U-Shaped Transformer with Frequency-Band Aware Attention for Speech Enhancement

Yi Li, *Student Member, IEEE,* Yang Sun, *Member, IEEE,*
Wenwu Wang, *Senior Member, IEEE,* and Syed Mohsen Naqvi, *Senior Member, IEEE*

***Abstract*—Recently, Transformer shows the potential to exploit the long-range sequence dependency in speech with self-attention. It has been introduced in single channel speech enhancement to improve the accuracy of speech estimation from a noise mixture. However, the amount of information represented across attention-heads is often huge, which leads to increased computational complexity. To address this issue, the axial attention is proposed i.e., to split a 2D attention into two 1-D attentions. In this paper, we develop a new method for speech enhancement by leveraging the axial attention, where we generate time and frequency sub-attention maps by calculating the attention map along time- and frequency-axis. Different from the conventional axial attention, the proposed method provides two parallel multi-head attentions for time- and frequency-axis, respectively. Moreover, the frequency-band aware attention is proposed i.e., high frequency-band attention (HFA), and low frequency-band attention (LFA), which facilitates the exploitation of the information related to speech and noise in different frequency bands in the noisy mixture. To re-use high-resolution feature maps from the encoder, we design a U-shaped Transformer, which helps recover lost information from the high-level representations to further improve the speech estimation accuracy. Extensive experiments on four public datasets are used to demonstrate the efficacy of the proposed method. The code of the proposed method is available at https://github.com/Yukino-3/U-shaped-Transformer-SE.**

***Index Terms*—Transformer, speech enhancement, time-frequency attention, U-shaped, frequency-band aware**

I. Introduction

SPEECH enhancement, aiming to improve the quality of the desired speech, is a crucial topic of audio signal processing, useful in many real-world applications, including automatic speech recognition (ASR), teleconferencing, hearing aids, and robotics [1]. Recently, numerous deep learning approaches have been proposed [2]–[4], giving state of the art performance.

Convolutional neural networks (CNNs) have been applied to speech enhancement by taking a two-dimensional spectrogram as an input [5]. The U-net was applied in speech enhancement where the receptive field is increased via successive down-sampling operations to improve the enhancement performance [6]. The deep residual U-net (ResU-net) was proposed by incorporating deep residual learning and dilated convolutions into the U-Net architecture [7], which aggregates contextual information by expanding receptive fields.

Following its success in natural language processing (NLP), the Transformer has been introduced for speech enhancement [8], which is based on the encoder and decoder architecture with stacked self-attention and point-wise feed-forward layers [9]–[12]. In the Transformer-based methods, the attention map is extracted from the spectrogram to guide the models to focus on important frames or channels. However, the Transformer architecture suffers from a limitation that the network training can be computationally expensive because significant amount of information needs to be represented across attention-heads [13]. To reduce its computational complexity, in another Transformer model [14], the attention maps in weights and heights of the feature maps are interleaved, which enables the features to be extracted along the individual axes.

Most deep learning architectures for speech enhancement are formulated in the full-band time-frequency (T-F) representation of the noisy mixture [11], [12], [15]–[18]. By using short-time Fourier transform (STFT), the state-of-the-art methods estimate the spectrogram of the desired speech signal from the noisy mixture spectrogram [19]–[21]. Moreover, some recent works focus on the time domain to avoid the long latency in calculating the spectra [9]. Tang et al. use speech signals in the time-frequency domain and time domain jointly to further improve the estimation accuracy [10]. However, it has been shown that most of the background noises, e.g., factory noise, tend to be uniformly distributed across the full band, while human speech mostly occupies in the lower frequency band [22], [23].

In this work, we leverage the advantage of the Transformer architecture for speech enhancement, and construct the attention maps along the time and frequency directions. We then introduce skip connections in Transformer to reduce the loss of feature information at each convolution [24]. We propose to divide the whole T-F attention map into three sub attention maps, i.e. time attention (TA), high frequency-band attention (HFA), and low frequency-band attention (LFA), respectively. Since most of the speech energy of the mixture is contained in the lower band 0-4000 Hz [22], the LFA, such as the 16-head attention with different learnable vectors for keys, values, and queries, is weighted more to exploit the desired source information, while the HFA is only trained with small weights and an overall learnable vector to improve the efficiency.

The contributions of this paper are summarized as follows:

- A U-shaped Transformer, simplified as U-Transformer, is introduced for the first time to address the speech enhancement

Y. Li and S.M. Naqvi are with the Intelligent Sensing and Communications Group, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (e-mails: [y.li140, mohsen.naqvi]@newcastle.ac.uk)

Y. Sun is working with the Big Data Institute, University of Oxford, Oxford OX3 7LF, U.K. (e-mail: yang.sun@bdi.ox.ac.uk)

W. Wang is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk)

problem. The proposed method can address the limitation of U-Nets and offers advantages in modeling long-range contextual and spatial information. Furthermore, the skip connections are added between the sub-layers in the encoder and decoders, which can reduce the degradation caused by the increasing depth of the network.

• The 2-D attention map is split into two 1-D sub-attention maps over time and frequency, which enables parallel calculations of the attention maps and thus facilitates the training process. In addition, independent learnable vectors for query, keys and values are exploited as local constraints between the frames of the sub-attention maps. For each location on the feature map, a local squared region is extracted to serve as a memory bank for computing the attention map.

• The multi-head attentions in time- and frequency-axis are further refined into three attentions i.e., time attention (TA), high frequency-band attention (HFA), and low frequency-band attention (LFA). Furthermore, a different number of heads and learnable vectors are used for the three multi-head attentions to efficiently train the sub-attention maps.

II. Related Work

A. Network Architectures in Speech Enhancement

Various network models for speech enhancement have been developed in the deep learning community. Over the past several years, increasing research efforts have been devoted to improve the inference efficiency of DNNs for speech enhancement. For example, in [15], three different techniques are systematically investigated with feed-forward DNN-based pipelines. The magnitude spectra of the clean speech signal and the noisy mixture signal are used as the network input. Moreover, as a natural choice for learning the temporal dynamics of speech, recurrent blocks are widely used to model the speech signal in regression tasks in various methods. For example, Leglaive et al. presented a generative approach to speech enhancement based on a recurrent variational autoencoder (RVAE) with a variational expectation-maximization algorithm [3]. In a way, a recurrent neural network (RNN) can be viewed as a DNN with an infinite depth [25].

Different from RNNs, the Transformer processes the input in parallel and does not necessarily depend on the inputs from the previous frames to be processed. The Transformer adopts the scaled dot-product attention of the query Q with all keys K, followed by division of a constant [26]. After applying the softmax function, the weights W on the values V are obtained. The attention of each head is the dot product of W and V. The attentions of all heads are concatenated and linearly projected again to obtain the final output [26].

U-shaped neural networks have recently been introduced in speech enhancement [6], where the U-net architecture is used to supplement a usual contracting network with successive layers, where pooling operations are replaced by upsampling operators [6]. Therefore, these layers increase the resolution of the output. A successive convolutional layer can then learn to assemble a precise output based on this information. The network is based on the fully convolutional network and its architecture has been modified and extended to work with a reduced number of training samples and to yield performance improvement [6]. A modified U-Net and a temporal activation layer (TAU-Net) have been jointly optimized to boost the speech enhancement performance in unseen noise environments [27].

In order to further improve the performance, the residual connection is introduced in speech enhancement [28], [29]. Instead of fitting each few stacked layer directly with a desired underlying mapping, the layers are designed to fit with a residual mapping. The deep residual networks are found to be easy to optimize, which provide significant performance gain with the greatly increased depth [28]. Moreover, in [7], the residual connections are combined with the U-net to aggregate contextual information by expanding the receptive fields. The residual blocks are summed to yield high-level features, which preserve and integrate the knowledge learned by all the stacked blocks of ResU-net.

The U-shaped architecture is combined with self- and cross-attention from Transformers known as U-Transformer for image segmentation [30]. The conventional U-nets are ineffective in modelling the long-range contextual interactions and spatial dependencies, which, however, are crucial for accurate segmentation in challenging contexts. The U-Transformer augments the U-shaped fully connected layers with Transformers [30]. To this end, attention mechanisms are incorporated at two main levels: a self-attention module leverages global interactions between encoder features, while the cross-attention module in the skip connections allows a fine spatial recovery in the U-Net decoder by filtering out the non-semantic features. In this work, we leverage the U-shaped Transformer, and adapt it for the speech enhancement problem.

B. Attention Based Speech Enhancement

Inspired by the huge success in natural language processing (NLP), attention based network models have been introduced to solve speech enhancement problem [31]. As a Squeeze-and-Excitation (SE) block, an attention mechanism operates in two steps. In the first step, it squeezes the input tensor over the time and frequency axis to output a one-dimensional vector [32]. The squeezing operation is an average pooling that enables the whole spatial information to be compressed into one bin. It embeds the input data into a global vector so that contextual information can be exploited in the second step. In the second step, the one-dimensional vector is passed to a multi-layer perception module composed of two fully-connected layers.

In recent years, different kinds of attention based variants are developed. Self-attention is a core building block of the Transformer, which not only enables parallelization of sequence computation, but also provides the paths of constant length between the symbols that are essential to learning long-range dependencies [33]. Compared to the conventional attention mechanism, the self-attention minimizes the total computational complexity per layer and maximizes the amount of parallelizable computations [17].

Rather than estimating the attention block only once, the scaled dot-product attention is utilized for the parallel calculation of the multi-head attention multiple times [34]. The

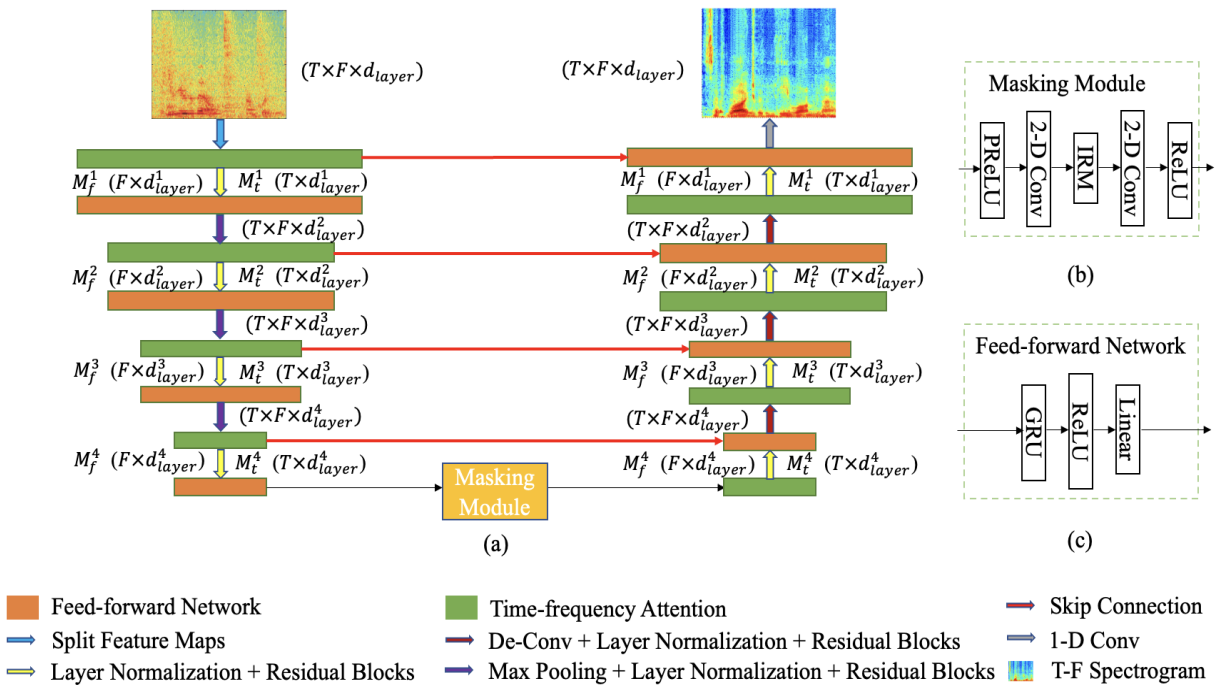

Fig. 1. The overall architecture of the proposed U-Transformer. As the input of the encoder, the features extracted from the noisy spectrogram are split into time- and frequency-axis, and masked by multi-head attentions. The weights for time and frequency W_t and W_f are separately calculated in time attention and frequency attention, respectively. The attention maps of the two attentions M_t and M_f are output from each time-frequency attention. However, the dimensions of M_t and M_f from each attention block are halved compared with the previous block. For example, d^2_{layer} is half of d^1_{layer}. After the feature is recovered by the De-Conv layers in the decoder, the reconstructed spectrogram is obtained as the output. The masking module and feed-forward network are presented in (b) and (c), respectively.

independent attention outputs are simply concatenated and linearly transformed to match the expected dimensions [35]. In addition, for each multi-head attention block, the feed-forward layer, followed by a layer normalization, is used to process the output from the attention layer, which allows the input to be rescaled and fit for the next sub-layer [33]. The two linear transformations exploited in the feed-forward layer share the same architecture across different positions, however, they employ different parameters between the two-layer normalization [36].

As aforementioned, Transformer has the limitation that the network training requires huge computational cost due to massive information represented across attention-heads [13]. To address this issue, Wang et al. proposed an axial method attention where the 2D self-attention is factorized into two 1-D self-attentions for panoptic segmentation [37]. The authors employed two axial-attention layers consecutively for the height-axis and width-axis of the feature map. Motivated by this idea, in this paper, we develop a method for speech enhancement based on axial attention where the attention map is calculated along the time- and frequency-axis. However, different from the conventional axial attention, in our work, independent learnable vectors are exploited for query, keys and values, and used as local constraints added between the frames of the sub-attention maps e.g., for each location on the feature map, a local squared region is extracted to serve as a memory bank for computing the attention map. This has an advantage in further reducing the computational cost involved in the conventional axial attention.

III. Proposed Method

In this section, we present the T-F attention based U-Transformer with the frequency-band aware attentions. Each block in the overall architecture of the network is introduced in the first subsection, followed by the description of the encoder-decoder U-Transformer, T-F attention and frequency-band aware attention with different multi-head attention parameters in the remaining subsections.

A. Speech Enhancement U-Transformer

The overall architecture of the proposed U-Transformer is presented in Fig. 1. The aim of speech enhancement is to estimate the desired speech signal $\mathbf{Y}^t = \{\mathbf{y}^t, \dots, \mathbf{y}^t\}$ from the t-length inputs $\mathbf{X}^t = \{\mathbf{x}^t, \dots, \mathbf{x}^t\}$, where $\mathbf{x}^t$ and $\mathbf{y}^t$ represent the t-th frame of the magnitude spectrogram of noisy mixture and estimated speech, respectively.

Initially, the noisy mixture is generated with the clean speech signal and the noise interference. The encoder consists of four Transformer blocks, and each block has a T-F multi-head attention, a feed-forward network, and two layer-normalizations.

The conventional Transformer encoder consists of three important modules: positional encoding, multi-head attention, and position-wise feed-forward network. However, in speech enhancement, the positional encoding part is removed since it is not suitable for acoustic sequence [33]. As shown in Fig. 1, the encoder is comprised of four sub-layers and each encoder layer has a multi-head attention and a feed-forward network. The residual connection [28] is exploited

in both multi-head attention mechanism and feed-forward network. In the proposed T-F attention method, two multi-head attention blocks are applied on sub attention maps to extract the desired information at time- and frequency-axis, respectively. Different from the conventional Transformer encoder, in the feed-forward network, the first fully connected (FC) layer is replaced by a gated recurrent unit (GRU) [38] layer because the GRU shows a better performance in recent speech enhancement works [39]. In addition, the GRU layer has a simpler structure, and thus, it is easier to implement, and also faster to train [38]. Moreover, the same dimension of attentions maps is obtained from the input and output of one sub-layer in the U-Transformer, e.g., $d_{\text{layer}} = 512$ in the first sub-layer, to facilitate the residual connections. A layer normalization is followed by both the multi-head attention and the feed-forward network before and after the operations.

The output from the encoder stack is provided to the masking module which consists of two 2-D convolutional layers with ReLU and PReLU activation functions. We use two different activation functions here for the following reasons. With PReLu, which has the slope as a parameter of the model, the speech source can be estimated more accurately because of the flexible range of values of the attention map, including both negative and non-negative values. However, the ideal ratio mask (IRM) is estimated with the energy of the attention maps of clean speech and noisy mixture, which involves only non-negative values. Therefore, we use ReLu for the outputs from the IRM estimation, but use PReLu for the estimation of the speech source.

Because information is propagated along time- and frequency-axis, to utilize the conditioning information, i.e., the relationship between each time and frequency point, the masking module is exploited between the encoder and decoder. The encoded representation is shifted up for the causality of the conditioning information with the IRM to estimate the target speech signal from the noisy mixture as [40]:

$$IRM = \left(\frac{S^2}{S^2 + I^2} \right)^{\beta} \tag{1}$$

where S^2 and I^2 are speech energy and interference energy, respectively, which can be calculated from each T-F point. According to [40], the tunable parameter β is typically set to 0.5 as an appropriate choice. In the sub-layer of the decoder, the feed-forward network obtains the masked attention maps from two inputs: (1) The attention map from the corresponding sub-layer in the encoder is introduced by a skip connection, which helps reduce the loss of feature information at each convolution [24]. (2) The output from the T-F attention block of the decoder. A concatenation operation is applied in the feed-forward network at each sub-layer of the decoder to integrate two inputs. Different from the conventional Transformer [26], the proposed U-Transformer benefits from the concatenation as combining the information from different compression levels is found empirically to improve performance in speech enhancement, as shown later in our experiments. Moreover, similar to the encoder, a layer normalization and residual blocks are added between the multi-head attention and the feed-forward network. The enhanced speech can be obtained from the output layer after the 1-D convolutional layer.

B. Time-frequency Attention

In the proposed U-Transformer, a self-attention layer is implemented in each sub-layer to take an L-length sequence of embeddings as input and to produce a same size output sequence. The output of the attention matrix is represented as [26]:

$$\mathcal{A}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{L}} \right) V \tag{2}$$

where $\mathcal{A}$ is attention and $\top$ is the symbol for matrix transpose. Initially, the full attention map of dimension $T \times F \times d_{\text{layer}}$ is the input to each sub-layer of the encoder. Different from the axial attention [14], due to the requirement of the speech signal, the proposed T-F attention applies global average pooling along time- and frequency-axis to split the 2-D attention map into two 1-D sub maps as $T \times d_{\text{layer}}$ and $F \times d_{\text{layer}}$, respectively. Each sub attention map propagates information along one specific axis. Moreover, the dimensions of Q, K, and V are $T \times F \times d_{\text{layer}}$, and are changed to $T \times d_{\text{layer}}$ or $F \times d_{\text{layer}}$ after the axial transformation. Two sub attention maps are constructed for parallel calculations on multiple GPUs to optimize the training process. The multi-head attention for the time direction can be represented as:

$$\begin{aligned} \text{multihead}(Q_t, K_t, V_t) &= [\text{h}_1; \ldots; \text{ h }_8]\, W_t^O \\ \text{with } \text{h}_i &= \mathcal{A}\left(Q_t W_t^Q, K_t W_t^K, V_t W_t^V \right) \end{aligned} \tag{3}$$

where $W_t^O \in \mathbb{R}^{8 \times d_v \times d_{\text{layer}}}$, $W_t^Q \in \mathbb{R}^{d_{\text{layer}} \times d_k}$, $W_t^K \in \mathbb{R}^{d_{\text{layer}} \times d_k}$, and $W_t^V \in \mathbb{R}^{d_{\text{layer}} \times d_v}$ are the weights required to be trained. In the time attention map, the query, keys and values are denoted as Q_t, K_t, V_t, respectively. The number of heads is set empirically to 8 and the index of each head is denoted as i in the proposed T-F attention method similar to the original Transformer [26]. The dimension of the hidden layers in each head is set empirically to 512 in our work. Similarly, for frequency attention, we use same equations but different notation f:

$$\begin{aligned} \text{multihead}(Q_f, K_f, V_f) &= [\text{h}_1; \ldots; \text{ h }_8]\, W_f^O \\ \text{where } \text{h}_i &= \mathcal{A}\left(Q_f W_f^Q, K_f W_f^K, V_f W_f^V \right) \end{aligned} \tag{4}$$

The multi-head attention mechanism shows high efficacy to learn the long-term dependencies because a direct connection between the frames is used. The weights of the multi-head attention layer are computed by pooling over the query-key affinities $Q_t K_t$, and the key-dependent bias term $K_c r_{c-t}^{K_t}$:

$$W_t^Q = \sum_{c \in \mathcal{N}_{1 \times n}(t)} (Q_t K_t + K_c r_{c-t}^{K_t}) r_{c-t}^{V_t} \tag{5}$$

$$W_t^K = \sum_{c \in \mathcal{N}_{1 \times n}(t)} Q_t r_{c-t}^{Q_t} r_{c-t}^{V_t} \tag{6}$$

$$W_t^V = \sum_{c \in \mathcal{N}_{1 \times n}(t)} (Q_t K_t + Q_t r_{c-t}^{Q_t} + K_c r_{c-t}^{K_t}) V_c \tag{7}$$

where $\mathcal{N}_{1\times n}(t)$ is the local $1 \times n$ region around the frame c. The location similarity is estimated by the inner product $Q_t r_{c-t}^{Q_t}$ between the frames (t, c). Then, r_{c-t}s are learnable vectors to update the weights and the superscripts refer to the multi-head attention parameters. The outputs of time and frequency attentions can be written as:

$$M_t = \text{multihead}(Q_t, K_t, V_t) \tag{8}$$

$$M_f = \text{multihead}(Q_f, K_f, V_f) \tag{9}$$

Then, the masked attention maps are integrated and processed by a feed-forward network to obtain the output of the improved Transformer decoder at time t, where residual connections and layer normalization $h(\cdot)$ are added as well.

$$\mathbf{y}_i^t = \text{ReLU}(h(M_t + M_f + M_p))W_i + b_i \tag{10}$$

where the i-th weight $W_i \in \mathbb{R}^{d_{layer}\times T}$ and the i-th bias $b_i \in \mathbb{R}^T$ are trained with the output from the previous layer M_p. The desired speech signal $\hat{\mathbf{Y}}^t$ is estimated by integrating L frames. The pseudo-code of the proposed T-F attention is summarized in Algorithm 1.

Algorithm 1: Time-frequency Attention Algorithm.

input : Extracted feature map as $T \times F \times d_{\text{layer}}$,
Attention map from the last sub-layer M_p,
learning rate η, epoch $E_{\max}$
output: Attention map M_{sum} as $T \times F \times d_{\text{layer}}$

Initialize learning vectors;
for $E = 1, 2, ..., E_{\max}$ **do**
  for $t \in [1, T]$ **do**
    Calculate $r_{c-t}^{Q_t}$, $r_{c-t}^{K_t}$ and $r_{c-t}^{V_t}$;
    W_t^O, W_t^Q, W_t^K, $W_t^V \leftarrow \mathbf{r}_{c-t}^{Q_t}$, $r_{c-t}^{K_t}$, $r_{c-t}^{V_t}$;
    Update the time attention map M_t;
  end
  for $f \in [1, F]$ **do**
    Calculate $r_{g-f}^{Q_f}$, $r_{g-f}^{K_f}$ and $r_{g-f}^{V_f}$;
    W_f^O, W_f^Q, W_f^K, $W_f^V \leftarrow \mathbf{r}_{g-f}^{Q_f}$, $r_{g-f}^{K_f}$, $r_{g-f}^{V_f}$;
    Update the frequency attention map M_f;
  end
  $M = M_f \times M_t$;
  $\mu_i \leftarrow x_{ij}$ //mini-batch mean;
  $\sigma_i \leftarrow x_{ij}, \mu_i$ //mini-batch variance;
  $\hat{x}_{ij} \leftarrow x_{ij}, \mu_i, \sigma_i, \epsilon$ //normalize with error ϵ;
  $M_{sum} = M + M_p$;
end

C. Frequency-band Aware Attention

To fully exploit the desired speech information, T-F attention is further divided into three multi-head attentions, time attention (TA), high frequency-band attention (HFA), and low frequency-band attention (LFA), as shown in Fig. 2.

The proposed attention block has three frequency-band aware multi-head attention mechanisms. The input of the block is a $T \times F \times d_{\text{layer}}$ attention map which is divided into two sub attention maps by a 1-D convolution to shuffle the features in time- and frequency-axis. The frequency attention map is further divided into two $\frac{F}{2} \times d_{\text{layer}}$ sub-maps based on the critical frequency f_c. According to [22], [41], the critical frequency $f_c = 4000$ Hz is found to be the best choice due to significant difference of the power spectral density (PSD) between the lower frequency band and higher frequency band of the mixture spectra. When the sampling rate and the maximum frequency of the speech signal are set to 16 kHz and 8 kHz, respectively, the frequency band [4000 - 8000 Hz] is assumed as the high frequency-band and paid with smaller computation costs during the training because it only includes unvoiced speech and limited voiced speech energy. The other frequency-band in the mixture, i.e., the lower band [0 - 4000 Hz], which is composed of mostly the voiced speech and is the major focus of the frequency-band aware attention. Because the target speech signal is composed of both voiced and unvoiced speech components, the unvoiced speech in the high frequency-band may affect the speech enhancement performance. However, the vocal folds are the primary sound source and the average pitch frequency is about 125 Hz for an adult male, 210 Hz in adult females, and over 300 Hz in children [42]. Therefore, the proposed method uses different multi-head attentions for the sub-bands.

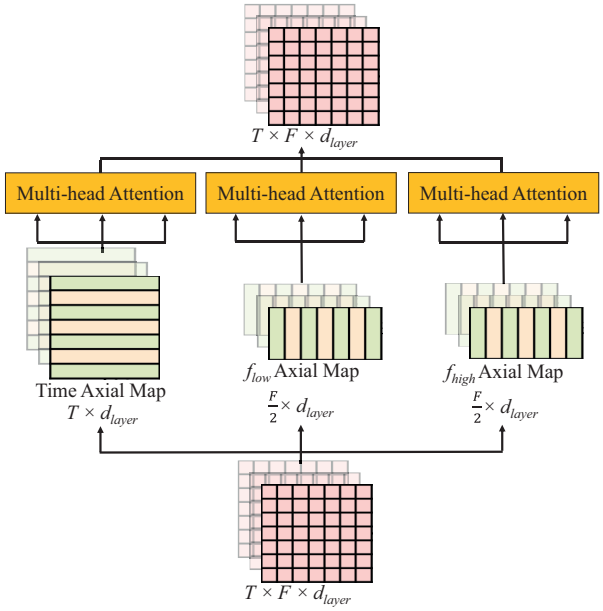

Fig. 2. The frequency-band aware attention. The input and output maps have the same size as $T \times F \times d_{\text{layer}}$ at each sub-layer in the proposed U-Transformer. In four sub-layers, $d_{\text{layer}} = 512, 256, 128, 64$, respectively.

In the proposed T-F attentions method, we use eight-head attentions for both time- and frequency-axis. However, in the frequency-band aware attention method, the numbers of the heads in TA, HFA, and LFA are set to 8, 2, and 16, respectively. In the conventional multi-head attention [14], [26], the number is set to 8 for the full frequency-band of the spectrogram. As the energy of the desired speech signal is distributed at the lower frequency-band [0-4000 Hz], the LFA is trained with 16-head attention and independent learnable vectors for time attention, which incurs more computation loads than the HFA. However, in the HFA, we only exploit an overall vector r_{g-f}^{hf} and the output of the multi-head attention

is represented as:

$$W_f^Q = \sum_{g \in \mathcal{N}_{1\times n}(f)} (Q_f K_f + K_g) r_{g-f}^{hf} \tag{11}$$

$$W_{hf}^K = \sum_{g \in \mathcal{N}_{1\times n}(f)} Q_f r_{g-f}^{hf} \tag{12}$$

$$W_{hf}^V = \sum_{g \in \mathcal{N}_{1\times n}(f)} (Q_f K_f + (Q_f + K_g) r_{g-f}^{hf}) V_g \tag{13}$$

The three multi-head attention blocks are trained with different sub attention maps of the extracted features and provide a combined and masked attention map which share the same size as the feature map. The integrated attention map is added to the layer normalization with a residual connection. The pseudo-code of the high frequency-band attention is summarized in Algorithm 2.

Algorithm 2: High Frequency-band Attention.

input : High frequency-band attention map as $\frac{F}{2} \times d_{\text{layer}}$, learning rate η, epoch $E_{\max}$, Estimated LFA M_{lf}, time attention map M_t from Algorithm 1

output: Attention map as $T \times F \times d_{\text{layer}}$

Initialize learning vectors;
for $E = 1, 2, ..., E_{\max}$ **do**
  for *each column* $f \in [1, \frac{F}{2}]$ **do**
    Calculate r_{c-t}^{hf};
    W_{hf}^O, W_{hf}^Q, W_{hf}^K, $W_{hf}^V \leftarrow r_{g-f}^{hf}$;
    Update the frequency attention map M_{hf};
  end
  $M = M_{hf} + M_{lf} + M_t$ //integrate three sub attention maps ;
end

IV. Experimental Results

A. Datasets

We extensively perform experiments on several public datasets, including DEMAND [43], IEEE [44], TIMIT [45], VOICE BANK (VCTK) [46], and Deep Noise Suppression (DNS) challenge [47].

1) DEMAND: Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [43] provides a set of recordings from real-world noise. We randomly collect and use 6 of 15 recordings for the speech enhancement experiments with noise interferences, and the noises are $psquare$, $dliving$, $dkitchen$, $nriver$, $tcar$ and $pstation$. Each noise interference has a unique case and lasts four minutes long, and it is divided into two clips with an equal length. One is used to match the lengths of the speech signals to generate training data and the other is used to generate development and inference data.

2) IEEE & TIMIT: The IEEE dataset [44] contains speech data of American English speakers. The TIMIT dataset [45] contains broadband recordings from 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. In the training and development stages, 600 recordings from 60 speakers and 60 recordings from 6 speakers are randomly selected in each dataset, respectively.

3) VCTK: The VOICE BANK dataset [46] already constitutes the largest datasets of British English. According to [6], [7], 11572 noisy mixtures are generated with 6 background noises at one of 4 SNR levels (15, 10, 5, and 0 dB) in the training stage.

4) DNS: The clean speech set includes over 500 hours of clips from 2150 speakers and the noise set includes over 180 hours of clips from 150 classes in the DNS challenge [47]. In the training stage, 75% of the clean speeches are mixed with the background noise but without reverberation. In the testing stage, 150 noisy clips are randomly selected from the blind test dataset without reverberations.

B. Baselines and Model Configuration

In this work, three baseline models, including U-net [6], ResU-net [7], and SETransformer [8] with attention [48] are implemented for the comparison and ablation experiments.

Both U-net and ResU-net baselines use 1D convolution and zero-padded blocks. The number of layers are set to 10 as the best configuration reported in [6]. Different from the U-net, the downsampling and upsampling blocks are constructed as residual units in ResU-net. The ResU-net consists of an identity mapping and two 1-D convolution blocks. Each convolution block includes a dilated convolution layer, a batch normalization (BN), and a Leaky ReLU activation function. Dilation is applied to both the time direction and the frequency direction in the convolution operation, which can aggregate contextual information over both time and frequency dimensions. The identity mapping with 1x1 convolution connects the input and output of the unit, which is only used to ensure the same dimensions of two tensors that are passed to an addition operation [7].

In addition, six state-of-the-art speech enhancement methods [9]–[11], [33], [39], [49] are reproduced as the original implementations and compared with the proposed method. The first one, TSTNN [9], is a two-stage Transformer network for speech enhancement in the time domain which uses four stacked two-stage transformer blocks to extract local and global information from the speech latent representation stage by stage. The second method is a cross-domain framework named TFT-Net [10], which exploits time-frequency spectra as input to six dual-path attention blocks and produces time domain waveform as output. The third method is a dual-path Transformer network (DPTNet) for end-to-end speech separation [11], and we use the background noise as the interference for fair comparison. The fourth method, named DPT-FSNet [49], combines the full-band and sub-band fusion (FullSubNet) method in [23] and DPTNet in [11]. The inter and intra parts of the dual-path Transformer model the full-band and sub-band information, respectively. The T-GSA method [33]

TABLE I
SPEECH ENHANCEMENT PERFORMANCE COMPARISONS WITH BASELINES ON THE IEEE AND TIMIT DATASETS. EACH RESULT IS THE AVERAGE OF 2160 EXPERIMENTS (120 SIGNALS × 3 SNR LEVELS × 6 BACKGROUND NOISES). **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Computation		IEEE			TIMIT		
Method	Para. (M)	FLOPs (B)	STOI (%)	PESQ	fwSNRseg (dB)	STOI (%)	PESQ	fwSNRseg (dB)
Unprocessed	-	-	42.3	1.52	3.11	41.5	1.44	3.04
U-net [6]	30.00	4.44	68.2	1.93	8.81	71.3	1.96	10.66
Unet+Attention [6] [48]	35.52	4.46	68.5	1.92	8.96	67.8	1.97	8.42
ResU-net [7]	13.34	1.80	72.6	2.14	10.77	72.1	2.09	9.50
ResU-net+Attention [7] [48]	18.95	1.82	74.0	2.20	12.51	73.0	2.08	9.38
Transformer [8]	6.44	2.31	78.4	2.32	13.08	77.9	2.25	12.70
Transformer+Attention [8] [48]	11.92	2.33	78.8	2.41	13.24	78.3	2.33	12.84
TSTNN [9]	0.92	0.94	78.5	2.30	13.22	78.1	2.30	12.80
TFT-Net [10]	8.96	1.22	78.7	2.35	12.70	77.8	2.27	12.93
DPTNet [11]	2.69	1.39	82.3	2.38	14.34	81.1	2.28	13.06
FullSubNet [23]	5.75	2.98	82.5	2.42	15.07	81.9	2.30	13.54
DPT-FSNet [49]	**0.88**	0.60	83.2	2.49	15.88	82.2	2.40	13.75
Unet+TF	36.81	3.89	72.0	1.99	11.46	70.4	1.99	9.84
ResU-net+TF	20.12	1.01	80.5	2.38	13.21	78.6	2.31	13.22
U-Transformer+TF	4.21	**0.33**	81.6	2.41	15.47	80.7	2.39	14.05
Unet+FAT	37.11	3.91	78.1	2.21	12.55	75.2	2.16	10.62
ResU-net+FAT	20.93	1.05	82.6	2.49	14.92	80.7	2.40	13.16
U-Transformer+FAT	4.31	0.34	**85.0**	**2.74**	**17.39**	**82.6**	**2.59**	**14.81**

uses Gaussian-weighted self-attention (GSA) in Transformer in the encoder and a complex fully-connected layer in the decoder. The self-adaptation based multi-head self-attention (SA-MS) method [39] uses the multi-head self-attention to capture long-term dependencies in the speech and noise with a DNN backbone.

Three state-of-the-art benchmarks in the DNS challenge are reproduced and compared with the proposed method. The first one is the full-band and sub-band fusion model (FullSubNet) which captures the full-band spectral information and the long-distance cross-band dependencies, meanwhile retaining the ability to modeling signal stationarity and attending the local spectral pattern [23]. The full-band network contains three long short-term memory (LSTM) layers with 512 hidden units for each layer and the sub-band model includes two LSTM layers (384 / 256 units) and one dense layer [50]. The second one is the dual-path recurrent neural network (DPRNN) [51]. Similar to [51], we apply 6 DPRNN blocks with 128 hidden units in each direction bidirectional LSTM (BLSTM) on the time-domain audio separation network (TasNet) that contains a linear 1-D convolutional encoder, a separator, and a linear 1-D transposed convolutional decoder [52]. The third one is the deep complex convolutional recurrent network (DCCRN) [53]. In the implementation, both the CNN and RNN structures can handle the complex-valued operation. The network is an essentially causal convolutional encoder-decoder (CED) architecture with two LSTM layers between the encoder and the decoder [53].

Moreover, all the speech utterances are resampled to 16 kHz. They are converted to spectrogram using fast Fourier transform (FFT), with a window of 512 samples (32ms) with an overlap of 256 samples (16ms) between the neighboring windows. Since the input and the output of the proposed method and baselines are both magnitude spectrogram and the dimension of single axis is set to 257. A linear processing layer is stacked when splitting the feature map to convert the speech spectrogram to feature vectors of dimensions d_{layer} = 512. All the experiments are run on a work station with four Nvidia GTX 1080 GPUs and 16 GB of RAM. The proposed method is trained by using the Adam optimizer with a learning rate set empirically to 0.0008. The batch size is set to 16. We train the networks for 100 epochs, due to the use of a large amount of training data, i.e., 11572 speech signals mixed with 6 background noises. The training and validation loss curves are plotted in Fig. 3. According to these loss values, we set the number of training epochs as 100 to avoid potential overfitting.

Fig. 3. Training and validation loss curves.

C. Evaluations on the IEEE and TIMIT datasets

To evaluate and compare the quality of the enhanced speech with various methods, we use the short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and frequency-weighted segmental signal-to-noise ratio (fwSNRseg) as performance measures on the IEEE and TIMIT datasets. The STOI and the PESQ are bounded in the range of [0, 1] and [-0.5, 4.5], respectively [54]. The fwSNRseg is estimated by computing the segmental signal-to-noise ratios (SNRs) in each spectral band and summing the weighed SNRs from all bands [55] in the range of [-10,

35] dB. The proposed T-F attention and frequency-band aware attention are abbreviated as TF and FAT, respectively. We also compare the parameters and floating-point operations (FLOPs) of the models.

Table I shows the averaged speech enhancement performance of the proposed method as compared with those of the baselines using the IEEE and the TIMIT datasets, with three SNR levels (-5, 0, 5 dB) and six noise interferences i.e., $psquare$, $dliving$, $dkitchen$, $nriver$, $tcar$ and $pstation$. From Table I, it can be observed that: (1) The conventional attention block has limited improvement in speech enhancement performance. However, the proposed T-F attention and frequency-band aware attention significantly improve the inference performance in all standard models. In terms of PESQ, the proposed T-F attention and frequency-band aware attention obtain 7.3% and 22.8% improvements compared with the standard Transformer model [8], respectively. The proposed T-F attention mechanism adopts both global connection and efficient computation on time and frequency directions. In addition, the learnable vectors $r_{c-t}^{V_t}$, $r_{c-t}^{K_t}$, and $r_{c-t}^{Q_t}$ for V, K, and Q utilize the positional information between the frames (t, c) to update the weights W^V, W^K, and W^Q, respectively. The 2-D attention map is split into two 1-D sub-maps in time- and frequency-axis, which allows the parallel calculation to facilitate training [14]. (2) In all the evaluated models, the proposed frequency-band aware attention U-Transformer offers the best effectiveness. The reason is that the proposed U-Transformer inherits advantages from both ResU-net and Transformer. Moreover, the desired information at the lower frequency-band is fully used by a 16-head attention and independent learnable vectors. However, in the HFA, only an overall learnable vector r_{g-f}^{hf} is applied to further improve the performance.

Furthermore, the visualizations are given in Fig. 4 which are related to the estimated spectra of the desired speech signals from different methods. The target speech signal is randomly selected from the testing set. After comparing the estimated spectra with the spectrogram of target speech signal, it can be observed that the spectrogram obtained via the proposed U-Transformer with frequency-band aware attention is closer to the clean speech signal, which again confirms that the frequency-band aware attention U-Transformer method outperforms the baselines.

In this work, the proposed T-F attention method produces two 1-D attention maps to guide the models to focus on the time frame or frequency channel, respectively. Consequently, the feature maps along time- and frequency-axis are combined to generate a 2-D attention map enabling the models to capture the speech distribution in the T-F domain. Furthermore, by using the 16-head attention on the lower band spectra with more desired feature information, the speech enhancement performance is further improved.

D. Evaluations on the VCTK and DNS challenge datasets

In these experiments, the VCTK and DNS challenge datasets are used to further evaluate the proposed methods as compared with the state-of-the-art methods.

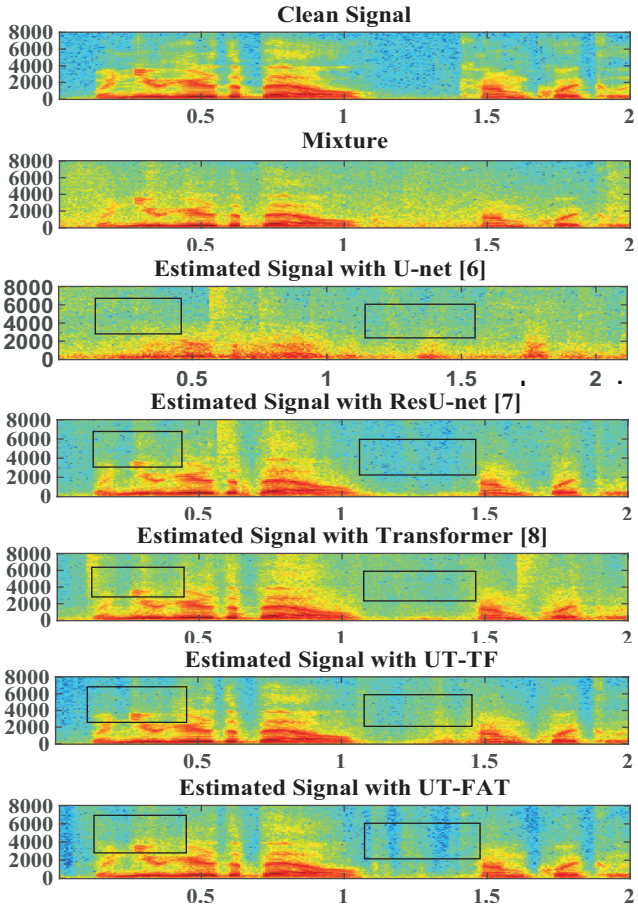

Fig. 4. The spectra of different signals: *TF* refers to T-F attention and *FAT* denotes the proposed frequency-band aware attention method. The x- and y-axis are time (s) and frequency (Hz), respectively. The experiment is implemented with *dliving* and -5 dB SNR level. The proposed UT-FAT method offers 0.37 and 0.09 improvements over the best-performing baseline, i.e. the original Transformer [8], in terms of PESQ and STOI, respectively.

The testing set with 2 speakers, unseen during training, consists of a total of 20 different noise conditions: 5 types of noise sourced from the DEMAND dataset at one of 4 SNRs each (17.5, 12.5, 7.5, and 2.5 dB). This yields 824 test items, with approximately 20 different sentences in each condition per test speaker. To evaluate and compare the quality of the enhanced speech with various methods, we use mean opinion score (MOS) predictor of signal distortion (CSIG), MOS predictor of background intrusiveness (CBAK), MOS predictor of overall speech quality (COVL) to map the enhancement between [1, 5] [56]. Furthermore, similar to [6], [7], PESQ and segmental signal-to-noise ratio (SSNR) are used as well. Table II shows the averaged speech enhancement results on the VCTK dataset [46]. From this table, we can see that the proposed method outperforms the state-of-the-art methods in terms of all performance measures.

The proposed method is further evaluated on the DNS challenge benchmark and compared with the state-of-the-art methods. In these experiments, the averaged STOI (%), wide-band PESQ (WP), narrow-band PESQ (NP), and scale-invariant source-to-distortion ratio (SI-SDR) (dB) performances are presented in Table III. In the training stage, the noisy mixtures

TABLE II
SPEECH ENHANCEMENT PERFORMANCE COMPARISON ON THE **VCTK** DATASET. **BOLD** INDICATES THE BEST RESULTS AND *italic* INDICATES THE PROPOSED METHOD.

Method	PESQ	CSIG	CBAK	COVL	SSNR
Unprocessed	1.97	3.35	2.44	2.63	1.68
U-net [6]	2.39	3.48	3.15	2.94	9.43
Unet+Attention [6] [48]	2.40	3.50	3.30	3.03	10.01
ResU-net [7]	2.79	3.99	3.38	3.32	10.03
ResU-net+Attention [7] [48]	2.83	4.04	3.49	3.45	10.08
TSTNN [9]	2.91	4.20	3.47	3.59	9.14
TFT-Net [10]	2.71	3.86	3.39	3.25	10.00
DPTNet [11]	2.78	3.92	3.40	3.36	10.83
FullSubNet [23]	2.96	4.21	3.57	3.62	11.03
T-GSA [33]	2.92	4.01	3.47	3.55	10.02
SA-MS [39]	2.91	4.07	3.36	3.50	9.31
DPT-FSNet [49]	3.02	**4.37**	3.58	**3.81**	11.14
Unet+TF	2.67	3.70	3.41	3.26	10.45
ResU-net+TF	2.88	4.12	3.53	3.58	10.38
U-Transformer+TF	2.89	4.20	3.69	3.64	10.61
Unet+FAT	2.75	3.82	3.50	3.34	11.58
ResU-net+FAT	2.96	4.20	3.59	3.61	11.27
U-Transformer+FAT	**3.08**	4.23	**3.63**	3.68	**11.69**

are generated with a random SNR in between -5 and 20 dB as [23].

TABLE III
SPEECH ENHANCEMENT PERFORMANCE COMPARISON ON THE **DNS CHALLENGE** DATASET WITHOUT REVERBERATIONS. BESIDES, *italic* INDICATES THE PROPOSED METHOD AND **BOLD** INDICATES THE BEST RESULTS.

Method	FLOPs (B)	WP	NP	STOI	SI-SDR
Unprocessed	-	1.56	2.45	91.2	9.03
TSTNN [9]	0.94	2.55	2.61	91.9	10.92
DPRNN [51]	1.37	2.57	2.68	92.5	11.05
TFT-Net [10]	1.22	2.60	2.74	92.7	11.64
DCCRN [53]	2.75	2.64	3.17	92.9	12.21
FullSubNet [23]	2.98	2.72	3.28	95.3	16.17
DPT-FSNet [49]	0.72	**2.72**	**3.28**	**95.3**	**16.17**
U-Transformer+TF	**0.33**	2.65	3.18	92.9	12.60
U-Transformer+FAT	0.34	2.67	3.25	94.1	13.36

It can be observed from Table III that the DPT-FSNet method offers the best speech enhancement performance on the DNS challenge dataset. This is probably because the DPT-FSNet method is designed to not only capture the global (full-band) spectral information and the long-distance cross-band dependencies, but also retain the ability to model and attend the local spectral pattern, which matches well with the DNS challenge in [23]. However, for the results on the IEEE, TIMIT, and VCTK datasets, as shown in Tables I and II, the proposed method outperforms DPT-FSNet. It is noteworthy that DPT-FSNet, FullSubNet and DCCRN are causal speech enhancement methods whose output depends on the present and the previous inputs, while the proposed method is non-causal where the output depends only on the future inputs. Some recent evidence shows that causal inference may perform slightly better than the non-causal inference [57]. If an interference is given, it is possible to measure the causal effect, and enhancing speech performance can be achieved by changing the causal effect. However, the proposed method has been demonstrated to have an advantage in computational efficiency due to its non-causal structure.

E. Ablation study and model parameters

In this experiment, we first show speech enhancement performance of the proposed frequency-band aware attention U-Transformer with different numbers of heads. The models are trained and tested on the IEEE dataset with three SNR levels (-5, 0, 5 dB) and six noise interferences, i.e., $psquare$, $dliving$, $dkitchen$, $nriver$, $tcar$ and $pstation$. Comparisons of speech enhancement performance are showed in Table IV.

TABLE IV
ABLATION STUDY ON DIFFERENT NUMBERS OF HEADS FOR HFA + LFA.

No. of Heads	Para. (M)	STOI	PESQ	fwSNRseg (dB)
2 + 2	2.96	78.4	2.27	12.55
2 + 8	3.58	82.3	2.46	15.78
2 + 16	4.31	85.0	2.74	17.39
8 + 2	3.61	78.5	2.30	12.93
8 + 8	4.21	82.8	2.51	16.06
8 + 16	4.99	85.0	2.76	17.51
16 + 2	4.34	78.9	2.33	13.04
16 + 8	5.06	82.9	2.55	16.37
16 + 16	5.80	85.2	2.77	17.49

We set the number of heads in the proposed HFA and LFA as 2 and 16, respectively. According to Table IV, this offers the best trade-off between performance and model size. On the one hand, compared to the models with more heads in HFA, the proposed method significantly reduces the model size but with only a slight performance degradation. On the other hand, compared to the models with fewer heads in LFA, the proposed method has an enormous improvement on three performance measures due to the focus on the low frequency-band. Therefore, we choose '2+16' setting in the experiments. In addition, the proposed method saves $\mathcal{O}(N(d_{\text{layer}}-1)/d_{\text{layer}})$ factor of resources over standard self-attention. We also calculated the FLOPs in the IEEE dataset experiment. The FLOPs of the original Transformer are 2.3B, while those of the proposed method are only 0.3B.

The above detailed experimental results confirm that the proposed U-Transformer with the frequency-band aware attention can further improve speech enhancement performance both with noise and speech interferences compared to the baselines. With the comparison and ablation experiments, it can be observed that: (1) The proposed U-Transformer based method provides very good improvements. (2) Both T-F and frequency-band aware attentions significantly outperform the conventional attention mechanism. (3) According to the ablation experiments in Tables I-III, the frequency-band aware attention could further improve the speech enhancement performance, as compared with T-F attention. (4) The proposed U-Transformer with the frequency-band aware attention achieves better enhancement performance as compared with other state-of-the-art baselines. The reason is that the proposed attention method splits the feature map in time and frequency directions, and utilizes multi-head attention to mask the attention map over each direction. Different from axial attention, each multi-head attention has a set of learning vectors for its own query, keys, and values to fully use the

positional information. Furthermore, the proposed frequency-band aware method trains three sub-attention with different computational cost and provides an efficient computation. The lower frequency-band where the desired information is intensively distributed is trained with the learning vectors, therefore, speech enhancement performance is further improved. Moreover, the computation cost is reduced because the 2D attention map is factorized into two 1D attentions along time- and frequency-axis. The proposed time-frequency attention saves a $\mathcal{O}(N(d_{\text{layer}}-1)/d_{\text{layer}})$ factor of resources over standard self-attention on each tensor with shape $N = N^{1/d_{\text{layer}}} \times \cdots \times N^{1/d_{\text{layer}}}$.

V. Conclusion

In this paper, we have presented a novel U-Transformer with the frequency-band aware attention for speech enhancement problems. The T-F attention split the feature map obtained from the previous sub-layer to the time and frequency directions and exploited the multi-head attention to mask sub attention maps. Consequently, the 2-D attention map was factorized into two 1-D attentions and allowed parallel computations. Moreover, in order to fully use the information of the desired speech signal, the frequency-band aware attention was proposed to split the full band into two sub-bands and different learning vectors were allocated to TA, HFA, and LFA, respectively. The experimental results confirmed that the proposed U-Transformer outperformed the state-of-the-art models and the frequency-band aware attention could help to achieve further performance improvement. In the future, we will investigate the potential of incorporating the phase information from the complex spectrogram to further improve the performance.

Acknowledgement

We would like to thank the associate editor and the anonymous reviewers for their valuable comments and suggestions in improving this article. For the purpose of open access, the authors have applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising.

References

[1] Y. Luo, C. Han, and N. Mesgarani, "Group communication with context codec for lightweight source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[2] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 125–138, 2019.

[3] S. Leglaive, X. A. Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[4] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359 – 369, 2019.

[5] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *Interspeech*, 2017.

[6] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[7] F. Deng, T. Jiang, X. R. Wang, C. Zhang, and Y. Li, "NAAGN: noise-aware attention-gated network for speech enhancement," *Interspeech*, 2020.

[8] W. W. Yu, J. Zhou, H. B. Wang, and L. Tao, "SETransformer: speech enhancement transformer," *Cognitive Computation*, 2021.

[9] K. Wang, B. B. He, and W.-P. Zhu, "TSTNN: two-stage transformer based neural network for speech enhancement in the time domain," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[10] C. X. Tang, C. Luo, Z. Y. Zhao, W. X. Xie, and W. J. Zeng, "Joint time-frequency and time domain learning for speech enhancement," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[11] J. J. Chen, Q. R. Mao, and D. Liu, "Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation," *Interspeech*, 2020.

[12] G.Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[13] T. Y. Lin, Y. X. Wang, X. Y. Liu, and X. P. Qiu, "A survey of transformers," *arXiv preprint arXiv:2106.04554*, 2021.

[14] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.

[15] K. Tan and D. L. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1785 – 1794, 2021.

[16] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," *Interspeech*, 2020.

[17] Y. Zhao and D. L. Wang, "Noisy-reverberant Speech Enhancement Using DenseUNet with Time-frequency Attention," *Interspeech*, 2020.

[18] Y. Xian, Y. Sun, W. W. Wang, and S. M. Naqvi, "Convolutional fusion network for monaural speech enhancement," *Neural Networks*, vol. 143, pp. 97 – 107, 2021.

[19] S. Kumawat and S. Raman, "Depthwise-STFT based separable convolutional neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[20] Y. Li, Y. Sun, K. Horoshenkov, and S. M. Naqvi, "Domain adaptation and autoencoder based unsupervised speech snhancement," *Submitted to IEEE Transactions on Artificial Intelligence*, 2021.

[21] A. Pandey and D. L. Wang, "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization," *Interspeech*, 2020.

[22] Y. Li, Y. Sun, and S. M. Naqvi, "Single-channel dereverberation and denoising based on lower band trained SA-LSTMs," *IET Signal Processing*, vol. 14, no. 10, pp. 774 – 782, 2021.

[23] X. Hao, X. D. Su, R. Horaud, and X. F. Li, "FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[25] D. L. Wang and J. T. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *The 31st International Conference on Neural Information Processing Systems*, pp. 600 – 610, 2017.

[27] K. M. Jeon, G. W. Lee, N. K. Kim, and H. K. Kim, "TAU-Net: temporal activation U-Net shared with nonnegative matrix factorization for speech enhancement in unseen noise environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3400 – 3414, 2021.

[28] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, 2016.

[29] E. J. Nustede and J. Anemüller, "Towards speech enhancement using a variational U-Net architecture," *arXiv preprint arXiv:2012.03594*, 2021.

[30] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-Net Transformer: self and cross attention for medical image segmentation," *arXiv preprint arXiv:2103.06104*, 2021.

[31] B. J. Borgström and M. Brandstein, "Speech enhancement via attention masking network (SEAMNET): an end-to-end system for joint suppression of noise and reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 515 – 526, 2020.

[32] N. Furnon, R. Serizel, S. Essid, and I. Illina, "Attention-based distributed speech enhancement for unconstrained microphone arrays with varying

number of nodes," *European Signal Processing Conference (EUSIPCO)*, 2021.
[33] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: transformer with gaussian-weighted self-attention for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[34] M. L. Xu, S. Q. Li, and X. L. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," *arXiv preprint arXiv:2010.12155*, 2020.
[35] Z. N. Zhang, B. S. He, and Z. J. Zhang, "TransMask: a compact and fast speech separation model based on transformer," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
[36] K. Ramesh, C. Xing, W. Wang, D. Wang, and X. Chen, "Vset: a multi-modal transformer for visual speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
[37] H. Y. Wang, Y. K. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: stand-alone axial-attention for panoptic segmentation," *European Conference on Computer Vision (ECCV)*, 2020.
[38] J. Abdulbaqi, Y. Gu, S. Chen, and I. Marsic, "Residual recurrent neural network for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[39] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[40] Y. Wang, A. Narayanan, and D. L. 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.
[41] C. Crawl, "Definition: voice frequency," *Wayback Machine*, 2020.
[42] N. A. George, F. F. M. D. Mul, Q. J. Qiu, G. Rakhorst, and H. K. Schutte, "Depth-kymography: high-speed calibrated 3D imaging of human vocal fold vibration dynamics," *Physics in Medicine and Biology*, vol. 53, no. 10, pp. 2667 – 2675, 2008.
[43] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591 – 3591, 2013.
[44] IEEE Audio and Electroacoustics Group, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. AE-17, no. 3, pp. 225–246, 1969.
[45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic phonetic continuous speech corpus [CD-ROM]," *Linguistic Data Consortium*, 1993.
[46] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: design, collection and data analysis of a large regional accent speech database," *IEEE Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.
[47] C. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," *Interspeech*, 2021.
[48] X. Hao, C. H. Shan, Y. Xu, S. N. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
[49] F. Dang, H. T. Chen, and P. Y. Zhang, "DPT-FSNet: dual-path transformer based full-band and sub-band fusion network for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
[50] X. F. Li and R. Horaud, "Online monaural speech enhancement using delayed subband LSTM," *arXiv preprint arXiv:2005.05037*, 2020.
[51] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[52] 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.
[53] Y. X. Hu, Y. Liu, S. B. Lv, M. T. Xing, S. M. Zhang, Y. H. Fu, J. Wu, B. H. Zhang, and L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," *Interspeech*, 2020.
[54] 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, 2008.
[55] Z. X. Liu, H. T. Ma, and F. Chen, "A new data-driven band-weighting function for predicting the intelligibility of noise-suppressed speech," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017.
[56] 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, 2008.
[57] T. Hsieh, C.-H. H. Yang, P.-Y. Chen, S. M. Siniscalchi, and Y. Tsao, "Invfereence and denoise: causal inference-based neural speech enhancement," *arXiv preprint arXiv:2211.01189*, 2022.

Yi Li (Student Member, IEEE) recently defended his Ph.D. thesis within the Intelligent Sensing and Communications (ISC) Research Group, School of Engineering, Newcastle University, UK. Currently, he is a Senior Research Associate at the Security Lancaster, Lancaster University U.K. His is research areas of interest include audio signal processing and adversarial attack detection based on deep learning.

Yang Sun (Member, IEEE) received the M.Sc. and the Ph.D. degrees from Newcastle University U.K., in 2015 and 2019, respectively. Currently, he is a Post-doctoral Researcher of Department of Statistics and Big Data Institute, University of Oxford, UK, focusing on methods development with applications to brain lesion segmentation from MRI scans. His research areas include audio signal processing, speech source separation and medical image processing based on deep learning.

Wenwu Wang (Senior Member, IEEE) received the B.Sc. degree in 1997, the M.E. degree in 2000, and the Ph.D. degree in 2002, all from Harbin Engineering University, China. He is a Professor of Signal Processing and Machine Learning, within the Centre for Vision Speech and Signal Processing, University of Surrey, UK. He is an AI Fellow at the Surrey Institute for People Centred AI. His current research interests include signal processing, machine learning and perception. He authored over 300 papers in these areas. He has been involved in more than 30 research projects, funded by UK and EU research councils and industry (e.g., BBC, NPL, and Samsung). He is the elected Chair of the IEEE Signal Processing Society Technical Committee on Machine Learning for Signal Processing. He is a Senior Area Editor for IEEE Transactions on Signal Processing, and an Associate Editor for IEEE/ACM Transactions on Audio Speech and Language Processing.

Syed Mohsen Naqvi (Senior Member, IEEE) received Ph.D. from Loughborough University UK in 2010. He is a Reader in Multimodal Signal and Information Processing, Director of the Intelligent Sensing Laboratory, and Deputy Head of the Intelligent Sensing and Communications Research Group at Newcastle University, UK. His research contributions have been in human action, activity, behaviour analyses, multiple human target detection, localisation, and tracking, human speech enhancement and separation, and explainable AI, all for defence and healthcare applications. Dr Naqvi has 130+ publications in peer-reviewed articles in high impact journals and proceedings of leading international conferences. He was involved in above 15 research projects, funded by UKRI and Industry (e.g., EPSRC, BBSRC, MoD, Thales, Innovate UK, NHS). He also successfully graduated above 20 PhDs including the first two authors of this paper. He is an Associate Editor (AE) for IEEE/ACM Transactions on Audio Speech and Language Processing and an AE for IEEE Transactions on Signal Processing.