INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Temporal Convolutional Network with Frequency Dimension Adaptive Attention for Speech Enhancement

Qiquan Zhang[1] *, Qi Song*[2] *, Aaron Nicolson*[3]*, Tian Lan*[2]*, Haizhou Li*[1]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Alibaba Group, China
[3]Australian eHealth Research Centre, CSIRO, Herston, QLD, 4006, Australia

zhangqiquan_hit@163.com, qiqi.sq@alibaba-inc.com, aaron.nicolson@csiro.au,
congshi.lt@alibaba-inc.com, haizhou.li@nus.edu.sg

Abstract

Despite much progress, most temporal convolutional networks (TCN) based speech enhancement models are mainly focused on modeling the long-term temporal contextual dependencies of speech frames, without taking into account the distribution information of speech signal in frequency dimension. In this study, we propose a frequency dimension adaptive attention (FAA) mechanism to improve TCNs, which guides the model selectively emphasize the frequency-wise features with important speech information and also improves the representation capability of network. Our extensive experimental investigation demonstrates that the proposed FAA mechanism is able to consistently provide significant improvements in terms of speech quality (PESQ), intelligibility (STOI) and three other composite metrics. More promisingly, it has better generalization ability to real-world noisy environment.

Index Terms: Speech enhancement, adaptive attention mechanism, temporal convolutional networks

1. Introduction

Speech enhancement (SE) seeks to remove the background noise from the noisy speech signal, so as to improve the perceptual quality and intelligibility of speech. SE plays an important role in many applications, such as hearing aids, speech recognition, and speaker verification. Statistical model-based SE has been an attractive research direction for a long time, with representative methods including Wiener filtering [1], minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [2], and MMSE log-spectral amplitude (MMSE-LSA) estimator [3]. These methods often make some underlying assumptions [4], and lack the ability to handle the non-stationary noise. [1]

In recent years, deep neural networks (DNNs)-based supervised SE methods have demonstrated great progress. In the supervised SE methods, the design of training targets plays an important role. In the early study, the ideal binary mask (IBM) [5] is inspired by the concept of time-frequency (T-F) masking in computational auditory scene analysis. Subsequently, other training targets in the T-F domain are proposed, including masking-based targets and mapping-based targets. The most popular masking-based targets include ideal ratio mask (IRM) [6], phase-sensitive mask (PSM) [7], and complex IRM (cIRM) [8]. The most widely used mapping-based targets are target magnitude spectrum, log-power spectrum, and complex spectrum. More recently, time domain solutions [9–11] become popular, where the DNN is trained to directly predict the raw waveform of clean speech from raw waveform of noisy speech. The most prominent neural solution of this kind is the Conv-Tasnet [10], which achieves the state-of-the-art performance in speaker separation tasks. Taking the speech quality [12], robustness, and computational efficiency into consideration, the T-F domain methods remain the most attractive solution today.

Deep learning for statistical model-based methods has also been a popular research topic recently [13–16]. The performance of statistical model-based methods is directly impacted by the *a priori* signal-to-noise ratio (SNR). Motivated by the fact, in [14, 15] we proposed a DNN-based *a priori* SNR estimator (Deep Xi)[2] that bridges the gap between deep learning and statistical model-based methods. In Deep Xi framework, the mapped *a priori* SNR is presented as the training target to improve the convergence rate of the gradient descent algorithm. In this study, we use two training targets to conduct speech enhancement: IRM [6] and mapped *a priori* SNR [14, 15].

For DNN-based SE, temporal contextual information is able to facilitate speaker generalization [17]. Multi-layer perceptrons (MLPs) model typically exploits a context window of consecutive frames to obtain the contextual information. However, the context window cannot provide long-term contextual information. Subsequently, Chen *et. al* [17] proposed to use a recurrent neural network (RNN) with four hidden long short-term memory (LSTM) layers to capture the long-term dependencies. The LSTM model shows a significant superiority than MLPs model, but the high latency and the training complexity also limit its applicability. Recently, a model called TCN [18], which uses residual learning incorporated with dilated convolution units, can capture long-term contextual dependencies and outperforms LSTM models in several sequence modeling tasks. TCNs have become tremendously popular in speech enhancement [11,15,19] with the advantages of parallel processing, low training complexity and memory required.

However, TCN is unable to model the speech distribution along the frequency dimension, which is equally important to preserve the structure of speech signals. Recently, the study of channel attention [20, 21] has brought significant performance gains in computer vision tasks. Such attention module tells the model where the important channel features are. Motivated by the channel attention study in computer vision [20], we

[1]This research work was supported by the Science and Engineering Research Council, Agency of Science, Technology and Research, Singapore, through the National Robotics Program under Grant No. 1922500054.

[2]*Availability*: The mentioned Deep Xi framework is available at: https://github.com/anicolson/DeepXi.

http://dx.doi.org/10.21437/Interspeech.2021-46

present a frequency dimension adaptive attention module under the TCN framework, which dynamically generates differentiated weights to spectral elements according to their significance in speech.

The remainder of this paper is structured as follows. In Section 2 we describe the training targets used in this work. The description of our proposed model is given in Section 3. We present the experiment set up and results in Section 4. The conclusion is given in Section 5.

2. Training Targets

We assume additive noise in the short-time Fourier transform (STFT) domain:

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

where $Y[l,k]$, $S[l,k]$, 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.

We consider two training targets to conduct speech enhancement, i.e. the IRM [6] and the mapped *a priori* SNR [14, 15].

IRM is one of most popular masking-based training targets, and it 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.

In Deep Xi SE framework [14,15], we proposed the mapped *a priori* SNR as training targets, which is a mapped version of the instantaneous *a priori* SNR (in dB), $\xi_{dB}\,[l,k] = 10\log_{10}(\xi[l,k])$ and $\xi[l,k]$ is given by:

$$\xi[l,k] = \frac{|S[l,k]|^2}{|D[l,k]|^2} \tag{3}$$

Deep Xi utilizes the cumulative distribution function (CDF) of $\xi_{dB}\,[l,k]$ as the map [14, 15]. The estimated *a priori* SNR $\widehat{\xi}[l,k]$ can be flexibly applied to gain functions of statistical model-based SE methods to reconstruct the clean speech. In this study, we exploit the most widely used MMSE-STSA gain function [2] for performance evaluation. The MMSE-STSA gain function is given by:

$$G_{\text{MMSE-STSA}}[l,k] = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v[l,k]}}{\gamma[l,k]}\exp\left(\frac{-v[l,k]}{2}\right) \left((1+v[l,k])I_0\left(\frac{v[l,k]}{2}\right) + v[l,k]I_1\left(\frac{v[l,k]}{2}\right)\right) \tag{4}$$

where I_0 and I_1 are the modified Bessel function of zero and first order, respectively, and $v[l,k] = (\xi[l,k]/(\xi[l,k]+1))\gamma[l,k]$[1].

3. Model Description

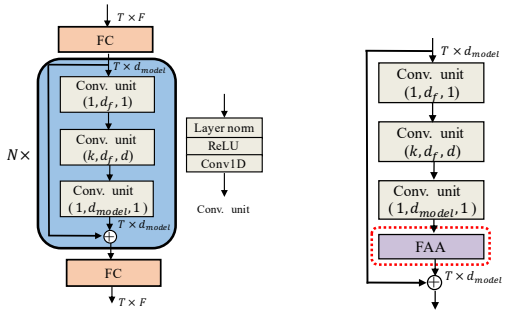
A square-root-Hann window function is used for spectral analysis and synthesis, with a frame-length of 32 ms (512 time-domain samples) and frame shift of 16 ms (256 time-domain samples). We use the 257-point noisy speech magnitude spectrum as the network input, which include both the DC frequency component and the Nyquist frequency component.

[1]In Deep Xi framework, the maximum likelihood *a posteriori* SNR estimator is used: $\hat{\gamma}[l,k] = \hat{\xi}[l,k] + 1$

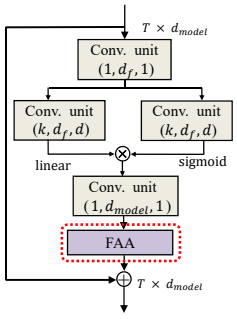
3.1. Proposed TCN-FAA and GaTCN-FAA Models

TCN is known to be effective in speech enhancement [11, 15, 19]. Let's recap the bottleneck TCN model [15] in Fig. 1(a), and its workflow next. The network input is the noisy speech magnitude spectrum with shape $T \times F$, where T is the number of time frames and $F = 257$ is that of the frequency bins. The input is first transformed by FC, a fully-connected layer of size $d_{\text{model}} = 256$ that includes layer normalisation [22] followed by the rectified linear unit (ReLU) [23] activation function, leading to a time-frequency representation of $T \times d_{model}$. The FC layer is followed by $N = 40$ widely used bottleneck TCN blocks, and each block contains three one-dimensional causal dilated convolutional units.

Each convolutional unit is pre-activated by layer normalisation followed by the ReLU activation function. The first and third convolutional units in each block have a kernel size of 1, whilst the second convolutional unit has a kernel size of $k = 3$. 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 long-range contextual field over previous speech frames. The dilation rate d is cycled as the block index n increases:$d = 2^{(n-1 \bmod (\log_2(D)+1)}$ [10, 15], where mod is the modulo operation, $n = 1, 2, ..., N$ and $D = 16$ is the maximum dilation rate. The last block is followed by the output layer, which is a FC layer with sigmoidal units.



(a) The bottleneck TCN model (b) The proposed bottleneck TCN model with a FAA block (TCN-FAA)



(c) The proposed GaTCN model with a FAA block (GaTCN-FAA)

Figure 1: *Illustration of the bottleneck TCN model, bottleneck TCN-FAA model and GaTCN-FAA model. Conv1D denotes one-dimension convolutional unit. The kernel size, output size, and dilation rate for each convolutional unit is denoted as (**kernel size**, **output size**, **dilation rate**).*

While TCN effectively models the long-term contextual information of speech, it doesn't model the energy distribution

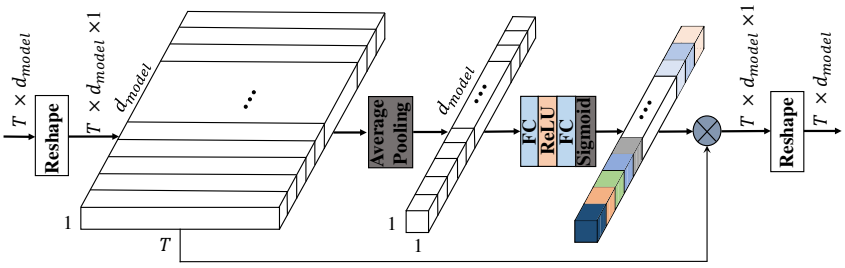


Figure 2: *Illustration of the FAA module, and the symbol $\otimes$ represents the element-wise multiplication.*

along the frequency dimension. To alleviate the limitation, we propose a frequency dimension adaptive attention (FAA) module under the TCN framework, which assigns differentiated weights dynamically to the d_{model} frequency elements. The FAA module is shown in Fig. 1(b)-(c) (highlighted in red dotted box) is easily integrated into TCN blocks. In [19] a gated bottleneck TCN residual block is proposed by incorporating a gating mechanism to alleviate the vanishing gradient problem.

We study the effectiveness of the proposed FAA module in two TCN frameworks, namely bottleneck TCN block (denoted as TCN) [15] and the gated bottleneck TCN block (denoted as GaTCN) [19], for speech enhancement.

3.2. Frequency Dimension Adaptive Attention (FAA)

Recently, channel attention has achieved impressive performance in many computer vision applications [20, 21]. One of the representative attention modules is Squeeze-and-Excitation attention module [20], which consists of a squeeze operation and an excitation operation. The squeeze operation utilizes global averaging pooling to access global spatial information. The excitation operation utilizes the accessed information to generate the weights to perform feature recalibration. In speech enhancement, we consider that the energy distribution in frequency dimension is as informative as the temporal contextual information. Inspired by the Squeeze-and-Excitation module, we introduce a frequency dimension adaptive attention (FAA) module to the TCN frameworks, which allows the TCN models to adaptively assign differentiated weights to different frequency elements. The proposed FAA module is illustrated in Fig. 2.

Given an intermediate input $\mathbf{F} \in \mathbb{R}^{T \times d_{\text{model}}}$, $\mathbf{F}$ is firstly reshaped into the shape of $1 \times T \times d_{\text{model}}$. Subsequently, the squeeze operation aggregates the global information of frequency bins using global averaging pooling, generating the statistics of frequency bins $\mathbf{F}^{\text{c}}_{\text{avg}}$, which denotes averaged-pooled features. The aggregated features in squeeze operation are then forwarded to an MLP network with two FC layers to generate the weights $\boldsymbol{\omega} \in \mathbb{R}^{1 \times 1 \times d_{\text{model}}}$ for frequency bins. To reduce parameters and computational cost, the size of hidden layer is set to $\mathbb{R}^{1 \times 1 \times d_{\text{model}}/r}$, where r denotes the reduction ratio. In short, the weights of frequency bins are computed as

$$\boldsymbol{\omega} = \sigma\left(f_{\{\mathbf{W}_1, \mathbf{W}_2\}}(\mathbf{F}^{\text{c}}_{\text{avg}})\right) \tag{5}$$

where σ denotes the sigmoid function and $f_{\{\mathbf{W}_1, \mathbf{W}_2\}}(\mathbf{F}^{\text{c}}_{\text{avg}})$ takes the form

$$f_{\{\mathbf{W}_1, \mathbf{W}_2\}}(\mathbf{F}^{\text{c}}_{\text{avg}}) = \mathbf{W}_2 \eta\left(\mathbf{W}_1 \mathbf{F}^{\text{c}}_{\text{avg}}\right) \tag{6}$$

where η denotes the ReLU activation function, $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{model}}/r \times d_{\text{model}}}$, and $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}/r}$ denote the MLP weights. In this work, the reduction ratio r is set to 4. The obtained attention weights are assigned to corresponding frequency bins to recalibrate features. The FAA is a lightweight and general module, which can be easily integrated into TCNs.

4. Experiments

4.1. Dataset

The training data set consists of 74,250 clean speech recordings and 2,382 noise recordings. Each clean speech recording is mixed with a random section of a randomly selected noise recording at a random SNR (-10 dB to 20 dB, in 1 dB increments). The clean speech recordings come from the train-clean-100 set of LibriSpeech [24] (28,539 utterances) and CSTR VCTK corpus [25] (42,015 utterances), and the *si** and *sx** training sets of TIMIT corpus [26] (3,696 utterances). The noise recordings from the following noise datasets are included in the training set: the Environmental Background Noise dataset [27, 28], the Nonspeech dataset [29], multiple FreeSound packs[2], the QUT-NOISE dataset [30], the noise set

[2]Freesound packs that are used: 147, 199, 247, 379, 622, 643, 1 133, 1 563, 1840, 2432, 4366, 4439, 15,046, 15,598, 21,558.

Table 1: *Performance comparisons of neural speech enhancement networks in terms of PESQ for Test Set 1.*

Network	SNR level (dB)																								
	IRM													Mapped a *Priori* SNR											
	Voice babble			Street music			F16			Factory			Ave	Voice babble			Street music			F16			Factory			Ave
	-5	5	15	-5	5	15	-5	5	15	-5	5	15		-5	5	15	-5	5	15	-5	5	15	-5	5	15	
Noisy speech	1.00	1.83	2.60	1.15	1.88	2.59	1.29	1.88	2.58	1.05	1.68	2.49	1.83	-	-	-	-	-	-	-	-	-	-	-	-	-
ResLSTM	1.23	2.26	2.92	1.50	2.40	3.00	1.63	2.41	2.97	1.08	2.21	2.76	2.21	1.11	2.20	2.99	1.34	2.28	2.90	1.44	2.41	2.98	1.09	2.40	2.88	2.17
ResBLSTM	1.25	2.27	2.96	1.53	2.41	3.02	1.61	2.44	2.95	1.10	2.24	2.79	2.22	1.24	2.24	3.01	1.43	2.31	2.91	1.54	2.45	3.01	1.11	2.38	2.98	2.22
TCN	1.30	2.35	3.09	1.50	2.30	3.01	1.58	2.40	2.97	1.26	2.21	2.88	2.26	1.29	2.33	3.02	1.36	2.22	2.93	1.68	2.5	3.06	1.21	2.31	3.01	2.26
TCN-FAA	**1.42**	**2.44**	**3.15**	**1.72**	**2.55**	**3.17**	**1.87**	**2.67**	**3.23**	**1.54**	**2.39**	**3.11**	**2.46**	**1.42**	**2.48**	**3.25**	**1.72**	**2.64**	**3.24**	**1.99**	**2.82**	**3.35**	**1.66**	**2.49**	**3.10**	**2.53**
GaTCN	1.35	2.39	3.10	1.60	2.40	2.96	1.61	2.52	3.11	1.33	2.24	3.03	2.33	1.34	2.43	3.14	1.59	2.50	3.09	1.64	2.56	3.13	1.27	2.22	3.08	2.35
GaTCN-FAA	**1.46**	**2.45**	**3.19**	**1.68**	**2.54**	**3.15**	**1.92**	**2.67**	**3.23**	**1.61**	**2.53**	**3.14**	**2.48**	**1.24**	**2.41**	**3.17**	**1.47**	**2.54**	**3.15**	**1.88**	**2.71**	**3.31**	**1.26**	**2.44**	**3.16**	**2.42**

Table 2: *Performance comparisons of neural speech enhancement networks in terms of STOI (in %) for Test Set 1.*

Network	SNR level (dB)																								
	IRM													Mapped a *Priori* SNR											
	Voice babble			Street music			F16			Factory			Ave	Voice babble			Street music			F16			Factory			Ave
	-5	5	15	-5	5	15	-5	5	15	-5	5	15		-5	5	15	-5	5	15	-5	5	15	-5	5	15	
Noisy speech	60.2	83.0	95.5	59.0	81.9	95.6	60.4	82.4	95.7	57.8	80.8	94.5	79.6	-	-	-	-	-	-	-	-	-	-	-	-	-
ResLSTM	63.0	87.9	96.0	67.9	88.8	96.8	66.1	87.3	95.2	59.0	85.0	95.2	82.3	58.0	85.8	96.4	64.5	86.1	96.2	62.8	86.9	95.9	58.1	84.5	94.2	80.8
ResBLSTM	62.7	87.0	96.2	68.1	88.5	96.9	66.3	87.9	95.6	59.2	85.4	95.0	82.4	57.6	85.7	96.5	64.6	86.2	96.0	65.0	87.0	96.0	58.5	85.4	94.6	81.1
TCN	64.7	88.3	97.1	68.5	87.9	96.7	66.8	87.9	95.5	61.2	85.7	95.1	84.1	62.1	87.1	96.8	66.3	86.8	96.5	66.5	88.2	96.1	59.3	85.7	94.3	83.4
TCN-FAA	**67.4**	**90.1**	**97.3**	**71.5**	**90.5**	**97.4**	**74.0**	**91.2**	**97.6**	**67.5**	**87.3**	**96.2**	**86.7**	**63.7**	**89.1**	**97.3**	**71.1**	**90.0**	**97.4**	**72.7**	**90.8**	**97.4**	**67.5**	**87.1**	**94.8**	**86.0**
GaTCN	65.7	89.1	97.1	69.4	88.9	96.5	68.8	89.3	97.0	62.5	86.4	96.0	85.1	62.7	88.2	97.2	67.4	89.4	97.1	67.2	89.2	96.5	59.6	84.0	96.1	84.0
GaTCN-FAA	**68.0**	**89.6**	**97.4**	**71.7**	**90.5**	**97.5**	**74.3**	**91.3**	**97.6**	**68.4**	**89.0**	**96.8**	**87.0**	**63.4**	**89.2**	**97.4**	**71.2**	**90.1**	**97.2**	**69.8**	**89.8**	**97.2**	**60.1**	**86.7**	**96.6**	**85.1**

of the MUSAN corpus [31], and coloured noise recordings (with an α value ranging from -2 to 2 in increments of 0.25). Each extracted 5% of the dataset to form the validation set. For testing, firstly to maintain the continuity of work, we adopted the same test set as [14, 32] as the Test Set 1. Also, we mix the ITU standard utterances with recorded noise in the real environment and obtain 200 utterances to form the Test Set 2, to study whether the proposed models generalize well to real noisy environment. All audio recordings used in the experiment are single-channel, with a sampling rate of 16 kHz.

Table 3: *Performance comparisons in terms of three composite metrics for Test Set 1.*

Network	IRM			Mapped a *Priori* SNR		
	CSIG	CBAK	COVL	CSIG	CBAK	COVL
Noisy speech	2.45	1.83	1.75	-	-	-
ResLSTM	2.92	2.27	2.21	2.89	2.28	2.14
ResBLSTM	3.00	2.28	2.21	2.97	2.30	2.19
TCN	3.06	2.34	2.26	3.07	2.40	2.29
TCN-FAA	**3.27**	**2.50**	**2.45**	**3.29**	**2.57**	**2.51**
GaTCN	3.14	2.41	2.33	3.12	2.42	2.34
GaTCN-FAA	**3.28**	**2.52**	**2.47**	**3.18**	**2.48**	**2.41**

4.2. Baselines and Training Methodology

We compare our proposed TCN-FAA and GaTCN-FAA networks with four baselines, which include a residual LSTM (ResLSTM) network [14], a residual bidirectional LSTM (ResBLSTM) network [14], a TCN model [15] and a GaTCN model [19]. The ResLSTM consists of 5 residual blocks, with each block containing a LSTM cell with a size of 512. The ResBLSTM is identical to the ResLSTM, except that the residual blocks include both a forward and backward LSTM cell, each with a cell size of 512. TCN and GaTCN models contain 40 residual blocks. For mapped *a priori* SNR training target, the last layer applies a sigmoid activation, and uses cross entropy as the cost function, and for IRM, the last layer applies a linear activation, and uses the mean square error as the cost function. In the training stage, we used the Adam algorithm with default hyper-parameters settings [33] for gradient descent optimisation and the mini-batch size was set to 10.

4.3. Experimental Evaluation

In this work, we evaluate the speech enhancement performance in terms of five widely used metrics: perceptual evaluation of speech quality (PESQ) [34], short-time objective intelligibility (STOI) [35] , a composite measure for signal distortion (CSIG), a composite measure for noise distortion (CBAK) and a composite measure for overall speech quality (COVL) [36]. For Test Set 1, the PESQ and STOI results are shown in Table 1 and Table 2. Similar performance trend is observed under two training targets, we take the IRM training target as an example. It is obvious that the proposed TCN-FAA and GaTCN-FAA achieve significant improvements over all the baseline models. According to Table 1, the proposed FAA module yields 0.2 and 0.15 PESQ improvements over the TCN and GaTCN baselines on average, respectively. In terms of STOI, 2.6% and 1.9% improvements are obtained on average. The three composite metrics results are shown in Table 3, it also can be seen that TCN-FAA and GaTCN-FAA consistently outperform the original baseline TCN and GaTCN in all metrics under both training targets, which demonstrates the effectiveness of the proposed FAA module.

Table 4: *Performance comparisons in terms of CSIG, CBAK, COVL, PESQ and STOI metrics for Test Set 2.*

Network	IRM					Mapped a *Priori* SNR				
	CSIG	CBAK	COVL	PESQ	STOI	CSIG	CBAK	COVL	PESQ	STOI
Noisy speech	1.92	1.73	1.51	1.90	86.7	-	-	-	-	-
ResLSTM	3.07	2.30	2.38	2.54	90.0	3.19	2.41	2.53	2.62	89.1
ResBLSTM	3.05	2.32	2.39	2.55	90.1	3.15	2.37	2.47	2.62	88.9
TCN	3.10	2.33	2.42	2.57	91.0	3.21	2.43	2.56	2.64	90.2
TCN-FAA	**3.22**	**2.46**	**2.57**	**2.67**	**92.6**	**3.37**	**2.51**	**2.68**	**2.74**	**92.1**
GaTCN	3.14	2.38	2.48	2.61	91.3	3.28	2.49	2.62	2.69	90.8
GaTCN-FAA	**3.28**	**2.47**	**2.59**	**2.70**	**91.6**	**3.39**	**2.54**	**2.70**	**2.79**	**91.0**

Table 4 presents the comparison results on Test Set 2, it can be clearly observed that TCN-FAA and GaTCN-FAA outperform all the baseline models across five metrics by a comfortable margin under two training targets. For example, under IRM training target, TCN-FAA provides 0.12, 0.13, 0.15, 0.10 and 1.6% improvements over TCN for CSIG, CBAK, COVL, PESQ and STOI, respectively. These evaluation results prove the excellent generalization ability of the proposed frequency dimension adaptive attention mechanism to real noisy environment.

5. Conclusions

In this paper, we proposed a frequency dimension adaptive attention module (FAA) to improve TCN models for monaural speech enhancement task. The FAA module is integrated into TCN and GaTCN to perform speech enhancement. Moreover, the experiment evaluations are conducted on two training targets. The results of objective metrics show that the proposed TCN-FAA and GaTCN-FAA consistently outperform their base models without FAA module. More promisingly, it also shows excellent generalization ability to real noisy environment.

6. References

[1] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] 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.

[5] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[6] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[8] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

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

[10] 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.

[11] 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.

[12] R. Liu, B. Sisman, G. l. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021.

[13] Y. Xia and R. Stern, "A priori snr estimation based on a recurrent neural network for robust speech enhancement." in *INTERSPEECH*, 2018, pp. 3274–3278.

[14] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[15] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[16] Y.-H. Tu, I. Tashev, S. Zarar, and C.-H. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2531–2535.

[17] 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.

[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] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 189–198, 2018.

[20] 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.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[22] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[27] 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 *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2204–2208.

[28] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 736–739.

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

[30] 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," *Proceedings of Interspeech 2010*, 2010.

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

[32] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, "Deep residual-dense lattice network for speech enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8552–8559.

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

[34] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[35] 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.

[36] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.