INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A no-reference speech quality assessment method based on neural network with densely connected convolutional architecture

Wuxuan Gong[1], *Jing Wang*[1], *Yitong Liu*[1], *Hongwen Yang*[1]

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China

gongwuxuan@bupt.edu.cn, wangjing18@bupt.edu.cn, liuyitong@bupt.edu.cn, yanghong@bupt.edu.cn

Abstract

Most speech quality assessment methods require a perfect reference signal to evaluate the damaged speech's quality. However, it is challenging to obtain clean reference signals due to various types and levels of noise in reality. Meanwhile, no-reference speech quality assessment is less accurate than full-reference method. To address these issues, we propose a novel no-reference speech quality assessment model that improves evaluation accuracy with lower complexity. The model is primarily composed of three densely connected convolutional (DCC) modules and a bidirectional long short-term memory (BLSTM) module. Experiment results demonstrate that our method outperforms the baselines, achieving state-of-the-art on the no-reference speech quality assessment task. When using PESQ as optimization targets, the MSE, PLCC and SRCC reach 0.0389, 0.9695 and 0.9715, whereas when using STOI, these metrics reach 0.0019, 0.9608, and 0.9630, respectively.

Index Terms: speech quality assessment, deep learning, neural network, non-intrusive, objective assessment

1. Introduction

Speech quality assessment plays an crucial role in speech-related fields as speech degraded by environmental noise, electronic noise and other factors can drastically impair human intuitive perception. Over the years, different methods have been proposed to convert the implicit auditory perception into explicit scores to quantify speech quality. For the subjective auditory perception methods, professionals give speech quality assessment results through direct listening to the speech audio. Common subjective speech assessment methods, including mean opinion score (MOS) [1], ABX Test and so on, are able to get accurate and reasonable results at the cost of time and labor. By contrast, objective quality calculation usually uses mathematical methods to calculate different parameters related to speech quality and then come to quality scores, such as perceptual evaluation of speech quality (PESQ) [2], short-time objective intelligibility (STOI) [3], perceptual objective listening quality analysis (POLQA) [4], ITU-T P.563 [5], etc. Most of the assessment methods mentioned above are fully referenced, which means they require a clean reference speech. However, clean reference speeches are not always available in real speech testing scenarios, and the testing process is also hard to replicate. Therefore, no-reference assessment methods are more suitable for this situation. Unfortunately, the accuracy and generalization of such methods are not always optimal [6].

In recent years, deep learning has been widely utilized in speech quality assessment. Since the backpropagation mechanism can automatically optimize parameters along the gradient in differentiable mathematical models, human can construct more complex neural networks. Consequently, it is possible to further improve the accuracy of speech quality assessment algorithms under no-reference conditions. Yoshimura et al. [7] used a fully connected neural network and a convolutional neural network (CNN) to build a model. They took manually extracted features as input to train the model and found the correlation between the prediction results of the model and the MOS scored by humans in their experiments. Fu et al. [8] proposed a deep learning model based on bidirectional long short-term memory (BLSTM) network to extract the serialized frequency domain features of the speech and predict the PESQ score of the speech without reference. The high correlation between the prediction score and the PESQ score are demonstrated in the experimental results. It has proved the effectiveness and feasibility of the network based on the BLSTM structure for the speech quality assessment task. CNN and BLSTM have been combined in the work of Chen et al. [9]. They took the short-time Fourier transform (STFT) of the speech signal as the input feature, and then trained the model to learn the correlation between the speech and the MOS score. Subsequently, STOI-Net [10], HASA-Net [11], MOSA-Net [12] and NISQA [13] introduced the attention mechanism to CNN-BLSTM based network, trained and applied the model to different speech assessment tasks. At the same time, [14–18] to employ the neural network of temporal convolutional network (TCN) structure to directly take the time domain information as the input feature to optimize the speech's quality score.

Currently, many deep learning-based speech quality assessment models use the traditional CNN structure, which consists of serial connection convolution layers. However, this structure struggles to efficiently extract high-dimensional and low-dimensional features simultaneously and may encounter issues such as gradient disappearance or explosion. To address this problem, we propose a densely connected convolutional (DCC) structure that aims to improve feature extraction performance by mixing high-dimensional and low-dimensional features. Then, based on the DCC module, we add BLSTM to fuse the timing features, and then extracted the quality score through a fully connection layer and a average pooling layer. After that, we train and verify our model on speech with multiple types and levels of impairments under different optimization targets. Our results show that our proposed model is able to effectively learn multi-dimensional feature representations and achieve accurate speech quality scores with robust generalization.

The remainder of this paper is organized as follows. we introduce our novel proposed model for speech quality assessment and its corresponding objective function for training in Section II. In Section III, we describe the experimental process and display the experimental results. Our findings are also discussed. Finally, we conclude our work in Section IV.

10.21437/Interspeech.2023-811

2. Proposed method

2.1. Backbone architecture

The backbone of the traditional CNN-based speech quality assessment network is mainly composed of multiple serial concatenated convolutional layers. The features calculated by the final convolutional layer are directly fed to the subsequent network, while the effective low-dimensional features calculated by the intermediate convolutional layers fail to be effectively transferred and only evolve to high-dimensional features as the network progresses. With the premise that low-dimensional features have a greater impact on optimizing the objective function, the model backbone output tends to evolve towards the low-dimensional features generated by the intermediate convolutional layer. At this point, it is equivalent to the identity mapping between the output features and the low-dimensional features. Thus, a conflict arises between preserving low-dimensional features and extracting high-dimensional feature information. Furthermore, as the number of layers increases, the gradient may exponentially increase or decay, which poses a risk of gradient explosion or vanishing.

To address these issues, we draw upon Gao et al.'s [19] techniques for solving similar problems in the computer vision field and introduce DCC blocks from DenseNet to the speech quality assessment task, with some modifications to the basic architecture. The computation process of a DCC block is described as follows:

$$\mathbf{y} = concat\left(D\left(g^{n}\left(\mathbf{x}\right), D\left(\mathbf{x}\right)\right)\right) \tag{1}$$

where $concat(\cdot)$ stands for concatenation of inputs and $D(\cdot)$ means downsampling operation on matrices. $g(\cdot)$ are given as:

$$g(\mathbf{x}) = ReLU(Conv(\mathbf{x})) \tag{2}$$

where $ReLU(\cdot)$ represents ReLU activation function and $Conv(\cdot)$ indicates convolution calculation.

The overall framework of our model is shown in Figure 1. We employ a feature extracting network with three DCC blocks as our backbone. A DCC block mainly consists of four convolutional layers and a cascade structure, where an activation function layer is applied to activate the output data after each convolutional layer. The input features are concatenated with the output features after passing through three feature convolutional layers. Then we set the step size of the fourth convolutional layer to 2 (non-temporal dimension) to reduce the size of the feature. In this way, the calculation complexity of the model is reduced and the receptive field is expanded at the same time. By cascading the features before and after convolution, the low-dimensional feature information can be transferred to the convolved features without identity mapping. Through multiple densely connected CNN blocks, our model can extract both low-dimensional and high-level features effectively.

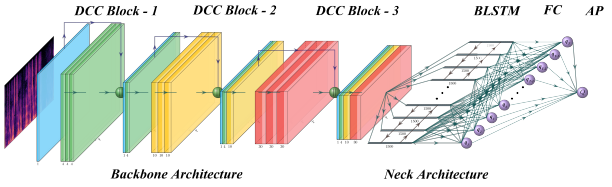

Figure 1: *The overall framework of our proposed model. FC and AP respectively represent fully connected layer and average pooling layer.*

As speech signals are time-series data, the quality of each time frame contributes to the overall speech assessment results. Due to the local calculation property of convolutional layers, the original timing characteristics of the speech signal will not be destroyed. Thus, the output features obtained through the backbone will still retain the timing correlation of the speech signal. In this subsection, we expect that the DCC-based backbone will combine the input information of each time frame to extract frame-level quality features. Then, these features are scored through the neck architecture to obtain frame-level scores and system-level scores in sequence, which will be discussed in the next subsection.

2.2. Neck architecture

The system-level score of speech signal is not only related to the features of the current time frame, but also to the features of the preceding and subsequent time frames. Therefore, a sequential analysis of the features of all frames in the temporal dimension is required. The DCC-based backbone architecture extracts the temporal frame-level features of the speech signal, and the neck architecture is responsible for analyzing and calculating of these frame-level features sequentially and combining them for system-level representation. The BLSTM network has the ability to process information bidirectionally based on the time step, which can be comprehensively represented by combining the temporal features. Therefore, we selected the BLSTM network as the temporal features analysis module to extract the overall features of speech quality. A standard BLSTM consists of an input gate, a forget gate, an output gate, an input modulation gate, a memory cell state, and a common BLSTM unit at a time step can be expressed as follows:

$$\begin{cases} \mathbf{i}_f^t = Sigmod\left(\sigma_{i_f}\left(\mathbf{x}^t, \mathbf{h}^{t-1}\right)\right), \\ \mathbf{f}_f^t = Sigmod\left(\sigma_{f_f}\left(\mathbf{x}^t, \mathbf{h}^{t-1}\right)\right), \\ \mathbf{o}_f^t = Sigmod\left(\sigma_{o_f}\left(\mathbf{x}^t, \mathbf{h}^{t-1}\right)\right), \\ \mathbf{g}_f^t = Tanh\left(\sigma_{g_f}\left(\mathbf{x}^t, \mathbf{h}^{t-1}\right)\right), \\ \mathbf{c}_f^t = \mathbf{f}_f^t \odot \mathbf{c}_f^{t-1} + \mathbf{i}_f^t \odot \mathbf{g}_f^t, \\ \mathbf{h}_f^t = \mathbf{o}_f^t \odot Tanh\left(\mathbf{c}_f^t\right). \end{cases} \tag{3}$$

$$\begin{cases} \mathbf{i}_b^t = Sigmod\left(\sigma_{i_b}\left(\mathbf{x}^t, \mathbf{h}^{t+1}\right)\right), \\ \mathbf{f}_b^t = Sigmod\left(\sigma_{f_b}\left(\mathbf{x}^t, \mathbf{h}^{t+1}\right)\right), \\ \mathbf{o}_b^t = Sigmod\left(\sigma_{o_b}\left(\mathbf{x}^t, \mathbf{h}^{t+1}\right)\right), \\ \mathbf{g}_b^t = Tanh\left(\sigma_{g_b}\left(\mathbf{x}^t, \mathbf{h}^{t+1}\right)\right), \\ \mathbf{c}_b^t = \mathbf{f}_b^t \odot \mathbf{c}_b^{t+1} + \mathbf{i}_b^t \odot \mathbf{g}_b^t, \\ \mathbf{h}_b^t = \mathbf{o}_b^t \odot Tanh\left(\mathbf{c}_b^t\right). \end{cases} \tag{4}$$

$$\mathbf{y}^t = concat\left(\mathbf{h}_f^t, \mathbf{h}_b^t\right) \tag{5}$$

where $\mathbf{i^t}, \mathbf{f^t}, \mathbf{o^t}, \mathbf{g^t}, \mathbf{c^t}, \mathbf{h^t}$ represent input gate, forget gate, output gate, input modulation gate, memory cell state and hidden state in the BLSTM unit at time t, respectively. A group of formulas in (3) represents the forward processing condition of one BLSTM unit, denoted by the subscript f. Similarly, the group of formulas in (4) indicates the backward processing condition of one BLSTM unit, denoted by the subscript b. The results in (3) and (4) are concatenated to compose the final output $\mathbf{y}^t$ at time t in equation (5). $Sigmoid(\cdot)$ stands for sigmoid activation function. $Tanh(\cdot)$ stands for tanh activation function. $\odot$ means for Hadamard product. And $\sigma(\cdot)$ denotes a computational process with a learnable matrix, which can be described as:

$$\sigma_s\left(\mathbf{x}, \mathbf{y}\right) = \mathbf{W}_{sx} \cdot \mathbf{x} + \mathbf{W}_{sy} \cdot \mathbf{y} + bias_s \tag{6}$$

where $\mathbf{W}_{sx}$ and $\mathbf{W}_{sy}$ represent a learnable matrix for the input $\mathbf{x}$ and $\mathbf{y}$ under the condition of s , respectively. $bias_s$ represents the learnable parameter under the condition of s .

Then, after passing through the BLSTM module, we use a fully connected layer to regress the features and obtain the frame-level score. Finally, based on these frame-level scores, a global average pooling layer is applied to calculate the final system-level score.

2.3. Objective function

In this paper, we regard the no-reference quality assessment based on deep learning as a regression task. Under the no-reference condition, a non-invasive method of our model is used to regress the quality scores with reference speech calculated. Considering that the speech is distorted by the non-stationary noise at different time scales, it is unreasonable to use the system-level score for training directly which may leads to inaccurate estimates. Therefore, we combined the error of frame-level scores and system-level scores as the objective function to iteratively optimize the model. The objective function can be describe as :

$$O = \frac{1}{N}\sum_{i=1}^{N}\left[\left(S_i - S_i'\right)^2 + \frac{\alpha}{T}\sum_{t=1}^{T}\left(S_i - s_{i,t}\right)^2\right] \quad (7)$$

where S_i is the system-level ground truth score of the i-**th** speech, and S_i' is the system-level prediction score of the i-**th** speech. $s_{i,t}$ is the frame-level score of the i-**th** speech at time t . T represents the length of time frames of one speech. N represents the number of speech training sets. α is the balance factor between frame-level scores and system-level scores, and we set the balance factor α to 1 in this study.

3. Experiment

3.1. Experimental setup

MUSAN dataset [20], which is a corpus of music, speech, and noise recordings, is adopted in our experiments. We selected 173 speech signals in MUSAN which are resampled to 16kHz. Each resampled speech signal was divided into 8-second slices, resulting in 9169 clean speech slices that were used as a reference dataset. To simulate the impact of noise on speech signal transmission under stationary and burst noise conditions, we add varying levels of Gaussian white noise to the speech signal slices. As for the stationary noise, we randomly add 10 different types of Gaussian white noise to each speech slice, with a signal-to-noise ratio (SNR) uniformly distributed between -30dB and 40dB (interval 1dB). For burst noise, we first randomly add 10 kinds of background Gaussian white noise with SNR uniformly distributed in the range of 20dB-40dB to each signal slice (interval 1dB). Then, we add burst Gaussian white noise of a duration of 1 second to each speech slice with background noise which is uniformly distributed in the range of signal-to-noise ratio -15dB-15dB (interval 1dB). Apart from the noise mentioned above, no other type of impairment was considered. Among the obtained 183,380 speech slices with noise, we randomly select 160,000 signal slices as the training set, 15,000 signal slices as the test set, and 8,380 signal slices as the cross-validation set without repetition.

The input information of the model is the spectral feature of the signal slice. Hence, we perform a STFT with a Hamming window of 32ms and a hop of 16ms on the speech slices. The STFT has 512 points, and the length of the sliding window is 256. The model predicts the quality score of the speech slice based on its STFT information, and the objective function calculates the error between the prediction score and the ground-truth score. The backpropagation algorithm optimizes the model based on the error.

We evaluated the accuracy of our model for speech quality scores using three evaluation metrics: mean square error (MSE), Pearson linear correlation coefficient (PLCC), and Spearman rank-order correlation coefficient (SRCC). MSE directly quantifies the error between the model's prediction score and the ground-truth score, PLCC evaluates the degree of linear correlation between the two scores, and SRCC measures the dependence of the model's prediction score and the ground-truth score. Lower MSE, higher PLCC and higher SRCC are inclined to better performance of the model.

3.2. Performance evaluation of model architecture

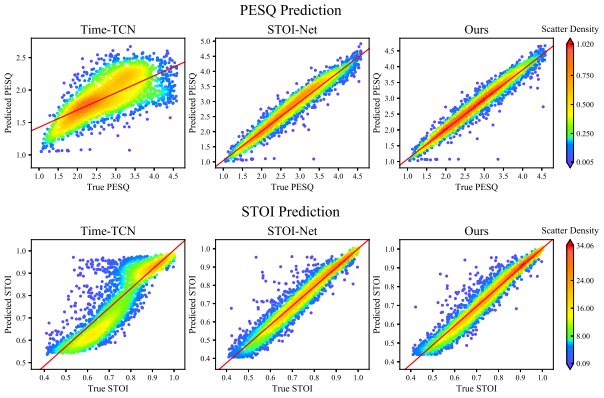

Figure 2: *The scatter diagram of prediction scores and ground truth scores of Time-TCN, STOI-Net and our proposed model under the condition that PESQ scores and STOI scores are used as optimization metrics. Red line depicts the fitted linear curve.*

We selected PESQ and STOI as the objective assessment algorithms to score speech slices, and trained our model using these scores as reference. It is worth noting that we regress the two assessment algorithms separately under the same conditions, rather than making the model learn two assessment algorithms at the same time. This helps us compare the generalization of our model in different algorithms. We then compare our proposed model with Xupeng Jia's Time-TCN model [17] and

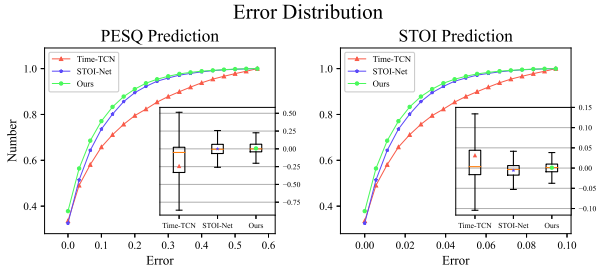

Figure 3: *The error's CDF curve and box plot of prediction scores of Time-TCN, STOI-Net and our proposed model under the condition that PESQ scores and STOI scores are used as optimization metrics. Orange line is the median of error.* ▲, ★ *and* ● *represent the mean of error in different models.*

Ryandhimas E's STOI-Net [10]. To keep other variables unchanged, the dataset is shuffled with the same random seed. The training epoch, the learning rate and the batch size are set to 2, 1e-4 and 1, respectively. We choose the model which minimizes the loss value of the objective function in the cross-validation set as the final test model. The PLCC, SRCC, and MSE in the experiment are shown in Table 1 . The scatter diagram of prediction scores and ground truth scores is shown in Figure 2, and the error's CDF curve and box plot are shown in Figure 3. It can be seen that in Table 1 , our model outperforms the other two models in terms of lower MSE, higher PLCC, and higher SRCC, regardless of whether PESQ or STOI is used as the optimization metric. This is also evident from Figure 2, where prediction scores of our model have a more obvious correlation with the ground truth scores. Additionally, Figure 3 exhibits prediction scores of our model have a higher proportion in the lower-error area.

Table 1: *MSE, PLCC and SRCC results of Time-TCN, STOI-Net and our proposed model on PESQ and STOI prediction scores. ↑ or ↓ is better.*

	PESQ			STOI		
Model	**MSE ↓**	**PLCC ↑**	**SRCC ↑**	**MSE ↓**	**PLCC ↑**	**SRCC ↑**
Time-TCN	1.1653	0.6805	0.6936	0.0069	0.8700	0.8790
STOI-Net	0.0448	0.9651	0.9675	0.0022	0.9569	0.9590
Ours	**0.0389**	**0.9695**	**0.9715**	**0.0019**	**0.9608**	**0.9630**

3.3. Performance evaluation of DCC module

In the experiments in Section 3.2, compared with the Time-TCN and STOI-Net models, our model shows superior performance when PESQ scores and STOI scores were used as optimization metrics. In order to verify the effect of the DCC module in our model, the following experiments are carried out. The neck architecture remains unchanged. And we replace the 3-layer DCC module with the traditional CNN module as comparison model 1 (CM-1). Then we remove the backbone architecture, modify the BLSTM to match the dimension of the input features, and get the comparison model 2 (CM-2). Finally, we use PESQ scores and STOI scores as optimization metrics, and train, cross-validate and test the two models under the same conditions as the experiment in Section 3.2. The scores predicted by the three models are used for comparison and analysis. PLCC, SRCC, and MSE in the experiments are shown in Table 2. The scatter diagrams of the prediction scores and the reference score are shown in Figure 4, and the error's CDF curve and box plot are shown in Figure 5. The results show that the proposed model with the DCC module performs better than the other two models in terms of MSE, PLCC, and SRCC. The scatter diagrams also indicate that our model exhibits a higher correlation with the reference scores than the other two models.

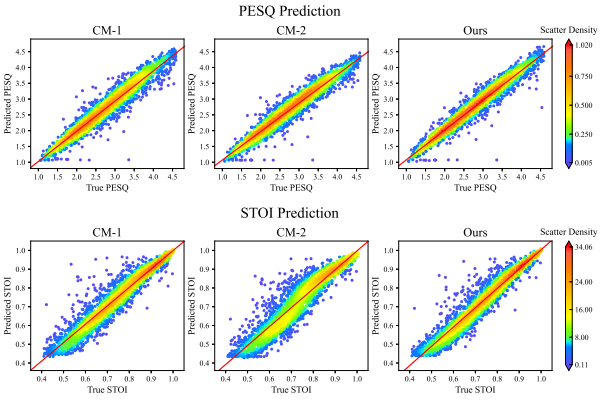

Figure 4: *The scatter diagram of prediction scores and ground truth scores of CM-1, CM-2 and our proposed model under the condition that PESQ scores and STOI scores are used as optimization metrics. Red line depicts the fitted linear curve.*

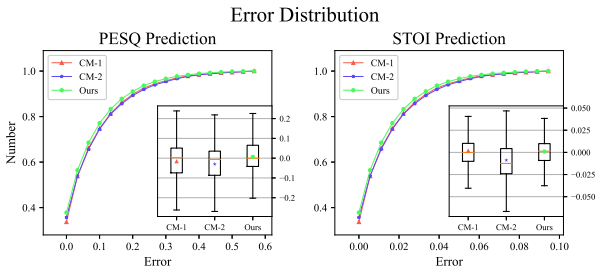

Figure 5: *The error's CDF curve and box plot of prediction scores of CM-1, CM-2 and our proposed model under the condition that PESQ scores and STOI scores are used as optimization metrics. Orange line is the median of error. ▲, ★ and ● represent the mean of error in different models.*

Table 2: *MSE, PLCC and SRCC results of CM-1, CM-2 and our proposed model on PESQ and STOI prediction scores. ↑ or ↓ is better.*

	PESQ			STOI		
Model	**MSE ↓**	**PLCC ↑**	**SRCC ↑**	**MSE ↓**	**PLCC ↑**	**SRCC ↑**
CM-1	0.0536	0.9599	0.9613	0.0021	0.9510	0.9533
CM-2	0.0581	0.9625	0.9643	0.0031	0.9354	0.9419
Ours	**0.0389**	**0.9695**	**0.9715**	**0.0019**	**0.9608**	**0.9630**

4. Conclusions

In this paper, we propose a novel deep learning-based method for speech assessment. We replace the multi-layer concatenated convolutional module with a 3-layer DCC block. When the reference speech is difficult to obtain, our proposed method has a wider application prospect and can predict more accurate scores. Experimental results show that compared with other deep learning-based methods, our model predicts speech quality scores closer to ground truth scores. Additionally, we investigate the effectiveness of the DCC block in the speech quality assessment task and show that our model, which incorporates this structure, yields more accurate results compared to models without the DCC block. Moving forward, we plan to evaluate our model's generalization ability by testing the prediction results under the condition of different datasets. Moreover, we intend to investigate the performance of deep learning approaches on multiple speech quality assessment tasks.

5. References

[1] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *MULTIMEDIA SYSTEMS*, vol. 22, no. 2, pp. 213–227, MAR 2016.

[2] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for

speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[3] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[4] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[5] L. Malfait, J. Berger, and M. Kastner, "P. 563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[6] H. Takahashi and K. Kondo, "On reference signal estimation from noisy speech using deep learning for intelligibility estimation," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2018, pp. 347–348.

[7] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks." in *Interspeech*, 2016, pp. 342–346.

[8] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.

[9] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[10] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "Stoi-net: A deep learning based non-intrusive speech intelligibility assessment model," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 482–486.

[11] H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, and Y. Tsao, "Hasa-net: A non-intrusive hearing-aid speech assessment network," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 907–913.

[12] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[13] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[14] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[15] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[16] P. Manocha, B. Xu, and A. Kumar, "Noresqa: A framework for speech quality assessment using non-matching references," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 363–22 378, 2021.

[17] X. Jia and D. Li, "A deep learning-based time-domain approach for non-intrusive speech quality assessment," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 477–481.

[18] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 631–635.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[20] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.