Convolutional Transformer based Local and Global Feature Learning for Speech Enhancement

Chaitanya Jannu[1], Sunny Dayal Vanambathina[2]
School of Electronics Engineering, VIT-AP University, Amaravati, India[1, 2]

***Abstract*—Speech enhancement (SE) is an important method for improving speech quality and intelligibility in noisy environments where received speech is severely distorted by noise. An efficient speech enhancement system relies on accurately modelling the long-term dependencies of noisy speech. Deep learning has greatly benefited by the use of transformers where long-term dependencies can be modelled more efficiently with multi-head attention (MHA) by using sequence similarity. Transformers frequently outperform recurrent neural network (RNN) and convolutional neural network (CNN) models in many tasks while utilizing parallel processing. In this paper we proposed a two-stage convolutional transformer for speech enhancement in time domain. The transformer considers global information as well as parallel computing, resulting in a reduction of long-term noise. In the proposed work unlike two-stage transformer neural network (TSTNN) different transformer structures for intra and inter transformers are used for extracting the local as well as global features of noisy speech. Moreover, a CNN module is added to the transformer so that short-term noise can be reduced more effectively, based on the ability of CNN to extract local information. The experimental findings demonstrate that the proposed model outperformed the other existing models in terms of STOI (short-time objective intelligibility), and PESQ (perceptual evaluation of the speech quality).**

Keywords*—Convolutional neural network; recurrent neural network; speech enhancement; multi-head attention; two-stage convolutional transformer; feed-forward network

I. Introduction

In the area of speech processing, speech enhancement is crucial. The main objective is to enhance speech that has been impaired by background noise in terms of clarity and quality. Numerous applications including powerful speech recognition, teleconferencing, and hearing aids, employ it as a pre-processor. Recent advances in deep learning have made it possible to develop several data driven methods to solve traditional estimation problems without having to rely on supervision. The majority of existing deep learning models for speech enhancement, such as convolutional neural network (CNN) and recurrent neural network (RNN), are implemented in the time-frequency (T-F) domain. Those methods use short-time Fourier transforms (STFT) to train on spectral magnitude. To reconstruct the time-domain signal using the inverse short-time Fourier transform (iSTFT), it is necessary to take into account the phase of noisy speech along with the improved speech magnitude. Despite some impressive results [1-4], T-F domain methods still have two key drawbacks. In the first place, Fourier transforms add an additional overhead to fast speech denoising. Second, during the denoising process, the noisy phase is generally ignored. Nevertheless, phase information has been shown to be significant for improving speech quality [5]. To achieve better enhancement results, there are some studies that look at magnitude as well as phase simultaneously during training [6]. Several recent works have directly estimated the clean speech in time domain from noisy raw data [7-11].

II. Related Work

A fully convolutional network (FCN) has been proposed by Fu et al. [12] for raw waveform-based speech enhancement which improves the quality compared to masking-based methods [13-14]. In distinction to both masking and mapping techniques that use noisy phase to reconstruct enhanced speech, FCN provides speech enhancement by simply mapping from a noisy speech to the matching clean speech. Many studies have looked into CNN or RNN-based encoder-decoder frameworks. CNN requires more convolutional layers to enlarge the receptive field when modelling long-range sequences such as speech. The Convolutional recurrent networks (CRNs) [15] are used to extract long-context information using CNN's which are familiar for feature extraction and RNNs' which are familiar for temporal modelling. For processing long-term temporal sequences, a dilated convolutional neural network has been proposed [16]. K. Tan et al. [17] proposed GRN with Dilated convolutions for supervised SE. The convolutions with dilation enlarge the receptive field without sacrificing resolution compared to regular convolutions by maintaining the same kernel size and network depth. At various SNR levels the GRN exhibits higher generalization ability to untrained speakers when compared to the LSTM model in [18]. A CNN model in time domain is presented by considering frequency domain loss to enhance the quality of corrupted speech [19]. Although the work in [19] can achieve cutting-edge performance, the issue of real-time enhancement is not addressed. Later the same authors proposed densely connected network [20] for real time SE and achieved better performance than [19].

The RNNs such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are usually used to model long-term sequences with order information. Although introducing temporal convolutional network (TCN) blocks [10] or LSTM layers between encoder and decoder can improve denoising performance by extracting higher-level features and expanding receptive fields [11], the contextual information of speech is generally ignored, limiting denoising performance. Recently Li, Qinglong, et al. [21] proposed a DCTCRN model for speech enhancement where discrete cosine transform is used as input to the model for processing noisy speech and uses

LSTM as bottleneck. Even though the performance of the model is better, the LSTMs easily prone to the problem of overfitting and it also requires large time to train and are also sensitive to random weight initializations. Kai Zhen et al. proposed a DCCRN [22] network for SE in time domain where densenets and GRUs are used for context aggregation. This model yields better intelligibility scores compared to convolutional base line models but quality is less compared to Wave-U-Net model [7]. Recently Lin et al. proposed a SASE model [23] with self-attention for speech enhancement which extracts speaker information using LSTM layers and also CRN model is used for speech enhancement based on embedding of speaker information which performs better than CRNs and LSTM models but the quality is not satisfactory. A unique ARN for time-domain speech improvement was recently proposed by Pandey A., et al. [24] in order to promote cross-corpus generalization. Self-attention and feedforward blocks are added to RNN to create ARN. It has been shown that transformer neural networks are capable of resolving the long-dependency problem effectively and are capable of operating in parallel, so they are good at tackling a wide range of natural language processing tasks [25]. As the transformer only uses the attention mechanism, the vanishing gradient and exploding gradient problem of RNNs are also solved. A transformer model for speech enhancement was proposed in [26], which has a comparatively large size of model. Transformer enhances the training speed and prediction accuracy compared to RNNs [25]. By combining the advantages of LSTM and multi-head attention mechanism, Yu et al. [27] proposed SETransformer. When compared with a standard transformer and an LSTM model, SETransformer showed better denoising performance. Recently Wang et al. [28] introduced an end-to-end speech enhancement two-stage transformer neural network (TSTNN).

The limitations of existing frameworks are the CNN models [12,19] requires more convolutional layers to enlarge the receptive field when modelling long-range sequences such as speech. The disadvantage of RNN-based models [18] is that they cannot perform parallel processing, resulting in high computation complexity. The limitation of TCN [10] is that the contextual information of speech is generally ignored, limiting denoising performance. The CRN model [21] suffers with computational complexity. The transformer models present in [26], [27] performs better than convolutional and RNN baselines but they are implemented in the time-frequency (T-F) domain. To achieve better enhancement results, there are some studies that look at magnitude as well as phase simultaneously during training. Recently Wang et al. [28] introduced an end-to-end speech enhancement two-stage transformer neural network (TSTNN) which performs better than [26] and [27]. But in TSTNN model same structure of FFN is used for extracting local and global features. And also, they have not concentrated on the local (short-term) noises present at the output of encoder.

Motivated by recent success of transformer neural networks in natural language processing tasks [25] and SE [26-28] we propose a novel two-stage convolutional transformer neural network that enhances monaural speech in a time domain. The transformers are capable of resolving the long-range dependency problem effectively and are capable of operating in parallel. As the transformer only uses the attention mechanism, the vanishing gradient and exploding gradient problem of RNNs are also solved. In this work, we propose a two-stage convolutional transformer neural network that enhances monaural speech in a time domain from end to end which differs from the existing transformer models such as T-gsa [26], SETransformer [27] and two-stage transformer neural network (TSTNN) [28]. Based on the transformer's ability to model sequences and the dual-path network's ability to extract contextual information [29], we propose a two-stage convolutional transformer neural network.

In this paper the proposed transformer considers global information as well as parallel computing, resulting in a reduction of long-term noise. Unlike the transformer proposed in [28], we proposed a novel transformer structure for intra and inter transformers to extract the local as well as global features of noisy speech using CNN and RNN layers [30]. A 1D-Conv layer and temporal convolution module (TCM) are used in intra transformer where local features are extracted and a 1D-Conv layer and Bi-directional long short-term memory (Bi-LSTM) are used in inter transformer where global features are extracted.

Moreover, in the proposed model, a CNN module is also added to intra and inter transformers so that short-term noise can be reduced more effectively, based on the ability of CNN to extract local information. The proposed model enhances the local information modelling capability of the traditional transformer model by adding a convolution layer. The convolutional module consists of Depth and Point wise convolutions instead of normal convolution to increase the speed of operation of model and both of them have less parameters compared to normal convolution. The details of convolutional module and modified layers of feed forward network are explained in the following sections.

Our contributions:

- In contrast to RNNs and CNNs, the proposed transformer model uses parallel processing and the long-term dependencies can be modelled more efficiently with multi-head attention (MHA) by using sequence similarity.
- The novelty of the proposed work is different transformer structures for intra and inter transformers are used for extracting the local as well as global features of noisy speech using CNN and RNN layers. In the proposed model a CNN module is also added to intra and inter transformers so that short-term noise can be reduced more effectively, based on the ability of CNN to extract local information.
- The proposed model enhances the local information modelling capability of the traditional transformer model by adding a CNN module.

The remainder of this work is structured as follows. The presents the related works. Section II presents related works and we explained the details of proposed two-stage convolutional transformer in Section III. Section IV presents

the experimental findings. Section V presents discussion and the paper is concluded in Section VI.

III. Architecture of Proposed Two-Stage Convolutional Transformer

The overall architecture of proposed model is shown in Fig. 1. The model consists of four modules: Encoder, Transformer module, Masking module and Decoder. There are two convolution layers in the encoder among them the first one is to increase the channels to 64 and the second one is used to halve the size of the frame. The transformer module internally consists of four transformer blocks responsible for feature extraction. The detailed transformer module is explained in following Section 3.C. The masking module is used to obtain the mask using 2-way convolution and nonlinear activations. In the decoding phase the reconstruction of features is done by using dilated dense block and sub-pixel convolution. Finally, we will get the enhanced speech after normalization and PReLU.

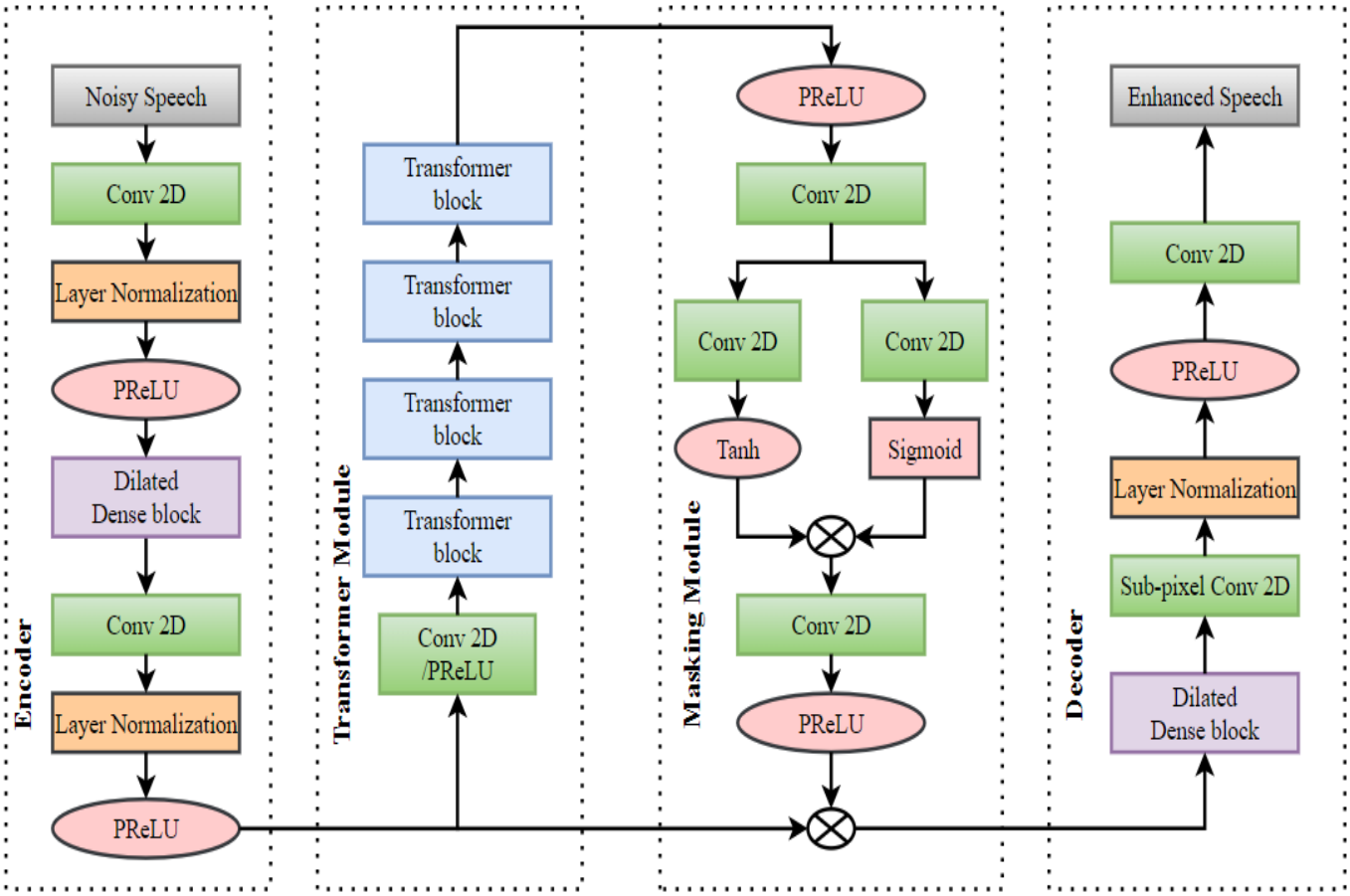


Fig. 1. Architecture of proposed convolutional transformer

A. Segmentation and Overlap-Add

In this stage the input noisy speech $Y \in R^{1\,X\,L}$ is splitted into frames with length of each frame as F and hop size as H. The 3D tensor $I \in R^{1\,X\,N\,X\,F}$ is then created by stacking all frames as input to the encoder. The length of input noisy speech is denoted by L and total frame count is denoted by N.

$$N = [(L - F)/(F - H) + 1] \qquad (1)$$

To recover the enhanced speech waveform the overlap-add operation is used at decoder.

B. Encoder

There are two convolution layers in the encoder among them the first one is to increase the channels to 64 with filter size of (1,1) and the second one is used to halve the frame size with filter and stride of (1,3) and (1,2). And a 4-layer dilated dense block [20] is inserted in between them. The Layer normalization and PReLU nonlinearity [30-31] are applied to all convolutional layers.

C. Transformer Module

The transformer module internally consists of four transformer blocks stacked together where feature extraction is performed to extract local as well as global features. Before giving the output of encoder to input of transformer, we use convolution with a kernel of size (1, 1) followed by PReLU activation to halve the channel dimension to reduce the computational complexity of the transformer network. The transformer block is shown in below Fig. 2.

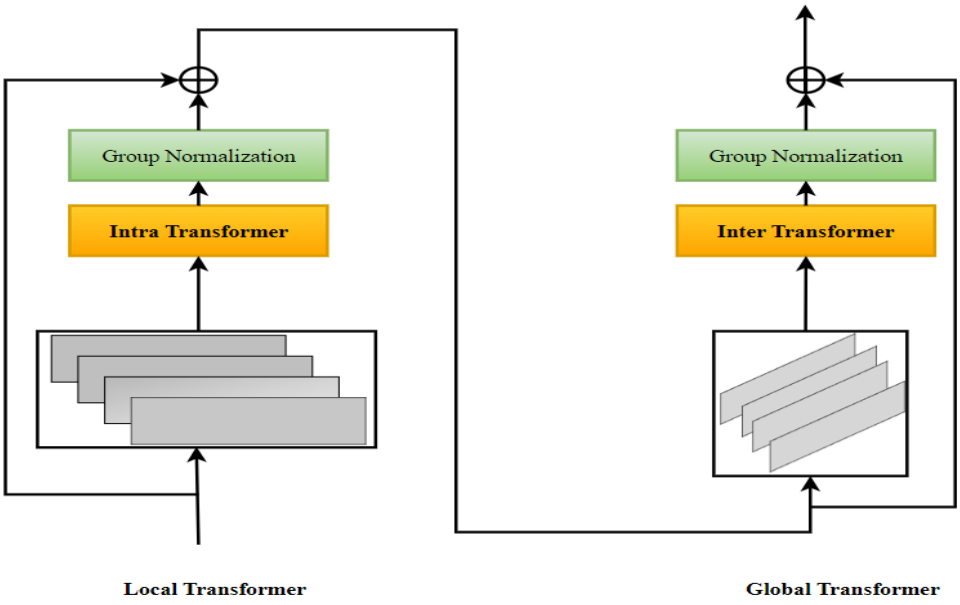


Fig. 2. Transformer block

1) Improved transformer: The transformer block consists of intra and inter transformers as shown in above Fig. 2. To independently model the local chunks intra transformer is used and inter transformer is used to extract global dependencies by summarizing information of all blocks.

Three important modules are included in the original transformer encoder [25]: positional encoding, multi-head attention, and position-wise feed-forward. Due to the fact that the positional encoding is not appropriate for acoustic sequences the authors of baseline model [28] designed GRUs to learn positional information by replacing the first fully connected layer of feed-forward networks with RNNs to track order information [32-33]. The structure of feed forward network used in baseline model [28] is shown in Fig. 3.

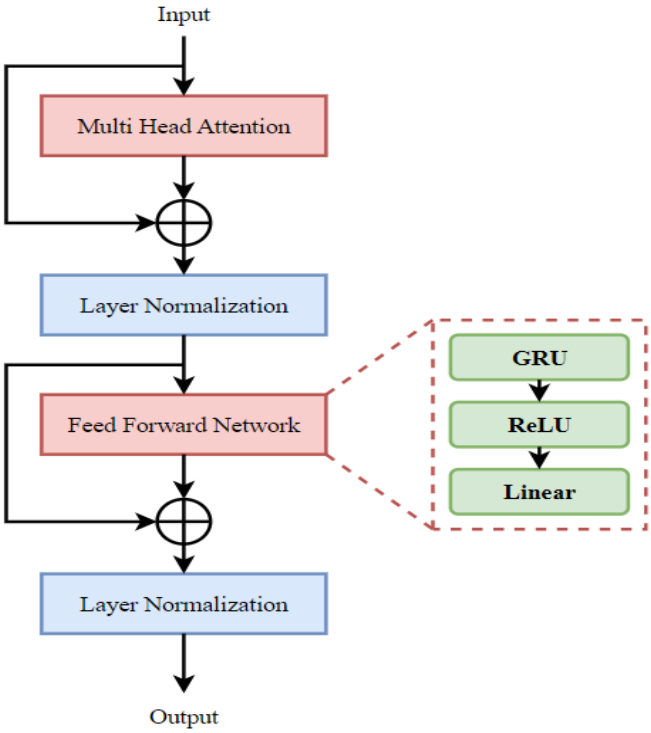
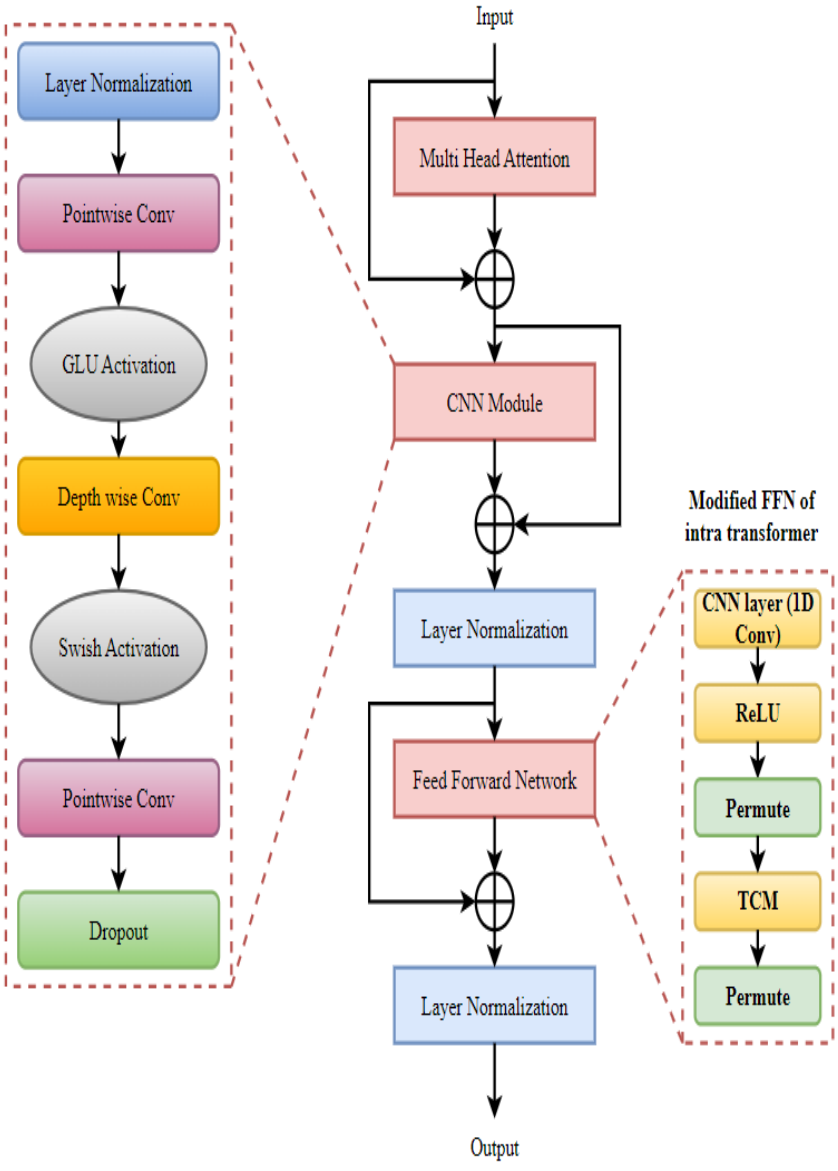


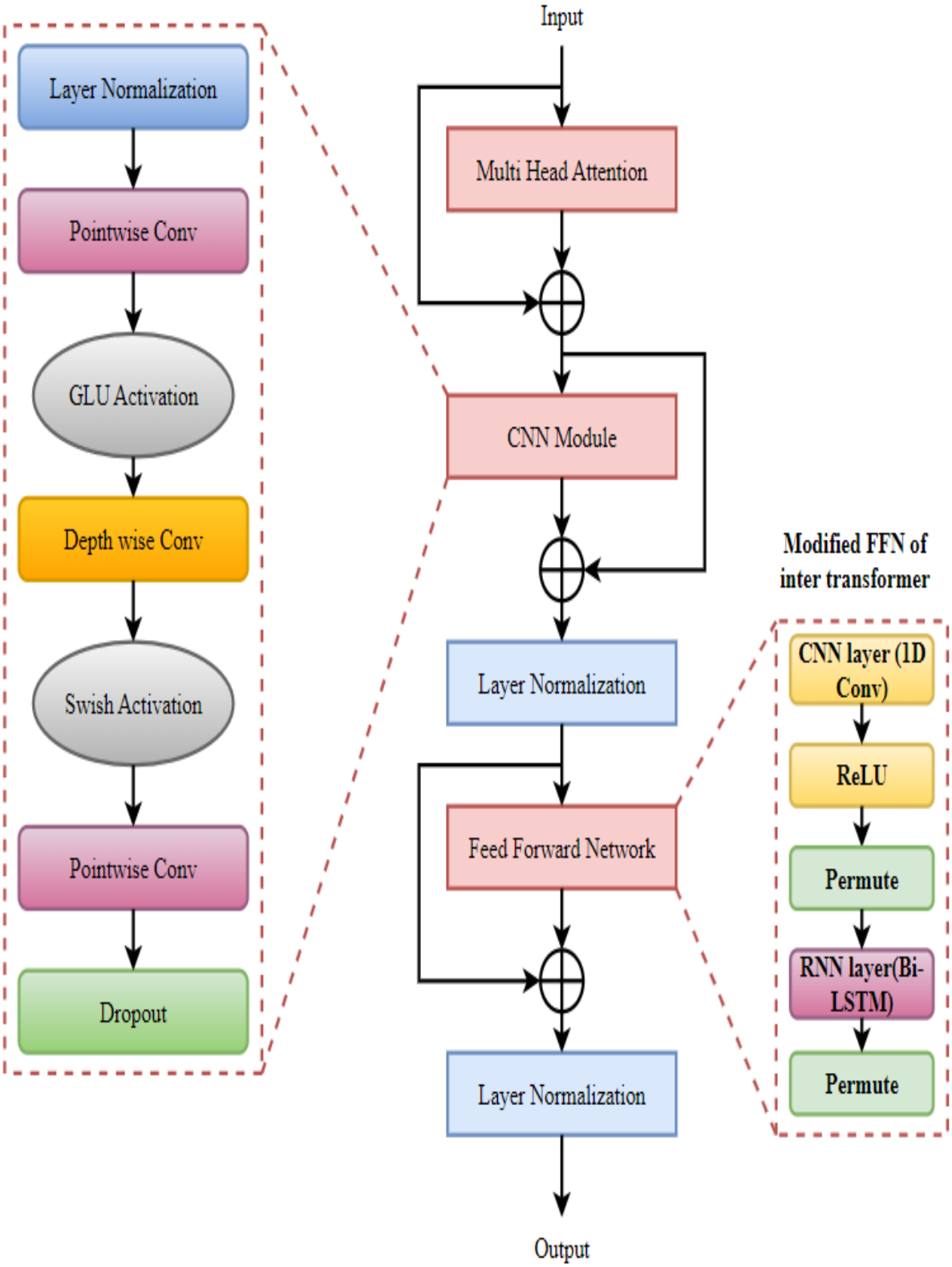
Fig. 3. The structure of feed forward network in baseline model

In the baseline model same structure of feed forward network is used in both intra and inter transformers of transformer block. In the proposed method unlike baseline, we introduced two different feed forward network layer structures in the intra and inter transformers of transformer block to well model the local and global features. In addition to this we have also added a convolutional module to the baseline intra and inter transformers of transformer block to further filter the local (short-term) noises present at the output of encoder. Due to the powerful feature self-learning ability of the CNN model, it will increase the local feature extraction ability of the transformer model. The proposed model enhances the local information modelling ability of the traditional transformer model by adding a convolution layer. The structure of proposed intra and inter transformers along with convolutional module are shown in below Fig. 4(a) and 4(b). The convolutional module consists of Depth and Point wise convolutions instead of normal convolution to increase the speed of operation of model and both of them have less parameters compared to normal convolution.

2) Proposed intra and inter transformer: The key distinction between intra and inter transformer is in its structure of Feed Forward Network (FFN). A CNN layer (1D-Conv) and temporal convolution module (TCM) are used in intra transformer to learn long-term dependency of speech from the encoder output. The layer details of TCM are shown in below Fig. 5. A CNN layer (1D-Conv) and Bi-directional long short-term memory (Bi-LSTM) [34] are used in inter transformer to extract long-distance global features.



(a)



(b)

Fig. 4. The structure of inter transformer in transformer block

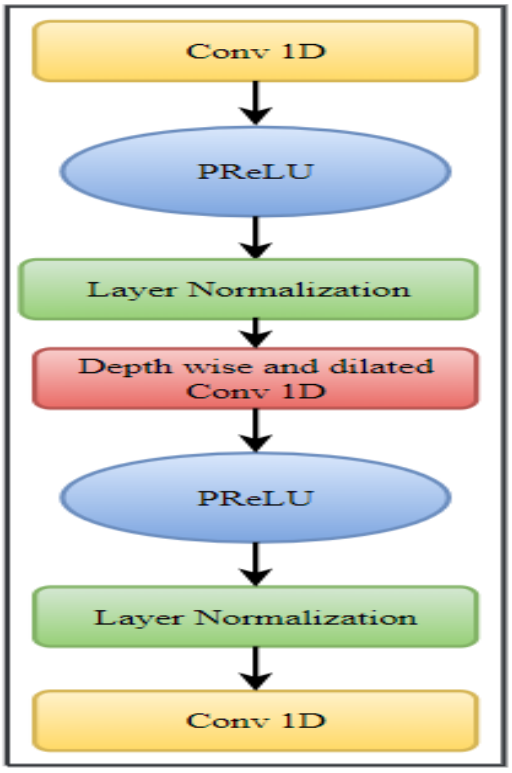


Fig. 5. The Layers of TCM

In general, Multi-Head Attention (MHA) works as follows:

In the MHA block, the input (Y) is first mapped *h*-times with different, learnable linear transformations to obtain queries, keys, and values representations, respectively.

$$Q_i = YW_i^Q \ , K_i = YW_i^K \ , V_i = YW_i^V \quad (2)$$

In the above Eq. 1 $Q_i, K_i, V_i \in R^{l\ X\ d/h}$ are mapped query, key and value. $W_i^Q, W_i^K, W_i^V \in R^{d\ X\ d/h}$ represents i^{th} linear projection matrix for query, key and value.

The query will be dot produced with all keys and a constant is divided by the dot product. After that a softmax is applied to the values to obtain weights. According to Eq.2, each head's attention is a dot product of its weight and value.

$$Head_i = Attention(Q_i, K_i, V_i) \ = softmax\left(\frac{Q_i {K_i}^T}{\sqrt{d}}\right) V_i \ , \ i = 1,2, \dots h \quad (3)$$

The number of parallel attention heads are represented by h. In our proposed work h is considered as 4. The final output is obtained by concatenating the attentions of all heads and linearly projecting them again.

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{Head}_1, \dots \text{Head}_h)W^O \quad (4)$$

In the above Eq. 3 $W_i^O \in R^{d\ X\ d}$ is linear projection matrix.

Next, it follows residual connection as shown in Eq. (4).

$$out = Y + \text{MultiHead}(Q, K, V) \quad (5)$$

In the above Eq. 4 $Y \in R^{l\ X\ d}$ represents input. Where l is sequence length and d is dimension.

To get the final output of intra and inter transformers of transformer block, the output of multi-head attention block is then processed by convolutional module and corresponding feed-forward networks, as well as residual connections and layer normalization [18].

The output of convolutional module is given as

$$Conv\ Module\ (out) = Dropout(PointwiseConv(Swish(DepthWiseConv\ (GLU(PointwiseConv\ (Layernormalization(out))))))) \quad (6)$$

$$CM_{out} = Layernormalization(out + (Conv\ Module\ (out))) \quad (7)$$

The final output of intra transformer is given as

$$FFN_{intra}(CM_{out}) = \text{TCN}(\text{ReLU}(\text{Conv1D}(CM_{out}))) \quad (8)$$

$$Finaloutput_{intra} = Layernormalization(CM_{out} + FFN_{intra}(CM_{out})) \quad (9)$$

The final output of inter transformer is given as:

$$FFN_{inter}(CM_{out}) = \text{Bi} - \text{LSTM}(\text{ReLU}(\text{Conv1D}(CM_{out}))) \quad (10)$$

$$Finaloutput_{inter} = Layernormalization(CM_{out} + FFN_{inter}(CM_{out})) \quad (11)$$

3) Proposed two-stage transformer block: In the proposed 2-stage transformer block there are two transformers called intra and inter transformers as shown in Fig. 2, which are used for extracting local and global context information respectively. The local transformer with the input as a 3-D tensor with dimensions as [C, N, F] is first applied to each chunk and local information is processed in parallel on the last dimension F of the input tensor. For learning global dependency, the global transformer uses the information of output from the local transformer which is implemented on tensor dimension N. Moreover, each transformer undergoes a group normalization and residual connections are also used.

D. Masking Module

A masking network obtains denoising masks by utilizing transformer module's output features. First to match the output of the encoder, transformer module output is doubled along the channel dimension using PReLU and convolution. Afterwards, it undergoes two-way 2D convolution and nonlinearity, with the outputs multiplying together to form the input of two-dimensional convolution and PReLU. In order to obtain the final masked encoder feature, the mask and the encoder's output are multiplied element-wise.

E. Decoder

At this stage decoding the encoder feature into enhanced speech features is accomplished via dilated dense blocks and sub-pixel convolutions [35]. Using 2D Conv with filter size (1, 1), the enhanced speech feature's channel dimension is recovered into one and the enhanced waveform is produced by overlap-add.

F. Loss Function

The loss in the T-F domain can direct the model to acquire extra information, resulting in improved intelligibility as well as quality [19]. To train the model, we combine two losses. First, a waveform is created using the overlap-and-add approach utilizing the improved frames. Using the mean squared error among the enhanced and clean utterances, an utterance level loss is computed in the temporal domain.

It is said that the time-domain loss is:

$$L_t(y, \hat{y}) = \frac{1}{M}\sum_{n=0}^{M-1}(y_i[n] - \hat{y_i}[n])2 \quad (12)$$

where M is the utterance length, y[n] is the nth sample of the clean utterance and $\hat{y}[n]$ is nth sample of enhanced utterance.

The loss in frequency-domain is given as:

$$L_f(y, \hat{y}) = \frac{1}{T \cdot F}\sum_{t=1}^{T}\sum_{f=1}^{F}|(|Y(t,f)_r| + |Y(t,f)_i|) - (|\hat{Y}(t,f)_r| + |\hat{Y}(t,f)_i|)| \quad (13)$$

where $Y(t,f)$ and $\hat{Y}(t,f)$ denotes T-F units of STFT's of Y and $\hat{Y}$ respectively. Where T denotes total frames number and F denotes frequency bins count. The real and imaginary components of a complex variable Y is denoted by $Y_r$and Y_i, respectively.

Finally, the T-F domain lose is given by

$$L(y, \hat{y}) = \alpha * L_t(y, \hat{y}) + (1 - \alpha) * L_f(y, \hat{y}) \quad (14)$$

The hyper-parameter α is set to 0.2.

IV. Experiments

A. Datasets

We conducted the experiments using a 2 data sets one is public dataset published by Valentini et al. in [36]. This database contains 30 utterers from the Voice Bank corpus [37], among them 28 utterers are utilized to train the model and 2 are utilized for testing. There are 11,572 pairs of clean-noisy speeches in the training set. The noisy environments comprise 10 types of noises. Among them 8 noises are from the DEMAND dataset [38] and 2 are synthetic sounds at SNRs of 0, 5, 10 and 15 dB. The test set contains 824 mixtures using 5 noises from [38] which are not present in training set at SNRs of 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB.

Additionally, we assess the performance of our model in unseen environments. To test the generalization capability of proposed model we considered large Librispeech dataset [39]. To train the model we have chosen clean speech with duration of 50h from train-clean-100. A randomly selected set of noises from DEMAND dataset [38] and music samples from MUSAN [40] was used in the training set, with SNR ranges from −10dB to 10 dB. For each of the six DEMAND categories, we used two out of three types of noise. The test set of 500 samples are randomly taken from test-clean and the babble noise from NOISEX-92 dataset [41] and the noises river and restaurant are taken from DEMAND which are different from training set. The SNR are −5dB, 0dB, 5dB, 10dB, and 15dB. The sampling rate of all utterances is 16KHz.

B. Experimental Setup

Each speech is resampled to a 16 kHz frequency. A rectangle window with a 32 ms (512 samples) size and overlap length of 16 ms is used to extract the frames. If an utterance lasts more than 4 seconds, we chunk a random 4 second chunk from it during each training session. To make the smaller utterances in the batch the same size as the largest utterance, zero padding is used. we train the model over the course of 100 epochs and optimization employs the Adam optimizer. Gradient clipping is used with a maximum L2-norm of 5 to prevent gradient explosion. During the training phase, dynamic strategies are used to determine learning rate [13]. In this experiment, learning rates are linearly increased during training and then reduced by 0.98 every two epochs.

C. Evaluation Metrics

In this study, speech quality is measured by using STOI [42], whose values normally fall between 0 and 1, and PESQ [43], whose ranges are -0.5 to 4.5. Speech quality improves with a higher PESQ value. Speech intelligibility increases with increased STOI.

D. Results and Analysis

1) Comparison with existing works on Valentini et al. dataset [36]: The experiment was carried out with two different datasets. The proposed convolutional transformer model was compared with various existing models such as SEGAN [8], CGAN [44], Wave-U-Net [7], MMSE-GAN [2], Metric GAN [3], DCUnet-16 [6], DEMUCS-small [11], SE-Transformer [27], TSTNN [28]. All the existing models used for comparison are trained with similar dataset used to train our model. All the models are reproduced for comparison. As a comparison to the number of existing models stated in the original papers, we calculated the metric scores of our model. The proposed model was assessed by using the STOI and PESQ. The PESQ and STOI values for Valentini et al. dataset [36] are given in Table I.

TABLE I. ASSESSMENT OF PESQ AND STOI OF PROPOSED MODEL WITH PREVIOUSLY PUBLISHED BASELINE SCORES USING THE VALENTINI ET AL. DATASET

MODEL	PESQ	STOI (%)
Noisy	1.97	91
SEGAN [8]-2017	2.16	93
CGAN [44]-2018	2.34	93
Wave-U-Net [7]-2018	2.40	-
MMSE-GAN [2]-2018	2.53	93
Metric GAN [3]-2019	2.86	-
DCUnet-16 [6]-2019	2.93	-
DEMUCS-small [11]-2020	2.93	95
SE-Transformer [27]-2022	2.62	93
TSTNN [28]-2021	2.96	95
Convolutional TSTNN (proposed)	3.12	96

The above Table I gives the PESQ and STOI scores for proposed and various existing models such as SEGAN [8], CGAN [44], Wave-U-Net [7], MMSE-GAN [2], Metric GAN [3], DCUnet-16 [6], DEMUCS-small [11], SE-Transformer [27], and TSTNN [28].

The PESQ score for SEGAN [8] is 2.16 and for the proposed is 3.12. The STOI score for SEGAN [8] is 93% and for the proposed is 96%. The SEGAN is an end-to-end SE model where only strided convolutions are used in the generator and discriminator. The PESQ score for CGAN [44] is 2.34 and for the proposed is 3.12. The STOI score for CGAN [44] is 93% and for the proposed is 96%. The CGAN is a CNN based GAN operates in T-F domain. In both the GAN models only, ordinary convolutional layers are used. The PESQ score for DEMUCS-small [11] is 2.93 and for the proposed is 3.12. The STOI score for DEMUCS-small [11] is 95% and for the proposed is 96%. The DEMUCS model is a time domain U-Net model with ordinary convolutions and LSTM bottle neck. The CNN alone cannot well model the long-range dependencies of speech signal. In the proposed model to further enhance the performance of the model in addition to ordinary convolution a dilated dense block is used in both encoder and decoder. At various resolutions, the dilated convolutions aid with context aggregation and the dense connectivity provides feature map with more precise target information by passing through multiple layers. A transformer block is used for extracting the local and global features from a noisy speech signal. The PESQ score for SE-Transformer [27] is 2.62 and for the proposed is 3.12. The STOI score for SE-Transformer [27] is 93% and for the proposed is 96%. The SE-transformer uses only LSTMs, multi-head attention and 1D convolution for SE. The PESQ score for TSTNN [28] is 2.96 and for the proposed is 3.12. The STOI score for TSTNN [28] is 95% and for the proposed is 96%.

2) Comparison with existing works on librispeech dataset [39]: To test the generalization ability of our proposed model we conducted experiments with large Librispeech dataset. The average PESQ and STOI scores for three types of noises at SNRs of −5dB, 0dB, 5dB, 10dB, and 15dB are presented in Table II. The proposed convolutional transformer model was compared with DNS-baseline [45], DEMUCS [11] and TSTNN [28].

TABLE II. EVALUATION OF PESQ AND STOI OF OUR MODEL AND EXISTING MODELS IN THE EXISTENCE OF BABBLE NOISE ON LIBRISPEECH DATASET

Model	Noisy		DNS [45]		DEMUCS [11]		TSTNN [28]		Proposed	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-5dB	1.11	0.66	1.23	0.70	1.50	0.82	1.58	0.83	1.69	0.84
0dB	1.19	0.77	1.48	0.82	1.92	0.90	2.10	0.91	2.32	0.92
5dB	1.40	0.86	1.86	0.90	2.36	0.94	2.49	0.94	2.68	0.94
10dB	1.75	0.92	2.29	0.94	2.76	0.96	2.88	0.97	3.04	0.96
15dB	2.25	0.96	2.71	0.96	3.07	0.98	3.19	0.98	3.58	0.98

From Table II, in the existence of babble noise at -5dB input SNR, the PESQ value for the proposed method is 1.69 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 1.23, 1.50 and 1.58. The STOI score of the proposed method is 0.84 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.70, 0.82 and 0.83.

From Table II, in the existence of babble noise at 5dB input SNR, the PESQ value for the proposed method is 2.68 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 1.86, 2.36 and 2.49. The STOI score of the proposed method is 0.94 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.90, 0.94 and 0.94.

From Table II, in the existence of babble noise at 15dB input SNR, the PESQ value for the proposed method is 3.58 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.71, 3.07 and 3.19. The STOI score of the proposed method is 0.98 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.96, 0.98 and 0.98.

From Table II in the existence of babble noise at all input SNR conditions it is noted the proposed model performs better than that of DNS baseline [45], DEMUCS [11] and TSTNN [28] models. The proposed convolutional transformer delivers improved performance because the local and global features are well modelled by CNN and RNN layers. Moreover, a CNN module is added to deal with short-term noise.

TABLE III. EVALUATION OF PESQ AND STOI OF PROPOSED MODEL AND EXISTING MODELS IN THE EXISTENCE OF RIVER NOISE ON LIBRISPEECH DATASET

Model	Noisy		DNS [45]		DEMUCS [11]		TSTNN [28]		Proposed	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-5dB	1.23	0.80	1.81	0.86	2.05	0.90	2.12	0.91	2.28	0.92
0dB	1.45	0.87	2.17	0.91	2.37	0.94	2.46	0.94	2.67	0.95
5dB	1.78	0.92	2.55	0.95	2.66	0.96	2.84	0.96	2.96	0.96
10dB	2.23	0.96	2.92	0.96	3.98	0.97	3.12	0.97	3.42	0.97
15dB	2.78	0.98	3.23	0.98	3.33	0.98	3.41	0.98	3.64	0.98

From Table III, in the existence of river noise at 0dB input SNR, the PESQ value for the proposed method is 2.67 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.17, 2.37 and 2.46. The STOI score of the proposed method is 0.95 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.91, 0.94 and 0.94.

From Table III, in the existence of river noise at 5dB input SNR, the PESQ value for the proposed method is 2.96 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.55, 2.66 and 2.84. The STOI score of the proposed method is 0.96 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.95, 0.96 and 0.96.

From Table III, in the existence of river noise at 10dB input SNR, the PESQ value for the proposed method is 3.42 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.92, 3.98 and 3.12. The STOI score of the proposed method is 0.97 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.96, 0.96 and 0.97.

From Table III in the existence of river noise at all input SNR conditions it is noted the proposed model performs better than that of DNS baseline [45], DEMUCS [11] and TSTNN [28] models. The proposed convolutional transformer delivers improved performance because the local and global features are well modelled by CNN and RNN layers. Moreover, a CNN module is added to deal with short-term noise.

TABLE IV. EVALUATION OF PESQ AND STOI OF PROPOSED MODEL AND EXISTING MODELS IN THE EXISTENCE OF RESTAURANT NOISE ON LIBRISPEECH DATASET

Model	Noisy		DNS [45]		DEMUCS [11]		TSTNN [28]		Proposed	
SNR	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-5dB	1.09	0.63	1.21	0.68	1.38	0.78	1.49	0.79	1.62	0.81
0dB	1.15	0.75	1.43	0.80	1.68	0.88	1.85	0.91	1.98	0.92
5dB	1.31	0.84	1.76	0.88	1.99	0.93	2.29	0.94	2.41	0.94
10dB	1.62	0.91	2.16	0.93	2.33	0.95	2.67	0.95	2.87	0.96
15dB	2.07	0.95	2.59	0.96	2.70	0.97	2.94	0.97	3.18	0.97

From Table IV, in the existence of restaurant noise at 0dB input SNR, the PESQ value for the proposed method is 1.98 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 1.43, 1.68 and 1.85. The STOI score of the proposed method is 0.92 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.80, 0.88 and 0.91.

From Table IV, in the existence of restaurant noise at 10dB input SNR, the PESQ value for the proposed method is 2.67 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.16, 2.33 and 2.67. The STOI score of the proposed method is 0.96 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.93, 0.95 and 0.95.

From Table IV, in the existence of restaurant noise at 15dB input SNR, the PESQ value for the proposed method is 3.18 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 2.59, 2.70 and 2.94. The STOI score of the proposed method is 0.97 and for existing DNS baseline [45], DEMUCS [11] and TSTNN [28] are 0.96, 0.97 and 0.97.

From Table IV in the existence of restaurant noise at all input SNR conditions it is noted the proposed model performs better than that of DNS baseline [45], DEMUCS [11] and TSTNN [28] models. The proposed convolutional transformer delivers improved performance because the local and global features are well modelled by CNN and RNN layers. Moreover, a CNN module is added to deal with short-term noise.

From Table II, Table III, and Table IV, it is noted the transformer-based models performs better than that of DNS baseline [45] and DEMUCS [11]. The proposed convolutional transformer delivers improved performance in terms of enhanced signal quality than all existing methods for all types of noises at all SNR conditions.

V. DISCUSSION

The results from experimental studies show that the proposed convolutional transformer consistently improves speech quality as well as intelligibility. The performance of proposed transformer is better than the existing models such as SEGAN, CGAN, Wave-U-Net, MMSE-GAN, Metric GAN, DCUnet-16 and TSTNN. In the existing models CNN alone cannot well model the long-range dependencies of speech signal. In the proposed model to further enhance the performance of the model in addition to ordinary convolution a dilated dense block is used in both encoder and decoder. At various resolutions, the dilated convolutions aid with context aggregation and the dense connectivity provides feature map with more precise target information by passing through multiple layers. A transformer block is used for extracting the local and global features from a noisy speech signal. In TSTNN model same structure of FFN is used in both intra and inter transformers. And also, they have not concentrated on the local (short-term) noises present at the output of encoder. To further enhance the performance in the proposed model we used TCM in the intra transformer and Bi-LSTM in the inter transformer as CNN's are good at extracting local features and the RNN's are good at extracting global features from the noisy speech signal. In the proposed model the transformer considers global information as well as parallel computing, resulting in a reduction of long-term noise. In contrast to RNNs and CNNs, the proposed transformer model uses parallel processing and the long-term dependencies can be modelled more efficiently with multi-head attention (MHA) by using sequence similarity. Moreover, a CNN module is added to the transformer so that short-term noise can be reduced more effectively, based on the ability of CNN to extract local information. We observe that the performance drop is more without CNN module that means on the basis of the transformer model CNN module is able to capture locally relevant context information based on global contextual information.

VI. CONCLUSION

In this paper we proposed a two-stage convolutional transformer for speech enhancement in time domain. The transformer considers global information as well as parallel computing, resulting in a reduction of long-term noise. In the proposed work unlike two-stage transformer neural network (TSTNN) different transformer structures for intra and inter transformers are used for extracting the local as well as global features of noisy speech. Moreover, a CNN module is added to the transformer so that short-term noise can be reduced more effectively, based on the ability of CNN to extract local information. The experimental findings demonstrate that the proposed model outperformed the other existing models on both the datasets in terms of STOI (short-time objective intelligibility), and PESQ (perceptual evaluation of the speech quality). In future, we would like to analyze the performance of the transformer in T-F domain by adding new layers like Time Frequency attention (TFA) and also apply it in speech separation and multi-channel speech enhancement.

REFERENCES

[1] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2017, pp. 1–6

[2] M. H. Soni, N. Shah, and H. A. Patil, "Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5039–5043.

[3] S.-W. Fu et al., "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in ICML, 2019.

[4] N. Shah, H. A. Patil, and M. H. Soni, "Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network," in Proceedings, APSIPA Annual Summit and Conference, 2018, vol. 2018, pp. 12–15.

[5] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation," in Proc. Interspeech 2018, 2018, pp. 2713–2717.

[6] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware Speech Enhancement with Deep Complex UNet," Mar. 2019.

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

[8] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," arXiv preprint arXiv:1703.09452, 2017.

[9] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5069–5073.

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

[11] Defossez A, Synnaeve G, Adi Y, "Real Time Speech Enhancement in the Waveform Domain". arXiv preprint arXiv:2006.12847, 2020.

[12] Fu S, W Tsao, Y Lu, X & Kawai H., "Raw waveform-based speech enhancement by fully convolutional networks," In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 006-012, IEEE, 2017.

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

[14] Y Wang, A Narayanan, and D Wang., "On training targets for supervised speech separation," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] Tan K, & Wang D "A convolutional recurrent neural network for real-time speech enhancement,". In Interspeech, Vol. 2018, pp. 3229-3233, 2018.

[16] Yu F, Koltun V, "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122, 2015.

[17] K Tan, J Chen, and D L Wang., "Gated residual networks with dilated convolutions for monaural speech enhancement," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, pp. 189–198, 2019.

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

[19] Pandey and D Wang., "A new framework for supervised speech enhancement in the time domain," In Proceedings of Interspeech, pp. 1136–1140, 2018.

[20] Pandey A, & Wang D., "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," In International Conference on Acoustics Speech and Signal Processing, pp. 6629-6633, IEEE,2020.

[21] Li Qinglong, Fei Gao, Haixin Guan, and Kaichi Ma., "Real-time monaural speech enhancement with short-time discrete cosine transform,". arXiv:2102.04629,2021.

[22] K Zhen, M S Lee and M Kim., "A Dual-Staged Context Aggregation Method towards Efficient End-to-End Speech Enhancement," International Conference on Acoustics, Speech and Signal Processing, pp. 366-370,2020.

[23] Lin Ju, Adriaan J. Van Wijngaarden, Melissa C. Smith, and Kuang-Ching Wang., "Speaker-Aware Speech Enhancement with Self-Attention, "In 29th European Signal Processing Conference (EUSIPCO), pp. 486-490. IEEE, 2021.

[24] Pandey, A., & Wang, D. (2022). Self-attending RNN for speech enhancement to improve cross-corpus generalization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1374-1385.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[26] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6649–6653.

[27] W. Yu, J. Zhou, H. Wang, and L. Tao, "SETransformer: Speech enhancement transformer," Cogn. Comput., vol. 2021, pp. 1–7, Feb. 2021.

[28] K. Wang, B. He, and W. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) Conf., 2021, pp. 7098–7102.

[29] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 46–50.

[30] Zhao, J., Mao, X., & Chen, L, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical signal processing and control, *47*, 312-323.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[32] M. Sperber, J. Niehues, G. Neubig, S. Stuker, and A. Waibel, "Self-attentional acoustic models," Proc. Interspeech 2018, pp. 3723–3727, 2018.

[33] J Chen, Q Mao, D. Liu "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation". arXiv preprint arXiv:2007.13975, 2020.

[34] Li, X., Li, Y., Dong, Y., Xu, S., Zhang, Z., Wang, D. and Xiong, S., 2020. Bidirectional LSTM Network with Ordered Neurons for Speech Enhancement. In *Interspeech* (pp. 2702-2706).

[35] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[36] C. Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," 2017.

[37] Veaux C, Yamagishi J, King S. "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database." In 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation(O-COCOSDA/CASLRE). IEEE, 2013: 1-4.

[38] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environ- mental noise recordings," The Journal of the Acoustical Society of America, vol. 133, no. 5, pp. 3591–3591, 2013.

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

[40] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015.

[41] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, no. 3, pp. 247–251, 1993.

[42] C H Taal, R C Hendriks, R Heusdens, and J Jensen., "An algorithm for intelligibility prediction of time– frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[43] W Rix, J G Beerends, M P Hollier, and A P Hekstra., "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," In ICASSP, pp. 749–752, 2001.

[44] Shah N, Patil A, Soni H. Time-frequency mask-based speech enhancement using convolutional generative adversarial network, Asia-Pacifc Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2018; p. 1246–51.

[45] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.