IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Temporally Dynamic Spiking Transformer Network for Speech Enhancement

MANAL ABDULLAH ALOHALI[1], NASIR SALEEM[2], DELEL RHOUMA[3], MOHAMED MEDANI[4], HELA ELMANNAI[5], AND SAMI BOUROUIS[6]

[1]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh-11671, Saudi Arabia
[2]Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University, D.I.Khan-29050, Pakistan
[3]Department of Computer Science, College of Computer, Qassim University, Saudi Arabia
[4]Department of Computer Science, College of Science and Art at Mahayil, King Khalid University, Muhayil Aseer, 62529, Saudi Arabia
[5]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh-11671, Saudi Arabia
[6]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia

Corresponding Author: Delel Rhouma: d.rhouma@qu.edu.sa

ABSTRACT Speech enhancement (SE) aims to improve the quality and intelligibility of speech signals, particularly in the presence of noise or other distortions, to ensure reliable communication and robust speech recognition. Deep neural networks (DNNs) have shown remarkable success in SE due to their ability to learn complex patterns and representations from large amounts of data. However, they face limitations in handling long-term temporal sequences. Spiking neural networks and transformers inherently manage temporal data and capture fine-grained temporal patterns in speech signals. This paper proposes a model that integrates self-attention with spiking neural networks for speech enhancement. The proposed model employs a convolutional encoder-decoder architecture with a spiking transformer acting as a bottleneck network. The spiking self-attention mechanism in this framework represents features using spike-based queries, keys, and values. This approach enhances features by effectively capturing temporal dependencies and contextual relationships in speech signals. The spiking transformer is divided into two branches to capture comprehensive global dependencies across the temporal and spectral dimensions. The encoder-decoder incorporates a multi-scale feature extractor, which extracts features at various scales, enabling the model to build a comprehensive hierarchical representation. This representation significantly enhances the model's ability to learn and process noisy speech, leading to excellent SE performance. Experiments are conducted using two publicly available benchmark datasets: WSJO-SI84 and VCTK+DEMAND. The proposed model demonstrated improved SE performance, showing significant progress with notable improvements of 33.69% in ESTOI, 1.05 in PESQ, and 11.36 dB in SDR over the noisy mixtures.

INDEX TERMS Speech Enhancement, Deep Learning, Spiking Transformer, Temporal Dynamics, Spiking Self-Attention (SSA), Speech Recognition, Convolutional Encoder-Decoder

I. INTRODUCTION

SPEECH enhancement is aimed at improving the overall quality and intelligibility of speech signals, particularly in the presence of background noise, reverberation, or other distortions. This process is fundamental in various applications, such as telecommunications, where it ensures more explicit voice communication over phones; hearing aids, which require enhanced speech signals for the hearing impaired; voice-controlled and ASR systems [1], [2], where accurate speech recognition is required for reliable operation. It also enhances the user experience in multimedia applications, such as video conferencing and streaming services, by providing clear and intelligible audio.

Classical speech enhancement methods, such as spectral subtraction [3], Wiener filtering [4], and statistical model-based techniques [5], have long been used to improve the intelligibility and quality of speech signals. Spectral subtraction estimates the noise spectrum during silent periods and subtracts it from the noisy speech, but it often introduces artifacts like musical noise. Wiener filtering optimizes the signal-to-noise ratio by applying a filter based on estimated speech and noise spectra, though it requires accurate noise estimation and can be less effective with non-stationary noise. Statistical model-based methods, such as Hidden Markov

Models (HMMs) or Gaussian Mixture Models (GMMs), leverage probabilistic models to separate speech from noise but typically require extensive training data and computational resources. Despite their widespread use, these classical methods face limitations in handling dynamic and complex noise environments, often resulting in residual noise and speech distortion, prompting the need for more advanced approaches [6]–[10].

Recent advances in deep neural networks (DNNs) have significantly transformed speech enhancement, offering superior performance compared to classical methods [11], [12]. Techniques such as Convolutional Neural Networks (CNNs) [13], [14], Recurrent Neural Networks (RNNs) [15], and Long Short-Term Memory (LSTM) [16] networks have been widely adopted to address the complex, non-linear relationships between noisy and clean speech signals. These networks are capable of learning intricate features and temporal dependencies from large datasets, leading to more effective noise suppression and improved speech intelligibility. Moreover, approaches like Generative Adversarial Networks (GANs) [17] and transformers [18] have further enhanced speech quality by refining the generation of clean speech from noisy inputs. Despite their computational demands, these deep learning models have demonstrated remarkable robustness in diverse and challenging noise environments.

Spiking Neural Networks (SNNs) [19]–[21] represent a promising paradigm for speech processing, leveraging principles inspired by biological neural networks to address key challenges in speech enhancement [7], [22], recognition [23], [24], and synthesis. Unlike conventional ANNs that use continuous-valued activations, SNNs operate based on the timing of discrete spikes, emulating the asynchronous and event-driven nature of the neurons. This allows SNNs to efficiently encode and process temporal information, making them particularly well-suited for tasks involving time-dependent data, such as speech signals. In speech processing, SNNs offer several advantages over traditional ANNs [25]. Firstly, their event-driven processing enables energy-efficient computation, making them suitable for deployment in low-power devices like smartphones and IoT devices. Additionally, the temporal coding of information in spike timings allows SNNs to capture fine-grained temporal patterns in speech signals and can improve speech processing in noisy environments.

There have been a few studies on speech enhancement in literature utilizing spiking neural networks such as [7], [22], [26]–[28]. The study in [28] introduces a three-layer SNN architecture with lateral inhibition and generates a training dataset by adding three levels of Gaussian white noise and computing the Short Time Fourier Transform (STFT) of the signal. The study encodes the log-scaled STFT magnitude into discrete spike timing using the Bens Spiker algorithm [29]. The SNN model processes the encoded input spikes using masking to eliminate uncorrelated spikes. This involves element-wise multiplication of the complex noisy STFT with the SNN's output spike train to obtain an enhanced STFT. Experimental results show favourable performance. However, the SNN lacks learning capabilities, and the architecture is relatively shallow. Additionally, further testing with diverse noise types is necessary to validate generalizability. The study in [26] proposes a similar approach, utilizing a three-layer SNN architecture with lateral inhibition. The study uses the log-scaled STFT magnitude as input to the SNN model, which produces a binary mask. The enhanced STFT is obtained through element-wise multiplication with the binary mask. Their experimental setup incorporates five distinct real-world noise types, demonstrating good performance in terms of various SNR. However, the speech enhancement process depends on eliminating uncorrelated noise components, and the SNN architecture lacks integration of any learning strategies. In a recent study [27], Intel introduced a basic SNN-based solution as a baseline for the Intel Neuromorphic Deep Noise Suppression Challenge (Intel N-DNS Challenge). This model uses a three-layer feedforward sigma-delta neural network to mask the STFT Magnitude. The delta-encoded STFT magnitude serves as the input to the SNN, which generates a multiplicative mask for enhancing the STFT. The baseline SNN undergoes training using the surrogate gradient method. Another study in [22] introduces a single-channel speech enhancement method employing a U-Net SNN architecture. The study reveals the capability of SNNs to handle large-scale regression tasks like speech enhancement. The objective evaluations show that the proposed approach outperforms multiple state-of-the-art ANN-based models and outperforms the baseline solution of the Intel N-DNS Challenge. Additionally, the model attains competitive results compared to an equivalent ANN architecture, highlighting the potential of SNN for speech enhancement. A very recent study [7] presents Spiking-S4, a lightweight SNN model specifically formulated for speech enhancement. The study draws on the pioneering model of integrating a structured state space model with spiking neural networks for speech enhancement. Through the evaluation of two benchmark datasets, the model validates that the Spiking-S4 model achieves competitive results with traditional ANN models while demonstrating excellent computational efficiency.

In traditional neural networks, information is typically processed continuously. In spiking neural networks, information is encoded in discrete, asynchronous spikes, similar to neurons communicating in the brain. ANNs have shown outstanding results in speech enhancement, especially when equipped with substantial computational power. However, they often show limitations in processing long-term temporal sequences effectively. On the other hand, spiking neural networks have a natural ability to manage temporal data and are excellent at capturing complex temporal patterns encountered in speech signals. This paper introduces a novel model integrating self-attention with a spiking neural network to enhance speech quality and intelligibility. The model architecture consists of a convolutional encoder-decoder framework, with a spiking transformer serving as a bottleneck. In this framework, the spiking self-attention mechanism is employed to encode sparse features using spike-based representations of query (**Q**), key (**K**), and value (**V**). This strategy effectively captures temporal dependencies and contextual relationships in the speech signal, thereby enhancing the feature representation. Moreover, the encoder-decoder architecture incorporates a multi-scale feature extractor, which extracts features across different scales. This enables the model to build a comprehensive hierarchical representation of the input signal. By leveraging this representation, the model significantly improves its ability to learn and process noisy speech, resulting in superior speech enhancement performance. The following are the contributions to this study:

- We present a novel approach to speech enhancement using a temporal dynamic spiking transformer as a bottleneck network connecting a convolutional codec. The temporal dynamic spiking transformer model integrates the temporal dynamics of speech signals with the powerful processing capabilities of spiking neural networks and transformers, implying considerable improvements in speech quality and intelligibility. Unlike traditional DNNs that may overlook the temporal characteristics of speech, our model prioritizes capturing and processing temporal dependencies, leading to more contextually relevant enhancements.
- The bottleneck is divided into two spiking transformer branches to capture comprehensive global dependencies across the temporal and spectral dimensions. This dual-branch approach ensures the model comprehensively understands and processes the complex patterns in temporal and spectral features.
- We provide a detailed examination of the computational load of the proposed speech enhancement model, presenting a detailed comparison of model complexity, inference time, and memory footprint. Through a detailed evaluation, we provide an understanding of the trade-offs between performance improvement and resource requirements. By quantifying the architectural complexities, benchmarking its real-world performance, and analyzing its memory utilization patterns, we show the balance between speech enhancement performance and managing resource limitations.
- we present a comprehensive evaluation of the proposed speech enhancement model using two widely recognized benchmark datasets: WSJO-SI84 and VCTK+DEMAND. Our evaluation contains a comprehensive analysis of model performance across two datasets, leveraging their high-quality and phonetically balanced attributes. By conducting experiments on these benchmark datasets, we aim to provide a robust review of the model under diverse conditions and contexts.

The subsequent sections of this paper are organized as follows: Section 2 presents the problem formulation. Section 3 explains the proposed speech enhancement model, providing detailed descriptions of its various modules. In Section 4, we outline the experimental settings. The findings and analyses are presented in Section 5. Finally, Section 6 concludes the study.

II. PROBLEM FORMULATION

The single-channel speech enhancement problem aims to improve the quality and intelligibility of speech recorded in noisy or degraded acoustic environments using one microphone. Given a noisy speech signal $y(n)$, where n denotes the discrete time index, the objective is to estimate the clean speech signal $s(n)$ by minimizing noise signals $d(n)$; $y(n) = s(n) + d(n)$. The SE problem can be expressed as finding an optimal complex mapping $\hat{s}(n)$ that minimizes the difference between the estimated clean speech and the observed noisy speech while considering the characteristics of both the speech signal and the background noise. Mathematically, this can be formulated as:

$$\hat{s}(n) = \underset{\hat{s}(n)}{argmin} \mathbb{E}\left\{ \|s(n) - \hat{s}(n)\|^2 \right\} \tag{1}$$

Where $\mathbb{E}$ denotes the expectation operator. This is subject to the constraint that the estimated speech signal $\hat{s}(n)$ accurately represents the underlying clean speech while mitigating the impact of the additive noise $d(n)$. This problem is challenging due to the presence of various types of noise sources, non-stationary acoustic environments, and the need to maintain the temporal and spectral characteristics of the speech during enhancement.

III. PROPOSED SPEECH ENHANCEMENT

The network architecture of the proposed speech enhancement model is illustrated in Fig. 1(A). The encoder processes the combined real and imaginary parts of the mixture spectrogram $Y \in \mathbb{R}^{(2\times T\times F)}$ as input and outputs an estimated complex spectrum $\hat{S} \in \mathbb{R}^{(2\times T\times F)}$, where T denotes time frames and F represents frequency bins. The encoder consists of four convolutional layers composed of multi-scale feature extraction blocks, followed by a downsampling operation that reduces the frequency dimension of the feature maps by half $(1/2)$ [30]. Afterwards, the feature maps are processed by a dual-branch spiking transformer network (STN), which sequentially captures contextual information along the time and frequency dimensions. These reshaped feature maps are then passed to the decoder. Skip connections are incorporated between the encoder and decoder to improve the flow of gradients and information throughout the network. The decoder comprises a sub-pixel convolutional layer [31] and four layers of multi-scale feature extraction blocks. Ultimately, the decoder outputs the real and imaginary components ($\hat{S}r$, $\hat{S}i$) of the estimated spectrum, which are used to reconstruct the target speech waveform via inverse STFT.

A. MULTI-SCALE FEATURE EXTRACTION

The encoder has four convolutional layers, each with a multi-scale feature block (MSFB), as shown in Fig.1(B) [30], [32]. Each MSFB includes two individual convolutional layers and

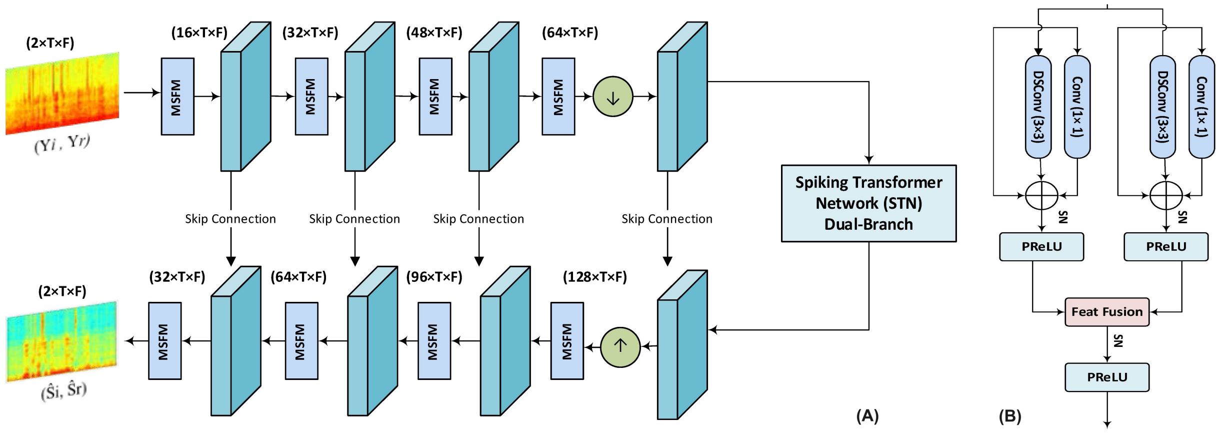

FIGURE 1. (A): Network structure of the Proposed Speech Enhancement Model. ↓ in the circle indicates downsampling containing a 2D convolutional layer followed by switchable normalization (SN) and PReLU, whereas ↑ in the circle indicates upsampling. (B) The framework of the multi-scale feature extraction block.

a feature fusion block (FFB). This is followed by a parametric ReLU (PReLU) activation and switchable normalization (SN) operations. Unlike traditional normalization, switchable normalization computes means and variances from different proportions. It dynamically switches between them by learning the importance of weights for each proportion. The convolutional unit employs a multi-branch structure that includes 2D depthwise separable convolutions (DSConv), pointwise convolutions, and residual connections. DSConv breaks the convolution into two separate processes and reduces the computational complexity. Pointwise convolution is a (1×1) convolution that adjusts the number of channels. These convolutions are followed by SN and PReLU activation. The depthwise separable convolutions use kernels of sizes (3×3) and (5×5) respectively. By using branches in the multi-scale feature blocks with varying levels of complexity (different kernel sizes), it captures features at multiple scales and with different receptive fields. This multi-branch approach improves the feature space, implying it can detect and represent a wider variety of features. It enhances the expressive power of the convolutions, allowing the model to better understand and process complex patterns.

To efficiently integrate features of distinct scales from the two convolutional units, this study uses the feature fusion block (FFB) [33]. The features $fb_1^{(C\times T\times F)}$ and $fb_2^{(C\times T\times F)}$ of two convolutional units are fed to the feature fusion block which selectively processes the feature maps. The feature fusion block (FFB) is depicted in Fig. 2. Feature maps FB1 and FB2 are concatenated along the channel dimension, and then processed through a linear layer followed by a sigmoid activation. This produces a gating parameter w for fb_1 and (1-w) for fb_2, respectively; given as:

$$w = \sigma(Linear(Concat[fb_1, fb_2])) \tag{2}$$

$$f = PReLU((w \otimes fb_1 + (1-w) \otimes fb_2) + fb_1 + fb_2) \tag{3}$$

Here $f^{(C\times T\times F)}$ denotes the output of the FFB.

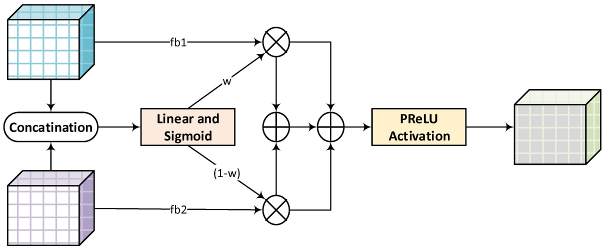

FIGURE 2. Diagram of the feature fusion block (FFB) which selectively processes features from two convolutional units. Where w indicates a weighted factor.

B. SPIKING TRANSFORMER NETWORK (STN)

Traditional deep learning models represent information through continuous decimal values, whereas Spiking Neural Networks (SNNs) operate using discrete spike sequences for processing and transmitting data. These spiking neurons receive continuous inputs and convert them into spike patterns. This study uses the spiking transformer, which integrates the self-attention and transformer architecture into spiking neural networks to enhance temporal learning capabilities. The traditional multi-head self-attention (MHSA) is replaced by spiking self-attention (SSA). This mechanism substituted the conventional activation function with spiking neurons and eliminated the softmax function usually used before attention calculation [34]. The STN framework is depicted in Fig. 4. The central component of the spiking transformer architecture is its encoder, which integrates the spiking self-attention with an MLP block. Since the softmax function and the floating-point matrix multiplication of query (Q) and key (K) disregard SNN computational rules, the MHSA (Q and V) computations are not viable in SNNs. Further, the efficient computation required by SNNs is undermined by the quadratic space and time complexity related to the sequence length of MHSA. On the other hand, the spiking self-attention uses spike-based Q and K values. Firstly, the Q, K, and V are computed using learnable matrices. These

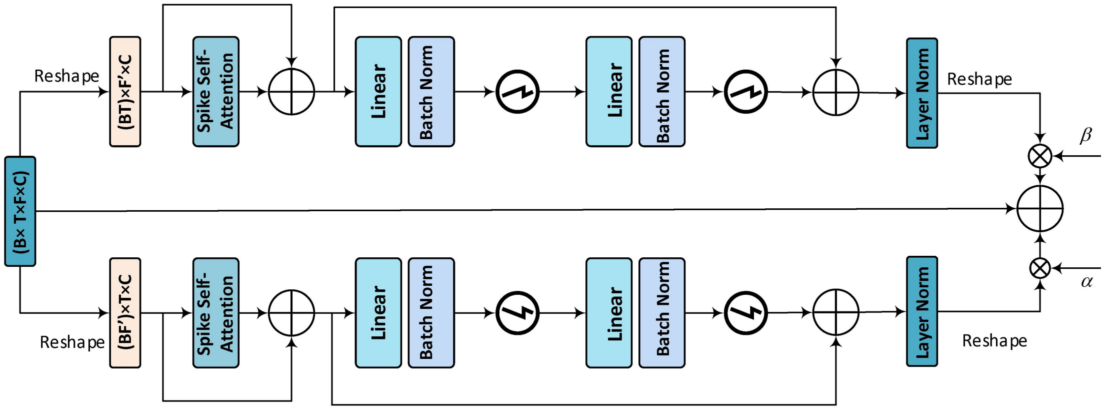

FIGURE 3. Diagram of the dual-STN with Temporal Attention Branch (TAB) and Frequency Attention Branch (FAB), controlled by α and β.

are subsequently transformed into spiking sequences through distinct layers of spike neurons; given as:

$$Q = SN_Q(BN(XW_Q)) \tag{4}$$

$$K = SN_K(BN(XW_K)) \tag{5}$$

$$V = SN_V(BN(XW_V)) \tag{6}$$

Where SN is the spike neuron layer and BN shows the batch normalization. The computations of the attention matrix purely contain spike-formed Q and V containing only 1 and 0. The spike-based self-attention can be defined as:

$$SSA'(Q, K, V) = SN(QK^T V) \tag{7}$$

$$SSA(Q, K, V) = SN(BN(Linear(SSA'(Q, K, V)))) \tag{8}$$

Based on Equations (4-6), the spike neuron layers (SN) produce spike sequences Q and K, inherently yielding non-negative values (0 or 1), thereby leading to a non-negative attention map. Spiking self-attention selectively aggregates relevant features while disregarding irrelevant details.

Similar to the study in [35], the bottleneck is divided into two spiking transformer branches to capture comprehensive global dependencies across the temporal and spectral dimensions. This dual-branch approach ensures the model comprehensively understands and processes the complex patterns in temporal and spectral features. As shown in Fig. 3, the bottleneck consists of two separate sub-branches that function simultaneously along the time and frequency dimensions; a Temporal Attention Branch (TAB) and the Frequency Attention Branch (FAB). These branches are effective in capturing extensive global dependencies across temporal and spectral dimensions by incorporating two adaptive weights, labelled as α and β. In Fig. 3, B, T, F, C, denote the batch size, the frame number, the frequency dimension, and the channel number, respectively. The final output is followed by PReLU activation and convolutional 2D layer; given as:

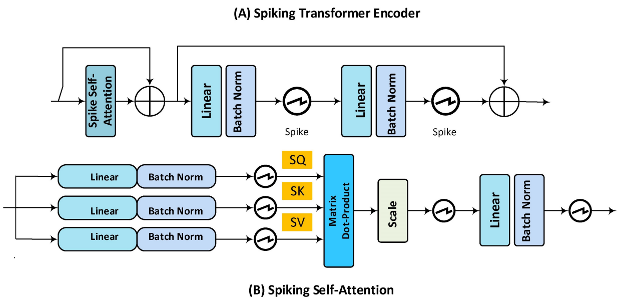

FIGURE 4. Diagram of the Spiking Transformer Network (STN) with Spiking Self-Attention (SSA). Here SQ, SK, and SV denote spiked query, key, and value.

$$f_{output} = f_{in} + \alpha Out_{FAB} + \beta Out_{TAB} \tag{9}$$

Where α and β are initialized as 1, which are adjusted adaptively to suitable values.

IV. EXPERIMENTS

A. DATASETS

We evaluate the performance of the proposed speech enhancement model called TD-STNet on the WSJ0-SI84 dataset, which includes 7,138 clean sentences from 83 speakers (42 male and 41 female). For this study, we randomly selected 3,000 training sentences and 1,000 validation sentences from these 80 speakers. Additionally, we create two test sets, each containing 200 sentences from 3 male and 3 female untrained speakers. To generate clean-noisy pairs, we use various noise types sourced from the Perception and Neurodynamics Laboratory and the Laboratory for Recognition and Organization of Speech and Audio. During the mixing process, a random section of noise is extracted and mixed with a randomly selected sentence at SNRs ranging from -5dB to 5dB in 1dB increments. This results in clean-noisy pairs for both training and validation, with the total duration of the training set amounting to approximately 150 hours. To show the generalization abilities of the model, we use two challenging untrained noise types (babble and factory1 from NOISEX-92). For network evaluation, we employ four testing SNRs: -3dB, 0dB, 3dB, and 6dB, yielding 300 clean-noisy pairs for each SNR.

We evaluate the performance of the proposed TD-STNet framework using the publicly available VCTK+DEMAND dataset. The training set includes sentences from 28 speakers, while the testing set features sentences from 2 speakers, each contributing approximately 400 sentences. During training, the sentences are mixed with 10 different noise types at four SNR levels (0 dB, 5 dB, 10 dB, and 15 dB), creating a total of 11,572 clean-noisy mixtures. For testing, the sentences are mixed with 5 noise types at SNRs of 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB, resulting in a testing set with 824 clean-noisy mixtures. Importantly, both the speakers and noise types in the testing set are untrained (unseen) in the training set.

B. NETWORK SETTINGS AND TRAINING

The convergence of a deep neural network significantly relies on optimal weight initialization. In this study, we utilise the Glorot-Uniform (Xavier) initializer to initialize the network. The encoder consists of four convolutional layers with MSFM. The decoder mirrors the encoder's structure, containing four sub-pixel deconvolutional layers. We use a batch size of 16 for training. The network is optimized using the Adam optimizer with carefully tuned default parameters: $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$, where α, β_1, and β_2 are the step size and the exponential decay rates for the first and second-moment estimates, respectively. The initial learning rate γ is set at 0.001 and dynamically decays based on the average training loss. When the average loss decreases by a factor of ten, the learning rate is successively reduced to 0.0005, 0.0002, and 0.0001. The model is trained for 100 epochs with a dropout rate of 20%. The noisy mixtures are sampled at 16 kHz with a window length of 512 samples. With a 50% overlap between frames, the frame-shift is set to 16 ms.

Estimating the complex mapping function F_ϕ requires optimizing a weighted combination of time and frequency loss, as described by [36], which is given by:

$$L = \upsilon * L_{time} + (1-\upsilon) * L_{frequency} \qquad (10)$$

Where υ indicates an adaptable parameter, which is fixed empirically as 0.4. The loss functions are provided as:

$$L_{time} = \frac{1}{N}\sum_{m=0}^{N-1}(s[n]) - \hat{s}[n])^2 \qquad (11)$$

$$L_{freq} = \frac{1}{TF}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}[(|S_r| + |S_i|) - (|\hat{S}_r| + |\hat{S}_i|)] \qquad (12)$$

Where $s(n)$ and $\hat{s}(n)$ indicate the clean speech and enhanced version of noisy speech, with N denoting the sample number. Whereas (we omit (t, f) due to space limitations) $S^{(t,f)}$ and $S^{(\widehat{t,f})}$ symbolise clean and enhanced spectrograms with real (r) and imaginary (i) parts.

C. METRICS AND BENCHMARKS

This section outlines the metrics and benchmarks used to quantitatively assess SE performance. The metrics include PESQ (Perceptual Evaluation of Speech Quality) [37], ESTOI (Extended Short-Time Objective Intelligibility) [38], and SI-SDR (Scale-Invariant Signal-to-Distortion Ratio). PESQ, with scores ranging from -0.5 to 4.5, measures speech quality, where higher scores indicate better quality. In this study, narrow-band PESQ is applied to the WSJ0-SI84+DNS dataset, while wide-band PESQ is used for the VCTK+DEMAND dataset. ESTOI scores range from 0 to 1, with 1 indicating perfect intelligibility. Higher SDR values denote better performance. Segmental SNR (SNRSeg) and frequency-weighted SNRSeg (FW-SNRSeg) are also used to evaluate the quality and intelligibility [39]. For the WSJ0-SI84 dataset, this study compares the proposed SE against several benchmarks: CRN [40], DPRNN [41], GCRN [42], DCCRN [43], AECNN [44], CTS-Net [45], GaGNet [46], and CPB-Net [35]. For the VCTK+DEMAND dataset, the baselines include PHASEN [47], RDL-Net [48], DEMUCS [49], GaGNet [46], TSTNN [50], MSSA-TCN [51], FAF-Net [52], PFRNet [53], CTS-Net [45], U-shaped transformer (UT-FAT) [54], dual-branch state space (DB-S4D)-based SE [55], DB-CRN [56], and SADN-UNet [57].

V. RESULTS AND ANALYSIS

A. EVALUATION ON WSJ0-SI84 DATASET

This section assesses the performance of the proposed speech enhancement model against recent benchmarks.

Table 1 compares the speech enhancement performance of TD-STNet and benchmarks in settings with seen speakers and noise backgrounds. On average, AECNN shows the lowest SE performance (PESQ and ESTOI) among benchmarks, with a ΔESTOI of 31.81% and ΔPESQ of 0.85, but a better ΔSI-SDR of 9.99 dB than CRN (7.69dB). GCRN, with its improved convolutional encoder-decoder (CED) in the complex domain, achieves better results: ΔSI-SDR of 10.11dB, ΔESTOI of 36.22%, and ΔPESQ of 1.05. DPRNN, using dual-path processing, outperforms CRN and AECNN, showing a ΔSI-SDR of 10.30dB, ΔESTOI of 33.21%, and ΔPESQ of 0.98. DCCRN, with a complex CED, also improves over CRN and AECNN, achieving a ΔSI-SDR of 10.16dB, ΔESTOI of 32.73%, and ΔPESQ of 0.98. CTS-Net and CPB-Net, with advanced architectures and feature

TABLE 1. Speech enhancement performance of the TD-STNet and benchmarks in seen speakers and seen noisy background settings. Seen conditions include stationary and non-stationary noises: babble, street, F16, and airport noise.

Measure	Input Feature	Year	ESTOI (%)				SI-SDR (dB)				PESQ			
SNR (dB)	–	–	-3dB	0dB	3dB	Avg	-3dB	0dB	3dB	Avg	-3dB	0dB	3dB	Avg
Mixture	–	–	31.80	40.47	49.36	40.54	-2.87	0.13	3.12	0.13	1.66	1.87	2.08	1.87
CRN [40]	Magnitude	2018	66.13	73.50	79.20	72.94	5.38	7.74	9.96	7.69	2.49	2.75	2.96	2.73
AECNN [44]	Waveform	2019	64.54	73.20	79.33	72.36	7.71	10.13	12.14	9.99	2.43	2.76	2.99	2.72
GCRN [42]	Real-Imag	2019	69.74	77.59	82.95	76.76	7.77	10.28	12.29	10.11	2.64	2.94	3.18	2.92
DPRNN [41]	Waveform	2020	65.93	74.21	81.11	73.75	7.94	10.39	12.56	10.30	2.55	2.89	3.14	2.86
DCCRN [43]	Real-Imag	2020	65.40	73.79	80.62	73.27	7.75	10.33	12.41	10.16	2.54	2.88	3.12	2.85
CTS-Net [45]	Real-Imag+Mag	2021	74.48	80.76	84.84	80.03	9.44	11.67	13.48	11.53	2.84	3.12	3.32	3.09
GaGNet [46]	Real-Imag	2022	68.65	76.61	82.22	75.83	8.97	11.23	13.32	11.17	2.58	2.90	3.17	2.88
CPB-Net [35]	Real-Imag	2022	72.38	76.39	84.04	77.60	9.27	11.59	13.03	11.30	2.77	3.07	3.28	3.04
TD-STNet	Real-Imag	2024	74.44	81.13	85.09	80.22	9.28	11.80	13.49	11.52	2.80	3.16	3.30	3.09

TABLE 2. SE performance of the TD-STNet and benchmarks in unseen speakers and unseen noisy background settings. Unseen conditions include stationery and non-stationary noises: babble1, Car, and factory1 noise.

Measure	Input Feature	Year	ESTOI (%)				SI-SDR (dB)				PESQ			
SNR (dB)	–	–	-3dB	0dB	3dB	Avg	-3dB	0dB	3dB	Avg	-3dB	0dB	3dB	Avg
Mixture	–	–	30.58	39.16	48.18	39.31	-3.07	-0.08	2.95	-0.07	1.42	1.65	1.89	1.65
CRN [40]	Magnitude	2018	61.96	70.63	76.80	69.80	5.16	7.71	9.99	7.62	2.17	2.49	2.73	2.46
AECNN [44]	Waveform	2019	62.21	71.54	78.10	70.62	7.13	9.89	11.95	9.66	2.17	2.51	2.82	2.50
GCRN [42]	Real-Imag	2019	66.53	76.28	81.95	74.92	7.54	10.06	12.05	9.88	2.38	2.76	2.97	2.70
DPRNN [41]	Waveform	2020	65.25	73.90	80.89	73.35	7.66	10.13	12.29	10.03	2.30	2.68	2.96	2.65
DCCRN [43]	Real-Imag	2020	63.88	72.78	79.29	71.98	7.51	10.06	12.16	9.91	2.30	2.62	2.91	2.61
CTS-Net [45]	Real-Imag+Mag	2021	72.79	79.13	84.14	78.69	9.10	11.45	13.21	11.25	2.61	2.93	3.11	2.87
GaGNet [46]	Real-Imag	2022	67.89	76.05	81.89	75.28	8.73	11.00	13.09	10.94	2.34	2.69	2.95	2.66
CPB-Net [35]	Real-Imag	2022	71.68	76.72	83.90	77.43	9.07	11.38	12.83	11.09	2.54	2.88	3.09	2.84
TD-STNet	Real-Imag	2024	73.63	79.08	84.89	79.20	9.03	11.53	13.23	11.26	2.56	2.95	3.12	2.89

types, demonstrate excellent SE performance, with CTS-Net achieving a ΔSI-SDR of 11.53dB, ΔESTOI of 39.48%, and ΔPESQ of 1.21, and CPB-Net showing a ΔSI-SDR of 11.30dB, ΔESTOI of 37.06%, and ΔPESQ of 1.17.

TD-STNet surpasses most benchmarks, except for CTS-Net, where it falls behind in some metrics and SNRs. TD-STNet achieves average values of ΔSI-SDR=11.52dB, ΔESTOI=39.67%, and ΔPESQ =1.22, outperforming CRN by 3.83dB in ΔSI-SDR, 7.28% in ΔESTOI, and 0.35 in ΔPESQ. It also outperforms GCRN by 1.41dB in ΔSDR, 3.46% in ΔESTOI, and 0.17 in ΔPESQ, whereas exceeds GaGNet by 0.35dB in ΔSI-SDR, 4.39% in ΔESTOI, and 0.2 in ΔPESQ. TD-STNet exceeds two time-domain SE benchmarks, with improvements of 1.53dB in ΔSI-SDR, 7.86% in ΔESTOI, and 0.36 in ΔPESQ over AECNN, whereas 1.23dB in ΔSI-SDR, 6.47% in ΔESTOI, and 0.23 in ΔPESQ over DPRNN. At low SNR conditions of -3dB, TD-STNet achieves better speech quality than GaGNet, CPB-Net, DCCRN, and GCRN, with ΔSI-SDR improvements of 0.31, 0.01, 1.53, and 1.51, ΔESTOI improvements of 5.79%, 2.06%, 9.04%, and 4.7%, and ΔPESQ improvements of 0.22, 0.03, 0.26, and 0.16. TD-STNet and CTS-Net show competitive performance; TD-STNet shows overall speech enhancement performance through multi-scale feature learning and effective global information capturing.

Table 2 presents the average SE performance of TD-STNet and benchmarks in scenarios with unseen speakers and noisy backgrounds (car, factory1, and babble1). The results demonstrate that TD-STNet wildly outperforms time-domain benchmarks (AECNN and DPRNN) and single-branch convolutional encoder-decoder benchmarks (CRN, GCRN, and DCCN) across all metrics in unseen noisy and speaker scenarios. An average improvement of approximately ΔPESQ=1.22, ΔESTOI = 38.39%, and ΔSI-SDR=11.20dB is observed over noisy mixtures with unknown speakers. At -3dB low SNR, TD-STNet shows an average improvement of ΔPESQ=0.18, ΔESTOI=7.10%, and ΔSDR=1.49dB over GCRN. Additionally, at a favourable 3dB SNR, TD-STNet surpasses GaGNet by ΔPESQ=0.17, ΔESTOI = 3%, and ΔSI-SDR=0.14dB. CTS-Net performs marginally better in a few specific conditions, such as achieving PESQ=2.61 at -3dB which is better than the proposed model by factor ΔPESQ=0.05, ESTOI=74.48% (ΔESTOI=0.04% greater than CTS-Net), and SI-SDR = 9.44dB (ΔSDR=0.07dB). Overall, Table 2 highlights the superior performance of TD-STNet and benchmarks in real-world applications where SE must handle various unseen noise types and speaker characteristics.

Additionally, segmental SNR (SNRSeg) and frequency-weighted SNRSeg (FW-SNRSeg) are employed to evaluate the quality and intelligibility of noisy speech. SNRSeg provides a detailed analysis by assessing segments of the speech signal rather than the entire signal at once. FW-SNRSeg expands SNRSeg by incorporating frequency weighting to better reflect human auditory perception. The additional metrics are as follows:

TABLE 3. Performance on the VCTK+DEMAND database. "–" means no results were supplied in the original study. The symbol "Δ" means improvements.

Models	Para#	STOI	PESQ	Covl	Csig	Cbak	SNRSeg	ΔSTOI	ΔPESQ	ΔCovl	ΔSNRSeg
Mixture	–	91.6	1.97	2.63	3.34	2.44	1.69	–	–	–	–
PHASEN [47]	8.76M	–	2.99	3.62	4.21	3.55	7.66	–	1.02	0.99	5.97
RDL-Net [48]	3.91M	93.8	3.02	3.72	4.38	3.43	–	2.2	1.05	1.09	–
DEMUCS [49]	128M	95.1	3.07	3.63	4.31	3.40	8.53	3.5	1.1	1	6.84
GaGNet [46]	5.94M	94.7	2.94	3.59	4.36	3.45	9.24	3.1	0.97	0.96	7.55
TSTNN [50]	0.92M	95.1	2.96	3.49	4.17	3.53	9.72	3.5	0.99	0.86	8.03
MSSA-TCN [51]	9.91M	94.0	3.02	3.67	4.29	3.50	–	2.4	1.05	1.04	–
FAF-Net [52]	6.90M	95.0	3.19	3.66	4.13	3.38	–	3.4	1.22	1.03	–
PFRNet [53]	4.61M	95.0	3.24	3.90	4.48	3.70	–	3.4	1.27	1.27	–
CTS-Net [45]	4.35M	–	2.92	3.59	4.25	3.46	–	–	0.95	0.96	–
UT-FAT [54]	4.31M	–	3.08	3.68	4.23	3.63	11.69	–	1.11	1.05	10
DB-S4D [55]	10.80M	93.4	2.55	3.23	3.94	3.00	–	1.8	0.58	0.6	–
DB-CRN [56]	8.31M	94.0	3.16	3.62	4.07	3.68	10.98	2.4	1.19	0.99	9.29
SADN-UNet [57]	2.63M	95.0	2.82	3.51	4.18	3.47	–	3.4	0.85	0.88	–
TD-STNet	3.14M	95.0	3.13	3.82	4.29	3.66	11.12	3.4	1.16	1.19	9.43

$$SNRSeg = \frac{10}{W}\sum_{w=0}^{W-1} log_{10}\frac{s^2(n)}{(s(n)-\hat{s}(n))^2} \tag{13}$$

$$FWS = \frac{10}{W}\sum_{w=0}^{W-1}\frac{\sum_{i=1}^{L} M(i,n)log_{10}\frac{S(i,n)^2}{(S(i,n)-\hat{S}(i,n))^2}}{\sum_{i=1}^{K} M(i,n)} \tag{14}$$

Here FWS denotes FW-SRNSeg. The weight $M(i,n)$ indicates the emphasis on the i^{th} frequency bin, where L is the total number of bands, and W is the total number of frames in the signal. $S(i,n)$ represents the excitation spectrum of the clean signal in i^{th} frequency bin at n^{th} frame.

Figure 5 illustrates the average speech enhancement performance of TD-STNet in seen and unseen conditions for speakers and SNRs, evaluated using SNRSeg and FW-SNRSeg. In seen noisy mixtures for known speakers at low SNR of -3dB, an average improvement of ΔSNRSeg = 3.74dB and ΔFW-SNRSeg=5.71dB is observed. In unseen noisy mixtures for unknown speakers at -3dB SNR, the average improvement is ΔSNRSeg=5.01dB and ΔFW-SNRSeg=7.27dB.

B. EVALUATION ON VCTK+DEMAND DATASET

This section compares the performance of the proposed method with recent time and time-frequency domain benchmarks using PESQ, STOI, Csig, Cbak, Covl, and SNRSeg. Csig, Cbak, and Covl suggest the mean opinion score (MOS) for speech distortion, background noise intrusiveness, and overall speech quality, respectively. For the VCTK+DEMAND dataset, the benchmarks include PHASEN [47], RDL-Net [48], DEMUCS [49], GaGNet [46], TSTNN [50], MSSA-TCN [51], FAF-Net [52], PFRNet [53], CTS-Net [45], U-shaped transformer (UT-FAT) [54], dual-branch state space (DB-S4D)-based SE [55], DB-CRN [56], and SADN-UNet [57]. Table 3 outlines the results for the VCTK+DEMAND dataset. "–" denotes missing results in the original paper, and "Δ" indicates improvement in PESQ, STOI, Covl, and SNRSeg.

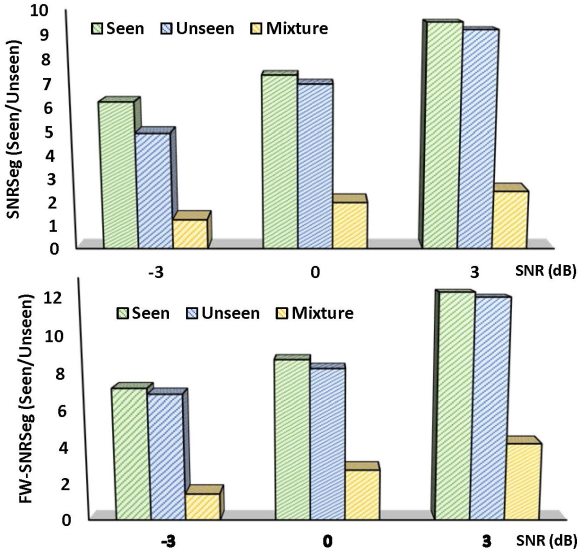

FIGURE 5. SNRSeg and FW-SNRSeg for Seen and Unseen conditions.

Based on the results in Table 3, TD-STNet demonstrates competitive performance compared to benchmarks across multiple metrics (PESQ, STOI, Csig, Cbak, Covl, and SNRSeg). TD-STNet yields average ΔPESQ=0.15 (over PHASEN) and ΔPESQ=0.12 (over DEMUCS), whereas ΔCovl=0.19 and ΔCovl=0.2 over PHASEN and DEMUCS, respectively. Compared to RDL-Net, TD-STNet shows ΔPESQ=0.11, ΔSTOI=1.2%, ΔCbak=0.23, and ΔCovl=0.10. From GaGNet, TD-STNet results in ΔPESQ=0.19, ΔSTOI=0.3%, ΔCovl=0.23, and ΔSNRSeg= 1.88dB. Additionally, TD-STNet improves upon MSSA-TCN and CTS-Net with ΔPESQ=0.11 and 0.21, whereas ΔCovl=0.15 and 0.23, respectively. These findings highlight the ability of TD-STNet to enhance quality, intelligibility, and noise reduction. Compared to state-of-the-art time-domain baselines, TD-STNet consistently outperforms across all objective metrics. Compared to TSTNN, TD-STNet achieves ΔPESQ=0.17, ΔCsig=0.12, ΔCovl=0.33,

TABLE 4. Computational efficiency for real-time processing vs performance evaluating metrics on the VCTK+DEMAND database. since no future frames are involved during processing, this amounts to a causal SE model. "–" denotes no results provided in the original paper. The symbol "Δ" indicates improvement in PESQ and STOI.

Model	Year	Feat	MACs	RTFs	Para #	PESQ	STOI	ΔPESQ	ΔSTOI
Mixture	–	–	–	–	–	1.97	91.6	–	–
DCCRN [43]	2020	Real-Imaginary	14.36[G/s]	2.19	3.67M	2.54	93.8	0.57	2.2
NSNet [58]	2021	Magnitude	0.43[G/s]	0.02	6.16M	2.47	92.3	0.5	0.7
GaGNet [46]	2021	RI+Magnitude	1.65[G/s]	0.05	5.95M	2.94	–	0.97	–
DPT-FSNet [59]	2022	Magnitude	—	—	0.88M	3.20	95.0	1.23	3.4
DPTGAN [60]	2022	Magnitude	—	—	24.74M	2.86	94.0	0.89	2.4
FRCRN [61]	2022	Real-Imaginary	12.30[G/s]	–	10.27M	3.21	–	1.24	–
FullSubNet+ [62]	2022	RI+Magnitude	30.06[G/s]	0.55	8.67M	2.88	94.0	0.91	2.4
DF-Net [63]	2022	Magnitude	0.35[G/s]	0.11	2.31M	2.81	94.2	0.84	2.6
DPT-ECA [64]	2023	Magnitude	17.85[G/s]	—	2.17M	3.17	95.0	1.2	3.4
TD-STNet	2024	Real-Imaginary	9.64[G/s]	0.18	3.14M	3.13	95.0	1.16	3.4

TABLE 5. Comparisons of parameter size (Million), MACs (G/s), CPU-PT (sec), MFP, and TBT (sec).

Models	Model Size	MACs	CPU-PT	FMFP	BMFP	TBT
CRN [40]	17.58M	2.57 (G/s)	0.27 sec	0.18	0.33	0.07 sec
GCRN [42]	9.77M	2.42 (G/s)	0.17 sec	0.19	0.22	0.15 sec
DCCRN [43]	3.67M	14.36 (G/s)	0.40 sec	0.66	0.67	0.27 sec
TD-STNet	3.14M	9.64 (G/s)	0.20 sec	0.28	0.29	0.25 sec

TABLE 6. Generalization ability of the proposed speech enhancement towards unknown speakers and background noises.

Noise	32-Talker Babble Noise				Exhibition Hall Noise				Crowd Laughter Noise				Rain Thunder Noise			
Measure	PESQ		ESTOI		PESQ		ESTOI		PESQ		ESTOI		PESQ		ESTOI	
SNR	-3dB	3dB	-3dB	3dB	-3dB	3dB	-3dB	3dB	-3dB	3dB	-3dB	3dB	-3dB	3dB	-3dB	3dB
Mixture	1.61	1.95	33.01	50.91	1.54	1.92	32.52	50.39	1.65	1.95	33.59	51.02	1.64	1.94	33.57	51.00
CRN [40]	2.26	2.74	65.67	78.87	2.18	2.69	65.24	78.39	2.31	2.76	66.19	78.94	2.30	2.75	66.17	78.92
GCRN [42]	2.45	2.97	69.72	82.97	2.39	2.93	69.26	82.51	2.48	2.98	70.27	83.02	2.47	2.97	70.25	83.00
DCCRN [43]	2.33	2.89	64.87	79.54	2.26	2.86	64.36	79.08	2.37	2.89	65.47	79.59	2.36	2.88	65.45	79.57
TD-STNet	2.51	3.07	71.89	83.98	2.47	3.05	71.47	83.58	2.52	3.06	72.40	83.97	2.51	3.05	72.38	83.95

and ΔSNRSeg=1.4dB. PFRNet and FAF-Net show improvements in PESQ (0.11 and 0.06) over TD-STNet, but they show higher computational complexity with an additional 1.47M and 3.76M parameters. UT-FAT reaches superior SNRSeg (ΔSNRSeg=10dB)m which is 0.57dB higher than TD-STNet but with 1.18M additional parameters. From DBCRN, TD-STNet results in average improvements of 0.14dB in SNRSeg, 0.1% in STOI, and 0.2 in Covl, respectively.

C. COMPUTATIONAL COMPLEXITY ANALYSIS

We evaluate the computational efficiency of the proposed speech enhancement model for real-time processing, emphasizing its causal nature (processing only current frames). The evaluation includes analysis of trainable parameters, multiply-accumulate operations (MACs), and real-time factor (RTF), measured on a Core i5-1135G7@ 2.4GHz CPU. The computational efficiency of TD-STNet for real-time processing is compared against several benchmarks: DPT-FSNet [59], DPTGAN [60], DPT-ECA [64], NSNet [58], FullSubNet+ [62], DCCRN [43], GaGNet [46], FRCRN [61], and DF-Net [63]. Table 4 displays the parameter count, MACs, and RTF for TD-STNet and benchmarks. TD-STNet features approximately 3.14 million parameters, 9.64(G/s) MACs, and 0.18 RTF, outperforming DCCRN (0.53M parameters, 4.72G/s MACs, 2.01 RTF) and FullSubNet (5.53M parameters, 20.42G/s MACs, 0.37 RTF). NSNet achieves superior MACs (0.43G/s) and RTF (0.02) with 3.02M additional parameters, a small PESQ (2.47), and STOI (0.923) compared to TD-STNet. DF-Net shows lower MACs (0.35G/s), RTF (0.11), and parameters (2.31M); however, TD-STNet exceeds in ΔPESQ=0.32 and ΔSTOI=0.80%. TD-STNet requires approximately 10.12MB of memory with 3.14M parameters. The inference speed was evaluated using an Intel® Core™ i5-1135G7 CPU@2.40 GHz for processing 16000 audio samples. Training batch time (TBT) and memory usage were assessed on an NVIDIA GeForce GTX 1650Ti with 4 samples in a batch. The further information is shown in Table 5. It shows the parameter size in millions, the MACs in gigabits per second, the CPU processing time (CPU-PT) in seconds, the forward and backward memory footprint (MFP), and the training batch time (TBT) for three benchmark speech enhancement models.

We further analyze the time-spectral representation of speech generated using TD-STNet and benchmarks. Figure 6 presents the time-spectral analysis of speech spoken by a female speaker from the WSJO-SI84 dataset. The spectrogram of the produced speech by TD-STNet effectively preserves harmonic structures, demonstrating robustness during speech activity regions. During the speech pauses, TD-STNet attenuates residual noise while retaining faint harmonics at higher

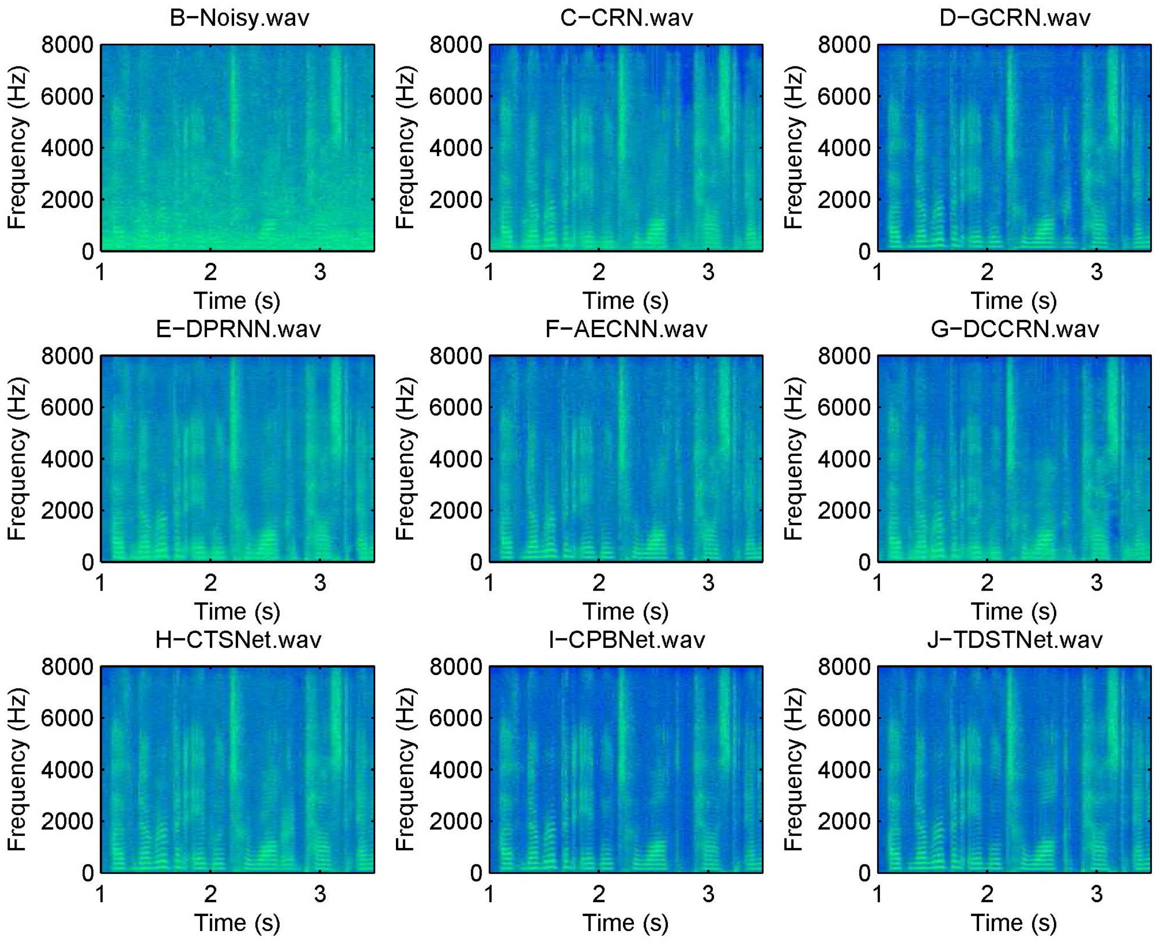

FIGURE 6. Spectrogram visualization at -3dB babble noise. Noisy speech (PESQ=1.54, ESTOI=31.3%), CRN (PESQ=2.43, ESTOI=68.9%), GCRN (PESQ=2.51, ESTOI=69.1%), DCCRN (PESQ=2.42, ESTOI=65%), DPRNN (PESQ=2.52, ESTOI=65.8%), AECNN (PESQ=2.29, ESTOI=64.2%), CTS-Net (PESQ=2.59, ESTOI=68.5%), CPB-Net (PESQ=2.74, ESTOI=72.21%), and TD-STNet (PESQ=2.78, ESTOI=74.32%).

frequencies. This increases speech quality and reduces background distortions, thereby improving speech intelligibility. The spectrograms of the GCRN, DCCRN, and AECNN also exhibit enhanced spectral representations, highlighting the advancements in recent DNN-based speech enhancement models. The spectrograms include the PESQ and ESTOI scores at -3dB babble noise.

D. SPEAKERS AND NOISES: ROBUST GENERALIZATION

The generalization of a speech enhancement model is the ability to improve speech signals in real-world scenes, despite differences in speakers and background noises from the training data. A speech enhancement must manage various speakers and noises, requiring robust generalization for effective performance in diverse environments. We experiment with the generalization of the proposed TD-STNet with unknown speakers and noises, including crowd laughter, exhibition halls, 32-talker babble, and rain thunders (the PSD of background noises are depicted in Fig. 7). Training uses clean data (clean-trainset-56spk from the VCTK database) and background noises are mixed at SNRs: -5dB, 0dB, and +5dB. Testing involves four unknown noises and speakers at -3dB and +3dB SNRs. Table 6 shows that PESQ and STOI metrics evaluate the performance of TD-STNet and three benchmark models (CRN, GCRN, and DCCRN). Real-world noise characteristics vary over time, requiring adaptive tracking of noise power spectral density (PSD). Table 6 demonstrates the performance of TD-STNet over benchmarks in unknown conditions, achieving ΔESTOI improvements of 5.73%, 1.66%, and 5.89% whereas ΔPESQ improvements of 0.28, 0.07, and 0.17 over CRN, GCRN, and DCCRN, respectively. The benchmark GCRN performs better with a ΔPESQ of 0.93 and a ΔESTOI of 34.7% over noisy mixtures, outperforming CRN (ΔPESQ=0.72, ΔESTOI=30.63%), and DCCRN (ΔPESQ = 0.83, ΔESTOI=30.47%). At low SNRs (-3dB), the proposed model generalizes well and obtains better PESQ and ESTOI.

E. ABLATION STUDIES

Ablation studies are performed to understand the contributions of individual modules in the proposed model by sys-

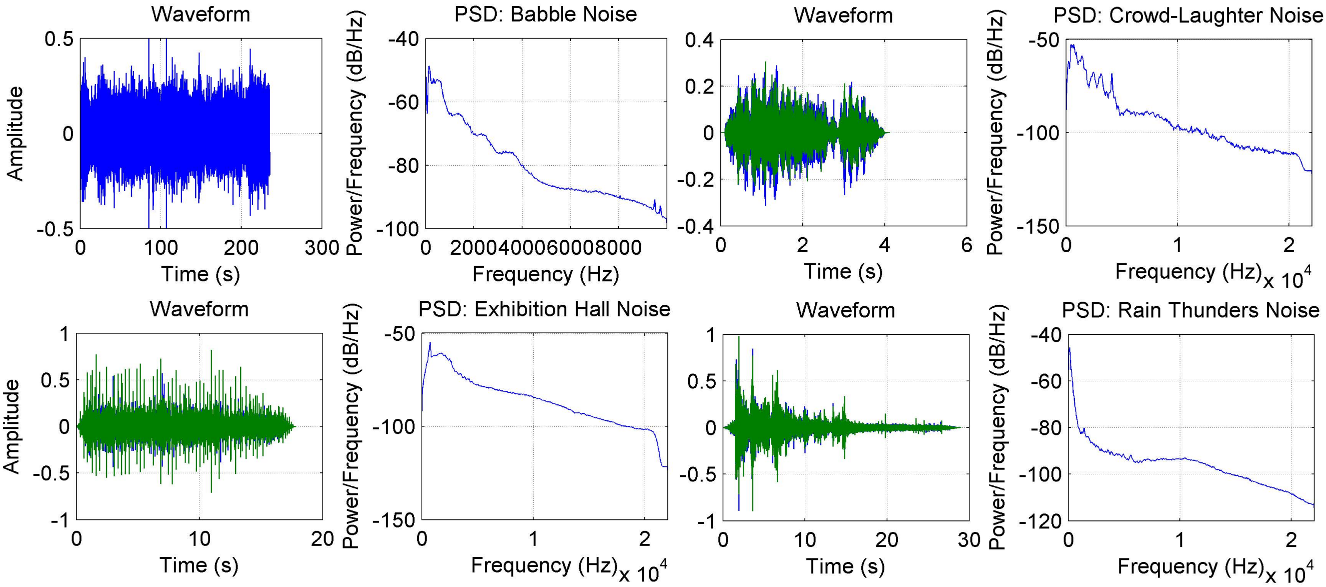

FIGURE 7. PSD and Waveforms of four background noises to examine the generalization ability.

TABLE 7. Ablation studies on the VCTK+DEMAND database to understand the contributions of individual modules. $w/$ and w/o denote *with* and *Without*, respectively.

Performance	Speech Enhancement Performance						Computational Complexity		
Models	STOI	PESQ	SNRSeg	Covl	Csig	Cbak	Para#	RTF	MACs
Mixture	91.6	1.97	1.69	2.63	3.34	2.44	–	–	–
TD-STNet $w/$ TN	94.7	2.98	10.88	3.76	4.18	3.45	3.16M	0.25	9.77 (G/s)
TD-STNet w/o STN	93.1	2.78	9.32	3.38	3.87	3.11	1.11M	0.07	4.01 (G/s)
TD-STNet $w/$ STN+FAB	94.7	3.02	10.94	3.78	4.23	3.52	2.88M	0.18	8.21 (G/s)
TD-STNet $w/$ STN+TAB	94.1	3.06	11.01	3.81	4.27	3.58	2.88M	0.18	8.21 (G/s)
DCCRN [43]	93.8	2.54	8.62	3.27	3.88	3.18	3.70M	2.19	14.36 (G/s)
GaGNet [46]	94.7	2.94	9.24	3.59	4.26	3.45	5.95M	0.05	1.65 (G/s)
TD-STNet (Proposed)	95.0	3.13	11.12	3.82	4.29	3.66	3.14M	0.18	9.64 (G/s)

tematically altering them and observing the impact on overall speech enhancement performance. This process helped validate design options and confirm the importance of specific modules. For ablation studies, we have evaluated the configurations: (i) TD-STNet representing the complete proposed model, (ii) TD-STNet $w/$ TN representing the SE model replacing the STN with traditional transformer, (iii) TD-STNet w/o representing the model without STN bottleneck, (iv) TD-STNet $w/$ STN+FAB representing the proposed model with STN bottleneck using the frequency attention branch (FAB), and (v) TD-STNet $w/$ STN+TAB representing the proposed model with STN bottleneck using the frequency attention branch (FAB). Additionally, we provide three benchmark models for comparison. Table 7 shows the results of ablation studies to examine the contributions of individual modules. Using the STN bottleneck enhances speech performance more effectively than the traditional transformer model, demonstrating the success of the spiking transformer. However, removing the bottleneck reduces computational complexity but significantly degrades speech enhancement performance. Incorporating both time and frequency attention modules improves speech enhancement while maintaining acceptable computational complexity (3.14M para#, 0.18 RTF, and 9.64G/s MACs).

VI. CONCLUSION

This paper presents a model for speech enhancement by integrating self-attention with spiking neural networks. It uses a convolutional encoder-decoder architecture with a spiking transformer bottleneck network (STN). In the proposed SE model, the spiking self-attention in STN uses spike-based queries, keys, and values to capture temporal dependencies and contextual relationships. The spiking bottleneck network has two branches to capture global dependencies across temporal and spectral dimensions. The encoder-decoder includes a multi-scale feature extractor for hierarchical representation, improving the model's ability to process noisy speech. In conclusion, our approach to speech enhancement employs a temporal dynamic spiking transformer as a bottleneck network in a convolutional codec. By integrating the temporal dynamics of speech with the advanced processing capabilities of spiking neural networks and transformers, our model significantly improves speech quality and intelligibility (achieves average values of ΔSI-

SDR=11.52dB, ΔESTOI=39.67%, and ΔPESQ =1.22, outperforming benchmark by 3.83dB in ΔSI-SDR, 7.28% in ΔESTOI, and 0.35 in ΔPESQ). It effectively captures and processes temporal dependencies, resulting in more contextually relevant enhancements. Our dual-branch spiking transformer bottleneck effectively captures global dependencies across temporal and spectral dimensions (represented by Table 7 where results support this conclusion). This comprehensive approach enhances the model's ability to understand and process complex patterns in speech, leading to superior speech enhancement performance. we thoroughly examine the computational load of the proposed speech enhancement model in unknown background noises and speakers (as given in Table 6 where model generalization to unknown conditions is effective). Comparing model complexity, inference time, and memory footprint, our evaluation highlights the performance improvements and resource requirements, demonstrating a balance between enhanced speech performance and resource management. The proposed model shows better real-time performance in terms of many metrics (featuring approximately 3.14 million parameters, 9.64(G/s) MACs, and 0.18 RTF, outperforming many benchmark models, confirmed by Table 4). Integrating time and frequency attention modules improves SE while maintaining adequate computational complexity.

Effective extracting and processing of temporal features are essential for time series data such as highly non-stationary speech and noise signals. The spiking transformer in the proposed model creates an attention matrix at each time step. However, this attention process only relates to the current input, resulting in the underutilized information from different time steps. The future study can include a robust module within the query structure, which aims to better utilize historical information, thereby improving the temporal features.

REFERENCES

[1] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, “Automatic speech recognition (asr) systems for children: A systematic literature review,” Applied Sciences, vol. 12, no. 9, p. 4419, 2022.

[2] A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, “Arabic speech recognition: Advancement and challenges,” IEEE Access, 2024.

[3] M. Gupta, R. Singh, and S. Singh, “Analysis of optimized spectral subtraction method for single channel speech enhancement,” Wireless Personal Communications, vol. 128, no. 3, pp. 2203–2215, 2023.

[4] H. Nguyen, T. V. Ho, M. Akagi, and M. Unoki, “Phase-aware speech enhancement with complex wiener filter,” IEEE Access, 2023.

[5] S. Shi, K. Paliwal, and A. Busch, “On dct-based mmse estimation of short time spectral amplitude for single-channel speech enhancement,” Applied Acoustics, vol. 202, p. 109134, 2023.

[6] B. J. Borgström and M. S. Brandstein, “A multiscale autoencoder (msae) framework for end-to-end neural network speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[7] Y. Du, X. Liu, and Y. Chua, “Spiking structured state space model for monaural speech enhancement,” in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 766–770.

[8] G. W. Lee, H. K. Kim, and D.-J. Kong, “Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition,” IEEE Access, 2024.

[9] F. E. Wahab, Z. Ye, N. Saleem, and R. Ullah, “Compact deep neural networks for real-time speech enhancement on resource-limited devices,” Speech Communication, vol. 156, p. 103008, 2024.

[10] J. Wang, N. Saleem, and T. S. Gunawan, “Towards efficient recurrent architectures: A deep lstm neural network applied to speech enhancement and recognition,” Cognitive Computation, pp. 1–16, 2024.

[11] K. Mraihi and M. A. Ben Messaoud, “Deep learning-based empirical and sub-space decomposition for speech enhancement,” Circuits, Systems, and Signal Processing, pp. 1–31, 2024.

[12] H. J. Park, W. Shin, J. S. Kim, and S. W. Han, “Leveraging non-causal knowledge via cross-network knowledge distillation for real-time speech enhancement,” IEEE Signal Processing Letters, 2024.

[13] N. Saleem, T. S. Gunawan, S. Dhahbi, and S. Bourouis, “Time domain speech enhancement with cnn and time-attention transformer,” Digital Signal Processing, p. 104408, 2024.

[14] J. Ali, N. Saleem, S. Bourouis, E. Alabdulkreem, H. El Mannai, and S. Dhahbi, “Spatio-temporal features representation using recurrent capsules for monaural speech enhancement,” IEEE Access, 2024.

[15] R. R. Rai and M. Mathivanan, “Recalling-enhanced recurrent neural network optimized with chimp optimization algorithm based speech enhancement for hearing aids,” Intelligent Decision Technologies, vol. 18, no. 1, pp. 123–134, 2024.

[16] Z. Li, A. Basit, A. Daraz, and A. Jan, “Deep causal speech enhancement and recognition using efficient long-short term memory recurrent neural network,” Plos one, vol. 19, no. 1, p. e0291240, 2024.

[17] S. Abdulatif, R. Cao, and B. Yang, “Cmgan: Conformer-based metric-gan for monaural speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[18] M. Hadwan, H. A. Alsayadi, and S. AL-Hagree, “An end-to-end transformer-based automatic speech recognition for qur’an reciters.” Computers, Materials & Continua, vol. 74, no. 2, 2023.

[19] K. Yamazaki, V.-K. Vo-Ho, D. Bulsara, and N. Le, “Spiking neural networks and their applications: A review,” Brain Sciences, vol. 12, no. 7, p. 863, 2022.

[20] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, “Spiking neural networks and online learning: An overview and perspectives,” Neural Networks, vol. 121, pp. 88–100, 2020.

[21] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, “Deep learning in spiking neural networks,” Neural networks, vol. 111, pp. 47–63, 2019.

[22] A. Riahi and É. Plourde, “Single channel speech enhancement using u-net spiking neural networks,” in 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, 2023, pp. 111–116.

[23] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, “Deep spiking neural networks for large vocabulary automatic speech recognition,” Frontiers in neuroscience, vol. 14, p. 513257, 2020.

[24] D. Auge, J. Hille, F. Kreutz, E. Mueller, and A. Knoll, “End-to-end spiking neural network for speech recognition using resonating input neurons,” in International Conference on Artificial Neural Networks. Springer, 2021, pp. 245–256.

[25] D. Auge, J. Hille, E. Mueller, and A. Knoll, “A survey of encoding techniques for signal processing in spiking neural networks,” Neural Processing Letters, vol. 53, no. 6, pp. 4693–4710, 2021.

[26] Y. Xing, W. Ke, G. Di Caterina, and J. Soraghan, “Noise reduction using neural lateral inhibition for speech enhancement,” International Journal of Machine Learning and Computing, 2019.

[27] J. Timcheck, S. B. Shrestha, D. B. D. Rubin, A. Kupryjanow, G. Orchard, L. Pindor, T. Shea, and M. Davies, “The intel neuromorphic dns challenge,” Neuromorphic Computing and Engineering, vol. 3, no. 3, p. 034005, 2023.

[28] J. Wall, C. Glackin, N. Cannings, G. Chollet, and N. Dugan, “Recurrent lateral inhibitory spiking networks for speech enhancement,” in 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 2016, pp. 1023–1028.

[29] B. Schrauwen and J. Van Campenhout, “Bsa, a fast and accurate spike train encoding scheme,” in Proceedings of the International Joint Conference on Neural Networks, 2003., vol. 4. IEEE, 2003, pp. 2825–2830.

[30] W. Wei, Y. Hu, H. Huang, and L. He, “Iifc-net: A monaural speech enhancement network with high-order information interaction and feature calibration,” IEEE Signal Processing Letters, 2023.

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

[32] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “Repvgg: Making vgg-style convnets great again,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13 733–13 742.

[33] L. Wang, W. Wei, Y. Chen, and Y. Hu, “D 2 net: A denoising and dereverberation network based on two-branch encoder and dual-path transformer,” in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2022, pp. 1649–1654.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[35] G. Yu, A. Li, H. Wang, Y. Wang, Y. Ke, and C. Zheng, “Dbt-net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2629–2644, 2022.

[36] A. Pandey and D. Wang, “Dense cnn with self-attention for time-domain speech enhancement,” IEEE/ACM transactions on audio, speech, and language processing, vol. 29, pp. 1270–1279, 2021.

[37] 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,” Journal of the Audio Engineering Society, vol. 50, no. 10, pp. 755–764, 2002.

[38] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, “A non-intrusive short-time objective intelligibility measure,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5085–5089.

[39] J. G. Beerends, N. M. Neumann, E. L. van den Broek, A. L. Casanovas, J. T. Menendez, C. Schmidmer, and J. Berger, “Subjective and objective assessment of full bandwidth speech quality,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 440–449, 2019.

[40] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement.” in Interspeech, vol. 2018, 2018, pp. 3229–3233.

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

[42] K. Tan and D. Wang, “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 380–390, 2019.

[43] 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,” arXiv preprint arXiv:2008.00264, 2020.

[44] A. Pandey and D. Wang, “A new framework for cnn-based speech enhancement in the time domain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1179–1188, 2019.

[45] 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 Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1829–1843, 2021.

[46] A. Li, C. Zheng, L. Zhang, and X. Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” Applied Acoustics, vol. 187, p. 108499, 2022.

[47] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “Phasen: A phase-and-harmonics-aware speech enhancement network,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 9458–9465.

[48] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K. K. Paliwal, and F. Shang, “Deep residual-dense lattice network for speech enhancement,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 8552–8559.

[49] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” arXiv preprint arXiv:2006.12847, 2020.

[50] K. Wang, B. He, and W.-P. Zhu, “Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain,” in ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2021, pp. 7098–7102.

[51] J. Lin, A. J. d. L. van Wijngaarden, K.-C. Wang, and M. C. Smith, “Speech enhancement using multi-stage self-attentive temporal convolutional networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3440–3450, 2021.

[52] H. Yue, W. Duo, X. Peng, and J. Yang, “Reference-based speech enhancement via feature alignment and fusion network,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11 648–11 656.

[53] R. Yu, Z. Zhao, and Z. Ye, “Pfrnet: Dual-branch progressive fusion rectification network for monaural speech enhancement,” IEEE Signal Processing Letters, vol. 29, pp. 2358–2362, 2022.

[54] Y. Li, Y. Sun, W. Wang, and S. M. Naqvi, “U-shaped transformer with frequency-band aware attention for speech enhancement,” IEEE/ACM transactions on audio, speech, and language processing, 2023.

[55] L. Sun, S. Yuan, A. Gong, L. Ye, and E. S. Chng, “Dual-branch modeling based on state-space model for speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[56] Y. Li, M. Sun, X. Zhang et al., “Scale-aware dual-branch complex convolutional recurrent network for monaural speech enhancement,” Computer Speech & Language, vol. 86, p. 101618, 2024.

[57] X. Xiang, X. Zhang, and H. Chen, “A nested u-net with self-attention and dense connectivity for monaural speech enhancement,” IEEE Signal Processing Letters, vol. 29, pp. 105–109, 2021.

[58] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, “Towards efficient models for real-time deep noise suppression,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 656–660.

[59] F. Dang, H. Chen, and P. Zhang, “Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 6857–6861.

[60] D. Zhang, A. Dong, J. Yu, Y. Cao, C. Zhang, and Y. Zhou, “Speech enhancement generative adversarial network architecture with gated linear units and dual-path transformers,” in 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2022, pp. 2563–2568.

[61] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, “Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 9281–9285.

[62] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, “Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7857–7861.

[63] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 7407–7411.

[64] Z. Zheng, Y. Liu, J. Liu, K. Niu, and Z. He, “Dual-path transformer based on efficient channel attention mechanism for speech enhancement,” in 2023 International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2023, pp. 7–12.

MANAL ABDULLAH ALOHALI received the Ph.D. degree in computer science from the University of Plymouth, U.K. She is an Associate Professor with the Information Systems Department, CCIS, Princess Nourah bint Abdulrahman University (PNU), Saudi Arabia, where she is the Dean of CCIS. Her research interests reside in the areas of information systems, machine learning, and cyber security. She received the PNU’s Research Excellence Award.

NASIR SALEEM received a B.S. in Telecommunication Engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2008. He earned his M.S. in Electrical Engineering from CECOS University, Peshawar, Pakistan, in 2012, and completed his Ph.D. in Electrical Engineering with a specialization in digital speech processing and deep learning from the University of Engineering and Technology, Peshawar, Pakistan, in 2021. Following his PhD, he was a postdoctoral fellow at the Islamic International University Malaysia (IIUM), where he researched modern artificial intelligence-based speech processing algorithms. From 2008 to 2012, he served as a lecturer at the Institute of Engineering Technology (IET), Gomal University, engaging in both teaching and research. Currently, he is an assistant professor in the Department of Electrical Engineering within the Faculty of Engineering and Technology (FET) at Gomal University. He also holds the position of Deputy Director of the Quality Assurance Directorate at the same institution. His research interests include human-machine interaction, speech enhancement, speech recognition, speech and video processing, and machine learning applications. He has published several research papers in renowned journals and conferences, including those by Elsevier, Springer, and IEEE. In addition to his research, he actively participates in academic activities such as guest editing and paper reviewing.

HELA ELMANNAI received the Ph.D. degree in information technology from SUPCOM, Tunisia. She is currently an Associate Professor with the Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. Her research interests include artificial intelligence, networking, blockchain, and engineering applications.

DELEL RHOUMA

MOHAMMAD MEDANI received the B.S. degree in computer science from the Faculty of Applied Sciences and Computer Science, Omdurman Ahlia University, Sudan, in 1999, the M.S. degree in computer science and Information from the Faculty of Engineering and Technology, University of Gezira, Sudan, in 2003, and the Ph.D. degree in computer science from the Faculty of Graduate Studies and Scientific Research, National Ribat University, Sudan, in 2014. From 2007 to 2010, he was the Head of the Computer Science Department, Faculty of Applied Sciences and Computer Science. He was the External Examiner assigned to Alzaiem Alazhari University, from 2014 to 2016. He was an Assistant Professor at the Faculty of Computer Science, The National Ribat University, Sudan, in 2014. He has been an Assistant Professor at the Computer Science Department, King Khalid University, Saudi Arabia, since 2015. His research interests include compiler design and operating systems.

SAMI BOUROUIS received the Engineer, M.Sc., and Ph.D. degrees in computer science from the University of Tunis, Tunisia, in 2003, 2005, and 2011, respectively. He is currently a Professor at the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include data mining, image processing, statistical machine learning, cybersecurity, and pattern recognition applied to several real-life applications.