Objective speech quality assessment tools are vital for evaluating speech processing devices and algorithms. It is well-known that most of the existing objective speech quality assessment tools, such as perceptual evaluation of speech quality (PESQ), segmental signal-to-noise ratio, and log-spectral distance, rely on the knowledge of the clean speech, which is not always available. In this paper, we study to predict the PESQ score without clean speech as the reference by using Deep Neural Network (DNN). Both the regression-based and classification-based methods are considered in training the DNN model. In the regression training, all frame-based features extracted from each sentence are mapped to only one PESQ score, which is estimated by the PESQ tool recommended by ITU-T. In the classification training, the PESQ score ranging from -0.5 to 4.5 is equally divided into 100 subintervals, and then all frame-based features extracted from one sentence are mapped to one part of these subintervals. In the training phase, the TIMIT database are chosen as the clean speech signals, which are mixed with different types of noises at different levels of SNR to cover all the range of PESQ score. Experimental results show that the proposed DNN-based PESQ prediction is promising to provide a non-intrusive speech quality assessment for practical applications.

Keywords: DNN, PESQ, Speech quality

1. Introduction

Since the large-scale introduction of telephony networks, efforts have been made to guarantee high-quality and reliable services to human users [1]. Thus, a reliable estimation of speech quality is important to assess the speech devices or algorithms. The most reliable way to assess the speech quality is conducting listening tests, which serves as a subjective quality assessment measure. However, subjective tests in general are expensive and time consuming and thus it is an attractive topic to find an efficient objective quality assessment of speech which can correlate well with the subjective ones.

Objective speech quality assessment has attracted much attention in the last two decades, and these methods can be broadly classified as intrusive or non-intrusive depending on whether these algorithms need reference signals. Intrusive algorithms need both reference and degraded signals to estimate the distortion introduced by the system under test, while non-intrusive algorithms rely on the degraded signals only. In most real-world scenarios, the clean speech signals are not available, for example, recording the speech in noisy and/or reverberant environments, non-intrusive algorithms are more preferable.
Non-intrusive evaluation, which is also termed as no-reference or single-ended, is challenging since the measurement of speech quality has to be performed with only the output speech signal of the system under test. The P.563 Recommendation released by ITU-T describes an objective non-intrusive method for predicting the subjective quality of 3.1k Hz (narrow-band) telephony applications [2], but it cannot correlate well with the subjective MOS (Mean Opinion Score). Several non-intrusive methods using machine learning methods for evaluating the quality of audio signals have been proposed [3, 4]. The method described in [3] uses classification and regression trees (CART) to predict the speech quality score of the degraded speech based on handcrafted features. In [4], a neural network model that consists of an auto-encoder and linear regressor with augmented feature set as input has been proposed. But many of these methods do not generalize well on unseen noises and different signal-to-noise ratio (SNR) conditions.

In this paper, we proposed a non-intrusive method to predict the speech quality based on deep neural network. Our work only considers speech signals which have been degraded by the additive noise signals without any nonlinear processing. To make the DNN model generalize well on noises, the noise dataset used in our experiment includes over 100 different types of noises. The complementary feature set described in [5] is used as the DNN input and the related PESQ score of the degraded speeches is used as the DNN output. Also, both regression and classification methods are considered in training the DNN model.

The remainder of the paper is organized as follows. In Section 2, we present the proposed DNN-based speech quality prediction system. Experiments are provided in Section 3. Finally, some conclusions are presented in Section 4.

2. Algorithm description

This section describes the DNN system which is used to predict speech quality non-intrusively. At the beginning, we present an overview of the proposed system. We then describe the feature extraction and the DNN-based classification model. At the end, the DNN details are given.

2.1 System overview

A block diagram of the proposed speech quality prediction system is illustrated in Fig. 1. Our work aims to map one entire utterance to a speech quality score. Considering the different lengths of speech utterances, for the regression model, we map the frame features extracted from one sentence to the related speech quality score calculated by the PESQ algorithm. In the prediction stage, the speech features are processed by the well-trained DNN model to predict the speech quality scores frame by frame. Subsequently, the final estimated speech quality score is obtained by combining the frame scores through a global average. The classification model will be described in the Section 2.3.

![Figure 1: A block diagram of the proposed speech quality prediction system.](image-url)
2.2 Feature extraction

A complementary feature set is computed from the noisy speech signals [5]. This set includes amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) and mel-frequency cepstral coefficients (MFCC). All these features are extracted at the frame level and concatenated with the corresponding delta features. The STFTs in the feature extraction process are computed by dividing a signal into 20 ms time frames with 50% overlap between adjacent frames. Because the sampling rate of the speech is 16k Hz, the frame length equals 320 with the frame shift 160.

Temporal dynamics is employed to capture the correlations between adjacent frames of the feature set, \( F \). Specifically, we join adjacent frames into a single feature vector. The augmented feature vector, \( \hat{F} \), centered at the \( t^{th} \) time frame is as follows:

\[
\hat{F}(t) = [F(t - p), ..., F(t), ..., F(t + p)]^T
\]

where \( p \) (set to 5 empirically) denotes the number of adjacent frames to include on each side. The augmented feature set is then normalized to have zero mean and unit variance that will be used as the DNN input.

2.3 DNN-based classification model

In the DNN-based regression model, DNN is adopted as a mapping function from frame-features to frame quality scores. While for the classification model, DNN needs to map the features to different class labels. So the speech quality scores need to be allocated to different categories. To achieve this goal, we divide the PESQ score ranging from -0.5 to 4.5 into 100 subintervals equally and then number them from 1 to 100 accordingly where the interval numbers are used as quality score classes. Finally, the interval number is encoded to one-hot code, which is used for the DNN output in the classification model. For example, supposing one speech quality score is in the interval labeled as \( i \), then its one-hot code \( l_i \) is

\[
l_i = [0, 0, ..., 0, 1, 0, ..., 0] \in \mathbb{Z}^{100},
\]

where the one-hot code \( l_i \) contains 100 elements, but only the \( i^{th} \) is one, and others are all zeros. Given the class label \( i \), the estimated frame score \( y \) will be

\[
y = i \times 0.05 - 0.5.
\]

And given the quality score \( s \), the related class label \( i \) will be

\[
i = \begin{cases} 
1, & s = -0.5, \\
\left\lfloor \frac{s + 0.5}{0.05} \right\rfloor, & \text{otherwise.}
\end{cases}
\]

where the \( \left\lfloor x \right\rfloor \) rounds \( x \) to the nearest integer greater than or equals to it. For examples, \( \left\lfloor 4.0 \right\rfloor \) equals 4 and \( \left\lfloor 4.1 \right\rfloor \) is 5.

2.4 DNN structure

The DNN is trained to map a single frame of the augmented feature vector to a speech quality score. This is accomplished with a four layer DNN, where each of the hidden layers has 1024 units. Sigmoid activation functions are used in the hidden layer. For the regression model, the output layer contains one unit and uses linear activation function. For the classification model, the output layer contains 100 units and uses softmax function. The network structure of DNN is shown in Fig. [2].
For the DNN training process, the standard backpropagation algorithm coupling with dropout regularization [6] (with the dropout rate 0.2) is used to train the networks. We use the adaptive gradient descent [7] along with a momentum term as the optimization technique. A momentum rate of 0.5 is used for the first 5 epochs, after which the rate increases to and is kept as 0.9. And the cost function in classification training is cross-entropy, while for the regression model, mean square error function is chosen as the cost function.

3. Experiments

3.1 Database

In our experiments, the clean speech data was derived from the TIMIT database [8]. And the noise set contains 100 environment noises taken from [9] and other 36 noises from both the NOISEX-92 corpus [10] and Aurora2 database [11]. Among them, 112 noises were randomly chosen to synthesize the noisy speech signals for DNN training and the remaining noise signals were used for DNN testing. Each utterance in the training set (4620 utterances in total) of the TIMIT database was mixed with a noise randomly chosen from the above-mentioned 112 noises at a certain SNR level. In addition, 13 SNR levels (from -30 dB to 40 dB with steps of 5 dB) were used in the noisy speech utterance synthesis. This resulted in a collection of about 54 hours of noisy training data (including one condition of clean speech utterances) prepared to train the DNN. And the test set was synthesized with the TIMIT test set in a similar way, which contains 1680 noisy speech utterances for each SNR level.

3.2 Metrics

To evaluate the performance of our DNN model, three common criteria are used: Pearson linear correlation coefficient (PCC), Spearman rank order correlation coefficient (SRCC) and Root Mean Squared Error (RMSE), computed between the predicted and true PESQ scores. Besides, the "Bin Error" mentioned in [3] is also used to analyze our predicted results, which evaluates the absolute mean residual error between the true and estimated PESQ scores in bins of size 0.3.
As defined in [12], PCC is given by
\[ PCC = \frac{\sum_{i=1}^{N} (s_i - \bar{s})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (s_i - \bar{s})^2 \sum_{i=1}^{N} (y_i - \bar{y})^2}}, \]
where \( \bar{s} \) is mean of the PESQ score \( s_i \), \( \bar{y} \) is mean of predicted quality score \( y_i \) and \( N \) denotes the number of test samples.

SRCC is defined as
\[ SRCC = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \]
where \( d_i \) is the difference between the ranks of \( i^{th} \) speech sample in PESQ and predicted scores.

RMSE is given by
\[ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - s_i)^2}. \]

Before describing the Bin Error, a collection \( A = \{ y_i \mid |y_i - s_i| \leq \alpha \} \) and \( \alpha \in \{0.3, 0.6, 0.9\} \) need to be defined. Then, the Bin Error can be given by
\[ Bin Error \leq \alpha : \frac{M}{N} \times 100\%, \]
where \( M \) is the number of elements in \( A \). It should be noted that a better objective quality metric has higher PCC, SRCC and Bin Error values but lower RMSE value.

### 3.3 PESQ estimation results

#### 3.3.1 Effects of regression and classification methods

The test noisy speech utterances used in this experiment contains 10 SNR levels (from -5 dB to 40 dB with steps of 5 dB) with time length about 3 s. And all of the noisy signals were used to test the regression model and the classification model separately. The results are shown in Table 1. As shown, we can observe that both the regression model and the classification model can generate results which correlate well with the PESQ scores, but the classification model is able to predict quality scores with less root mean square error than the regression model. The values of Bin Error further explain the findings: the classification model is able to generate more accurate quality scores than the regression model. For the absolute bias less than 0.3, the predicted quality scores of classification model takes 76% percentage of overall test set results, which is about 5% more than the regression model. From the above, it can be concluded that the classification model performs better than the regression model in the speech quality score prediction experiment. And in the following subsections, only the classification model is taken into consider.

Fig. 3 shows the overall PESQ estimation results on the whole test set. The X-axis denotes the mean of true PESQ scores of the test samples in each subinterval mentioned in Section 2.3 and the Y-axis is the mean of predicted scores of related samples. It can be seen that the predicted scores correlate well with the PESQ. Few test samples locate in the interval \([0.5, 1]\), causing larger bias for this part.

#### 3.3.2 Model performances on unseen SNRs

In this experiment, the test dataset was synthesized under nine SNR levels (from -3 dB to 37 dB with steps of 5 dB) which do not appear in the DNN training dataset, resulting in \( 1680 \times 9 \) noisy sentences,
Table 1: Effects of regression and classification method.

<table>
<thead>
<tr>
<th>Method</th>
<th>PCC</th>
<th>SRCC</th>
<th>RMSE</th>
<th>Bin Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>Regression</td>
<td>0.95</td>
<td>0.95</td>
<td>0.29</td>
<td>71.2%</td>
</tr>
<tr>
<td>Classification</td>
<td>0.95</td>
<td>0.94</td>
<td>0.27</td>
<td>76.1%</td>
</tr>
</tbody>
</table>

Figure 3: Results for PESQ prediction on test set.

to assess the model robustness. The results are shown in Table 2. In this table, one can see that all of the assessment metrics are comparable to the findings in Section 3.3.1. This demonstrates that our DNN-based classification model generalize well on unseen SNRs.

We also evaluated our DNN model on the whole test clean database, and the mean of predicted quality scores is 4.31 with standard deviation 0.13. Fig. 4 shows a clean speech utterance and two noisy speech signals (with SNR 15 dB and 10 dB, respectively) as well as their respective frame-level quality assessments by our DNN model. It is clear that most of the frames in the clean speech utterance are regarded as clean and are labeled with 4.5. As for the noisy speeches, the frame scores concentrate upon the mean score.

Table 2: Prediction results on unseen SNRs.

<table>
<thead>
<tr>
<th>PCC</th>
<th>SRCC</th>
<th>RMSE</th>
<th>Bin Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>0.96</td>
<td>0.95</td>
<td>0.28</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

3.3.3 Effects of time lengths

The time length of noisy speech utterance in Section 3.3.1 is about 3 s. In this section, we extended the time length of test signals to 10 s and 20 s by randomly splicing 3 or 6 clean speeches in the TIMIT
Figure 4: Example of (a) clean speech, and (d) its corresponding frame-level quality assessment by DNN model; (b) noisy speech with SNR 15 dB, PESQ score 2.24 and (e) its corresponding frame-level quality assessment, estimated score 2.28; (c) noisy speech with SNR 10 dB, PESQ score 2.25 and (f) its corresponding frame-level quality assessment, estimated score 2.54.

test database separately. The test noisy speech signals were synthesized with 6 SNR levels (from -5 dB to 20 dB with step of 5 dB) containing 1000 sentences for each SNR. And these noisy signals were used as test database in this experiment. The results are shown in Table 3. As shown, with time length of noisy speech increasing from 3 s to 20 s, the PCC improves 0.04 and the Bin Error less than 0.3 improves about 7%. The findings show that our trained DNN model can better predict the speech quality with a longer duration of the (noisy) speech.

Table 3: Prediction results of speeches with different time lengthes.

<table>
<thead>
<tr>
<th>Time lengths</th>
<th>PCC</th>
<th>SRCC</th>
<th>RMSE</th>
<th>Bin Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>3s</td>
<td>0.91</td>
<td>0.93</td>
<td>0.29</td>
<td>75.1%</td>
</tr>
<tr>
<td>10s</td>
<td>0.94</td>
<td>0.95</td>
<td>0.25</td>
<td>81.0%</td>
</tr>
<tr>
<td>20s</td>
<td>0.95</td>
<td>0.96</td>
<td>0.24</td>
<td>82.4%</td>
</tr>
</tbody>
</table>

4. Conclusions

This paper proposed a non-intrusive and end-to-end DNN-based speech quality evaluation method. Both the regression and the classification methods are considered in training the DNN model. Our experi-
mental results show that the DNN-based classification model performs better than the regression model in the speech quality score prediction. And the well-trained DNN model can generalize well on the unseen SNR levels and the unseen noisy signals. The experimental results also show that the DNN model can yield a high correlation to PESQ and it is recommended to evaluate the speech quality with the speech that has its duration over 20 seconds.

ACKNOWLEDGMENTS

This work was supported by the National Science Fund of China (NSFC) under Grant 61571435 and 61801468.

REFERENCES


