Circuits, Systems, and Signal Processing (2024) 43:5682–5710
https://doi.org/10.1007/s00034-024-02721-2

A Multi-scale Subconvolutional U-Net with Time-Frequency Attention Mechanism for Single Channel Speech Enhancement

Sivaramakrishna Yechuri[1] · Thirupathi Rao Komati[2] · Rama Krishna Yellapragada[3] · Sunnydaya Vanambathina[4]

Received: 18 January 2023 / Revised: 27 April 2024 / Accepted: 1 May 2024 /
Published online: 28 May 2024

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 novel time-frequency attention (TFA) for speech enhancement that includes a multi-scale subconvolutional U-Net (MSCUNet). The TFA extracts valuable channels, frequencies, and time information from the feature sets and improves speech intelligibility and quality. Channel attention is first performed in TFA to learn weights representing the channels' importance in the input feature set, followed by frequency and time attention mechanisms that are performed simultaneously, using learned weights, to capture both frequency and time attention. Additionally, a U-Net based multi-scale subconvolutional encoder-decoder model used different kernel sizes to extract local and contextual features from the noisy speech. The MSCUNet uses a feature calibration block acting as a gating network to control the information flow among the layers. This enables the scaled features to be weighted in order to retain speech and suppress the noise. Additionally, central layers are employed

✉ Sivaramakrishna Yechuri
sivaramakrishna.y@mlrit.ac.in

Thirupathi Rao Komati
tkomati@gitam.edu

Rama Krishna Yellapragada
yramakrishna@kluniversity.in

Sunnydaya Vanambathina
sunny.dayal@vitap.ac.in

[1] Department of Electronics and Communication Engineering, MLR Institute of Technology, Hyderabad, Telangana 500043, India

[2] Department of CSE, GITAM University, Vishakapatnam, Andhra Pradesh 530045, India

[3] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh 522302, India

[4] School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India

to exploit the interdependency among the past, current, and future frames to improve predictions. The experimental results show that the proposed TFAMSCUNet mode outperforms several state-of-the-art methods.

Keywords Speech enhancement · Feature calibration · Time-frequency attention · Multi-scale subconvolutional layers

1 Introduction

The background noise and other residual noises reduce the quality and intelligibility of the speech signal captured in a real acoustic scenario. The goal of speech enhancement (SE) is to recover the intended speech by eliminating distracting ambient noise and noisy speech mixtures. Single channel speech enhancement refers to the scenario, where only a single mix is available, which is an extreme case of the under-determined problem, i.e. the number of sources is greater than the number of mixtures. This problem can be found in many real-world applications, such as mobile communication, automatic speech recognition, and robotics [21, 25, 32, 33, 42].

There are many different techniques that have been proposed for SE. Conventional techniques include statistical techniques based on statistical modeling of spatial, spectral, or temporal properties generated from the sensor signals, such as adaptive wiener filtering [19] and minimal mean square error estimation (MMSE) [5] model. For instance, by modeling the speech and noise spectral components as statistically independent Gaussian random variables, the MMSE estimator accomplishes enhancement.

In terms of enhancing speech, deep neural networks (DNNs) are now thought to be state-of-the-art. In contrast to conventional approaches, DNN-based algorithms [7, 39, 43] seek to learn, through training based on mask or map relationship between the noisy speech and target speech. Then, using either an ideal binary mask (IBM) or an ideal ratio mask (IRM) as the training target, the trained model is utilized to predict the target speech through the T-F mask [14, 26, 31] or mapping [49]. According to recent findings, mapping-based models perform better than masking-based models [38].

Vanilla DNN and Recurrent neural networks (RNNs) have been employed for temporal modeling of speech [1], which is different compared to traditional DNNs. Long short-term memory (LSTM) [10] 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 [44]. The bi-directional LSTM (Bi-LSTM) has been proposed to replace the LSTM. According to earlier findings, it enhances performance under unseen speakers [1, 38]. Bi-LSTM considers the future frames into account and preserves the long-term interdependence between the past, present, and future frames of noisy speech [38].

Convolutional neural networks (CNN) [28] exploitation has been another potential area of SE research. Convolutional encoder-decoder (CED) is proposed to estimate the mapping relationship between the noisy and target speech. Multi-resolution convolutional encoders (MRCE) model has been proposed to improve the SE performance by increasing the receptive fields of the network in Wavenet with extended convolutions and employing a gated mechanism to regulate the information flow among each layer [6, 30]. Furthermore, to increase the receptive fields in the time-frequency (T-F) domain, the gated recurrent network (GRN) approach is applied with 1-D dilated convolutions [38].

In practice, non-causal systems have predictive algorithms such as machine learning and deep learning algorithms. They predict future characteristics based on past input data. Many predictive models are quite complicated to understand and use. Such models are generally used in complex domains such as quantum computing and computational biology to perform longer computations and analyze the complex results as quickly as possible. Machine learning serves as a computational engine for data mining and analytics, where it is used for information extraction, data pattern recognition and prediction.

Convolutional and recurrent models have been combined to enhance SE performance even more. Convolutional recurrent network (CRN) [37] is a combination of the CED and LSTM models, which offers better performance compared to LSTM models. In CRN, the CED is used to locate the T-F patterns and the LSTM is used to record long-term interdependence between the past and present frames.

Predictive analysis uses AI algorithms to assess potential delays and bottlenecks. Taking into account historical data, weather patterns and unforeseen variables, AI predicts potential disruptions and enables proactive adjustments. Key deep learning techniques for predictive maintenance anomaly detection and fine-tuning reduce computing time.

Compared to RNN, GRU has promising characteristics in terms of the balance between fast computation and the ability to map the relationship between time series data sets. Compared with the LSTM network structure, GRU can solve the problem of predicting time series with long interval and long delay. GRU can outperform LSTM units both in terms of convergence in CPU time and in terms of parameter updates and generalization [3].

In [50], a low-dealy SE method is presented that aims to achieve a single-channel speech enhancement with minimal delay and complexity in the time-frequency domain. This method introduces the concept of utilizing the perceptually optimal magnitude spectrum as the training target, thereby enhancing the quality of the speech signal. It is well recognized that dimensionality reduction plays a crucial role in optimizing the computational efficiency and accuracy of machine learning algorithms [40]. By reducing the number of input features, the authors in [8] specifically concentrate on mitigating the computational cost associated with the dimensionality of the problem at hand. Through their research, they explore various techniques and strategies to effectively reduce the dimensionality of the input space, thereby improving the overall performance of the system.

In [27], proposed a temporal convolutional neural network (TCNN) to improve the SE performance in the time domain. TCNN uses a series of 1D causal and dilated convolutions to capture long-range speech context from past and previous frames. In [12], proposed a deep complex convolutional recurrent neural network (DCCRN). It uses a complex convolutional encoder and decoder model that utilizes 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. A multi-scale feature recalibration convolutional bidirectional GRU network (MCGN) [46] model for SE. Local and contextual features can be extracted from the signal using multi-scale recalibration convolutional layers. In the recalibration network, gating is used to control information flow between layers, thus retaining speech and suppressing noise by weighting the rescaled features.

The importance weighting (IW) [17] method proposed by the authors is a model in the form of a U-network containing encoders and decoders. The shared weights between encoder and decoder and source to target are removed. The model is trained with worst-case weights and the loss is minimized using the min-max method. Removing the skip connections between source and target may result in a loss of the original speech information, speech quality and intelligibility.

The authors proposed the U-transformer model [18] in which a frequency band aware attention block (FAT block) is used to train the model. The model assumes 4–8kHz as unvoiced speech and 0–4kHz as voiced speech. The FAT uses three multi-head attentions for time attention, a higher frequency band and a lower frequency band. The FAT focuses only on 0–4kHz, but we cannot estimate the noise content in either the higher or lower frequency bands.

A multi-stage SE framework [13] was proposed in using a multistage structure in which time-frequency attention (TFA) blocks are followed by stacks of squeezed temporal convolutional networks (S-TCN) with exponentially increasing dilation rates. This model is shown to outperform self-attention based temporal convolutional networks and convolutional recurrent network (CRN) baseline models with less computational complexity. The limitation of the above model is its sequential nature, i.e., the performance is highly dependent on its previous results. As a result of such cascaded dependencies, the second stage of the model should be able to correct the estimation error left over from the previous stage.

Some deep learning-based SE methods have also used attention mechanisms to control the computation cost and total parameters. Attention networks optimizing the weights of input features can be accomplished by using a neural attention module to minimize loss. Information can be enhanced and interference from irrelevant information can be reduced in learning-based enhanced frameworks. The squeeze-and-excitation attention (SEA) model has been proposed in [11]. The algorithm utilizes 2D global pooling to compute channel attention and provides impressive performance gains. A convolutional block attention module [45] is proposed, which sequentially improves significant parts of input features through channel attention and spatial attention. MASENet [47] is a combination of convolutional multi-scale and temporal convolutional attention (TCA) models to extract local and global feature information from speech. MASENet encoder block group outputs are recalibrated by the attention block and emphasize informative details. In SADNUNet [48] model, the encoder and decoder model uses nested UNet and dense block to extract local and context features from speech. All encoder group outputs are recalibrated by the self-attention (SA) block, emphasizing informative details and also reducing the unwanted features.

The above-discussed methods are cutting-edge at this time and show promising improvement in SE. But, there are still a number of restrictions in those models. These are fixed kernel (filter) size frequently employed for the CED and CRN approaches. From the noisy speech, the contextual features must be extracted with a bigger size kernel, and the local features can be extracted with a small size kernel. It would be ideal to have a technique that can extract both local and contextual data. To design the causal systems LSTM and CRN models used, which is frequently taken into account only past and present samples of the speech. However, in terms of [30], future data frames including into account will be helpful for the model's prediction results much more. As a result, in our work, the performance of the enhancement is thought to be improved by the non-causal system.

Further, the LSTM models are used on devices with limited resources, and the computational loads required to calculate the input, output, forget gates, and cell memory can occasionally be difficult [4, 10]. It would be preferable to utilize more memory-efficient RNN models, such as gated recurrent unit (GRU)/ Bidirectional GRU (Bi-GRU), whose performance is comparable to that of LSTM/ Bi-LSTM. The properties of various scales are additionally immediately combined and given equal weight in the Inception network [34]. This implies that features be given equal weight, which could be difficult, particularly if the features are brought about by noise.

The limitations of the state-of-the-art models lie in the fact that they only extract important features from the channel-wise perspective and rely on a fixed kernel size to extract both local and global feature sets. This approach fails to consider the diverse scales at which features can exist and leads to a loss of information that significantly impacts the intelligibility and quality of enhanced speech.

To address these drawbacks, the proposed model aims to extract features across various scales and emphasize the importance of features in channels, time, and frequencies in order to minimize information loss.

The motivation behind this proposed model is to overcome the aforementioned limitations by introducing a multi-scale subconvolutional U-Net combined with a time-frequency attention mechanism for speech enhancement.

The novelty of this model lies in its combination of the multi-scale subconvolutional U-Net with the time-frequency attention mechanism (TFA). The multi-scale subconvolutional U-Net employs different kernel sizes to extract both local and global features, thus capturing information at diverse scales. On the other hand, the time-frequency attention mechanism is a combination of spatial and channel attention mechanisms. By incorporating this mechanism, the model is able to reduce information loss by extracting valuable channels, frequencies, and time information from the feature sets, resulting in improved speech intelligibility and quality.The Information loss in TFA is very minimal compared to SEA, TCA, and SA models.

Overall, the proposed model represents a significant advancement in the field of speech enhancement, as it addresses the limitations of existing models and introduces a novel approach that combines the benefits of multi-scale feature extraction and attention mechanisms. The model's ability to extract features in diverse scales and highlight important features in channels, time, and frequencies is crucial in reducing information loss and improving the overall quality of enhanced speech.

Specific contributions of the proposed multi-scale subconvolutional UNet (MSC-UNet) with TFA mechanism for speech enhancement(TFAMSCUNet) model is following:

- Basically MSCUNet is a convolutional encoder-decoder model, which uses different-sized kernels in each convolutional layer to produce features at various scales. This makes it possible to apply a distinct weight to each feature in each scale, making it possible to keep the speech-related components while suppressing the noise-related ones, as well as to capture the interdependence between local and global contextual information within the speech.

- A bottleneck convolutional layer is utilized, which contains a 1-D convolutional layer with kernels of size (1,1) to compress the information flow inside the proposed model.
- TFA is used to extract valuable channels, frequencies, and times from spectrogram images and to improve SE performances. Channel attention is first performed to learn weights representing the channels' importance in the input feature map, followed by frequency and time attention mechanisms that are performed simultaneously, using learned weights, to capture both frequency and time attention.
- A fully connected (FC) network and two Bi-GRU layers are employed in between the MSCUNet encoder and decoder. The FC layer is used to minimize the encoder output dimension. The Bi-GRU layers are capable of capturing the relationships between the past, present, and future frames. It provides similar results compared to Bi-LSTM but it requires only a few parameters.
- Finally, the output layer summed multi-scale outputs and accelerate convergence. Enhanced speech output is estimated by the output layer by providing access to several scales of convolutional operators, which facilitate the training of the network.

The remainder of the paper is organized as follows: Section 2 describes the proposed TFAMSCUNet method. Section 3 describes the experimental result analysis. Section 4 states the conclusions.

2 Proposed Model

2.1 Problem Statement

In the monaural SE model, the noisy speech can be written as:

$$y_t = x_t + n_t \tag{1}$$

where y_t indicates the mixture of noisy speech, x_t and n_t indicate the pure speech signal and additive noise signal respectively. The Short time fourier transform (STFT) employed on noisy speech is defined as:

$$Y_{t,f} = X_{t,f} + N_{t,f} \tag{2}$$

where $X_{t,f}$ and $N_{t,f}$ indicate the pure speech and additive noise in the T-F domain respectively. The proposed model network discovers the relation(M) between the magnitude spectrum of pure speech and the mixture of noisy speech. M estimates the loss function is defined as:

$$
\begin{aligned}
Loss &= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F} M\left(\|Y_{t,f}\| - \|X_{t,f}\|^2\right) \\
&= \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\left(\|\hat{X}_{t,f}\| - \|X_{t,f}\|^2\right)
\end{aligned}
\tag{3}
$$

where the estimated pure speech magnitude spectrum is indicated with $\hat{X}_{t,f}$, the pure speech magnitude spectrum is indicated with $X_{t,f}$. $\hat{X}_{t,f}$, is also having the noisy speech phase values to recover the pure speech.

2.2 Architecture of the Proposed Mode

The proposed TFAMSCUNet model is shown in Fig. 1. A multi-scale sub-convolutional encoder and decoder, central layers, and an output layer comprise the TFAMSCUNet model.

The proposed model's input is the magnitude spectrum of noisy speech $\|Y_{t,f}\|$, which produces enhanced clean speech in the T-F domain $\|S_{t,f}\|$. The multi-scale subconvolutional encoder (MSCE) contains an input layer, a bottleneck layer, a time-frequency attention block, and a multi-scale convolution block. In a multi-scale subconvolution block (MSB), there are seven different sub-convolutions with different kernel sizes to extract the features. A features calibration block (FCB) is followed by an MSB. It allows the network to be selective when utilizing these rescaled features by assigning different weights. Each FCB output is passed through the TFA block. The novel TFA architecture extracts significant frequency, time, and channel information from the input. The multi-scale subconvolutional decoder (MSCD) block is a mirror version of the MSCE block. The output of the MSCE block is fed to the central layer. The central layer contains an FC layer and two Bi-GRU. After processing is completed in the central layer, the output is fed into the MSCD block. Additionally, skip connections are used to improve the flow of information between the MSCE and MSCD blocks. The stride value is (1,2) in all layers of MSCE and MSCEs, except the output layer. The stride value i.e., (1,1) is fixed in the output layer.

2.3 Multi-scale Subconvolutional Block

During CNN training, a high-level feature can be affected 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 [38]. Traditionally, CNNs use a fixed kernel size, which balances the extraction of local and contextual information. A multi-scale subconvolutional block (MSB) addresses this limitation by capturing information on different scales and generating multi-scaled features.

In the top of Fig. 2 shows the MSB architecture. To capture information at varying scales, MSB uses different convolutional operators of different sizes on the encoder side. Small kernel sizes of convolutional operators can capture the local dependency between adjacent T-F points in the short-duration speech. By employing the smallest kernel size (1,2), two adjacent T-F points can be extracted as features. Feature extrac-

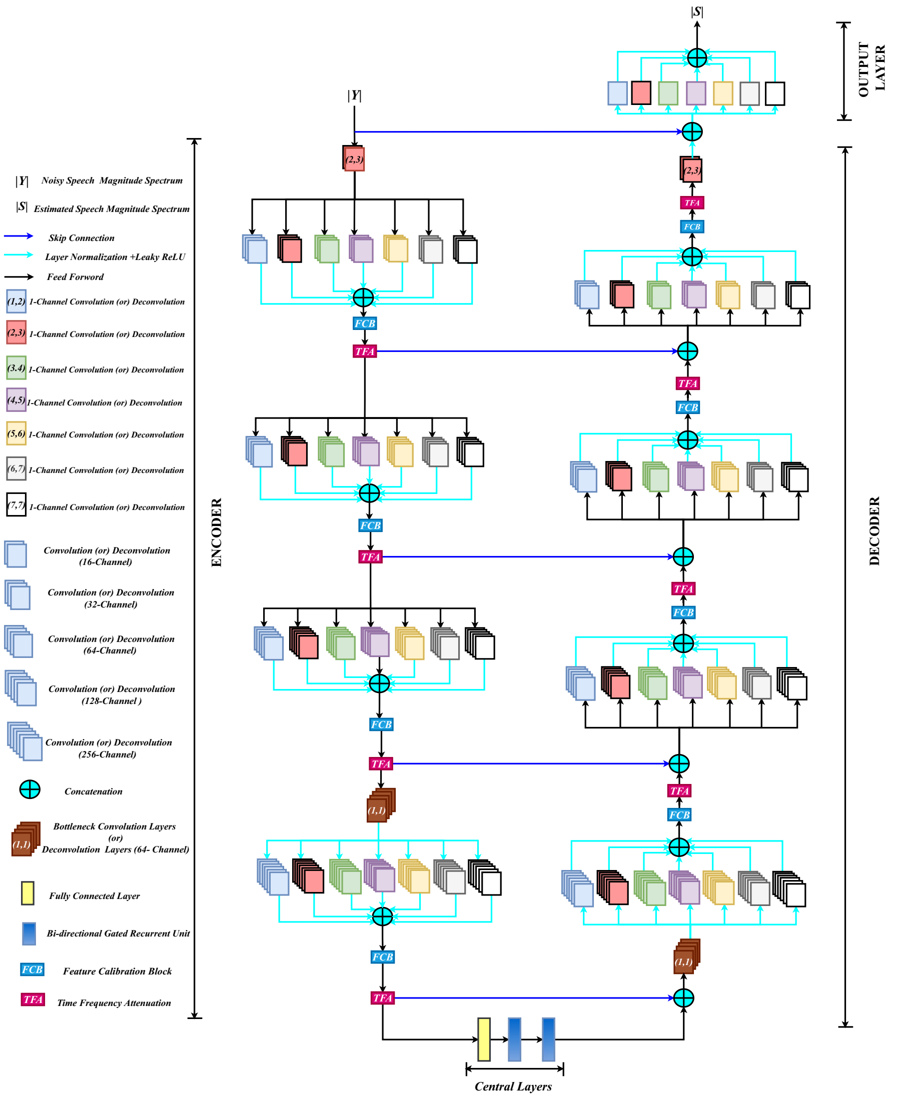
|Y| Noisy Speech Magnitude Spectrum
|S| Estimated Speech Magnitude Spectrum
Skip Connection
Layer Normalization +Leaky ReLU
Feed Forword
(1,2) 1-Channel Convolution (or) Deconvolution
(2,3) 1-Channel Convolution (or) Deconvolution
(3.4) 1-Channel Convolution (or) Deconvolution
(4,5) 1-Channel Convolution (or) Deconvolution
(5,6) 1-Channel Convolution (or) Deconvolution
(6,7) 1-Channel Convolution (or) Deconvolution
(7,7) 1-Channel Convolution (or) Deconvolution
Convolution (or) Deconvolution (16-Channel)
Convolution (or) Deconvolution (32-Channel)
Convolution (or) Deconvolution (64-Channel)
Convolution (or) Deconvolution (128-Channel)
Convolution (or) Deconvolution (256-Channel)
Concatenation
(1,1) Bottleneck Convolution Layers (or) Deconvolution Layers (64- Channel)
Fully Connected Layer
Bi-directional Gated Recurrent Unit
FCB Feature Calibration Block
TFA Time Frequency Attenuation
ENCODER
DECODER
OUTPUT LAYER
|Y|
|S|
(2,3)
FCB
TFA
(1,1)
Central Layers

Fig. 1 Proposed TFAMSCUNet model architecture

tion from the long-duration speech is possible through convolutional operators with large kernel sizes. In comparison to smaller kernels, these features contain contextual information. After each convolutional operator, layer normalization and LReLU [22] operations are performed. Then, as shown in Fig. 2, we concatenate the outputs of each convolutional operation to produce the input for the next steps. The multi-scale deconvolutional block is similar to the MSB, but instead of convolutional operators, it uses deconvolutional ones.

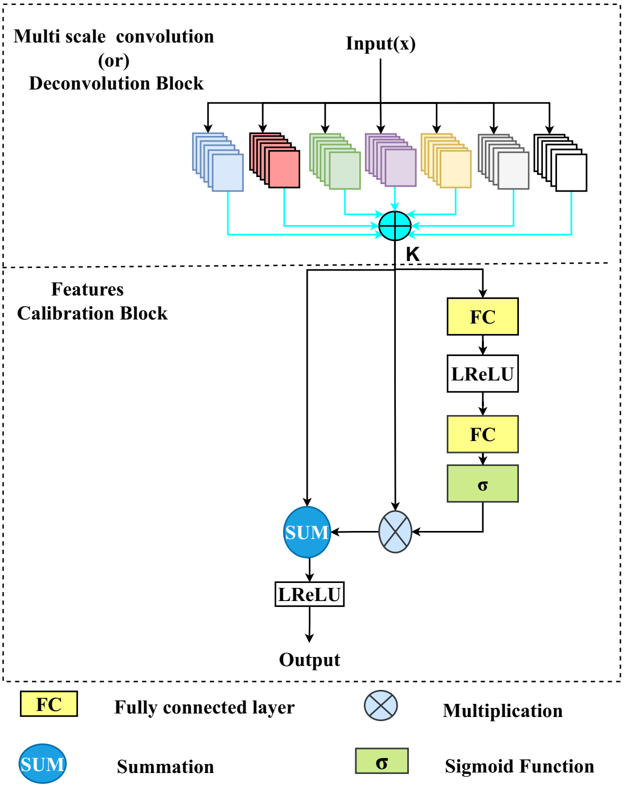

Fig. 2 Architecture of multi-scale convolutional and features calibration blocks

MSB contains m subconvolutional blocks. Each one has the same amount of channels but distinct kernel sizes are used to extract the features. X and K represent MSB input and output, respectively. The output $K = [k_1, k_2, .., k_m]$ represents the mth 2-D subconvolutional block that has different sized kernel.

2.4 Features Calibration Block

At the bottom of Fig. 2, a feature calibration block (FCB) is introduced after the MSB. The calibration coefficients are extracted by using two criteria: determining the nonlinear relationship within the multi-scaled features and assigning a relatively higher weight to speech components and a lower weight to noise components within the feature. In order to meet these criteria, we use FC layers, sigmoid (σ), and LReLU activation units. The work of FCB is

$$FC_{1m} = W_{1m} \odot k_n + B_{1m} \tag{4}$$

$$A_m = Max[0, FC_{1m}] \tag{5}$$

$$FC_{2m} = W_{2m} \odot A_m + B_{2m} \tag{6}$$

$$R_n = \frac{e^{FC_{2m}}}{e^{FC_{2m}} + J} \tag{7}$$

where W and B indicate weight and base parameters. Equations (4) and (6) indicate the fully connected layers FC_{1m} and FC_{2m} operations. In Eq. (7), we divide FC_{2m} element-wise as well as apply the exponential function (e). The calibration coefficient for a mth scaled feature is contained in a vector R_m. The size of J=[1,1,...,1] is the same as FC_{2m}. Based on empirical evidence, the LReLU [24] function is chosen as Eq. (5), used to find the non-negative constraint. Sigmoid is a gating function that is inspired by the success of gating to control the flow of information and assigns different weights for speech and noise components. As a result of the rescaling, the nth feature is as follows:

$$P_m = k_m \odot R_m \tag{8}$$

The rescaled MSB features are indicated with $P = [P_1, P_2, ..., P_m]$. The residual learning [9] takes place inside the FCB layer through a deep skip connection. As a result of residual learning, no additional parameters are introduced. Mathematically, the FCB output is equal to P. By using residual learning and activation, the FCB output is

$$Output = Max[0, K + P] \tag{9}$$

2.5 Time-Frequency Attention

The TFA architecture extracts significant frequency, time, and channel information from the magnitude spectra of input noisy speech. The TFA architecture allows a much higher degree of computing power to be dedicated to this small but critical part of the data. Figure 3 shows the TFA model architecture. TFA uses the output of the FCB rescaled feature set as its input, then refines it using the refined features set. Three sub-modules are included in TFA: the channel attention module (CAM), the frequency attention module (FAM), and the time attention module (TAM). With CAM, general information about channel importance is extracted from an input features set based on the inter-channel relationship between features. The TAM focuses on where the time information is most relevant to the channel attention refined feature map, while the FAM focuses on where the frequency information is most relevant.

The input of TFA is $F \in B^{H \times W \times C}$ which is the output of previous convolutional layers. Here C indicates the channels, W is the width, and H is the height. In TFA, the input features set is first passed through the CAM, which is indicated with M_c. CAM produces the refined channel features set F_c. Then, at a time, TAM and FAM are applied to F_c, which is indicated by M_t and M_f, respectively. After performing the parallel operation, TAM provides the refined time features set F_t and FAM provides the refined frequency features set F_f. With the concatenation operation, both feature sets, F_f and F_t are combined. At the final stage, the concatenated features are passed through a 1 × 1 sized convolutional layer to produce the final refined output features set $\hat{F} \in B^{H \times W \times C}$, which is considered as input to the next layers.

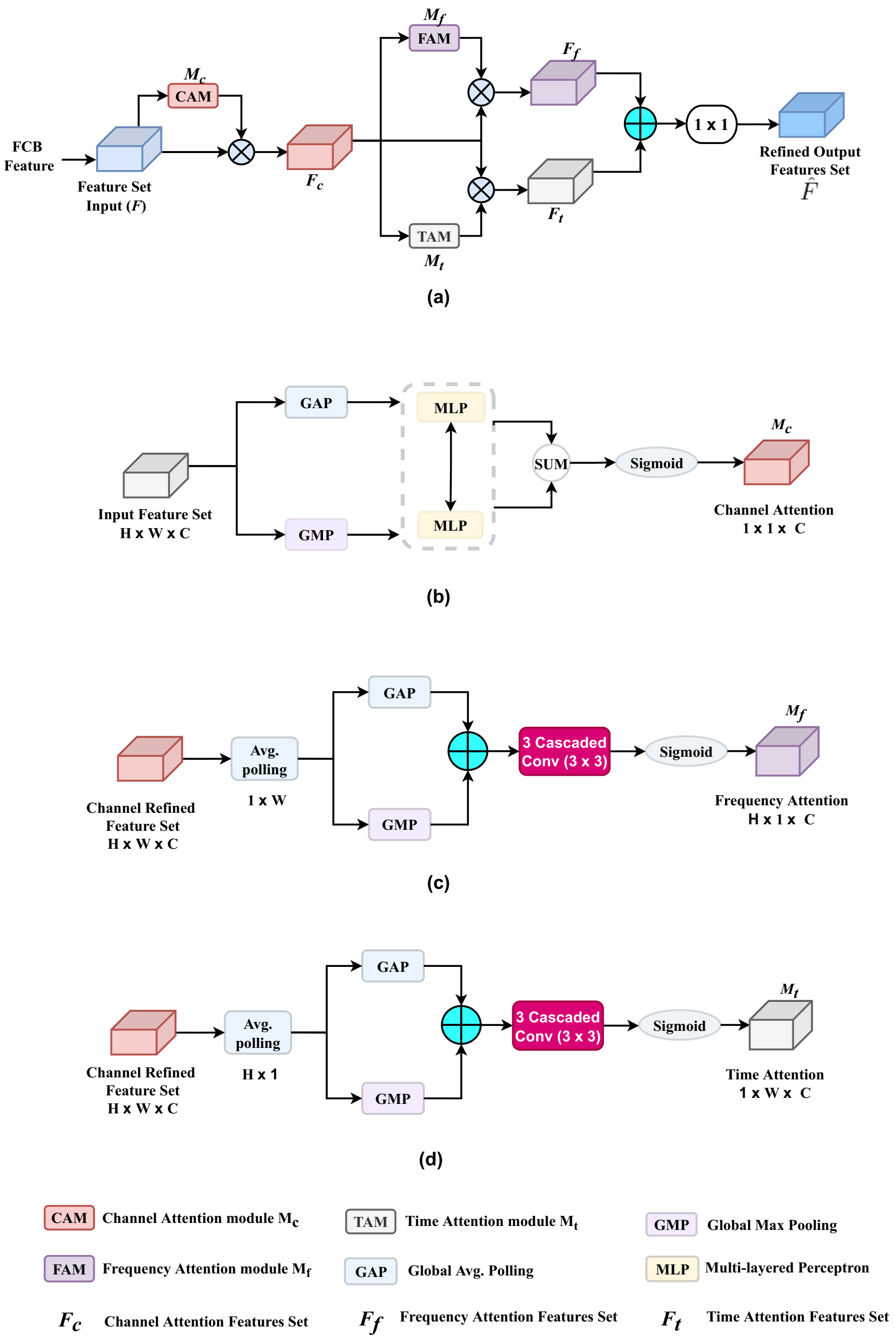

Fig. 3 **a** TFA model architecture. **b** Channel attention module. **c** Frequency attention module. **d** Time attention module

Figure 3 shows the detailed structure of TAM, FAM, and CAM. The global max pooling (GMP) and global average pooling (GAP) operations are used in CAM to generate two distinct feature sets from the input feature set F.The feature sets are fed into a shared multi-layered perceptron (MLP) with $C/8$ and C hidden connected layers. The shared MLP network outputs are added with an element-wise addition operation. In the end, the sigmoid activation function (σ) is used to produce the channel attention features set $M_c(F) \in B^{1\times1\times C}$. By multiplying the M_c by the input features set F, the refined channel features F_c are obtained. The operation of CAM is defined as:

$$M_c(F_t) = \sigma[\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F))] \tag{10}$$

The TAM and FAM have similar structures. TAM concentrates on the input spectra of the time axis, and FAM concentrates on the input spectra of the frequency axis. The refined channel features set is taken as an input to TAM and FAM. In TAM, first, perform average pooling (AP) on F_c along the frequency axis to produce the time features F_t. Next, apply the GAP and GMP to F_f and then concatenate both outputs. Thereafter using cascaded three convolution layers $\mathrm{CNV}_3^{3\times3}$ with kernel size is 3×3 and a sigmoid function (σ) to produce the time attention set features $M_t(F_c)$. By multiplying the M_t by the input features set F, the refined time features F_t are obtained. The operation of TAM is defined as:

$$M_t(F_c) = \sigma[\mathrm{CNV}_3^{3\times3}(\mathrm{GAP}(F_t) + \mathrm{GMP}(F_t))] \tag{11}$$

Unlike TAM, FAM first performs average pooling on F_c along the time axis to produce the frequency features F_f, and then performs the same operations like in TAM to produce the frequency attention feature sets M_f. By multiplying the M_f by the input features set F, the refined frequency features F_f are obtained. The operation of FAM is defined as:

$$M_f(F_c) = \sigma[\mathrm{CNV}_3^{3\times3}(\mathrm{GAP}(F_f) + \mathrm{GMP}(F_f))] \tag{12}$$

Concatenated the refined time (F_t) and frequency (F_f) attention features are passed through a 1 sized convolutional layer to produce the final refined output features set $\hat{F} \in B^{H\times W\times C}$.

2.6 Bottleneck Convolution Layers

As per the practical aspect of multi-scale subconvolutional encoder and decoder blocks, there is a problem with concatenating the multi-scale features, which would increase the dimension of the features and increase the computational cost. As a result, we need a structure that can preserve the information while minimizing complexity (such as dimensions). Based on embedding techniques that provide sufficient information about relatively large patches in low-dimensional embeddings [34, 35], we incorporate bottleneck convolution layers in our TFAMSCUNet model. Following layer normalization and LReLU, the bottleneck convolutional layer utilizes (1,1) kernels and 64 channels. According to Fig. 1, it appears before the last convolutional encoder layer and the first in the decoder layer.

2.7 Central Layers

LSTM is used in the CRN method to track long-term temporal interdependency. According to [20], SE performance is improved by utilizing future frames. As result, Bi-GRUs are used to represent the long-term interdependency between the past, present, and future temporal frames in our work. In comparison, Bi-GRU performs similarly to Bi-LSTM algorithms [2, 4, 15], but has a parameter efficiency advantage. The dimension would inevitably increase as a result of the multi-scaled sub-convolutional blocks being combined. Therefore, a method for keeping the information while also reducing the size and computing cost must be developed. We chose a fully connected (FC) layer to solve this because it consists of fewer parameters than an RNN, thereby reducing the dimension of the FC layer's output compared to the output of the encoder layer.

2.8 Output Layer

As shown in Fig. 1, the skip connection is used to add input to the output layer. From the magnitude of the noisy mixture input and the flow of the previous layer's information, the output layer can predict clean speech. There are seven sub-2D deconvolution layers in the output layer, and each sub-layer has a different kernel size. As a result of concatenating and summing the scaled features, an output matrix with the same size as the input matrix is generated. In this way, the output layer utilizes local and contextual information. In the output layer, the stride size is set to (1,1). Layer normalization and linear activation are followed.

3 Experimental Result Analysis

3.1 Datasets

We use the $Common\ Voice$ [23] corpus to test our system, which is a publicly available voice dataset powered by the voices of volunteer contributors around the world. People who want to build voice applications can use the dataset to train machine learning models. The database contains 1.6 million utterances from 84659 speakers. From that, we select the English corpus and randomly choose 5000 utterances for the training set and 1000 utterances for the validation set, respectively. The test set is also taken from Common Voice, which consists of 1000 utterances. We built training and validation sets using 125 different types of noise and varying signal SNR values from − 5 to + 5 dB. Clean words, noise, and SNR are all chosen at random in each mixed procedure. There are 50,000 training and 4000 validation utterances made in total.

Two test sets are created, one set used for seen noises and the second for unseen noises, to determine the model's noise generalization capability. The seen noises are collected from $NOIZEUS$ [20], consisting of Street, Restaurant, and Babble noises. The unseen noises are Train, Airport, and Exhibition hall noise. Three SNR levels are used to test the noise mixture, i.e., − 5 dB, 0 dB, and 5 dB.

Speech enhancement performance is measured using the following metrics: signal-to-distortion ratio (SDR) [41], perceptual evaluation of speech quality (PESQ) [29], and short-time objective intelligibility (STOI) [36]. 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 Baseline Models

Each utterance is resampled at 16 kHz. In this model, the input is the time frame, with a hop length is 256 and a rectangle window length is 512. Adam[16] optimizer, trained over 50 epochs for TFAMSCUNet, has an initialized learning rate of 0.01. The batch size at the utterance level is set to 16 throughout each epoch.

TFAMSCUNet model performance is compared with eight baseline models: skip connections with DNN (SDNN) [39], MRCE [6], CRN [37], TCNN [27], DCCRN [12], MCGN [46], MASENet [47], SADNUNet [48],TFA-S-TCN [13], U-Transformer + FAT [18], IW-Minimax [17].

3.3 Ablation Study of TFAMSCUNet 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. Here, the U-Net is used as a baseline model to compare the performance of our proposed TFAMSCUNet model. The U-Net is a basic encoder-decoder model, having convolutions and deconvolutions with the same kernel size. Additionally, we increased the number of channels in U-Net to match TFAMSCUNet in terms of parameters, and the remaining configurations are used as before. The SDR, STOI, and PESQ values of U-Net were significantly improved over the noise mixture.

Table 1 Ablation study of the proposed model is shown in terms of averaged SDR, STOI, and PESQ metrics

Model	Trainable parameters (Millions)	SDR (dB)	STOI	PESQ
Noise Mixture	–	3.96	54.21	1.37
UNet (N=4)	4.58	4.31	62.29	1.52
MCBNet	15.2	6.45	70.63	1.77
MCBNet + FCB	20.25	7.74	74.47	1.87
MCBNet + FCB + BNC	15.5	7.58	71.26	1.84
MCBNet + FCB + BNC + CL	21.03	7.81	79.64	1.99
MCBNet + FR + CL + BNC + TFA (Proposed method)	**10.12**	**9.53**	**81.04**	**2.39**

The proposed model is indicated in the BOLD Italic text. N indicated the depth of UNet

Next, we replaced the UNet encoder convolutions and decoder deconvolutions with MCB, which we named MCBNet. The MCB contains seven subconvolutional layers with the same size and different kernel sizes. MCBNet provides a significant improvement compared to UNet, i.e., 2.14 in SDR, 8.34 in STOI, and 0.25 in PESQ. The computational cost is higher because the trainable parameters are much larger than UNet. Next, the FCB is incorporated into MCBNet. FCB acts as a gating function to control the flow of information. The FCB model assigns different weights to speech and noise components, which can help in the reconstruction of clean speech from the noisy mixture. MCBNet+FCB is more effective than MCBNet alone, and it shows performance improvement, i.e., 1.29 in SDR, 3.84 in STOI, and 0.10 in PESQ. FCB increases the trainable parameters a little more, which increases the computational cost.

Next, bottleneck convolution (BNC) layers are incorporated before the last convolutional encoder and the first decoding layer. It compresses the dimensions of features and retains the required information with a small amount of loss. Using BNC, the trainable parameters are reduced, and the computational cost is low. By incorporating BNC, the model's performance is slightly reduced compared to MCBNet+FCB.

Next, the central layers (CL) are incorporated between the MCBNet encoder and decoder. The fully connected layer is used to minimize encoder output dimensions. The bottleneck and FC layers help capture global information from the mixture. The bidirectional GRU layers are capable of capturing the interdependency relationships between the past, present, and future frames. As a result, model performance improved by 0.23 in SDR, 8.38 in STOI, and 0.15 in PESQ.

Next, TFA is incorporated into MSBNet+FCB+BNC+CL. TFA produces the refined feature set from the multi-scale feature set. TFA extracts useful channels, frequencies, and times from the multi-scale feature set and improves performance. In TFA, channel attention is first performed to learn weights representing the channels' importance in the input feature map. This is followed by frequency and time attention mechanisms that are performed simultaneously, using learned weights, to capture both frequency and time attention. By incorporating TFA, the model parameters are reduced to 10.12 million. The model performance also improves significantly, i.e., 1.72 in SDR, 1.40 in STOI, and 0.40 in PESQ.

3.4 Objective Comparison of Baseline Models Under Seen Noises

The SDR, STOI and PESQ results are presented in Tables 2, 3 and 4 for both the baseline and proposed methods in real-world noise. The speakers used in testing are seen in the training data. The noises used in testing include Babble, Street, and Restaurant.

Based on all compared methods, SDNN produces the lowest enhancement performance with an average of 5.06 dB of SDR, 72.72% of STOI, and 1.98 PESQ. As a result, SDNN is still insufficiently effective. The MRCE outperforms the SDNN slightly since it uses a multi-resolution convolutional encoder-decoder and shows a slight increase in SDR, STOI, and PESQ over the SDNN. CRN achieves 5.63 dB SDR, 75.33% STOI, and 2.16 PESQ, all of which are significantly higher than SDNN and MRCE. Due to its ability to capture local spatial patterns, CRN can make use of the

Table 2 SDR values of all baseline models under seen noises

Metric	SDR											
Noise	Babble				Street				Restaurant			
SNR (dB)	– 5	0	5	Avg.	– 5	0	5	Avg.	– 5	0	5	Avg.
Noisy mixture	2.35	4.35	6.35	4.35	3.45	5.35	5.62	4.81	3.11	4.05	5.86	4.34
SDNN [39]	3.41	4.61	6.41	4.81	3.51	6.41	7.41	5.78	3.32	4.61	5.99	4.64
MRCE [6]	3.65	4.95	6.75	5.12	3.63	6.75	7.75	6.04	3.56	4.85	6.19	4.87
CRN [37]	3.89	5.19	7.39	5.49	3.87	7.09	8.09	6.35	3.67	4.99	6.49	5.05
TCNN[27]	4.02	5.35	7.62	5.66	4.12	7.22	8.22	6.52	4.05	5.23	7.26	5.51
DCCRN [12]	4.49	5.59	7.89	5.99	4.57	7.59	8.59	6.92	4.14	6.14	7.69	5.99
MCGN [46]	4.88	6.01	8.3	6.40	5.23	8.1	9.1	7.48	4.47	6.81	8.25	6.51
MASENet [47]	5.03	6.33	8.63	6.66	5.31	8.23	9.23	7.59	4.72	7.43	8.55	6.90
SADNUNet [48]	5.25	6.89	8.99	7.04	5.61	8.45	9.85	7.97	4.99	7.89	8.94	7.27
IW-Minimax [17]	6.13	7.19	9.39	7.57	6.32	8.71	10.36	8.46	5.49	8.44	9.27	7.73
U-Transformer+FAT [18]	6.61	7.87	9.53	8.00	6.91	9.04	10.91	8.95	6.02	8.79	9.51	8.11
TFA-S-TCN [13]	6.94	8.12	10.22	8.43	7.13	9.23	11.27	9.21	6.71	9.16	10.21	8.69
TFAMSCUNet	**7.83**	**9.22**	**12.54**	**9.86**	**7.98**	**9.84**	**12.69**	**10.17**	**7.73**	**9.38**	**12.45**	**9.85**

Proposed model represent in bold

Table 3 STOI values of all baseline models under seen noises

Metric	STOI											
Noise	Babble				Street				Restaurant			
SNR (dB)	– 5	0	5	Avg.	– 5	0	5	Avg.	– 5	0	5	Avg.
Noisy mixture	52.75	62.04	70.55	61.78	60.57	66.47	71.17	66.07	53.28	63.96	72.15	63.13
SDNN [39]	65.79	72.55	77.57	71.97	67.26	70.43	80.14	72.61	68.66	72.35	79.76	73.59
MRCE [6]	66.12	72.94	78.05	72.37	67.96	71.05	80.98	73.33	69.12	73.97	80.23	74.44
CRN [37]	69.09	75.84	79.77	74.9	70.19	73.91	81.68	75.26	71.54	74.87	81.08	75.83
TCNN[27]	71.13	77.54	80.98	76.55	72.96	74.95	83.06	76.99	72.61	76.74	82.64	77.33
DCCRN [12]	72.41	78.16	81.33	77.3	73.39	75.67	83.74	77.6	73.18	77.51	83.19	77.96
MCGN [46]	73.32	79.04	82.15	78.17	74.18	76.24	84.45	78.29	74.06	78.34	83.97	78.79
MASENet [47]	73.94	79.86	82.96	78.92	75.67	77.14	85.03	79.28	74.97	79.15	84.56	79.56
SADNUNet [48]	74.51	80.41	83.34	79.42	76.31	78.17	85.76	80.08	75.38	80.15	85.16	80.23
IW-Minimax [17]	74.81	80.61	83.49	79.63	76.47	78.64	85.93	80.34	75.63	80.47	85.64	80.58
U-Transformer+FAT [18]	75.05	81.24	83.81	80.03	76.71	78.92	86.14	80.59	75.63	80.76	85.91	80.84
TFA-S-TCN [13]	75.23	81.47	84.24	80.22	76.94	79.26	86.57	80.91	76.21	81.14	86.14	81.16
TFAMSCUNet	**76.12**	**82.15**	**85.54**	**81.27**	**78.69**	**81.97**	**86.36**	**82.34**	**77.92**	**82.47**	**86.24**	**82.21**

Proposed model represent in bold

Table 4 PESQ values of all baseline models under seen noises

Metric	PESQ											
Noise	Babble				Street				Restaurant			
SNR (dB)	– 5	0	5	Avg.	– 5	0	5	Avg.	– 5	0	5	Avg.
Noisy mixture	1.17	1.41	1.74	1.44	1.64	1.88	2.15	1.89	1.41	1.67	1.93	1.67
SDNN [39]	1.58	1.99	2.19	1.92	1.93	2.11	2.41	2.15	1.64	1.89	2.14	1.89
MRCE [6]	1.61	2.03	2.27	1.97	2.07	2.24	2.53	2.28	1.69	1.94	2.28	1.97
CRN [37]	1.73	2.12	2.36	2.07	2.18	2.34	2.59	2.37	1.75	2.06	2.34	2.05
TCNN[27]	1.83	2.28	2.55	2.22	2.35	2.53	2.77	2.55	1.88	2.21	2.54	2.21
DCCRN [12]	1.92	2.34	2.67	2.31	2.43	2.61	2.85	2.63	1.96	2.29	2.62	2.29
MCGN [46]	1.99	2.41	2.74	2.38	2.51	2.69	2.93	2.71	2.02	2.36	2.73	2.37
MASENet [47]	2.16	2.52	2.85	2.51	2.59	2.78	3.06	2.81	2.11	2.45	2.85	2.47
SADNUNet [48]	2.27	2.63	2.99	2.63	2.66	2.97	3.19	2.94	2.26	2.64	3.08	2.66
IW-Minimax [17]	2.31	2.65	3.03	2.66	2.68	3.01	3.21	2.96	2.27	2.67	3.10	2.68
U-Transformer+FAT [18]	2.33	2.67	3.04	2.67	2.70	3.04	3.24	2.99	2.99	2.68	3.12	2.69
TFA-S-TCN [13]	2.34	2.68	3.07	2.70	2.71	3.06	3.26	3.01	2.30	2.71	3.13	2.71
TFAMSCUNet	**2.48**	**2.88**	**3.19**	**2.85**	**2.86**	**3.16**	**3.31**	**3.11**	**2.53**	**2.89**	**3.25**	**2.89**

The bold values represent the proposed method values, and it shows the significant improvement over the baselines

T-F structure of magnitude spectra in its input. Additionally, LSTM layers incorporate past and current temporal frames into the CRN to exploit temporal dependency. The limitation of CRN is that it has mor e trainable parameters. Each LSTM requires four linear layers (MLP layer) per cell to run at each time step. Linear layers require large amounts of memory bandwidth to be computed. During the training, LSTM faces the “vanishing gradient” problem.

With dilated convolution layers, the TCNN performs better than the CRN. The TCNN model generates 5.89 dB of SDR, 76.95% of STOI, and 2.32 PESQ values on average. Because it uses a series of 1D causal and dilated convolutions to capture long-range speech contexts from the past. This demonstrates that the TCNN model performs better compared to the SDNN, MRCE, and CRN models. To cover the more receptive area, TCNN requires more 1D-causal and dilated convolutions, which increases computation cost and complexity.

The DCCRN model generates 6.3 dB of SDR, 77.62% of STOI, and 2.41 PESQ values on average. The DCCRN model uses a complex convolutional encoder and decoder model with LSTM and dense layers. With a dense layer, the receptive area will be increased, and the large temporal dependencies will be extracted from the complex encoder-decoder structure. As a result, when compared to all previous models, the DCCRN model outperforms them all. The limitation of DCCRN is that kernel sizes increase exponentially in dense blocks, which can lead to aliasing.

The MCGN model produces an average of 6.69 dB of SDR, 78.41% of STOI, and 2.48 PESQ values. Local and contextual features can be extracted from the signal using multi-scale recalibration convolutional layers. In the recalibration network, gating is used to control information flow between layers, thus improving speech quality. The limitation of MCGN is that it has more trainable parameters (around 77 million), which require large amounts of memory bandwidth to be computed.

MASENet is a combination of convolutional multi-scale and temporal convolutional attention (TCA) models to extract local and global feature information from speech. MASENet encoder block group outputs are recalibrated by the attention block and emphasize informative details. As a result, the model generates 7.05 dB of SDR, 79.25% of STOI, and 2.59 PESQ values on average. The limitation of the model is that more features are reduced based on temporal channel attention, which affects speech intelligibility.

The SADNUNet model generates 7.42 dB of SDR, 79.91% of STOI, and 2.74 PESQ values on average. SADNUNet is a nested UNet encoder-decoder model. Each encoder and decoder use the dense block to extract local and contextual features from speech. All encoder group outputs are recalibrated by the self-attention (SA) block, emphasizing informative details as well as reducing unwanted features. A limitation of this model is that the dense block increases the kernel size exponentially to cover large receptive areas; as a result, the large kernel size increases, leading to aliasing.

The IW method generates 7.92 dB of SDR, 80.18% of STOI, and 2.76 PESQ values on average. The shared weights between encoder and decoder and source to target are removed. The model is trained with worst-case weights and the loss is minimized using the min-max method. Removing the skip connections between source and target may result in a loss of the original speech information, speech quality and intelligibility.

The U-transformer model generates 8.35 dB of SDR, 80.48% of STOI, and 2.78 PESQ values on average. The model assumes 4–8 kHz as unvoiced speech and 0–4 kHz as voiced speech. The FAT uses three multi-head attentions for time attention, a higher frequency band and a lower frequency band. The FAT focuses only on 0–4 kHz, but we cannot estimate the noise content in either the higher or lower frequency bands.

A multi-stage SE framework generates 8.77 dB of SDR, 80.76% of STOI, and 2.80 PESQ values on average. The model using a multistage structure in which time-frequency attention (TFA) blocks are followed by stacks of squeezed temporal convolutional networks (S-TCN) with exponentially increasing dilation rates. This model is shown to outperform self-attention based temporal convolutional networks and convolutional recurrent network (CRN) baseline models with less computational complexity. The limitation of the above model is its sequential nature, i.e., the performance is highly dependent on its previous results. As a result of such cascaded dependencies, the second stage of the model should be able to correct the estimation error left over from the previous stage.

In comparison with the baseline methods, the proposed TFAMSCUNet model achieves, on average, 9.96 dB of SDR, 81.95% of STOI, and 3.05 PESQ, which are 1.2 dB, 1.18%, and 0.25 higher relative to the multi-stage SE framework. The magnitude spectrum of the input is encoded at different scales by TFAMSCUNet. Small kernel sizes of sub-convolutional layers capture local interdependency. Sub-convolutional layers with a large kernel size are employed to determine the interdependency between larger regions. By using small and large kernel sizes, the receptive field of TFAMSCUNet is enlarged, and the different scaled features are assigned different weights. Additionally, TFA extracts useful channels, frequencies, and times from the multi-scale feature set, which produces a refined feature set that improves performance. In TFA, channel attention is first performed to learn weights representing the channels' importance in the input feature set, followed by frequency and time attention mechanisms that are performed simultaneously, with learned weights, to capture both frequency and time attention. As well, central layers are introduced to link the multi-scale encoder and decoder, which can exploit the interdependence between the past, present, and future frames. TFAMSCUNet is also used to learn residual mapping relationships from the raw data.

3.5 Objective Comparison of Baseline Models Under Unseen Noises

The results are presented in Tables 5, 6 and 7 for both the baseline and proposed methods in real-world noise. The speakers used in testing are unseen in the training data. The noises used in testing include Train, Airport, and Exhibition hall.

The skip connection in SDNN boosts enhancement performance over the noisy mixture. The SDNN produces the enhancement performance with an average of 4.91 dB of SDR, 70.44% of STOI, and 2.02 PESQ. As a result, SDNN is still insufficiently effective. The MRCE outperforms the SDNN slightly since it uses a multi-resolution convolutional encoder-decoder and shows a slight increase in SDR, STOI, and PESQ over the SDNN. MRCE's performance is limited by shallow structures and small channel numbers. Moreover, large-sized filters are more expensive to compute. CRN

Table 5 SDR values of all baseline models under seen noises

Metric	SDR											
Noise	Babble				Street				Restaurant			
SNR (dB)	− 5	0	5	Avg.	− 5	0	5	Avg.	− 5	0	5	Avg.
Noisy mixture	1.35	3.15	5.35	3.28	2.54	4.65	6.35	4.51	3.58	5.86	6.51	5.32
SDNN [39]	2.91	3.21	6.41	4.18	2.91	4.71	6.71	4.78	3.91	5.99	7.31	5.74
MRCE [6]	3.45	3.75	6.75	4.65	3.25	4.93	6.98	5.05	4.65	6.35	7.85	6.28
CRN [37]	3.99	4.19	7.39	5.19	3.49	5.37	7.69	5.52	4.89	6.52	8.34	6.58
TCNN[27]	4.52	5.15	7.72	5.80	4.02	5.32	7.92	5.75	5.02	6.85	8.59	6.82
DCCRN [12]	4.59	5.49	8.59	6.22	4.49	5.64	8.39	6.17	5.49	6.97	8.84	7.10
MCGN [46]	4.66	5.73	9.1	6.50	4.61	5.91	8.8	6.44	5.71	7.11	9.14	7.32
MASENet [47]	4.78	5.99	9.23	6.67	4.93	6.43	9.53	6.96	6.03	7.53	9.37	7.64
SADNUNet [48]	5.15	6.19	9.45	6.93	5.24	6.72	9.65	7.20	6.25	7.89	9.67	7.94
IW-Minimax [17]	5.95	7.24	10.51	7.90	6.63	7.59	10.53	8.25	6.79	8.41	10.81	8.67
U-Transformer+FAT [18]	6.45	7.46	10.78	8.23	6.93	7.97	10.93	8.61	7.19	8.73	10.97	8.96
TFA-S-TCN [13]	7.08	8.11	11.02	8.74	7.25	8.43	11.47	9.05	7.68	8.96	11.83	9.49
TFAMSCUNet	**7.63**	**9.04**	**12.05**	**9.57**	**7.79**	**9.64**	**12.59**	**10.01**	**8.13**	**9.88**	**12.85**	**10.29**

Proposed model represents in bold letters

Table 6 STOI values of all baseline models under unseen noises

Metric	STOI											
Noise	Train				Airport				Exhibition hall			
SNR (dB)	– 5	0	5	Avg.	– 5	0	5	Avg.	– 5	0	5	Avg.
Noisy mixture	55.43	63.32	71.45	63.4	58.57	65.34	72.14	65.35	55.17	65.61	74.28	65.02
SDNN [39]	63.57	71.24	75.34	70.05	65.14	69.32	78.24	70.90	64.44	71.14	75.53	70.37
MRCE [6]	66.12	72.94	78.05	72.37	66.35	71.44	79.98	72.59	66.14	72.98	76.31	71.81
CRN [37]	68.78	74.73	79.63	74.38	67.32	74.34	80.28	73.98	67.43	73.76	77.66	72.95
TCNN[27]	70.84	76.43	79.98	75.75	69.06	75.15	81.03	75.08	68.86	75.04	79.66	74.52
DCCRN [12]	71.39	76.96	80.31	76.22	70.39	75.96	81.29	75.88	69.26	76.13	80.18	75.19
MCGN [46]	71.81	77.33	80.89	76.67	71.18	76.24	84.45	77.29	70.04	76.73	81.95	76.24
MASENet [47]	72.24	77.51	81.26	77.00	71.65	76.84	84.93	77.81	70.36	77.15	82.55	76.69
SADNUNet [48]	72.56	77.83	81.56	77.31	71.91	77.26	85.12	78.10	70.77	77.65	82.94	77.12
IW-Minimax [17]	72.81	78.02	81.89	77.57	72.21	77.67	85.45	78.44	71.03	78.13	83.16	77.44
U-Transformer+FAT [18]	72.94	78.24	82.32	77.83	72.58	77.93	85.87	78.79	71.41	78.58	83.34	77.78
TFA-S-TCN [13]	73.04	78.55	82.63	78.07	72.87	78.17	86.19	79.08	72.14	78.79	84.47	78.47
TFAMSCUNet	**74.42**	**81.04**	**84.51**	**79.99**	**74.67**	**80.64**	**87.36**	**80.89**	**74.04**	**80.47**	**87.23**	**80.58**

Proposed model represents in bold letters

Table 7 PESQ values of all baseline models under unseen noises

Metric	PESQ											
Noise	Train				Airport				Exhibition hall			
SNR (dB)	– 5	0	5	Avg.	– 5	0	5	Avg.	– 5	0	5	Avg.
Noisy mixture	1.25	1.46	1.67	1.46	1.64	1.88	1.97	1.83	1.66	1.88	2.01	1.85
SDNN [39]	1.62	1.93	2.06	1.87	1.93	2.11	2.41	2.15	1.88	2.11	2.16	2.05
MRCE [6]	1.71	2.11	2.27	2.03	2.07	2.24	2.53	2.28	1.93	2.21	2.37	2.17
CRN [37]	1.77	2.24	2.35	2.12	2.18	2.34	2.59	2.37	2.05	2.38	2.53	2.32
TCNN[27]	1.89	2.47	2.69	2.35	2.35	2.53	2.77	2.55	2.22	2.63	2.74	2.53
DCCRN [12]	1.95	2.53	2.72	2.40	2.43	2.61	2.85	2.63	2.31	2.79	2.85	2.65
MCGN [46]	2.05	2.59	2.83	2.49	2.51	2.69	2.93	2.71	2.39	2.88	3.08	2.78
MASENet [47]	2.17	2.62	2.99	2.59	2.59	2.78	3.06	2.81	2.44	2.97	3.13	2.85
SADNUNet [48]	2.21	2.64	3.03	2.63	2.66	2.84	3.11	2.87	2.52	3.05	3.19	2.92
IW-Minimax [17]	2.26	2.67	3.14	2.69	2.71	2.89	3.18	2.93	2.58	3.09	3.25	2.97
U-Transformer+FAT [18]	2.31	2.71	3.19	2.74	2.74	2.95	3.22	2.97	2.64	3.13	3.34	3.04
TFA-S-TCN [13]	2.37	2.74	3.22	2.78	2.78	3.01	3.29	3.03	2.67	3.18	3.46	3.10
TFAMSCUNet	**2.62**	**2.91**	**3.59**	**3.04**	**2.95**	**3.39**	**3.68**	**3.34**	**2.89**	**3.47**	**3.87**	**3.41**

The bold values represent the proposed method values, and it shows the significant improvement over the baselines achieves 5.76 dB SDR, 73.77% STOI, and 2.27 PESQ, all of which are significantly higher than SDNN and MRCE. The CRN model has more trainable parameters because it uses LSTM. During the training, LSTM faces the “vanishing gradient” problem.

With dilated convolution layers, the TCNN performs better than the CRN. The TCNN model generates 6.12 dB of SDR, 75.11% of STOI, and 2.47 PESQ values on average. It uses a series of 1D causal and dilated convolutions to capture long-range speech contexts from the past. TCNN model needs more 1D-causal and dilated convolutions to cover large receptive, which leads to computation cost and complexity.

The DCCRN model generates 6.49 dB of SDR, 75.76% of STOI, and 2.56 PESQ values on average. The DCCRN model uses a complex convolutional encoder and decoder model with LSTM and dense layers. As a result, when compared to all previous models, the DCCRN model outperforms them all. The dense block in DCCRN increases the kernel sizes exponentially which leads to increases in the trainable parameters and aliasing effect.

The MCGN model produces an average of 6.75 dB of SDR, 76.78% of STOI, and 2.67 PESQ values. Local and contextual features can be extracted from the signal using multi-scale recalibration convolutional layers. In the recalibration network, gating is used to control information flow between layers, thus improving speech quality. The limitation of MCGN is that it has more trainable parameters (around 77 million), which require large amounts of memory bandwidth to be computed.

MASENet is a combination of convolutional multi-scale and temporal convolutional attention (TCA) models to extract local and global feature information from speech. As a result, the model generates 7.09 dB of SDR, 77.16% of STOI, and 2.75 PESQ values on average. The limitation of the model is that more features are reduced based on temporal channel attention, which affects speech enhancement performance.

The SADNUNet model generates 7.35 dB of SDR, 78.26% of STOI, and 2.80 PESQ values on average. SADNUNet is a nested UNet encoder-decoder model. Each encoder and decoder use the dense block to extract local and contextual features from speech. A limitation of this model is that the dense block increases the kernel size exponentially to cover large receptive areas; as a result, the large kernel size increases, leading to aliasing.

The IW method generates 8.27 dB of SDR, 77.81% of STOI, and 2.86 PESQ values on average. The shared weights between encoder and decoder and source to target are removed. The model is trained with worst-case weights and the loss is minimized using the min-max method. Removing the skip connections between source and target may result in a loss of the original speech information, speech quality and intelligibility.

The U-transformer model generates 8.6 dB of SDR, 78.13% of STOI, and 2.91 PESQ values on average. The model assumes 4–8kHz as unvoiced speech and 0–4 kHz as voiced speech. The FAT uses three multi-head attentions for time attention, a higher frequency band and a lower frequency band. The FAT focuses only on 0–4 kHz, but we cannot estimate the noise content in either the higher or lower frequency bands.

A multi-stage SE framework generates 9.09 dB of SDR, 78.51% of STOI, and 2.97 PESQ values on average. The model using a multistage structure in which time-frequency attention (TFA) blocks are followed by stacks of squeezed temporal convolutional networks (S-TCN) with exponentially increasing dilation rates. This model is shown to outperform self-attention based temporal convolutional networks and convolutional recurrent network (CRN) baseline models with less computational complexity. The limitation of the above model is its sequential nature, i.e., the performance is highly dependent on its previous results. As a result of such cascaded dependencies, the second stage of the model should be able to correct the estimation error left over from the previous stage.

In comparison with the baseline methods, the proposed TFAMSCUNet model achieves, on average, 9.95 dB of SDR, 80.49% of STOI, and 3.26 PESQ, which are 0.86 dB, 1.98%, and 0.29 higher relative to the multi-stage SE framework. The magnitude spectrum of the input is encoded at different scales by TFAMSCUNet. Small kernel sizes of sub-convolutional layers capture local interdependency. Sub-convolutional layers with a large kernel size are employed to determine the interdependency between larger regions. By using small and large kernel sizes, the receptive field of TFAMSCUNet is enlarged, and the different scaled features are assigned different weights. Additionally, TFA extracts useful channels, frequencies, and times from the multi-scale feature set, which produces a refined feature set that improves performance. In TFA, channel attention is first performed to learn weights representing the channels' importance in the input feature set, followed by frequency and time attention mechanisms that are performed simultaneously, with learned weights, to capture both frequency and time attention. As well, central layers are incorporated to link the MSCE and MSCD, which can exploit the interdependence between the past, present, and future frames. TFAMSCUNet is also used to learn residual mapping relationships from the raw data (Tables).

3.6 Convergence Lines of All Models

In Fig. 4, we compare the MSEs of the baseline methods with those of TFAMSCUNet. The results show that TFAMSCUNet converges faster compared to other baseline methods. After 20 epochs of training, the TFAMSCUNet model offers the lowest MSEs. According to this, a TFA with multi-scale sub-convolution features and calibration model representation should improve the algorithm's convergence speed and enhancement performance.

3.7 Multi-kernel Analysis

Our next experiment analyses how performance is affected by kernel size with unseen noises. With kernel sizes ranging from 1×1 to 10×10, the experiments exploit different receptive fields in the T-F domain. Detailed experimental results are presented in Table 8 in terms of SDR, STOI, and PESQ. Performance increases as the kernel size increases, e.g., from 1×2 to 7×7, but then begins to saturate at 11×11. Despite this, there is only a small performance difference (Table 8).

A larger kernel size, such as 7×7, can provide a larger receptive field, which generates the T-F feature map from a larger region (i.e., contextual information), which may be effective in reducing noise. Conversely, a smaller kernel size, like 1×2 captures the feature map from a smaller region (i.e., local information), which is thus effective in maintaining the detailed T-F structure. These results seem consistent with those

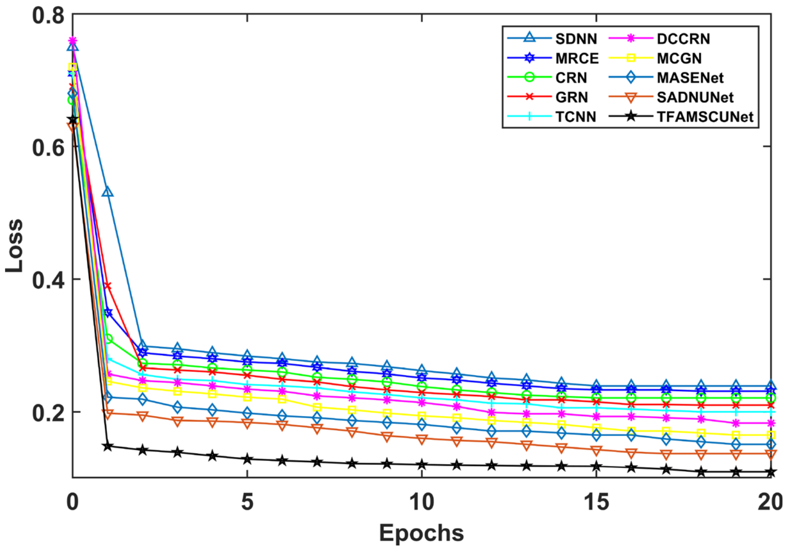

Fig. 4 Convergence plot of all models on the testing set

Table 8 Multi-kernel size analysis

Kernel size	SDR	STOI	PESQ
1 × 2	10.43	71.06	1.59
2 × 2	10.61	71.14	1.61
2 × 3	10.65	71.48	1.62
3 × 4	10.76	71.79	1.63
4 × 5	10.94	72.12	1.65
5 × 6	11.05	72.67	1.67
6 × 7	11.14	72.94	1.69
7 × 7	11.32	73.34	1.75
10 × 10	11.41	73.16	1.73
Multi-Kernel	12.03	75.32	2.01

described in [15, 22]. Contrary to BGRU layers, which capture time dependency (i.e., time domain), 2D-convolutional layers allow for frequency and time expansion.

As shown in Table 8, performance is also dependent on the choice of kernel size. When the kernel size is larger than 7×7, performance may decrease in terms of STOI and PESQ. Parallelization of multi-kernel allows the model to capture features at different scales, thereby exploiting both local and contextual information and, as in our method, increasing performance with unseen noises. The smoothing effect becomes stronger with larger kernel sizes, thereby mitigating noise, whereas smaller kernel sizes preserve finer spectrum structures. With a bank of kernels, the model has a greater probability of capturing and differentiating features from noise and speech, thereby enhancing speech enhancement.

4 Conclusion

In this paper, a novel framework has been proposed for single-channel speech enhancement. Several novel strategies were incorporated into the proposed TFAMSCUNet model to improve the performance of speech enhancement. First, we incorporated MSCUNet. The subconvolutional 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 the speech. Additionally, a feature calibration model is used after each multi-scale sub-convolution block. It acts as a gating function to control the flow of information. The feature calibration model assigns different weights to speech and noise components, which can help in the reconstruction of clean speech from the noisy mixture. Second, we incorporate bottleneck convolutional and deconvolutional layers. It can reduce information flow within the proposed model while retaining information. Thirdly, we incorporated TFA after the feature calibration model. It extracts useful channels, frequencies, and times from the multi-scale feature set and improves performance. IN TFA, Channel attention is first performed to learn weights that represent the importance of the channels in the input feature map, followed by frequency and time attention mechanisms that are performed simultaneously, using learned weights, to capture both frequency and time attention. Fourthly, we incorporate the central layers. The fully connected layer is used to minimize encoder output dimensions. The bi-directional GRU layers are more capable to capture past, present, and future frames relationship. Finally, we incorporated the output layer. The multi-scale outputs were summed to accelerate convergence. For evaluating the efficacy of the proposed method, unseen speakers with seen and unseen noises were used. Compared with state-of-the-art baseline methods, the proposed method's performance is significantly improved.

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

Data Availability The data that support the findings of this study are available in 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

Conflict of interest There is no conflict of interest.

References

1. J. Chen, D. Wang, Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. **141**(6), 4705–4714 (2017)
2. K. Cho, B. Van Merriënboer, C. Gulcehre et al., Learning phrase representations using RNN encoder–decoder for statistical machine translation (2014). arXiv preprint arXiv:1406.1078
3. J. Chung, C. Gulcehre, K. Cho et al., Empirical evaluation of gated recurrent neural networks on sequence modeling (2014a). arXiv preprint arXiv:1412.3555
4. J. Chung, C. Gulcehre, K. Cho et al., Empirical evaluation of gated recurrent neural networks on sequence modeling (2014b). arXiv preprint arXiv:1412.3555

5. 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)
6. E.M. Grais, D. Ward, M.D. Plumbley, Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders, in *2018 26th European Signal Processing Conference (EUSIPCO)* (IEEE, 2018), pp. 1577–1581
7. K. Han, Y. Wang, D. Wang et al., Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(6), 982–992 (2015)
8. C. Haruta, N. Ono, A low-computational DNN-based speech enhancement for hearing aids based on element selection, in *2021 29th European Signal Processing Conference (EUSIPCO)* (IEEE, 2021), pp 1025–1029
9. K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778 (2016)
10. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 7132–7141 (2018)
12. Y. Hu, Y. Liu, S. Lv et al., DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement (2020). arXiv preprint arXiv:2008.00264
13. C. Jannu, S.D. Vanambathina, Multi-stage progressive learning-based speech enhancement using time-frequency attentive squeezed temporal convolutional networks. Circuits Syst. Signal Process. **42**(12), 7467–7493 (2023)
14. Y. Jiang, D. Wang, R. Liu et al., Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2112–2121 (2014)
15. R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in *International Conference on Machine Learning* (PMLR, 2015), pp 2342–2350
16. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
17. Y. Li, Y. Sun, K. Horoshenkov et al., Domain adaptation and autoencoder-based unsupervised speech enhancement. IEEE Trans. Artif. Intell. **3**(1), 43–52 (2022)
18. Y. Li, Y. Sun, W. Wang et al., U-shaped transformer with frequency-band aware attention for speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. (2023). https://doi.org/10.1109/TASLP.2023.3265839
19. J. Lim, A. Oppenheim, All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. **26**(3), 197–210 (1978)
20. P. Loizou, Y. Hu, NOIZEUS: a noisy speech corpus for evaluation of speech enhancement algorithms. Speech Commun. **49**, 588–601 (2017)
21. P.C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007)
22. A.L. Maas, A.Y. Hannun, A.Y. Ng et al., Rectifier nonlinearities improve neural network acoustic models, in *Proc. ICML*, (Atlanta, 2013), p 3
23. Mozilla (2017) Commonvoice. https://commonvoice.mozilla.org/en
24. V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *ICML* (2010)
25. 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)
26. A. Narayanan, D. Wang, Ideal ratio mask estimation using deep neural networks for robust speech recognition, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. (IEEE, 2013), pp.7092–7096
27. A. Pandey, D. Wang, TCNN: temporal convolutional neural network for real-time speech enhancement in the time domain, in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2019), pp.6875–6879
28. S.R. Park, J. Lee, A fully convolutional neural network for speech enhancement (2016). arXiv preprint arXiv:1609.07132
29. Recommendation IT, 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)
30. D. Rethage, J. Pons, X. Serra, A wavenet for speech denoising, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2018), pp.5069–5073
31. S. Rickard, O. Yilmaz, On the approximate w-disjoint orthogonality of speech, in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. (IEEE, 2002), pp I–529

32. B. Rivet, W. Wang, S.M. Naqvi et al., Audiovisual speech source separation: an overview of key methodologies. IEEE Signal Process. Mag. **31**(3), 125–134 (2014)
33. Y. Sun, Y. Xian, W. Wang et al., Monaural source separation in complex domain with long short-term memory neural network. IEEE J. Sel. Top. Signal Process. **13**(2), 359–369 (2019)
34. C. Szegedy, W. Liu, Y. Jia et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 1–9 (2015)
35. C. Szegedy, V. Vanhoucke, S. Ioffe, et al., Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 2818–2826 (2016)
36. C.H. Taal, R.C. Hendriks, R. Heusdens et al., An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
37. K. Tan, D. Wang, A convolutional recurrent neural network for real-time speech enhancement, in *Interspeech*, pp 3229–3233 (2018)
38. 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)
39. M. Tu, X. Zhang, Speech enhancement based on deep neural networks with skip connections, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2017), pp 5565–5569
40. S. Velliangiri, S. Alagumuthukrishnan et al., A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. **165**, 104–111 (2019)
41. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
42. D. Wang, Deep learning reinvents the hearing aid. IEEE Spectr. **54**(3), 32–37 (2017)
43. Y. Wang, D. Wang, Towards scaling up classification-based speech separation. IEEE Trans. Audio Speech Lang. Process. **21**(7), 1381–1390 (2013)
44. F. Weninger, H. Erdogan, S. Watanabe et al., Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR, in *International conference on latent variable analysis and signal separation*. (Springer, 2015), pp. 91–99
45. S. Woo, J. Park, J. Lee et al., CBAM: convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19 (2018)
46. Y. Xian, Y. Sun, W. Wang et al., 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)
47. 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)
48. 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)
49. Y. Xu, J. Du, L.R. Dai et al., A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2014)
50. X. Zhang, X. Ren, X. Zheng et al., Low-delay speech enhancement using perceptually motivated target and loss, in *Interspeech*, pp. 2826–2830 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.