Interspeech 2022
18-22 September 2022, Incheon, Korea

Non-intrusive Speech Quality Assessment with a Multi-Task Learning based Subband Adaptive Attention Temporal Convolutional Neural Network

Xiaofeng Shu, Yanjie Chen, Chuxiang Shang, Yan Zhao, Chengshuai Zhao, Yehang Zhu, Chuanzeng Huang and Yuxuan Wang

Speech, Audio and Music Intelligence (SAMI) group, ByteDance

{shuxiaofeng, chenyanjie, shangchuxiang.19, zhao.yan, zhaochengshuai, zhuyehang, huangchuanzeng, wangyuxuan.11}@bytedance.com

Abstract

In terms of subjective evaluations, speech quality has been generally described by a mean opinion score (MOS). In recent years, non-intrusive speech quality assessment shows an active progress by leveraging deep learning techniques. In this paper, we propose a new multi-task learning based model, termed as subband adaptive attention temporal convolutional neural network (SAA-TCN), to perform non-intrusive speech quality assessment with the help of MOS value interval detector (VID) auxiliary task. Instead of using fullband magnitude spectrogram, the proposed model takes subband magnitude spectrogram as the input to reduce model parameters and prevent overfitting. To effectively utilize the energy distribution information along the subband frequency dimension, subband adaptive attention (SAA) is employed to enhance the TCN model. Experimental results reveal that the proposed method achieves a superior performance on predicting the MOS values. In ConferencingSpeech 2022 Challenge, our method achieves a mean Pearson's correlation coefficient (PCC) score of 0.763 and outperforms the challenge baseline method by 0.233.

Index Terms: Speech quality assessment, MOS, VID, multi-task learning, SAA-TCN, PCC

1. Introduction

The effective assessment of speech quality and intelligibility is a vital issue in modern speech communication systems. Typically, the speech assessment metrics are divided into two categories, subjective and objective metrics. For the evaluation of subjective metrics, it involves human subjects to participate in listening tests and provide feedbacks. One of the most widely used subjective metrics is mean opinion score (MOS), whose value range is from 1 to 5. Higher value means better speech quality. Although MOS value is considered as a gold standard in many scenarios including the naturalness assessment of synthesized speech and the quality estimation of conference speech, it is time-consuming and labor-intensive to perform the evaluations. Hence, objective assessment metrics are usually used as alternatives including perceptual evaluation of speech quality (PESQ) [1], short-time objective intelligibility (STOI) [2], signal-to-distortion ratio (SDR), perceptual objective listening quality assessment (POLQA) [3]. Although showing good correlations with subjective evaluations, the need for a reference signal limits the applications of the mentioned objective metrics in real world, where typically no reference signal is available.

In recent years, deep neural network (DNN)-based learning architectures have been successfully applied to many speech processing tasks, which greatly motivates the investigations of DNN in non-intrusive speech quality assessment to overcome the limitation of conventional objective metrics [4–8]. In [4], a convolutional and recurrent neural network combined model was adopted as a MOS predictor for voice conversion. In [5], quality-Net used one bidirectional long short term memory (Bi-LSTM) layer followed with two fully connected (FC) layers to predict frame-level PESQ of input audio utterances. Taking rating bias of listeners into consideration, [6] added a listener-bias branch of the system to model listener preferences. Besides, different training architectures were utilized to improve the prediction accuracy. To evaluate the performance of noise suppressors, Deep Noise Suppression Mean Opinion Score (DNS-MOS) API was used as a scoring tool [7], and training of DNS-MOS applied a multi-stage self-teaching framework to average out human bias. In [8], DNN based speech enhancement system was also used to assist non-intrusive speech quality assessment, which can be considered as an surrogate subjective metric and does not need a reference signal. Although DNN based non-intrusive evaluation systems have achieved tremendous progress, some drawbacks persist and need to be further addressed. First, most systems do not generalize well in real environments, for example, in online conference conditions, which involve multiple speech distortions such as background noise, room reverberation, and packet loss. Second, the prediction accuracy of current evaluation systems tended to be limited on unseen data [9]. In [10], the multi-task learning mechanism was also used to effectively estimate multiple objective assessment metrics including PESQ, STOI, Hearing-Aid Speech Quality Index (HASQI), and SDR.

Recently, temporal convolutional neural network (TCN) based models have shown great successes in many speech processing tasks, including noise suppression [11], acoustic echo cancellation [12], dereverberation [13], and speech separation [14]. Specifically, TCN employs residual learning with dilated convolution units, which is able to capture wide contextual dependencies and outperforms long short term memory (LSTM) models in several sequence modeling tasks.

Motivated by the channel-wise subband study [15] and attention based TCN [16], in this study, we propose a multi-task network, termed as subband adaptive attention temporal convolutional neural network (SAA-TCN), to perform non-intrusive speech quality assessment. The proposed model utilizes subband magnitude spectrogram as the input feature. Compared with fullband model, this design can improve computation efficiency while mitigating overfitting by reducing model parameters. The introduction of a subband adaptive attention (SAA) module helps the TCN to model the energy distribution along the frequency dimension to focus more on important feature channels. Finally, MOS value interval detector (VID) as an auxiliary task is proposed to improve the MOS predictions. In

10.21437/Interspeech.2022-10315

addition, an auxiliary module is deployed to provide additional information for the main MOS prediction task. The experimental results demonstrate that our approach provides reliable MOS predictions for unknown speech samples.

The rest of the paper is organized as follows. Section 2 describes the non-intrusive speech quality assessment system as well as their sub-modules. In Section 3, the experimental and comparison results are presented to evaluate the performance of the proposed method. The conclusions are given in Section 4.

2. Proposed method

2.1. System overview

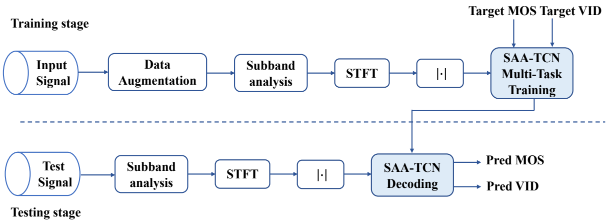

Figure 1: *Schematic diagram of the multi-task SAA-TCN based non-intrusive speech quality assessment system.*

The schematic diagram of the multi-task SAA-TCN based non-intrusive speech quality assessment system is illustrated in Figure 1. The whole system consists of two stages. First, in the "training stage", data augmentation is used for input signals to prevent model training from overfitting. After that, the signals are decomposed into subbands introduced in [15]. Then, the subband magnitude spectrogram is fed into SAA-TCN as the input feature. During multi-task training, MOS and VID are used as our optimization targets. Our experiments show that multi-task learning can help to obtain more accurate predictions on MOS branch.

In the "testing stage", similar to training stage, we perform subband analysis and extract subband magnitude spectrogram. The well-trained SAA-TCN then takes subband feature as input to predict the real MOS and VID.

2.2. Data augmentation

As described in [9], we explore two data augmentation strategies: perturbing the audio speed by a randomly chosen factor between 0.95 and 1.05; adding silence by a random time from 0.1s to 1s. For all training samples, we apply the aforementioned data augmentation online during training, which helps the system reduce overfitting.

2.3. Subband analysis and subband feature

Liu [15] proposed to use the channel-wise subband feature to reduce resource consumption and improve separation performance. In this study, only the subband decomposition is used in the analysis procedure. We utilize a set of analysis filters $H_k\left(e^{j\omega}\right)$, $k = 1, 2, 3, 4$ to decompose subbands.. The analysis filters we used are uniform filterbanks with a length of 64, which are referred to the configuration in [17]. We show the subband magnitude spectrogram extraction in Figure 2. The interested readers are recommended to the open source toolbox [1] for more details. The usage of such subband features not only speeds up computation but also helps to avoid overfitting during training by reducing the number of parameters in the model.

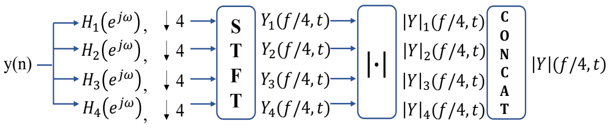

Figure 2: *Subband analysis and subband feature extractions.*

2.4. The SAA-TCN model

Figure 3 shows the network architecture of our proposed mulit-task SAA-TCN system for speech quality assessment. The details of each component of the system are described in the following subsections.

2.4.1. *Encoder block*

In this work, six encoder blocks are applied to the subband magnitude spectrograms followed by a batch normalization layer to learn suitable features. Each encoder block consists of one convolution layer with kernel size 3 × 3 followed by a batch normalization layer and rectified linear unit (ReLU) activation. The max pooling layer with kernel sizes 1 × 2 is used to downsample the feature map in subband frequency. The detailed design of the encoder block is shown in Table 1.

Table 1: *Design of encoder blocks.*

	Layer	Size	Stride
	Conv, 16 ch Batch normalization Relu	3×3	1×1
	Maxpool	1×2	1×2
	Conv, 32 ch Batch normalization Relu	3×3	1×1
	Maxpool	1×2	1×2
4×	Conv, 64 ch Batch normalization Relu	3×3	1×1
	Maxpool	1×2	1×2

2.4.2. *TCN block*

Motivated by the successful TCN for sequence modeling [18] and the effectiveness of adding a frequency dimension adaptive attention (FAA) module in TCN [16], N TCN blocks with a subband frequency dimension adaptive attention module are used in our model, as shown in the area encompassed by red dash lines in Figure 3. Each TCN block contains three one-dimensional causal dilated convolutional units. d_{model} and d_f denote the dimension of encoder blocks output and the size of first convolutional unit, respectively. In our

[1] https://www.mathworks.com/matlabcentral/fileexchange/40128-filter-bank-design

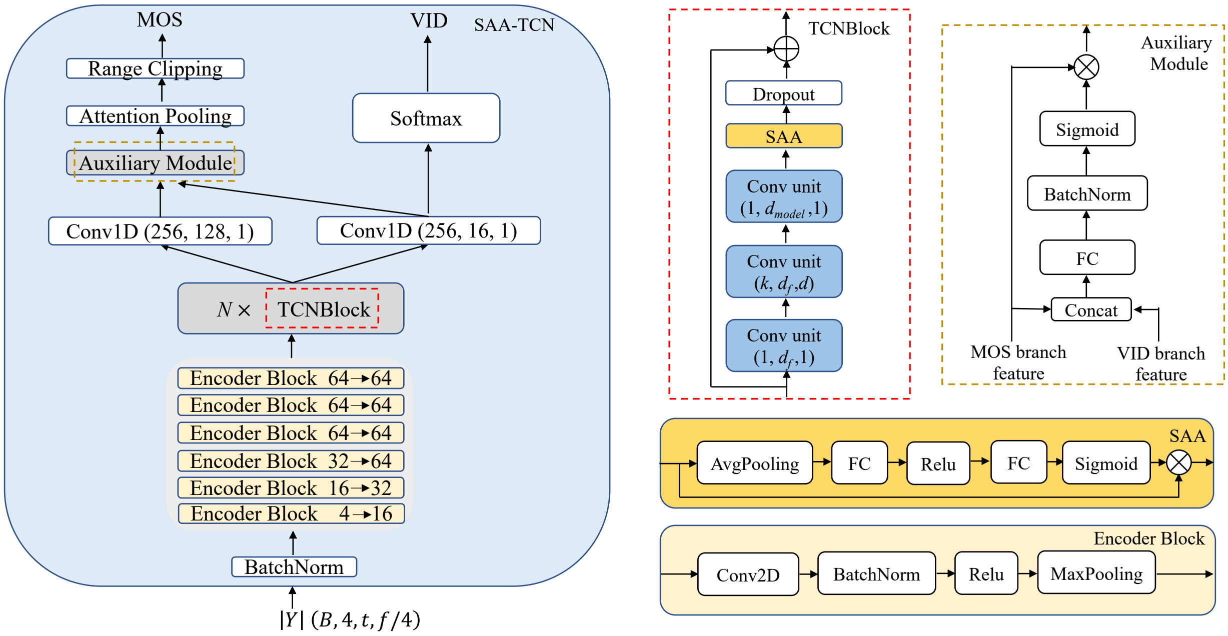

Figure 3: *Overview of the SAA-TCN architecture. We use the subband magnitude spectrogram as the input feature to perform the MOS and MOS value interval detection (VID) multi-task training.*

model, each convolutional unit is pre-activated by the ReLU activation followed by batch normalization. In Figure 3, we use $(kernel\ size, output\ size, dilation\ rate)$ to describe detailed parameters of each convolutional unit. The first and third convolutional units in each block have a kernel size of 1, and the second convolutional unit has a kernel size of 3. The first and third convolutional units have a dilation rate of 1, while the second convolutional unit employs a dilation rate of d, providing a larger context on temporal dimension. As mentioned in [14], the dilation rate d is cycled as the block index b increases, given by

$$d = 2^{(b-1 \bmod \log_2(D)+1)}, \tag{1}$$

where mod is the modulo operation, and D is the maximum dilation rate. In the ConferencingSpeech 2022 Challenge, we choose $d_{model} = 256$, $d_f = 64$, $D = 16$ and $N = 20$ as the final configuration.

Recently, Squeeze-and-Excitation attention module has been successfully applied to deep neural network based speech processing. In [16], the authors proposed a FAA module to obtain significant performance gains for speech enhancement. In this study, we also add a SAA module under the TCN framework to model the energy distribution along the subband frequency dimension. After obtaining the output of three dilated convolutional units, we first reshape it into the shape of $1 \times T \times d_{model}$, and then global averaging pooling in SAA module is used to generate the statistics of downsampled subband frequency bins, denoted by g_s. The average-pooled features are then fed into two fully connected layers. With two FCs, the linear information among subband frequency domain can be combined more efficiently. The attention weight of the subband frequency bins can be written as

$$\mathrm{w} = \sigma(W_1\delta(W_0(g_s))), \tag{2}$$

where the symbol δ represents the ReLU operation, $W_0 \in \mathbb{R}^{d_{model}\times d_{model}/r}$ and $W_1 \in \mathbb{R}^{d_{model}/r\times d_{model}}$ represent weights of the FC layers. To reduce parameters and computation cost, the size of r is selected to be 4 in this study. The symbol σ represents the excitation function, and a Sigmoid function is used. The SAA can be easily integrated into TCN and bring significant gain for system performance. Finally, dropout layer is used to further mitigate overfitting issue during model training.

2.4.3. *Auxiliary module*

MOS value interval detector (VID) estimation is used as an auxiliary task to improve the accuracy of MOS prediction, which is described in multi-task learning section 2.5 in details. The auxiliary module, shown in the area encompassed by golden dash lines in Figure 3, aims at helping to incorporate the information from the VID branch to the MOS prediction branch.

2.4.4. *Attention pooling*

Attention mechanism [20] has been successfully applied in many learning tasks. Here we use the same attention pooling module as proposed in [21]. The attention scores are first computed by the feedforward networks, then masked at the zero-padded time steps and applied to a softmax function to yield the normalized attention weights, which is given by

$$Q = \mathrm{softmax}((H_1\delta(YH_0))), \tag{3}$$

where $Y \in \mathbb{R}^{L\times d_{tf}}$, $H_0 \in \mathbb{R}^{d_{tf}\times M}$, $H_1 \in \mathbb{R}^{1\times L}$ and $Q \in \mathbb{R}^{1\times M}$, d_{tf} and M denote the feature vectors of dimension and the size of one FC layer in the feedforward networks, respectively. In this study, M is set to 128, and L is the effective frame-level dimensionality. These attention weights are applied to the input matrix Y using a matrix multiplication operation, given by

$$h = YQ^T. \tag{4}$$

Finally, we pass through a FC layer to estimate the overall score.

2.4.5. *Range clipping*

To fit the human scoring distributions, we apply hyperbolic tangent function to ensure a fixed range of network prediction score. That is,

$$s = 2\tanh(\bar{s}) + 3 \tag{5}$$

Where $\bar{s}$ is the output of attention pooling layer. The value range of s is constrained between 1 and 5, which guarantees the model to always give reasonable scores.

2.5. Multi-task learning

Multi-task learning simultaneously optimizes multiple objectives of different tasks using a shared backbone model. The benefits come from auxiliary information and cross regularization from different tasks. In our study, the MOS interval detector is used as an auxiliary task to improve the performance of MOS predictions in the context of multi-task learning. The VID state is obtained according to the MOS value interval given by

$$\begin{aligned} &For\ i = 0, 1, ...15: \\ &\quad \mathrm{VID} = i,\ if\ \ 1 + 0.25 * i \leq \mathrm{MOS} < 1 + 0.25 * (i + 1), \end{aligned} \tag{6}$$

where the VID state i corresponds to different MOS value intervals. 16 states are chosen based on the experimental results described in Section 3.

The loss function of SAA-TCN for multi-task training is

$$L = L_{\mathrm{VID}} + L_{\mathrm{MOS}}, \tag{7}$$

where L_{VID} and L_{MOS} are the loss components for VID and MOS prediction, respectively. We use the cross-entropy loss as L_{VID}, and L_{MOS} is defined as,

$$\begin{cases} L_{\mathrm{MOS}} = L_{MSE} + L_{Pearson} \\ L_{MSE} = (\hat{s} - s)^2 \\ L_{Pearson}(s, \hat{s}) = 1 - \rho^2_{Pearson}(s, \hat{s}) \\ \rho_{Pearson}(s, \hat{s}) = \frac{Cov(s, \hat{s})}{\sigma_s \sigma_{\hat{s}}}, \end{cases} \tag{8}$$

where s and $\hat{s}$ are the desired scores and the estimated scores, respectively; Cov and σ denote the covariance and variance, respectively.

3. Experimental results

ConferencingSpeech 2022 Challenge datasets are used to evaluate the proposed system, which include Tencent Corpus, NISQA Corpus, IU Bloomington Corpus, PSTN Corpus and their corresponding MOS labels. From these datasets, only Tencent Corpus, NISQA Corpus and PSTN Corpus were used as the training data because the IU Bloomington Corpus adopted ITU-R BS.1534 for subjective test resulting in a rating range of $0 \sim 100$ instead of $1 \sim 5$. Validation sets are selected from Tencent Corpus and PSTN Corpus. We used 3000 randomly selected speech clips from PSTN Corpus, 800 randomly selected speech clips from Tencent WithoutReverb Corpus, and 256 randomly selected speech clips from Tencent Reverb Corpus as validation sets. All the waveforms are resampled to 48 kHz. We calculate the STFT of the subband signals with a window length of 512 and window shift of 256. The model is trained with the Adam optimizer for 100 epochs with an initial learning rate of 1e-3. We use early stopping to select the best models.

Table 2: *Model results in terms of MSE, PCC, and SROCC for different validation sets (Tencent Corpus and PSTN).*

Validation set	Tencent corpus			PSTN		
Model	**MSE**	**PCC**	**SROCC**	**MSE**	**PCC**	**SROCC**
NISQA	0.1678	0.9424	0.9321	0.2492	0.8229	0.8071
TCN	0.1483	0.9452	0.9372	0.2599	0.8121	0.7943
SAA-TCN	0.1142	0.9575	0.9494	0.2473	0.8236	0.8087
Multi-Task SAA-TCN	**0.0971**	**0.9649**	**0.9591**	**0.2405**	**0.8273**	**0.8115**

Table 3: *The performance of Multi-Task SAA-TCN using different VID states configuration for different validation sets (Tencent Corpus and PSTN).*

Validation set	Tencent corpus			PSTN		
VID states	**MSE**	**PCC**	**SROCC**	**MSE**	**PCC**	**SROCC**
4	0.1149	0.9581	0.9508	0.2462	0.8230	0.8081
8	0.1086	0.9619	0.9545	0.2455	0.8235	0.8086
16	**0.0971**	**0.9649**	**0.9591**	**0.2405**	**0.8273**	**0.8115**
32	0.1035	0.9626	0.9556	0.2477	0.8217	0.8065

Table 4: *Proposed method and ConferencingSpeech 2022 Challenge baseline results in terms of RMSE, PCC for different test sets (NISQA Corpus testset and Challenge testset).*

Test set	NISQA-TEST		Challenge-TEST	
Model	**RMSE**	**PCC**	**RMSE**	**PCC**
Challenge Baseline	0.6311	0.8337	0.543	0.724
Ours	0.4960	0.8575	0.474	0.781

The NISQA speech quality prediction model proposed in [21] is used as the baseline system. As shown in Table 2, our method outperforms the baseline method in terms of mean squared error (MSE), Pearson's correlation coefficient (PCC), and Spearman rank order correlation coefficient (SROCC) metrics on the validation sets, indicating the capability of SSA-TCN for assessing speech quality. The baseline system, TCN system and SAA-TCN without multi-task learning (using L_{MOS}) are also presented for comparisons.

Table 3 shows the validation results for different VID states configuration. It can be seen that the best performance is achieved by 16 VID states. Consequently, in the final submission to the ConferencingSpeech 2022 Challenge, we adopt the following setup, TCN: $d_{model} = 256$, $d_f = 64$, $D = 16$, $N = 20$; SAA: $r = 4$; Multi-task: VIDStates $= 16$.

The challenge testset results released by the organizer are shown in Table 4, and for overall performance, we obtain the better result than baseline in terms of all eval metrics. And NISQA-TEST (including NISQA-TEST-FOR, NISQA-TEST-LIVETALK and NISQA-TEST-P501) is also used as the other testset to evaluate the performances. We obtained higher scores for both NISQA-TEST and Challenge-TEST in terms of root mean squared error (RMSE) and PCC. In the future, data augmentation module will be updated according to the real data distribution to further improve the performance of our proposed non-intrusive speech quality assessment system.

4. Conclusion

In this paper, a novel non-intrusive speech quality assesment method based on multi-task SAA-TCN is proposed to provide reliable MOS predictions. The proposed SAA-TCN model uses the subband magnitude spectrogram as the input feature, which reduces model parameters and prevents overfitting. SAA module assists TCN model to obtain energy distribution along the subband frequency dimension. The MOS value VID as an auxiliary task improves the performance of MOS prediction main task. The experimental results show that the proposed model outperforms the NISQA method in terms of all test metrics. On the test sets of the ConferencingSpeech 2022 Challenge, our model produces a superior performance compared with baseline and is among the top five models in this challenge.

5. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[7] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.

[8] X. Dong and D. S. Williamson, "Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 3348–3359, 2020.

[9] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," *arXiv preprint arXiv:2110.02635*, 2021.

[10] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 911–915.

[11] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6628–6632.

[12] X. Shu, Y. Zhu, Y. Chen, L. Chen, H. Liu, C. Huang, and Y. Wang, "Joint echo cancellation and noise suppression based on cascaded magnitude and complex mask estimation," *arXiv preprint arXiv:2107.09298*, 2021.

[13] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1598–1607, 2020.

[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] H. Liu, L. Xie, J. Wu, and G. Yang, "Channel-wise subband input for better voice and accompaniment separation on high resolution music," *arXiv preprint arXiv:2008.05216*, 2020.

[16] Q. Zhang, Q. Song, A. Nicolson, T. Lan, and H. Li, "Temporal convolutional network with frequency dimension adaptive attention for speech enhancement," *Proc. Interspeech 2021*, pp. 166–170, 2021.

[17] H. Liu, Q. Kong, and J. Liu, "Cws-presunet: Music source separation with channel-wise subband phase-aware resunet," *arXiv preprint arXiv:2112.04685*, 2021.

[18] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] 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.