



Non-intrusive speech quality assessment with attention-based ResNet-BiLSTM

Kailai Shen¹ · Diqun Yan¹ · Zhe Ye¹ · Xianbo Xu¹ · JinXing Gao¹ · Li Dong¹ · Chengbin Peng¹ · Kun Yang¹

Received: 16 October 2022 / Revised: 26 November 2022 / Accepted: 12 March 2023 / Published online: 10 April 2023
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

Abstract

Speech quality is frequently affected by a variety of factors in online conferencing applications, such as background noise, reverberation, packet loss and network jitter. In real scenarios, it is impossible to obtain a clean reference signal for evaluating the quality of the conferencing speech. Therefore, an effective non-intrusive speech quality assessment (NISQA) method is necessary. In this paper, we propose a new network framework for NISQA based on ResNet and BiLSTM. ResNet is utilized to extract local features, while BiLSTM is used to integrate representative features with long-term time dependencies and sequential characteristics. Considering that ResNet may result in the loss of context information when applied to the NISQA task, we propose a variant of ResNet which can preserve the time series information of the conferencing speech. The experimental results demonstrate that the proposed method has a high correlation with the mean opinion score of clean, noisy and processed speech.

Keywords NISQA · ResNet · BiLSTM · Attention mechanism · MOS

1 Introduction

In recent years, speech-based human–computer interaction and speech communication have grown commonplace due to the growth of the telecommunications industry and the popularity of speech communication technologies such as online conferencing. Especially since COVID-19 epidemic, many people have to work from home, and they can communicate and cooperate seamlessly through online conferencing software. During online conferencing, however, speech quality is frequently degenerated by a variety of factors, such as background noise, reverberation, packet loss and network jitter. For example, delays in speech quality can lead to a sharp decline in communication efficiency during an online conference. However, there is a high demand for speech quality assessment. This requires speech communication service providers to estimate the perceived speech quality to monitor the speech services offered to their customers [1].

Depending on the availability of the clear reference, speech quality assessment techniques can be categorized as

intrusive or non-intrusive [2]. Although intrusive methods are more accurate, the requirement of a clean reference limits them to many practical applications in realistic scenarios because clean reference is typically not present in quality of service (QoS) monitoring such as online conferencing applications. Without the intervention of human listeners, existing intrusive objective speech quality measurements can automatically assess the effectiveness of communication systems, but such methods require a reference signal with excellent sound quality. In contrast, it is necessary to investigate non-intrusive methods that simply need an evaluation signal to provide a quality score without the need for a sonically superior, unimpaired reference signal. It is consistent with the user's perception when using online conferencing applications and is more relevant for research.

A few indicators, such as signal-to-noise ratio (SNR) [3], signal-to-distortion ratio (SDR) [4] and signal-to-interference ratio (SIR) [5], have been used to estimate the speech quality. These indicators provide an objective measure of target speech. The ITU-T Recommendation P.808's MOS is the most widely used SQA indicator of user opinion. Using the absolute category rating (ACR) approach, a speech corpus is rated on a scale of 1–5 by human listeners. Even though MOSs are thought to be the most effective method for evaluating speech quality on subjective listening tests, they

✉ Diqun Yan
yandiqun@nbu.edu.cn

¹ Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China

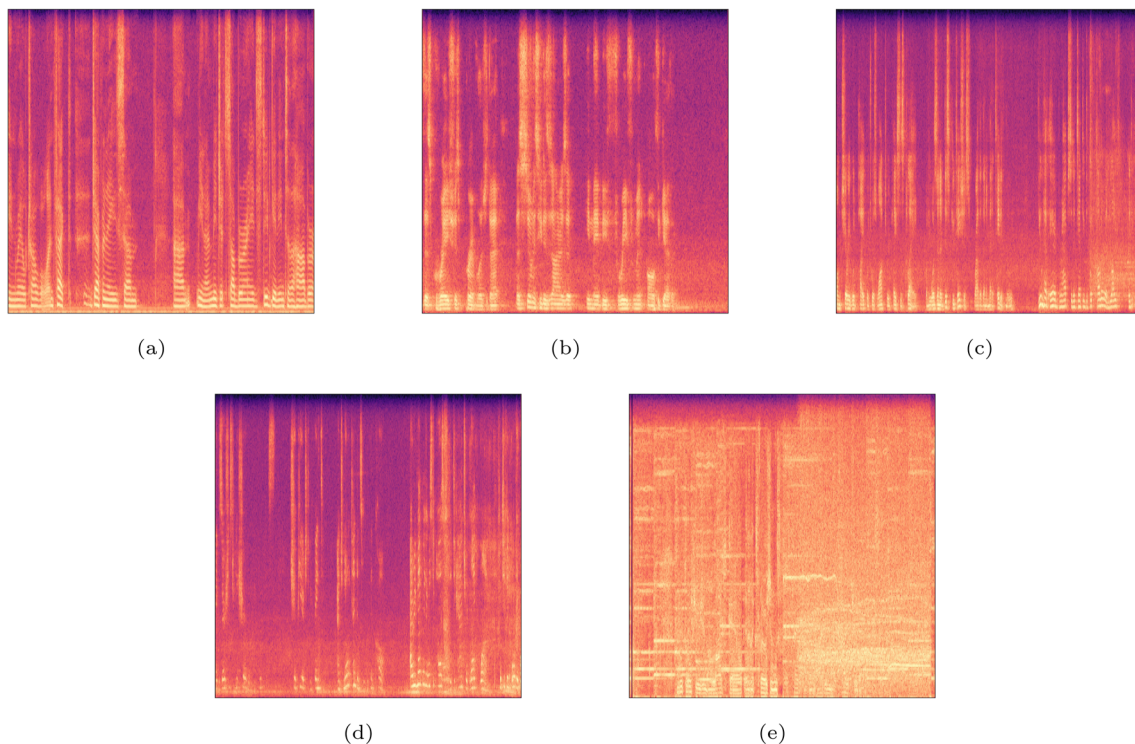


Fig. 1 Spectrograms of different MOS in **a** 5, **b** 4, **c** 3, **d** 2 and **e** 1

involve having a lot of participants do listening exams and giving perceptual evaluations. To address the aforementioned issues, numerous assessment models have been proposed [1, 6–15]. For example, Fu et al. [12] predicted frame by frame the quality of augmented speech by using a BLSTM-based quality assessment model called Quality-Net. Yoshimura et al. [9] proposed a fully connected neural network and CNN-based synthetic speech naturalness predictor to predict MOS. Lo et al. [13] proposed MOSNet to predict the quality of synthetic speech. Tseng et al. [14] utilizing self-supervised representations for MOS prediction. In order to make better use of the scores of each judge in the MOS dataset, Leng et al. [15] proposed a MOS predictor MBNet with a mean subnet and a bias subnet. It is important to note that different listeners may provide different ratings for the same corpus. As a result, it is still a challenge to develop a model with high relevance to the subjective human evaluation.

Inspired by the successful combination of CNN and RNN and the ResNet's powerful ability to extract local features, this paper introduces a non-intrusive speech quality evaluation method based on ResNet and BiLSTM. In addition, attention mechanisms are employed to focus on different parts of the input [16]. Specifically, the multi-head attention is used to obtain weights that can be scored according to the BiLSTM output, resulting in scores that are more relevant to human scores. In summary, the contributions of this work are as follows: (1) A variant ResNet model for non-intrusive

speech quality assessment is designed. In this model, a time invariant convolution is adopted to generate feature maps while keeping the time series information. (2) We propose an effective method based on ResNet-BiLSTM for non-intrusive speech quality assessment, which outperforms the state-of-the-art models in terms of accuracy on PSTN Corpus [17] and ITU-T P Supplement-23.

The remainder of this paper is organized as follows. Section 2 presents the proposed methods. The experimental results are discussed in Sect. 3. Section 4 provides an explanation of the discussions. Finally, the conclusions and recommendations for further research are summarized in Sect. 5.

2 Methods

2.1 Feature extraction

Spectral representation is the basis of digital signal processing. Magnitude spectrogram is commonly used to represent speech signals. Figure 1 shows the spectrograms at different MOS values. It can be found that there are differences in spectrograms with different scores. Thus, we used sequence of frame-based spectral feature to preprocess speech signals. For every 256 sample points, each speech file was subjected to a short-time Fourier transform (STFT), resulting in 257-

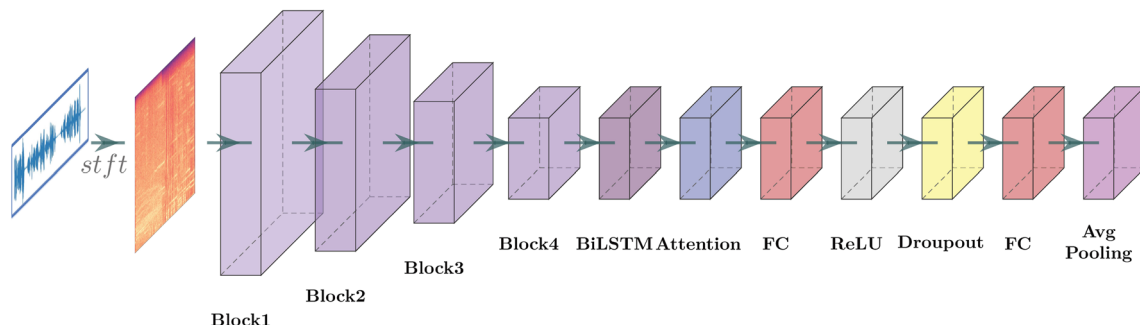


Fig. 2 Model architecture. The 257-dimensional 2D audio spectrogram as input into the four residual blocks. Subsequently, the features extracted by ResNet are sent to BiLSTM with attention. Finally two FC layers and an average pooling layer are used to obtain the final prediction

dimensional spectral features. The STFT of signal $x(t)$ is defined as follows:

$$\text{STFT}_x(t, f) = \int_{-\infty}^{+\infty} x(\tau)g(\tau - t)e^{-j2\pi f\tau} d\tau \quad (1)$$

where $g(t)$ is the window function. From the above formula, the STFT of signal $x(t)$ is the signal multiplied by an analysis window $g(\tau - t)$ centered on t . Multiplying $x(t)$ by the analysis window function $g(\tau - t)$ is equivalent to taking out a slice of the signal near the analysis time point t . For a given time t , $\text{STFT}_x(t, f)$ can be regarded as the spectrum at that time. Next, the magnitude spectrogram is calculated by taking the mode of $\text{STFT}_x(t, f)$, and the result is that the signal x .

2.2 Model architecture

The architecture of the proposed ResNet-BiLSTM model is shown as Fig. 2. Given a speech sample, first, we calculate the magnitude spectrogram x for the speech by STFT with librosa [18] tool. The next step is to obtain local feature maps by ResNet. To capture long-term dependencies in speech signal, bidirectional LSTM is used where input is x_t and output is h . The model performs better due to the self-attention mechanism [16]. To align the dimension of the learning function with a frame-level dimension, two fully connected layers are added. Finally, the utterance-level MOS is derived by average pooling. All the activation functions are rectified linear units (ReLU) [19]. Batch normalization (BN) [20] is applied to convolution operations. The architecture of ResNet-BiLSTM is detailed in Table 1.

2.3 ResNet

In time series data modeling, CNN is frequently employed and performs effectively [21, 22]. By stacking additional convolution layers, CNN increases the size of its receptive field. However, due to the disappearance or explosion of gradients, deeper models are not always better. Therefore, He et al.[23]

proposed ResNet in 2015, which contains a skip connection. ResNet training network output is defined as follows:

$$x_{l+1} = F(x_l) + x_l \quad (2)$$

where x_{l+1} and x_l denote the residual block output and input, and $F(x)$ is residual mapping function.

In this work, we propose an enhanced residual block that only requires a convolution operation along the time dimension to generate a feature map with time series information. The residual block for this work is shown in Fig. 3. For the SQA task, the information of the entire speech is useful, and the above operation preserves the integrity of the time series information, so it is beneficial to predict the speech quality. To extract local features for each local region in the input spectral features, ResNet utilizes a set of convolutional filters to generate feature maps. This is indicated as follows:

$$(h_k)_{ij} = (W_k \otimes x)_{ij} + b_k \quad (3)$$

where x stands for the input feature maps, W_k and b_k stand for the k th filter and bias, respectively, and $(h_k)_{ij}$ signifies the (i, j) element of the k th output feature map. The 2D spatial convolution operation is represented by the symbol \otimes .

There are four residual blocks, and each block has a different number of layers compared to ResNet-18 and ResNet-50. To minimize the number of the trainable parameters, we use fewer residual blocks in the proposed ResNet-BiLSTM. Each residual block is configured with the same number of layers.

A BN layer is added to each residual module. BN layer makes all samples in a minibatch associated. Therefore, the network will not generate a definite result from a given training sample, which means that the output of the same sample no longer depends only on the sample itself but also on the other samples in the same batch. And when the network takes a random batch, the whole network will not learn in this direction. Therefore, over-fitting is avoided to a certain extent.

Table 1 Configuration of ResNet-BiLSTM architectures

Layers	ResNet-BiLSTM
Input layer	<i>mag spectrogram</i> ($N \times 257$)
ResNet Block(1)	$\left\{ \begin{array}{l} \text{Conv2D} - \text{filters} = 16, \text{strides} = (1, 1) \\ \text{BatchNormalization}() \\ \text{Conv2D} - \text{filters} = 16, \text{strides} = (1, 1) \\ \text{Conv2D} - \text{filters} = 16, \text{strides} = (1, 3) \end{array} \right.$
ResNet Block(2)	$\left\{ \begin{array}{l} \text{Conv2D} - \text{filters} = 32, \text{strides} = (1, 1) \\ \text{BatchNormalization}() \\ \text{Conv2D} - \text{filters} = 32, \text{strides} = (1, 1) \\ \text{Conv2D} - \text{filters} = 32, \text{strides} = (1, 3) \end{array} \right.$
ResNet Block(3)	$\left\{ \begin{array}{l} \text{Conv2D} - \text{filters} = 64, \text{strides} = (1, 1) \\ \text{BatchNormalization}() \\ \text{Conv2D} - \text{filters} = 64, \text{strides} = (1, 1) \\ \text{Conv2D} - \text{filters} = 64, \text{strides} = (1, 3) \end{array} \right.$
ResNet Block(4)	$\left\{ \begin{array}{l} \text{Conv2D} - \text{filters} = 128, \text{strides} = (1, 1) \\ \text{BatchNormalization}() \\ \text{Conv2D} - \text{filters} = 128, \text{strides} = (1, 1) \\ \text{Conv2D} - \text{filters} = 128, \text{strides} = (1, 3) \end{array} \right.$
Recurrent layer	BiLSTM-128
FC layer	Attention FC-128 ReLU dropout(0.3) FC-1
Output layer	Average pool

The number of frames is denoted by N

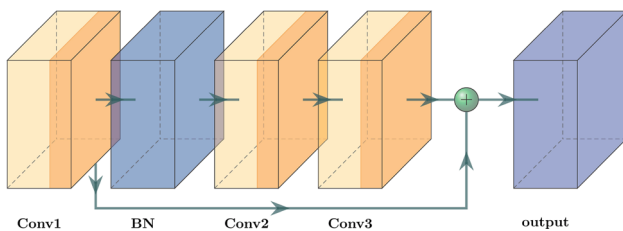


Fig. 3 Residual block used by the proposed model. Each block uses the output of the previous block as input

2.4 BiLSTM with attention

Long short-term memory (LSTM) [24], a type of temporal recurrent neural network, was developed to address the problem of long-term dependence that all RNNs share and which all RNNs have in the form of a chain of repeating neural network modules. For each layer, LSTM processes at time t by computing as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{4}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{6}$$

$$c_t = f_t c_{t-1} + i_t \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{7}$$

$$h_t = o_t \sigma_h(c_t) \tag{8}$$

where σ_g denotes sigmoid function, σ_c and σ_h denote hyperbolic tangent function, c is the internal cell states, h is the hidden states, and f , i and o are gates.

BiLSTM, which is composed of forward LSTM and backward LSTM, can better capture forward and backward information than LSTM which can only encode front to back information. The backward pass performs the opposite of the forward pass, turning all $t - 1$ to $t + 1$ in Eq. (3)–Eq. (7) to provide feature maps. Then we concatenate the outputs of forward and backward as follows:

$$h = [h_{fw} : h_{bw}] \tag{9}$$

where h_{fw} and h_{bw} represent the output of forward and backward, respectively.

Because speech is a kind of time series signal, it is closely related to its time sequence. Thus, the BiLSTM mechanism has been shown to be effective in speech quality assessment [12, 13, 25].

Utilizing attention mechanism has produced a number of outcomes in various domains [26, 27]. Consequently, we use self-attention mechanism to obtain weights that can be scored based on the BiLSTM output, where it takes the same input to derive the query Q_h , key K_h and value V_h pairs. This is defined as follows:

$$f_{\text{self-attention}}(Q_h, K_h, V_h) = \text{softmax} \left\{ \frac{Q_h K_h^T}{\sqrt{d_h}} \right\} V_h \quad (10)$$

where d_h is the dimension of Q_h, K_h, V_h . $Q_h = hW^Q$ is the query, which W^Q is the corresponding projection weight. Similar results can be obtained for K_h and V_h using the appropriate projection weights W^K and W^V .

2.5 Objective function

We regard MOS prediction as a regression task. The MOS evaluation of each dataset was used as the ground truth for training the model. Fu et al. and Lo et al. [12, 13] introduced frame-level prediction error to obtain a utterance-level prediction more related to human score. The objective function maps the final utterance-level MOS score by averaging the frame-level scores over each frame. Specifically, the ground-truth MOS is used for all frames in the speech utterance to calculate the frame-level MSE. And the frame-level MSE helps the model converge with better prediction accuracy [12]. Thus, we set the objective function as follows:

$$O = \frac{1}{S} \sum_{s=1}^S [(\hat{X}_s - X_s)^2 + \frac{1}{T_s} \sum_{t=1}^{T_s} (\hat{X}_s - x_{s,t})^2] \quad (11)$$

where \hat{X}_s is the truthful MOS, X_s is the predicted MOS for the s th speech sample, besides, T_s is the total number of frames in the s th speech sample, S is the total number of speech samples for training and $x_{s,t}$ is the frame-level prediction at time t .

3 Datasets

In this section, the available datasets are presented, some of which come from the datasets recommended with ConferencingSpeech 2022 [28]. Another set of narrowband datasets are taken from the ITU-T P Supplement-23.

3.1 ConferencingSpeech 2022

ConferencingSpeech 2022 challenges are proposed to stimulate the development of NISQA technology in online conferencing applications. This challenge will provide four

datasets: NISQA Corpus [29], PSTN Corpus, IU Bloomington Corpus [30] and Tencent Corpus¹. Of these, all three datasets are publicly available for the first time, with the exception of the NISQA corpus. It is worth noting that IU Bloomington Corpus uses ITU-R BS.1534 for subjective test, and the score range is 0–100 instead of 1–5. Therefore, it is not taken into account in the experiment.

NISQA Corpus This dataset is a publicly available SQA dataset containing more than 14,000 speech samples with simulated (e.g., codecs, packet loss, background noise) and lives (e.g., mobile phone, Zoom, Skype, WhatsApp) conditions.

PSTN Corpus This dataset comes from the public audiobook dataset Librivox and is filtered to get good quality speeches. It is then divided into 10-second segments and DNS Challenge 2021 is used to add noise. There are a total of 58,709 speech samples, of which 40,739 are based on noise reference files and 17,970 on clean reference files.

Tencent Corpus This dataset contains samples with and without reverberation. There are around 10,000 Chinese corpus without reverberation, and each speech exhibits the simulated degradation commonly encountered during online conferencing. A total around 4000 simulated impairment and live-recorded speech samples are considered in the case of reverberation.

3.2 ITU-T P supplement-23

Speech samples from the ITU-T P Supplement-23 were utilized in the characterization tests of the G.729 8 kbit/s codec. Ten datasets make up this corpus; however, only the seven that rated on the ACR scale are employed in this work. Each dataset contains coded speech with frame loss and noise circumstances and is scored in a narrowband context.

4 Experiments

4.1 Experimental settings

For the ResNet-BiLSTM, the dropout [31] rate is set to 0.3. Adam [32] Optimizer is used to train the model with a learning rate of 0.0001. Early stopping is used for the validation set's MSE with 10 epochs patience. We use Tensorflow [33] as the framework on an Nvidia RTX-2080Ti GPU to conduct our experiments.

We train and evaluate our model on four speech datasets with MOS labels (trained and evaluated on each individual datasets). And we use zero padding to ensure that all audio has the same length. Root mean squared error (RMSE), linear

¹ <https://github.com/ConferencingSpeech/ConferencingSpeech2022/tree/main/Training/Dev%20datasets>.

Table 2 Prediction results on NISQA Corpus with and without attention mechanism

ResNet-BiLSTM	RMSE	LCC	SRCC
<i>w/o</i> attention	0.5168	0.8938	0.8928
With attention	0.5068	0.8910	0.8900

correlation coefficient (LCC) and Spearman’s rank correlation coefficient (SRCC) are taken as the evaluation indicators. For more details, root mean square error (RMSE) is the mean square difference between the value that was actually observed and the value that the model anticipated. It can be obtained through:

$$RMSE = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - Y_i)^2} \tag{12}$$

where X_i indicates subjective score MOS, Y_i represents the predicted objective score MOS and N represents the total number of speech samples.

Linear correlation coefficient (LCC) describes the linear correlation between subjective score and algorithm score (normal distribution). It can be obtained through:

$$LCC = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \tag{13}$$

Spearman’s rank correlation coefficient (SRCC) calculates the correlation coefficient of the monotonic relationship between two variables. It can be obtained through:

$$SRCC = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \tag{14}$$

where N raw data X_i, Y_i are converted into hierarchical data x_i, y_i .

4.2 Impact of attention mechanism

Table 2 shows the RMSE, LCC and SRCC values of models with and without attention mechanism. Table 2 demonstrates that the performance of RMSE improved from 0.5168 to 0.5068 after the attention mechanism was used in the proposed ResNet-BiLSTM model.

4.3 Impact of reverberation

In the real world, non-intrusive speech quality assessment typically occurs while there is reverberation. In light of this, robustness of SQA method is a crucial requirement. Tencent Corpus contains speech examples with and without

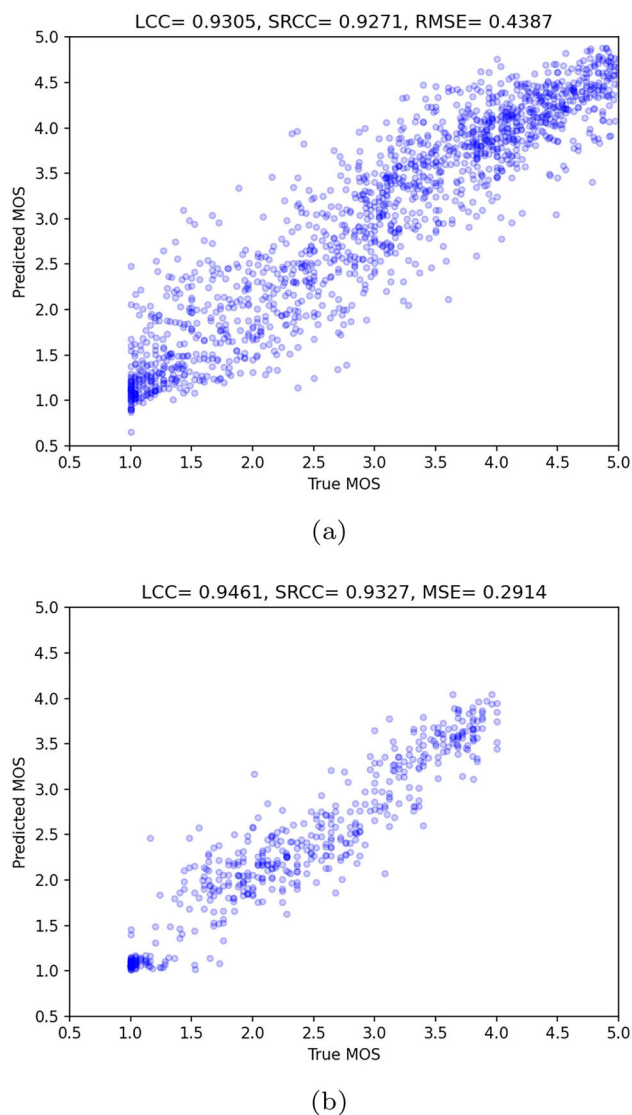


Fig. 4 Scatter plot of predictions on the Tencent Corpus in **a** without reverberation and **b** with reverberation

reverberation. Figure 4a, b shows the scatter plot predictions by ResNet-BiLSTM on Tencent Corpus without and with reverberation, respectively. Figure 5 shows that the proposed ResNet-BiLSTM model maintains good performance in the presence of reverberation and predictions are highly correlated with human evaluations.

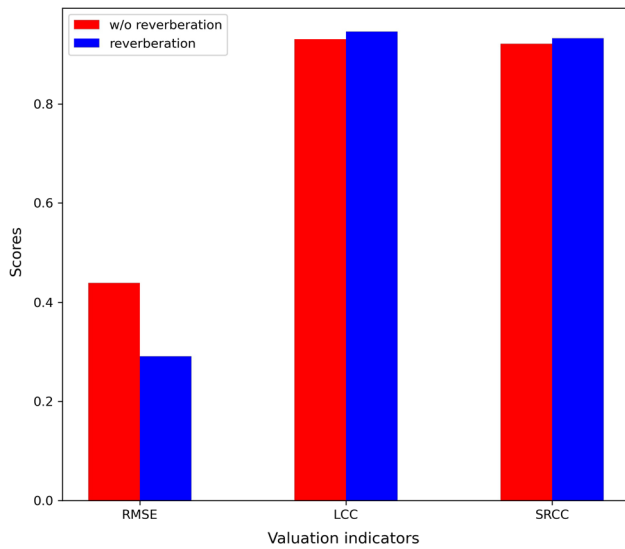
4.4 Comparison of different models

In this section, the effectiveness of various models is compared. Table 3 shows the RMSE, LCC and SRCC values of different models on PSTN Corpus and ITU-T P Supplement-23. From Table 3, ResNet-BiLSTM yielded 0.5127 in RMSE, 0.8087 in LCC and 0.8078 in SRCC, which are all superior to other models on PSTN Corpus. In addition, on ITU-T P Supplement-23, the LCC and SRCC performance of ResNet-

Table 3 Prediction results with different models

Models	PSTN			ITU-T P Supplement-23		
	RMSE	LCC	SRCC	RMSE	LCC	SRCC
P.563 [34]	0.9659	0.2196	0.2178	0.5502	0.7703	0.7553
Quality-Net [12]	0.5510	0.7745	0.7769	0.6820	0.7374	0.7460
MOSNet [13]	0.5299	0.7937	0.7924	0.5267	0.7825	0.7728
NISQA [29]	0.5191	0.7719	0.7700	0.5793	0.8122	0.7999
ResNet-BiLSTM (Ours)	0.5127	0.8087	0.8078	0.4921	0.8118	0.7977

Best performances are indicated with bold

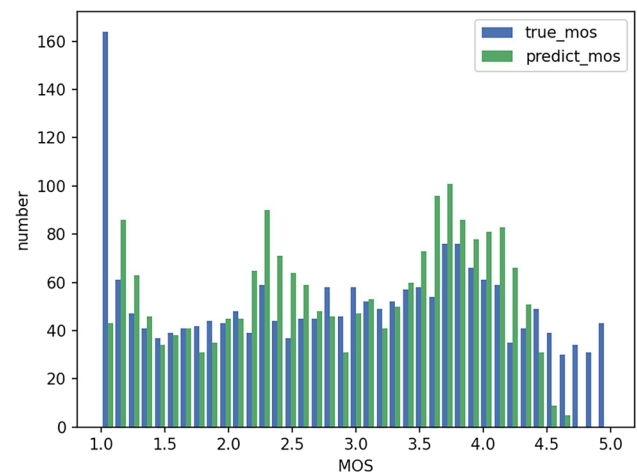
**Fig. 5** Histogram of the predictions on Tencent Corpus with and without reverberation

BiLSTM is comparable to that of NISQA, while RMSE is 6.5% lower than MOSNet.

5 Discussions

In general, ResNet is unsuitable for speech signals because it has a lot of trainable parameters and pooling processes, which may destroy contextual information of speech signals. In this work, we employ a variant ResNet which makes it effectively extract local feature maps while keeping the time series data with a manageable amount of trainable parameters. Therefore, it is possible to investigate various intricate CNNs in combination with RNNs to improve NISQA's accuracy even further.

The proposed ResNet-BiLSTM model has a limitation in that it rarely predicts MOS with low and high scores. It can be seen from Fig. 6, most of the predicted MOS from the ResNet-BiLSTM model are in the range of 1.5 and 4.5. The use of MSE-based objective function may cause this issue and it can be improved by modifying the objective function in the future research.

**Fig. 6** Histogram of the predictions of ResNet-BiLSTM

6 Conclusions

In this study, we investigated the ResNet and BiLSTM (ResNet-BiLSTM) combination for non-intrusive speech quality assessment tasks. And an attention mechanism is used to obtain weights that can be scored based on the BiLSTM output. Based on the results of the extensive evaluation of MOS for human perception, the experimental results show that the prediction score generated by our model is close to the human score. In the future, we will improve the structure of the network to improve the accuracy of the prediction scores to better serve online conferencing applications. In addition, we expect that the proposed model can also identify the causes of quality loss when giving MOS, and provide timely feedback to the service providers so they can adjust the call quality of real-time meetings in time to provide users with excellent usage experience.

Author Contributions Methodology and experiments were contributed by KS; formal analysis were contributed by JG, XX, ZY and KY; writing of the original draft was contributed by KS; and review and editing were contributed by DY, LD and CP.

Funding This work was supported by the National Natural Science Foundation of China (Grant No. 61300055, 61901237), Zhejiang Natural Science Foundation (Grant No. LY20F020010), Ningbo Science and

Technology Innovation Project (Grant No. 2022Z075), Ningbo Natural Science Foundation (Grant No. 202003N4089) and K.C. Wong Magna Fund in Ningbo University.

Data availability All accompanying data are provided in the manuscript.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical approval This declaration is not applicable.

References

- Falk, T.H., Chan, W.-Y.: Single-ended speech quality measurement using machine learning methods. *IEEE Trans. Audio Speech Lang. Process.* **14**(6), 1935–1947 (2006). <https://doi.org/10.1109/TASL.2006.883253>
- Union, I.T.: Single-ended method for objective speech quality assessment in narrowband telephony applications. *ITU-T Recommendation P. 563* (2004)
- Wong, J.W.: Signal-to-noise ratio (SNR). *Encycl. Neurosci.* 3690–3690 (2009)
- Weik, M.H.: *Signal-to-Distortion Ratio*. Springer, New York (2000)
- Rao, V., Rao, P., Tun, G.N.: Signal to interference ratio (SIR)-10 dB (2012)
- Cernak, M., Rusko, M.: An evaluation of synthetic speech using the PESQ measure. In: *Proceedings of Forum Acusticum*, pp. 2725–2728 (2005)
- Huang, D.-Y.: Prediction of perceived sound quality of synthetic speech. *Asia-Pacific Signal and Information Processing Association (APSIPA)* (2011)
- Remes, U., Karhila, R., Kurimo, M.: Objective evaluation measures for speaker-adaptive HMM-TTS systems. In: *Eighth ISCA Workshop on Speech Synthesis (SSW)* (2013)
- Yoshimura, T., Henter, G.E., Watts, O., Wester, M., Yamagishi, J., Tokuda, K.: A hierarchical predictor of synthetic speech naturalness using neural networks. In: *INTERSPEECH*, pp. 342–346 (2016)
- Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R.A., Sculley, D.: AutoMOS: learning a non-intrusive assessor of naturalness-of-speech. *arXiv preprint arXiv:1611.09207* (2016)
- Avila, A.R., Gamper, H., Reddy, C., Cutler, R., Tashev, I., Gehrke, J.: Non-intrusive speech quality assessment using neural networks. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 631–635 (2019). *IEEE*
- Fu, S.-W., Tsao, Y., Hwang, H.-T., Wang, H.-M.: Quality-Net: an end-to-end non-intrusive speech quality assessment model based on BLSTM. *arXiv preprint arXiv:1808.05344* (2018)
- Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., Wang, H.-M.: MOSNet: deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352* (2019)
- Tseng, W.-C., Huang, C.-Y., Kao, W.-T., Lin, Y.Y., Lee, H.-Y.: Utilizing self-supervised representations for MOS prediction. In: *INTERSPEECH*, pp. 2781–2785 (2021)
- Leng, Y., Tan, X., Zhao, S., Soong, F., Li, X.-Y., Qin, T.: MBNet: MOS prediction for synthesized speech with mean-bias network. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391–395 (2021). <https://doi.org/10.1109/ICASSP39728.2021.9413877>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017)
- Mittag, G., Cutler, R., Hosseinkashi, Y., Revow, M., Srinivasan, S., Chande, N., Aichner, R.: DNN no-reference PSTN speech quality prediction. In: *INTERSPEECH* (2020)
- McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: Librosa: audio and music signal analysis in python. In: *Proceedings of the 14th Python in Science Conference (SciPy2015)*, vol. 8, pp. 18–25. Citeseer (2015)
- Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: *International Conference on Machine Learning (ICML)* (2010)
- Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning (ICML)*, pp. 448–456 (2015)
- Yousefi, M., Hansen, J.H.L.: Block-based high performance CNN architectures for frame-level overlapping speech detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 28–40 (2021). <https://doi.org/10.1109/TASLP.2020.3036237>
- Chun, C., Jeon, K.M., Leem, C., Lee, B., Choi, W.: Comparison of CNN-based speech dereverberation using neural vocoder. In: *International Conference on Artificial Intelligence in Information and Communication (ICAIC)*, pp. 251–254 (2021). <https://doi.org/10.1109/ICAIC51459.2021.9415259>
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
- Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: *INTERSPEECH* (2012)
- Zhang, Z., Vyas, P., Dong, X., Williamson, D.S.: An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 316–320 (2021). <https://doi.org/10.1109/ICASSP39728.2021.9414182>
- Miao, H., Cheng, G., Gao, C., Zhang, P., Yan, Y.: Transformer-based online CTC/attention end-to-end speech recognition architecture. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6084–6088 (2020). <https://doi.org/10.1109/ICASSP40776.2020.9053165>
- Yang, C.-H., Qi, J., Chen, P.-Y., Ma, X., Lee, C.-H.: Characterizing speech adversarial examples using self-attention U-Net enhancement. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3107–3111 (2020). <https://doi.org/10.1109/ICASSP40776.2020.9053288>
- Yi, G., Xiao, W., Xiao, Y., Naderi, B., Möller, S., Wardah, W., Mittag, G., Culter, R., Zhang, Z., Williamson, D.S., Chen, F., Yang, F., Shang, S.: Conferencing Speech 2022 challenge: non-intrusive objective speech quality assessment (NISQA) challenge for online conferencing applications. In: *INTERSPEECH* (2022)
- Mittag, G., Naderi, B., Chehadi, A., Möller, S.: NISQA: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In: *INTERSPEECH*, pp. 2127–2131 (2021)
- Dong, X., Williamson, D.S.: A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals, pp. 4631–4635 (2020)
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)

33. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283 (2016)
34. Malfait, L., Berger, J., Kastner, M.: P. 563—the ITU-T standard for single-ended speech quality assessment. *IEEE Trans. Audio Speech Lang. Process.* **14**(6), 1924–1934 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.