IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 May 2023, accepted 14 June 2023, date of publication 29 June 2023, date of current version 7 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3290908

RESEARCH ARTICLE

NSE-CATNet: Deep Neural Speech Enhancement Using Convolutional Attention Transformer Network

NASIR SALEEM[1,2], TEDDY SURYA GUNAWAN[2,3], (Senior Member, IEEE), MIRA KARTIWI[4], (Member, IEEE), BAMBANG SETIA NUGROHO[3], (Member, IEEE), AND INUNG WIJAYANTO[3], (Member, IEEE)

[1]Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University, Dera Ismail Khan 29050, Pakistan
[2]Electrical and Computer Engineering Department, International Islamic University Malaysia (IIUM), Kuala Lumpur 53100, Malaysia
[3]School of Electrical Engineering, Telkom University, Bandung 40257, Indonesia
[4]Information Systems Department, International Islamic University Malaysia (IIUM), Kuala Lumpur 53100, Malaysia

Corresponding authors: Teddy Surya Gunawan (tsgunawan@iium.edu.my) and Nasir Saleem (nasirsaleem@gu.edu.pk)

ABSTRACT Speech enhancement (SE) is a critical aspect of various speech-processing applications. Recent research in this field focuses on identifying effective ways to capture the long-term contextual dependencies of speech signals to enhance performance. Deep convolutional networks (DCN) using self-attention and the Transformer model have demonstrated competitive results in SE. Transformer models with convolution layers can capture short and long-term temporal sequences by leveraging multi-head self-attention, which allows the model to attend the entire sequence. This study proposes a neural speech enhancement (NSE) using the convolutional encoder-decoder (CED) and convolutional attention Transformer (CAT), named the NSE-CATNet. To effectively process the time-frequency (T-F) distribution of spectral components in speech signals, a T-F attention module is incorporated into the convolutional Transformer model. This module enables the model to explicitly leverage position information and generate a two-dimensional attention map for the time-frequency speech distribution. The performance of the proposed SE is evaluated using objective speech quality and intelligibility metrics on two different datasets, the VoiceBank-DEMAND Corpus and the LibriSpeech dataset. The experimental results indicate that the proposed SE outperformed the competitive baselines in terms of speech enhancement performance at -5dB, 0dB, and 5dB. This suggests that the model is effective at improving the overall quality by 0.704 with VoiceBank-DEMAND and by 0.692 with LibriSpeech. Further, the intelligibility with VoiceBank-DEMAND and LibriSpeech is improved by 11.325% and 11.75% over the noisy speech signals.

INDEX TERMS Neural speech enhancement, T-F attention, convolutional encoder-decoder, convolutional attention transformer, T-F masking.

I. INTRODUCTION

Speech enhancement refers to the process of improving the quality of speech signals by reducing noise and other unwanted distortions. Speech signals can be corrupted by various sources of interference, including background noise, reverberation, and signal distortions caused by recording or transmission processes. SE techniques aim to enhance the quality and intelligibility of speech signals by removing or suppressing these unwanted components while preserving the underlying speech information. SE has applications in many areas, including telecommunications, hearing aids, audio forensics, and speech recognition systems. Conventional SE such as spectral subtraction (SS) [1], [2], Wiener filtering (WF) [3], and statistical methods [4] have been proposed; however, these conventional SE are computationally efficient but fails in dealing nonstationary background noises.

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

Deep learning (DL) has become a popular paradigm for SE [5], [6], [7], addressing the limitations of conventional SE methods [1], [2], [3], [4]. DL-based SE uses models for speech and noise, where the model parameters are estimated through training on speech and/or noise signals. With a hidden layers framework between the input and output layers, DL models can construct complex nonlinear relations and generate feature representations from lower-level input data. Given a dataset of clean-noisy speech pairs, a neural model can learn to transform noisy magnitude spectra into their clean counterparts through mapping-based SE [8], [9], or estimate time-frequency (T-F) masks [10], [11] such as the ideal binary mask (IBM) [12], ideal ratio mask (IRM) [13], Complex ratio mask (cRM) [14], and spectral magnitude mask (SMM) [15] through masking-based SE. In spectral mapping, the model learns a direct mapping rule between the noisy and clean spectral features. However, this can sometimes result in overly smoothed output spectra. On the other hand, spectral masking has been shown to be a more successful learning method, where the gain parameters of the target speech are multiplied by the input noisy magnitude spectrum. This helps to preserve the fine spectral details in the output speech.

A number of DL models, such as feed-forward DNNs (FDNNs) [8], [9], [13], [16], [17], convolutional neural network (CNN) [18], [19], recurrent neural network (RNN) [20], gated recurrent unit (GRU) [21], generative adversarial network (GAN)-based SE [22], [23], a very first deep learning-based SE [7], and conformer-based SE [24], [25], [26] are successfully used for SE. The proposed model differs from the studies [24], [25], [26] since these studies are based purely on the transformer neural networks in the time and frequency domain; however, the proposed SE uses a convolutional transformer as a bottleneck between convolutional encoder-decoder structure. The CAT module differs from the conformer networks [24], [25], [26] since it uses multi-head attention, a TFA attention module, and all the layers are replaced with convolutional layers. One of the main challenges in training RNNs is the vanishing or exploding gradient problem, which can occur during backpropagation through time (BPTT). This problem arises due to the repeated multiplication of gradients over many time steps, which can cause the gradients to either vanish or explode exponentially. When the gradients vanish, the model cannot effectively learn long-term dependencies, and when the gradients explode, the model can become unstable and fail to converge. To address this problem, several variants of RNNs have been developed, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs). These models use gating mechanisms to selectively update and pass information through the network, which helps to mitigate the vanishing and exploding gradient problem and allows the models to learn long-term dependencies more effectively. Another approach to modeling temporal dependencies in speech signals is to use convolutional neural networks (CNNs), which can capture local patterns in the input data and are more computationally efficient than RNNs. Transformer neural network has also been shown to be effective for modeling long-term temporal dependencies in speech signals. The key innovation of the Transformer is the self-attention mechanism, which allows the model to attend to different parts of the input sequence during each layer of computation. This attention mechanism enables the model to capture long-term dependencies in the input sequence without the need for recurrent connections. One advantage of the Transformer over RNN-based models is that it can process input sequences in parallel, making it more computationally efficient for long input sequences.

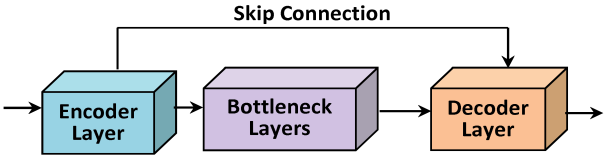

FIGURE 1. A typical Convolution Encoder-Decoder (CED) framework with a bottleneck layer.

In the field of speech enhancement, deep neural networks are used as a supervised learning problem to enhance noisy speech. Two types of approaches exist in this context: time-frequency (T-F) domain and time-domain approaches. Direct regression, a time-domain approach, involves learning a regression function directly from the waveform of a speech-noise mixture to the target speech, without using an explicit signal front-end. This approach typically involves the use of 1-D convolutional neural networks (Conv1d). A convolution encoder-decoder (CED) or a U-Net framework resembles the short-time Fourier transform (STFT) and it is inverse (iSTFT), which are common signal processing techniques used in SE. The enhancement network is then inserted between the encoder and decoder, and this can be accomplished using networks with temporal modeling capacity, such as temporal convolutional networks (TCNs), LSTM, or GRU networks. The Time-Frequency domain works on the spectrogram with an assumption that fine-detailed structures of speech and noise can be effectively discriminated using T-F representations following STFT. Convolution recurrent network (CRN) is a recent approach that employs a CED structure similar to the time-domain approaches however extracts high-level features enabling improved separation of noise from speech spectrograms using 2-D CNN (Conv2d).

Several convolution encoder-decoder (CED) frameworks have been proposed that use a bottleneck layer, typically the LSTM or GRU networks, to model the temporal dependencies in the speech signal. A typical CED framework with a bottleneck is demonstrated in Fig 1. Originally, the CED architecture with two LSTM layers is proposed by Tan and Wang [27]. Two LSTM layers are used to capture long-term dependencies. By incorporating convolutional and recurrent layers, a study [28] proposed the recurrent convolutional encoder-decoder (R-CED) network to model

time and frequency domains of the speech signal while capturing long-term dependencies for enhancing noisy speech. In a CED architecture [29], the fully-connected LSTM is replaced with convolutional LSTM (ConvLSTM). A multi-scale CED framework with two BiLSTM layers is proposed by Yang [30]. In a CED frame, the study [31] proposed a convolutional fusion network (CFN) and incorporated a group convolutional fusion unit (GCFU) into CED. An augmented Kalman filtering (AKF) is added to the CED framework for SE [32] where the network estimates the instantaneous noise spectrum for determining the linear prediction coefficients (LPCs) of noise. An LSTM-Convolutional-BLSTM Encoder-Decoder network for SE is proposed in the CED framework [33] to balance the complexity of the model and to improve the model capacity to capture T-F features. The study [34] proposed a gated convolutional recurrent network (GCRN) in the CED framework for complex spectral mapping. A temporal convolutional module (TCM) is added between the encoder-decoder for time-domain SE [35]. An end-to-end (E2E) CED model is proposed for SE where RNNs are integrated between the encoder-decoder structure [36]. A progressive Learning-based CED framework is proposed for SE [37] where two LSTM layers are added to capture temporal dependencies. A study [38] proposed a CED framework for noise-independent and speaker-independent SE in complex spectral mapping-domain. A technique is incorporated to reduce the trainable parameters and the computational load. An E2E WaveCRN model [39] proposed CED where a CNN captured the local features and simple recurrent units (SRU) modeled the sequential properties of the local features. An extension of WaveCRN [40] is proposed in [41] where LSTM/GRU/SRU are added between the CED framework. A Convolutional Recurrent U-net for Speech Enhancement (CRUSE) model for SE is proposed where parallel GRUs are added to the CED firework. A DCCRN [42] is proposed for SE which added complex LSTM into the CED. Li et al. [65], [66] showed the importance of using compressed complex spectrum as the input feature and using it as the training target where temporal convolutional modules (TCM) are used as a bottleneck in [66]. Table 1 summarises different CED approaches for SE.

Unlike the CEDs in Table 1 where bottleneck layers are mostly the RNNs [27], [34], [37], this study incorporates temporal modeling in the convolutional layers by using a multi-head attention (MHA) module. The bottleneck layers in this study are motivated by the success of the transformer models for speech processing. The bottleneck block is composed of convolution layers with 1-D kernels and MHA modules. Furthermore, a T-F attention with time-frame attention and frequency-channel attention is applied to generate a 2-D attention map to quantify the important T-F speech distributions. The T-F attention module is incorporated into the bottleneck block for effective speech enhancement. The performance of the proposed SE is evaluated using objective speech quality and intelligibility metrics on two different datasets, the VoiceBank-DEMAND Corpus [43] and the LibriSpeech dataset [44]. The contributions of this study are summarized as:

- A neural SE (NSE) is proposed by using a convolutional encoder-decoder (CED) framework and convolutional attention transformer (CAT), named the NSE-CATNet.
- Unlike conventional bottlenecks for temporal modeling in CEDs, this study uses a bottleneck composed of convolutional layers-based temporal modeling and a multi-head attention module.
- A T-F attention with time-frame and frequency-channel attention is applied to generate an attention map for quantifying the important T-F speech distributions in the estimated T-F Mask.

The remainder of the paper is organized as follows. Section II presents the proposed SE based on the CED with the CAT module. Section III presents experiments and settings. Results and discussions are presented in Section IV. Finally, Section 6 concludes this study.

TABLE 1. Summary of the different CED approaches for SE.

Ref#	Year	Domain	Bottleneck
[27]	2018	Spectral Mapping	LSTM
[28]	2021	Spectral Mapping	RNN
[29]	2020	Spectral Mapping	Conv-LSTM
[30]	2021	Spectral Mapping	BiLSTM
[31]	2021	Spectral Mapping	GCFU
[32]	2020	Spectral Mapping	AKF
[33]	2021	Spectral Mapping	BiLSTM-LSTM
[34]	2019	Complex Spectral Mapping	GLU
[35]	2019	Time-Domain Mapping	TCN
[36]	2018	Spectral Mapping	Bi-RNN
[37]	2020	Spectral Masking	LSTM
[38]	2019	Complex Spectral Mapping	Grouped-LSTM
[39]	2020	Time Domain Mapping	BiSRU
[40]	2020	Time-Domain Mapping	GRU/SRU/LSTM
[41]	2021	Spectral Masking	Parallel-GRU
[42]	2020	Complex Spectral Mapping	Complex LSTM
[65]	2021	Compressed Complex Spectrum	–
[66]	2022	Complex Spectral Mapping	TCM

II. PROPOSED NEURAL SPEECH ENHANCEMENT

A. PROBLEM FORMULATION: SPEECH ENHANCEMENT IN STFT DOMAIN

A noisy speech signal $y(n)$ can be modeled as a combination of the underlying clean speech signal $s(n)$ and the background noise signal $v(n)$. This mixture can be represented mathematically as follows:

$$y(n) = s(n) + v(n) \tag{1}$$

The goal of SE is to estimate the clean speech signal $s(n)$ from the observed mixture $y(n)$ while minimizing the distortion caused by the noise signals $v(n)$. Taking the short-time Fourier transform (STFT) of both sides of equation $y(n) = s(n) + v(n)$, we obtain:

$$Y_{m,k} = S_{m,k} + V_{m,k} \tag{2}$$

where $Y_{k,m}$, $S_{k,m}$, and $V_{k,m}$ are the STFT coefficients of $y(n)$, $s(n)$, and $v(n)$, respectively, at frequency bin k and time frame m. This can be accomplished by estimating a spectral-masking $M^{IRM}_{m,k}$ or $M^{SSM}_{m,k}$ where estimated mask $\hat{M}^{IRM}_{m,k}$ or $\hat{M}^{SSM}_{m,k}$ is multiplied with magnitude of the observed noisy mixture. The resulting enhanced speech signal can then be transformed back to the time domain, using inverse STFT (ISTFT) and noisy phase.

B. PROPOSED SE

The proposed SE is depicted in Fig. 2 where the encoder in the CED framework is designed to extract high-level features from the input speech signal by applying convolutional and pooling operations. The decoder, on the other hand, performs the inverse operations of the encoder in a symmetric manner. It maps the low-level features produced by the encoder back to the original input size by applying deconvolutional (or transposed convolutional) and upsampling operations. These operations help to reconstruct the input signal by gradually increasing its resolution and complexity. The symmetric structure of the CED framework ensures that the shape of the inputs and outputs are preserved, which is important for tasks such as speech enhancement where the original input signal needs to be reconstructed as accurately as possible. Additionally, by using a symmetric structure, the network can be trained end-to-end, which allows the model to learn to extract and reconstruct features in an optimized way.

C. CED FRAMEWORK

The CED model consists of five convolutional (Conv2D) and deconvolutional (Deconv2D) layers that form the encoder-decoder network. The use of exponential linear rectified unit (ELU) activation and batch normalization (BN) helps in achieving better convergence and generalization, respectively, while also reducing model complexity. The final layer uses a softplus activation function, which ensures that the output of the network is always positive. To improve the context that the model considers, a stride of 2 is applied along the frequency direction for all convolutional and deconvolutional layers. At the same time, the time dimension of the features remains the same. This allows the network to leverage larger context information, which can be particularly useful for speech enhancement. Skip connections are added to improve the flow of gradients and information through the network. These connections connect the output of the encoder to the input of the decoder, allowing the network to bypass certain layers and more directly connect the low-level features learned by the encoder to the high-level features produced by the decoder. A detailed structure of CED for encoder-decoder is depicted in Fig. 3, where F_{in} and F_{out} represent input and output feature maps of the encoder.

D. BOTTLENECK: CONVOLUTIONAL ATTENTION TRANSFORMER (CAT)

The success of the transformer and its variants [45] for speech processing has inspired the design of bottleneck layers, which are a type of neural network layer that can be used to reduce the dimensionality of the input signal while preserving its important features. The transformer architecture has proven to be very effective for processing sequential data, such as speech signals. However, the transformer architecture is computationally intensive and requires a large number of parameters, which can make it challenging to deploy in real-time or low-resource scenarios.

In this paper, the bottleneck layers are designed to address these challenges by reducing the dimensionality of the input signal before feeding it to the transformer layers. These layers typically use a convolutional neural network (CNN) to extract high-level features. This allows the model to operate on a lower-dimensional representation of the input signal, reducing the computational cost and the number of parameters required. Convolution layers with 1-D kernels and multi-head attention (MHA) modules are used as an alternative approach to RNN-based bottlenecks. Convolutional layers with 1-D kernels capture local patterns in sequential data (speech signals). The MHA module can capture long-range dependencies in speech signals. They work by attending to different parts of the input sequence and computing a weighted sum of the input vectors. By using multiple attention heads, the model can attend to different aspects of the input and capture complex patterns in the data. By combining convolutional layers and MHA modules in a block, we can capture both local and global patterns in speech signals. While RNN-based bottlenecks have been popular in the past, convolutional layers and attention modules offer a viable alternative that can achieve comparable or better performance while being computationally efficient.

Figure 4 shows the structure of the convolutional Transformer bottleneck. The input to the block is first processed by a convolutional layer, which applies a set of 1-D kernels to the input sequence to extract local features. The output of the convolutional layer is then passed through a PReLU activation function, which introduces non-linearity to the model and helps to address the vanishing gradient problem. Layer normalization is then applied to the output of the PReLU activation function. Layer normalization normalizes the values of each feature independently, allowing the model to learn more efficiently and generalize better. The intermediate results are then fed to a multi-head attention module, which captures long-range dependencies in the input sequence by attending to different parts of the sequence. The output of the attention module is then normalized using layer normalization. In a Transformer, the context vectors of the input features map are encoded as a set of key-value pairs (K, V) with a dimension similar to the input sequence length. The output at the previous timestep is computed into a query Q, and the next term in the output sequence of the decoder is a

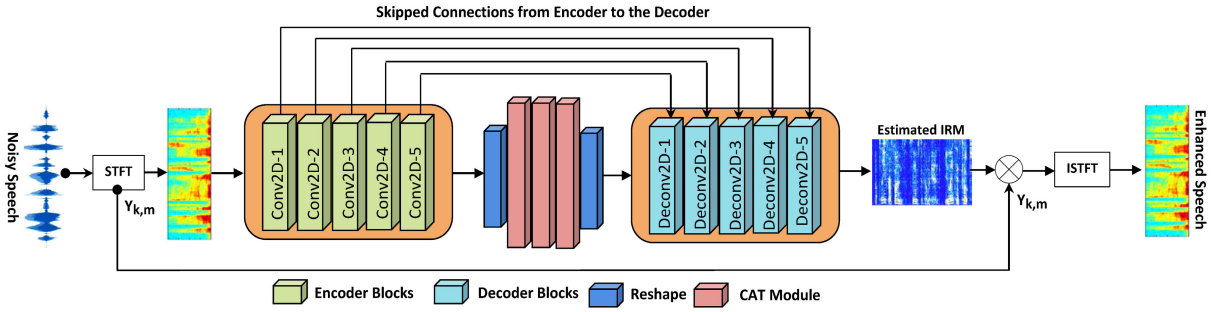

FIGURE 2. Proposed NSE-CATNet speech enhancement framework.

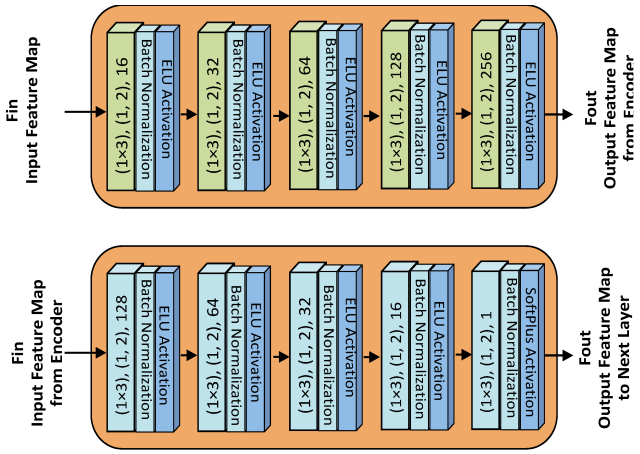

FIGURE 3. A detailed structure of the Encoder (upper panel)-Decoder (bottom panel).

mapping from the K,V pairs with Q as (Q,K,V). The outputs of the decoder are the weighted sum of all values from the (K,V) encoded representation of the inputs. The MHA in a transformer assigns the alignment weights to each hidden state as a sequence-length-scaled dot-product of the query with all the keys, given as:

$$Attention_h(Q_h, K_h, V_h) = softmax(\frac{Q_h K_h^T}{\sqrt{d_k}})V_h \quad (3)$$

The scaled dot product is scaled by the dimension h of hidden states for sequence output at timestep t. The layer operates on the encoded latent space regardless of the number of attention heads, and a softmax is computed from a weighted sum of all layers in the bottleneck architecture (demonstrated in Fig. 4), given as:

$$MH(Q, K, V) = Concat(A_1, A_2, \ldots, A_H]W^0 \quad (4)$$

$$Head_m = Attention(QW_h^Q, QK_h^K, QV_h^V) \quad (5)$$

where QW_h^Q, QK_h^K, QV_h^V are the learnable parameters. Four identical bottleneck layers are applied in this study.

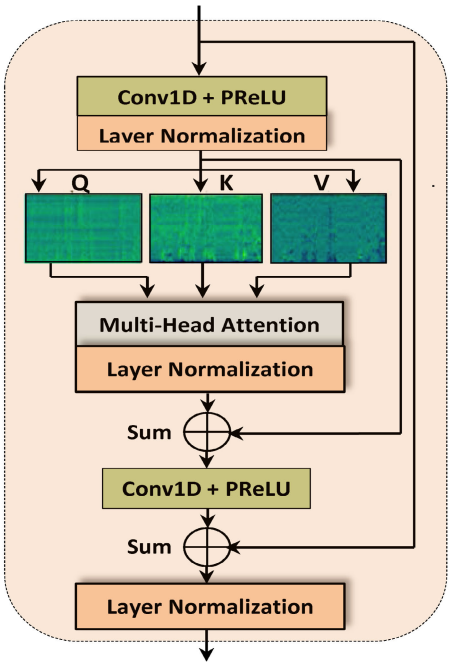

FIGURE 4. The bottleneck structure.

E. TIME-FREQUENCY ATTENTION INTO CAT MODULE

The attention process has been extensively studied in the field of speech processing. The studies [46], [47] have investigated attention mechanisms for modeling speech distribution along the frequency and time dimensions to demonstrate its effectiveness. Time-frequency attention (TFA) [47] functional neural module is incorporated into the convolutional transformer bottleneck, named as CATNet. The TFA module is composed of time-dimension attention (TDA) and frequency-dimension attention (FDA) to create 1-D attention maps such that the model focuses on the time-frames and frequency-wise channels. The TDA creates a 1-D attention map $TD_A \in \mathbb{R}^{(1\times L)}$ whereas FDA creates a 1-D attention map $FD_A \in \mathbb{R}^{(d_{model}\times 1)}$. After creating the 1-D attention maps, the TDA and FDA are infused to create a final 2-D attention map $TF_A \in \mathbb{R}^{(L\times d_{model})}$, thereby assigning labeled attention

weights to all Time-frequency spectral components. This allows the neural network to grasp the distributions of speech signals in the time-frequency representation.

The time-frame index and frequency-channel index of the speech signals determine their distributions across the time-frequency plane. The TDA provides time-frame-wise statistics $B_T \in \mathbb{R}^{(1\times L)}$ by performing global average pooling along the frequency dimension on the given input Y:

$$B_T(m) = \frac{1}{d_{model}}\left(\sum_{k=1}^{d_{model}} \mathbf{Y}_{m,k}\right) \tag{6}$$

where $B_T(m)$ shows the m^{th} element of B_T. On the other hand, FDA provides frequency-wise statistics $B_F \in \mathbb{R}^{(d_{model}\times 1)}$ by performing global average pooling along the time-frame dimension on the given input Y:

$$B_F(k) = \frac{1}{L}\left(\sum_{l=1}^{L} \mathbf{Y}_{m,k}\right) \tag{7}$$

The final 2-D time-frequency attention map is given as:

$$TFA_{m,k} = TDA_m \times FDA_k \tag{8}$$

The output of the TFA module $\tilde{Y}$ is given as:

$$\tilde{Y} = Y \odot TFA \tag{9}$$

where $\odot$ is the element-wise multiplication operator.

The TFA module is integrated into the convolutional transformer bottleneck, as depicted in Fig. 5. For the intermediate latent time-frequency tensor $\mathbf{Z} \in \mathbb{R}^{L\times d_{model}}$ as input, the bottleneck projects tensor $\mathbf{Z}$ to the query $\mathbf{Q} \in \mathbb{R}^{L\times d_{model}}$, key $\mathbf{K} \in \mathbb{R}^{L\times d_{model}}$, and value $\mathbf{V} \in \mathbb{R}^{L\times d_{model}}$ such that:

$$\mathbf{Q} = \mathbf{U}\mathbf{W}^{Q} \tag{10}$$

$$\mathbf{K} = \mathbf{U}\mathbf{W}^{K} \tag{11}$$

$$\mathbf{V} = \mathbf{U}\mathbf{W}^{V} \tag{12}$$

where $\{\mathbf{W}^{Q}, \mathbf{W}^{K}, \mathbf{W}^{V}\} \in \mathbb{R}^{d_{model}\times d_{model}}$ show different learnable projections, respectively. These projections are segmented into H attention heads $h = \{1, 2, 3, \ldots, H\}$ as d_k, d_q, d_v dimensions. This allows the model to focus on various elements of information. To create the outputs, the scaled dot-product attention is applied to each head in parallel, as in Eq. (3). The outputs of all attention heads are aggregated and projected linearly to produce the output of the bottleneck module, as in Eq. (4), where $W^0 \in \mathbb{R}^{d_{model}\times d_{model}}$. The time-frequency attention module receives outputs from the prior bottleneck and performs a time-frequency attention operation to update the model to pay attention to the spectral components. For additional detailed descriptions of the TFA module, we refer the reader to the study [46], [47].

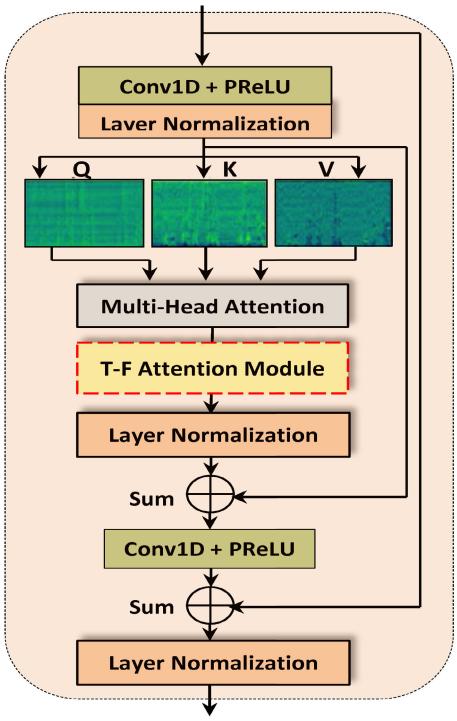

FIGURE 5. The bottleneck structure with time-frequency attention.

III. EXPERIMENTS AND SETUP

A. DATASETS

This section first describes the clean speech and noise data. This study uses *train-clean-100* training set from the LibriSPeech dataset [44] containing 28539 speech sentences uttered by 251 speakers. LibriSpeech is a large-scale corpus of read English speech that was created by a collaboration between the University of Maryland, the University of Edinburgh, and the Karlsruhe Institute of Technology. The corpus consists of approximately 1,000 hours of speech data sampled at 16kHz, which was extracted from audiobooks from the LibriVox project (https://www.openslr.org/12). In addition, the VoiceBank-DEMAND corpus [43] is used to train the proposed model. The VoiceBank-DEMAND corpus is a dataset of clean speech and noise that was created for research in speech processing and enhancement. It consists of 10 speakers, each providing approximately 4 hours of speech, for a total of 40 hours of speech data. In addition to the clean speech and noise data, the VoiceBank-DEMAND corpus also includes a set of artificially mixed speech and noise signals, which were created by adding the noise signals to the clean speech signals at various signal-to-noise ratios (SNRs). However, this study uses only clean sentences from the dataset. The noises in the training set are taken from the QUT-NOISE dataset [48], the Nonspeech dataset [49], and the RSG-10 dataset [50]. The noise duration over 30 seconds is divided into 30 seconds segments. To create noisy sentences in the training, the clean speech sentences are mixed with noises randomly at SNRs between −10dB and 10dB with a 5dB incremental step. For model evaluation, this study adopts the voice babble and factory noise from the NOISEX-92 dataset [51] whereas street noise from the Urban Sound dataset [52]. A colored noise source that is F16 is selected from the RSG-10 noise dataset [50]. For each noise source, the clean speech sentences are selected randomly from the test-clean-100

of LibriSpeech and are mixed with the noises at SNR levels −10dB, −5dB, 0dB, 5dB, and 10dB, respectively. In addition, two noise sources (cafeteria and factory2) are selected as unseen noises from the DEMAND dataset [43].

B. TRAINING METHODOLOGY

This section describes the detailed training methodology. A mini-batch of 32 samples is used for training iterations. Each selected clean speech sentence for the mini-batch is mixed with selected noise at SNRs. For the two masking-based training objectives including IRM and SMM, this study adopts the mask approximation, where the mean-square error (MSE) is the loss function. The two training objectives and mask approximation-based MSE loss are given as;

$$M_{m,k}^{IRM} = \sqrt{\frac{|S_{m,k}|^2}{|S_{m,k}|^2 + |V_{m,k}|^2}} \tag{13}$$

$$M_{m,k}^{SSM} = \frac{|S_{m,k}|^2}{|Y_{m,k}|^2} \tag{14}$$

$$Loss_{MSE}^{MA} = \frac{1}{2N}\sum_{m=1}^{N-1}(M_{m,k} - \hat{M}_{m,k})^2 \tag{15}$$

where $M_{m,k}$ is the ground-truth mask and $\hat{M}_{m,k}$ is the estimated mask. Each sentence in the mini-batch is zero-padded, such that it gives a similar quantity of time frames as the longest noisy sentence. The Adam optimizer with default hyper-parameters ($\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$) [54] and 0.001 learning rate is used for gradient descent optimization. The gradient clipping is also adopted in the model, where the gradient is clipped between [1, 1].

C. MODEL ARCHITECTURE

Table 2 provides the architecture of the proposed model. The input-output size of each layer in the CED is described as (Feature-Map × Time-Step × Frequency-Channel) whereas The hyperparameters are described as) Kernel-Size × Stride × Output-Channel). The CED model consists of five convolutional (Conv2D) and deconvolutional (Deconv2D) layers that form the encoder-decoder network. The use of exponential linear rectified unit (ELU) activation and batch normalization (BN) helps in achieving better convergence and generalization, respectively, while also reducing model complexity. The final layer uses a softplus activation function, which ensures that the output of the network is always positive. This study assumes 16kHz sampled speech signals. To create 50% overlapping time-frames, a 20 milliseconds Hamming window is adopted. The input to the model is 161-dimensional spectra, corresponding to the 320-point STFT (16kHz×20 milliseconds=320 points).

D. EVALUATION METRICS

In experiments, this study adopts five widely used metrics for evaluating speech enhancement, including short-time objective intelligibility (STOI) [53], extended short-time objective intelligibility (ESTOI) [54], perceptual evaluation of speech quality (PESQ) [55], and three composite measures [56]. PESQ is a standard objective measurement algorithm used to assess the quality of speech. The output of PESQ is a single score that represents the overall quality of the speech signal. This score ranges from −0.5 (worst quality) to 4.5 (best quality) and is often reported in Mean Opinion Score (MOS) units. STOI and ESTOI are measures of the intelligibility of speech signals. The STOI score ranges from 0 to 1, with higher scores indicating better intelligibility. An STOI/ESTOI score of 1 indicates perfect intelligibility, while an STOI/ESTOI score of 0 indicates no intelligibility. The composite measures are MOS (mean opinion score) predictors. C_{SIG} (predicted MOS for signal distortion), C_{BAK} (predicted MOS for background noise intrusiveness), and C_{OVL} (overall speech quality), respectively. The composite measures range from 0 to 5. A higher score of all these mentioned metrics shows better SE performance. In addition, two other measures, segmental SNR (SNRSeg) and Source-to-Distortion Ratio (SDR), are used to examine the performance of the proposed NSE-CATNet.

TABLE 2. CED model architecture.

Layer	Input Size	Output Size	Hyperparameters
Conv2D-1	2×T×161	16×T×80	(1×3), (1,2), 16
Conv2D-2	16×T×80	32×T×39	(1×3), (1,2), 32
Conv2D-3	32×T×39	64×T×19	(1×3), (1,2), 64
Conv2D-4	64×T×19	128×T×9	(1×3), (1,2), 128
Conv2D-5	128×T×9	256×T×4	(1×3), (1,2), 256
Deconv2D-5	512×T×4	128×T×9	(1×3), (1,2), 128
Deconv2D-4	256×T×9	64×T×19	(1×3), (1,2), 64
Deconv2D-3	128×T×19	32×T×39	(1×3), (1,2), 32
Deconv2D-2	64×T×39	16×T×80	(1×3), (1,2), 16
Deconv2D-1	32×T×80	1×T×161	(1×3), (1,2), 1
Linear	1×T×161	1×T×161	161

IV. RESULTS AND DISCUSSIONS

This section presents the experimental results obtained with the proposed NSE-CATNet in seen and unseen noises. An interpretation is adopted to represent the SE systems. NSE-CATNet+IRM means the ideal ratio mask is estimated with the proposed NSE-CATNet whereas NSE-CATNet+SSM indicates that the spectral magnitude mask is estimated with the proposed NSE-CATNet.

A. SPEECH ENHANCEMENT PERFORMANCE IN SEEN NOISY CONDITIONS

Table 3 provides the PESQ scores obtained with two models (NSE-CATNet+IRM and NSE-CATNet+SSM) in four example background noises (voice babble, factory, street, and F-16) for two training objectives (IRM and SSM). Compared to unprocessed noisy speech (UnP), the proposed model provides considerable improvements in terms of the PESQ for both training objectives. By taking the voice babble noise with −10dB SNR as a first case, the proposed NSE-CATNet with IRM achieves 0.38 gain on PESQ whereas the proposed NSE-CATNet with SSM achieves 0.42 gain on

TABLE 3. Speech enhancement performance using PESQ for two training objectives.

Noise Type	Voice Babble Noise					Factory Noise					Street Noise					F-16 Noise				
SNR (in dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy (UnP)	1.29	1.40	1.71	2.03	2.27	1.13	1.39	1.70	1.97	2.33	1.29	1.41	1.72	2.02	2.29	1.12	1.28	1.45	1.85	2.09
NSE-CATNet+IRM	1.67	2.17	2.66	2.89	2.96	1.46	2.16	2.65	2.87	2.93	1.53	2.11	2.61	2.88	2.96	1.42	2.09	2.55	2.81	2.94
NSE-CATNet+SSM	1.71	2.21	2.72	2.92	3.01	1.51	2.22	2.68	2.90	2.99	1.55	2.15	2.69	2.91	3.01	1.48	2.16	2.61	2.90	3.02

TABLE 4. Speech enhancement performance using STOI for two training objectives.

Noise Type	Voice Babble Noise					Factory Noise					Street Noise					F-16 Noise				
SNR (in dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy (UnP)	49.6	58.4	69.3	81	89.5	49.1	59	70.4	80.5	89.5	48.2	59.6	69.8	81.6	89.6	50.1	61.4	70.6	82.2	89.8
NSE-CATNet+IRM	61.1	73.2	84.0	91.4	94.4	60.6	74.2	83.7	91.2	94.1	62.2	73.4	84.6	91.5	94.2	64.3	74.3	83.2	91.7	94.5
NSE-CATNet+SSM	62.3	74.1	84.7	92.1	94.8	61.2	74.9	84.1	91.9	95.5	63	73.9	83.1	90.1	91.8	64.9	74.8	83.9	91.6	95.3

TABLE 5. Speech enhancement performance using ESTOI for two training objectives.

Noise Type	Voice Babble Noise					Factory Noise					Street Noise					F-16 Noise				
SNR (in dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy (UnP)	39.4	50.1	61.9	74.7	84	38.9	50.7	63	74.2	84	38	51.3	62.4	75.3	84.1	39.9	53.1	63.2	75.9	84.3
NSE-CATNet+IRM	50.9	64.9	76.6	83.1	85.9	50.4	65.9	76.3	82.9	86.6	52	65.1	75.2	83.2	85.7	54.1	66	75.8	83.4	86
NSE-CATNet+SSM	52.1	65.8	77.3	83.8	86.3	51	66.6	76.7	83.6	87	52.8	65.6	75.7	83.8	86.3	54.7	66.5	76.5	84.3	86.8

TABLE 6. Speech enhancement performance using CSIG, CBAK, and COVL for two training objectives.

Metric	CSIG					CBAK					COVL				
SNR (in dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy (UnP)	1.46	1.83	2.28	2.76	3.18	1.43	1.64	1.89	2.21	2.49	1.41	1.63	1.88	2.2	2.5
NSE-CATNet+IRM	2.51	2.96	3.31	3.85	4.18	2.02	2.33	2.69	2.97	3.24	2	2.32	2.68	2.96	3.25
NSE-CATNet+SSM	2.58	3.02	3.44	3.94	4.32	2.1	2.42	2.78	3.07	3.4	2.08	2.41	2.77	3.06	3.41

TABLE 7. Speech enhancement performance using SNRSeg and SDR for two training objectives.

Metric	SNRSeg					SDR				
SNR (in dB)	-10	-5	0	5	10	-10	-5	0	5	10
Noisy (UnP)	-6.37	-5.41	-2.28	0.95	4.88	-7.55	-3.84	-0.89	3.05	5.13
NSE-CATNet+IRM	1.66	1.81	3.09	4.97	6.94	0.92	4.66	6.34	9.68	14.03
NSE-CATNet+SSM	1.71	1.92	3.24	5.03	7.04	1.09	4.81	6.45	9.77	14.10

PESQ, respectively. Further, by taking the factory noise with -5dB SNR as a second case, the proposed NSE-CATNet with IRM improves the PESQ by 0.77 over the unprocessed noisy speech while the NSE-CATNet with SSM obtains 0.83 gain over the unprocessed noisy speech on PESQ, respectively. In addition, by considering the street noise with 0dB SNR, the NSE-CATNet with IRM obtains 0.89 gain over unprocessed noisy speech whereas the NSE-CATNet with SSM improves the PESQ score by 0.97. On average (all SNRs and noises), the NSE-CATNet+IRM improves the PESQ by 0.74 and the NSE-CATNet+SSM improves the PESQ by 0.78 over unprocessed noisy speech.

Table 4 provides the STOI scores obtained with two models (NSE-CATNet+IRM and NSE-CATNet+SSM) in voice babble, factory, street, and F-16 for IRM and SSM training objectives. In contrast to unprocessed noisy speech, the proposed models provide considerable improvements in terms of the STOI for both training objectives. By taking the F-16 noise with -10dB SNR as a first case, the proposed NSE-CATNet with IRM achieves 0.14 gain on STOI whereas the proposed NSE-CATNet with SSM achieves 0.15 gain on STOI, respectively. Further, by taking the voice babble noise with -5dB SNR as a second case, the proposed NSE-CATNet with IRM improves the STOI by 0.15 over the unprocessed noisy speech while the NSE-CATNet with SSM obtains 0.16 gain over the unprocessed noisy speech on STOI, respectively. In addition, by considering the factory noise with 0dB SNR, the NSE-CATNet with IRM obtains 0.133 gain over unprocessed noisy speech whereas the NSE-CATNet with SSM improves the STOI score by 0.14. On average (all SNRs and noises), the NSE-CATNet+IRM improves the STOI by 0.11 and the NSE-CATNet+SSM improves the STOI by 0.107 over unprocessed noisy speech.

Table 5 provides the ESTOI scores obtained with two models (NSE-CATNet+IRM and NSE-CATNet+SSM) in voice babble, factory, street, and F-16 for IRM and SSM training objectives. The proposed NSE-CATNet+IRM improves the overall ESTOI by 9.81% whereas the NSE-CATNet+SSM improves the overall ESTOI by 0.992 over unprocessed noisy speech. Table 6 reports the average C_{SIG}, C_{BAK}, and C_{OVL} scores for each SNR which covers voice babble, factory, street, and F-16, respectively. It is notable that the proposed NSE-CATNet with IRM and SSM significantly improves the performance in terms of the three composite metrics. At -10dB SNR, for example, the NSE-CATNet with IRM improves C_{SIG} by 1.05, C_{BAK} by 0.73, and C_{OVL} by

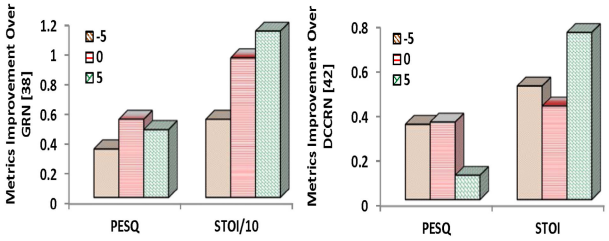

FIGURE 6. Improvements of the NSE-CATNet over unprocessed noisy speech.

TABLE 8. Seen noisy speech enhancement performance in terms of all evaluation metrics.

Noise Condition	Seen Noisy Conditions					
SNR (in dB)	PESQ	STOI	ESTOI	CSIG	CBAK	COVL
Noisy (UnP)	1.68	69.96	68.39	2.3	1.93	1.92
NSE-CATNet+IRM	2.42	81.09	78.35	3.36	2.65	2.64
NSE-CATNet+SSM	2.47	81.75	78.99	3.46	2.75	2.74

0.68 whereas the NSE-CATNet with SSM improves C_{SIG} by 1.16, C_{BAK} by 0.82, and C_{OVL} by 0.80, respectively. Table 7 reports the average SNRSeg and SDR scores for each SNR which covers the voice babble, factory, street, and F-16 noise, respectively. To increase the readability of the obtained results, Table 8 presents the overall average scores obtained with all performance metrics. The experimental results validate that the proposed NSE-CATNet model consistently obtains notable improvements to the unprocessed seen noisy speech in terms of PESQ, STOI, ESTOI, C_{SIG}, C_{BAK}, and C_{OVL}. To show the success of the proposed SE, two strong baselines including GRN [38] and DCCRN [42] are selected, and the improvement in PESQ and STOI at various input SNRs is compared. Figure 6 shows the improvements of the NSE-CATNet (with both training objectives) over GRN and DCCRN.

B. SPEECH ENHANCEMENT IN UNSEEN NOISY CONDITIONS

To further examine the performance of the proposed NSE-CATNet model, Table 9 shows the speech enhancement performance in two unseen noisy conditions (factory2 and cafeteria) in terms of PESQ, STOI, ESTOI, PESQ, C_{SIG}, C_{BAK}, and C_{OVL}. The effective model architecture indicates that the speech performance is not drastically altered in unseen noisy conditions. For example, the STOI and ESTOI improvements over unprocessed noisy speech are 0.11 and 0.99 (by NSE-CATNet with IRM) whereas 0.12 and 0.11 (by NSE-CATNet with SSM). Further, the PESQ gain over unprocessed noisy speech is 0.74 (by NSE-CATNet with IRM) and 0.79 (by NSE-CATNet with SSM). Finally, the NSE-CATNet with IRM improves C_{SIG} by 1.06, C_{BAK} by 0.72, and C_{OVL} by 0.66 whereas the NSE-CATNet with SSM improves C_{SIG} by 1.14, C_{BAK} by 0.80, and C_{OVL} by 0.63, respectively. The results in unseen noisy conditions confirm the success of the proposed SE.

TABLE 9. Unseen noisy speech enhancement performance in terms of all evaluation metrics.

Noise Condition	Unseen Noisy Conditions					
SNR (in dB)	PESQ	STOI	ESTOI	CSIG	CBAK	COVL
Noisy (UnP)	1.606	68.87	67.51	2.234	1.876	1.863
NSE-CATNet+IRM	2.346	79.99	77.47	3.294	2.596	2.503
NSE-CATNet+SSM	2.396	80.66	78.11	3.394	2.696	2.683

TABLE 10. ANOVA analysis at 95% confidence interval.

Metric	PESQ		STOI		ESTOI	
SNR	0dB		0dB		0dB	
Analysis	Pvalue	Fvalue	Pvalue	Fvalue	Pvalue	Fvalue
NSE-CATNet+IRM →UnP	0.0020	88.91	0.0001	85.12	0.0016	95.13
NSE-CATNet+SSM →UnP	0.0001	89.34	0.0001	86.74	0.0001	96.02

C. SPECTRO-TEMPORAL ANALYSIS

To visually examine the processed speech along with its spectral regions and residual noise, Fig. 7 shows the spectro-temporal representations. A clean speech (Fig 7(a)) from the VoiceBank-DEMAND dataset is mixed with voice babble noise at 0dB SNR level to create a noisy speech (Fig. 7(b)). Voice babble is a challenging noise type created by multiple people speaking at the same time, resulting in indistinct noise. Figure 7(c) illustrates the spectro-temporal representation of the noisy speech enhanced by the NSE-CATNet+IRM where negligible residual noise is evident. Further, no significant speech distortion is observed in the spectro-temporal representation. Also, the speech processed by NSE-CATNet+SSM (Fig. 7(d)) shows a fine spectral structure. Less residual noise and speech distortion indicate better speech quality and intelligibility.

D. ANOVA ANALYSIS

The experimental scores indicate that the proposed NSE-CATNet performs better at each SNR level. Therefore, to confirm the significance of the results at a favorable SNR (5dB), this study conducts a one-way ANalysis-of-VAriance (ANOVA) statistical test. The statistical test is performed at 95% confidence interval. The difference between scores is deemed statistically significant when the probability is less than 0.05 ($p < 0.05$) and the value is larger than the critical value, which is $f_{value} > f_{critical}$. Table 10 shows the statistical analysis at a critical value of 3.09. The P_{value} of the proposed NSE-CATNet is larger than 0.05, and the critical value is greater than 3.09. The statistical test suggests that the results are statistically significant at all SNR levels.

E. COMPUTATIONAL LOAD AND SE PERFORMANCE

The computational load of the proposed NSE-CATNet model is measured in terms of trainable parameters and MACs (Multiply-Accumulate operations), useful metrics for estimating computational complexity and optimizing the performance on specific hardware platforms. Table 11 shows the total trainable parameters and MACs for the proposed and related convolutional recurrent networks in the CED domain. The total number of trainable parameters of the NSE-CATNet model is around 3.57M and is advantageous

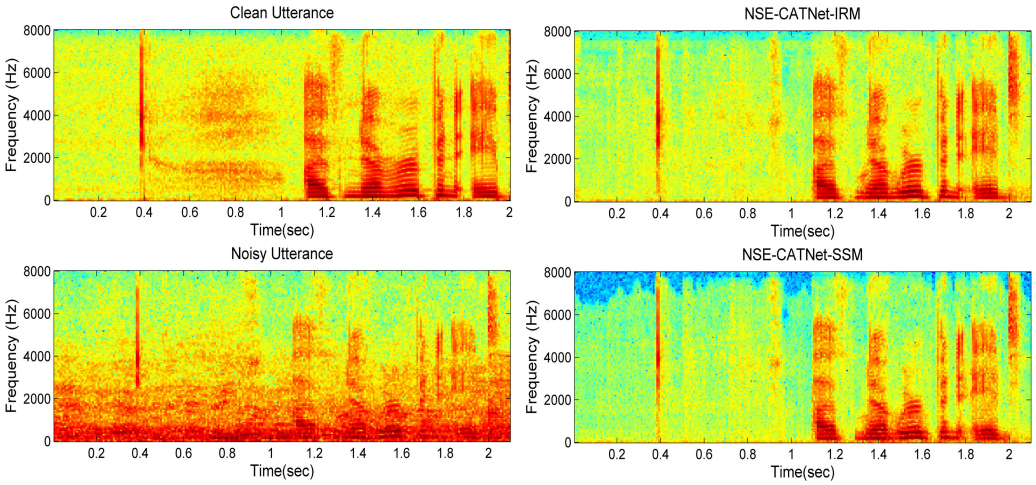

FIGURE 7. Spectro-Temporal analysis. Noisy speech degraded by voice babble at 0dB SNR. Clean speech "The problems are a result of that shortfall". Noisy speech processed by NSE-CATNet+IRM. Noisy speech processed by NSE-CATNet+SSM.

in MACS (2.725 G/s). Since this study uses a convolutional MHA bottleneck, the trainable parameters are significantly reduced. With regular MHA architecture [22] ($d_{model} = 256$, H=8, and $d_{ff} = 1024$), the trainable parameters are around 4.18M which are reduced to 1.96M with the convolutional bottleneck. The time-frequency attention module adds additional 0.41K trainable parameters. With such computational complexity, the proposed model achieves better results in terms of PESQ and STOI (average of -5dB, 0dB, and 5dB SNR levels), as given in Table 10. The symbol "↑" indicates the improvement in PESQ and STOI. Note that the proposed method shows reduced trainable parameters and MACs except CRN (2.57 G/s); however, the evaluation metrics (PESQ and STOI) are better than CRN.

F. COMPARISON WITH OTHER STUDIES

To showcase the superiority of the proposed NSE-CATNet, this study compares the performance with multiple recent studies from the speech enhancement literature using the VoiceBank-DEMAND dataset. Three SNR levels (-5dB, 0dB, 5dB) are selected for the comparison. The comparative results are reported in Table 12, where it can be observed that the proposed NSE-CATNet model with IRM and SSM training objectives shows highly competitive performance to the multiple state-of-the-art models in terms of PESQ and STOI evaluation metrics. Except for CRN-TCS which performs marginally better at favorable SNR (5dB), the remaining models underperform the proposed NSE-CATNet. For instance, the average PESQ gain of the E2E-BLSTM-CRN over unprocessed noisy speech is 0.84 which is 1.81% less than the proposed NSE-CATNet. Further, the average PESQ gain of the DeepXi over unprocessed noisy speech is 0.54 which is 10.85% less than the proposed NSE-CATNet. Similarly, by taking the -5dB adverse SNR level as a case, the proposed NSE-CATNet outperforms the related models by reasonable margins, such as the PESQ is improved by a factor of 0.27 higher with NSE-CATNet+SSM over state-of-the-art CRN model. Further, by taking the 0dB SNR level as another case, the proposed NSE-CATNet outperforms the related studies by reasonable margins, such as the STOI is improved by 0.446 with NSE-CATNet+IRM over the CRN-BLSTM model. To show the overall performance over related models, Table 13 summarizes the average PESQ and STOI improvements of all competing models over unprocessed noisy speech. Where the symbol "↑" indicates the improvements.

G. CROSS CORPUS AND TRAINING OBJECTIVE ANALYSIS

To further examine the performance, this section performs the cross-corpus and training objectives analysis. Speech datasets typically consist of recordings of speech utterances produced by different speakers, often in controlled environments to ensure high-quality recordings. However, even within controlled environments, there can be variations in recording quality due to factors such as the type of microphones, the recording equipment, and the room acoustics. In addition to the LibriSpeech and VoiceBank-DEMAND databases, this study selects speech sentences from IEEE-Male [64] and IEEE-Female [64] databases. Clean speech sentences are selected from all databases and the proposed NSE-CATNet model with IRM and SSM training objectives is trained individually. To analyze the effect of a speech dataset on speech enhancement performance, this section presents Table 14 which shows the PESQ and STOI scores across five SNRs (−10dB to +10dB)for four databases (LibriSpeech [44], VoiceBank-DEMAND [43], IEEE-Male, and IEEE-Female). The proposed model shows almost equal performance at four databases indicating the

TABLE 11. Computational load in terms of para and MACs.

Metric	Year	Domain	Para#	MACs	PESQ	STOI	↑PESQ	↑STOI
GCRN [34]	2020	Time-Frequency	9.770 M	2.420 G/s	2.48	87.23	0.82	16.92
CRN [27]	2018	Time-Frequency	17.58 M	**2.570 G/s**	2.33	81.73	0.72	11.42
DCCRN [42]	2020	Time-Frequency	3.670 M	11.13 G/s	2.54	85.59	0.93	15.28
AECNN [35]	2021	Time-Frequency	4.820 M	36.56 G/s	2.62	87.28	1.01	16.97
NSE-CATNet (Proposed)	2023	Time-Frequency	**3.570 M**	2.725 G/s	**2.64**	**87.43**	**1.03**	**17.12**

TABLE 12. Comparison with the state-of-the-art se models.

Metric	**PESQ**				**STOI**			
SNR (in dB)	**-5**	**0**	**5**	**Average**	**-5**	**0**	**5**	**Average**
Noisy Unprocessed	1.37	1.64	1.97	1.66	59.61	70.02	81.32	70.31
E2E-BLSTM-CRN [40]	2.13	2.57	2.81	2.5	70.17	79.85	87.37	79.13
E2E-BGRU-CRN [40]	2.14	2.59	2.83	2.52	72.92	83.76	89.43	82.03
E2E-BSRU-CRN [40]	2.13	2.62	2.85	2.53	73.56	85.09	90.31	82.98
DeepResGRU [21]	2.09	2.29	2.49	2.29	74.13	81.81	85.51	80.48
CFN-GCFU [31]	1.98	2.24	2.62	2.28	71.61	78.19	86.21	78.67
MCBNet [30]	2.01	2.32	2.52	2.28	72.81	79.15	84.15	78.71
CRN-BLSTM [33]	1.93	2.23	2.51	2.22	70.31	77.08	81.96	76.45
PL-CRNN [37]	2.06	2.51	2.85	2.47	73.16	84.42	90.15	82.57
CNN-GRU [57]	2.01	2.34	2.65	2.33	74.61	83.11	90.11	82.61
DTLN [58]	1.91	2.34	2.67	2.31	72.72	85.19	90.68	82.86
DCCRN [42]	1.85	2.34	2.78	2.32	74.51	85.87	92.38	84.25
DNN-TGSA [59]	2.01	2.31	2.58	2.3	74.41	81.21	84.12	79.91
DeepXi [60]	1.99	2.21	2.41	2.2	72.01	81.21	91.99	81.73
GRN [38]	1.86	2.16	2.42	2.15	69.76	76.89	81.42	76.02
AECNN [35]	1.92	2.19	2.45	2.19	72.01	77.78	82.51	77.43
CRN [27]	1.92	2.22	2.49	2.21	70.11	76.95	81.88	76.31
GAN [61]	1.72	2.15	2.44	2.11	65.01	75.71	82.61	74.44
LSTM [62]	1.82	2.15	2.44	2.14	68.78	75.81	81.54	75.37
CRN-TCS [63]	2.15	2.67	3.05	2.62	74.63	85.09	90.83	83.51
GaGNet [67]	2.39	2.61	3.09	2.72	76.53	82.13	86.20	78.37
NSE-CATNet+IRM (Proposed)	2.16	2.68	2.86	2.57	74.77	85.73	92.52	84.34
NSE-CATNet+SSM (Proposed)	**2.19**	**2.69**	2.88	2.59	**75.01**	**86.27**	**93.13**	**84.81**

TABLE 13. PESQ and STOI improvements of all related studies over unprocessed noisy speech.

Models	↑PESQ	↑STOI
E2E-BLSTM-CRN [40]	0.84	08.82
E2E-BLSTM-CRN [40]	0.48	11.72
E2E-BLSTM-CRN [40]	0.87	12.67
DeepResGRU [21]	0.63	10.17
CFN-GCFU [31]	0.62	08.36
MCBNet [30]	0.62	08.40
CRN-BLSTM [33]	0.56	06.14
PL-CRNN [37]	0.81	12.26
CNN-GRU [57]	0.67	12.30
DTLN [58]	0.65	12.55
DCCRN [42]	0.66	13.94
DNN-TGSA [59]	0.64	09.60
DeepXi [60]	0.54	11.42
GRN [38]	0.49	05.71
AECNN [35]	0.53	07.12
CRN [27]	0.55	06.00
GAN [61]	0.45	04.13
LSTM [62]	0.48	05.06
CRN-TCS [63]	**0.96**	13.20
NSE-CATNet+IRM (Proposed)	0.91	14.03
NSE-CATNet+SSM (Proposed)	0.93	**14.50**

generalization towards various speech databases. The proposed NSE-CATNet with SSM training objective performs slightly better than the IRM training objective. It is important to mention that the results with the VoiceBank-DEMAND dataset are different in Table 14 since this set of experiments uses different SNR levels whereas Table 15 follows the exact remedy available in many SOTA models.

Further examine the proposed NSE-CATNet with existing baseline models for SE in time-domain and time-frequency-domain, this study uses the publicly available VoiceBank-DEMAND dataset with an exact remedy followed by baseline studies during the model evaluations. The training set (composed of 11572 speech utterances) consists of 28 speakers with four SNRs (15dB, 10dB, 5dB, and 0dB). The test sets (composed of 824 speech utterances) consist of 2 speakers with four SNRs (17.5dB, 12.5dB, 7.5dB, and 2.5dB). The results are presented in Table 15 to validate the performance of the proposed NSE-CATNet with baseline models [22], [23], [25], [42], [67], [68], [69], [70]. With the VoiceBank-DEMAND dataset, the proposed NSE-CATNet with IRM and SSM training objectives achieves the best results as compared to the baseline models except C_{sig} where SE-Conformer shows the best results (4.45). The proposed NSE-CATNet achieves competitive performance as compared to the baseline models. For example, from RDL-Net to NSE-CATNet, average 0.17, 2.3%, 0.03, 0.23, and 0.10 improvements are achieved in terms of PESQ, STOI, C_{sig}, C_{bak} and C_{ovl}, respectively. Similarly, from DCRNN to NSE-CATNet, average 0.51 PESQ, 2.4% STOI, 0.53 C_{sig}, 0.48 C_{bak}, 0.55 C_{ovl},

TABLE 14. Cross corpus and training objective analysis.

Database	LibriSpeech		VB-DEMAND		IEEE-Male		IEEE-Female	
Metric	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
NSE-CATNet+IRM	2.356	80.95	2.341	81.54	2.364	80.41	2.364	81.24
NSE-CATNet+SSM	2.412	81.62	2.403	81.88	2.424	81.12	2.423	81.78

TABLE 15. Performance evaluation on VoiceBank+DEMAND database. denotes that the result is not provided in the original paper.

Models	Year	Domain	Para#	PESQ	STOI	Csig	Cbak	Covl	SNRSeg
Noisy	-	-	-	1.97	91.6	3.34	2.44	2.63	1.69
SEGAN [22]	2017	Time-Domain	97.47 M	2.16	93.1	3.47	2.93	2.84	7.66
MetricGAN+ [23]	2021	Time-Frequency	–	3.15	–	4.14	3.16	3.64	–
DCCRN [42]	2019	Time-Frequency	3.70 M	2.68	93.7	3.88	3.18	3.27	8.62
GAGNet [67]	2021	Time-Frequency	5.94 M	2.94	94.7	4.36	3.45	3.59	9.24
RDL-Net [68]	2020	Time-Frequency	3.91 M	3.02	93.8	4.38	3.43	3.72	–
DEMUCS [69]	2020	Time-Domain	128.0 M	3.07	95.1	4.31	3.40	3.63	8.53
TSTNN [70]	2021	Time-Domain	0.92 M	2.96	95.1	4.17	3.53	3.49	9.72
SE-Conformer [25]	2021	Time-Domain	–	3.13	95.1	**4.45**	3.55	**3.82**	–
NSE-CATNet (Proposed)	2023	Time-Frequency	3.57 M	**3.19**	**96.1**	4.41	**3.66**	**3.82**	**9.97**

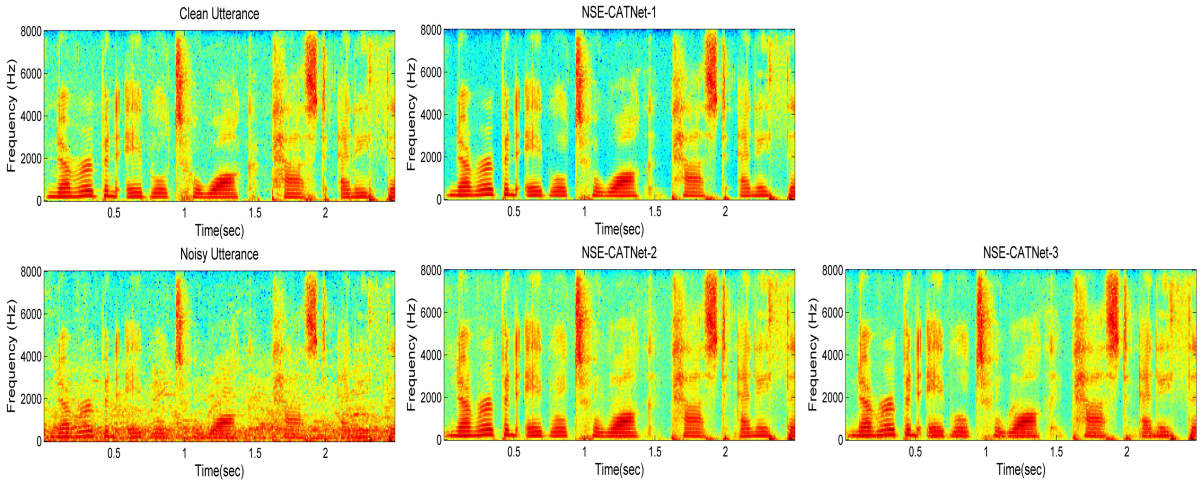

FIGURE 8. Spectro-Temporal analysis. A Clean speech "I have never been able to walk passed anything I believed to be wrong without saying something" is degraded with voice babble at -5dB SNR. Noisy speech processed by NSE-CATNet-1 (with CED only). Noisy speech processed by NSE-CATNet-2 (with CED+Bottleneck).Noisy speech processed by NSE-CATNet-3 (with CED+CAT).

and 1.35dB SNRSeg improvements over achieved in terms of PESQ, STOI, C_{sig}, C_{bak}, C_{ovl}, and SNRSeg, respectively.

H. ABLATION STUDY

This section conducts the ablation study to show the performance of different modules in the proposed NSE-CATNet model. The models are examined as (a) the NSE-CATNet model applying CED without bottleneck (denoted by NSE-CATNet-1); (b) the NSE-CATNet model applying CED with bottleneck (denoted by NSE-CATNet-2); and (c) NSE-CATNet model applying CED with bottleneck and TFA module (denoted by NSE-CATNet-3, the full model). To examine the models in ablation studies, the experiments use clean utterances from the LibriSpeech database mixed with voice babble at two SNRs (0dB and 5dB). Table 16 gives the overall results in terms of PESQ and STOI values. The full model shows the best performance as expected. The inclusion of CAT into CED significantly improves the PESQ and STOI values. The integration of the time-frequency attention module into the bottleneck greatly improves the values. The TFA module adds additional trainable parameters, but the overall computational load is less as compared to the related studies (given in Table 11). Figure 8 shows the spectrograms of the ablation studies. The spectrograms show the impacts of different modules in the proposed speech enhancement.

TABLE 16. Ablation study.

Models	PESQ	STOI	↑PESQ	↑STOI
NSE-CATNet-1	2.19	73.65	0.51	3.69
NSE-CATNet-2	2.35	80.02	0.67	10.06
NSE-CATNet-3	2.43	81.42	0.75	11.46

TABLE 17. ASR analysis in terms of CER (in %) after speech enhancement.

Database	-10dB	-5dB	0dB	5dB	10dB	Clean
NSE-CATNet+IRM	54.11	37.82	23.35	15.93	11.22	9.96
NSE-CATNet+SSM	54.08	37.77	23.29	15.78	11.01	9.96

I. AUTOMATIC SPEECH RECOGNITION PERFORMANCE

Speech enhancement techniques can be used as a front-end to automatic speech recognition (ASR) systems to improve their accuracy and robustness in noisy environments. In a noisy environment, the quality of the speech signal is degraded due to various sources of interference such as background noise, reverberation, and competing talkers. This degradation can negatively affect the performance of ASR systems, making it difficult for them to accurately recognize speech. By using NSE-CATNet as a front-end to an ASR system, the accuracy and robustness of the system are significantly improved. This study applies the state-of-the-art end-to-end speech transformer with self-attention as a speech recognition component [71]. The ASR performance is measured in terms of character error rate (CER). Table 17 shows that in the presence of noise, the performance of the ASR system is significantly impaired. However, as for the NSE-CATNet, it significantly improves the ASR robustness. We examine that both training objectives obtained almost similar CER results.

V. SUMMARY AND CONCLUSION

The paper proposes a novel neural speech enhancement (NSE) system based on the convolutional encoder-decoder (CED) framework where conventional recurrent networks are replaced with a convolutional attention transformer (CAT) module to extract high-level features. This allows the model to operate on a lower-dimensional representation of the input signal, reducing the computational cost and the number of parameters required. The bottleneck contains the convolution layers with 1-D kernels and multi-head attention (MHA) modules. Furthermore, to quantify the important time-frequency speech distributions in the speech signals, a time-frequency attention (TFA) module with time-frame attention and frequency-channel attention is added to the convolutional transformer that generates a 2-D attention map. The time-frequency attention module shows effectiveness for neural speech enhancement. The performance of the proposed speech enhancement is evaluated using objective speech quality (PESQ) and intelligibility (STOI) metrics on the VoiceBank-DEMAND and the LibriSpeech databases. Compared to unprocessed noisy speech (UnP), the proposed SE model provides considerable improvements in terms of the PESQ and STOI for both training objectives (IRM and SSM). With voice babble noise with -10dB SNR, the NSE-CATNet with IRM achieves 16.23% gain on PESQ whereas achieves 17.83% gain on PESQ with NSE-CATNet+SSM. With factory noise with -5dB SNR, the NSE-CATNet with IRM improves the PESQ by 27.22% while obtaining 29.79% gain over the unprocessed noisy speech on PESQ the NSE-CATNet with SSM. On average, the NSE-CATNet+IRM improves STOI by 10.08% and the NSE-CATNet+SSM improves the STOI by 10.74% over unprocessed noisy speech. The results in seen and unseen noisy conditions confirm the success of the proposed speech enhancement. The speech processed by NSE-CATNet+IRM and NSE-CATNet+SSM concludes a fine spectral structure where less residual noise and speech distortion indicates better speech quality and intelligibility. The P_{value} of the proposed NSE-CATNet is larger than 0.05, and the critical value is greater than 3.09 which concludes that values obtained with the proposed model are statistically significant. The total number of trainable parameters of the NSE-CATNet model is around 3.57M and is advantageous in MACS (2.725 G/s). Since this study uses a convolutional MHA bottleneck, the trainable parameters are significantly reduced concluding the less computational load with better quality and intelligibility. The comparative results (in Table 11) conclude that the proposed NSE-CATNet model with IRM and SSM training objectives shows highly competitive performance to the multiple state-of-the-art models in terms of PESQ and STOI evaluation metrics. During cross-corpus analysis, the proposed model concludes better performance at four databases indicating the generalization towards various speech databases. The ablation study concludes the success of the CAT module in the CED framework. Finally, the proposed NSE-CATNet is examined against baseline models for SE in time-domain and time-frequency-domain using the publicly available VoiceBank-DEMAND dataset with an exact remedy followed by the original study and the proposed model showed a significant performance on the dataset.

Our future studies focus on converting the proposed NSE-CATNet into a complex spectral mapping-based personalized speech enhancement by adding speaker embeddings to the bottleneck layers. The audio samples[1] can be found for reference.

REFERENCES

[1] M. Gupta, R. K. Singh, and S. Singh, "Analysis of optimized spectral subtraction method for single channel speech enhancement," *Wireless Pers. Commun.*, vol. 128, no. 3, pp. 2203–2215, Feb. 2023.
[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[3] C. Jannu and S. D. Vanambathina, "Weibull and Nakagami speech priors based regularized NMF with adaptive Wiener filter for speech enhancement," *Int. J. Speech Technol.*, vol. 26, no. 1, pp. 197–209, Mar. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10772-023-10020-5
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[5] D. Mukhutdinov, A. Alex, A. Cavallaro, and L. Wang, "Deep learning models for single-channel speech enhancement on drones," *IEEE Access*, vol. 11, pp. 22993–23007, 2023.
[6] T. Rosenbaum, I. Cohen, E. Winebrand, and O. Gabso, "Differentiable mean opinion score regularization for perceptual speech enhancement," *Pattern Recognit. Lett.*, vol. 166, pp. 159–163, Feb. 2023.

[1] http://nasirsaleem.website3.me/

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, Aug. 2013, pp. 436–440.
[8] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[9] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[10] S. Chakrabarty and E. A. P. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.
[11] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.
[12] Y. Jiang, H. Zhou, and Z. Feng, "Performance analysis of ideal binary masks in speech enhancement," in *Proc. 4th Int. Congr. Image Signal Process.*, vol. 5, Oct. 2011, pp. 2422–2425.
[13] F. Bao and W. H. Abdulla, "A new ratio mask representation for CASA-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 7–19, Jan. 2019.
[14] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[16] N. Saleem and M. I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 1, pp. 84–91, 2020.
[17] N. Saleem, M. I. Khattak, and A. B. Qazi, "Supervised speech enhancement based on deep neural network," *J. Intell. Fuzzy Syst.*, vol. 37, no. 4, pp. 5187–5201, Oct. 2019.
[18] R. Soleymanpour, M. Soleymanpour, A. J. Brammer, M. T. Johnson, and I. Kim, "Speech enhancement algorithm based on a convolutional neural network reconstruction of the temporal envelope of speech in noisy environments," *IEEE Access*, vol. 11, pp. 5328–5336, 2023.
[19] S. Girirajan and A. Pandian, "Real-time speech enhancement based on convolutional recurrent neural network," *Intell. Autom. Soft Comput.*, vol. 35, no. 2, pp. 1987–2001, 2023.
[20] Y. Xia and J. Wang, "Low-dimensional recurrent neural network-based Kalman filter for speech enhancement," *Neural Netw.*, vol. 67, pp. 131–139, Jul. 2015.
[21] N. Saleem, J. Gao, M. I. Khattak, H. T. Rauf, S. Kadry, and M. Shafi, "DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107914.
[22] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," 2021, *arXiv:2104.03538*.
[23] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Aug. 2017, pp. 3642–3646.
[24] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," 2022, *arXiv:2209.11112*.
[25] E. Kim and H. Seo, "SE-conformer: Time-domain speech enhancement using conformer," in *Proc. Interspeech*, Aug. 2021, pp. 2736–2740.
[26] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "DF-conformer: Integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust. (WASPAA)*, Oct. 2021, pp. 161–165.
[27] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, Sep. 2018, pp. 3229–3233.
[28] A. Karthik and J. L. MazherIqbal, "Efficient speech enhancement using recurrent convolution encoder and decoder," *Wireless Pers. Commun.*, vol. 119, no. 3, pp. 1959–1973, Aug. 2021.
[29] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6674–6678.
[30] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "Multi-scale residual convolutional encoder decoder with bidirectional long short-term memory for single channel speech enhancement," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 431–435.
[31] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "Convolutional fusion network for monaural speech enhancement," *Neural Netw.*, vol. 143, pp. 97–107, Nov. 2021.
[32] S. K. Roy and K. K. Paliwal, "Causal convolutional encoder decoder-based augmented Kalman filter for speech enhancement," in *Proc. 14th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Dec. 2020, pp. 1–7.
[33] Z. Wang, T. Zhang, Y. Shao, and B. Ding, "LSTM-convolutional-BLSTM encoder–decoder network for minimum mean-square error approach to speech enhancement," *Appl. Acoust.*, vol. 172, Jan. 2021, Art. no. 107647.
[34] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.
[35] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6875–6879.
[36] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2401–2405.
[37] A. Li, M. Yuan, C. Zheng, and X. Li, "Speech enhancement using progressive learning-based convolutional recurrent neural network," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107347.
[38] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6865–6869.
[39] T. Hsieh, H. Wang, X. Lu, and Y. Tsao, "WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 2149–2153, 2020.
[40] R. Ullah, L. Wuttisittikulkij, S. Chaudhary, A. Parnianifard, S. Shah, M. Ibrar, and F.-E. Wahab, "End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement," *Sensors*, vol. 22, no. 20, p. 7782, Oct. 2022.
[41] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 656–660.
[42] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, Oct. 2020, pp. 2472–2476.
[43] C. Valentini, "Noisy speech database for training speech enhancement algorithms and TTS models," Univ. Edinburgh, School Inform. Centre Speech Res., Edinburgh, Scotland, Tech. Rep., 2016.
[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[45] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 641–652.
[46] Q. Zhang, Q. Song, A. Nicolson, T. Lan, and H. Li, "Temporal convolutional network with frequency dimension adaptive attention for speech enhancement," in *Proc. Interspeech*, Aug. 2021, pp. 166–170.
[47] Q. Zhang, X. Qian, Z. Ni, A. Nicolson, E. Ambikairajah, and H. Li, "A time-frequency attention module for neural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 462–475, 2023.
[48] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Sep. 2010, pp. 3110–3113.
[49] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[50] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG-10 noise database," Tech. Rep., IZF 1988-3, 1988. [Online]. Available: https://www.steeneken.nl/wp-content/uploads/2014/04/RSG-10_Noise-data-base.pdf and https://www.steeneken.nl/7-noise-data-base/

[51] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[52] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[53] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[54] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5085–5089.

[55] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part I—Time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, 2002.

[56] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[57] M. Hasannezhad, Z. Ouyang, W. Zhu, and B. Champagne, "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 764–768.

[58] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," 2020, *arXiv:2005.07551*.

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

[60] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[61] N. Shah, H. A. Patil, and M. H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1246–1251.

[62] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, May 2016.

[63] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[64] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[65] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Exp. Lett.*, vol. 1, no. 1, Jan. 2021, Art. no. 014802.

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

[67] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, Feb. 2022, Art. no. 108499.

[68] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, "Deep residual-dense lattice network for speech enhancement," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8552–8559.

[69] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," 2020, *arXiv:2006.12847*.

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

[71] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen, "Gated recurrent fusion with joint training framework for robust end-to-end speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 198–209, 2021.

NASIR SALEEM received the B.S. degree in telecommunication engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2008, the M.S. degree in electrical engineering from CECOS University, Peshawar, in 2012, and the Ph.D. degree in electrical engineering (digital speech processing and deep learning) from the University of Engineering and Technology, Peshawar, in 2021. Currently, he is a Postdoctoral Fellow with Islamic International University Malaysia (IIUM), where he is researching modern artificial intelligence-based speech processing algorithms. From 2008 to 2012, he was a Senior Lecturer with the Institute of Engineering Technology (IET), Gomal University, where he was involved in teaching and research. He is currently an Assistant Professor with the Department of Electrical Engineering, Faculty of Engineering and Technology (FET), and the Deputy Director of the Quality Assurance Directorate, Gomal University. He has published several research articles in well-known venues so far, such as Elsevier, Springer, and IEEE. His current research interests include human–machine interaction, speech enhancement, speech and video processing, and machine learning applications. He is involved in academic activities, such as a Guest Editor and reviewing papers from several well-known venues, including IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, IEEE SIGNAL PROCESSING, IEEE MULTIMEDIA, IEEE ACCESS, *Expert Systems with Applications*, *Applied Acoustics*, and *Neural Networks* journal.

TEDDY SURYA GUNAWAN (Senior Member, IEEE) received the B.Eng. degree (cum laude) in electrical engineering from Institut Teknologi Bandung (ITB), Indonesia, in 1998, the M.Eng. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001, and the Ph.D. degree from the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia, in 2007. He has been a Professor with International Islamic University Malaysia, since 2019, where he was the Head of the Department of Electrical and Computer Engineering, from 2015 to 2016, and the Head of the Program Accreditation and Quality Assurance, Faculty of Engineering, from 2017 to 2018. He was a Visiting Research Fellow with UNSW, from 2010 to 2021. He was an Adjunct Professor with Telkom University, from 2022 to 2023. His current research interests include speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award from IIUM, in 2018. He was the Chairperson of the IEEE Instrumentation and Measurement Society Malaysia Section, in 2013, 2014, 2020, and 2021. He has been a Chartered Engineer of IET, U.K., since 2016; an Insinyur Profesional Utama of PII, Indonesia, since 2021; a Registered ASEAN Engineer, since 2018; and an ASEAN Chartered Professional Engineer, since 2020.

MIRA KARTIWI (Member, IEEE) is currently a Professor with the Department of Information Systems, Kulliyyah of Information and Communication Technology, and the Deputy Director of e-learning with the Centre for Professional Development, International Islamic University Malaysia (IIUM). She is an experienced consultant specializing in the health, financial, and manufacturing sectors. Her current research interests include health informatics, e-commerce, data mining, information systems strategy, business process improvement, product development, marketing, delivery strategy, workshop facilitation, training, and communications. She was one of the recipients of the Australia Postgraduate Award (APA), in 2004. For her achievement in research, she was awarded the Higher Degree Research Award for Excellence, in 2007. She has also been appointed as an editorial board member in local and international journals to acknowledge her expertise.

BAMBANG SETIA NUGROHO (Member, IEEE) received the bachelor's degree in telecommunications from the Department of Electrical Engineering, Institut Teknologi Bandung (ITB), Bandung, Indonesia, in 1999, the master's degree from the Graduate School of Electrical Engineering, ITB, in 2004, and the Ph.D. degree from the Graduate School of Electrical Engineering, Universitas Indonesia, in 2015. Since 1999, he has been a Lecturer and a Researcher with the School of Electrical Engineering, Telkom University, Indonesia. He has authored more than 25 scientific publications in the telecommunication area. His current research interests include telecommunication and antenna engineering. He is a member of the IEEE Communication Society and the Antenna and Propagation Society. He is a Reviewer of the International Conference on Telecommunications.

INUNG WIJAYANTO (Member, IEEE) received the bachelor's and master's degrees in telecommunication engineering from Institut Teknologi Telkom (now Telkom University), Bandung, Indonesia, in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from the Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia. He was a Teaching Staff with the School of Electrical Engineering, Telkom University, in 2010. He has published more than 40 research articles with Elsevier, IEEE, and Springer. His current research interests include audio and image processing, biomedical signal and image processing and analysis, computer vision, and medical instruments. He serves as a reviewer for several Springer and Elsevier journals.