Liu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:20
https://doi.org/10.1186/s13636-021-00209-4

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Neural network-based non-intrusive speech quality assessment using attention pooling function

Miao Liu[1], Jing Wang[1*], Weiming Yi[2] and Fang Liu[2]

Abstract

Recently, the non-intrusive speech quality assessment method has attracted a lot of attention since it does not require the original reference signals. At the same time, neural networks began to be applied to speech quality assessment and achieved good performance. To improve the performance of non-intrusive speech quality assessment, this paper proposes a neural network-based assessment method using attention pooling function. The proposed systems are based on the convolutional neural networks (CNNs), bidirectional long short-term memory (BLSTM), and CNN-LSTM structure. Comparing four types of pooling functions both theoretically and experimentally, we find the attention pooling function performs the best among the four. Experiments are conducted in a dataset containing various degraded speech signals with corresponding subjective quality scores. The results show that the proposed CNN-LSTM model using attention pooling function achieves state-of-the-art correlation coefficient (R) and root-mean-square error (RMSE) of 0.967 and 0.269, outperforming the performance of standardization ITU-T P.563 and autoencoder-support vector regression method.

Keywords: Speech quality assessment, Non-intrusive, Neural network, Attention pooling, CNN-BLSTM

1 Introduction

Speech quality assessment has become an important part of speech systems, which can be used to detect the quality of speech enhancement [1], speech synthesis [2], and other speech systems. Therefore, it is necessary to develop an effective, reliable, and flexible speech quality assessment method.

At present, the main challenge facing the speech quality assessment task is how to improve the prediction accuracy of non-intrusive methods to approach or even surpass the intrusive methods on the basis of objective evaluation. So far, P.563 [3] is the only published standard in ITU-T to evaluate no-reference speech quality. It was proposed relatively early and its accuracy is far from intrusive methods. With the rapid development of deep learning technology, many researchers have applied deep neural networks to speech quality assessment [4–7], which greatly improved the accuracy of non-intrusive methods. But none of them paid attention to the pooling function before the output of neural networks in speech quality assessment task.

In this paper, we propose a neural network-based non-intrusive speech quality assessment using attention pooling function [8]. We analyzed four existing pooling functions on speech quality assessment task and conducted experiments on the convolutional neural network (CNN), bidirectional long short term memory (BLSTM), and CNN-LSTM structure. The experiment results verified that the CNN-LSTM structure using the attention pooling function has a great performance on this task. As far as we know, this is the first analysis of the pooling function on non-intrusive speech quality assessment.

*Correspondence: wangjing@bit.edu.cn
[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
Full list of author information is available at the end of the article

The rest of this paper is organized as follows. Section 2 introduces the related works about speech quality assessment. Section 3 introduces the related neural networks. Section 4 presents the pooling function used in speech quality assessment system. Section 5 introduces the contrast methods. Section 6 introduces the experiment setups. Section 7 shows the evaluation experiments to assess the system performance. Finally, Section 8 states the conclusions.

2 Related works

The speech quality assessment methods contain subjective tests and objective tests. Subjective tests, based on listeners' feeling to the heard speech, generally use the mean opinion score (MOS) described in ITU-T P.800 [9] to measure speech quality. This method is accurate but time-consuming and labor-intensive. Objective evaluation methods can be divided into two categories (intrusive and non-intrusive) according to the presence or absence of reference signals. Intrusive method not only requires the speech signal to be evaluated, but also the original clean signal without damage, as introduced in ITU-T P.862 [10] and P.863 [11] . The non-intrusive method doesn't require a clean signal and directly provides a quality score based on the signal to be evaluated. The ITU-T P.563 [3] standard algorithm is one of non-intrusive methods, which is widely used in the evaluation of narrowband speech. Although the non-intrusive method is not as accurate as the intrusive method, it has developed rapidly in recent years because of its simple implementation.

Many non-intrusive evaluation algorithms of speech quality have been proposed. ANIQUE [12] is based on the functional roles of human auditory systems and the characteristics of human articulation systems. Tiago H. Falk et al. [13] used gaussian mixture models (GMMs) to model the behavior of clean speech and compared features extracted from degraded speech signals to the artificial reference model signals. D. -S. Kim et al. [14] proposed a perceptually motivated algorithm based on a temporal envelope representation of speech to assess speech quality. Meet H. Soni [15] used the ideal ratio mask (IRM) for non-intrusive quality assessment of noise suppressed speech. Wang [16] applied autoencoder to extract bottleneck features of speech signals and mapped the features to the predicted MOS using support vector regression (SVR) [17].

Recently, deep learning methods have been well applied in the field of speech quality assessment due to their nonlinear fitting performance. Haemin et al. [4] proposed a deep neural network (DNN) based non-intrusive speech quality estimation method in real-time voice communication systems. Hakami and Kleijn [5] used augmented feature set and the neural network to improve the prediction accuracy of the single-ended quality assessment approach. Quality-Net [6], based on bidirectional long short term memory (BLSTM), combined the frame-level scores to the final estimated utterance-level quality score using average pooling method. Lo et al. [7] adopted the convolutional and recurrent neural network models to build a mean opinion score predictor. Gabriel and Sebastian [2] proposed a TTS naturalness prediction model which achieved promising results on unseen datasets.

3 Related neural networks

Non-intrusive speech quality assessment can be regarded as a weak labeled regression task. Only the utterance-level speech quality labels will be provided. The non-intrusive speech quality assessment system based on neural network is shown in Fig. 1. Many neural network based methods such as convolutional neural networks (CNNs) and long short term memory (LSTM) have been used to predict the speech quality scores. In this section, we will introduce the related neural networks.

3.1 CNNs

CNNs were first proposed in image classification [18]. Compared with traditional back-propagation NN, CNNs use the local connectivity and weight sharing methods to retain important parameters and remove a large number of redundant parameters in order to achieve better learning results. Because of its outstanding ability to characterize shallow features, CNNs have been introduced to speech related tasks such as speech recognition [19] and speech quality assessment [7, 20]. A conventional CNN

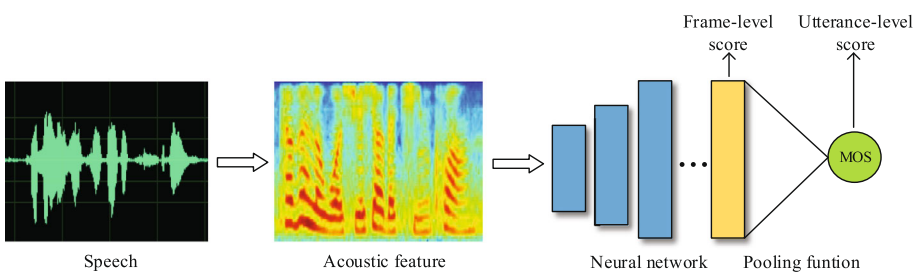

Fig. 1 Non-intrusive speech quality assessment system based on neural network

consists of convolution layers, pooling layers, and fully connected (FC) layers. When CNNs are used to process audio signals, the input data is often a two-dimensional or three-dimensional array. Before features are sent to the convolutional layer, they need to be normalized in time or frequency dimension. Each convolution layer is composed of multiple convolution kernels and each element of convolution kernels corresponds to a weight coefficient and a bias value. Convolution layers apply convolution operation to the input and pass the output to the next layer. The output of a convolutional layer is called feature maps. There are three main parameters of the convolution layer, namely the size of convolution kernels, the step size, and the padding. These three parameters determine the size of feature maps after the convolution operation. ReLU activation [21] is usually used to increase the nonlinearity of models. Recently, batch normalization [22] is adopted in CNN architectures after convolutional layers to stabilize training. Pooling layers can effectively reduce the size of feature maps, thereby reducing the number of parameters in networks. A time distributed fully connected layer is applied to the output of the last convolutional layer to predict the quality scores of each frames in time axis. Finally, the predicted scores are aggregated over the time axis to obtain the utterance-wise score.

3.2 LSTM

Speech is a continuous signal in time axis that changes according to the context. It is not correct to consider it at a single moment in time axis. CNNs cannot capture long time dependency in a speech utterance, while recurrent neural networks (RNNs) [23] are types of neural networks that can store history information in their hidden states and thus capture long-term dependency of sequential data. Therefore, RNNs are more conducive to modeling the time information of speech signals than CNNs. A problem with the traditional RNN is that it cannot distinguish whether the information from previous moments is useful. In other words, any information from previous moments will be passed down, which may cause the gradient disappear or explode in training. Long short term memory (LSTM) [24] is a variation of RNN. It can filter information from previous moments by forget gates, input gates, and output gates, which makes it possible to overcome the problem of long time dependency. Bidirectional long short-term memory (BLSTM) considers the past and future information at the same time when calculating, that is, the output is determined by the previous inputs and the following inputs. BLSTM is applied to our speech quality assessment systems.

3.3 CNN-LSTM

CNN-LSTM is a neural network structure that combines CNN and LSTM and it has been recently used for speech quality assessment [2, 7, 25]. In this structure, CNNs extract deep features of speech and the CNN feature vectors are then used as input for LSTM network that models time dependencies, which means that CNN-LSTM has the advantages of both CNN and LSTM.

4 Pooling functions

As shown in Fig. 1, in the neural network-based speech quality assessment systems, the role of pooling function is to aggregate frame-level quality scores into utterance-level quality scores. Therefore, the choice of pooling functions has a great influence on final results. However, most researchers only used the average pooling function [6, 7] and did not make further attempts. Pooling function has been extensively experimented and applied to the weakly labeled sound event detection task [8, 26–28], where linear softmax [27] and attention pooling function [8] achieve a strong performance. In this section, we will introduce and analyze the max, average, linear softmax, and attention pooling function in speech quality assessment system.

4.1 Definition of the pooling functions

Let $y_i \in [0, 5]$ be the prediction of a frame-level quality score at the ith frame and $y \in [0, 5]$ be the aggregated utterance-level score. We list the definitions of four pooling functions to be compared in Table 1.

The max pooling function takes the maximum of all frame-level scores y_i's as the utterance-level score y, which

Table 1 Definition and gradient of four pooling functions

Pooling function	Definition	Gradient
Max pooling	$y = \max_i y_i$	$\frac{\partial y}{\partial y_i} = \begin{cases} 1, i = \arg\max_i y_i \\ 0, \text{else} \end{cases}$
Average pooling	$y = \frac{1}{n}\Sigma_i y_i$	$\frac{\partial y}{\partial y_i} = \frac{1}{n}$
Linear softmax	$y = \frac{\Sigma_i (y_i)^2}{\Sigma_i y_i}$	$\frac{\partial y}{\partial y_i} = \frac{2y_i - y}{\sum_j y_j}$
Attention	$y = \frac{\Sigma_i y_i w_i}{\Sigma_i w_i}$	$\frac{\partial y}{\partial y_i} = \frac{w_i}{\sum_j w_j}, \frac{\partial y}{\partial w_i} = \frac{y_i - y}{\sum_j w_j}$

n is the number of frames in a utterance

Table 2 The specific information of the database

Conditions	Background noise
MNRU (Q = 5, 6, 10, 12, 15, 18, 24, 25, 30, 35, 36)	No background noise
MNRU (Q = 8, 14, 20, 26, 32)	Car noise: 15dB
G.729.1	No background noise
	Office noise: 40dB
	Babble noise: 40dB
	Babble noise: 128dB
	Car noise: 15dB
	Street noise: 20dB
G.711, Codec1, Codec2	No background noise
	Background music
	Office noise
	Babble noise
	Interfering talker
	Car noise
	Office noise: 40dB
G.729, G.729E, Codec3, Codec4, Codec5, Codec6	No background noise
	Car noise: 15dB

means that only the frame with the largest score will have an impact on the final utterance-level score.

The average pooling function [26] takes the average of all frame-level quality scores y_i's to get the utterance-level score y, which means it assigns an equal weight to all frames.

The linear softmax pooling function computes y as a weighted average of y_i's, where the weights are equal to the frame-level scores y_i's themselves. In this way, larger y_i's receive larger weights. Compared to average pooling, the utterance-level score is mainly determined by frames with the larger frame-level scores and the affect of frames with smaller scores will be reduced.

Finally, the attention pooling function is also a weighted average. Unlike linear softmax, the weights w_i for each frame are learnable and modeled by a dedicated layer in neural network. The utterance-level score y is then computed using the general weighted average formula of y_i's. The attention pooling function appears to be most favored by researchers because of its flexibility in sound event detection task [29, 30].

4.2 Analysis of the pooling functions

As stated before, we only have the utterance-level speech quality labels. When the overall quality of a speech utterance is good, listeners will give it a high score. But when only part of a speech utterance is bad, listeners will give a lower score. This means that a speech utterance with a high score should have high scores for each frame, and a speech utterance with a low score must have bad frames but may also have good frames. Based on this concept, we will analyze the gradient of the loss function w.r.t. the frame-level quality scores y_i's. And the weights w_i's also will be analyzed in the case of attention pooling.

Let $t \in [0, 5]$ be the utterance-level ground truth. The loss function we used is the mean squared error (MSE):

$$L = (y - t)^2 \tag{1}$$

The gradient of the loss function w.r.t. the utterance-level quality scores is represented as:

$$\frac{\partial L}{\partial y} = 2(y - t) \tag{2}$$

It does not depend on the choice of the pooling function. It is negative when the utterance-level predicted score is smaller than the utterance label ($y < t$) and positive when the utterance-level predicted score is larger than the utterance label ($y > t$).

According to the chain rule, we can get the loss function w.r.t. the frame-level scores y_i and the frame-level weights w_i respectively:

$$\frac{\partial L}{\partial y_i} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial y_i} \tag{3}$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial w_i} \tag{4}$$

We can divide it into two terms to analyze. The second item, $\partial y/\partial y_i$ (and $\partial y/\partial w_i$), is calculated for four pooling function in Table 1.

With the max pooling function, $\partial y/\partial y_i$ equals 1 for the frame with the largest score and 0 elsewhere. It will cause

Table 3 Configuration of different model structures. The symbol *C* indicates the number of convolutional kernels and the symbol - - - indicates that the model structure does not have this part

Model	BLSTM	CNN	CNN-LSTM
Input layer	Log mel spectrogram (bs ×800 frames ×64 mel bins)		
Convolution layer	- - -	$\left\{ \begin{pmatrix} 3 \times 3@C \\ \text{BN,ReLU} \end{pmatrix} \times 2 \right\} \times 4$ avg. pooling	
Recurrent layer	BLSTM-32	- - -	BLSTM-32
FC layer	FC-1, ReLU (frame-level score)		
Output layer	Max pooling, average pooling, attention pooling, linear softmax (utterance-level score)		

only this frame to receive a non-zero gradient during the back propagation of the network. Since we want to evaluate the utterance-level quality as a whole, it seems unreasonable that only the parameters related to the frame with the largest score are updated.

With the average pooling function, $\partial y/\partial y_i$ is always positive and equals $1/n$ for each frame-level score y_i, which means the gradient is distributed evenly across all frames. When the utterance predicted score is smaller than the utterance label ($y < t$), the gradient $\partial L/\partial y_i$ is negative, and this will boost the scores y_i's of all frames. This is in line with our requirement for good speech utterances to have good quality in each frame. When the utterance predicted score is larger than the utterance label ($y > t$), the gradient $\partial L/\partial y_i$ is positive and the scores of all frames have to be suppressed, which is not what we expected. It may cause the scores of many good frames to be incorrectly dropped.

With the linear softmax pooling function, $\partial y/\partial y_i$ is positive when $y_i > y/2$, and negative when $y_i < y/2$. When the utterance predicted score is smaller than the utterance label ($y < t$), the gradient $\partial L/\partial y_i$ is negative when $y_i > y/2$, and positive when $y_i < y/2$. As a result, larger y_i's will be boosted, while smaller y_i's will be suppressed. It is wrong to make good frames better and bad frames worse. When the utterance label is larger than the utterance predicted score ($y > t$), the gradient is positive when $y_i > y/2$, and negative when $y_i < y/2$. As a result, larger y_i's will be suppressed, while smaller y_i's will be boosted. This is different from what we expected for low-score speech.

With the attention pooling function, the second term $\partial y/\partial y_i$ is always positive because w_i is always larger than zero. Therefore, the attention pooling function will boost all frames when $y < t$ and suppress all frames when $y > t$. The strength of the boosting or suppression depends on the learned weight, which is different from the average pooling function. Because the weights w_i's are learned, we should also consider the gradient of the loss function w.r.t. the weights , $\partial y/\partial w_i$. The second term $\partial y/\partial w_i$ is positive where $y_i > y$, and negative where $y_i < y$. When the utterance predicted score is smaller than the utterance label ($y < t$), the gradient $\partial L/\partial w_i$ is negative when $y_i > y$, and positive when $y_i < y$. This will cause the weight w_i to rise where the frame-level score y_i is large and to drop where y_i is small, which means frames with larger scores

Table 4 Performance of the proposed systems on the test set

Stytems	R	RMSE
LSTM-max	0.961	0.283
LSTM-avgerage	0.953	0.365
LSTM-linear softmax	0.949	0.335
LSTM-attention	0.962	0.283
CNN-max	0.955	0.302
CNN-avgerage	0.959	0.370
CNN-linear softmax	0.961	0.317
CNN-attention	0.963	0.299
CNN-LSTM-max	0.964	0.273
CNN-LSTM-avgerage	0.965	0.283
CNN-LSTM-linear softmax	0.957	0.315
CNN-LSTM-attention	**0.967**	**0.269**

y_i's should get larger weights w_i's. It can help the weighted average result y to rise faster. When the utterance predicted score is larger than the utterance label ($y > t$), the weight w_i will rise where the frame-level score y_i is small and drop where y_i is large, which means that larger weights will concentrate upon frames with smaller scores. In the process of declining scores in all frames, this opposite phenomenon will cause scores of bad frames drop faster, but scores of good frames may avoid too much drop. This agrees with what we expected for low-score speech.

5 Contrast methods

5.1 ITU-T P.563

ITU-T P.563 [3] is a non-intrusive speech quality evaluation standard proposed by ITU-T. P.563 includes three main modules, simulation module, speech reconstruction module, and estimation module. The simulation module extracts the feature parameters using the principles of speech and auditory perception. The speech reconstruction module uses parameters extracted from distorted speech to reconstruct speech in order to generate quasi-pure speech. The role of estimation module is to determine the type of distortion and give evaluation scores according to the gap between the input speech and the generated quasi-pure speech. If the input is a severely disturbed speech signal, the difference between the input and output signal will be large and the quality score will be low. On the contrary, if the input is a clean speech signal, the quality score will be high.

5.2 Autoencoder-SVR

Autoencoder-SVR, proposed in [16], uses autoencoder to extract bottleneck features of speech signals and then maps the features to the predicted MOS using SVR. The method trains the autoencoder and SVR in turn at first. First, autoencoder is trained from training speech signals represented by the log-power spectra features. Then, the parameters of the autoencoder are fixed. Next, bottleneck features extracted from the well-trained autoencoder and the corresponding MOS values are used to train the mapping model SVR. Autoencoder-SVR is not an end-to-end trained model since autoencoder and SVR are trained separately. Therefore, its final performance depends on the two parts of autoencoder and SVR and its training process will be more complicated and difficult than end-to-end methods'.

6 Experiments

6.1 Database

We evaluate the proposed method on a narrowband MOS-labeled database including both clean and degraded speech signals, which come from subjective Chinese listening tests designed by Beijing Institute of Technology. All speech signals are processed from data in the NTT-AT Chinese corpus. The database consists of 1248 speech pairs with subjective MOS ranging from 1 to 5. All the speech utterances are sampled at 8 kHz rate with 16 bits resolution and in the length of 8 s. Six professional listeners scored each sentence in the professional acoustics laboratory. After each speech utterance is scored, the final MOS is the average of scores of the six individuals. In the whole corpus, the average variance of all scores for each speech utterance is 0.7. The database contains many processing conditions including different standard codecs, acoustic noise background, and modulated noise reference unit (MNRU) of various levels. Table 2 shows the detail of the conditions. The number of speech utterances in each condition with each background noise is 24. Without considering the background noise, approximately 90% speech files (1100 samples) under each distortion condition were randomly selected as the training set, while the remaining data (148 samples) is used for testing. In [16] and [31], Shan and Wang conducted experiments on this database and achieved some results.

6.2 Feature

We use log mel spectrogram as input feature following previous work on deep learning-based speech quality assessment [2]. The short time Fourier transform (STFT) with a Hanning window of 256 samples with a hop size of 80 samples is applied to extract spectrogram. We apply 64 mel filter banks on the spectrogram to obtain log mel spectrogram. The mel filter banks have a lower cut-off frequency of 50 Hz to remove low frequency noise. We use the torchlibrosa [32] package to build log mel spectrogram extraction.

6.3 Data augmentation

We use SpecAugment [33] as our data augmentation method to prevent systems from overfitting. SpecAugment, a simple data augmentation method, is applied to the feature inputs of a neural network. The augmentation policy consists of warping the features, masking blocks of frequency channels and masking blocks of time steps. In our speech quality assessment systems, SpecAugment is applied to the log mel spectrogram of a speech utterance using frequency masking and time masking. Frequency masking is applied so that f consecutive mel frequency bins $[f_0, f_0 + f)$ are masked, where f is chosen from a uniform distribution from 0 to a frequency mask parameter f', and f_0 is chosen from $[0, F - f)$, where F is the number of mel frequency bins [33]. More than one frequency mask can be applied to each log mel spectrogram. The frequency mask can improve the robustness of our systems to frequency distortion of speech utterances [33]. Time masking is

applied in the time domain, which is similar to frequency masking.

6.4 Model

The detailed configuration of different structures in our system, including CNN, BLSTM, and CNN-LSTM structure, is shown in Table 3. They have shown to perform well on speech quality assessment [7]. In CNN and CNN-LSTM structure, the convolution layer part includes 4 convolutional blocks. Each convolutional block consists of 2 convolutional layers with kernel sizes of 3×3. Batch normalization and ReLU function is applied after each convolutional layer. The convolutional block consists of 8, 16, 32, and 64 kernels, respectively. The symbol C following @ represents the number of kernels in Table 3. A 2×2 average pooling is applied after the first three convolutional blocks. A 1×8 average pooling is applied after the last convolutional block to average out frequency axis. In BLSTM and CNN-LSTM structure, BLSTM with 32 hidden states is applied in the recurrent layer part. Then, in three model structures, time distributed fully connected layer with ReLU function is applied to predict the quality score of each time frame. To obtain the utterance-level prediction for supervised learning, aggregation functions including max, average, linear softmax, and attention pooling along time frames are applied. For attention pooling function, a separate fully connected layer with softmax activation is used to generate the weights.

6.5 Training

In order to avoid experimental contingency, for each model structure, we trained ten models using the 10-fold cross-validation and got 10 corresponding results on the test set. We took the average of all the 10 results as the final result of each model structure.

During model training, we use the Adam [34] optimizer with the initial learning rate of 0.001. The learning rate is scaled by 0.1 times if there is no more decrease on the loss of validation set within 5 epochs and training stops if there is no more decrease on the loss of validation set within 20 epochs. The total number of training epochs is 80. The mini batch size is 32. The network was trained using the PyTorch toolkit.

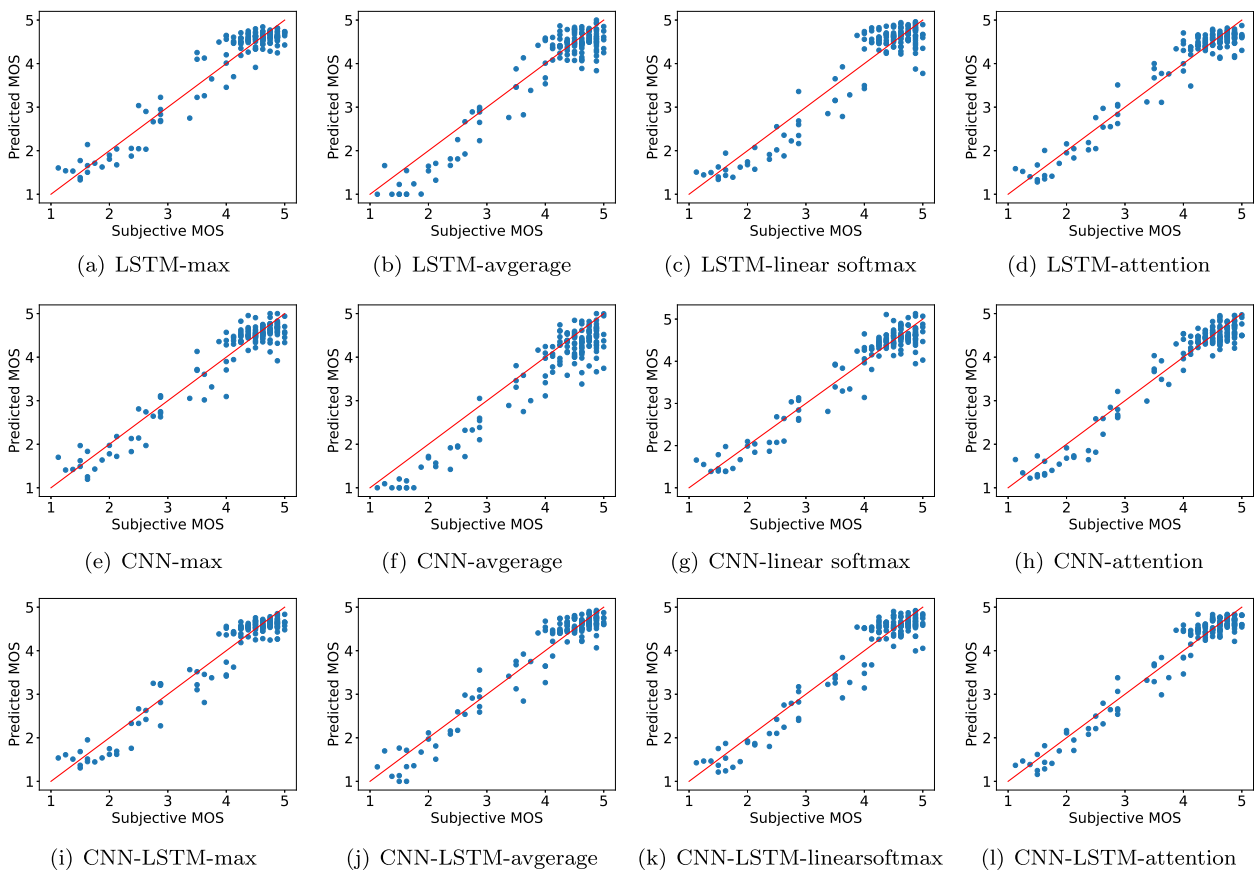

Fig. 2 Scattered plots of predicted MOS versus subjective MOS obtained from twelve models in Table 4

Table 5 Performance of different methods on the test set

Methods	R	RMSE
P.563	0.610	1.430
Autoencoder-SVR [16]	0.954	0.308
Cnn-LSTM-attention	**0.967**	**0.269**

6.6 Evaluation metrics

To evaluate the performance of systems, we use the correlation coefficient (R) and root-mean-square error (RMSE) between the predicted score S_k and the subjective score S'_k of each speech utterance k. The definition of correlation coefficient is as follows:

$$R = \frac{\sum_{k=1}^{N}(S_k - \overline{S})(S'_k - \overline{S'})}{\sqrt{\sum_{k=1}^{N}(S_k - \overline{S})^2 \sum_{k=1}^{N}(S'_k - \overline{S'})^2}} \tag{5}$$

where $\overline{S}$ is the average of S_k and $\overline{S'}$ the average of S'_k. N is the number of MOS labeled utterances in test set.

RMSE of MOS is defined as:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(S_k - S'_k)^2}{N}} \tag{6}$$

The R is the larger the better while the RMSE is the smaller the better.

7 Results and discussions

7.1 Comparison of different systems

Table 4 shows R and RMSE results for different model structures. Comparing the results of the combined experiment of the twelve models, we can find that the performance of attention pooling is better than the other three pooling functions regardless of the model structure. This shows that attention pooling has great robustness in different model structures. The CNN-LSTM structure is slightly better than CNN and LSTM structures as a whole because of its good learning both in time domain and frequency domain. The highest R of 0.967 and the lowest RMSE of 0.269 can be achieved by CNN-LSTM model using attention pooling function.

Figure 2 shows the scattered plots of predicted MOS versus subjective MOS of the test speech signals obtained from twelve models. The red diagonal line is the ideal situation that the objective MOSs are equal to the subjective MOSs. The blue dots represent the distribution of each test sample. Observing the alignment degree between data points and the diagonal line, we can see that the result distribution from the model using attention pooling function is closer to the diagonal line than that from the model using max, average and linear softmax pooling function.

7.2 Comparison of different methods

The results of different methods on the test set are shown in Table 5. On the one hand, the performance of the proposed method is much better than P.563, which means our proposed neural network-based non-intrusive assessment method has significantly improvement compared to traditional signal processing methods. On the other hand, the proposed method outperforms autoencoder-SVR method [16] with 1.4% relative increase in R and with 12.7% relative reduction in RMSE. This shows that our method has advantages over machine learning-based methods.

8 Conclusion

In this paper, we propose a neural network-based non-intrusive speech quality assessment using attention pooling function. We conduct experiments to compare four pooling functions among which attention pooling proved to be the best among them. From the experiment results, it can be seen that the proposed method has significant improvement in performance compared with the standardization ITU-T P.563 and autoencoder-SVR method. Specifically, the CNN-LSTM model using attention pooling function achieves the highest R of 0.967 and the lowest RMSE of 0.269. In the future, we will continue to research more on non-intrusive speech quality assessment methods considering the effects of different conditions and languages.

Abbreviations

MOS: Mean opinion score; GMMs: Gaussian mixture models; IRM: Ideal ratio mask; SVR: Support vector regression; DNN: Deep neural network; LSTM: Long short term memory; BLSTM: Bidirectional long short term memory; CNN: Convolutional neural network; MSE: Mean squared error; FC: Fully connected; RNN: Recurrent neural networks; MNRU: Modulated noise reference unit; STFT: Short time Fourier transform; R: Correlation coefficient; RMSE: Root-mean-square error; ReLU: Rectified linear units; BN: Batch normalization

Acknowledgements

We gratefully thank the reviewers and editors for their effort in the improvement of this work.

Authors' contributions

ML and JW conceived the algorithm. ML wrote the software, executed the experiments, and drafted the document. JW directed the project. ML and JW completed the final manuscript. All authors read and approved the final manuscript.

Funding

Support received from National Nature Science Foundation of China (No.62071039 and No.61620106002).

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Information and Electronics, Beijing Institute of Technology, Beijing, China. [2]Key Laboratory of Language, Cognition and Computation Ministry of Industry and Information Technology, School of Foreign Languages, Beijing Institute of Technology, Beijing, China.

Received: 29 November 2020 Accepted: 22 April 2021
Published online: 17 May 2021

References

1. Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(1), 229–238 (2008). https://doi.org/10.1109/TASL.2007.911054
2. G. Mittag, S. Möller, in *Proc. Interspeech 2020*. Deep learning based assessment of synthetic speech naturalness, (2020), pp. 1748–1752. https://doi.org/10.21437/Interspeech.2020-2382
3. I. T. Union, Single ended method for objective speech quality assessment in narrow-band telephony applications. ITU-T Recommendation P.563 (2004). Geneva
4. H. Yang, K. Byun, H. Kang, Y. Kwak, in *2016 IEEE International Conference on Digital Signal Processing (DSP)*. Parametric-based non-intrusive speech quality assessment by deep neural network, (2016), pp. 99–103. https://doi.org/10.1109/ICDSP.2016.7868524
5. M. Hakami, W. B. Kleijn, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Machine learning based non-intrusive quality estimation with an augmented feature set, (2017), pp. 5105–5109. https://doi.org/10.1109/ICASSP.2017.7953129
6. S.-w. Fu, Y. Tsao, H.-T. Hwang, H.-M. Wang, in *Proc. Interspeech 2018*. Quality-net: an end-to-end non-intrusive speech quality assessment model based on blstm, (2018), pp. 1873–1877. https://doi.org/10.21437/Interspeech.2018-1802
7. C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, H.-M. Wang, in *Proc. Interspeech 2019*. MOSNet: deep learning-based objective assessment for voice conversion, (2019), pp. 1541–1545. https://doi.org/10.21437/Interspeech.2019-2003
8. Q. Kong, Y. Xu, W. Wang, M. D. Plumbley, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio set classification with attention model: a probabilistic perspective (IEEE, Calgary, 2018), pp. 316–320. https://doi.org/10.1109/ICASSP.2018.8461392
9. I. T. Union, Methods for subjective determination of transmission quality. ITU-T Recommendation P.800 (1996). Geneva
10. I. T. Union, Perceptual evaluation of speech quality (PESQ): an objective method for end-to end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation P.862 (2001). Geneva
11. I. T. Union, Perceptual objective listening quality assessment (POLQA). ITU-T Recommendation P.863 (2011). Geneva
12. D.-S. Kim, Anique: an auditory model for single-ended speech quality estimation. IEEE Trans. Speech Audio Process. **13**(5), 821–31 (2005). https://doi.org/10.1109/TSA.2005.851924
13. T. H. Falk, Q. Xu, W.-Y. Chan, in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 vol. 1.* Non-intrusive GMM-based speech quality measurement, (2005), pp. 125–1281. https://doi.org/10.1109/ICASSP.2005.1415066
14. D.-S. Kim, Anique: an auditory model for single-ended speech quality estimation. IEEE Trans. Speech Audio Process. **13**(5), 821–831 (2005). https://doi.org/10.1109/TSA.2005.851924
15. M. H. Soni, H. A. Patil, in *2017 25th European Signal Processing Conference (EUSIPCO)*. Effectiveness of ideal ratio mask for non-intrusive quality assessment of noise suppressed speech, (2017), pp. 573–577. https://doi.org/10.23919/EUSIPCO.2017.8081272
16. J. Wang, Y. Shan, X. Xie, J. Kuang, Output-based speech quality assessment using autoencoder and support vector regression. Speech Commun. **110**, 13–20 (2019)
17. A. J. Smola, B. Schlkopf, A tutorial on support vector regression. Stats Comput. **14**(3), 199–222 (2004)
18. A. Krizhevsky, I. Sutskever, G. Hinton, in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Imagenet classification with deep convolutional neural networks (Curran Associates Inc., Red Hook, 2012), pp. 1097–1105
19. T. N. Sainath, O. Vinyals, A. Senior, H. Sak, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Convolutional, long short-term memory, fully connected deep neural networks, (2015), pp. 4580–4584. https://doi.org/10.1109/ICASSP.2015.7178838
20. A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, J. Gehrke, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Non-intrusive speech quality assessment using neural networks, (2019), pp. 631–635. https://doi.org/10.1109/ICASSP.2019.8683175
21. V. Nair, G. E. Hinton, *Rectified linear units improve restricted Boltzmann machines*. (Omnipress, Madison, 2010), pp. 807–814
22. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR. **abs/1502.03167** (2015). http://arxiv.org/abs/1502.03167. Accessed: 20 Sept 2020
23. T. Mikolov, M. Karafiát, L. Burget, J. Černocky, S. Khudanpur, Recurrent neural network based language model. Proc. 11th Ann. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2010. **2**, 1045–1048 (2010)
24. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
25. G. Mittag, S. Möller, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Full-reference speech quality estimation with attentional siamese neural networks, (2020), pp. 346–350. https://doi.org/10.1109/ICASSP40776.2020.9053951
26. A. Shah, A. Kumar, A. G. Hauptmann, B. Raj, A closer look at weak label learning for audio events. CoRR. **abs/1804.09288** (2018). http://arxiv.org/abs/1804.09288. Accessed: 22 Sept 2020
27. Y. Wang, J. Li, F. Metze, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling (IEEE, Brighton, 2019), pp. 31–35. https://doi.org/10.1109/ICASSP.2019.8682847
28. Y. Xu, Q. Kong, W. Wang, M. D. Plumbley, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Large-scale weakly supervised audio classification using gated convolutional neural network, (2018), pp. 121–125. https://doi.org/10.1109/ICASSP.2018.8461975
29. S. Hong, Y. Zou, W. Wang, M. Cao, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention, (2020), pp. 296–300
30. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans Audio Speech Lang Process. **28**, 2880–2894 (2020). https://doi.org/10.1109/TASLP.2020.3030497
31. Y. Shan, J. Wang, X. Xie, L. Meng, J. Kuang, in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Non-intrusive speech quality assessment using deep belief network and backpropagation neural network, (2018), pp. 71–75. https://doi.org/10.1109/ISCSLP.2018.8706696
32. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020). https://doi.org/10.1109/TASLP.2020.3030497
33. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: a simple data augmentation method for automatic speech recognition. Interspeech 2019 (2019). https://doi.org/10.21437/interspeech.2019-2680

34. D. P. Kingma, J. Ba, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, ed. by Y. Bengio, Y. LeCun. Adam: a method for stochastic optimization, (2015). http://arxiv.org/abs/1412.6980. Accessed: 19 Sept 2020

Publisher's Note