Yecchuri and Vanambathina
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:8
https://doi.org/10.1186/s13636-024-00331-z

EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Sub-convolutional U-Net with transformer attention network for end-to-end single-channel speech enhancement

Sivaramakrishna Yecchuri[1] and Sunny Dayal Vanambathina[1*]

Abstract

Recent advancements in deep learning-based speech enhancement models have extensively used attention mechanisms to achieve state-of-the-art methods by demonstrating their effectiveness. This paper proposes a transformer attention network based sub-convolutional U-Net (TANSCUNet) for speech enhancement. Instead of adopting conventional RNNs and temporal convolutional networks for sequence modeling, we employ a novel transformer-based attention network between the sub-convolutional U-Net encoder and decoder for better feature learning. More specifically, it is composed of several adaptive time—frequency attention modules and an adaptive hierarchical attention module, aiming to capture long-term time-frequency dependencies and further aggregate hierarchical contextual information. Additionally, a sub-convolutional encoder-decoder model used different kernel sizes to extract multi-scale local and contextual features from the noisy speech. The experimental results show that the proposed model outperforms several state-of-the-art methods.

Keywords Speech enhancement, Adaptive time-frequency attention transformers, Adaptive hierarchical attention, Transformer attention network and Sub-convolutional encoder

1 Introduction

Background noise and other residual sounds reduce the quality and intelligibility of recorded speech signal in a real-time. The goal of speech enhancement (SE) is to restore the intended speech by eliminating distracting ambient noise and noisy speech mixes. Single-channel speech enhancement refers to the scenario where only a single mix is available, which is an extreme case of the undetermined problem, i.e., the number of sources is greater than the number of mixes. This problem is found in many real-world applications, such as mobile communications, automatic speech recognition, and robotics [1–5].

Data scarcity is a major challenge when we train the deep learning (DL) models. DL demands a large amount of data to achieve exceptional performance. Federated learning is a distributed deep learning approach that allows institutions or hospitals to train a model on their data without sharing it, addressing privacy and regulatory concerns [6]. Each institution trains a model locally and shares the parameters with a central server, which aggregates them to create a global model. This process is repeated until convergence, improving model performance and generalizability by combining data from multiple institutions. Self-supervised learning [6] is another technique that uses unannotated data and a small amount of annotated data to train models, pre-training them on large datasets and fine-tuning on smaller datasets. Knowledge distillation involves training a smaller model to mimic a larger model's behavior, addressing data scarcity. Loss functions are critical in DL models, and in the case of data scarcity, selecting an appropriate

*Correspondence:
Sunny Dayal Vanambathina
sunny.dayal@vitap.ac.in
[1] School of Electronics Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India

one becomes crucial. Mean squared error, mean absolute error, cross-entropy loss, and hinge loss are commonly used loss functions for regression, multi-class classification, image classification, and binary classification problems, respectively.

The low complexity spectral enhancement methods are very suitable for hearing aids users [7]. The spectral subtraction technique, initially introduced by Boll [8], uses the assumption of uncorrelated speech and noise to remove noise in speech. This approach was further enhanced by Berouti et al. [9]. to minimize the artifacts caused by noise reduction. These methods can be generalized to enhance quality by appropriately adjusting the parameters [10]. In line with this concept, Sim et al. [11] proposed a method for optimal parameter selection based on minimum mean squared error. Additionally, Hu and Yu [12] suggested an adaptive noise estimation method to improve quality.

The multiband spectral subtraction method [13], which takes advantage of the non-uniform distribution of noise in different frequency bands, allows for adaptive noise attenuation in each band, resulting in improved speech quality; however, its application in hearing aids is not feasible due to their strict low-power and low-latency requirements.

Traditional noise reduction designs, although effective, are limited in their application to hearing aids due to their complexity and latency; however, a study in [14] presents a sample-based perceptual multiband spectral subtraction with a multiplication-based entropy voice activity detection, specifically tailored for low-power and low-latency requirements of completely-in-the-canal hearing aids.

A hearing device's spectral enhancement requires a filter bank with equally spaced narrow frequency bands and a stopband attenuation of at least 60 dB, low computational complexity, and a small time delay of less than 10 ms, which can be achieved using a uniform polyphase DFT filter bank implemented through the FFT, with a suggested 32-channel filter bank with a time delay of 8 ms under a sampling rate of 16 kHz [15–17].

In [18], a hearing device filter bank is proposed along with a spectral enhancement algorithm, and [19] provides a description of a low-complexity method for subband decomposition of audio signals in digital hearing aids for audibility restoration applications, making it an ideal choice for the design of a digital hearing aid. Moreover, the use of a modified discrete Fourier transform (MDFT) method with moderate hardware complexity [20] can also achieve sound wave decomposition.

There are many different techniques that have been proposed for SE. Traditional techniques include statistical techniques based on statistical modeling of spatial, spectral, or temporal properties of the sensor signals, such as adaptive Wiener filtering [21] and minimum mean square error (MMSE) estimation [22] model. For example, by modeling the spectral components of speech and noise as statistically independent Gaussian random variables, the MMSE estimator achieves an improvement.

In terms of speech enhancement, deep neural networks (DNNs) are now considered the state of the art. DNN-based algorithms [23–25] to learn the relationship between the noisy speech and the target speech through training based on masks or maps. Using an ideal binary mask (IBM) or an ideal ratio mask (IRM) as the training target, the trained model is then used to predict the target speech through the T-F mask [26–28] or mapping [29]. According to recent findings, mapping-based models perform better than masking-based models [30].

Vanilla DNNs and recurrent neural networks (RNNs) have been used for temporal modeling of speech [31], which is different from traditional DNNs. Long short-term memory (LSTM) [32] employed the input, output, and reset gates to record the interdependence between the past and present frames of noisy speech. This increases the estimation accuracy for the mask and mapping relations [33]. The bi-directional LSTM (Bi-LSTM) has been proposed to replace the LSTM. According to earlier findings, it enhances performance under unseen speakers [30, 31]. Bi-LSTM considers future frames and preserves the long-term interdependence between the past, present, and future frames of noisy speech [30].

The use of convolutional neural networks (CNN) [34] is another potential area of SE research. Convolutional encoder-decoder (CED) is proposed to estimate the mapping relationship between the noisy and the target speech. In [35], a deep complex recurrent convolutional neural network (DCCRN) was proposed. It uses a complex convolutional encoder and decoder model that employs complex LSTM and dense layers between the center of the encoder and decoder blocks. A complex LSTM and dense layer are used to extract the temporal dependencies from the complex encoder-decoder structure. The multi-resolutional convolutional encoder (MRCE) model has been proposed to improve the performance of SE by increasing the receptive fields of the network in WaveNet with extended convolutions and using a gated mechanism to regulate the information flow [36, 37]. To enlarge the receptive fields in the time-frequency (T-F) domain, the gated residual network (GRN) [30] and dilated convolutions (DCN) [38] approaches used with 1-D dilated convolutions.

The raw waveform is used directly to regenerate the enhanced speech without using a T-F representation [35, 39–43], which avoids the problem of explicit phase estimation. For example, speech enhancement generative

adversarial network (SEGAN) [40] proposed a generative adversarial network based SE method, in which a denoising generator directly maps the raw waveform of the clean speech from the mixed raw waveform by adversarial training. In [43], a temporal convolutional neural network (TCNN) is proposed to improve the performance of SE in the time domain. The TCNN utilizes a series of 1D causal and dilational convolution to capture the long-range speech context from past and previous frames. A multi-scale feature recalibration convolutional GRU network (MCGN) [42] model for SE. Local and contextual features can be extracted from the signal using multi-scale convolutional layers for recalibration. In the recalibration network, the information flow between the layers is controlled by gating, preserving speech and suppressing noise by weighting the rescaled features.

Some deep learning-based SEs have also employed attention mechanisms to control computational costs and overall parameters. Attention networks that optimize the weights of input features can be achieved with a neural attention module to minimize losses. In learning-based enhanced frameworks, information can be improved and interference from irrelevant information can be reduced. The squeeze-and-excitation attention (SEA) model was proposed in [44]. The algorithm uses global 2D pooling to calculate channel attention and offers impressive performance improvements. In [45], a convolutional block attention module is proposed that sequentially improves key parts of the input features by channel attention and spatial attention. Multi-scale and attention mechanisms for end-to-end single-channel speech enhancement (MASENet) [46] is a combination of multi-scale convolutional models and temporal convolutional attention (TCA) to extract local and global feature information from speech. The outputs of the MASENet encoder blocks are recalibrated by the attention block and highlight informative details. In the nested U-Net with self-attention and dense connectivity (SADNUNet) [47] model, the encoder and decoder model uses nested U-Net and dense blocks to extract local and contextual features from the speech. All encoder group outputs are recalibrated by the self-attention (SA) block, which highlights informative details and reduces unwanted features.

In the state-of-the-art attention-based methods of SE described above, different attention modules are used to determine significant features either in the spatial domain or in the channel domain. Attention models generate a strong loss of information that affects speech intelligibility and quality. To avoid this, we use the transformer attention network (TAN).

Transformer-based attention networks, renowned for their exceptional performance in the domain of speech enhancement, have established their efficacy in parallel computation. These networks, as evidenced by their impressive results [41, 48, 49], possess the unique ability to address the challenge of long-dependency more effectively than traditional recurrent neural networks (RNN) or convolutional neural networks (CNN). The distinguishing feature of transformer-based attention networks lies in their ability to model speech sequences directly, thereby incorporating contextual information for a more comprehensive understanding of the data.

More specifically, it consists of several adaptive transformer-based spectro-temporal attention modules and an adaptive hierarchical attention module that aims to capture long-term time-frequency dependencies and further aggregate intermediate hierarchical context information. The loss of information in TAN is very low compared to the SEA, TCA, and SA models.

To solve these problems, we propose a sub-convolutional U-Net (SCUNet) with a TAN mechanism for speech enhancement (TANSCUNet).

The specific contributions of the proposed sub-convolutional U-Net (SCUNet) with TAN mechanism for speech enhancement (TANSCUNet) are as follows.

- SCUNet is basically a convolutional encoder-decoder model that uses different sized kernels in each convolutional layer to generate features in different orders of magnitude. This allows each feature in each scale to be assigned its own weight, so that the speech-related components are preserved while the noise-related ones are suppressed, and the interdependence between local and global contextual information in speech can be captured.
- The TAN is equipped with three adaptive time-frequency attention (ATFA) transformers and an adaptive hierarchical attention (AHA) module. The ATFA transformers can capture local and global context information in both the time and frequency dimensions, while the AHA module can flexibly summarize all the output feature maps of the ATFA modules by a global attention weight.
- Finally, the output layer sums the multiscale outputs and accelerates convergence. The output layer appreciates the improved speech output by providing access to multiple scales of convolutional operators that facilitate the training of the network.

The rest of the paper is organized as follows. Section 2 describes the proposed TANSCUNet method. Section 3 describes the analysis of the experimental results. Section 4 contains the conclusions.

2 Proposed model

The proposed TANSCUNet model is shown in Fig. 1. The TANSCUNet model consists of a sub-convolutional encoder, a decoder, central layers, and an output layer. The input of the proposed model is a noisy waveform, which is divided into frames using the Hanning window in to-Frame block. The output of the to-Frame block is fed to the input layer to extract the intermediate features. From the output of the input layer, we can extract the context information at different scales by using the sub-convolutional U-Net. The depth of the U-Net model is five (i.e., five sub-convolutional encoder and decoder blocks are used). Each block of the sub-convolutional encoder (SCE) contains seven different sub-convolutionals with different kernel sizes to extract the multiscale features. The sub-convolutional decoder (SCD) block is a mirror version of the SCE block. The output of the last SCE block is fed into the transformer attention network.

The TAN consists of an adaptive hierarchical attention module (AHA) and three adaptive time-frequency attention modules (ATFA). Together, the ATFA and AHA modules create an "attention-in-attention" structure based on the adaptive attention weights. The results of ATFA can be further improved and integrated by AHA. In addition, skip connections are used to improve the information flow between the SCE and SCD blocks. The stride value is (1,2) in all layers of SCEs and SCDs, except in the output layer. In the output layer, the stride value is set to (1,1).

2.1 Subconvolutional encoder and decoder block

During CNN training, a high-level feature can be influenced by the receptive field. Local information can be extracted from a small receptive field, while contextual information can be extracted from a large receptive field [30]. Traditionally, CNNs use a fixed kernel size that balances the extraction of local and contextual information. A subconvolutive encoder (SCE) block addresses this limitation by capturing information at different scales and generating multi-scaled features.

Figure 2 shows the architecture of the SCE block. To capture information at different scales, SCE uses different convolutional operators of different sizes on the encoder side. Small kernel sizes of convolution operators can capture the local dependency between neighboring T-F points in the short duration. By using the smallest kernel size (1,2), two neighboring T-F points can be extracted as features. The extraction of features from the long-duration speech is possible using convolutional operators with large kernel sizes. Compared to smaller kernels, these features contain contextual information. After each convolutional operator, the layer normalization and LReLU [50] operations are performed. Then, as shown in Fig. 2, concatenate outputs of each individual convolutional operation block to generate the input for the next steps. The subconvolutional decoder block (SCD) is similar to the SCE but uses deconvolution operators instead of convolution operators.

The SCE block contains m subconvolution operators. Each has the same number of channels, but different kernel sizes are used to extract the features. X and K represent the SCE input and output respectively. The output $K = [k_1, k_2, ..., k_m]$ represents the m^{th} 2-D sub-convolution that has a different sized kernel.

2.2 Transformer attention network

According to the findings presented in Fig. 1, the proposed TAN module consists of not only an adaptive hierarchical attention (AHA) module but also three adaptive time-frequency attention (ATFA) modules. As it was previously stated in the study conducted by [51] , each ATFA module has the potential to strengthen long-range spectro-temporal relationships with minimal computational cost. This means that the ATFA modules have the capability to reinforce connections between different points in time and frequency in an efficient manner. On the other hand, the AHA module plays a crucial role in collecting comprehensive contextual information by combining numerous intermediate characteristics. By doing so, it is able to gather global multi-scale contextual data, which is then utilized to enhance and integrate the output of the ATFA modules.

The combination of these ATFA and AHA modules results in the formation of an intricate "attention-in-attention" structure, which is primarily based on adaptive attention weights. This structure allows for a more flexible and dynamic allocation of attention to different elements within the input data. Moreover, it enables the ATFA modules to benefit from the contextual information provided by the AHA module, leading to further improvements in their individual outputs. Consequently, the AHA module acts as a facilitator in enhancing the performance and effectiveness of the ATFA modules. Overall, this proposed TAN module exhibits a sophisticated mechanism of adaptive attention that optimizes the utilization of contextual information and reinforces spectro-temporal relationships).

2.2.1 Adaptive time-frequency attention

To mitigate the substantial computational complexity of traditional self-attention methods, we propose the utilization of an innovative adaptive time-frequency attention (ATFA) mechanism as an efficient solution to capture the extensive long-range correlations present in

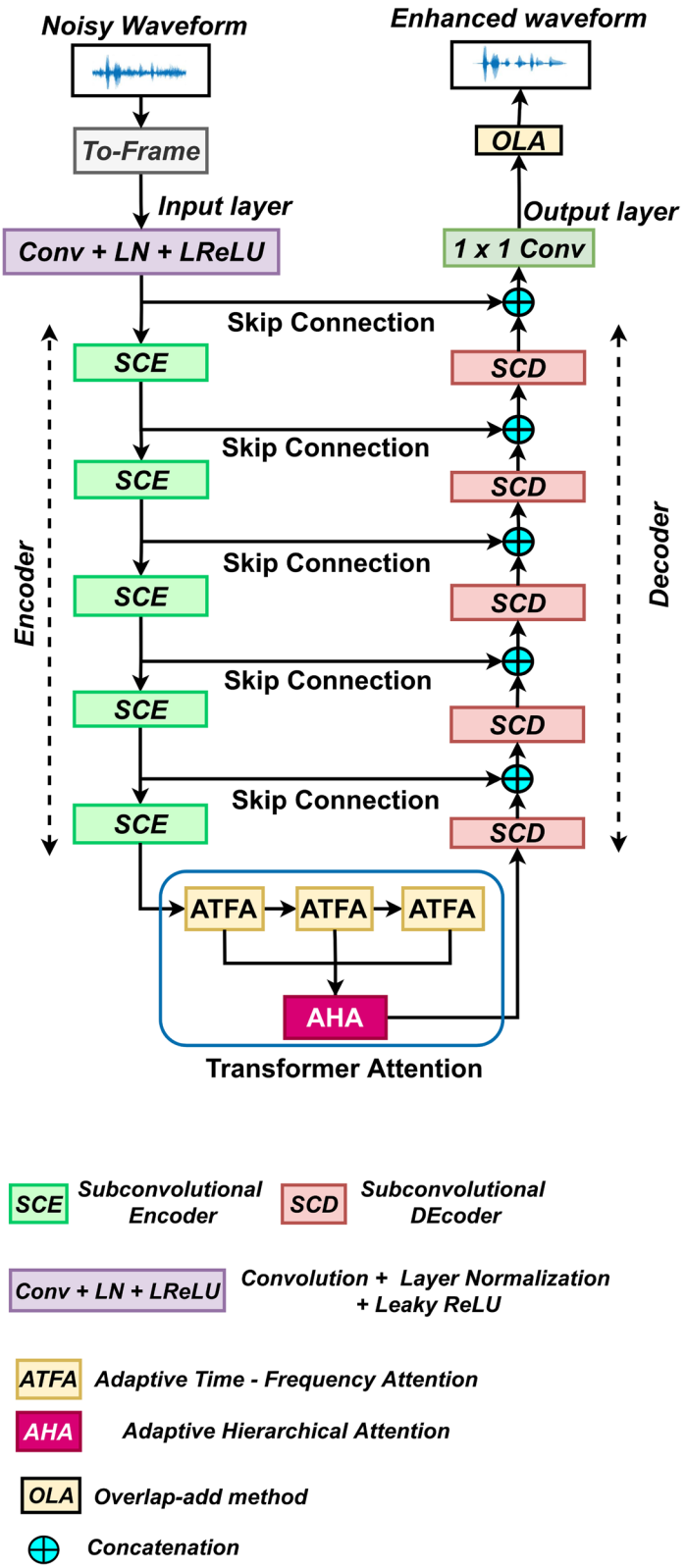

Fig. 1 Proposed TANSCUNet model architecture

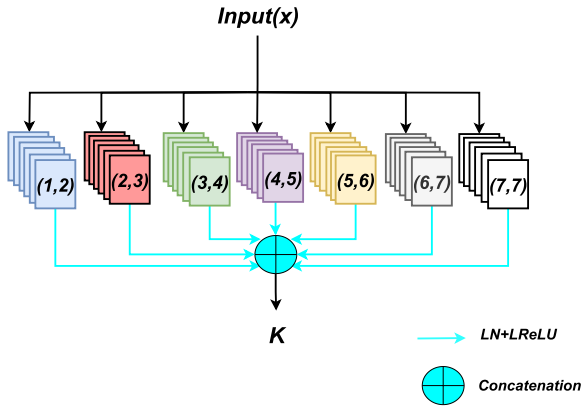

Fig. 2 Architecture of subconvolutional encoder blocks

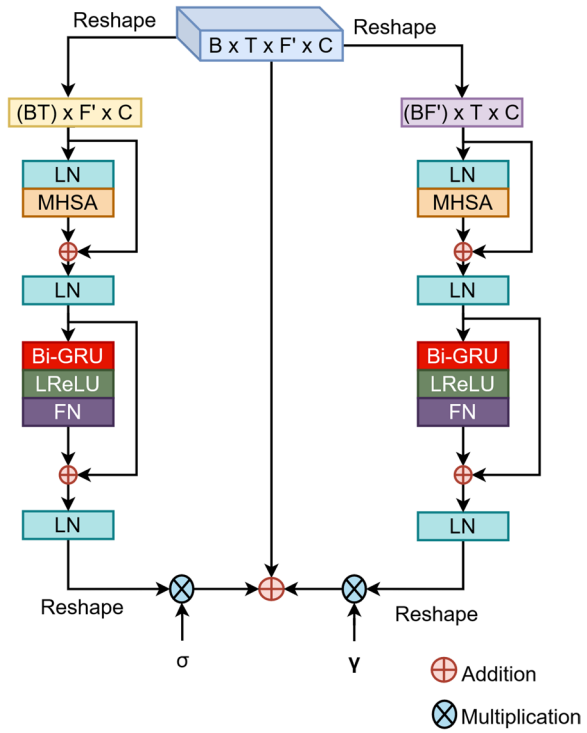

Fig. 3 Architecture of ATFA block

both the temporal and spectral dimensions, as delineated in [51, 52].

As clearly depicted in Fig. 3, the ATFAT is bifurcated into two distinct sub-branches that operate concurrently in the time and frequency axes, namely the adaptive temporal attention branch (ATAB) and the adaptive frequency attention branch (AFAB). These branches are adept at capturing comprehensive global dependencies along the temporal and spectral dimensions due to the incorporation of two adaptive weights, denoted as α and β. In each branch, unlike the conventional transformer, we employ a Bi-GRU-based enhanced transformer [41], which comprises of multi-head self-attention (MHSA) components and a Bi-GRU-based position-wise network. This is followed by the integration of residual connections and layer normalization (LN). The utilization of multi-head self-attention has been widely recognized and employed in the realms of natural language processing and speech processing due to its ability to effectively leverage contextual information contained within feature maps (Fig. 4).

In the MHSA modules, the input features undergo a series of linear projections h times, resulting in the generation of queries (Q), keys (K), and values (V) representations. Here, h denotes the number of heads present in the MHSA modules. Subsequently, the scaled dot-product attention mechanism is executed for each head, leading to the acquisition of a weighted sum of the values. The weights are obtained through an attention function that takes into account the query and the corresponding keys. Finally, the attentions of all heads are concatenated and linearly transformed to produce the ultimate output.

$$F_{in} = Reshape(IN) \tag{1}$$

$$h_i = Attention(q_i, k_i, v_i), \tag{2}$$

$$MHSA = Concatenation(h_1, h_2, .., h_i)W^O \tag{3}$$

$$F_{MHSA} = LN(F_{in} + MHSA) \tag{4}$$

$$FN(F_{MHSA}) = LeakyReLU(Bi - GRU(F_{MHSA})W_1 + B_1 \tag{5}$$

$$Output_{AFAB} = LN(F_{MHSA} + FN(F_{MHSA})) \tag{6}$$

The resulting output of the SCE block serves as an input for the AFAB block, denoted as $IN \in R_{B\times T\times F'\times C}$. The notation $F_{in} \in (B \times T) \times F' \times C$ represents the reshaping of the output, where B, T, F, and C signify the batch size, frame number, frequency dimension, and channel number, respectively.

Our model works with four heads in this context. Following the effectiveness of Bi-GRU-based transformers in speech separation and denoising highlighted in previous studies [41, 48], we introduce a modification of the feed-forward network (FN) in the vanilla transformer by replacing its first fully connected layer with a Bi-GRU. The final output is computed by feeding the output of the multi-head self-attention block (MHSA) into the Bi-GRU-based feed-forward network, followed by the inclusion of residual connections and layer normalization (LN). The notation $FN()$ denotes the output of the Bi-GRU-based linear feed-forward network, and W_1 stands for the weight of the linear transformation and B_1 for the

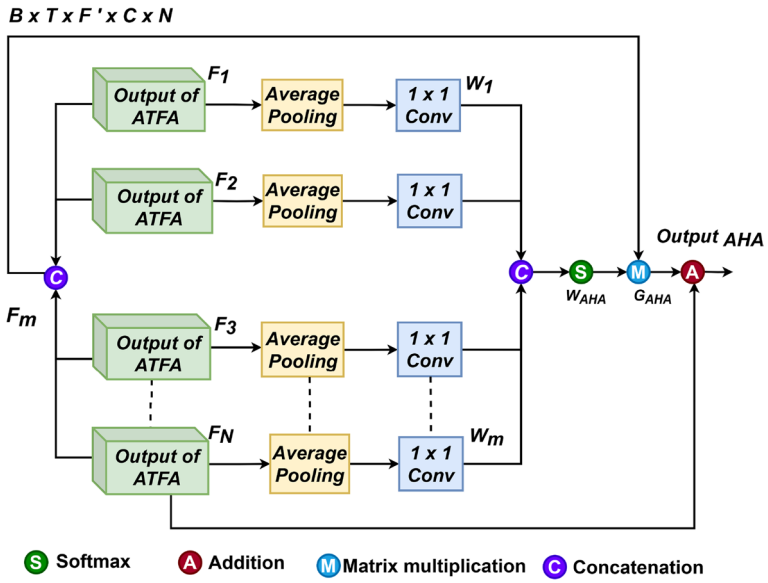

Fig. 4 Architecture of AHA block

bias. It is important to note that C is set to a value of 64 in this module. Then, the final output of the AFAB module is transformed back to the original size, represented as $Output_{AFAB} \in R_{B\times T\times F'\times C}$.

Likewise, the compressed input features undergo a transformation process, resulting in $B \times T$ vectors of dimension $F^{'} \times C$, which are then fed into ATAB to calculate the output, denoted as $Output_{ATAB}$, in parallel along the temporal axis. Finally, the output features from the two branches, as well as the original features, are combined using two adaptive weights σ and γ in order to derive the ultimate output of the ATFA module. Mathematically, this can be formulated as follows:

$$Output_{ATFA} = F_{in} + \sigma\, Output_{ATAB} + \gamma\, Output_{AFAB} \tag{7}$$

where σ and γ are initialized to 1 and automatically assigned appropriate values.

2.2.2 Adaptive hierarchical attention

In the AHA, a technique is used to obtain comprehensive global context information by cascading all intermediate results of the individual ATFA modules. This global context information is denoted by the symbol $F_{m=1,2,3...N}$, where N stands for the number of ATFA modules, which is set to 3 in the proposed method. To ensure efficient compression of the output features of each ATFA, a two-step process is performed. First, an average pooling layer is applied to compress the output feature of each ATFA into a compact representation. Second, a 1×1 convolutional layer is applied to further compress the information. These compressed representations are then cascaded with the outputs of the 1×1 convolutional layers. The extraction of the hierarchical attention information is facilitated by using a softmax function that results in the hierarchical attention weights, denoted by W_{AHA}. These attention weights play a crucial role in capturing the importance of the different features in the hierarchical structure. The definition of W_{AHA} is derived from the softmax function, which ensures that the attention weights sum to one and effectively emphasize the relevant features.

$$W_{AHA} = \frac{exp(Conv_{1\times1}(Avg.Pool(F_m)))}{\Sigma_{m=1}^{N}(Conv_{1\times1}(Avg.Pool(F_m)))} \tag{8}$$

A weighted pooled output, denoted as W_{AHA}, has been established for the m^{th} value which ranges from 1 to N. Following this, a matrix multiplication is performed between the hierarchical attention weight, W_{AHA}, and the global contextual information model, F_m. This ensures that the relationship between the two variables is accounted for and their interaction is taken into consideration.

$$G_{AHA} = W_{AHA} \odot F_m$$

$$G_{AHA} = \Sigma_{m=1}^{N}\left(\frac{exp(Conv_{1\times1}(Avg.Pool(F_m)))}{\Sigma_{m=1}^{N}(Conv_{1\times1}(Avg.Pool(F_m)))}\right) \odot F_m \tag{9}$$

The variable $G_{AHA} \in R_{B\times T\times F'\times C}$ denotes the aggregation of the global contextual feature map. To obtain the final output, denoted as $Output_{AHA} \in R_{B\times T\times F'\times C}$ the

output of the last ATFA block F_N is added to G_{AHA}. This combination of the output of the last ATFA block and the summation of the global contextual feature map results in the final output.

$$Output_{AHA} = F_N + G_{AHA} \tag{10}$$

2.3 Output Layer

As shown in Fig. 1, the skip connection is used to provide input to the output layer. Based on the size of the noisy input mixture and the information flow of the previous layer, the output layer can predict clean speech. In the output layer 1×1 convolution layers are used. By utilizing the overlap addition method, we predict the enhanced waveform.

3 Experimental result analysis

3.1 Datasets

To test our model, we use the *Common Voice* [53] corpus, a publicly available speech database. The database contains 1.6 million utterances from 84,659 speakers. From these, we select the *Common Voice Corpus* 13.0 under the English category. It consists of 3209 recorded hours, 2429 h of validation, and the total number of utterances is 86,942. We randomly select 70% of utterances for the training set and 30% of utterances for the validation set. The test set is also from the *CommonVoiceCorpus*13.0, which consists of 4000 utterances. We created training and validation sets with 125 different types of noise and different signal signal to noise ratios (SNR) values from − 5 to +5 dB. Clean words, noise, and SNR are randomly selected in each mixed method.

We created two test sets to evaluate the generalization capability of the model, one for seen noise conditions and the other for unseen noise conditions. From the *NOIZEUS* [54] database, we collected street, restaurant, and babble noises for seen noise condition test, while train, exhibition hall, and airport noises selected for unseen noise condition test. To test the noise mixture, we used three SNR levels: − 5 dB, 0 dB, and 5 dB.

Speech enhancement performance is measured using the following metrics: signal-to-distortion ratio (SDR) [55], perceptual evaluation of speech quality (PESQ) [56], and short-time objective intelligibility (STOI) [57]. The SDR is derived from the estimated speech SDR value minus the noisy mixture SDR value. A PESQ score ranges from − 0.5 to 4.5, indicating the quality of speech perception. STOI measures the quality of human speech intelligibility and ranges from 0 to 1. Higher values indicate better enhancement performance.

3.2 Experimental setup and baselines

All utterances are sampled at 16 kHz. For model building, individual utterances are converted into stacks of utterances and then employed the 512 length of a hanning window with a hop length of 256. The model is trained over 60 epochs, the optimizer is Adam [58], learning rate is 0.002, and batch size is 32 throughout each epoch.

Performance comparison the following baselines used namely Bi-LSTM [31], Bi-CRN [39], GRN [30],SEGAN [40], DCN [38], TSTNN [41], DCCRN[35], MCGN [42], MASENet [46], SADNUNet [47], and DBT-Net [51]. Note that we re-implement all baselines with non-causal configurations in order to ensure fair comparisons.

3.3 Ablation study of TANSCUNet model

Table 1 shows an ablation study of the proposed model. The performance of the proposed model is evaluated in terms of SDR, STOI, and PESQ metrics. The U-Net model is a basic encoder-decoder model, having convolutions and deconvolutions with the same kernel size. The depth (N) of U-Net varies from 2 to 7 when evaluating mean square error (MSE) loss for 50 epochs [59]. The model loss is significantly decreased when the depth of the model is chosen from 2 to 5. From $N = 6$ to 7 loss values are scattered. So, we chose the depth of the U-Net as 5.

Table 1 Ablation study of the proposed model is shown in terms of averaged SDR, STOI, and PESQ metrics. The proposed model is indicated in the BOLD Italic text. N indicated the depth of UNet

Metrics	TAN model			Par. (M)	PESQ				STOI (%)				SDR (in dB)			
SNR (in dB)	ATAB	AFAB	AHA	-	− 5.00	0.00	5.00	Avg.	− 5.00	0.00	5.00	Avg.	− 5.00	0.00	5.00	Avg.
Raw speech	x	x	x	x	1.48	1.66	1.87	1.67	32.14	41.24	50.17	41.18	− 2.98	0.14	3.15	0.10
SCUNet ($N = 5$)	x	x	x	13.20	2.18	2.41	2.68	2.42	62.01	69.46	76.04	69.17	5.78	8.03	10.56	8.12
TANSCUNet	✓	x	x	3.25	2.31	2.69	2.91	2.63	64.35	71.05	78.32	71.24	6.89	9.07	11.09	9.02
TANSCUNet	x	✓	x	3.25	2.53	2.78	3.02	2.78	66.16	73.26	80.72	73.38	7.83	10.18	11.57	9.86
TANSCUNet	✓	✓	x	3.51	2.66	2.91	3.16	2.90	68.33	75.54	82.26	75.38	8.55	10.91	12.13	10.53
TANSCUNet	✓	✓	✓	3.51	2.85	3.12	3.37	3.08	72.52	79.65	84.36	78.84	9.81	11.85	13.62	11.76

Next, we replaced the U-Net encoder and decoder with SCE and SCD, which we named SCUNet. The SCE contains seven sub-convolutional layers with the same size and different kernel sizes. SCUNet provides a significant improvement in SDR, PESQ, and STOI. The total trainable parameters of SCUNet is 13.20 million, so the computational cost is very high.

Next, TAN is incorporated into SCUNet, i.e., TANSCUNet. The TAN consist of three ATFA blocks and an AHA block. Each ATFA block is a combination of ATAB and ATFB, which are capable of capturing the global dependencies along the temporal and frequency axis. Case I: from Table 2, in TAN, select only ATAB to extract useful significant multi-scale temporal context. By incorporating ATAB, the model parameters are reduced to 3.25 million. The model performance also improves significantly over the SCUNet, i.e., 0.90 in SDR, 2.07 in STOI, and 0.21 in PESQ.

Case II: We select only AFAB in TAN. Now, the TAN is capable of capturing the global dependencies in frequency axis and also extracts the significant multi-scale context. The model performance significantly improves over the ATAB based TAN, i.e., 0.84 in SDR, 2.14 in STOI, and 0.15 in PESQ.

Case III: We select both ATAB and AFAB block to form a ATFA in TAN. Now, the ATFA is capable of capturing the global dependencies in temporal-frequency axis. By incorporating ATFA, the model parameters are increased 0.3 million, but the model performance improves significantly, i.e., 0.67 in SDR, 2 in STOI, and 0.12 in PESQ.

Case IV: Finally, we select the ATFA and AHA blocks. AHA module can combine many intermediate characteristics to collect global multi-scale contextual data. Together, the ATFA and AHA modules create a "attention-in-attention" structure based on the adaptive attention weights; the output of ATFA may be further improved and integrated by AHA. By incorporating AHA, model performance improves significantly, i.e., 1.23 in SDR, 3.46 in STOI, and 0.18 in PESQ.

Table 2 Multi-kernel size analysis

Kernel size	SDR	STOI	PESQ
1×2	10.43	71.06	1.74
2×2	10.61	71.79	1.81
2×3	10.65	72.16	1.88
3×4	10.76	73.54	1.96
4×5	10.94	74.38	2.01
5×6	11.05	75.67	2.12
6×7	11.14	76.04	2.37
7×7	11.32	77.41	2.41
10×10	11.21	75.16	2.33
Multi-kernel	12.03	79.65	2.73

3.4 Multi-kernel analysis

Our next experiment examines how kernel size affects performance under seen and unseen noise conditions at 0dB SNR. As shown in Table2, performance also depends on the choice of kernel size. We test different kernel sizes from 1×1 to 10×10 to exploit different receptive fields. When the kernel size is larger than 7×7, the performance in terms of SDR, STOI, and PESQ may decrease. Multi-kernel utilizes the diffrent kernels to allows the model to capture features at different scales, thereby exploiting both local and contextual information. The smoothing effect becomes stronger at larger kernel sizes, mitigating noise, while smaller kernel sizes preserve finer spectral structures. With a bank of kernels, the model has a greater probability of capturing and differentiating features of noise and speech, improving speech enhancement.

3.5 Performance comparison with baselines under seen condition

The model is already trained with test speeches and noises under seen conditions. Babble, street, and restaurant noises are used to test the model. Tables 3, 4, and 5 show the performance of the proposed method with baselines in terms of PESQ, STOI, and SDR metrics.

Bi-LSTM and Bi-CRN are magnitude-based methods, whereas the bi-directional RNN-based SE models adopt a typical CRN with an encoder-decoder model.

Bi-LSTM produces the lowest enhancement performance with an average of 6.23 dB of SDR, 73.53 % of STOI, and 2.15 PESQ. The Bi-CRN uses a multi-resolution convolutional encoder-decoder and shows a slight increase in SDR, STOI, and PESQ over the Bi-LSTM. Bi-CRN achieves 6.63 dB SDR, 75.32 % STOI, and 2.31 PESQ. Due to its ability to capture global spatial patterns. Additionally, LSTM layers incorporate past and current temporal frames into the CRN to exploit temporal dependency. CRN has more trainable parameters. Each LSTM requires four linear layers (MLP layers) per cell to run at each time step. Linear layers require large memory bandwidth. During training, LSTM faces the "vanishing gradient" problem.

The GRN model produces 7.42 dB of SDR, 77.83 % of STOI, and 2.51 PESQ values. The GRN model is constructed with residual and dilated convolution blocks and has been shown to perform well in many applications. The main drawback is that a deep network usually requires weeks of training, making it practically infeasible in real-time applications, and learning can be very inefficient if the network is too shallow.

Table 3 SDR values of all baseline models under seen noises. Proposed model represented by bold and italic letters

Metric	**SDR**											
Noise	**Babble**				**Street**				**Restaurant**			
SNR (dB)	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**
Noisy mixture	2.12	4.01	5.89	4.01	1.82	3.65	5.62	3.70	2.16	4.18	6.13	4.16
Bi-LSTM [31]	4.32	6.65	7.92	6.30	4.12	6.01	7.39	5.84	4.56	6.88	8.23	6.56
Bi-CRN [34]	4.79	6.94	8.37	6.70	4.51	6.43	7.88	6.27	4.87	7.26	8.6	6.91
SEGAN [40]	5.03	7.22	8.72	6.99	4.92	6.89	8.42	6.74	5.12	7.63	8.91	7.22
GRN [30]	5.25	7.59	9.22	7.35	5.23	7.26	8.87	7.12	5.97	7.97	9.44	7.79
DCN [38]	5.85	7.99	9.64	7.83	5.56	7.84	9.18	7.53	6.22	8.23	9.83	8.09
DCCRN [35]	6.13	8.34	9.96	8.14	5.89	8.26	9.53	7.89	6.79	8.59	10.25	8.54
TSTNN [41]	6.57	8.74	10.42	8.58	6.11	8.59	9.85	8.18	7.16	8.92	10.69	8.92
MASENet [46]	6.94	9.12	10.84	8.97	6.61	8.83	10.21	8.55	7.54	9.36	10.95	9.28
SADNUNet [47]	7.32	9.51	11.11	9.31	6.97	9.15	10.60	8.91	7.89	9.74	11.31	9.65
MCGN [42]	7.61	9.87	11.53	9.67	7.36	9.54	10.94	9.28	8.02	10.09	11.75	9.95
DBT-Net [51]	7.92	10.03	11.82	9.92	7.64	9.86	11.10	9.53	8.22	10.24	11.97	10.14
TANSCUNet	**8.62**	**10.69**	**12.57**	**10.63**	**8.37**	**10.48**	**11.58**	**10.14**	**8.81**	**10.86**	**12.71**	**10.79**

Table 4 STOI values of all baseline models under seen noises. Proposed model represented by bold and italic letters

Metric	**STOI**											
Noise	**Babble**				**Street**				**Restaurant**			
SNR (dB)	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**
Noisy mixture	56.75	62.04	68.55	62.45	55.41	61.26	69.17	61.95	58.57	66.47	73.15	66.06
Bi-LSTM [31]	67.79	74.22	78.57	73.53	67.26	71.35	77.14	71.92	68.26	76.43	80.76	75.15
Bi-CRN [34]	68.51	75.35	80.87	74.91	68.68	73.97	79.54	74.06	70.26	78.43	82.23	76.97
SEGAN [40]	69.93	76.69	81.69	76.10	69.59	75.03	81.19	75.27	72.26	79.13	83.08	78.16
GRN [30]	70.12	78.94	82.05	77.04	70.34	76.36	82.88	76.53	74.96	80.05	84.81	79.94
DCN [38]	72.09	80.11	83.77	78.66	71.26	78.91	83.68	77.95	75.19	81.74	85.64	80.86
DCCRN [35]	74.13	81.54	84.98	80.22	73.56	79.34	84.06	78.99	76.46	82.16	86.97	81.86
TSTNN [41]	75.41	83.16	86.53	81.70	74.59	81.23	85.14	80.32	77.39	83.69	87.56	82.88
MASENet [46]	77.32	84.04	87.15	82.84	76.68	82.24	86.85	81.92	78.18	84.47	88.18	83.61
SADNUNet [47]	78.51	86.43	88.11	84.35	77.61	84.41	87.56	83.19	79.52	86.63	90.21	85.45
MCGN [42]	80.31	87.78	90.03	86.04	78.54	85.69	88.94	84.39	80.13	88.33	91.83	86.76
DBT-Net [51]	80.92	88.03	91.12	86.69	79.64	86.71	89.40	85.25	81.22	89.24	92.47	87.64
TANSCUNet	**82.62**	**89.71**	**92.72**	**88.35**	**81.69**	**88.12**	**91.94**	**87.25**	**82.56**	**90.81**	**93.86**	**89.08**

The DCN model produces 7.82 dB of SDR, 79.15% of STOI, and 2.60 PESQ values. The DCN model builds on a stack of dilated convolutions that summarize contextual information at multiple levels without losing resolution. The dilated convolution is constructed by inserting zeros into the convolution kernel, which can increase the receptive field and the resolution of the outputs. However, a stack of dilated convolutions can lead to a "gridding" problem.

The DCCRN model produces 8.19 dB of SDR, 80.36 % of STOI, and 2.71 PESQ values. The model is constructed with complex CED and dense layers. With a dense layer, the receptive area is increased, and more temporal dependencies are extracted from the complex CED model. DCCRN's limitation is that kernel sizes increase exponentially in dense blocks, which can lead to aliasing.

The TSTNN model produces an average 8.56 dB of SDR, 81.63 % of STOI, and 2.84 of PESQ. The TSTNN utilizes a sequence of four two-stage transformer blocks to model local and global information from the encoder. The encoder uses the dilated dense block to exploit more receptive fields, which causes aliasing.

Table 5 PESQ values of all baseline models under seen noises. Proposed model represented by bold and italic letters

Metric	PESQ											
Noise	Babble				Street				Restaurant			
SNR (dB)	− 5	0	5	Avg.	− 5	0	5	Avg.	− 5	0	5	Avg.
Noisy mixture	1.23	1.52	1.83	1.53	1.51	1.83	2.02	1.79	1.66	1.88	2.01	1.85
Bi-LSTM [31]	1.85	1.97	2.44	2.09	1.84	2.02	2.49	2.12	1.98	2.11	2.66	2.25
Bi-CRN [34]	1.92	2.13	2.53	2.19	1.93	2.21	2.57	2.23	2.03	2.21	2.77	2.34
SEGAN [40]	1.99	2.21	2.66	2.29	2.05	2.29	2.68	2.34	2.15	2.38	2.83	2.45
GRN [30]	2.08	2.29	2.71	2.36	2.12	2.45	2.75	2.44	2.23	2.49	2.96	2.56
DCN [38]	2.17	2.38	2.85	2.47	2.22	2.49	2.87	2.52	2.31	2.63	3.04	2.66
DCCRN [35]	2.24	2.51	2.94	2.56	2.37	2.65	2.95	2.66	2.47	2.74	3.11	2.77
TSTNN [41]	2.36	2.62	3.07	2.68	2.48	2.73	3.09	2.76	2.55	2.99	3.25	2.93
MASENet [46]	2.45	2.76	3.13	2.78	2.59	2.83	3.16	2.87	2.68	3.08	3.37	3.04
SADNUNet [47]	2.58	2.83	3.24	2.88	2.66	2.94	3.27	2.96	2.72	3.16	3.46	3.11
MCGN [42]	2.64	2.90	3.32	2.95	2.79	3.11	3.35	3.08	2.81	3.27	3.53	3.20
DBT-Net [51]	2.69	2.97	3.38	3.01	2.84	3.16	3.40	3.13	2.87	3.34	3.59	3.23
TANSCUNet	**2.95**	**3.12**	**3.52**	**3.20**	**2.97**	**3.31**	**3.56**	**3.44**	**2.98**	**3.49**	**3.84**	3.44

MASENet is a combination of convolutional multi-scale and temporal convolutional attention models to extract local and global feature information from speech. MASENet encoder block group outputs are calibrated by the attention block and emphasize informative details. As a result, the model generates 8.93 dB of SDR, 82.79 % of STOI, and 2.94 PESQ values on average. The model limits more features depending on temporal channel attention, which affects speech intelligibility.

The SADNUNet model produces an average of 9.29 dB of SDR, 84.33 % of STOI, and 3.03 PESQ. SADNUNet is a nested U-Net model. Each encoder-decoder uses the dense block to extract local and contextual features from speech. The self-attention block calibrates the encoder output to improve the temporal context while reduce unwanted parameters. SADNUNet's limitation is that the dense block increases the kernel size exponentially to cover large receptive areas, which leads aliasing.

The MCGN model produces an average of 9.63 dB of SDR, 85.73 % of STOI, and 3.13 PESQ values. Local and contextual features can be extracted from the signal using multi-scale recalibration convolutional layers. In the calibration network, control the information flow between layers, thus improving the speech quality. MCGN has more trainable parameters (around 77 million), which require large amounts of memory bandwidth.

In comparison with the baseline methods, the proposed TANSCUNet model achieves, on average, 10.52 dB of SDR, 88.23 % of STOI, and 3.36 PESQ. These values are 0.66 dB, 1.72 %, and 0.23 higher relative to the DBT-Net model. TANSCUNet learns residual mapping relationships from raw data at different scales. Small kernel sizes of sub-convolutional layers capture local dependencies, while large kernel sizes determine the global dependency between larger regions. This allows us to enlarge TANSCUNet's receptive field and assign different weights to the various scaled features. In addition, TAN is introduced to link the sub-convolutional encoder and decoder, which exploits the interdependence between the past, present, and future frames.

3.6 Objective comparison of baseline models under unseen noises

The performance of the proposed method is shown in Tables 6, 7, and 8 under unseen noise conditions. The unseen speakers and noises were used for testing. Trains, airports, and exhibition hall noises are unseen noises. The proposed TANSCUNet model achieves, on average, 10.12 dB of SDR, 87.14 % of STOI, and 3.24 PESQ. These values are 0.85 dB, 1.6 %, and 0.15 higher relative to the DBT-Net model. Similarly, compared with all baselines, the proposed method shows significant improvement in terms of SDR, STOi, and PESQ metrics. In TANSCUNet, small kernel sizes of sub-convolutional layers capture local dependencies, while large kernel sizes determine the global dependency between larger regions. This allows us to enlarge TANSCUNet's receptive field and assign different weights to the various scaled features. In addition, TAN are introduced to link the sub-convolutional encoder and decoder, which exploits the interdependence between the past, present, and future frames.

Table 6 SDR values of baselines under unseen noise condition. Proposed model represented by bold and italic letters

Metric	**SDR**											
Noise	**Train**				**Airport**				**Exhibition hall**			
SNR (dB)	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**
Noisy mixture	1.83	3.52	5.53	3.63	2.11	3.83	5.82	3.92	2.66	3.98	5.94	4.19
Bi-LSTM [31]	3.71	5.75	6.41	5.29	3.81	5.93	6.81	5.52	3.93	5.59	6.89	5.47
Bi-CRN [34]	4.05	6.19	6.75	5.66	4.25	6.37	7.08	5.90	4.27	6.94	7.21	6.14
SEGAN [40]	4.45	6.63	7.29	6.12	4.69	6.71	7.69	6.36	4.71	7.21	7.59	6.50
GRN [30]	4.99	6.89	7.72	6.53	5.12	6.97	7.92	6.67	5.12	7.52	7.89	6.84
DCN [38]	5.52	7.13	8.02	6.89	5.69	7.33	8.29	7.10	5.43	7.85	8.34	7.21
DCCRN [35]	5.81	7.39	8.36	7.22	5.91	7.62	8.67	7.41	5.71	8.11	8.59	7.47
TSTNN [41]	6.15	7.58	8.80	7.51	6.23	7.95	8.93	7.70	6.03	8.43	8.84	7.77
MASENet [46]	6.32	7.87	9.23	7.81	6.54	8.24	9.33	8.03	6.35	8.79	9.14	8.09
SADNUNet [47]	6.61	8.24	9.65	8.17	6.86	8.51	9.65	8.34	6.77	9.02	9.77	8.52
MCGN [42]	6.95	8.62	10.01	8.53	7.17	8.47	10.23	8.62	7.19	9.37	10.47	9.01
DBT-Net [51]	7.42	9.27	10.74	9.14	7.53	8.84	10.80	9.05	7.82	10.14	10.90	9.62
TANSCUNet	**7.42**	**9.81**	**11.21**	**9.48**	**7.73**	**10.13**	**11.64**	**10.89**	**7.89**	**10.17**	**11.93**	**10.00**

Table 7 STOI values of baselines under unseen noise condition. Proposed model represented by bold and italic letters

Metric	**STOI**											
Noise	**Train**				**Airport**				**Exhibition hall**			
SNR (dB)	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**
Noisy mixture	53.54	59.14	66.55	59.74	57.57	63.16	68.84	63.19	59.17	65.61	69.28	64.69
Bi-LSTM [31]	68.31	72.47	76.19	72.32	69.25	73.11	78.05	73.47	69.42	74.14	78.53	74.03
Bi-CRN [34]	69.59	73.55	77.45	73.53	70.76	75.53	79.34	75.21	70.14	75.98	80.31	75.48
SEGAN [40]	70.61	75.44	79.75	75.27	71.26	76.81	81.76	76.61	71.43	77.76	81.66	76.95
GRN [30]	71.57	76.51	81.58	76.55	73.39	78.63	82.37	78.13	73.13	79.36	82.22	78.24
DCN [38]	73.25	78.62	83.64	78.50	74.86	79.15	83.19	79.07	75.86	80.04	84.18	80.03
DCCRN [35]	74.62	79.35	84.03	79.33	75.29	80.73	84.54	80.19	77.04	81.23	86.55	81.61
TSTNN [41]	75.31	80.45	85.71	80.49	76.36	81.92	85.35	81.21	78.36	83.15	87.94	83.15
MASENet [46]	77.29	81.04	86.16	81.50	78.29	83.79	87.43	83.17	79.47	84.26	88.23	83.99
SADNUNet [47]	78.14	83.87	87.08	83.03	79.08	84.64	88.02	83.91	80.24	85.23	89.03	84.83
MCGN [42]	79.86	84.73	88.14	84.24	79.64	85.53	89.01	84.72	80.82	86.01	90.42	85.75
DBT-Net [51]	80.41	85.07	89.11	84.86	80.57	86.23	89.38	85.39	81.02	86.73	91.31	86.35
TANSCUNet	**81.91**	**86.59**	**90.95**	**86.48**	**82.61**	**87.85**	**91.06**	**87.17**	**83.14**	**87.97**	**92.21**	**87.77**

4 Conclusion

In this paper, a novel framework has been proposed for single-channel speech enhancement. Several novel strategies were incorporated into the proposed TANSCUNet model very effectively to control information loss and also improve the performance of speech quality and intelligibility. The sub-convolutional encoder and decoder model uses different-sized kernels in each convolutional layer and produces features at various scales. Therefore, it captures the interdependency between local and global contextual information within speech. The multi-kernel achieves 12.03 SDR, 79.65% STOI, and 2.73 PESQ. It indicates that multi-kernel provides significant improvement compared to individual kernel size analysis. The combination of ATFA and AHA blocks in the TAN model is made. Stack of ATFA blocks in TAN effectively extracts global context and highlighted information in temporal and spectral dimensions with the help of MHSA and Bi-GRU layers and also highlighted contextual information is controlled with adaptive factors (α and β). AHA cascades all ATFA block outputs and extracts hierarchical attention information. From the

Table 8 PESQ values of baselines under unseen noise condition. Proposed model represented by bold and italic letters

Metric	**PESQ**											
Noise	**Train**				**Airport**				**Exhibition hall**			
SNR (dB)	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**	**− 5**	**0**	**5**	**Avg.**
Noisy mixture	1.17	1.41	1.74	1.44	1.31	1.67	1.93	1.64	1.64	1.88	2.15	1.89
Bi-LSTM [31]	1.85	2.16	2.49	2.17	1.98	2.26	2.58	2.27	2.18	2.31	2.66	2.38
Bi-CRN [34]	1.94	2.21	2.57	2.24	2.06	2.34	2.64	2.35	2.27	2.44	2.77	2.49
SEGAN [40]	2.02	2.32	2.66	2.33	2.19	2.46	2.75	2.47	2.38	2.57	2.93	2.63
GRN [30]	2.08	2.44	2.71	2.41	2.26	2.55	2.84	2.55	2.47	2.63	3.06	2.72
DCN [38]	2.17	2.59	2.85	2.54	2.34	2.61	2.96	2.64	2.55	2.77	3.14	2.82
DCCRN [35]	2.29	2.65	2.94	2.63	2.42	2.69	3.04	2.72	2.61	2.86	3.20	2.89
TSTNN [41]	2.36	2.71	3.01	2.69	2.51	2.76	3.11	2.79	2.69	2.93	3.24	2.95
MASENet [46]	2.42	2..77	3.08	2.75	2.63	2.84	3.18	2.88	2.76	3.01	3.29	3.02
SADNUNet [47]	2.55	2.83	3.17	2.85	2.71	2.92	3.26	2.96	2.81	3.06	3.33	3.07
MCGN [42]	2.61	2.89	3.21	2.90	2.78	3.04	3.34	3.05	2.85	3.16	3.42	3.14
DBT-Net [51]	2.67	2.92	3.27	2.96	2.82	3.09	3.37	3.09	2.89	3.24	3.49	3.20
TANSCUNet	**2.89**	**3.07**	**3.46**	**3.14**	**2.93**	**3.25**	**3.58**	**3.25**	**3.01**	**3.35**	**3.62**	**3.33**

ablation study, the combination of ATFA and AHA provides significant improvement compared to individual ATFA and AHA block performance. Analyze the effectiveness of the proposed method under unseen speaker conditions, including both seen and unseen noise. The proposed TANSCUNet model achieves under seen noise conditions, on average, 10.52 dB of SDR, 88.23% of STOI, and 3.36 of PESQ. Similarly, under unseen noise conditions, on average, there was 10.12 dB of SDR, 87.14% of STOI, and 3.24 of PESQ. Compared with all baselines, the proposed method's performance is significantly improved in terms of STOI, PESQ, and SDR.

Abbreviations

SE	Speech enhancement
DL	Deep learning
MMSE	Minimum mean square error
DNN	Deep neural network
IBM	Ideal binary mask
IRM	Ideal ratio mask
RNN	Recurrent neural network
LSTM	Long-short-term memory
Bi-LSTM	Bi-directional LSTM
CNN	Convolutional neural network
CED	Convolutional encoder-decoder
MRCE	Multi-resolutional convolutional encoder
T-F	Time-frequency
GRN	Gated residual network
DCN	Dilated convolutions
DRNN	Deep recurrent neural network
TCNN	Temporal convolutional neural network
GRU	Gated recurrent unit
CRN	Convolutional recurrent network
SEA	Squeeze-and-excitation attention
TCA	Temporal convolutional attention
SA	Self attention
TAN	Transformer attention network
SCUNet	Sub-convolutional U-Net
ATFA	Adaptive time-frequency attention
AHA	Adaptive hierarchical attention
FC	Fully connected
SCE	Sub-convolutional encoder
SCD	Sub-convolutional decoder
LN	Layer normalization
LReLU	Leaky rectified linear unit
OLA	Overlap-add method
MHSA	Multi-head self
FN	Feed-forward network
SNR	Signal-to-noise ratio
SDR	Source to distortion ratio
PESQ	Perceptual evaluation of speech quality
STOI	Short-time objective intelligibility
MSE	Mean square error

Acknowledgements
This work is done at a high-performance computing research laboratory at VIT-AP university.

Authors' contributions
Sivaramakrishna Yecchuri: conceptualization, methodology, software, writing—original draft preparation. Sunny Dayal Vanambathina: data curation, validation, supervision.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of data and materials
• NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms. "http://ecs.utdallas.edu/loizou/speech/noizeus/".
• Common Voice. "https://commonvoice.mozilla.org/en".

Declarations

Competing interests
The authors declare no competing interests.

Received: 4 August 2023 Accepted: 22 January 2024
Published online: 03 February 2024

References

1. D. Wang, Deep learning reinvents the hearing aid. IEEE Spectr. **54**(3), 32–37 (2017)
2. P.C. Loizou, *Speech enhancement: theory and practice* (CRC Press, Boca Raton, 2007)
3. S.M. Naqvi, M. Yu, J.A. Chambers, A multimodal approach to blind source separation of moving sources. IEEE J. Sel. Top. Signal Process. **4**(5), 895–910 (2010)
4. Y. Sun, Y. Xian, W. Wang, S.M. Naqvi, Monaural source separation in complex domain with long short-term memory neural network. IEEE J. Sel. Top. Signal Process. **13**(2), 359–369 (2019)
5. B. Rivet, W. Wang, S.M. Naqvi, J.A. Chambers, Audiovisual speech source separation: An overview of key methodologies. IEEE Signal Process. Mag. **31**(3), 125–134 (2014)
6. L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. Albahri, B.S.N. Al-dabbagh, M.A. Fadhel, M. Manoufali, J. Zhang, A.H. Al-Timemy et al., A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. J. Big Data **10**(1), 46 (2023)
7. P.G. Patil, T.H. Jaware, S.P. Patil, R.D. Badgujar, F. Albu, I. Mahariq, B. Al-Sheikh, C. Nayak, Marathi speech intelligibility enhancement using i-ams based neuro-fuzzy classifier approach for hearing aid users. IEEE Access **10**, 123028–123042 (2022)
8. S. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
9. M. Berouti, R. Schwartz, J. Makhoul, in *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Enhancement of speech corrupted by acoustic noise, vol. 4 (IEEE, Washington, DC, 1979), pp. 208–211
10. J.S. Lim, A.V. Oppenheim, Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**(12), 1586–1604 (1979)
11. B.L. Sim, Y.C. Tong, J.S. Chang, C.T. Tan, A parametric formulation of the generalized spectral subtraction method. IEEE Trans. Speech Audio Process. **6**(4), 328–337 (1998)
12. H. Hu, C. Yu, Adaptive noise spectral estimation for spectral subtraction speech enhancement. IET Signal Process. **1**(3), 156–163 (2007)
13. S. Kamath, P. Loizou, et al., in *ICASSP*. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, vol. 4 (Citeseer, 2002), pp. 44164–44164
14. C.W. Wei, C.C. Tsai, Y. FanJiang, T.S. Chang, S.J. Jou, Analysis and implementation of low-power perceptual multiband noise reduction for the hearing aids application. IET Circ. Devices Syst. **8**(6), 516–525 (2014)
15. S.M. Kim, S. Bleeck, An open development platform for auditory real-time signal processing. Speech Commun. **98**, 73–84 (2018)
16. S.M. Kim, Hearing aid speech enhancement using phase difference-controlled dual-microphone generalized sidelobe canceller. IEEE Access **7**, 130663–130671 (2019)
17. S.M. Kim, Auditory device voice activity detection based on statistical likelihood-ratio order statistics. Appl. Sci. **10**(15), 5026 (2020)
18. S.M. Kim, Wearable hearing device spectral enhancement driven by non-negative sparse coding-based residual noise reduction. Sensors **20**(20), 5751 (2020)
19. T. Devis, M. Manuel, A low-complexity 3-level filter bank design for effective restoration of audibility in digital hearing aids. Biomed. Eng. Lett. **10**(4), 593–601 (2020)
20. S. Vellaisamy, E. Elias, Design of hardware-efficient digital hearing aids using non-uniform mdft filter banks. Signal Image Video Process. **12**, 1429–1436 (2018)
21. J. Lim, A. Oppenheim, All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. **26**(3), 197–210 (1978)
22. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
23. Y. Wang, D. Wang, Towards scaling up classification-based speech separation. IEEE Trans. Audio Speech Lang. Process. **21**(7), 1381–1390 (2013)
24. K. Han, Y. Wang, D. Wang, W.S. Woods, I. Merks, T. Zhang, Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(6), 982–992 (2015)
25. M. Tu, X. Zhang, in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Speech enhancement based on deep neural networks with skip connections (IEEE, New Orleans, 2017), pp. 5565–5569
26. S. Rickard, O. Yilmaz, in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. On the approximate w-disjoint orthogonality of speech, vol. 1 (IEEE, Orlando, 2002), pp. I–529
27. Y. Jiang, D. Wang, R. Liu, Z. Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2112–2121 (2014)
28. A. Narayanan, D. Wang, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ideal ratio mask estimation using deep neural networks for robust speech recognition (IEEE, Vancouver, 2013), pp. 7092–7096
29. Y. Xu, J. Du, L.R. Dai, C.H. Lee, A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2014)
30. K. Tan, J. Chen, D. Wang, Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 189–198 (2018)
31. J. Chen, D. Wang, Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. **141**(6), 4705–4714 (2017)
32. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
33. F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J.L. Roux, J.R. Hershey, B. Schuller, in *International conference on latent variable analysis and signal separation*. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr (Springer, Liberec, 2015), pp. 91–99
34. S.R. Park, J. Lee, A fully convolutional neural network for speech enhancement. arXiv preprint arXiv:1609.07132. (2016)
35. Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, L. Xie, Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv preprint arXiv:2008.00264. (2020)
36. E.M. Grais, D. Ward, M.D. Plumbley, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders (IEEE, Rome, 2018), pp. 1577–1581
37. D. Rethage, J. Pons, X. Serra, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A wavenet for speech denoising (IEEE, Calgary, 2018), pp. 5069–5073
38. S. Pirhosseinloo, J.S. Brumberg, in *Interspeech*. Monaural speech enhancement with dilated convolutions. (INTERSPEECH 2019, Graz, 2019), pp. 3143–3147
39. K. Tan, D. Wang, in *Interspeech*. A convolutional recurrent neural network for real-time speech enhancement, vol. 2018 (INTERSPEECH 2019, 2018, Hyderabad, 2018), pp. 3229–3233
40. S. Pascual, A. Bonafonte, J. Serra, Segan: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452. (2017)
41. K. Wang, B. He, W.P. Zhu, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain (IEEE, Toronto, 2021), pp. 7098–7102
42. Y. Xian, Y. Sun, W. Wang, S.M. Naqvi, A multi-scale feature recalibration network for end-to-end single channel speech enhancement. IEEE J. Sel. Top. Signal Process. **15**(1), 143–155 (2020)
43. A. Pandey, D. Wang, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain (IEEE, Brighton, 2019), pp. 6875–6879
44. J. Hu, L. Shen, G. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Squeeze-and-excitation networks (IEEE, Salt Lake City, 2018), pp. 7132–7141
45. S. Woo, J. Park, J. Lee, I.S. Kweon. in *Proceedings of the European Conference on Computer Vision (ECCV)*. CBAM: convolutional block attention module (Springer, Cham, 2018), pp. 3–19
46. X. Xiang, X. Zhang, H. Chen, A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement. IEEE Signal Process. Lett. **28**, 1455–1459 (2021)
47. X. Xiang, X. Zhang, H. Chen, A nested U-net with self-attention and dense connectivity for monaural speech enhancement. IEEE Signal Process. Lett. **29**, 105–109 (2021)
48. 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)

49. Y. Li, Y. Sun, W. Wang, S.M. Naqvi, U-shaped transformer with frequency-band aware attention for speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 1511–1521 (2023)
50. A.L. Maas, A.Y. Hannun, A.Y. Ng, et al., in *Proc. icml*. Rectifier nonlinearities improve neural network acoustic models, vol. 30 (Proceedings of Machine Learning Research, Atlanta, 2013), p. 3
51. G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, H. Wang, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Dual-branch attention-in-attention transformer for single-channel speech enhancement (IEEE, Singapore, 2022), pp. 7847–7851
52. C. Tang, C. Luo, Z. Zhao, W. Xie, W. Zeng, in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. Joint time-frequency and time domain learning for speech enhancement (International Joint Conferences on Artificial Intelligence Organization, 2021), pp. 3816–3822
53. CommonVoice. Mozilla. (2017). https://commonvoice.mozilla.org/en. Accessed 10 Jan 2023
54. P. Loizou, Y. Hu, Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms. Speech Commun. **49**, 588–601 (2017)
55. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
56. Recommendation, ITU-T., Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862. (2001)
57. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
58. D.P. Kingma, Ba Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014)
59. Y. Sivaramakrishna, S. Vanambathina, A nested U-net with efficient channel attention and D3Net for speech enhancement. Circ. Syst. Signal Process. **42**, 4051–4071 (2023)

Publisher's Note