A Time-Frequency Attention Module for Neural Speech Enhancement

Qiquan Zhang, *Member, IEEE*, Xinyuan Qian, *Member, IEEE*, Zhaoheng Ni, Aaron Nicolson, Eliathamby Ambikairajah, *Senior Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Speech enhancement plays an essential role in a wide range of speech processing applications. Recent studies on speech enhancement tend to investigate how to effectively capture the long-term contextual dependencies of speech signals to boost performance. However, these studies generally neglect the time-frequency (T-F) distribution information of speech spectral components, which is equally important for speech enhancement. In this paper, we propose a simple yet very effective network module, which we term the T-F attention (TFA) module, that uses two parallel attention branches, i.e., time-frame attention and frequency-channel attention, to explicitly exploit position information to generate a 2-D attention map to characterise the salient T-F speech distribution. We validate our TFA module as part of two widely used backbone networks (residual temporal convolution network and Transformer) and conduct speech enhancement with four most popular training objectives. Our extensive experiments demonstrate that our proposed TFA module consistently leads to substantial enhancement performance improvements in terms of the five most widely used objective metrics, with negligible parameter overheads. In addition, we further evaluate the efficacy of speech enhancement as a front-end for a downstream speech recognition task. Our evaluation results show that the TFA module significantly improves the robustness of the system to noisy conditions.**

***Index Terms*—Speech enhancement, time-frequency attention, ResTCN, transformer.**

Manuscript received 7 May 2022; revised 22 October 2022 and 19 November 2022; accepted 21 November 2022. Date of publication 30 November 2022; date of current version 15 December 2022. This work was supported in part by ARC Discovery under Grant DP1900102479, in part by the Internal Project of Shenzhen Research Institute of Big Data under Grant T00120220002, in part by the Guangdong Provincial Key Laboratory of Big Data Computing under Grant B10120210117-KP02, in part by University Development Fund under Grants UDF01002333 and UF02002333, in part by the Research Foundation of Guangdong Province under Grant 2019A050505001, in part by the Science and Engineering Research Council, Agency for Science, Technology and Research (A*STAR), Singapore, through the National Robotics Program under Human-Robot Interaction Phase 1 under Grant 1922500054, and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Xinyuan Qian.)*

Qiquan Zhang is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, also with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: qiquan.zhang@unsw.edu.au).

Xinyuan Qian is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, and also with The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: eleqian@nus.edu.sg).

Zhaoheng Ni is with Meta, New York, NY 10003 USA (e-mail: zni@fb.com).

Aaron Nicolson is with the Australian e-Health Research Centre, CSIRO, Black Mountain, ACT 2601, Australia (e-mail: aaron.nicolson@csiro.au).

Eliathamby Ambikairajah is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: e.ambikairajah@unsw.edu.au).

Haizhou Li is with the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen 518172, China, also with the Shenzhen Research Institute of Big data, Shenzhen, 51872, China, also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, 119077, also with the University of Bremen, 28359, Bremen, Germany, and also with Kriston AI Lab, Xiamen, China (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/TASLP.2022.3225649

I. Introduction

SPEECH signals in a real-world acoustic environment are inevitably corrupted by background noise, which can severely degrade speech quality and intelligibility. Speech enhancement seeks to separate the target speech signal from the background noise. It is an essential component in a number of speech processing systems, such as hearing aids, automatic speech recognition (ASR), speaker verification, and the brain-computer interface. Monaural speech enhancement represents one of the challenges. Traditional signal processing-based methods have been extensively studied for a long time, mainly including spectral subtraction [1], Wiener filtering [2], and statistical model-based methods [3], [4], [5], [6]. These methods perform well for stationary noise, however, fail to handle non-stationary noise.

With the advent of deep learning, speech enhancement has made remarkable progress [7]. Techniques can be grouped into time-frequency (T-F) domain methods and time-domain methods, according to the way input signals are handled. Specifically, time-domain methods perform speech enhancement directly on the raw waveform domain, where a deep neural network (DNN) is optimized to learn the mapping from the noisy raw waveform to the clean one [8], [9], [10], [11], [12] via some latent feature representation. T-F domain methods typically transform the noisy raw waveform into a T-F representation or spectrogram first before mapping to the clean one with well-designed training objectives. The most popular T-F domain training objectives include ideal ratio mask (IRM) [13], spectral magnitude mask (SMM) [13], complex IRM (cIRM) [14], phase-sensitive mask (PSM) [15], target magnitude spectrum [16], and log-power spectrum [17]. More recently, the instantaneous a priori signal-to-noise ratio (SNR), termed Xi, is proposed as a training objective to bridge the gap between deep learning and traditional statistical model-based methods [18], [19]. In this study, we adopt a novel T-F domain method, which allows for intuitive

time-frequency analysis. We also adopt the three most widely used training objectives, i.e., IRM, SMM, PSM and a recent one (Xi) in the experiments.

Advanced speech enhancement algorithms depend on a strong backbone network architecture. Multi-layer perceptrons (MLPs) are the most widely adopted backbone network architecture in early studies. Furthering the idea of T-F masking in the computational auditory scene analysis, Wang et al. [20] proposed to employ an MLP to predict the ideal binary mask (IBM) [21] to separate the speech from background noise. Subsequently, Xu et al. [17] adopted an MLP as a regression function to learn the mapping from the noisy log-power spectra to the clean one. In [22], Chen et al. formulated speech enhancement as a sequence-to-sequence mapping, that effectively addresses speaker generalization issue, and employed a recurrent neural network (RNN) with four long short-term memory (LSTM) layers to model the long-range contextual information of the speech. The LSTM-RNN model demonstrates substantial performance improvement over MLPs [22], [23], [24]. However, deep LSTM-RNN network architectures involve a large number of parameters, which significantly limits its scope of applications.

Deep convolutional neural networks (CNNs) represent another successful backbone network architecture. Unlike RNN that involves a sequential process, CNN performs the filter processing on speech frames in parallel. Meanwhile, CNN captures the contextual information by stacking multiple layers. Recently, the residual temporal convolution networks (ResTCNs) [25], which employ 1-D dilated convolutional modules and residual skip connections, have demonstrated impressive performance in modeling long-term dependencies and outperformed RNN across a broad range of sequence modeling tasks. ResTCNs have gained considerable success in speech enhancement [10], [19], [26], [27] and speaker separation [11], [28] as well. Self-attention based Transformer backbone network [29] has achieved state-of-the-art performance on many natural language processing tasks. More recently, Transformers have been successfully adopted for speech enhancement [27], [30], [31] and many other speech processing-related tasks such as speech synthesis and voice conversion. As the key component of Transformers, the multi-head self-attention (MHA) mechanism processes the whole sequence at once and computes the similarity between all time-steps to obtain the new representation, allowing the Transformer model for modeling long-term dependencies more efficiently. In [27] Zhao et al. proposed to employ a MHA module to produce dynamic representations followed by a ResTCN model to learn a nonlinear mapping for speech dereverberation.

The generation of human speech is subject to the constraint of the physiological structure of vocal production, and the phonetic and phonotactic rules of a spoken language. The success of ResTCN and Transformer in speech enhancement mainly stems from their ability to effectively model the long-range temporal context of speech. We also note that the energy concentration of a speech utterance in time or frequency varies from utterance to utterance. To preserve such a speech formant structure, we expect that a speech enhancement model performs according to the energy concentration in the T-F plane of a spectrogram. We are motivated to investigate a dedicated mechanism to characterise the salient T-F speech distribution.

The idea of attention has been well studied for the network to learn to attend to the salient features in computer vision [32], [33], [34], speech emotion recognition [35], [36], and speaker verification [37]. In a preliminary study [38], we investigated an attention mechanism to model the speech distribution along the frequency dimension and demonstrate its efficacy. We proposed a functional neural module, termed T-F attention (TFA), as part of the backbone networks to attend to the salient T-F representation for speech enhancement [39]. The proposed TFA module consists of two parallel attention branches, i.e., time-dimension (TA) and frequency-dimension attention (FA) that produce two 1-D attention maps to guide the models to focus on 'where' (which time frames) and 'what' (which frequency-wise channels), respectively. Then the TA and FA branches are combined to generate the final 2-D attention map, which assigns differentiated attention weights for each T-F spectral component, allowing the networks to capture the speech distribution in the T-F representation. In this paper, we further study the TFA module [39] across different backbone network architectures and training objectives, and evaluates its efficacy for a robust ASR system.

There have been attempts to capture long-range correlations in the T-F representation by applying self-attention operation along the time and frequency axes [40], [41], which was referred to as T-F attention [42]. In this paper, T-F attention (TFA) refers to a dedicated mechanism different from self-attention. It models the salient T-F speech distribution of speech signals. In particular, the T-F attention [42] is based on self-attention, and the learned attention scores represent the similarity among T-F vectors. However, the differentiated attention weights learned by our TFA represent how informative each T-F spectral component is. Such a T-F attention module can be used to augment existing neural speech enhancement solutions including self-attention. Therefore, it is different from that in [42] either in terms of motivation or implementation. The main contributions of this work are as follows:

- We propose a simple yet very effective network module (TFA) to characterise the salient T-F distribution for speech enhancement. It can be flexibly integrated with existing backbone networks to improve performance.
- We design time-dimension (TA) and frequency-dimension (FA) attention to enable the models to focus on informative frames and frequency-wise channels, respectively. Comprehensive ablation studies validate the efficacy.
- We extensively evaluate the TFA module across different backbone networks and training objectives. The results confirm that our TFA module consistently provides significant performance gains in speech enhancement, as well as in robust speech recognition.

The remainder of this paper is organized as follows. In Section II, we formulate the research problem. In Section III, we propose a novel time-frequency attention mechanism for speech enhancement. In Section IV, we describe the experimental setup. The experimental results are presented in Section V. Finally, Section VI concludes this study.

II. Problem Formulation

A. Signal Model

Let a noisy speech signal be $x[n]$,

$$x[n] = s[n] + d[n], \tag{1}$$

where $s[n]$ and $d[n]$ denote clean speech and uncorrelated additive noise, respectively, and n denotes the discrete-time index. The noisy speech, $x[n]$, is then analysed frame-wise using the short-time Fourier transform (STFT):

$$X[l,k] = S[l,k] + D[l,k], \tag{2}$$

where $X[l,k]$, $S[l,k]$, and $D[l,k]$ denote the complex-valued STFT coefficients of the noisy speech, the clean speech, and the noise components at time-frame index l and discrete-frequency index k.

B. Training Objectives

A backbone network architecture is trained to optimize the designed training objectives for speech enhancement. Studies show that by optimizing the network with respect to the T-F mask, we improve the intelligibility and quality of speech in speech enhancement. Without loss of generality, we have chosen four widely used training objectives in this study, as summarized next.

1) Ideal Ratio Mask: The ideal ratio mask (IRM) [13] is one of the most popular masking-based training objectives, and it is defined as:

$$\text{IRM}[l,k] = \sqrt{\frac{|S[l,k]|^2}{|S[l,k]|^2 + |D[l,k]|^2}} \tag{3}$$

where $|S[l,k]|$ and $|D[l,k]|$ denote the spectral magnitudes of clean speech and noise, respectively. The value of IRM ranges from 0 to 1.

2) Spectral Magnitude Mask: The spectral magnitude mask (SMM) [13] is defined on the STFT magnitude of clean speech and noisy speech:

$$\text{SMM}[l,k] = \frac{|S[l,k]|}{|X[l,k]|} \tag{4}$$

where $|X[l,k]|$ denotes the spectral magnitude of the noisy speech.

3) Phase-Sensitive Mask: The phase-sensitive mask (PSM) [15] is an extension of the SMM, which introduces a phase error item to compensate for the use of noisy speech phases:

$$\text{PSM}[l,k] = \frac{|S[l,k]|}{|X[l,k]|}\cos[\theta_{S-X}] \tag{5}$$

where θ_{S-X} denotes the difference between the clean speech phase and noisy speech phase.

From equations (4) and (5), we can find that the upper bound of SMM and PSM values exceeds 1. To fit the output range of the sigmoid activation function, we clip the SMM and PSM to between 0 and 1.

4) Instantaneous a priori SNR: The instantaneous a priori SNR (Xi) was proposed as the training objective [18], [19] and is employed by statistical model-based methods [19]. To form the training objective, the normal cumulative distribution function (CDF) is used to map $\xi_{\text{dB}}[l,k] = 10\log_{10}\xi[l,k]$ to the interval $[0,1]$, where $\xi[l,k] = \frac{|S[l,k]|^2}{|D[l,k]|^2}$ [43]:

$$\bar{\xi}[l,k] = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi_{\text{dB}}[l,k] - \mu_k}{\sigma_k\sqrt{2}}\right)\right], \tag{6}$$

where erf denotes the error function and mean μ_k and variance σ_k^2 are calculated from the training set (over 1000 randomly selected samples in this study). During inference, the a priori SNR estimate is computed as follows:

$$\hat{\xi}[l,k] = 10^{\left(\left(\sigma_k\sqrt{2}\text{erf}^{-1}\left(2\hat{\bar{\xi}}_l[k]-1\right)+\mu_k\right)/10\right)} \tag{7}$$

With IRM, SMM, and PSM as training objectives, we train DNNs to produce masks at run time. We then apply the resulting masks on the STFT spectral magnitude of noisy speech to obtain a clean version. The enhanced magnitude is then used with the noisy speech phase to reconstruct the clean speech waveform. For Xi, we adopt the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator as the statistical model, which uses $\hat{\xi}[l,k]$ to compute the spectral magnitude of enhanced speech [44].

III. Time-Frequency Attention for Speech Enhancement

A. Time-Frequency Attention

A TFA module is a computational unit, which takes an intermediate T-F representation $\mathbf{Y} \in \mathbb{R}^{L\times d_{model}}$ as the input, i.e., L frames of d_{model} frequency-wise feature channels, and generates an enhanced representation $\widetilde{\mathbf{Y}} \in \mathbb{R}^{L\times d_{model}}$ with differentiated T-F attention. The diagram of the proposed TFA module is illustrated in Fig. 1.

The distribution of speech signals over the T-F plane is defined by its time-frame index and frequency-wise channel index. We would generate a position-aware attention map, which assigns differentiated weights to position-specific speech components. In practice, we employ two parallel attention branches, termed TA and FA, which produces a 1-D time-frame attention map $\mathbf{T_A} \in \mathbb{R}^{1\times L}$ and a 1-D frequency-dimension attention map $\mathbf{F_A} \in \mathbb{R}^{d_{model}\times 1}$. Then, the two 1-D attention maps are combined via a tensor multiplication operation, resulting in a position-aware 2-D T-F attention map $\mathbf{TF_A} \in \mathbb{R}^{L\times d_{model}}$.

We next provide the detailed working of the proposed TFA module. As the long-range context correlations as shown in T-F representation are essential to locate the informative T-F regions, each attention branch adopts a two-step strategy to capture such correlations to generate the attention map: information aggregation and attention generation.

Information Aggregation: The TA and FA branches aggregate the whole utterance information along the time and frequency dimensions, respectively. The global average pooling and max pooling are typical techniques to aggregate global spatial information [32], [33]. We adopt the global average pooling, which

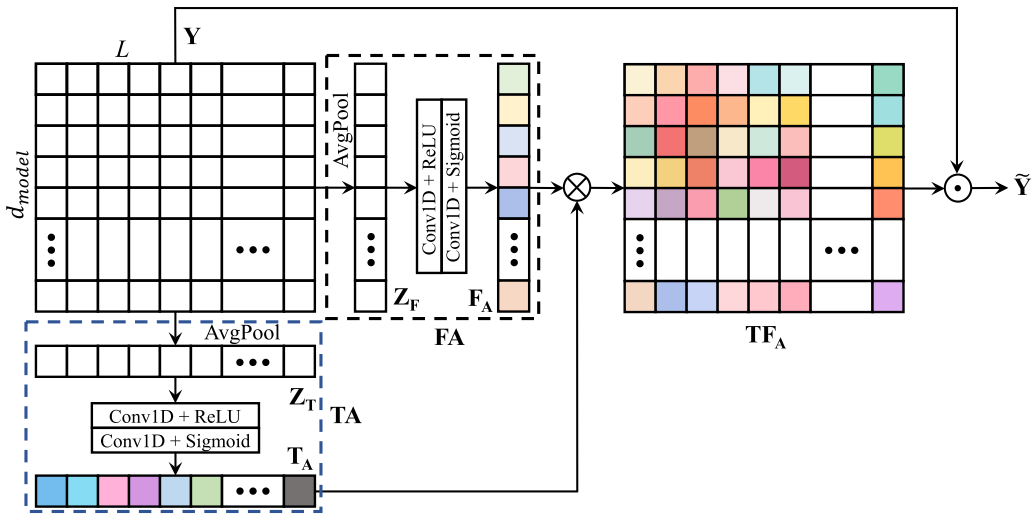


Fig. 1. A diagram of the proposed TFA module, where the TA and FA modules are shown in blue dotted box and black dotted box, respectively. Here, AvgPool and Conv1D represent the average pooling and dilated 1-D convolution operation, respectively. $\otimes$ and $\odot$ denote the matrix multiplication and element-wise product, respectively.

produces the global information descriptors that are expressive and general for the entire utterance. Specifically, the TA branch takes global average pooling along the frequency dimension on the given input $\mathbf{Y}$ and generates a time-frame-wise statistic $\mathbf{Z_T} \in \mathbb{R}^{1\times L}$ as follows:

$$\mathbf{Z_T}(l) = \frac{1}{d_{model}} \sum_{k=1}^{d_{model}} \mathbf{Y}(l,k), \tag{8}$$

where $\mathbf{Z_T}(l)$ is the l-th element of $\mathbf{Z_T}$. Similarly, the FA branch applies global average pooling along the time-frame dimension on the input $\mathbf{Y}$ and generates a frequency-wise statistic $\mathbf{Z_F} \in \mathbb{R}^{d_{model}\times 1}$. The k-th element of $\mathbf{Z_F}$ is given as:

$$\mathbf{Z_F}(k) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{Y}(l,k). \tag{9}$$

Attention Generation: A two-layer fully-connected (FC) network is often used to provide channel attention [32], [33]. However, the use of FC layer brings a large parameter overhead especially for long-duration speech. Alternatively, it has been suggested that an effective channel attention can be implemented in a more efficient way via 1-D convolution [34].

We adopt two stacked dilated 1-D convolution layers of kernel size k_{tfa} to capture the dependencies in the descriptors and learn a nonlinear interaction to produce the attention map. Specifically, given the descriptor $\mathbf{Z_T}$, the attention map in the TA branch is calculated as:

$$\mathbf{T_A} = \sigma(f_2^{TA}(\delta(f_1^{TA}(\mathbf{Z_T})))). \tag{10}$$

where f denotes a dilated 1-D convolution operation, δ and σ refer to the rectified linear module (ReLU) and the sigmoidal activation functions, respectively. The dilation rate d is set to 1 and 2 for the first and second convolution modules, respectively. A similar process is applied to the FA branch for generating the frequency-wise channel attention map:

$$\mathbf{F_A} = \sigma(f_2^{FA}(\delta(f_1^{FA}(\mathbf{Z_F})))) \tag{11}$$

Then, the attention maps obtained from the two attention branches interact with a tensor multiplication operation, resulting in our final 2-D T-F attention map $\mathbf{TF_A}$ written as:

$$\mathbf{TF_A} = \mathbf{T_A} \otimes \mathbf{F_A}, \tag{12}$$

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:

$$\mathbf{TF_A}(l,k) = \mathbf{T_A}(l) \times \mathbf{F_A}(k) \tag{13}$$

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. The output of our TFA module $\widetilde{\mathbf{Y}}$ is written as:

$$\widetilde{\mathbf{Y}} = \mathbf{Y} \odot \mathbf{TF_A}, \tag{14}$$

where $\odot$ denotes an element-wise multiplication.

B. Network Architecture

Our proposed TFA module is applicable to general neural speech enhancement architecture. To set the stage for this study, in Fig. 2, we illustrate a typical neural solution to speech enhancement, it is referred to as the backbone network in this paper. Here, two recently proposed backbone networks, ResTCN [19] and Transformer [29], [31], are employed. As the key component of the Transformer layer [29] is the MHA module, we term the Transformer layer as a MHANet block. The network takes the noisy spectral magnitude as input, $|\mathbf{X}| \in \mathbb{R}^{L\times K}$, where L and K denote the number of time frames and frequency bins, respectively. The first layer consists of a 1-D convolution layer with a frame-wise layer normalization followed by the ReLU activation function, that encodes the input into a latent T-F representation of $\mathbb{R}^{L\times d_{model}}$. The output of the first layer is

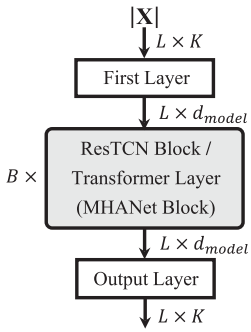
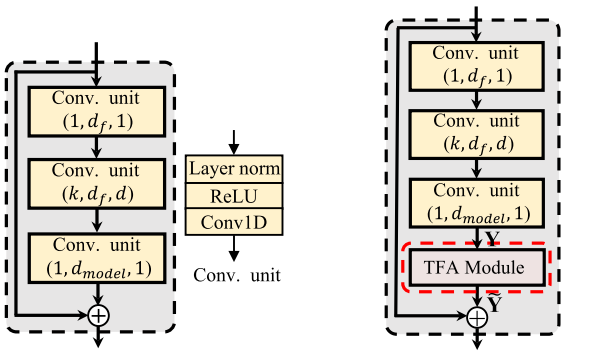


Fig. 2. Overall diagram of backbone networks.



(a) The original ResTCN block

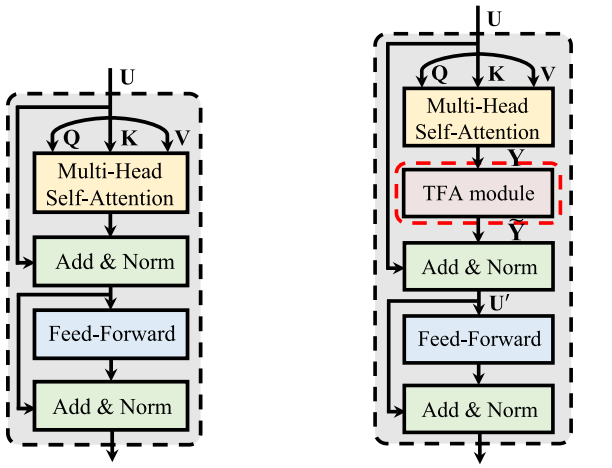
(b) The proposed ResTCN block with TFA module

Fig. 3. Illustration of (a) the ResTCN block and (b) the ResTCN block with the proposed TFA module. ⊕ denotes the element-wise summation operation.

then fed into B stacked ResTCN or MHANet blocks to perform T-F feature transformation. Following the last transformation block is the output layer, which is a 1-D convolution layer with a sigmoidal activation function that generates the estimates of IRM, SMM, PSM, and Xi.

1) Tfa in Restcn: In Fig. 3, we illustrate a backbone network based on a ResTCN block [19] and how the proposed TFA module works inside the ResTCN block. As shown in Fig. 3(a), the ResTCN block consists of three 1-D causal dilated convolutional modules. Each convolutional module employs a pre-activation design, where the input is pre-activated using frame-wise layer normalization followed by the ReLU activation function. We denote the kernel size, number of filters, and dilation rate for each convolutional module as a tuple of three elements. The first and third convolutional modules have a kernel size of 1, whilst the second convolutional module has a kernel size of k. The number of filters is d_f for the first and second convolutional modules, and d_{model} for the third convolutional module. The dilation rate, d, is employed in the second convolutional module, providing a contextual field over previous speech frames. The dilation rate is cycled as the block index $b = \{1, 2, 3, \ldots, B\}$ increases: $d = 2^{(b-1 \bmod (\log_2(D)+1))}$, where mod is the modulo operation and $D = 16$ is the maximum dilation rate.

As shown in Fig. 3(b), the ResTCN block is augmented by a TFA module to attend to the salient T-F representation. The residual connection [45] is applied between the input and output of the block to facilitate gradient optimization.



(a) The MHANet block (Transformer layer)

(b) The proposed MHANet block with TFA module

Fig. 4. Illustrations of (a) the MHANet block and (b) our proposed MHANet block with the TFA module.

2) Tfa in Mhanet: In Fig. 4, we illustrate how a TFA module is incorporated into the MHANet backbone [31]. As shown in Fig. 4(a), the MHANet block comprises two sub-blocks. The first is an MHA module, and the second is a two-layer fully connected feed-forward network (FFN). A residual connection is applied in each sub-block, followed by frame-wise layer normalization. To capture the T-F energy distribution of speech, as shown in Fig. 4(b), we propose to incorporate a TFA module into the MHA module.

Given an intermediate latent T-F tensor $\mathbf{U} \in \mathbb{R}^{L \times d_{model}}$ as the input to the block, the MHA module first projects the input $\mathbf{U}$ to queries ($\mathbf{Q} \in \mathbb{R}^{L \times d_{model}}$), keys ($\mathbf{K} \in \mathbb{R}^{L \times d_{model}}$), and values ($\mathbf{V} \in \mathbb{R}^{L \times d_{model}}$): $\mathbf{Q} = \mathbf{U}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{U}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{U}\mathbf{W}^V$, where $\{\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V\} \in \mathbb{R}^{d_{model} \times d_{model}}$ are different, learned linear projections. Then they are split into H attention heads, indexed by $h = \{1, 2, 3, \ldots, H\}$, and with dimensions d_k, d_k, and d_v, respectively, which enables the model to pay attention to different aspects of information. The scaled dot-product attention is applied to each head in parallel to generate the output,

$$\mathbf{A}_h = \text{Attention}\left(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h\right) = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_k}}\right)\mathbf{V}_h, \tag{15}$$

where $d_k = d_v = d_{model}/H$, and $\mathbf{K}_h^\top$ denotes the transpose of the h-th head of keys, $\mathbf{K}_h$. An upper triangular mask is used to mask out the similarities that include future frames. For more detailed descriptions about attention function, we refer the reader to the original study [29]. The outputs for each attention head are concatenated and linearly projected again, yielding the output of the MHA module as:

$$\text{MHA}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{Concat}(\mathbf{A}_1, \ldots, \mathbf{A}_H)\mathbf{W}^O, \tag{16}$$

where $\mathbf{W}^O \in \mathbb{R}^{d_{model} \times d_{model}}$. The TFA module takes the output from the previous MHA module, and conducts a T-F attention operation (described in Section III-A) to tell the model to focus

on the informative spectral components, resulting in an augmented T-F representation.

The two-layer FFN takes the output $\mathbf{U}' \in \mathbb{R}^{L \times d_{model}}$ from the first sub-block, and performs two linear transformations with a ReLU activation in the first layer:

$$\text{FFN}(\mathbf{U}') = \max(0, \mathbf{U}'\mathbf{W}^1 + \mathbf{b}^1)\mathbf{W}^2 + \mathbf{b}^2, \quad (17)$$

where $\mathbf{W}^1 \in \mathbb{R}^{d_{model} \times d_{ff}}$, $\mathbf{b}^1 \in \mathbb{R}^{d_{ff}}$, $\mathbf{W}^2 \in \mathbb{R}^{d_{ff} \times d_{model}}$, and $\mathbf{b}^2 \in \mathbb{R}^{d_{model}}$. The size of input and output is d_{model}, and the inner-layer has a size of d_{ff}.

Model Configuration: For ResTCN backbone, we adopt the parameter settings as in [19] to build the network: $d_{model} = 256$, $B = 40$, $d_f = 64$, and $k = 3$. For the MHANet backbone, we follow the parameter settings in [31]: $B = 5$, $d_{model} = 256$, $H = 8$, and $d_{ff} = 1024$. For our proposed TFA module, a kernel size $k_{tfa} = 17$ is adopted in the dilated 1-D convolution.

IV. Experimental Setup

A. Datasets and Feature Extraction

First, we describe the clean speech and noise data in this study. For clean speech recordings, we use the *train-clean-100* set from the Librispeech corpus [46] as the training set, which includes 28 539 utterances spoken by 251 speakers. The noise recordings in the training set are taken from the following datasets: the QUT-NOISE dataset [47], the Nonspeech dataset [48], the Environmental Background Noise dataset [49], [50], the RSG-10 dataset [51] (*voice babble*, *F16*, and *factory welding* are excluded for testing), the Urban Sound dataset [52] (*street music* recording no 26 270 is excluded for testing), the noise set from the MUSAN corpus [53], and colored noise recordings (with an α value ranging from -2 to 2 in increments of 0.25). Noise recordings that are over 30 seconds in length are split into segments of 30 seconds or less. This gives a total of 6 809 noise recordings, each with a length less than or equal to 30 seconds. For validation experiments, 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 recording is mixed with a random section of one noise recording at a randomly selected SNR level between -10 dB and 20 dB in 1 dB increments. This generates 1 000 noisy speech signals as the validation set.

For evaluation experiments, we adopt the recordings of four real-world noise sources (excluded from training set) including two non-stationary and two coloured. The two non-stationary noise sources are the *voice babble* from the the RSG-10 noise dataset [51] and *street music* from the Urban Sound dataset [52]. The two colored noise sources are *F16* and *factory welding* from RSG-10 noise dataset [51]. For each of the four noise recordings, ten clean speech recordings (without replacement) randomly selected from the *test-clean-100* of Librispeech corpus [46] 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 200 noisy mixtures for evaluation.

In this study, a square-root-Hann window function is used for spectral analysis and synthesis, with a frame length of 32 ms (512 samples) and a frame-shift of 16 ms (256 samples). 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.

B. Training Methodology

Here, we describe the details of training methodology used in this study. 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 selection order for the clean speech recordings is randomised for each epoch. For the three masking-based training objectives (i.e., IRM, SMM, and PSM), we adopt the mask approximation to learn the mask, where the mean-square error (MSE) is the loss function. For Xi, the cross-entropy is employed as the loss function [18], [19]. Each utterance in a mini-batch is padded with zeros, giving it the same number of time frames as the longest noisy utterance.

All models are trained from scratch. For ResTCN [19] and its TFA augmented variants, the *Adam* algorithm with default hyper-parameters [54] and a learning rate of 0.001 is used for gradient descent optimisation. For MHANet [31] and its TFA augmented variant, the *Adam* algorithm with parameters as in [29], i.e., $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$ is used for training. The gradient clipping technique is used for all models, where the gradients are clipped between $[-1, 1]$. As the training of MHANet is sensitive to the learning rate [29], [31], we adopt the warm-up training strategy in [29], where the learning rate is adjusted during the training process according to the rule:

$$lr = d_{model}^{-0.5} \cdot \min\left(n\_step^{-0.5}, n\_step \cdot w\_steps^{-1.5}\right) \quad (18)$$

where n_step and w_steps denote the number of training steps and warm-up training steps, respectively. Following [31], the number of steps $w\_steps = 40\,000$ is adopted for warm-up training in this work.

C. Evaluation Metrics

In our experiments, five widely used metrics are adopted for extensive speech enhancement evaluations, including perceptual evaluation of speech quality (PESQ) [55], extended short-time objective intelligibility (ESTOI) [56], and three composite metrics [57]. For the PESQ metric, we adopt the wide-band PESQ [55], which typically produces a lower score than the narrow-band counterpart [27]. The value range of PESQ is $[-0.5, 4.5]$ and the value of ESTOI is typically in $[0, 1]$. The three composite metrics are mean opinion score (MOS) predictors of the signal distortion (CSIG) [57], the background-noise intrusiveness (CBAK) [57], and the overall signal quality (COVL) [57], respectively. The value range of the three composite metrics is $[0, 5]$. For all of the above five metrics, a higher score indicates better enhancement performance. The word error rate (WER%) is adopted to evaluate the downstream ASR performance, and a lower WER% score means better speech recognition performance.

D. Comparative Models

We evaluate the proposed TFA module as part of two backbone networks (ResTCN [19] and MHANet [31]) on four training objectives. The proposed ResTCN and MHANet with the TFA module are denoted by "ResTCN+TFA" and "MHANet+TFA," respectively. In addition, we conduct an ablation study to validate the efficacy of each component (the TA and FA) in the TFA module. Similarly, we denote the ResTCN and MHANet with the TA and FA by "ResTCN+TA," "MHANet+TA," "ResTCN+FA," and "MHANet+FA," respectively. All models in this study are implemented using Tensorflow 1.13. The experiments were conducted on an NVIDIA Tesla V100 graphics processing unit (GPU) and an Intel Xeon Platinum 8163 CPU at 2.50 GHz (96 logical processors).

V. Experimental Results

A. Training and Validation Error

We first observe the training and validation errors across the models. The error curves produced by each of the models on the four training objectives (i.e., IRM, SMM, PSM, and Xi) are shown in Figs. 5–6, Figs. 7–8, 9–10, and 11–12, respectively, where each model is trained for 250 epochs. We observe similar trends of error curves on different training objectives. The purple curves are for ResTCN and MHANet, and the red curves are for ResTCN+TFA and MHANet+TFA. It can be easily observed that ResTCN+TFA and MHANet+TFA yield significantly lower training and validation errors than ResTCN and MHANet, which demonstrates the effect of the TFA module. In addition, the ablation study also confirms the efficacy of the FA and TA modules. One can observe that ResTCN+FA (blue curves) and ResTCN+TA (yellow curves) produce significantly lower training and validation error as compared to ResTCN. ResTCN+FA yields close error curves with ResTCN+TA. MHANet+FA (blue curves) and MHANet+TA (yellow curves) also achieve an obvious lower training and validation error than MHANet. For MHANet, applying the TA module achieves lower training and validation error than the FA module. Among the TFA, TA, and FA modules, the TFA module consistently produces the lowest training and validation error across different training objectives.

B. Experiment on Enhancement Performance

Tables I and II list the wide-band PESQ and ESTOI scores obtained by each of the models in all noisy test conditions, respectively, in four training objectives. The highest PESQ and ESTOI scores for each condition are highlighted in boldface. Compared to unprocessed noisy recordings, for all of the training objectives, our proposed models provide substantial improvements in terms of both PESQ and ESTOI scores. Taking the *street music* noise with SNR of 5 dB as a case, ResTCN+TFA and MHANet+TFA with the SMM achieve 0.70 and 0.65 gains on PESQ, and 19.40% and 19.43% gains on ESTOI, respectively. Among the four training objectives, overall, PSM and Xi show better performance than IRM and SMM in terms of PESQ. The obvious superiority in terms of ESTOI is not observed on any of the training objectives.

It is also easy to observe that for all training objectives, applying the TFA module significantly improves the PESQ and ESTOI scores of ResTCN and MHANet backbones with negligible parameter overheads (2.72 K and 0.34 K), demonstrating its effectiveness on speech enhancement. In the *F16* noise with SNR of 5 dB case, for instance, ResTCN+TFA and MHANet+TFA with the IRM provide 0.24 and 0.19 PESQ improvements, 3.98% and 2.84% ESTOI improvements, respectively, over the corresponding baselines. From the comparison results, among the two baselines, ResTCN benefits more from the TFA module in most cases.

In addition, the performance evaluations of the TA and FA modules are also reported in the ablation study. The TA and FA modules produce two 1-D attention maps to model the energy distribution of speech along time and frequency dimensions, respectively. As shown in Tables I and II, both ResTCN and MHANet achieve performance gains, in terms of PESQ and ESTOI, due to the TA and FA modules in most cases. Overall, the TA module provides more PESQ and ESTOI gains than the FA module. This could be explained by the fact that the temporal attention mechanism assigns the differentiated attention weights along the time axis, acting like a soft voice activity detector (VAD). The temporal information could be more informative than the spectral one in speech enhancement. The TFA module effectively combines the TA and FA modules to produce a 2-D attention map for modeling the T-F distribution of speech spectral components, which attains the highest PESQ and ESTOI scores in almost all cases.

Tables III–V report the average CSIG, CBAK, and COVL scores for each of the SNR levels (covering four noise sources), respectively, and the highest scores are highlighted in boldface. It is obvious that applying the TFA module to ResTCN and MHANet significantly improves their performance in terms of the three composite metrics, across different training objectives. In the -5 dB SNR case, for instance, ResTCN+TFA and MHANet+TFA with the IRM improve CSIG by 0.23 and 0.17, CBAK by 0.1 and 0.1, and COVL by 0.17 and 0.12, respectively. Again, compared to MHANet, ResTCN benefits more from the TFA module.

The TA and FA modules also provide substantial improvements to baselines in the three metrics. For instance, in the 5 dB case, the SMM is used as the training objective. For ResTCN applying FA and TA modules improves CISG by 0.21 and 0.21, CBAK by 0.12 and 0.12, and COVL by 0.19 and 0.20, respectively. For MHANet, applying FA and TA modules improves CISG by 0.07 and 0.12, CBAK by 0.08 and 0.10, and COVL by 0.09 and 0.13, respectively. Overall, the TA module performs slightly better than the FA. In the case of Xi as training objective and the SNR of 15 dB, MHANet+FA obtains the same CSIG scores (4.18) with MHANet+TFA. In all other cases, the TFA module obtains the highest CSIG, CBAK, and COVL scores.

1) The Number of Building Blocks: We evaluate the effectiveness of the TFA module across different numbers of building blocks with IRM as the training objective. ResTCN-20 and ResTCN-30 denote ResTCN models with 20 and 30 ResTCN blocks, respectively. MHANet-4 and MHANet-6

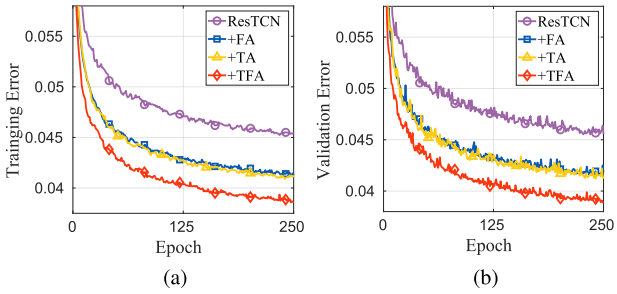


Fig. 5. The (a) training error and (b) validation error of ResTCN and proposed models with ResTCN as backbone on IRM training objective.

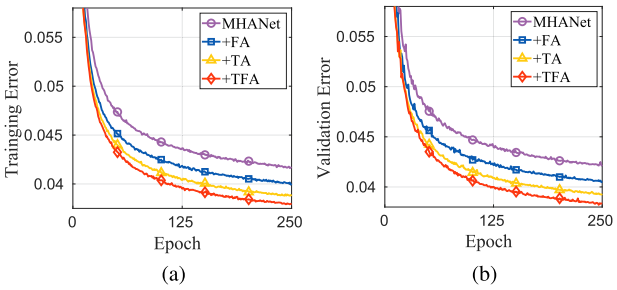


Fig. 6. The (a) training error and (b) validation error of MHANet and proposed models with MHANet as backbone on IRM training objective.

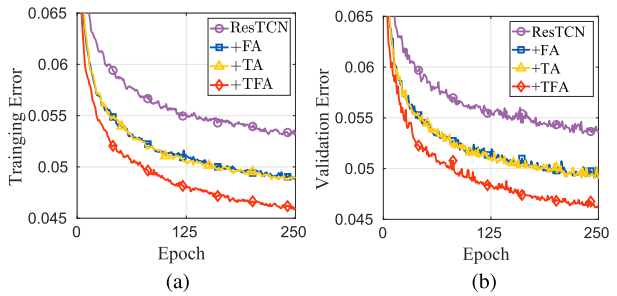


Fig. 7. The (a) training error and (b) validation error of ResTCN and proposed models with ResTCN as backbone on SMM training objective.

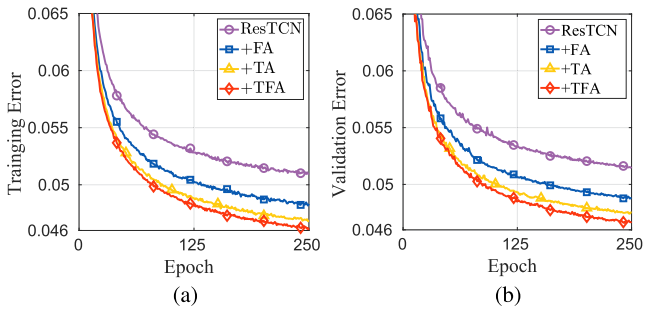


Fig. 8. The (a) training error and (b) validation error of MHANet and proposed models with MHANet as backbone on SMM training objective.

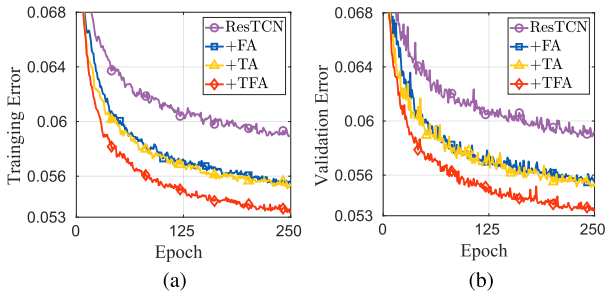


Fig. 9. Training error (a) and validation error (b) of ResTCN and the proposed models with ResTCN as backbone on PSM training objective.

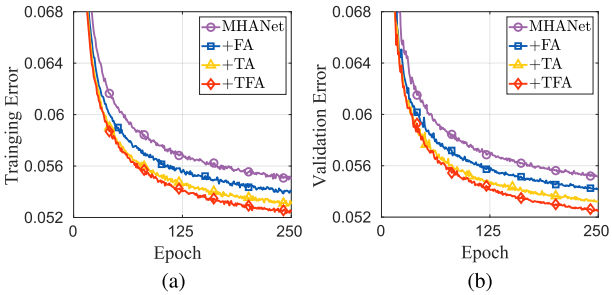


Fig. 10. Training error (a) and validation error (b) of MHANet and proposed models with MHANet as backbone on PSM training objective.

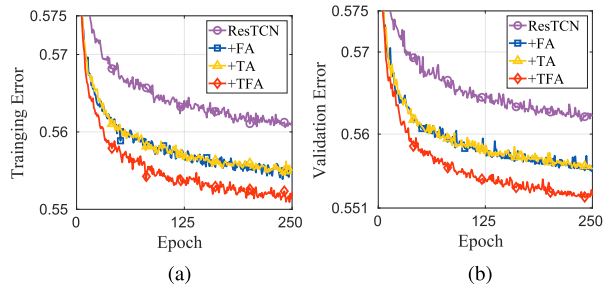


Fig. 11. Training error (a) and validation error (b) of ResTCN and the proposed models with ResTCN as backbone on Xi training objective.

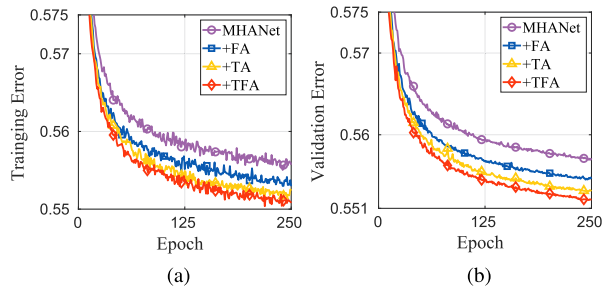


Fig. 12. The training error (a) and validation error (b) of MHANet and the proposed models with MHANet as backbone on Xi training objective.

denote MHANet models with 4 and 6 MHANet blocks, respectively. The experimental results are given in Table VI. It can be seen that the TFA module consistently affords substantial improvements to both ResTCN and MHANet. Here, we also report the evaluation results of ResTCN with self-attention (ResTCN+SA) [27], which further demonstrates the efficacy and efficiency of our TFA module. ResTCN+SA [27] employs a multi-head self-attention module as a pre-processing module followed by a ResTCN model. Compared to ResTCN+SA, ResTCN+TFA shows substantial superiority in terms of performance scores and parameter efficiency. In addition, we also study our TFA module in the recent Conformer [58], [59] and the

TABLE I
SPEECH ENHANCEMENT PERFORMANCE IN TERMS OF WIDE-BAND PESQ FOR DIFFERENT MODELS AND TRAINING TARGETS

Objective	Network	# Params	SNR level (dB)																			
			Voice babble					Street music					F16					Factory				
			-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
-	Noisy	–	1.07	1.12	1.23	1.47	1.89	1.03	1.05	1.10	1.25	1.56	1.04	1.06	1.11	1.27	1.58	1.05	1.05	1.10	1.24	1.52
IRM	ResTCN	1.976M	1.15	1.35	1.70	2.15	2.64	1.12	1.27	1.56	1.97	2.43	1.19	1.45	1.77	2.19	2.59	1.11	1.29	1.56	1.99	2.37
	+FA	+1.36K	1.19	1.43	1.81	2.27	2.77	1.15	1.32	1.63	2.05	2.54	1.26	1.52	1.91	2.34	2.76	1.15	1.36	1.70	2.17	2.62
	+TA	+1.36K	1.16	1.41	1.82	2.30	2.78	1.15	1.34	1.68	2.10	2.57	1.26	1.53	1.88	2.30	2.70	1.17	1.38	1.72	2.16	2.54
	+TFA	+2.72K	**1.22**	**1.48**	**1.89**	**2.35**	**2.83**	**1.18**	**1.39**	**1.75**	**2.16**	**2.60**	**1.32**	**1.61**	**1.97**	**2.39**	**2.79**	**1.21**	**1.41**	**1.80**	**2.24**	**2.65**
	MHANet	4.076M	1.17	1.39	1.75	2.21	2.71	1.13	1.30	1.59	1.97	2.43	1.22	1.48	1.86	2.28	2.70	1.11	1.27	1.56	1.96	2.38
	+FA	+0.17K	1.17	1.40	1.78	2.25	2.71	1.16	1.35	1.67	2.07	2.52	1.26	1.51	1.89	2.35	2.80	1.14	1.34	1.67	2.09	2.50
	+TA	+0.17K	1.17	1.43	1.81	2.28	2.75	1.16	1.36	1.67	2.08	2.52	1.26	1.55	1.91	2.35	2.78	1.15	1.38	1.70	2.12	2.50
	+TFA	+0.34K	**1.19**	**1.43**	**1.85**	**2.34**	**2.85**	**1.18**	**1.38**	**1.71**	**2.15**	**2.63**	**1.30**	**1.60**	**1.95**	**2.39**	**2.81**	**1.16**	**1.39**	**1.75**	**2.20**	**2.62**
SMM	ResTCN	1.98M	1.12	1.34	1.68	2.08	2.53	1.11	1.27	1.51	1.90	2.36	1.20	1.44	1.76	2.16	2.56	1.11	1.27	1.55	1.96	2.36
	+FA	+1.36K	1.19	1.45	1.83	2.31	2.80	1.15	1.34	1.66	2.08	2.55	1.27	1.55	1.91	2.35	2.77	1.15	1.34	1.68	2.14	2.59
	+TA	+1.36K	1.18	1.44	1.84	2.32	2.80	1.15	1.35	1.70	2.12	2.58	1.27	1.54	1.89	2.30	2.71	1.17	1.39	1.75	2.20	2.60
	+TFA	+2.72K	**1.21**	**1.51**	**1.93**	**2.39**	**2.86**	**1.18**	**1.40**	**1.80**	**2.25**	**2.71**	**1.34**	**1.63**	**2.01**	**2.45**	**2.81**	**1.21**	**1.45**	**1.82**	**2.28**	**2.74**
	MHANet	4.076M	1.16	1.39	1.75	2.22	2.71	1.13	1.31	1.62	2.02	2.46	1.24	1.50	1.84	2.26	2.70	1.12	1.28	1.57	1.99	2.41
	+FA	+0.17K	1.19	1.44	1.83	2.27	2.74	1.17	1.37	1.70	2.13	2.59	1.25	1.53	1.89	2.32	2.72	1.14	1.34	1.66	2.10	2.52
	+TA	+0.17K	1.19	1.46	1.86	2.35	2.83	1.17	1.38	1.72	2.14	2.59	1.28	1.59	1.92	2.33	2.76	1.15	1.37	1.74	2.14	2.54
	+TFA	+0.34K	**1.21**	**1.49**	**1.89**	**2.37**	**2.85**	**1.19**	**1.41**	**1.75**	**2.19**	**2.64**	**1.30**	**1.61**	**2.00**	**2.45**	**2.84**	**1.17**	**1.42**	**1.78**	**2.21**	**2.64**
PSM	ResTCN	1.98M	1.17	1.39	1.74	2.25	2.79	1.14	1.32	1.67	2.12	2.67	1.27	1.55	1.92	2.35	2.80	1.15	1.36	1.71	2.18	2.63
	+FA	+1.36K	1.21	1.43	1.86	2.37	2.86	1.18	1.40	1.75	2.21	2.73	1.30	1.59	2.01	2.52	2.97	1.22	1.47	1.88	2.35	2.82
	+TA	+1.36K	1.17	1.45	1.89	2.41	2.91	1.15	1.40	1.79	2.27	2.77	1.32	1.64	2.01	2.49	2.91	1.21	1.49	**1.89**	**2.36**	2.80
	+TFA	+2.72K	**1.25**	**1.55**	**1.98**	**2.48**	**2.96**	**1.22**	**1.47**	**1.87**	**2.32**	**2.79**	**1.35**	**1.66**	**2.11**	**2.59**	**3.01**	**1.25**	**1.50**	**1.89**	2.34	**2.86**
	MHANet	4.076M	1.19	1.45	1.87	2.36	2.91	1.18	1.40	1.74	2.18	2.66	1.29	1.59	2.01	2.49	2.92	1.16	1.37	1.74	2.26	2.69
	+FA	+0.17K	1.19	1.45	1.87	2.38	2.90	1.18	1.39	1.76	2.24	2.75	1.31	1.61	2.04	2.54	2.97	1.19	1.43	1.83	2.26	2.72
	+TA	+0.17K	1.19	1.46	1.87	2.43	**2.95**	1.20	1.44	1.80	2.25	2.74	1.32	1.63	2.03	2.53	3.01	1.18	1.41	1.80	2.27	**2.73**
	+TFA	+0.34K	**1.20**	**1.48**	**1.92**	**2.45**	**2.95**	**1.21**	**1.46**	**1.85**	**2.31**	**2.80**	**1.34**	**1.66**	**2.09**	**2.60**	**3.03**	**1.23**	**1.50**	**1.90**	**2.37**	2.72
Xi	ResTCN	1.98M	1.16	1.38	1.72	2.25	2.77	1.14	1.31	1.60	2.01	2.48	1.26	1.54	1.92	2.33	2.77	1.14	1.35	1.64	2.14	2.51
	+FA	+1.36K	1.21	1.43	1.82	2.37	2.88	1.19	1.42	1.75	2.23	2.73	1.31	1.60	2.04	2.52	2.96	1.19	1.46	1.84	2.35	2.78
	+TA	+1.36K	1.19	1.46	1.86	2.42	2.92	1.20	1.43	1.82	2.29	2.75	1.31	1.60	1.96	2.40	2.88	1.21	1.47	1.85	2.28	2.73
	+TFA	+2.72K	**1.22**	**1.52**	**1.93**	**2.48**	**2.95**	**1.24**	**1.48**	**1.86**	**2.32**	**2.77**	**1.38**	**1.69**	**2.10**	**2.60**	**3.01**	**1.26**	**1.53**	**1.93**	**2.43**	**2.83**
	MHANet	4.076M	1.18	1.43	1.77	2.30	2.84	1.18	1.38	1.71	2.16	2.67	1.28	1.60	2.06	2.53	2.98	1.16	1.38	1.78	2.25	2.66
	+FA	+0.17K	1.19	1.42	1.78	2.36	2.89	1.20	1.42	1.75	2.24	**2.74**	1.29	1.61	2.03	2.53	2.99	1.19	1.45	1.80	2.29	2.69
	+TA	+0.17K	1.06	**1.45**	1.80	2.39	2.91	1.19	1.41	1.76	2.16	2.59	1.31	1.62	2.03	2.52	2.96	1.19	1.44	1.82	2.24	2.67
	+TFA	+0.34K	**1.20**	**1.45**	**1.85**	**2.40**	**2.93**	**1.21**	**1.43**	**1.82**	**2.29**	2.73	**1.34**	**1.67**	**2.07**	**2.56**	**3.00**	**1.21**	**1.46**	**1.88**	**2.32**	**2.71**

self-attentive TCN (SA-TCN) [60, Fig. 1] network architectures. Here, we adopt the Conformer model (Conformer-5) with 5 Conformer building blocks [58, Fig. 1 (right)] and the 2-stage SA-TCN (2S-SA-TCN) [60] as the baseline backbones. A Conformer block consists of four modules stacked together, i.e., an FNN module, a self-attention module, a convolution module, and a second FNN module. Each stage of 2S-SA-TCN includes a self-attention module, followed by 24 ResTCN blocks. For Conformer, the TFA module is incorporated into the convolution module (following the third convolution unit) and the self-attention module (as shown in Fig. 4(b)). For 2S-SA-TCN, the TFA module is incorporated into the self-attention module and the ResTCN block as shown in Figs. 3(b) and 4(b), respectively. It is clear that the TFA module consistently provides significant improvements to both Conformer and 2S-SA-TCN.

C. Experiment on ASR Performance

In real-world environments, speech enhancement is often used as a front-end to improve the noise robustness of an ASR system. In this section, we investigate the effectiveness of our proposed model as the front-end for a robust ASR system. DeepSpeech[1] [61], an open-source ASR system developed using end-to-end deep learning technique, is used in this study to conduct ASR experiments for evaluating front-end performance. The RNN-based acoustic model and language model are used in DeepSpeech. Here, we treat the DeepSpeech ASR system as a black box, without fine-tuning during the experiment.

Table VII presents the average WER%[2] scores attained by all the models for each SNR condition and the WER% scores averaged across all SNR conditions. It can be seen that all front-end models achieve performance gains substantially in terms of WER% compared to the ASR performance on unprocessed noisy recordings. Overall, for two backbones (ResTCN and MHANet) and four training objectives, our proposed TFA module attains the lowest average WER% scores over all conditions, and performs the best under most SNR conditions. The evaluation shows that the proposed TFA module demonstrates substantial performance improvement to the two baselines, i.e., ResTCN and MHANet. For the four training objectives similar performance trends are observed, and on average, the IRM achieves the best performance. With the IRM as the training objective, the TFA module improves the ResTCN and MHANet baselines, with a relative WER% reduction of 15.78% and 8.58% over all conditions, respectively. In addition, the FA and TA modules provide a relative WER% reduction of 10.46% and 7.40% to the ResTCN, and that of 4.89% and 4.84% to

[1]The implementation of the DeepSpeech ASR system is available at:https://github.com/mozilla/DeepSpeech. The latest release model (0.9.3) is used in our experiment.

[2]The code implementation for WER% calculation is available at:https://github.com/jitsi/jiwer. Several pre-processing steps are applied before WER% calculation, including transforming a sentence into a list of words, removing multiple spaces between words and empty strings, and removing all leading and trailing spaces.

TABLE II
SPEECH ENHANCEMENT PERFORMANCE IN TERMS OF ESTOI (%) METRIC FOR DIFFERENT MODELS AND TRAINING TARGETS

Objective	Network	# Params	SNR level (dB)																			
			Voice babble					Street music					F16					Factory				
			-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
-	Noisy	–	28.76	44.42	60.67	74.97	85.37	30.39	44.03	58.15	71.13	81.80	27.45	41.89	56.70	70.27	81.30	25.03	38.45	53.30	68.09	80.42
IRM	ResTCN	1.98M	39.22	58.59	74.23	84.70	90.89	45.06	61.84	74.41	83.12	89.00	46.48	62.90	75.21	83.98	89.94	39.38	57.62	71.88	82.04	88.59
	+FA	+1.36K	42.58	61.54	76.28	85.53	91.38	47.00	62.73	75.25	83.81	89.56	49.51	64.79	76.59	85.16	90.51	41.96	59.65	73.83	83.46	89.59
	+TA	+1.36K	42.35	61.90	76.61	85.99	91.52	48.67	64.83	76.21	84.26	89.70	52.21	66.89	77.64	85.46	90.82	44.43	61.84	74.84	83.97	89.84
	+TFA	+2.72K	**44.72**	**63.56**	**77.79**	**86.39**	**91.80**	**50.40**	**66.01**	**77.23**	**84.78**	**89.96**	**53.97**	**67.97**	**78.50**	**86.37**	**91.36**	**46.22**	**62.88**	**75.86**	**84.72**	**90.48**
	MHANet	4.076M	42.83	61.36	75.47	85.27	91.28	47.38	63.13	75.49	83.81	89.56	47.70	64.36	76.68	85.00	90.55	40.46	57.74	72.34	82.20	88.60
	+FA	+0.2K	42.24	61.15	75.96	85.67	91.38	48.64	64.35	76.31	84.39	89.96	50.24	65.58	77.33	85.63	91.08	41.45	59.27	73.65	83.22	89.35
	+TA	+0.2K	42.95	62.53	76.65	86.02	91.69	48.56	64.35	76.03	84.00	89.68	50.27	66.23	77.73	85.69	91.16	43.10	60.64	74.36	83.50	89.41
	+TFA	+0.41K	**43.31**	**62.58**	**77.47**	**86.43**	**91.92**	**49.16**	**64.75**	**76.60**	**84.65**	**90.18**	**52.08**	**66.99**	**78.10**	**86.07**	**91.29**	**44.67**	**62.02**	**75.18**	**83.96**	**89.91**
SMM	ResTCN	1.98M	38.18	56.88	72.04	82.33	89.12	43.17	59.59	72.47	81.69	88.09	45.97	62.51	74.82	83.67	89.76	38.62	56.37	71.37	81.47	88.14
	+FA	+1.36K	41.65	62.27	76.17	85.64	91.35	46.84	63.20	75.60	83.95	89.64	50.29	65.37	76.66	85.07	90.52	41.95	59.49	73.62	83.19	89.43
	+TA	+1.36K	42.48	62.25	76.93	85.95	91.37	48.34	64.85	76.42	84.19	89.65	51.61	66.90	77.79	85.14	90.63	44.50	61.69	75.16	84.28	89.92
	+TFA	+2.72K	**43.74**	**64.28**	**78.47**	**86.57**	**91.73**	**48.96**	**65.06**	**77.55**	**85.18**	**90.09**	**54.15**	**68.48**	**78.65**	**86.13**	**91.01**	**47.57**	**64.12**	**76.39**	**84.50**	**90.12**
	MHA	4.076M	41.87	61.38	75.79	85.34	91.26	47.31	63.44	75.83	83.98	89.70	49.71	65.27	77.07	85.25	90.88	41.66	59.31	73.10	82.83	89.00
	+FA	+0.2K	43.20	62.25	76.62	85.99	91.71	48.34	64.02	76.46	84.76	90.18	50.28	65.82	77.58	85.73	91.13	41.68	59.68	73.63	83.39	89.56
	+TA	+0.2K	43.11	62.81	77.00	**86.45**	**91.99**	49.49	65.22	76.80	84.83	90.19	51.63	67.40	78.38	85.85	91.37	44.07	61.07	74.85	83.54	89.63
	+TFA	+0.41K	**44.70**	**64.06**	**77.58**	86.42	91.83	**50.47**	**65.92**	**77.58**	**85.19**	**90.35**	**52.97**	**67.60**	**78.66**	**86.33**	**91.71**	**44.88**	**61.81**	**75.55**	**84.27**	**90.23**
PSM	ResTCN	1.98M	40.46	59.53	74.71	85.03	90.92	45.69	61.95	75.31	83.96	89.67	46.66	63.14	75.45	84.16	89.91	38.22	57.17	72.41	82.48	88.85
	+FA	+1.36K	41.41	60.89	76.18	85.67	91.34	46.43	62.95	75.44	83.92	89.64	50.22	66.09	77.58	86.02	91.06	41.58	59.60	74.30	83.66	89.77
	+TA	+1.36K	41.28	61.08	76.64	86.05	91.61	47.03	63.73	76.22	84.49	90.02	51.39	66.82	77.75	85.53	90.97	43.11	61.36	75.36	84.31	90.20
	+TFA	+2.72K	**45.44**	**64.63**	**78.23**	**86.81**	**92.10**	**49.01**	**65.22**	**76.85**	**84.79**	**90.17**	**53.66**	**68.30**	**78.64**	**86.43**	**91.49**	**45.90**	**62.15**	**75.53**	**84.57**	**90.46**
	MHANet	4.076M	41.99	61.78	76.37	85.86	91.59	46.78	63.11	75.89	84.30	89.83	49.27	65.71	78.01	85.94	91.09	40.63	58.99	73.62	83.38	89.49
	+FA	+0.2K	41.84	61.56	76.31	85.83	91.51	47.79	63.92	76.06	84.57	90.11	48.48	65.53	78.88	85.88	91.06	41.06	58.98	73.73	83.39	89.50
	+TA	+0.2K	42.49	62.32	**77.22**	86.53	91.87	48.07	64.41	76.30	84.47	90.12	50.73	67.15	78.48	86.17	91.37	42.46	60.37	74.40	83.72	89.63
	+TFA	+0.41K	**42.76**	**62.35**	77.20	**86.65**	**91.96**	**48.60**	**64.57**	**76.74**	**84.84**	**90.33**	**51.97**	**67.19**	**78.62**	**86.65**	**91.60**	**43.74**	**61.43**	**74.99**	**84.27**	**90.25**
Xi	ResTCN	1.98M	38.43	57.56	73.29	84.47	90.61	42.79	60.15	73.74	82.91	88.81	46.90	63.07	75.39	83.76	89.78	37.06	56.63	71.88	82.30	88.69
	+FA	+1.36K	40.53	60.20	75.71	85.57	91.17	44.80	62.31	75.39	83.95	89.63	49.41	64.89	76.92	85.19	90.70	40.21	59.16	73.28	82.97	89.10
	+TA	+1.36K	42.05	61.59	76.18	85.75	91.44	46.85	63.28	76.04	84.23	89.81	50.92	66.32	77.41	84.87	90.64	42.23	60.25	73.81	83.14	89.30
	+TFA	+2.72K	**42.78**	**63.58**	**77.47**	**86.59**	**91.85**	**48.83**	**64.97**	**76.71**	**84.51**	**89.98**	**54.09**	**68.41**	**78.73**	**86.33**	**91.44**	**45.11**	**62.39**	**74.66**	**84.01**	**90.14**
	MHA	4.076M	40.44	60.80	75.26	85.23	91.21	45.71	62.86	75.34	83.83	89.71	48.48	65.65	77.68	85.53	90.91	38.99	58.05	73.10	83.13	89.31
	+FA	+0.17K	39.80	60.26	75.37	85.69	91.48	45.24	63.20	75.95	84.27	89.88	48.69	65.19	77.21	85.50	91.03	39.92	58.74	73.64	83.33	89.37
	+TA	+0.17K	41.24	**61.84**	76.24	85.79	91.49	45.74	62.84	76.02	84.00	89.51	49.37	66.17	**77.96**	85.66	90.97	41.31	59.45	73.54	83.20	89.39
	+TFA	+0.34K	**41.25**	61.22	**76.46**	**85.93**	**91.55**	**46.45**	**63.63**	**76.31**	**84.50**	**89.92**	**50.51**	**66.51**	77.94	**85.78**	**91.22**	**43.39**	**60.82**	**74.53**	**83.69**	**89.64**

TABLE III
AVERAGE CSIG SCORES FOR EACH SNR LEVEL AND THE HIGHEST CSIG SCORES ARE HIGHLIGHTED IN BOLDFACE

Objective	Network	Input SNR (dB)				
		-5	0	5	10	15
-	Noisy	1.49	1.80	2.20	2.65	3.16
IRM	ResTCN	2.17	2.66	3.12	3.58	4.00
	+FA	2.33	2.80	3.26	3.71	4.13
	+TA	2.30	2.79	3.26	3.71	4.11
	+TFA	**2.40**	**2.88**	**3.35**	**3.78**	**4.17**
	MHANet	2.20	2.68	3.17	3.63	4.05
	+FA	2.29	2.76	3.23	3.69	4.10
	+TA	2.33	2.83	3.28	3.72	4.12
	+TFA	**2.37**	**2.85**	**3.32**	**3.77**	**4.18**
SMM	ResTCN	2.10	2.59	3.06	3.51	3.94
	+FA	2.30	2.80	3.27	3.73	4.14
	+TA	2.29	2.80	3.28	3.74	4.14
	+TFA	**2.37**	**2.90**	**3.37**	**3.81**	**4.20**
	MHANet	2.21	2.70	3.17	3.63	4.06
	+FA	2.27	2.76	3.24	3.69	4.10
	+TA	2.30	2.81	3.29	3.74	4.15
	+TFA	**2.36**	**2.86**	**3.34**	**3.78**	**4.18**
PSM	ResTCN	2.15	2.65	3.13	3.61	4.06
	+FA	2.30	2.78	3.26	3.74	4.18
	+TA	2.29	2.81	3.30	3.77	4.20
	+TFA	**2.38**	**2.88**	**3.36**	**3.80**	**4.21**
	MHANet	2.22	2.73	3.23	3.73	4.16
	+FA	2.29	2.78	3.28	3.76	4.19
	+TA	2.30	2.80	3.30	3.78	4.22
	+TFA	**2.36**	**2.86**	**3.35**	**3.82**	**4.23**
Xi	ResTCN	2.19	2.70	3.16	3.63	4.05
	+FA	2.32	2.79	3.27	3.76	4.19
	+TA	2.29	2.80	3.28	3.75	4.18
	+TFA	**2.41**	**2.91**	**3.37**	**3.84**	**4.23**
	MHANet	2.26	2.77	3.25	3.72	4.16
	+FA	2.33	2.82	3.27	3.75	**4.18**
	+TA	2.31	2.83	3.30	3.76	4.17
	+TFA	**2.37**	**2.86**	**3.34**	**3.79**	**4.18**

TABLE IV
AVERAGE CBAK SCORES FOR EACH SNR CONDITION AND THE HIGHEST CBAK SCORES ARE HIGHLIGHTED IN BOLDFACE

Objective	Network	Input SNR (dB)				
		-5	0	5	10	15
-	Noisy	1.15	1.40	1.58	2.13	2.61
IRM	ResTCN	1.69	2.03	2.41	2.85	3.30
	+FA	1.75	2.10	2.50	2.94	3.39
	+TA	1.74	2.09	2.51	2.95	3.38
	+TFA	**1.79**	**2.15**	**2.56**	**2.99**	**3.42**
	MHANet	1.69	2.03	2.43	2.86	3.31
	+FA	1.75	2.10	2.50	2.93	3.36
	+TA	1.73	2.10	2.50	2.94	3.37
	+TFA	**1.79**	**2.14**	**2.55**	**2.98**	**3.42**
SMM	ResTCN	1.71	2.04	2.41	2.81	3.23
	+FA	1.78	2.14	2.53	2.96	3.39
	+TA	1.77	2.13	2.53	2.96	3.39
	+TFA	**1.85**	**2.21**	**2.60**	**3.04**	**3.46**
	MHANet	1.72	2.07	2.46	2.89	3.34
	+FA	1.77	2.14	2.54	2.97	3.39
	+TA	1.80	2.16	2.56	2.94	3.41
	+TFA	**1.83**	**2.20**	**2.60**	**3.03**	**3.45**
PSM	ResTCN	1.70	2.07	2.49	2.94	3.41
	+FA	1.78	2.15	2.58	3.03	3.49
	+TA	1.73	2.12	2.55	3.01	3.46
	+TFA	**1.80**	**2.19**	**2.62**	**3.08**	**3.53**
	MHANet	1.60	2.08	2.51	2.98	3.44
	+FA	1.78	2.15	2.57	3.02	3.48
	+TA	1.75	2.13	2.56	3.02	3.48
	+TFA	**1.80**	**2.19**	**2.61**	**3.06**	**3.49**
Xi	ResTCN	1.75	2.11	2.50	2.95	3.39
	+FA	1.85	2.21	2.61	3.06	3.50
	+TA	1.85	2.21	2.61	3.05	3.49
	+TFA	**1.90**	**2.26**	**2.67**	**3.11**	**3.53**
	MHANet	1.78	2.15	2.50	3.01	3.46
	+FA	1.85	2.21	2.52	3.06	3.50
	+TA	1.82	2.19	2.54	3.04	3.47
	+TFA	**1.88**	**2.24**	**2.59**	**3.09**	**3.51**

TABLE V
AVERAGE COVL SCORES FOR EACH SNR CONDITION AND THE HIGHEST COVL SCORES ARE HIGHLIGHTED IN BOLDFACE

Objective	Network	Input SNR (dB) -5	0	5	10	15
-	Noisy	1.17	1.32	1.58	1.92	2.37
IRM	ResTCN	1.56	1.94	2.34	2.81	3.26
	+FA	1.66	2.04	2.47	2.94	3.40
	+TA	1.65	2.04	2.48	2.95	3.37
	+TFA	**1.73**	**2.15**	**2.57**	**2.99**	**3.45**
	MHANet	1.58	1.95	2.39	2.84	3.30
	+FA	1.64	2.02	2.45	2.92	3.36
	+TA	1.67	2.07	2.49	2.94	3.38
	+TFA	**1.70**	**2.09**	**2.53**	**3.00**	**3.44**
SMM	ResTCN	1.53	1.90	2.30	2.75	3.19
	+FA	1.66	2.06	2.49	2.96	3.41
	+TA	1.65	2.06	2.50	2.97	3.41
	+TFA	**1.72**	**2.15**	**2.60**	**3.06**	**3.49**
	MHANet	1.59	1.97	2.39	2.89	3.31
	+FA	1.64	2.03	2.48	2.97	3.37
	+TA	1.67	2.08	2.52	2.98	3.42
	+TFA	**1.70**	**2.12**	**2.57**	**3.03**	**3.46**
PSM	ResTCN	1.55	1.95	2.40	2.89	3.38
	+FA	1.66	2.06	2.52	3.03	3.51
	+TA	1.64	2.08	2.55	3.05	3.51
	+TFA	**1.72**	**2.14**	**2.62**	**3.10**	**3.55**
	MHANet	1.60	2.02	2.49	3.00	3.47
	+FA	1.65	2.06	2.53	3.03	3.51
	+TA	1.65	2.07	2.54	3.05	3.53
	+TFA	**1.70**	**2.12**	**2.60**	**3.11**	**3.55**
Xi	ResTCN	1.59	1.99	2.40	2.89	3.34
	+FA	1.68	2.08	2.53	3.05	3.51
	+TA	1.67	2.09	2.54	3.03	3.49
	+TFA	**1.76**	**2.18**	**2.63**	**3.13**	**3.53**
	MHANet	1.63	2.05	2.50	3.00	3.47
	+FA	1.69	2.10	2.52	3.04	3.50
	+TA	1.68	2.10	2.54	3.03	3.48
	+TFA	**1.72**	**2.13**	**2.59**	**3.08**	**3.51**

TABLE VI
EVALUATION RESULTS ACROSS DIFFERENT NUMBER OF BUILDING BLOCKS WITH IRM AS TRAINING OBJECTIVE

Network	**# Param.**	**PESQ**	**ESTOI (%)**	**CSIG**	**CBAK**	**COVL**
Noisy	-	1.24	56.12	2.26	1.80	1.67
ResTCN-20+SA	1.32M	1.72	70.04	3.07	2.43	2.35
ResTCN-20	1.05M	1.70	68.92	3.05	2.41	2.33
+TFA	+1.36K	**1.82**	**72.17**	**3.22**	**2.51**	**2.47**
ResTCN-30+SA	1.78M	1.75	70.45	3.11	2.47	2.39
ResTCN-30	1.51M	1.73	69.57	3.09	2.43	2.36
+TFA	+2.04K	**1.90**	**73.04**	**3.28**	**2.56**	**2.55**
MHANet-4	3.29M	1.75	70.58	3.12	2.45	2.39
+TFA	+0.27K	**1.85**	**71.98**	**3.26**	**2.52**	**2.51**
MHANet-6	4.86M	1.78	71.22	3.15	2.45	2.41
+TFA	+0.41K	**1.91**	**73.37**	**3.30**	**2.59**	**2.56**
2S-SA-TCN	2.87M	1.78	71.64	3.19	2.48	2.44
+TFA	+3.2K	**1.94**	**74.43**	**3.38**	**2.62**	**2.63**
Conformer-5	6.97M	1.86	73.16	3.26	2.56	2.52
+TFA	+0.68K	**1.99**	**75.32**	**3.42**	**2.68**	**2.68**

The highest scores are highlighted in boldface.

the MHANet. The ablation results on ASR performance also illustrate the efficacy of the TA and FA modules.

In Table VIII, we compare the computation required by the models (ResTCN, ResTCN+TFA, MHANet, and MHANet+TFA), in terms of real-time factor (RTF) [62], which is the ratio of the time taken to process a speech utterance to the duration of the utterance. The RTFs are measured on an NVIDIA Tesla V100 GPU, averaged over 10 executions. We use a batch size of 20 noisy mixtures with a length of 7 seconds. It can be observed that the introduction of the TFA module does not significantly increase the computational cost.

TABLE VII
AVERAGE WER% RESULTS FOR EACH SNR CONDITION AND THE WER% SCORES AVERAGED OVER ALL NOISY CONDITIONS. THE LOWEST WER% SCORES ARE HIGHLIGHTED IN BOLDFACE

Objective	Network	Input SNR (dB) -5	0	5	10	15	Avg
-	Noisy	90.07	61.98	31.48	19.39	12.54	43.09
IRM	ResTCN	84.57	58.78	29.79	18.30	11.35	40.55
	+FA	81.80	53.63	27.50	15.62	9.18	37.55
	+TA	82.33	49.50	26.93	14.23	8.56	36.31
	+TFA	**75.49**	**48.48**	**25.37**	**12.91**	**8.50**	**34.15**
	MHANet	84.46	53.59	28.32	17.30	11.60	39.05
	+FA	83.33	50.02	26.46	16.50	9.18	37.14
	+TA	81.79	50.20	26.82	16.94	10.05	37.16
	+TFA	**79.58**	**48.48**	**26.44**	**14.87**	**9.13**	**35.70**
SMM	ResTCN	86.82	56.86	30.53	20.00	11.32	41.10
	+FA	80.67	52.95	29.63	17.58	10.27	38.22
	+TA	82.82	51.59	27.75	16.93	9.36	37.69
	+TFA	**79.77**	**46.88**	**26.67**	**13.61**	**9.00**	**35.19**
	MHANet	84.22	53.06	27.58	18.14	11.60	38.92
	+FA	83.15	50.29	26.81	16.12	9.49	37.17
	+TA	82.35	51.47	**25.71**	17.37	**8.60**	37.10
	+TFA	**81.21**	**50.21**	27.06	**14.48**	**8.60**	**36.33**
PSM	ResTCN	84.55	58.02	32.52	18.17	10.33	40.72
	+FA	80.74	56.56	27.23	17.16	9.65	38.37
	+TA	84.44	52.07	27.63	**12.89**	9.23	37.25
	+TFA	**80.64**	**46.31**	**26.59**	15.15	**8.47**	**35.43**
	MHANet	84.83	52.78	28.26	18.76	11.56	39.24
	+FA	84.71	54.71	27.60	17.45	**8.56**	38.60
	+TA	84.04	52.65	**26.65**	17.04	8.62	37.81
	+TFA	**80.30**	**50.79**	27.98	**15.64**	10.75	**37.09**
Xi	ResTCN	88.99	58.92	32.99	17.83	10.82	41.91
	+FA	83.98	53.92	28.00	18.46	10.56	38.98
	+TA	82.79	51.14	28.90	**14.76**	11.59	37.83
	+TFA	**80.75**	**47.23**	**25.57**	17.45	**9.43**	**36.09**
	MHANet	85.58	57.62	30.02	19.73	10.97	40.83
	+FA	84.95	**52.27**	29.27	17.71	**9.95**	38.83
	+TA	**82.44**	54.44	**27.35**	18.79	11.17	38.84
	+TFA	84.47	52.93	28.79	**17.28**	**9.95**	**38.68**

TABLE VIII
EVALUATION RESULTS OF INFERENCE SPEED (RTF)

Network	ResTCN	ResTCN+TFA	MHANet	MHANet+TFA
RTF	2.63×10^{-4}	3.14×10^{-4}	2.84×10^{-4}	3.02×10^{-4}

TABLE IX
COMPARISON TO MULTIPLE SPEECH ENHANCEMENT SYSTEMS ON THE VOICEBANK-DEMAND DATASET

Method	**# Param.**	**PESQ**	**STOI (%)**	**CSIG**	**CBAK**	**COVL**
Noisy	-	1.97	92	3.35	2.44	2.63
SEGAN [8]	43.2M	2.16	93	3.48	2.94	2.80
DSEGAN [63]	–	2.35	93	3.55	3.10	2.93
MetricGAN [64]	1.89M	2.86	–	3.99	3.18	3.42
PHASEN [65]	8.41M	2.99	–	4.21	3.55	3.62
TFT-Net [41]	–	2.75	–	3.93	3.44	3.34
T-GSA [30]	–	3.06	–	4.18	**3.59**	3.62
WaveCRN [66]	4.66M	2.64	–	3.94	3.37	3.29
DCCRN [67]	3.7M	2.68	93.5	3.88	3.18	3.27
DCCRN+ [68]	3.3M	2.84	–	–	–	–
S-DCCRN [69]	2.34M	2.84	94	4.03	2.97	3.43
DEMCUS [70]	58M	3.07	95	4.31	3.40	3.63
MHSA+SPK [71]	–	2.99	–	4.15	3.42	3.57
HiFi-GAN [72]	–	2.94	–	4.07	3.07	3.49
SADNUNet [73]	2.63M	2.82	95	4.18	3.47	3.51
CTS-Net [74]	4.35M	2.94	94.7	4.26	3.45	3.59
DCTCN [75]	9.7M	2.83	–	3.91	3.37	3.37
CleanUNet [62]	46.07M	2.91	**95.6**	**4.34**	3.42	3.65
SA-TCN [60]	3.76M	2.99	94.22	4.25	3.45	3.62
MetricGAN+ [76]	–	**3.15**	–	4.14	3.16	3.64
SEAMNET [77]	5.1M	–	–	3.87	3.16	3.23
ResTCN+TFA-Xi	1.98M	3.02	94.15	4.32	3.52	**3.68**

The highest scores are highlighted in boldface.

D. Comparative Study

In this section, we compare our proposed method with multiple state-of-the-art systems on the Voicebank-DEMAND

dataset [78]. As reported in Table IX, it can be observed that our proposed model ResTCN+TFA with Xi training objective (ResTCN+TFA-Xi) demonstrates highly competitive performance to those methods with respect to the five evaluation metrics. More significantly, our TFA module provides a simple and flexible way to improve existing network architectures for speech enhancement. It is worth noting that we focus on a small module rather than a system beyond existing enhancement systems.

VI. Conclusion

In this study, we propose the TFA module, a lightweight and flexible attention module designed to model the distribution of speech components in the T-F representation, improving the representational power of a network. Our TFA module consists of two parallel attention branches, i.e., the TA and FA modules, which produce two attention maps to model the speech distribution along time-frame and frequency dimensions, respectively. We evaluate the TFA module as part of ResTCN and Transformer backbone networks and adopt four widely used training objectives to conduct extensive speech enhancement experiments.

Our experimental results demonstrate that the TFA module consistently provides significant improvements to the baseline networks in terms of five metrics (PESQ, ESTOI, CSIG, CBAK, and COVL). Moreover, the evaluation results on the downstream ASR also demonstrate the effectiveness of the TFA module. This reveals the importance of the priori about the energy distribution of speech for speech enhancement, and the inability of the previous models to capture that priori. We believe that the success of the TFA module provides a new idea for the design of network architecture to boost speech enhancement. In future studies, we plan to further investigate the effectiveness of the TFA module across other commonly used datasets and other speech processing tasks such as speech recognition. In addition, we will also try to extend our proposed TFA module to multi-channel scenarios.

Acknowledgment

The model architecture design is done as a collaboration between Meta and other affiliations. All experiments are performed by National University of Singapore.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. Conf.*, 1996, vol. 2, pp. 629–632.

[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. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] M. Krawczyk-Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2251–2262, Dec. 2016.

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

[6] 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.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

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

[9] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[10] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6875–6879.

[11] 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, Aug. 2019.

[12] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 825–838, 2020.

[13] 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, Dec. 2014.

[14] 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, Mar. 2016.

[15] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013, pp. 436–440.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

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

[19] 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.

[20] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[21] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Boston, MA, USA: Springer, 2005, pp. 181–197.

[22] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoustical Soc. Amer.*, vol. 141, no. 6, pp. 4705–4714, 2017.

[23] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3709–3713.

[24] F. Weninger et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2015, pp. 91–99.

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

[26] 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, Jan. 2019.

[27] 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.

[28] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation," in *Proc. INTERSPEECH*, 2019, pp. 3183–3187.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[30] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6649–6653.

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

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

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

[34] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 531–11 539.

[35] S. Zhao et al., "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. Conf. Artif. Intell.*, vol. 34, no. 01, 2020, pp. 303–311.

[36] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6304–6308.

[37] C. Wang, J. Yi, J. Tao, Y. Bai, and Z. Tian, "Hierarchically attending time-frequency and channel features for improving speaker verification," in *Proc. IEEE 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[38] 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.

[39] Q. Zhang, Q. Song, Z. Ni, A. Nicolson, and H. Li, "Time-frequency attention for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7852–7856.

[40] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *Proc. AAAI Conf. Artif. Intell.* 2021, pp. 14549–14557.

[41] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 3816–3822.

[42] Y. Zhao and D. Wang, "Noisy-reverberant speech enhancement using DenseUNet with time-frequency attention," in *Proc. INTERSPEECH*, 2014, pp. 3261–3265.

[43] A. Nicolson and K. K. Paliwal, "On training targets for deep learning approaches to clean speech magnitude spectrum estimation," *J. Acoustical Soc. Amer.*, vol. 149, no. 5, pp. 3273–3293, 2021.

[44] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[47] 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. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 3110–3113.

[48] G. Hu, "100 nonspeech environmental sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, 2010.

[49] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 2204–2208.

[50] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *Proc. IEEE 38th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2016, pp. 736–739.

[51] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," TNO Inst. Percep., Soesterberg, The Netherlands, Rep. IZF 1988-3, 1988.

[52] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 1041–1044.

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

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[55] "862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs," ITU-Telecommun. Standardization Sector, Geneva, Switzerland, Rep. P.862.2, 2007.

[56] 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, Nov. 2016.

[57] 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, Jan. 2008.

[58] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[59] E. Kim and H. Seo, "SE-Conformer: Time-domain speech enhancement using conformer," in *Proc. INTERSPEECH*, 2021, pp. 2736–2740.

[60] J. Lin, A. J. d. L. V. Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3440–3450, 2021.

[61] A. Y. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.

[62] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7867–7871.

[63] H. Phan et al., "Improving GANs for speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1700–1704, 2020.

[64] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2031–2041.

[65] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9458–9465.

[66] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 2149–2153, 2020.

[67] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH*, 2020, pp. 2472–2476.

[68] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement," in *Proc. INTERSPEECH*, 2021, pp. 2816–2820.

[69] S. Lv et al., "S-DCCRN: Super wide band DCCRN with learnable complex feature for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7767–7771.

[70] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. INTERSPEECH*, 2020, pp. 3291–3295.

[71] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 181–185.

[72] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. INTERSPEECH*, 2020, pp. 4506–4510.

[73] X. Xiang, X. Zhang, and H. Chen, "A nested u-net with self-attention and dense connectivity for monaural speech enhancement," *IEEE Signal Process. Lett.*, vol. 29, pp. 105–109, 2022.

[74] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.

[75] R. Jigang and M. Qirong, "DCTCN: Deep complex temporal convolutional network for long time speech enhancement," in *Proc. INTERSPEECH*, 2022, pp. 5478–5482.

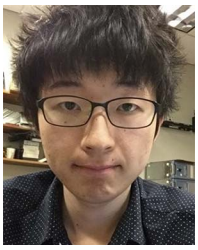
[76] S.-W. Fu et al., "MetricGAN : An improved version of MetricGAN for speech enhancement," in *Proc. INTERSPEECH*, 2021, pp. 201–205.

[77] B. J. Borgström and M. S. Brandstein, "Speech enhancement via attention masking network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 515–526, 2021.

[78] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 146–152.

Qiquan Zhang (Member IEEE) received the B.Sc. and Ph.D degrees in electronic science and technology from the Harbin Institute of Technology, China, in 2015 and 2020, respectively. Since March 2021, he has been a Postdoctoral Research Fellow with the Human Language Technology Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, under the supervision of Prof. Haizhou Li. He is currently a Postdoctoral Research Associate with the Signal Processing Lab, University of New South Wales, Sydney, NSW, Australia, under the supervision of Prof. Eliathamby Ambikairajah and Prof. Haizhou Li. His research interests include statistical signal processing, speech processing, audio-visual speech processing, speech enhancement, noise estimation, machine learning, and deep learning.

Xinyuan Qian (Member IEEE) received the B.Eng. (with first class Hons.) and M.Sc. (with Distinction) degrees from the University of Edinburgh, Edinburgh, U.K, and the Ph.D. degree from Queen Mary University of London, London, U.K. She has been a Research Fellow with the National University of Singapore, Singapore, since February 2020 and then become a Visiting Researcher with The Chinese University of Hong Kong, Shenzhen, China. She is currently an Associate Professor with the University of Science and Technology Beijing, Beijing, China. Her research interests include speech processing, audio-visual sound localization, speech enhancement, and deep learning.



Zhaoheng Ni received the B.Sc. degree in biology and medical engineering from Beihang University, Beijing, China, and the M.Sc. degree in computer science from the City University of New York, New York, NY, USA. He is currently a Research Scientist with Meta, United States. His research interests include single-channel and multi-channel speech enhancement, speech separation, and automatic speech recognition.

Aaron Nicolson received the B.Eng. degree (first class A Hons.) from Griffith University, Brisbane, QLD, Australia, in 2016, and the Ph.D. degree from the Signal Processing Laboratory of Griffith University, Brisbane, Australia under the supervision of Kuldip K. Paliwal. Since 2020, he has been a Postdoctoral Research Fellow with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), stationed, the Australia e-Health Research Centre, Brisbane,. He is also a part of CSIRO's machine learning and Artificial Intelligence Future Science Platform. His research interests include speech processing, speech enhancement, and medical image analysis.

Eliathamby Ambikairajah (Senior Member IEEE) received the B.Sc. (Eng.) (Hons.) degree from the University of Sri Lanka, Sri Lanka, and the Ph.D. degree in signal processing from Keele University, Keele, U.K. He was appointed as the Head of Electronic Engineering and later Dean of Engineering with the Athlone Institute of Technology, Athlone, Ireland, from 1982 to 1999. His key publications led to his repeated appointment as a short-term Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999. He was the Acting Deputy Vice-Chancellor Enterprise during 2020, after previously serving as the Head of School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Sydney, NSW, Australia from 2009 to 2019. He has authored or coauthored approximately 300 journal and conference papers and was the recipient of many competitive research grants. His research interests include speaker and language recognition, emotion detection and biomedical signal processing. He was a Faculty Associate with the Institute of Infocomm Research (A*STAR), Singapore from 2010 to 2018, and was an Advisory Board Member of the AI Speech Lab with AI Singapore (2019–2021). Prof. Ambikairajah was an Associate Editor for the IEEE TRANSACTIONS ON EDUCATION from 2012 to 2019. He was the recipient of the UNSW Vice-Chancellor's Award for Teaching Excellence in 2004 for his innovative use of educational technology and innovation in electrical engineering teaching programs, and in 2014, UNSW Excellence in Senior Leadership Award. In 2019 was the recipient of the People's Choice Award as part of the UNSW President's Awards. Prof. Ambikairajah was an APSIPA Distinguished Lecturer for the 2013–2014 term. He is a Fellow and a Chartered Engineer of the IET U.K. and Engineers Australia (EA) and a Life Member of APSIPA.

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronic engineering from the South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990 respectively. He is currently a Presidential Chair Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also an Adjunct Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Prior to that, he taught with the University of Hong Kong, Hong Kong, (1988–1990) and South China University of Technology, (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager with the AppleISS Research Centre (1996–1998), Research Director with Lernout & Hauspie Asia Pacific (1999–2001), a Vice President with InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore (2003–2016). His research interests include automatic speech recognition, speaker and language recognition, natural language processing. Dr. Li was an Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of Computer Speech and Language since 2012, an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019 and ICASSP 2022. Dr. Li is a Fellow of the ISCA, and a Fellow of the Academy of Engineering Singapore. He was the recipient of the National Infocomm Award 2002, and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.