and Gehrke, J. (**2019**). "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proceedings of WASPAA*, IEEE, pp. 85–89.

Garofolo, J. S. (**1993**). "TIMIT acoustic phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

Habets, E. (**2006**). "Room impulse response generator," technical report, Vol. 2, p. 1.

Hendriks, R., Heusdens, R., and Jensen, J. (**2010**). "MMSE based noise PSD tracking with low complexity," in *Proceedings of ICASSP*, pp. 4266–4269.
Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (**2016**). "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP*, IEEE, pp. 31–35.
Hu, Y., and Loizou, P. C. (**2008**). "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio Speech Lang. Process. **16**, 229–238.
ITU (**2012**). "P.1401–methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union.
Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE Trans. Audio Speech Lang. Process. **24**(11), 2009–2022.
Kim, D.-S., and Tarraf, A. (**2007**). "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," Bell Labs Tech. J. **12**(1), 221–236.
Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., and Yoshioka, T. (**2016**). "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *Proceedings of EURASIP*, Vol. **2016**(1), p. 7.
Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (**2019**). "SDR–half-baked or well done?," in *Proceedings of ICASSP*, IEEE, pp. 626–630.
Loizou, P. C. (**2013**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton).
Malfait, L., Berger, J., and Kastner, M. (**2006**). "P.563–The ITU-T standard for single-ended speech quality assessment," IEEE Trans. Audio Speech Lang. Process. **14**, 1924–1934.
Martin, R. (**1993**). "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Third European Conference on Speech Communication and Technology*.
Mittag, G. (**2019**). "NISQA—Non-intrusive speech quality and TTS naturalness assessment," Quality and Usability Lab, TU Berlin, https://github.com/gabrielmittag/NISQA (Last viewed 11/13/2020).
Mittag, G., and Möller, S. (**2019**). "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proceedings of ICASSP*, IEEE, pp. 7125–7129.
Pascual, S., Bonafonte, A., and Serra, J. (**2017**). "SEGAN: Speech enhancement generative adversarial network," arXiv:1703.09452.
Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R. A., and Sculley, D. (**2016**). "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," NIPS Workshop.
Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, Vol. 2, pp. 749–752.
Salehinejad, H., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (**2017**). "Recent advances in recurrent neural networks," arXiv:1801.01078.
Seetharaman, P., Mysore, G. J., Smaragdis, P., and Pardo, B. (**2018**). "Blind estimation of the speech transmission index for speech quality prediction," in *Proceedings of ICASSP*, pp. 591–595.
Sharma, D., Wang, Y., Naylor, P. A., and Brookes, M. (**2016**). "A data-driven non-intrusive measure of speech quality and intelligibility," Speech Commun. **80**, 84–94.
Soni, M. H., and Patil, H. A. (**2016**). "Novel subband autoencoder features for non-intrusive quality assessment of noise suppressed speech," in *Proceedings of INTERSPEECH*, pp. 3708–3712.
Sørensen, C., Xenaki, A., Boldt, J. B., and Christensen, M. G. (**2017**). "Pitch-based non-intrusive objective intelligibility prediction," in *Proceedings of ICASSP*, pp. 386–390.
Srinivasan, S., Roman, N., and Wang, D. L. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Commun. **48**, 1486–1501.
Stoller, D., Ewert, S., and Dixon, S. (**2018**). "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," arXiv:1806.03185.
Stupakov, A., Hanusa, E., Bilmes, J., and Fox, D. (**2009**). "COSINE-a corpus of multi-party conversational speech in noisy environments," in *Proceedings of ICASSP*, IEEE, pp. 4153–4156.
Suhadi, S., Last, C., and Fingscheidt, T. (**2010**). "A data-driven approach to *a priori* SNR estimation," IEEE Trans. Audio Speech Lang. Process. **19**(1), 186–195.
Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.
Varga, A., and Steeneken, H. J. M. (**1993**). "NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.
Vincent, E., Gribonval, R., and Févotte, C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines* (Springer, Berlin), pp. 181–197.
Wang, Y., Han, K., and Wang, D. L. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio Speech Lang. Process. **21**, 270–279.
Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE Trans. Audio Speech Lang. Process. **22**, 1849–1858.
Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (**2017**). "Sequence-to-sequence models can directly translate foreign speech," arXiv:1703.08581.
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (**2015**). "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proceedings of LVA/ICA*, Springer, pp. 91–99.
Williamson, D. S., and Wang, D. L. (**2017**). "Time-frequency masking in the complex domain for speech dereverberation and denoising," IEEE Trans. Audio Speech Lang. Process. **25**, 1492–1501.
Williamson, D. S., Wang, Y., and Wang, D. L. (**2016**). "Complex ratio masking for monaural speech separation," IEEE Trans. Audio Speech Lang. Process. **24**, 483–492.
Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (**2015**). "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proceedings of NIPS'15*, pp. 802–810.
Zhang, Y., Chan, W., and Jaitly, N. (**2017**). "Very deep convolutional networks for end-to-end speech recognition," in *Proceedings of ICASSP*, IEEE, pp. 4845–4849.