Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition

Sami Dhahbi[1], Nasir Saleem[2]*, Teddy Surya Gunawan[3], Sami Bourouis[4], Imad Ali[5], Aymen Trigui[6], Abeer D. Algarni[7]

[1] Department of Computer science, College of science and art at Mahayil, King Khalid University, Muhayil Aseer, 62529 (Saudi Arabia)
[2] Department of Electrical Engineering, FET, Gomal University, D.I. Khan-29050, KPK (Pakistan)
[3] Electrical and Computer Engineering Department, Islamic International University Malaysia, Kuala Lumpur (Malaysia)
[4] Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944 (Saudi Arabia)
[5] Department of Computer Science, University of Swat, Swat (Pakistan)
[6] Department of Computer Science, College of Computer Science, King Khalid University, Abha (Saudi Arabia)
[7] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671 (Saudi Arabia)

* Corresponding author. nasirsaleem@gu.edu.pk

Received 15 November 2023 | Accepted 30 January 2024 | Early Access 30 April 2024

Abstract

Traditional recurrent neural networks (RNNs) encounter difficulty in capturing long-term temporal dependencies. However, lightweight recurrent models for speech enhancement are important to improve noisy speech, while being computationally efficient and able to capture long-term temporal dependencies efficiently. This study proposes a lightweight hourglass-shaped model for speech enhancement (SE) and automatic speech recognition (ASR). Simple recurrent units (SRU) with skip connections are implemented where attention gates are added to the skip connections, highlighting the important features and spectral regions. The model operates without relying on future information that is well-suited for real-time processing. Combined acoustic features and two training objectives are estimated. Experimental evaluations using the short time speech intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and word error rates (WERs) indicate better intelligibility, perceptual quality, and word recognition rates. The composite measures further confirm the performance of residual noise and speech distortion. With the TIMIT database, the proposed model improves the STOI and PESQ by 16.21% and 0.69 (31.1%) whereas with the LibriSpeech database, the model improves STOI by 16.41% and PESQ by 0.71 (32.9%) over the noisy speech. Further, our model outperforms other deep neural networks (DNNs) in seen and unseen conditions. The ASR performance is measured using the Kaldi toolkit and achieves 15.13% WERs in noisy backgrounds.

Keywords

Real-Time Speech, Simple Recurrent Unit (SRU), Speech Enhancement, Speech Processing, Speech Quality.

DOI: 10.9781/ijimai.2024.04.003

I. Introduction

Building lightweight recurrent models for speech enhancement and Automatic Speech Recognition (ASR) involves designing models that can process audio data efficiently while improving speech quality and intelligibility, mostly in noisy or degraded environments. The speech enhancement (SE) is particularly significant as it reduces listener fatigue, especially when individuals are subjected to prolonged exposure to high noise. The SE positively impacts the efficiency of communication and multimedia systems. Furthermore, it improves the intelligibility of speech, thereby enhancing ASRs and interactions between humans and machines. Various proposals are available in the literature, encompassing methods such as spectral subtraction [1]-[2], Wiener filtering [3], and minimum mean square error (MMSE) [4]-[5].

To address the SE challenges, supervised learning models are considered. These models undergo training using large speech datasets [6]-[7]. Among successful models for SE are the regression-based deep neural networks (DNNs) [8]-[11]. Given that the relationship between input and target features is nonlinear, a multi-layer DNN incorporating nonlinear activation is a suitable option. Essential considerations for a DNN-based SE include the type of network, the learning objective, and

Please cite this article as: S. Dhahbi, N. Saleem, T. S. Gunawan, S. Bourouis, I. Ali, A. Trigui, A. D. Algarni, “Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition”, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 6, pp. 74-85, 2024, http://dx.doi.org/10.9781/ijimai.2024.04.003

the loss function [12]-[14]. For SE, the learning models are categorized into spectral mapping and masking. Mapping models entail training networks through direct mapping rules. They learn to estimate clean spectral features from noisy spectral features. However, these methods often result in excessively smooth spectra [9]. Conversely, masking-based learning algorithms have shown greater success in SE. They involve multiplying the estimated parameters of objectives (ideal ratio mask (IRM) or ideal binary mask (IBM)) with noisy magnitudes. Many deep-learning approaches have recently emerged time-frequency (T-F) masks as training objectives, yielding favorable results [15]-[21]. Fully connected feedforward DNNs (FDNNs) predict labels for individual time frames using small context windows. Yet, they lack control over the long-term context windows crucial for accurately tracking the target speaker. DNN-based SE algorithms employ multi-layer DNNs for learning nonlinear regression functions or estimate a spectral mask using noisy magnitudes. These models forego the requisite for statistical distributions assumption, yielding superior noise reduction when handling non-stationary noises. The SE system using recurrent neural networks (RNNs) with T-F masking is depicted in Fig. 1.

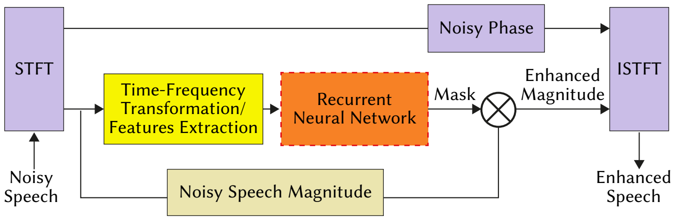


Fig. 1. SE with RNNs estimating T-F masking.

Recurrent neural network (RNN) has attainment significance in several challenging applications within Natural Language Processing, including neural machine translation [22], conversational/dialogue modeling [23], and ASR modeling [24]. Given that speech waveform is a sequential data type, it requires a temporal context for effective processing, and RNNs excel in capturing long-range temporal sequences. Previous research [10] [25] has recommended framing SE as a sequence-to-sequence procedure to manage long-term contextual window. Various models, including RNN, CNN [57], and other deep learning architectures, have been proposed and assessed on diverse noises and speakers. In a work [25], LSTM is introduced for speaker's generalization. The results indicate that the LSTM model demonstrates superior generalization to untrained speakers, significantly outperforming a DNN-based model in terms of speech intelligibility. Multiple studies have emphasized that employing the sequence-to-sequence approach enables LSTM to effectively control long-term context windows, leading to successful outcomes in speech enhancement [26]. To model the long input sequential data, RNNs face problems in capturing long-term temporal dependencies. Further, training an RNN with back propagation through time is exposed to vanish and explode the gradients. These challenges are addressed by proposing RNN variants using novel transition functional units and optimization techniques, such as LSTM [11] [27] and gated recurrent unit (GRU) [28]-[29]. Several approaches focused on connection architectures, including stacked RNNs [30] and skip RNNs [31]. In this paper, we have proposed efficient simple recurrent unit (SRU) models that are able to detain the long-term temporal dependencies and prevent the gradient from decaying. The contributions are highlighted below.

- A proposed SRU model takes on an hourglass shape, effectively capturing long-term temporal and sequential data. This results in reduced feature resolutions without sacrificing data in the layers.
- Skip connections are introduced between nonadjacent layers to mitigate decaying gradient. Additionally, attention gates within the skip connections are used to reduce irrelevant features and highlight crucial features across different spectral regions.
- Robust training of the proposed SRU-based model is achieved by extracting combined feature sets from the noisy speech.
- We estimate two distinct training objectives, Ideal Ratio Mask and Ideal Binary Mask, to attenuate noise in the noisy mixture. This approach aims to enhance speech quality, intelligibility, and reduce word error rates.

The rest of this study is structured as follows: Section II outlines the formulation of the SE problem. Section III introduces the proposed SE. Details of the experiments conducted are outlined in Section IV. Section V provides the results and discussions. Ultimately, Section VI presents the drawn conclusions.

II. Problem Formulation

Take into account that a clear speech signal $x(t)$ undergoes degradation due to presence of background noise $n(t)$. This leads to the generation of a noisy speech signal $s(t)$, which can be represented as:

$$s(t) = x(t) + n(t) \tag{1}$$

The noisy speech signal, denoted as $s(t)$, undergoes a transformation to the frequency domain through application of the short-time Fourier Transform (STFT). This results in the acquisition of the frequency domain depiction:

$$|S(f,t)| = |X(f,t)| + |N(f,t)| \tag{2}$$

Where t and f denote the frame and frequency indexes, respectively. A combined feature set is extracted to robustly train the proposed SRU model. During inference, the trained parameters estimate Ideal Ratio Mask and the Ideal Binary Mask as training objectives. The estimated magnitude masks are then multiplied by noisy magnitudes to suppress the background noises:

$$|\hat{X}(f,t)| = |M_x(f,t)| \otimes |S(f,t)| \tag{3}$$

Where $|M(t, f)|$ is the estimated T-F mask. The estimated magnitude and noisy phase reconstruct noise-free enhanced speech waveforms. The block diagram of the proposed SE is depicted in Fig. 2.

III. Proposed Speech Enhancement

SRUs can detain the information in speech waveforms which is a kind of long-term temporal sequence. The proposed network architecture effectively addresses the limitations of traditional RNNs using the following approaches. Our aim is to reduce the complexity (which is directly linked with neurons quantity and number of steps) without degrading the speech enhancement performance. Since equal number of neurons is each layer will introduce computational load, we have arranged neurons in increasing-decreasing order which forms a U-Shape layer. With this arrangement, the overall complexity of the model is reduced (with reduced neurons). Further, the same mechanism is adopted for time steps. By increasing the number of time steps, the computational complexity can indeed increase. This is because the model needs to process information across multiple time steps, leading to a higher demand for computational resources. More time steps may require more complex models to capture long-term dependencies in the data. Firstly, the network architecture has a unique shape with bottom and upper pyramids. For the upper pyramid, there is a decrease in time steps while the number of neurons increases. Conversely, the lower pyramid exhibits an increase in time steps paired with a decrease in number of neurons across layers. This architectural design enables the model to manage high-resolution

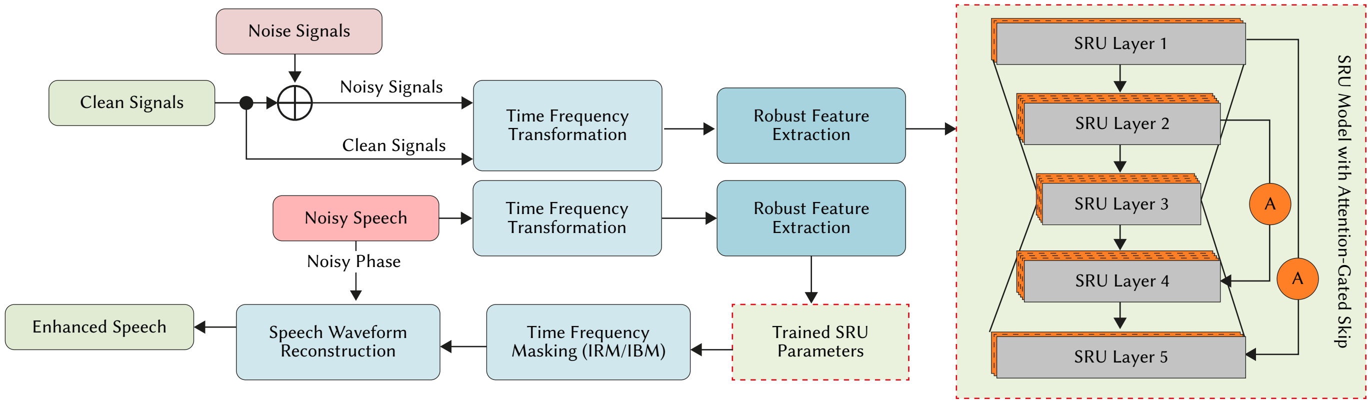

Noise Signals
Clean Signals
Noisy Signals
Clean Signals
Time Frequency Transformation
Robust Feature Extraction
Noisy Speech
Time Frequency Transformation
Robust Feature Extraction
Noisy Phase
Enhanced Speech
Speech Waveform Reconstruction
Time Frequency Masking (IRM/IBM)
Trained SRU Parameters
SRU Layer 1
SRU Layer 2
SRU Layer 3
SRU Layer 4
SRU Layer 5
A
A
SRU Model with Attention-Gated Skip

Fig. 2. The block diagram of the proposed Speech Enhancement System.

features without encountering memory overflow issues. Additionally, skip connections have been incorporated between layers of similar shapes, spanning from upper to lower pyramid. This inclusion enhances the mitigation of gradient decay throughout the layers. To further refine the skips, an attention gate is introduced, emphasizing crucial spectral regions. Given that the speech spectrum showcases dominant formants in low-frequency areas and a sparse distribution in high-frequency zones, it becomes imperative to employ attention weights for distinguishing these varied spectral regions through the attention process. The model consists of five SRU layers featuring two attention-gated skip connections, as illustrated in Fig. 3. Details regarding time steps and units can be found in Table I. The network adeptly learns the nonlinear relationship, effecting the transformation of noisy speech, denoted as *s*(t), into a clear and intelligible speech signal, denoted as *x*(t).

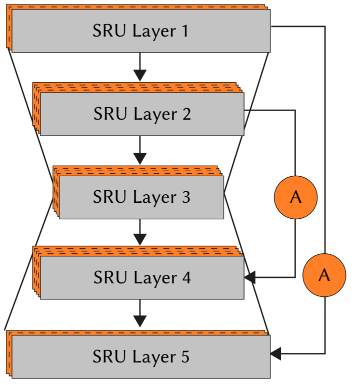


Fig. 3. The proposed SRU architecture.

TABLE I. Proposed SRU Details

Layer	Neurons in Layer	Time-Steps in Layers
1	256	512
2	512	256
3	1024	128
4	512	256
5	256	512

The SRU is a type of recurrent neural network (RNN) that incorporates parallelism analogous to convolutional and feedforward networks. It achieves an exact stability between sequential dependency and independency. While the state computation in SRU is time-dependent, each dimension within the state operates independently. Additionally, SRU improves the training of deep recurrent models through the integration of highway connections and a specialized parameter initialization technique tailored for effective gradient propagation in deep architectures. The computational process of SRU includes the following steps.

$$f_t = \sigma(W_f x_t + b_f) \tag{4}$$

$$r_t = \sigma(W_r x_r + b_r) \tag{5}$$

$$C_t = f_t \otimes C_{t-1} + (1 - f_t) \otimes W x_t \tag{6}$$

$$h_t = r_t \otimes g C_t + (1 - r_t) \otimes x_t \tag{7}$$

Where **W** and **b** are learnable weight matrices and bias terms. In the SRU architecture, the computation of forget and reset gates is independent, eliminating interdependence and simplifying the gating mechanism for faster training. Moreover, the candidate hidden state is determined through element-wise multiplication of the reset gate with the previous hidden state, enhancing SRU's ability to capture long-term dependencies more effectively compared to the traditional RNN. This streamlined SRU design enhances network capacity by sharing hidden states among similar and lower layers. The SRU architecture incorporates a strategy of decreasing time-steps and increasing neurons from the first layer to the mid-layer, and conversely, increasing time-steps and decreasing neurons from the mid-layer to the last layer. This approach facilitates a deeper representation. The shared hidden states among layers in SRU mean that the hidden states in layer '*l*' at time '*t*' are derived by combining hidden states from the (*l*-1) lower layer at time (*t*-1). Before the skips, the hidden states of upper and lower layers are combined, resulting in a final output similar to the input vectors:

$$h_t^l = SRU(h_t^{l-1}, h_{t-1}^l) \tag{8}$$

The output vector is generated through the combination of the hidden states across all layers, as follows:

$$Z = SRU(h_l^5, h_T^5) \tag{9}$$

Here, Z represents the output vector from the final layer in the SRU. To prevent gradient decay across the layers, two skips are introduced. These skips promote enhanced generalization by integrating low-level and high-level features. Given that speech spectra include various frequency components with formants dominating in the low-frequency regions and displaying sparse distributions in higher-frequency areas, it becomes crucial to employ an attention process that assigns attention weights to discern different frequency regions. This attention process focuses on crucial regions and features, thereby enhancing the output quality. Initially, the alignment vector is calculated for the output yout of the layer as follows:

$$V_{align} = tanh(y_{out} \otimes W) \tag{10}$$

Where **W** indicates trainable weights. The score λ for corresponding alignment vector is given as:

$$\lambda_{score} = \alpha(V_{align} \otimes \beta) \tag{11}$$

The dynamic range for λ is 0 to 1. To avoid weak scores, a controlling parameter β is incorporated. The parameter λ assigns different scores to different features in the feature space. The output of the attention process is given as:

$$\hat{y}out = \gamma_{score} \odot yout \tag{12}$$

Where $\odot$ denotes Hadamard product used to weigh all feature streams by using the obtained Scores. The feature-level computation process of attention weights is depicted in Fig. 4. The features are derived from input frames of the speech. The frame shift and duration remain at a consistent 10 milliseconds and 20 milliseconds. The features consist of 31-dimensional Mel-Frequency Cepstral Coefficients, 13-dimensional Relative Spectral Transformed Perceptual Linear Prediction Coefficient, 64-dimensional Gammatone Filter-bank, and 15-dimensional Amplitude Modulation Spectrogram, outlined as follows:

$$f_s = f_s^{MFCC} + f_s^{RASTA-PLP} + f_s^{AMS} + f_s^{GFE} \tag{13}$$

$$f_x = f_x^{MFCC} + f_x^{RASTA-PLP} + f_x^{AMS} + f_x^{GFE} \tag{14}$$

Where f_x and f_s show feature sets of clean x(t) and noisy speech s(t), respectively. GFE features are derived from a Cochleagram [32]. The delta features (Δfx and Δfs) are affixed to the features.

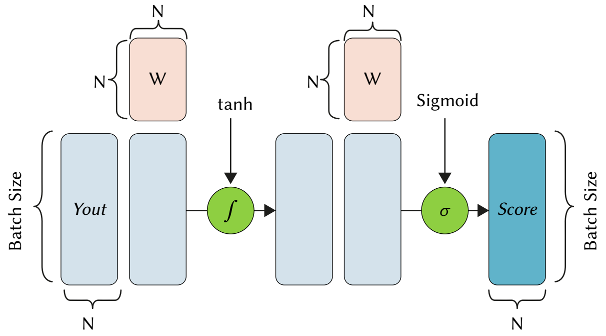


Fig. 4. Attention procedure over Features.

IV. Experiments

A. Dataset

In order to evaluate the proposed SE, a set of experiments is conducted exploiting speech utterances from the TIMIT [33] and LibriSpeech [34] databases. The TIMIT database comprises phonetically balanced speech waveforms sampled at a rate of 16 kHz, while LibriSpeech encompasses 1000 hours of speech waveforms sampled at the rate of 16 kHz. The experiments exclusively utilize clean speech utterances obtained from both databases. For evaluating the effectiveness of the proposed SE under varying noisy conditions, various background noises are selected from Noisex-92 and Aurora-4 databases [35]-[36]. To mix noises with clean utterances, four signal-to-noise ratios (SNRs) are utilized, ranging from -8dB to 4dB in increments of 4dB. To train the SRU network proposed in this study, sentences are selected from the TIMIT database. These utterances are used to generate an ideal ratio mask (IRM) and an ideal binary mask (IBM) for each SNR. To enhance the generalization of the model across different speakers, the training sentences belong to male/female speakers and degraded by all noises. This amounted to 10500 training sentences from TIMIT database. An additional 1500 utterances were randomly chosen for testing. All noise sources, except for two (factory2 and cafeteria), are used in both the training and testing. These two sources are reserved as unseen noises. Furthermore, 2000 sentences are extracted from the LibriSpeech dataset and used to estimate IRM and IBM across all SNRs (8dB to 4dB). This creates a total of 10500 training utterances from the LibriSpeech. In this case, five noise sources are introduced in experiments for training the models with LibriSpeech. These noise sources are airport, babble, street, cafeteria, and car noise, respectively.

B. Network Setting

This study uses five-layered SRU network to enhance speech degraded by noise. The input layer is provided with a 1408-d context window containing 11 frames. Each SRU layer comprises M neurons and N time steps, while the output layer encompasses 257 neurons. During the training process, backpropagation through time (BPTT) [37] is utilized. For optimization, the adaptive gradient descent [38] method with a momentum parameter m is employed, where a scaling factor of 0.0010 is set for AGD. The learning rate follows a linear reduction from 0.06 to 0.002 over the course of processing. Samples of 512 batch size are chosen for training. A total of 80 epochs are completed, during which m remains constant at 0.4 for initial epochs, and subsequently, it is raised to 0.8 for other epochs. Dropout regularization [39] with a 0.02 rate is used to mitigate overfitting. During mask estimation, the Mean Squared Error (MSE) is used as a loss function. Notably, the SRU model operates without exploiting future information, ensuring causality. To estimate current speech frame, a feature context window of 11 frames (comprising 10 previous and 1 current frame) is employed. This approach involves concatenating 11 frames of features into extended vectors, serving as the network's input for each time step, as depicted in Fig. 5. For further details regarding the deep model's hyperparameters, refer to Table II.

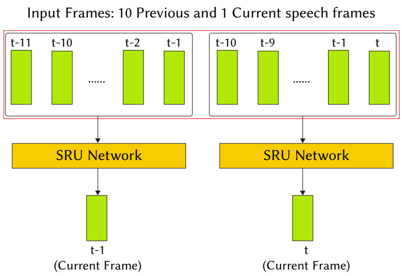


Fig. 5. Causal SRU with feature window of 11 frames.

TABLE II. Proposed SRU Details

Hyper Parameters	Baseline SRU	Baseline DNN	Proposed SRU
Hidden Layers	5	5	5
Layer 1 Neurons	1024	1024	256
Layer 2 Neurons	1024	1024	512
Layer 3 Neurons	1024	1024	1024
Layer 4 Neurons	1024	1024	512
Layer 5 Neurons	1024	1024	256
Learning Rate	0.0001	0.0001	0.0001
No. of Epochs	80	80	80
Momentum Rate	0.8	0.8	0.8
Dropout Rate	0.2	0.2	0.2
Loss Function	MSE	MSE	MSE
Activation	ReLU	ReLU	ReLU

C. Evaluation Metrics

The assessment of our SE involves the use of four objective metrics during the experiments. These metrics encompass the short-time objective intelligibility (STOI), the perceptual evaluation of speech quality (PESQ), and composite measures (CM). These metrics serve purpose of evaluating intelligibility, quality, distortion, and residual noise. The perceptual speech quality, as determined by PESQ [40] following ITU-T P.862 guidelines, is scored within -0.5 to 4.5. STOI [41] quantifies speech intelligibility from 0 to 1 with percentage. Further, the composite measures [42] consist of the C_{SIG} (indicating speech distortions) and the C_{BAK} (reflecting residual noise) [56].

D. Model Representation

To evaluate the proposed SE models, various configurations are considered, each with specific interpretations. The SRU-NoSkip-IRM model estimates IRM using the proposed SRU architecture without skip connections, while the SRU-NoSkip-IBM focuses on estimating IBM without skip connections. In contrast, the SRU-WithSkip-IRM and SRU-WithSkips-IBM models aim to estimate IRM and IBM, respectively, utilizing the proposed SRU with skip connections. Additionally, the SRU-AttSkip-IRM and SRU-AttSkip-IBM models incorporate attention skip connections in their quest to estimate IRM and IBM. The baseline SRU [25], denoted as SRU-IRM and SRU-IBM, employs IRM and IBM as training objectives. All models are trained using the TIMIT and LibriSpeech datasets.

V. Results and Discussions

A. Speech Enhancement in Seen Noises

Table III and Table IV illustrate a comparison of our speech enhancement (SE) algorithms across three distinct noise types, evaluated by STOI. The training objectives involve estimating the Ideal Ratio Mask (IRM) and the Ideal Binary Mask (IBM). The SRU model, incorporating mask estimation, combined feature sets, and attention skips, demonstrated superior performance in comparison to networks lacking skips or utilizing skips without attention. Enhanced intelligibility and quality were observed in the proposed models when applied to noisy speech. For instance, in Table III and Table IV, both SRU-AttSkip-IRM and SRU-AttSkip-IBM exhibited improvements in STOI by 7.7% and 6.9%, respectively, over noisy speech (UNP) at -8dB babble noise. Similarly, at -4dB car noise, these models improved STOI by 23.9% and 23.5%. At 0dB factory noise, the SRU-AttSkip-IRM and SRU-AttSkip-IBM showed STOI improvements of 20.2% and 19.7% over noisy speech. In comparison to the SRU-WithSkip-IRM and SRU-WithSkip-IBM, the proposed models with attention skips achieved a 2.1% and 2.5% improvement in STOI at -8dB babble noise. Additionally, these attention skip models outperformed SRU-NoSkip-IRM and SRU-NoSkip-IRM by 9.1% and 8.5% at -8dB babble noise. Overall, SRU-AttSkip-IRM exhibited notable advantages over SRU-AttSkip-IBM, displaying improved average STOI across noise types and Signal-to-Noise Ratios (SNRs) by 1.23%.

TABLE III. STOI in Seen Noise for IRM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	48.2	58.1	67.1	76.2
	SRU-NoSkips	52.7	66.7	77.0	84.8
	SRU-WithSkips	53.8	68.8	79.1	87.0
	SRU-AttenSkips	55.9	70.1	80.3	88.6
Car Noise	Noisy Mixture	51.8	58.9	68.6	77.1
	SRU-NoSkips	72.4	79.2	84.9	89.4
	SRU-WithSkips	74.5	86.9	86.9	91.6
	SRU-AttenSkips	75.7	88.3	88.3	93.2
Factory Noise	Noisy Mixture	55.2	61.7	69.8	78.0
	SRU-NoSkips	66.3	76.7	84.5	90.5
	SRU-WithSkips	68.3	78.8	86.5	92.6
	SRU-AttenSkips	69.5	79.9	87.9	93.7

TABLE IV. STOI in Seen Noise for IBM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	48.2	58.1	67.1	76.2
	SRU-NoSkips	51.6	66.7	77.0	84.8
	SRU-WithSkips	52.6	64.7	77.5	85.7
	SRU-AttenSkips	55.1	66.2	79.8	88.1
Car Noise	Noisy Mixture	51.8	58.9	68.7	77.1
	SRU-NoSkips	72.4	79.2	85.0	89.4
	SRU-WithSkips	71.5	84.3	84.3	88.8
	SRU-AttenSkips	75.3	87.7	87.7	92.2
Factory Noise	Noisy Mixture	55.2	61.7	69.8	78.0
	SRU-NoSkips	66.3	76.7	84.5	90.5
	SRU-WithSkips	67.2	77.5	84.4	89.5
	SRU-AttenSkips	68.2	78.8	86.8	91.7

TABLE-V. PESQ in Seen Noise for IRM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	1.17	1.52	1.86	2.10
	SRU-NoSkips	1.65	1.88	2.17	2.51
	SRU-WithSkips	1.68	1.90	2.20	2.54
	SRU-AttenSkips	1.79	2.06	2.25	2.65
Car Noise	Noisy Mixture	1.27	1.42	1.62	1.87
	SRU-NoSkips	1.97	2.26	2.57	2.84
	SRU-WithSkips	2.01	2.29	2.60	2.87
	SRU-AttenSkips	2.10	2.34	2.66	2.96
Factory Noise	Noisy Mixture	1.28	1.32	1.52	1.76
	SRU-NoSkips	1.52	1.86	2.16	2.53
	SRU-WithSkips	1.55	1.87	2.17	2.55
	SRU-AttenSkips	1.62	2.01	2.26	2.67

TABLE-VI. PESQ in Seen Noise for IBM Training-Objective

Noise	Model	-8dB	-4dB	0dB	4dB
Babble Noise	Noisy Mixture	1.17	1.52	1.86	2.10
	SRU-NoSkips	1.63	1.85	2.17	2.50
	SRU-WithSkips	1.66	1.87	2.20	2.53
	SRU-AttenSkips	1.69	1.91	2.23	2.56
Car Noise	Noisy Mixture	1.27	1.42	1.62	1.87
	SRU-NoSkips	2.01	2.24	2.57	2.83
	SRU-WithSkips	2.04	2.27	2.61	2.86
	SRU-AttenSkips	2.07	2.30	2.63	2.89
Factory Noise	Noisy Mixture	1.28	1.32	1.52	1.76
	SRU-NoSkips	1.49	1.83	2.14	2.51
	SRU-WithSkips	1.50	1.84	2.15	2.52
	SRU-AttenSkips	1.59	1.92	2.23	2.59

Table V and Table VI assess the performance of the proposed SE models across seen noise types. According to PESQ evaluations, our SRU model, incorporating combined features and attention skips, demonstrated superior performance compared to models lacking skips or utilizing skips without an attention gate. This resulted in enhanced perceptual speech quality relative to counterpart models when applied to both noisy and proposed model-processed speech. For example in Table V and Table VI, SRU-AttSkip-IRM and SRU-AttSkip-IBM improved the PESQ by 0.34 (20.98%) and 0.31 (19.49%) over the noisy

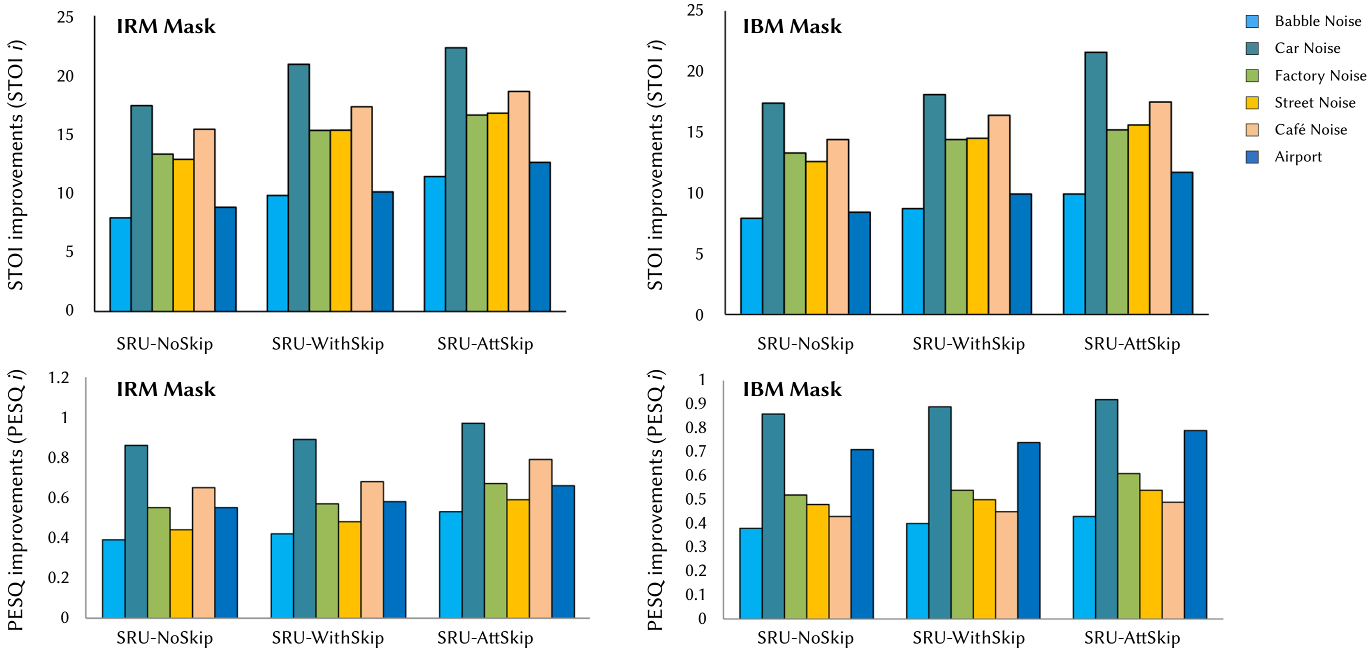


Fig. 6. Average enhancements (STOI*i* and PESQ*i*) across various noises.

TABLE-VII. STOI and PESQ Test Scores in All Example Noise Sources for TIMIT Dataset

Metric	Model	IRM				IBM			
		-8dB	-4dB	0dB	4dB	-8dB	-4dB	0dB	4dB
STOI	SRU-NoSkips	63.8	74.2	82.1	88.2	62.1	71.4	80.0	85.9
	SRU-WithSkips	65.5	78.2	84.2	90.4	63.8	75.5	82.1	88.0
	SRU-AttenSkips	67.3	79.5	85.6	91.8	66.2	77.6	84.8	90.7
PESQ	SRU-NoSkips	1.71	2.00	2.30	2.63	1.71	1.97	2.29	2.61
	SRU-WithSkips	1.75	2.02	2.32	2.65	1.73	1.99	2.32	2.64
	SRU-AttenSkips	1.84	2.14	2.39	2.76	1.78	2.05	2.36	2.68

mixture at -8dB factory noise. SRU-AttSkip-IRM and SRU-AttSkip-IBM improved the PESQ by 0.54 (26.21%) and 0.39 (20.41%) over noisy mixture at -4dB babble noise. Moreover, at 0dB car noise, the SRU-AttSkip-IRM and SRU-AttSkip-IBM enhanced PESQ by 1.04 (39.1%) and 1.01 (38.4%) over noisy mixture. Compare to other models, the proposed SRU-WithSkip-IRM and SRU-WithSkip-IBM with attention skips improved PESQ by 3.04% and 1.04% at 4dB in car noisy background. It shows that at high SNRs all the proposed SRU models perform nearly equally. Furthermore, models with attention skips improved the PESQ by 6.17% over SRU-NoSkip-IRM and 0.06 (2.34%) SRU-NoSkip-IRM at 4dB babble noise. For PESQ, SRU-AttSkip-IRM outscored SRU-AttSkip-IBM by 3.07%. Average STOI and PESQ for both masks can be found in Table VII and Table VIII, respectively. These scores are averaged across different seen noises. The obtained results validate that the proposed SRU-AttSkip achieved noteworthy results.

Average enhancements (STOI*i* and PESQ*i*) across various noises are illustrated in Fig. 6. Additional experimentation involved evaluating our SE models on the LibriSpeech dataset. This dataset, comprising 1000 hours of audiobook-derived utterances at 16 kHz sampling frequency, was chosen for evaluation. For this study, clean utterances were exclusively selected and mixed with noises (car, babble, airport, street, cafeteria). Table VIII shows PESQ and STOI for the LibriSpeech database. SRU-AttSkip-IRM and SRU-AttSkip-IBM configurations exhibited a significant 16.44% and 14.9% average STOI improvement over noisy speech. Correspondingly, these configurations led to an average PESQ improvement of 33.19% (0.78 factor) and 31.14% (0.71 factor) over the unprocessed speech. Cross-corpus comparisons highlighted the superior performance of the proposed models when trained on the LibriSpeech dataset in contrast to the TIMIT dataset.

Table IX presents the outcomes of the testing, focusing on CBAK and CSIG. The results clearly demonstrate that the robust feature sets and attention skips yielded superior performance in terms of both residual noise and distortion. In comparison, SRU-AttSkip-IRM and SRU-AttSkip-IBM effectively mitigated background noises and introduced less distortions when compared with SRU-NoSkip-IRM and SRU-NoSkip-IBM. Average C_{SIG} and C_{BAK} scores showed an enhancement from 2.02 and 1.73 with the noisy speech to 3.04, 3.01, 3.10, and 2.45 with SRU-NoSkip-IRM and SRU-NoSkip-IBM, marking progresses of 1.02 (33.55%) and 0.72 (29.4%), respectively. Correspondingly, C_{SIG} and C_{BAK} scores progressed from 3.04 and 2.45 with the SRU-NoSkip-IRM model to 3.10 and 2.49 with the SRU-WithSkip-IRM model. Lastly, C_{SIG} and C_{BAK} advanced from 3.05 and 2.43 with the SRU-WithSkip-IBM model to 3.13 and 2.48 with SRU-AttSkip-IBM.

Table X presents the performance in the seen noisy backgrounds. To evaluate our SE for noise generalization, Table XI presents the outcomes of PESQ and STOI tests for two unseen noises (cafeteria and factory2). Our SE models demonstrated significant performance over

TABLE-VIII. STOI and PESQ Test Scores in Five Example Noise Sources for LibriSpeech Dataset.

Metric	Model	IRM				IBM			
		-8dB	-4dB	0dB	4dB	-8dB	-4dB	0dB	4dB
STOI	SRU-NoSkips	63.9	75.0	81.7	88.3	62.9	71.1	80.4	86.3
	SRU-WithSkips	66.1	78.6	84.3	90.2	64.4	74.7	81.4	88.2
	SRU-AttenSkips	67.4	79.7	86.0	92.0	66.4	77.0	84.9	90.6
PESQ	SRU-NoSkips	1.72	2.04	2.30	2.66	1.66	1.98	2.28	2.61
	SRU-WithSkips	1.75	2.11	2.41	2.74	1.71	2.03	2.31	2.68
	SRU-AttenSkips	1.83	2.21	2.49	2.86	1.80	2.11	2.43	2.77

TABLE IX. C_{SIG} and C_{BAK} Test Scores in All Noise Sources at Four SNRs

Metric	C_{SIG}					C_{BAK}				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	1.38	1.78	2.22	2.69	2.02	1.35	1.59	1.83	2.14	1.73
SRU-NoSkips-IRM	2.35	2.85	3.22	3.74	3.04	1.95	2.27	2.65	2.92	2.45
SRU-NoSkips-IBM	2.33	2.82	3.19	3.71	3.01	1.90	2.21	2.60	2.86	2.39
SRU-WithSkips-IRM	2.40	2.89	3.31	3.81	3.10	1.97	2.32	2.68	2.98	2.49
SRU-WithSkips-IBM	2.35	2.84	3.23	3.76	3.05	1.94	2.26	2.63	2.90	2.43
SRU-AttenSkips-IRM	2.54	3.01	3.41	3.91	3.22	2.06	2.40	2.78	3.06	2.58
SRU-AttenSkips-IBM	2.47	2.96	3.28	3.82	3.13	1.98	2.32	2.66	2.97	2.48

TABLE X. STOI and PESQ Test Scores in Seen Noise Sources Against Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	51.7	59.6	68.5	77.1	64.2	1.24	1.48	1.67	1.91	1.58
SRU-AttenSkips-IRM	65.5	76.0	83.9	90.3	79.0	1.81	2.09	2.39	2.72	2.25
SRU-AttenSkips-IBM	64.6	74.9	83.2	89.1	78.0	1.77	2.03	2.35	2.67	2.21
LSTM-IRM [3]	63.5	74.2	82.4	88.9	77.3	1.71	2.00	2.33	2.67	2.18
LSTM-IBM [3]	62.7	72.1	81.7	87.8	76.1	1.67	1.86	2.29	2.63	2.11
DNN-IRM [43]	58.5	70.0	78.7	85.6	73.2	1.57	1.75	2.19	2.53	2.01
DNN-IBM [43]	56.1	67.3	76.5	83.1	70.8	1.49	1.70	2.11	2.45	1.94
CNN [14]	59.3	70.0	79.8	86.8	74.0	1.62	1.83	2.25	2.59	2.07
GAN [3]	54.3	65.0	75.7	82.6	70.0	1.53	1.72	2.15	2.44	1.96

TABLE XI. STOI and PESQ Test Scores in Unseen Noise Sources Against Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	50.3	58.3	67.5	76.3	63.1	1.15	1.39	1.58	1.88	1.50
SRU-AttenSkips-IRM	64.3	74.8	82.7	90.0	78.0	1.79	2.05	2.34	2.69	2.22
SRU-AttenSkips-IBM	63.4	72.7	82.0	88.9	77.8	1.76	1.98	2.28	2.65	2.17
LSTM-IRM [3]	62.0	72.9	81.3	88.2	76.1	1.62	1.91	2.24	2.64	2.10
LSTM-IBM [3]	61.3	70.8	80.6	87.0	75.0	1.58	1.77	2.20	2.60	2.04
DNN-IRM [43]	57.0	68.6	77.7	84.8	72.0	1.48	1.66	2.10	2.51	1.94
DNN-IBM [43]	55.0	66.0	75.5	82.4	69.7	1.40	1.61	2.02	2.42	1.86
CNN [14]	57.9	68.7	78.8	86.0	72.9	1.53	1.74	2.16	2.56	2.00
GAN [3]	52.9	63.8	74.6	81.9	68.3	1.44	1.63	2.06	2.41	1.89

TABLE XII. STOI and PESQ Test Scores Against Unsupervised Competing SE Algorithms

Metric	STOI					PESQ				
SNR (in dB)	-8	-4	0	4	Avg	-8	-4	0	4	Avg
Noisy Mixture	50.3	58.3	67.5	76.3	63.1	1.15	1.39	1.58	1.88	1.50
SRU-AttenSkips-IRM	64.3	74.8	82.7	90.0	78.0	1.79	2.05	2.34	2.69	2.22
SRU-AttenSkips-IBM	63.4	72.7	82.0	88.9	77.8	1.76	1.98	2.28	2.65	2.17
LRSD [36]	54.3	63.2	70.6	79.2	66.8	1.38	1.71	1.98	2.28	1.83
NRPCA [33]	55.8	63.3	70.6	80.2	67.5	1.41	1.78	2.02	2.32	1.88
MMSE [6]	50.8	60.5	68.8	78.2	64.6	1.28	1.51	1.81	2.10	1.68

both baseline and competing networks in situations involving unseen noises. The most noteworthy STOI and PESQ scores were achieved by SRU-WithSkip-IRM and SRU-WithSkip-IBM models, owing to their sophisticated network architecture. With robust acoustic features and architectural changes applied to the proposed SRU models, the performance remained relatively stable across both seen and unseen noises. For instance, the average STOI values observed improvements of 14.9% and 13.7% over noisy speech when using SRU-WithSkip-IRM and SRU-WithSkips-IBM, respectively, resulting in STOI values of 78% and 76.8%. Notably, at SNRs (-4dB and -8dB) SRU-WithSkip-IRM and SRU-WithSkip-IBM exhibited STOI improvements of 1.89% and 1.78% as compared to baseline SRU (SRU-IRM and SRU-IBM). Furthermore, PESQ scores observed significant improvements, reaching 2.22 (32.43%) and 2.17 (31.90%) with SRU-WithSkip-IRM and SRU-WithSkip-IBM, respectively, compared to score of 1.50. This represents a substantial improvement in PESQ in unseen noisy conditions, outperforming the noisy speech. Across various metrics, our SRU models showcased considerable performance improvements compared to baseline SRU (SRU-IRM and SRU-IBM) and related models. Specifically, our SRU models exhibited STOI enhancements of 1.80%, 2.90%, 5.90%, 8.20%, 5%, and 9.6%, respectively, while also enhancing PESQ by factors of 0.10 (4.54%), 0.16 (7.27%), 0.26 (11.8%), 0.34 (15.45%), 0.20 (9.1%), and 0.31 (14.1%).

We present the results of the proposed models and their competitive counterparts, evaluated using STOI and PESQ metrics. The findings indicate notable improvements in speech quality, intelligibility, noise suppression, and speech distortions attributable to the proposed SRU. These models also showed superior performance when compared to the baseline SRU [25], DNN [12], CNN [43], and GAN (employing a 3-layer ReLU MLP) [19]. The results encompassing the proposed and competing models are tabulated in Table X and Table XI. The obtained results are averaged over all SNRs. The obtained scores show improvements in intelligibility and quality of the proposed SRU models for SE. For STOI, SRU-AttSkip-IRM and SRU-AttSkip-IBM surpassed DNN-IRM and DNN-IBM by 3.99% and 6.7%, respectively.

SRU-AttSkip-IRM showcased STOI improvement of 5.10% and 9.7% over CNN and GAN, whereas SRU-AttSkip-IBM showed an improvement of 8.49% over GAN and 5.01% over CNN. For PESQ, SRU-AttSkip-IRM, and SRU-AttSkip-IBM enhanced by 9.09% and 14.09% as compared to CNN and GAN, respectively. Moreover, a comparison with three unsupervised techniques was conducted to show success of supervised learning over unsupervised counterparts. These unsupervised methods include Low-rank sparse decomposition (LRSD) [44], Nonnegative RPCA (NRPCA) [45], and MMSE [46] for SE. Table XII provides the results. STOI scores showed improvements of 11.2%, 10.5%, and 13.4% with SRU-AttSkip-IRM, and 11%, 103%, and 13.2% with SRU-AttSkip-IBM. Similarly, PESQ scores showed improvements of 0.39 (equal to 17.56%), 0.34 (equivalent to 15.31%), and 0.54 (equal to 24.32%) with SRU-AttSkip compared to LRSD, NRPCA, and MMSE, respectively.

B. Spectrogram Analysis

To illustrate the spectral parts of speech that have undergone processing, this section presents a spectro-temporal analysis. Fig. 7 shows the spectrograms representing different speech utterances. In Fig. 7(a), we observe the spectrogram of clean speech. Fig. 7(b) shows that a clean sentence is mixed at 0dB with babble noise, resulting in noisy speech. This specific noisy condition is noteworthy because the noise characteristics resemble those of the target speech. Fig. 7(c) shows the enhanced speech through SRU-IBM, where background noise is evident. Figure 7(d) portrays the signals of SRU-IRM-enhanced speech, showcasing even lower levels of residual noise and distortion when compared to SRU-IBM. Figure 7(e) shows a spectrogram of the enhanced speech generated by SRU-AttSkip-IRM. This presentation indicates reduced distortion and residual noise. Concluding this experiment, Fig. 7(f) illustrates the enhanced speech attributed to SRU-AttSkip-IBM, revealing less distortions and residual noise.

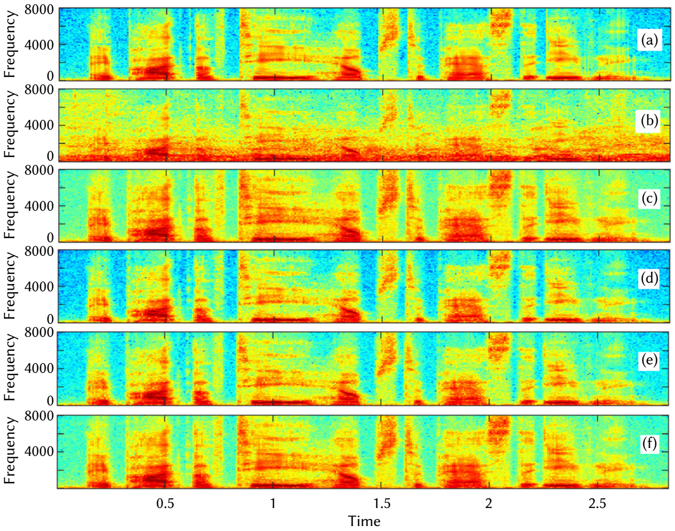


Fig. 7. Spectro-Temporal Analysis. (a) Clean, (b) Noisy, (c) SRU-IBM, (d) SRU-IRM, (e) SRU-AttSkip-IBM, and (f) SRU-AttSkip-IRM.

C. Computational Complexity

In practical applications, computational resources are frequently limited. Therefore, it is essential to strike a suitable balance between model performance and computational efficiency. Table XIII presents efficiency of parameters in the proposed SRU model. The assessment shows that implementing SRU-based SE with attention gates in skip connections considerably reduces the trainable parameters (2.607M) and parameter size (12.54MB), in comparison to similar models such as LSTM (17.384M, 65.42MB), GRU (13.33M, 52.33MB), and the baseline SRU (8.69M, 35.98MB). The introduction of attention gates into the skip connections leads to a slight increase in parameter count. In current DNN research for SE, there is a focus on enhancing model performance for hardware accelerators to execute these models rapidly and efficiently. To make the suggested SRU applicable in embedded systems, there is a need to minimize hardware memory consumption. This necessitates an examination of multiply–accumulate operations (MACs), where it is evident that the proposed SRU exhibits the lowest MACs (0.986 G/s with attention gates), ensuring efficient execution without compromising speech enhancement performance. Our SRU-based approach not only significantly reduces parameter size but also minimizes MACs. Additionally, the study evaluates the Real-Time Factor (RTF) in the proposed model, measuring the ratio of processing time to input audio data duration, which is crucial for real-time applications. Conducted on single-core Intel(R) Core (TM) i5-1135G7 CPU @ 2.40GHz processor, the experiment yields an RTF of 0.36.

TABLE XIII. Computational Complexity and Efficiency

Model	Para#	MACs	Memory
LSTM	17.38M	3.291 G/s	65.42 MB
GRU	13.33M	2.605 G/s	52.33 MB
SRU-(Baseline)	08.69M	1.554 G/s	35.98 MB
SRU-(Proposed)	2.607M	0.986 G/s	12.54 MB

D. Automatic Speech Recognition (ASR)

The results of the SE assessments demonstrate notable reduction of background noises and successfully recover the speech with a better intelligibility. Consequently, this study expects improved speech recognition capabilities, especially in presence of difficult noisy environments. Our speech enhancement model can be served as a front end preprocessor for obtaining lower word error rates (WER as ASR results). For ASR, this study implements the Kaldi toolkit [46], which adopts a GMM-HMM system and subsequently trains a DNN utilizing Mel-frequency filter-bank features. The training methodology draws inspiration from Tachioka [47]-[48]. Assessment of ASR performance is based on word error rates. For training the proposed SRU-driven speech enhancement models, a random selection of 2000 utterances was made from TIMIT and LibriSpeech. Following training the SRU models, the speech enhancement process was performed, leading to the synthesis of time-domain utterances. The synthesized time-domain utterances were used to train ASR models. As shown in Table XIV, ASR system presented improved performance when trained with data processed by SRU-AttSkip. The WERs showed a gradual reduction with more favorable SNR levels. On average, a WER of 14.25% was attained with the utilization of utterances processed by the proposed SRU-AttSkip. The results show that the proposed approach can effectively be utilized as a pre-processor to enhance ASR performance.

E. Performance Comparison Using VoiceBank+DEMAND

This section conducts experiments on the VoiceBank+DEMAND. The purpose of these experiments is to highlight the validation of the proposed SE approaches in comparison to contemporary benchmarks. The results of these experiments are detailed in Table XV. The analysis of results, as shown by PESQ, STOI, and Segmental SNR (SNRSeg), reveals certain observations. Higher values in these metrics signify enhanced performance. Notably, the experiments showcase results for SRU with skips, emphasizing a superior performance of attention skips with SRU. Our SRU perform better on VoiceBank+DEMAND and obtains competitive results: PESQ, STOI, SNRSeg, and trainable parameters. From GAGNet [49], the proposed SRU improves the metrics by 0.22 (PESQ), 0.5% STOI, and 0.64dB (SNRSeg), respectively. Further, from DCCRN [51], the proposed SRU improves the metrics by 0.47 (PESQ), 1.5% STOI, and 1.26dB (SNRSeg). The parameter count of the proposed SRU is better than all models except TSTNN [54]; however, PESQ = 2.96, STOI = 95.1%, and SNRSeg = 9.72dB are not reasonable to SRU. Our SRU obtains superior STOI, PESQ, and SNRSeg results with fewer trainable parameters (2.607M), MACs (0.986 G/s), and memory size (12.54MB).

F. Additional Experiments

We further examined the proposed SRU for different languages additional to English such as Urdu, Turkish, and Spanish. Although the speech enhancement aims to reduce the background noises from target speech, the results in Table XVI show that they are not severely degraded by different languages. However, since Urdu, Turkish, and Spanish are low-resource languages, the ASR performance is severely degraded in noisy environments. The STOI and PESQ values for all languages are marginally different which indicates that SE performance is not significantly affected by languages. To improve WERs, different strategies, such as speech augmentation and neural speech synthesis, can be used for ASR.

TABLE XIV. Computational Complexity and Efficiency

DNN Model	Noisy Mixture	SRU-AttSkip-IRM	SRU-AttSkip-IBM	SRU-IRM	SRU-IBM	DNN-IRM	DNN-IBM
WERs	55.35%	14.25%	14.75%	19.20%	19.95%	29.25%	30.02%

TABLE XV. Performance Evaluation on the VoiceBank+DEMAND Database

Model	Domain	Year	Parameter #	PESQ	STOI	SNRSeg
SEGAN [50]	Time	2017	97.5M	2.16	93.1	7.66
DCCRN [51]	Time-Frequency	2019	3.70M	2.68	93.7	8.62
GAGNet [49]	Time-Frequency	2021	5.64M	2.94	94.7	9.24
RDL-Net [52]	Time-Frequency	2020	3.91M	3.02	93.8	--
DEMUCS [53]	Time	2020	128M	3.07	95.1	8.53
TSTNN [54]	Time	2021	0.92M	2.96	95.1	9.72
SE-Conformer [55]	Time-Frequency	2021	--	3.13	95.1	--
PFR-Net [58]	Time-Frequency	2022	4.61M	3.19	95.0	--
FAF-Net [59]	Time-Frequency	2022	6.90M	3.24	95.0	--
MAB-CED [60]	Time-Frequency	2022	4.82M	2.84	85.0	--
SRU (Baseline)	Time-Frequency	2024	8.69M	3.09	94.4	9.11
SRU (Proposed)	Time-Frequency	2024	2.61M	3.15	95.4	9.88

TABLE XVI. ASR for Different Languages

Model	SRU-AttSkip-IRM			SRU-AttSkip-IBM		
Language	STOI	PESQ	WERs	STOI	PESQ	WERs
English	78.12	2.19	14.25	78.03	2.15	14.75
Urdu	78.10	2.14	30.14	77.94	2.10	31.21
Turkish	77.95	2.09	20.22	77.84	2.04	22.14
Spanish	77.47	2.01	26.11	77.34	1.98	27.25

VI. Conclusions

In this study, a novel speech enhancement (SE) system is introduced, utilizing lightweight recurrent neural networks (RNNs) that have been trained with robust features. The approach includes the development of an hourglass SRU model, which effectively captures temporal dependencies by decreasing feature resolutions. To counteract gradient decline across layers, nonadjacent symmetrical layers are connected through skip connections. Furthermore, an attention gate is integrated into these skips, aiming to emphasize significant features and spectral regions. Composite and robust features are derived from the noisy magnitude, enhancing training of the proposed models for improved SE and ASR performance. The model independently estimates two masks: ideal ratio mask and ideal binary mask. The findings show several key aspects of the proposed speech enhancement model. The incorporation of combined feature learning allows for the integration of additional information, enabling the network to grasp the intricate nonlinear relationship between noisy and clean speech. The adopted SRU architecture proficiently captures long-term temporal dependencies while downsizing the feature resolution for parameter estimation during testing, thereby preventing excessive memory usage. The introduction of the skips and attention gates within these skips significantly addressed gradient decay across layers, additionally highlighting important features and spectral regions. The proposed SRU strategy contributes to superior performance compared to the baseline, as evident from enhanced trainable parameter metrics. The proposed speech enhancement model outperforms the recent deep SE representations across diverse background noises, showing promising outcomes concerning speech distortion and residual noise. Notably, the models showcase superior results not only in seen noisy environments but also in unseen noise contexts. Results from GMM-HMM ASR show the potential of SRU-AttSkip SE model as a preprocessor for enhancing ASR performance in noisy environments. Our SRU model performs better on the VoiceBank+DEMAND and obtains competitive results: PESQ, STOI, SNRSeg, and trainable parameters. Further, the proposed SRU has the lowest MACs (0.986 G/s) with attention gates. SRU obtains superior STOI, PESQ, and SNRSeg results with fewer trainable parameters (2.607M), MACs (0.986 G/s), and memory size (12.54MB).

The SRU model can be sensitive to the noisy inputs. If the input data contains significant noise or errors, the model's performance may degrade. In the future, we anticipate the extension of the proposed network architecture to address this limitation and use for regression-based speech enhancement (SE), jointly optimized with application to automatic speech recognition and automatic speaker recognition. Additionally, the potential for devising more robust acoustic features is observed, offering the prospect for further work.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/383/44.

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding

The current work was supported by the Deanship of Scientific Research at King Khalid University through large group Research Project under grant number RGP2/383/44.

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113– 120, 1979.

[2] S. Nasir, A. Sher, K. Usman, U. Farman, "Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation", *Research Journal of Applied Sciences,* Engineering and Technology, vol. 6, no. 6, pp. 1081–1087, 2013.

[3] J. Lim, A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[4] Y. Ephraim, D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] N. Mohammadiha, P. Smaragdis, A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2140–2151, 2013.

[7] I. Tashev, M. Slaney, "Data driven suppression rule for speech enhancement," in: *2013 Information Theory and Applications Workshop (ITA)*, IEEE, 2013, pp. 1–6.

[8] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[9] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[10] M. Kolbæk, Z.-H. Tan, J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 1, pp. 153–167, 2016.

[11] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] Y. Wang, A. Narayanan, D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] N. Saleem, M.I. Khattak, "Deep neural networks for speech enhancement in complex-noisy environments", *International Journal of Interactive Multimedia and Artificial Intelligence,* vol. 6, no. 1, pp. 84–91, 2020.

[14] N. Saleem, M.I. Khattak, M. Al-Hasan, 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.

[15] N. Saleem, M.I. Khattak, "Multi-scale decomposition based supervised single channel deep speech enhancement," *Applied Soft Computing*, vol. 95, pp. 106666, 2020.

[16] N. Saleem, M.I. Khattak, M. Al-Hasan, A. Jan, "Multi-objective long-short term memory recurrent neural networks for speech enhancement," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 10, pp. 9037– 9052, 2021.

[17] S. Samui, I. Chakrabarti, S.K. Ghosh, "Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network," *Applied Soft Computing*, vol. 74, pp. 583–602, 2019.

[18] M.H. Soni, N. Shah, H.A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5039–5043.

[19] N. Shah, H.A. Patil, M.H. Soni, "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," in: *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2018, pp. 1246–1251.

[20] W. Yu, J. Zhou, H. Wang, L. Tao, "Setransformer: speech enhancement transformer," *Cognitive Computation,* vol. 14, pp. 1152-1158, 2022.

[21] J. Cadore, F.J. Valverde-Albacete, A. Gallardo-Antolín, C. Peláez-Moreno, "Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement," *Cognitive computation,* vol. 5, pp. 426–441, 2013.

[22] I. Sutskever, O. Vinyals, Q.V. Le, "Sequence to sequence learning with neural networks," in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, 2014, pp. 3104–3112.

[23] I. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[24] K. Zarzycki, M. Ławryńczuk, "LSTM and GRU neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors," *Sensors*, vol. 21, no. 16, pp. 5625, 2021.

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

[26] M. Sundermeyer, H. Ney, R. Schlüter, "From feedforward to recurrent lstm neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 3, pp. 517–529, 2015.

[27] M. Fernández-Díaz, A. Gallardo-Antolín, "An attention long short-term memory based system for automatic classification of speech intelligibility," *Engineering Applications of Artificial Intelligence,* vol. 96, pp. 103976, 2020.

[28] M. Q. Gandapur, E. Verdú, "ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 8, no. 4, 2023.

[29] N. Saleem, J. Gao, M.I. Khattak, H.T. Rauf, S. Kadry, M. Shafi, "Deepresgru: Residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition," *Knowledge-Based Systems,* vol. 238, pp. 107914, 2022.

[30] J. Ali Reshi, R. Ali, "An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network," *International Journal of Interactive Multimedia and Artificial Intelligence*, pp. 1-13, 10.9781/ijimai.2023.02.007.

[31] B. Chang, L. Meng, E. Haber, F. Tung, D. Begert, "Multi-level residual networks from dynamical systems view," arXiv preprint arXiv:1710.10348, 2017.

[32] Y. Shao, S. Srinivasan, Z. Jin, D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language,* vol. 24, no. 1, pp. 77–93, 2010.

[33] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom," NIST speech disc 1-1.1. NASA STI/Recon technical report, no. 93, 27403, 1993.

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

[35] D. Pearce, J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," *Inst. for Signal & Inform. Process.*, Mississippi State Univ., Tech. Rep, 2002,

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

[37] T. F. Damayanti, A. Wanto, H.S. Tambunan, "Prediction of Palm Oil Seed Stock Production Results with the Back-propagation Algorithm," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 2, no. 2, pp. 105-112, 2023.

[38] Q. Song, Y. Wu, Y.C. Soh, "Robust adaptive gradient-descent training algorithm for recurrent neural networks in discrete time domain," *IEEE Transactions on Neural networks,* vol. 19, no. 11, pp. 1841–1853, 2008.

[39] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[40] A.W. Rix, M.P. Hollier, A.P. Hekstra, 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.

[41] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 4214–4217.

[42] Y. Hu, P.C. Loizou, "Evaluation of objective measures for speech enhancement," in: *Ninth International Conference on Spoken Language Processing*, 2006.

[43] T. Kounovsky, J. Malek, "Single channel speech enhancement using convolutional neural network," in: *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics (ECMSM)*, IEEE, 2017, pp. 1–5.

[44] P. Sun, J. Qin, "Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1862–1866, 2016.

[45] W. Shi, X. Zhang, X. Zou, W. Han, G. Min, "Auditory mask estimation by RPCA for monaural speech enhancement," in: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, IEEE, 2017, pp. 179–184.

[46] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, 2011.

[47] Y. Tachioka, S. Watanabe, J. Le Roux, J.R. Hershey, "Discriminative methods for noise robust speech recognition: A chime challenge bench-mark," in: *The 2nd International Workshop on Machine Listening in Multisource Environments*, 2013, pp. 19–24.

[48] A. Shewalkar, D. Nyavanandi, S.A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research,* vol. 9, no. 4, pp. 235-245, 2019.

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

[50] S. Pascual, J. Serra, A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech communication*, vol. 114, pp. 10-21, 2019.

[51] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, ... & L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," arXiv preprint arXiv:2008.00264, 2020.

[52] M. Nikzad, A. Nicolson, Y. Gao, J. Zhou, K.K. Paliwal, 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.

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

[54] K. Wang, B. He, 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.

[55] E. Kim, H. Seo, "SE-Conformer: Time-Domain Speech Enhancement Using Conformer," in *Interspeech,* 2021, pp. 2736-2740.

[56] Z. Ye, N. Saleem, H. Ali, "Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023, doi: 10.9781/ijimai.2023.05.007.

[57] M.I. Khattak, A. Jan, N. Saleem, E. Verdú, N. Khurshid, "Automated detection of COVID-19 using chest X-ray images and CT scans through multilayer-spatial convolutional neural networks," *International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6*, no. 6, pp. 15-24, 2021.

[58] G. Yao, C. Wang, Y. Wu, Y. Wang, "Pyramid fully residual network for single image de-raining," *Neurocomputing*, vol. *456*, pp.168-178, 2021.

[59] H. Yue, W. Duo, X. Peng, 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. 11648-11656.

[60] N. Saleem, T.S. Gunawan, M. Shafi, S. Bourouis, A. Trigui, "Multi-Attention Bottleneck for Gated Convolutional Encoder-Decoder-Based Speech Enhancement," *IEEE Access*, vol. *11*, pp. 114172-114186, 2023

Sami Dhahbi

Sami Dhahbi received the engineering and M.S. degrees from the National School of Computer Science, University of Manouba, Tunisia, in 2005 and 2006, respectively, and the Ph.D. degree in computer science from the University of Tunis, El Manar, Tunisia, in 2016. He is currently an assistant professor of computer science at King Khalid University, Saudi Arabia. He is also a member of the LIMTIC Research Laboratory at the University of Tunis, El Manar. He is the author of several articles. His research interests include machine learning, medical imaging, and more recently, networks and cloud computing.

Nasir Saleem

Nasir Saleem received B.S. M.S. and PhD Engineering degree from University of Engineering & Technology, Peshawar-25000, Pakistan, in 2008; 2012, and 2021 with specialization in speech processing and deep learning. Did postdoctoral fellow, Islamic International University Malaysia (IIUM), researching the artificial intelligence-based speech processing algorithms. From 2008 to 2012, he was a senior lecturer at the Institute of Engineering Technology (IET), Gomal University, where he was involved in teaching and research. He is now an assistant professor in the Department of Electrical Engineering, Faculty of Engineering and Technology (FET), and Deputy Director of the Quality Assurance Directorate at Gomal University. Human-machine interaction, speech enhancement, speech recognition, speech and video processing, and machine learning applications are the areas he is currently researching.

Teddy Surya Gunawan

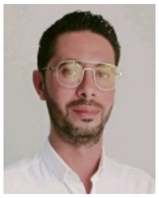
Teddy Surya Gunawan (Senior Member, IEEE) received his BEng degree in Electrical Engineering with cum laude award from Institut Teknologi Bandung (ITB), Indonesia, in 1998. He obtained his M.Eng degree in 2001 from the School of Computer Engineering at Nanyang Technological University, Singapore, and a Ph.D. degree in 2007 from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia. He was a Visiting Research Fellow (2010 to 2021) at UNSW and is currently an Adjunct Professor at Telkom University (2022-2023). His research interests are speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award in 2018 from IIUM. He is currently an IEEE Senior Member (since 2012), was chairperson of the IEEE Instrumentation and Measurement Society – Malaysia Section (2014, 2020, 2021), Professor (since 2019), Head of Department (2015-2016) at the Department of Electrical and Computer Engineering, and Head of Programme Accreditation and Quality Assurance for Faculty of Engineering (2017-2018), International Islamic University Malaysia. He has been a Chartered Engineer (IET, UK) since 2016 and Insinyur Profesional Utama (PII, Indonesia) since 2021, a registered ASEAN engineer since 2018, and ASEAN Chartered Professional Engineer since 2020.

Sami Bourouis

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.

Imad Ali

Imad Ali received B.Sc. Telecom Engineering degree from the University of Engineering and Technology, Peshawar, Pakistan in 2008, M.S. Electrical Engineering degree from CECOS University, Peshawar, Pakistan in 2012; and a Ph.D. degree in Social Networks and Human-Centered Computing, from National Tsing Hua University, Taiwan, in collaboration with Academia Sinica, Taiwan, in 2020. From 2009 to 2014, he served as a lecturer at different universities in Pakistan. He is currently working as an Assistant Professor in the Department of Computer Science, University of Swat, Pakistan. His research interest includes Question Answering Systems, Data Science, and Machine Learning.



Aymen Trigui

Aymen Trigui 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 an Associate Professor at the Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests include data mining, image processing, statistical machine learning, cybersecurity, and pattern recognition applied to several real-life applications.

Abeer D. Algarni

Abeer D. Algarni received the B.Sc. degree (Hons.) in computer science from King Saud University, Riyadh, Saudi Arabia, in 2007, and the M.Sc. and Ph.D. degrees from the School of Engineering and Computer Sciences, Durham University, U.K., in 2010 and 2015, respectively. She has been an Assistant Professor with the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, since 2008. Her current research interests include networking and communication systems, digital image processing, digital communications, and cyber security.