TIME-FREQUENCY ATTENTION FOR MONAURAL SPEECH ENHANCEMENT

Qiquan Zhang[1], Qi Song[2,], Zhaoheng Ni[3], Aaron Nicolson[4], Haizhou Li[1,5]*

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Alibaba Group, China   [3]Meta AI, United States
[4]Australian e-Health Research Centre, CSIRO, Australia
[5]Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Most studies on speech enhancement generally don't explicitly consider the energy distribution of speech in time-frequency (T-F) representation, which is important for accurate prediction of mask or spectra. In this paper, we present a simple yet effective T-F attention (TFA) module, where a 2-D attention map is produced to provide differentiated weights to the spectral components of T-F representation. To validate the effectiveness of our proposed TFA module, we use the residual temporal convolution network (ResTCN) as the backbone network and conduct extensive experiments on two commonly used training targets. Our experiments demonstrate that applying our TFA module significantly improves the performance in terms of five objective evaluation metrics with negligible parameter overhead. The evaluation results show that the proposed ResTCN with the TFA module (ResTCN+TFA) consistently outperforms other baselines by a large margin.

Index Terms— speech enhancement, time-frequency attention, energy distribution, temporal convolutional network

1. INTRODUCTION

Speech enhancement seeks to enhance the speech signal in the presence of background noise. It is a fundamental component for many speech processing applications, such as automatic speech recognition, speaker identification, hearing aids, and teleconference. Statistical model-based speech enhancement [1–3] has been extensively studied for decades, which performs well for stationary noises, however, fails to handle non-stationary noises [4].

Speech enhancement with supervised deep learning has achieved remarkable progress. Existing methods can be grouped into two categories by the way input signals are handled. Time-domain methods perform speech enhancement directly on the speech waveform, where a DNN is optimized to learn the mapping from the noisy waveform to the clean one [5, 6]. Time-frequency (T-F) domain methods typically train a DNN to predict a spectral representation of the clean speech or a T-F mask. The most popular T-F masks include ideal ratio mask (IRM) [7], phase-sensitive mask (PSM) [8], and complex IRM (cIRM) [9]. In this study, we adopt the IRM and PSM to perform speech enhancement.

In earlier studies, multi-layer perceptrons (MLPs) are the most widely adopted architectures, but they are limited in capturing the

long-term dependencies. To overcome the limitation, Chen *et al.* [10] employed a recurrent neural network (RNN) with four long short-term memory (LSTM) layers to perform speech enhancement, demonstrating obvious superiority over MLPs. However, LSTM network suffers from a slow and complex training procedure and requires a large number of parameters, which severely limits its applicability. Recently, the residual temporal convolution networks (ResTCNs) [11], which utilize dilated convolution and residual skip connections, have shown impressive performance in modeling long-term dependencies and gained considerable success in speech enhancement [12–14]. More recently, self-attention based Transformer [15] model has been successfully applied for speech enhancement and many other speech processing-related tasks for their capability of capturing long-range dependencies.

The existing models mainly focus on how to effectively model the long-range dependencies, while they generally ignore the energy distribution characteristics of speech in T-F representation, which is equally important to speech enhancement. The attention mechanisms [16–18] have been well studied to learn what is important to the learning task. Inspired by the idea of attention, we propose a novel architecture unit, termed T-F attention (TFA) module, to model the energy distribution of speech. Specifically, the TFA module consists of two parallel attention branches, i.e., time-dimension (TA) and frequency-dimension attention (FA) [19] that produce two 1-D attention maps to guide the models to focus on 'where' (which time frames) and 'what' (which frequency channels) respectively. The TA and FA modules are combined to generate a 2-D attention map enabling the models to capture the speech distribution in the T-F domain. To validate the idea, we use the recent ResTCN architecture as the backbone network and adopt two representative training targets, that will be discussed in Section 2, to perform extensive experiments.

The rest of this paper is organized as follows. Section 2 gives the introduction of T-F domain speech enhancement. In Section 3, we describe the proposed network. Section 4 presents the experimental setup and evaluation results. Section 5 concludes this paper.

2. PROBLEM FORMULATION

The noisy speech can be modeled as a combination of clean speech and additive noise in the Short-Time Fourier Transform (STFT) domain:

$$X[l, k] = S[l, k] + D[l, k] \tag{1}$$

where $X[l, k]$, $S[l, k]$, and $D[l, k]$ denote the STFT coefficients at time frame l and frequency bin k of the noisy speech, clean speech, and noise, respectively. For supervised speech enhancement, a DNN is typically trained to predict the pre-designed training targets. The

results are then applied to reconstruct the clean speech. To demonstrate the efficacy of our proposed TFA module, we adopt two widely used training targets to conduct extensive enhancement experiments. The details are given below.

The ideal ratio mask (IRM) [7] is defined as:

$$\text{IRM}[l, k] = \sqrt{\frac{|S[l, k]|^2}{|S[l, k]|^2 + |D[l, k]|^2}} \tag{2}$$

where $|S[l, k]|$ and $|D[l, k]|$ denote the spectral magnitudes of clean speech and noise, respectively. The phase-sensitive mask (PSM) [8] is defined on the STFT magnitude of clean and noisy speech. A phase error item is introduced to compensate for utilizing the noisy speech phases:

$$\text{PSM}[l, k] = \frac{|S[l, k]|}{|X[l, k]|} \cos[\theta_{S[l,k]-X[l,k]}] \tag{3}$$

where $\theta_{S[l,k]-X[l,k]}$ denotes the phase difference between clean speech and noisy speech. The PSM is truncated to between 0 and 1 to fit the output range of the sigmoid activation function.

3. SPEECH ENHANCEMENT WITH T-F ATTENTION

3.1. Network Architecture

Fig.1 (a) shows the architecture of the ResTCN backbone network [12], which takes as input the STFT magnitude of noisy speech $|\mathbf{X}| \in \mathbb{R}^{L \times K}$ of L frames, each having K frequency bins. The output layer is a fully-connected layer with a sigmoidal activation function that generates the output masks, that is IRM or PSM. Fig.1(b) shows how we plug our TFA module into the ResTCN block. The ResTCN block (shown in the black dotted box of Fig.1 (a)) includes three 1-D causal dilated convolution units. The kernel size, number of filters, and dilation rate for each convolutional unit are denoted as **kernel size**, **number of filters**, **dilation rate**. The dilation rate is cycled as the block index $b = \{1, 2, 3, ..., B\}$ increases: $d = 2^{(b-1 \mod (\log_2(D)+1))}$, where mod is the modulo operation and $D = 16$ is the maximum dilation rate. Each convolutional unit employs a pre-activation design, where the input is pre-activated using frame-wise alization (LN) followed by the ReLU activation function.
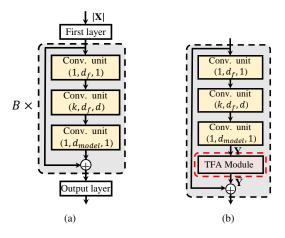

Fig. 1. Illustration of (a) the ResTCN backbone network and (b) our proposed ResTCN block with TFA module.

3.2. T-F Attention Module

In Fig. 2, we illustrate the proposed TFA module. We take a transformed T-F representation $\mathbf{Y} \in \mathbb{R}^{L \times d_{model}}$ as input of L frames and

d_{model} frequency channels. TFA utilizes two branches to generate a 1-D frequency-dimension attention map $\mathbf{F_A} \in \mathbb{R}^{1 \times d_{model}}$ and a 1-D time-frame attention map $\mathbf{T_A} \in \mathbb{R}^{L \times 1}$ in parallel, then combines them with a matrix multiplication to obtain the final 2-D T-F attention map $\mathbf{TF_A} \in \mathbb{R}^{L \times d_{model}}$. The refined output is written as:

$$\widetilde{\mathbf{Y}} = \mathbf{Y} \odot \mathbf{TF_A}, \tag{4}$$

where $\odot$ denotes the element-wise product. The detailed description of the proposed TFA attention is given below.

The energy distribution of speech along time and frequency dimension is essential for producing an accurate attention map. Each attention branch generates the attention map in two steps: global information aggregation and attention generation. Specifically, the FA module takes global average pooling along the time-frame dimension on given input $\mathbf{Y}$ and generates a frequency-wise statistic $\mathbf{Z_F} \in \mathbb{R}^{1 \times d_{model}}$, formulated as:

$$\mathbf{Z_F}(k) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{Y}(l, k), \tag{5}$$

where $\mathbf{Z^F}(k)$ is the k-th element of $\mathbf{Z_F}$. Similarly, the TA module takes global average pooling along the frequency dimension on input $\mathbf{X}$ and generates a time-frame-wise statistic $\mathbf{Z_T} \in \mathbb{R}^{L \times 1}$. The l-th element of $\mathbf{Z_T}$ is written as:

$$\mathbf{Z_T}(l) = \frac{1}{d_{model}} \sum_{k=1}^{d_{model}} \mathbf{Y}(l, k). \tag{6}$$

The two statistics $\mathbf{Z_T}$ and $\mathbf{Z_F}$ can be seen as two descriptors for the speech energy distributions in time-frame dimension and frequency dimension, respectively. To make full use of the two descriptors to produce the accurate attention weights, we stack two 1-D convolution layers of size $k_{tfa} = 17$ as the nonlinear transformation function. Specifically, the attention in the FA module is calculated as:

$$\mathbf{F_A} = \sigma(f_2^{FA}(\delta(f_1^{FA}(\mathbf{Z_F})))) \tag{7}$$

where f denotes a 1-D convolution operation, δ and σ refer to the ReLU and sigmoid activation functions, respectively. The same calculation process is applied in the TA module to generate the attention map:

$$\mathbf{T_A} = \sigma(f_2^{TA}(\delta(f_1^{TA}(\mathbf{Z_T})))). \tag{8}$$

Then, the obtained attention maps from two attention branches are combined with a tensor multiplication, producing our final 2-D attention map $\mathbf{TF_A}$ written as:

$$\mathbf{TF_A} = \mathbf{T_A} \otimes \mathbf{F_A}, \tag{9}$$

where $\otimes$ denotes the tensor multiplication operation. The (l, k)-th element of the final 2-D attention map $\mathbf{TF_A}$ is computed as:

$$\mathbf{TF_A}(l, k) = \mathbf{T_A}(l) \times \mathbf{F_A}(k) \tag{10}$$

where $\mathbf{T_A}(l)$ and $\mathbf{F_A}(k)$ denote the l-th element of $\mathbf{T_A}$ and the k-th element of $\mathbf{F_A}$, respectively.

4. EXPERIMENTS

4.1. Datasets and Feature Extraction

We use the *train-clean-100* set from the Librispeech [20] corpus as the clean speech recordings in the training set, which includes 28 539
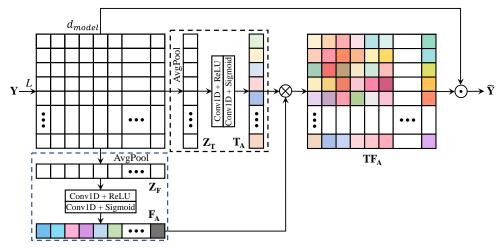
Fig. 2. Diagram of our proposed TFA module, where the TA and FA modules are shown in black and blue dotted boxes, respectively. AvgPool and Conv1D represent average pooling and 1-D convolution operation, respectively. $\otimes$ and $\odot$ denote the matrix multiplication and element-wise product, respectively.

utterances spoken by 251 speakers. The employed noise recordings in the training set are taken from the following datasets: the QUT-NOISE dataset [21], the Nonspeech dataset [22], the Environmental Background Noise dataset [23, 24], the RSG-10 dataset [25] (*voice babble*, *F16*, and *factory welding* are excluded for testing), the Urban Sound dataset [26] (*street music* recording no.26 270 is excluded for testing), the noise set from the MUSAN corpus [27], and coloured noise recordings (with an α value ranging from -2 to 2 in increments of 0.25). This gives a total 6 909 noise recordings. For the validation set, we randomly select 1 000 clean speech and noise recordings (without replacement) and remove them from the aforementioned clean speech and noise sets. Each clean speech is mixed with a random section of one noise recording at a random SNR level between -10 dB and 20 dB in 1 dB increments, which generates 1 000 noisy speech as the validation set. For the test set we utilize the four real-world noise recordings (*voice babble*, *F16*, *factory welding*, and *street music*) excluded from the RSG-10 dataset [25] and Urban Sound dataset [26]. For each of the four noise recordings, ten clean speech recordings (without replacement) randomly selected from the *test-clean-100* of Librispeech corpus [20] are mixed with a random segment of the noise recordings at the following SNR levels: $\{$-5 dB, 0 dB, 5 dB, 10 dB, 15 dB$\}$. This generates a test set of 200 noisy speech recordings. All clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz.

A square-root-Hann window function is used for spectral analysis and synthesis, with a frame-length of 32 ms and a frame-shift of 16 ms. The 257-point single-sided STFT magnitude spectrum of noisy speech, which includes both the DC frequency component and the Nyquist frequency component is used as the input.

4.2. Experimental Setup

The ResTCN model is used as the baseline backbone to validate the effectiveness of our TFA module. In addition, we also adopt two recent models as baselines, the ResTCN with self attention (ResTCN+SA) [28] and the multi-head self-attention network (MHANet) [29]. The ResTCN baseline employs the following the parameter as in [12], $k = 3$, $d_{model} = 256$, $d_f = 64$, and $B = 40$. ResTCN+SA [28] employs a multi-head self-attention module to produce dynamic representations followed by a ResTCN model

($B = 40$ stacked baseline ResTCN blocks are adopted to build the ResTCN model for a fair comparison) to perform nonlinear mapping. MHANet model [29] uses 5 stacked Transformer encoder layers [15] to perform speech enhancement, with the parameter setting as in [29]. To validate the efficacy of FA and TA components in TFA module, we conduct an ablation study, where ResTCN using FA and TA (termed ResTCN+FA and ResTCN+TA) are evaluated.

Training methodology: A mini-batch size of 10 noisy speech utterances is used for each training iteration. The noisy speech signals are created as follows: each clean speech recording selected for the mini-batch is mixed with a random section of a randomly selected noise recording at a randomly selected SNR level (-10 dB to 20 dB, in 1 dB increments). The mean squared error (MSE) between the target mask and the estimated mask is used as the objective function. For ResTCN, ResTCN+SA, and proposed models, the *Adam* optimizer with default hyper-parameters [30] and a learning rate of 0.001 are used for gradient descent optimisation. As MHANet has been found difficult to train [29, 31], we employ the training strategy as in [29]. Gradient clipping is applied to all the models, where the gradients are clipped between $[-1, 1]$.

4.3. Training & Validation Error

Fig. 3-4 show the curves of training and validation errors produced by each of the models for 150-epoch training. It can be seen that
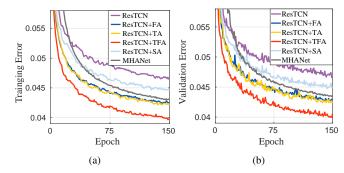

Fig. 3. The curves of training error (a) and validation error (b) on the IRM training target.

the ResTCN with our proposed TFA (ResTCN+TFA) yields significantly lower training and validation errors as compared to ResTCN,

which confirms the efficacy of the TFA module. Meanwhile, compared to ResTCN+SA and MHANet, ResTCN+TFA achieves the lowest training and validation errors, and shows obvious superiority. Among the three baselines, MHANet performs best, and ResTCN+SA outperforms ResTCN. In addition, the comparisons among ResTCN, ResTCN+FA, and ResTCN+TA demonstrate the efficacy of TA and FA modules.
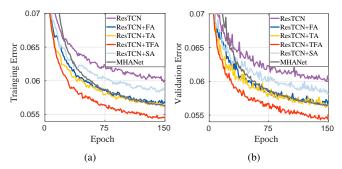

Fig. 4. The curves of training error (a) and validation error (b) on the PSM training target.

4.4. Results and Discussion

In this study, five metrics are used for extensive evaluations of enhancement performance, including wideband perceptual evaluation of speech quality (PESQ) [32], extended short-time objective intelligibility (ESTOI) [33], and three composite metrics [34], which are mean opinion score (MOS) predictors of the signal distortion (CSIG), background-noise intrusiveness (CBAK), and overall signal quality (COVL).

Table 1. Average PESQ (wideband version) scores for each SNR level. The highest PESQ scores are highlighted in boldface.

Target	Network	# Params	Input SNR (dB)				
			-5	0	5	10	15
-	Noisy		1.05	1.07	1.13	1.31	1.64
IRM	ResTCN+SA [28]	2.24M	1.15	1.34	1.64	2.07	2.51
	MHANet [29]	4.08M	1.16	1.36	1.68	2.10	2.56
	ResTCN [12]	1.98M	1.13	1.32	1.61	2.06	2.50
	ResTCN+FA	+1.36K	1.18	1.39	1.74	2.15	2.60
	ResTCN+TA	+1.36K	1.18	1.40	1.75	2.19	2.63
	ResTCN+TFA	+2.72K	**1.20**	**1.43**	**1.79**	**2.27**	**2.70**
PSM	ResTCN+SA [28]	2.24M	1.20	1.42	1.79	2.26	2.75
	MHANet [29]	4.08M	1.20	1.44	1.83	2.30	2.78
	ResTCN [12]	1.98M	1.19	1.40	1.76	2.23	2.72
	ResTCN+FA	+1.36K	1.20	1.47	1.87	2.36	2.85
	ResTCN+TA	+1.36K	1.21	1.49	1.90	2.38	2.85
	ResTCN+TFA	+2.72K	**1.26**	**1.54**	**1.96**	**2.44**	**2.91**

Tables 1 and 2 present the average PESQ and ESTOI scores for each SNR level (across four noise sources), respectively. The evaluation results show that our proposed ResTCN+TFA consistently achieves significant improvements over ResTCN in terms of PESQ and ESTOI on IRM and PSM, with negligible parameter overhead, which demonstrates the effectiveness of the TFA module. In the 5 dB SNR case, for instance, ResTCN+TFA with IRM improves the baseline ResTCN by 0.18 in PESQ and by 4.94% in ESTOI. Compared to MHANet and ResTCN+SA, ResTCN+TFA

Table 2. Average ESTOI scores (in %) for each SNR level. The highest ESTOI scores are highlighted in boldface.

Target	Network	# Params	Input SNR (dB)				
			-5	0	5	10	15
-	Noisy		27.91	42.14	57.21	71.11	82.22
IRM	ResTCN+SA [28]	2.24M	43.32	60.68	74.37	83.67	89.77
	MHANet [29]	4.08M	43.72	61.08	74.68	83.90	89.96
	ResTCN [12]	1.98M	42.47	59.93	73.67	83.22	89.44
	ResTCN+FA	+1.36K	44.87	61.73	75.06	84.27	90.09
	ResTCN+TA	+1.36K	46.11	62.96	75.86	84.48	90.21
	ResTCN+TFA	+2.72K	**47.61**	**64.05**	**78.61**	**85.45**	**90.83**
PSM	ResTCN+SA [28]	2.24M	43.03	61.13	75.07	84.24	90.07
	MHANet [29]	4.08M	44.66	62.32	75.74	84.69	90.43
	ResTCN [12]	1.98M	42.76	60.45	74.47	83.91	89.84
	ResTCN+FA	+1.36K	44.91	62.38	75.88	84.82	90.45
	ResTCN+TA	+1.36K	45.70	63.25	76.49	85.09	90.70
	ResTCN+TFA	+2.72K	**48.67**	**65.28**	**77.43**	**85.80**	**91.16**

performs best in all cases, and shows obvious performance superiority. Among the three baselines, overall, the performance rank is in order MHANet > ResTCN+SA > ResTCN. Meanwhile, ResTCN+FA and ResTCN+TA also provide substantial improvements over ResTCN, which confirms the efficacy of FA and TA modules. Table 3 lists the average CSIG, CBAK, and COVL scores across all of the test conditions. Similar performance trends are observed to those in Tables 1 and 2. Again, our proposed ResTCN+TFA significantly outperforms ResTCN in three metrics, and performs best among all models. On average, ResTCN+TFA with PSM improves CSIG by 0.21, CBAK by 0.12, and COVL by 0.18 over ResTCN. Compared to MHANet, ResTCN+TFA with PSM improves CSIG by 0.12, CBAK by 0.08, and COVL by 0.11.

Table 3. Average scores of CSIG, CBAK, and COVL across all test conditions. The highest scores are highlighted in boldface.

Network	IRM			PSM		
	CSIG	CBAK	COVL	CSIG	CBAK	COVL
Noisy	2.26	1.80	1.67	-	-	-
ResTCN+SA [28]	3.10	2.45	2.37	3.14	2.53	2.46
MHANet [29]	3.13	2.46	2.40	3.21	2.56	2.51
ResTCN [12]	3.08	2.43	2.35	3.12	2.52	2.44
ResTCN+FA	3.21	2.51	2.48	3.25	2.61	2.55
ResTCN+TA	3.23	2.53	2.49	3.27	2.58	2.57
ResTCN+TFA	**3.28**	**2.56**	**2.54**	**3.33**	**2.64**	**2.62**

5. CONCLUSION

In this study, we propose a light-weight and flexible attention unit termed TFA module, which is designed to model the energy distribution of speech in the T-F representation. Extensive experiments with ResTCN as a backbone on two training targets (IRM and PSM) demonstrate the effectiveness of the proposed TFA module. Among all the models, our proposed ResTCN+TFA consistently performs best and significantly outperforms other baselines in all of the cases. Future research work includes investigating the efficacy of TFA on more architectures (e.g., more recent Transformer) and more training targets.

6. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] Q. Zhang, M. Wang, Y. Lu, L. Zhang, and M. Idrees, "A novel fast nonstationary noise tracking approach based on mmse spectral power estimator," *Digital Signal Processing*, vol. 88, pp. 41–52, 2019.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[4] Q. Zhang, M. Wang, Y. Lu, M. Idrees, and L. Zhang, "Fast nonstationary noise tracking based on log-spectral power mmse estimator and temporal recursive averaging," *IEEE Access*, vol. 7, pp. 80 985–80 999, 2019.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," *Proc. INTERSPEECH*, pp. 3642–3646, 2017.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[9] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2015.

[10] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[11] 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.

[12] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[13] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, 2018.

[14] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. ICASSP*, 2019, pp. 6875–6879.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[18] V. A. Trinh, B. McFee, and M. I. Mandel, "Bubble cooperative networks for identifying important speech cues," in *Proc. Interspeech*, 2018.

[19] Q. Zhang, Q. Song, A. Nicolson, T. Lan, and H. Li, "Temporal Convolutional Network with Frequency Dimension Adaptive Attention for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 166–170.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[21] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. INTERSPEECH*, 2010.

[22] G. Hu, "100 nonspeech environmental sounds," *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[23] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," in *Proc. ICASSP*, 2016, pp. 2204–2208.

[24] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *Proc. EMBC*, 2016, pp. 736–739.

[25] H. J. Steeneken and F. W. Geurtsen, "Description of the rsg-10 noise database," *report IZF*, vol. 3, p. 1988, 1988.

[26] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM-MM*, 2014, pp. 1041–1044.

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

[28] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.

[29] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," in *Proc. EMNLP*, 2020, pp. 5747–5763.

[32] R. I.-T. P. ITU, "862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs. ITU-Telecommunication standardization sector, 2007."

[33] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[34] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. process.*, vol. 16, no. 1, pp. 229–238, 2007.