JOURNAL OF SCIENCE, TECHNOLOGY AND ENGINEERING RESEARCH

Bilim, Teknoloji ve Mühendislik Araştırmaları Dergisi

ISSN (Online) 2717-8404

Available *online* at *https://dergipark.org.tr/en/pub/jster*

RESEARCH ARTICLE

A Supervised Learning Approach With Residual Attention Connections

[1] Hamza Ali, [1] Fazal Muhammad, [2] Talha Ali, [3] Fazal-e-Wahab, [1] Muhammad Ismail,

[1]Department of Electrical Engineering, University of Engineering & Technology, Mardan, KPK (Pakistan)
engr.hamarwat@gmail.com, fazal.muhammad@uetmardan.edu.pk, m.ismail012018@gmail.com
[2]Department of Electrical Engineering, University of Engineering & Technology, Peshawar, KPK (Pakistan)
talhanisar849@gmail.com
[3]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026 Anhui (China)

Citation:
Hamza A., Fazal M., Talha Ali, Fazal-e-W., Muhammad I. (2024). *A supervısed learnıng approach wıth resıdual attentıon connectıons*, Journal of Science Technology and Engineering Research, 5(1): 78-85. 10.53525/jster.1469477

HIGHLIGHTS

- *Effect and important of this article in literature*
- *Exchange between sources in related subjects of this article*
- *Contribution and strongest impact on the related sobject of this article*
- *Examined study and obtained results why is important*

Article Info

Received **:** April 16, 2024
Accepted **:** May 27, 2024

DOI:
10.53525/jster.1469477

[1]Corresponding Author:

Hamza Ali
engr.hamarwat@gmail.com,
Phone: +92 333 9971656

ABSTRACT

The proposed study aims to improve speech quality despite background noise, which often disrupts clear communication. We focus on developing efficient and effective models that work well on devices with limited resources. We draw inspiration from computational auditory scene analysis techniques to train proposed models to differentiate speech from background noise while keeping computational demands low. We introduce two models: CRN-WRC (Convolutional Recurrent Network without Residual Connections) and CRN-RCAG (Convolutional Recurrent Network with Residual Connections and Attention Gates). Our thorough testing shows that proposed models significantly enhance speech quality and understanding, even in noisy environments with varying background noise levels. Notably, the CRN-RCAG model consistently outperforms the CRN-WRC, particularly in handling untrained noise types. We achieve impressive results by integrating residual connections and attention gates into proposed models while maintaining computational efficiency.

Keywords: *Speech enhancement, Convolutional Recurrent Network, supervised learning, Gated Recurrent Unit, Residual-connection*

I. INTRODUCTION

We are constantly surrounded by sound, whether indoors or outside. We are constantly surrounded by many sorts of noise like traffic noise, and street noise however, this noise can occasionally interfere with effective communication. Even automated speech recognition (ASR) [1] technologies face difficulties in understanding us. To fix this, speech enhancement becomes critical. It seeks to improve speech quality by converting noisy input into clearer output. However, typical voice-augmentation algorithms can be computationally intensive, particularly for devices with limited resources. The proposed research focuses on developing lightweight and efficient models designed primarily for low-power devices with limited memory and processing capability. We aim to remove background noise from your voice without draining your device's resources. To achieve this, we use supervised learning techniques inspired by computational auditory scene analysis (CASA) [2], Supervised methods [3] involve creating separate models for speech and noise signals, with the parameters of these models learned using training examples that include both types of signals. We use masking and mapping [4] targets in the T-F domain to guide proposed models toward achieving optimal results without overwhelming them with complexity. The CRNN identifies critical speech components [5], [6], which analyses the time-frequency bands. As this is going on, the AG serves to let important information through while blocking out unwanted noise. The result: significantly enhanced speech quality, even in challenging acoustic environments. This research has significant implications for various applications. Improved speech quality and intelligibility directly benefit ASR systems, voice communication, and assistive hearing devices. By leveraging the magic of deep learning [7], we are paving the way for a future where clear communication is no longer hindered by noise. This paper contributes to this domain by offering a lightweight yet effective approach to single-channel speech enhancement, promising better speech quality and intelligibility. The proposed work bridges the gap between speech enhancement and resource constraints. By incorporating skip connections and attention gates within the CRNN architecture, we achieve remarkable results while ensuring computational efficiency.

II. PROPOSED METHOD

In proposed research, we employ two distinct models to tackle the task at hand. The first model, referred to as a conventional convolutional recurrent network without residual connections (CRN-WRC), integrates a recurrent block network for speech enhancement (SE). Meanwhile, the second model incorporates residual connections with (1×1) kernel size and attention gates (CRN-RCAG) within the CRNN architecture to enhance its performance.

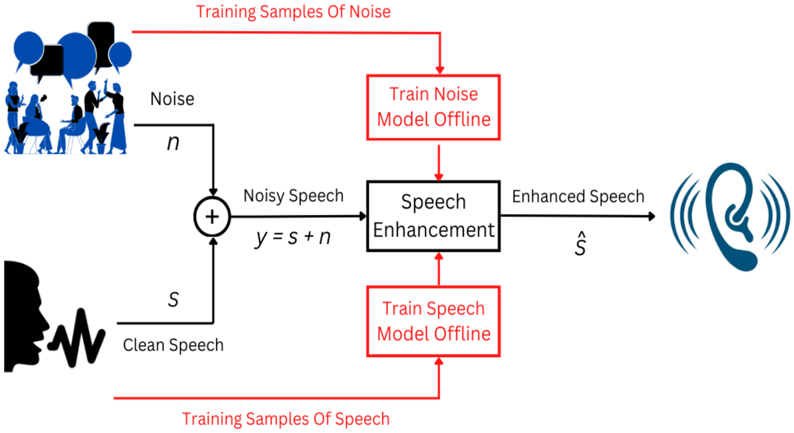

Training Samples Of Noise
Noise
n
Train Noise Model Offline
Noisy Speech
$y = s + n$
Speech Enhancement
Enhanced Speech
$\hat{s}$
s
Clean Speech
Train Speech Model Offline
Training Samples Of Speech

Fig. 1. Supervised Approach to Noise Reduction: Speech and Noise Model Interaction

A. *Conventual CRN-WRC Model*

The architecture of proposed Convolutional Recurrent Network is depicted in Fig.2. shows the network architecture, which consists of five layers for the causal encoder and decoder, respectively, that are convolutional (Conv2D) and deconvolutional (Deconv2D) [8], [9]. Exponential linear units are used in all layers but the output layer, which uses soft-plus activation [10]. A Gated Recurrent Unit (GRU) layer models the latent feature sequences after inputs are first encoded into a high-dimensional latent space [11]. Then the GRU layer's output sequences are converted back into their original shape by the decoder. This method combines two powerful topologies convolutional neural networks (CNNs) [12] for feature extraction and recurrent neural networks (RNNs) [13] for temporal modelling to produce better outcomes In proposed study, we leverage residual connections to enhance information flow and address redundancy in speech enhancement. Specifically, we connect the encoder outputs to the decoder inputs using these residual connections as depicted in Fig.3. Additionally, we incorporate attention gates (AGs) [14] within these connections to effectively reduce redundant regions while emphasizing important spectral features. Given the large number of frequency components in the spectra, we recognize that formant frequencies dominate low-frequency areas, while high-frequency regions exhibit a sparse distribution. In speech, certain frequencies are more important than others. For example, low-frequency areas usually have more dominant sounds, while high-frequency areas have fewer sounds. So, it's crucial to give more attention to the important parts of the speech signal. Therefore, it is crucial to differentiate distinct spectral locations with varying weights.

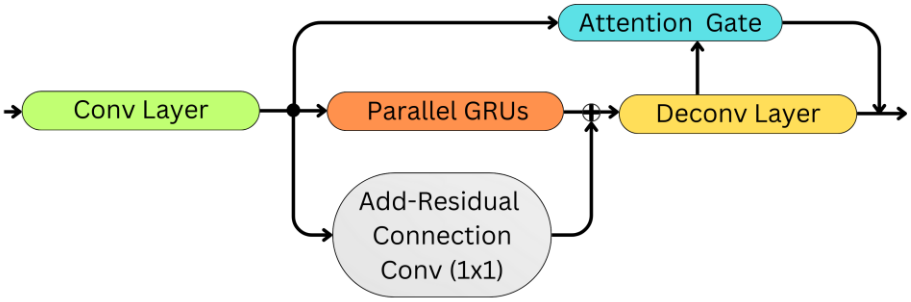


Fig. 3. CRN Architecture for Speech Enhancement with Residual Connections and Attention Gates

B. Extending the Proposed Model

The encoder's job is to take the input features and find the important parts while making the information easier to work with. It does this by passing the input features through five layers that squeeze them down and make them more manageable. Then, special layers called GRU layers help the model understand the order of the information over time, and another layer helps adjust the features even more. AGs help the model focus on the most important parts of the input, ignoring the rest. These AGs work by comparing the features from different parts of the model and deciding what's important. The decoder's job is to take the simplified features and turn them back into the original input size. This completes the cycle, making the whole process like a mirror image. By adding Attention Gates, the decoder becomes even better at generating accurate results. Additionally, residual connections (1 x 1) kernel size helps the flow of information by linking the output of one part of the model to the input of another. This ensures that important information isn't lost as it moves through the network.

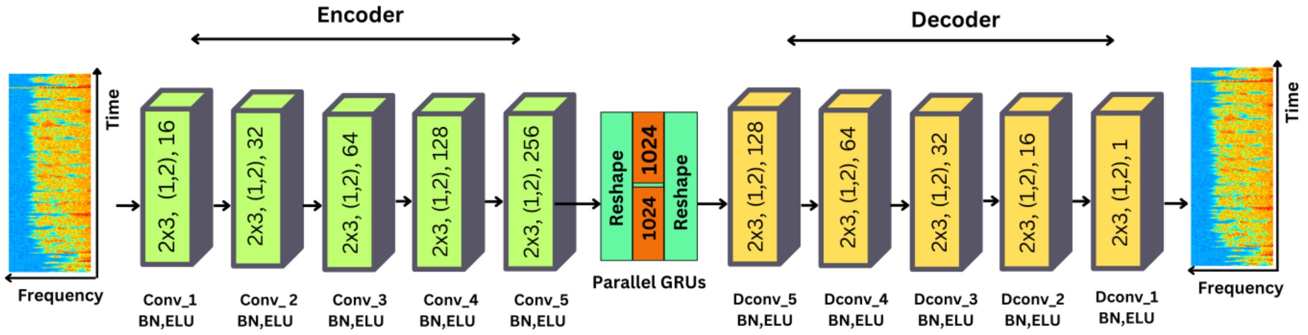


Figure. 2. The architecture of Convolutional Recurrent Network without residual connections

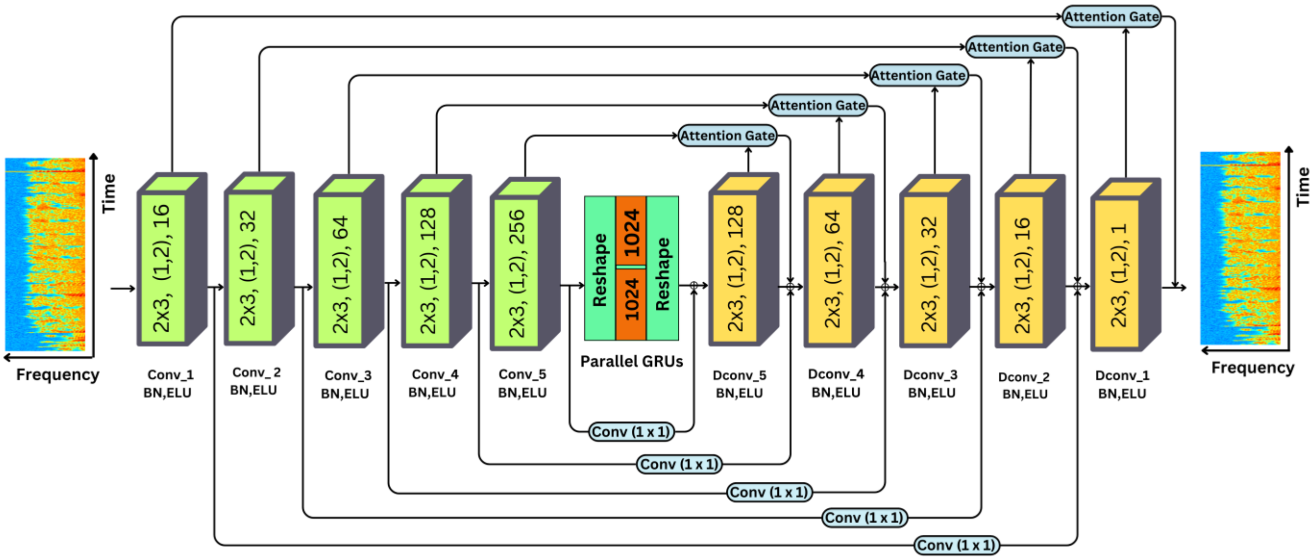


Fig. 4. CRN Architecture for Speech Enhancement with Residual Connections and Attention Gates

C. *Parallel GRUs For Temporal Modelling*

In proposed experiments, we use Gated Recurrent Unit (GRU) models to help us predict what a future speech frame might look like based on past frames. We used a window of 11 frames, including 10 past frames and 1 present frame, to make this prediction. We fed a long vector of these 11 frames into the network at each step. GRU is a newer type of recurrent neural network known for its memory cells, which are good at understanding patterns over time. To capture the timing of speech signals, we added a GRU layer between the encoder and decoder parts of proposed model. The equations describe how GRU works, including terms like z, r, and h, and illustrate the update gate and reset gate, which helps the model learn from the data.

$$z_l = \sigma(W_z[x_t, h_{t-1}] + b_z) \qquad (1)$$

$$r_t = \sigma(W_r[x_t, h_{t-1}] + b_r) \qquad (2)$$

$$\bar{h}_l = \tanh(W_h[r_l \odot h_{l-1}, x_l] + b_h) \qquad (3)$$

$$h_l = (I - z_l) \odot h_{l-1} + (z_l) \odot \bar{h}_t \qquad (4)$$

To handle the input shapes required by GRUs, we used a technique from a previous study to group the large network into smaller, parallel networks. This helps the model process information more efficiently.

III. EVALUATION METRICS

To evaluate the proposed networks, we used two primary performance metrics: Short-Time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ).

A. *STOI (Short-Time Objective Intelligibility)*

STOI is a metric used to measure the intelligibility of speech, particularly in noisy conditions. It is designed to predict the intelligibility of processed speech signals by comparing temporal envelopes of the original and degraded signals in short-time windows. Higher STOI values indicate better speech intelligibility. We chose STOI because it is highly correlated with human speech intelligibility scores, making it a reliable measure for our speech enhancement task.

B. PESQ (Perceptual Evaluation of Speech Quality)

PESQ is a widely used objective measure for assessing the quality of speech signals. It compares an original clean signal with a processed or degraded version, providing a score that reflects perceived speech quality. PESQ takes into account various perceptual aspects of speech quality, making it a comprehensive metric for our evaluation. Higher PESQ scores indicate better speech quality. PESQ was chosen due to its robustness and its high correlation with subjective listening tests, which is crucial for evaluating enhancements in real-world noisy environments.

C. Speech Dataset and Experimental Setup

In proposed research study, we tested speech enhancement networks using a large dataset called LibriSpeech [15], which includes recordings of people reading audiobooks that contain 0.22 million spoken sentences from 2,000 speakers. We focused on a subset of this dataset called LibriClean, which contains clean recordings, consisting of 104,015 high-quality utterances from 92 speakers. To evaluate proposed networks, we randomly selected 5000 speech samples from 40 speakers. Out of these speakers, we used 2 males and 2 females for testing whom the network hadn’t heard before, and the rest were used for training. We also used various types of noise during training to make sure the networks could handle different environments. For testing, we used challenging noise types like multi-talker babble, street noise, and cafeteria noise. We tested proposed models with both trained and untrained speakers to see how well they could handle different voices. The proposed networks were trained using a method called Adam optimizer and optimized to reduce prediction errors.

IV. FINDINGS AND DISCUSSION

Table 1 compares the performance of two speech enhancement models: CRN-WRC (Convolutional Recurrent Network without Residual Connections) and CRN-RCAG (Convolutional Recurrent Network with Residual Connections and Attention Gates). For each model, the results are presented for different noise conditions: babble noise, street noise, and cafeteria noise, across various signal-to-noise ratio (SNR) levels ranging from -6 dB to 6 dB.

Under each noise condition and SNR level, two sets of results are provided: 1. noisy: Represents the speech signal before enhancement. 2. enh: Represents the speech signal after enhancement using the respective model. Comparing the results between CRN-WRC and CRN-RCAG models, it is observed that the CRN-RCAG model consistently outperforms the CRN-WRC model across all noise conditions and SNR levels, as indicated by higher STOI and PESQ scores. This suggests that CRN-RCAG with residual connections and attention gates provides superior speech enhancement compared to CRNWRC without residual connections. Both models effectively enhance speech quality, but CRN-RCAG stands out as the stronger performer across different noise conditions and SNR levels as illustrated in Fig. 5.

A. Speech Enhancement Performance in Familiar Noisy Environments

The table 1 compares the performance of two different speech enhancement models, denoted as CRN-WRC and CRN-RCAG. For each model, two metrics are evaluated: PESQ and STOI. PESQ measures the perceived speech quality, while STOI evaluates the speech intelligibility. Tables 2 and 3 display the results for various signal-to-noise ratios (SNR), ranging from -6 dB to 6 dB.

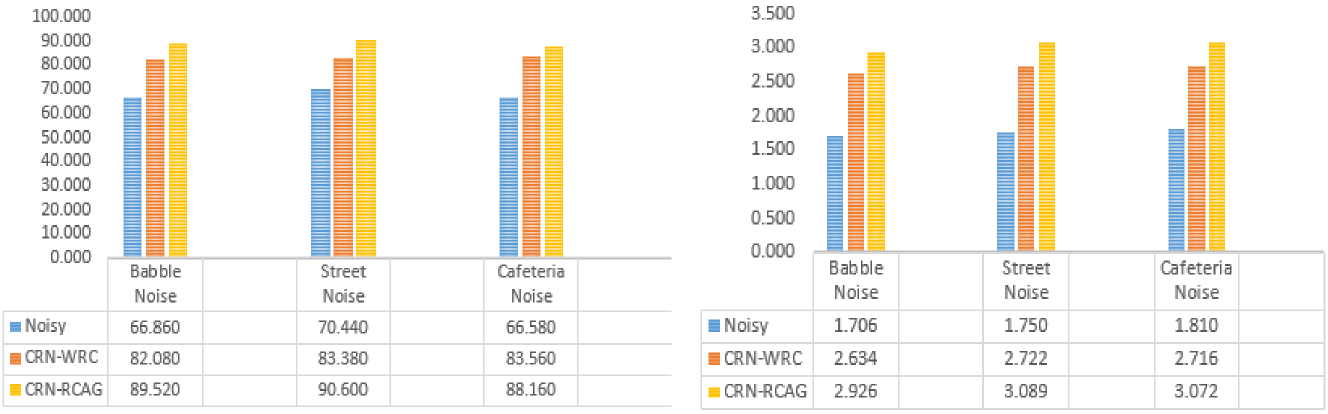
Under each SNR condition, two scenarios are considered: “noisy,” representing the original speech signal before enhancement, and “enh,” representing the speech signal after enhancement using the respective model. Looking at the results, we observe that both CRN-WRC and CRN-RCAG models significantly improve speech quality and intelligibility compared to the original noisy speech across all SNR levels. For the CRN-WRC model, the PESQ scores range from 1.39 to 2.18 for noisy speech signals, and the STOI percentages range from 53.7% to 78.3%. After enhancement using CRN-WRC, there is a significant improvement in both PESQ scores (ranging from 2.26 to 3.01) and STOI percentages (ranging from 72.6% to 88.9%). Similarly, for the CRN-RCAG model, the PESQ scores range from 2.57 to 3.40 for enhanced speech signals, and the STOI percentages range from 83.7% to 95.3%. The enhancement provided by CRN-RCAG demonstrates superior performance compared to CRN-WRC, particularly evident in higher PESQ and STOI scores across different noise levels.

B. *Speech Enhancement Performance in Unfamiliar Noisy Environments*

In Table 3, The models were tested on two different types of untrained noise environments, airport noise, and factory noise. We compared the performance of proposed models in noisy conditions with no training on those specific noises. We found that proposed models improved speech quality and intelligibility compared to the original noisy speech, especially in untrained noisy environments like factory noise and cafeteria noise. For noisy signal the PESQ scores range from 1.12 to 1.39, and the STOI scores range from 52.4% to 64.2% in the airport noise environment. In the factory noise environment, the PESQ scores range from 1.34 to 1.83, and the STOI scores range from 56.7% to 73.9%. After enhancement using CRN-WRC, the PESQ scores improve significantly to a range of 1.83 to 2.26 in the airport noise environment and 2.13 to 2.26 in the factory noise environment. Similarly, the STOI scores increase to a range of 67.8% to 73.9% in the airport noise environment and 67.5% to 73.9% in the factory noise environment. When using CRN-RCAG, the PESQ scores further improve to a range of 2.19 to 2.48 in the airport noise environment and 2.31 to 2.48 in the factory noise environment. Correspondingly, the STOI scores increase to a range of 72.6% to 78.3% in the airport noise environment and 77.1% to 79.3% in the factory noise environment. However, CRN-RCAG demonstrates better performance, achieving higher PESQ scores and STOI percentages in both noise environments, indicating its effectiveness in enhancing speech signals corrupted by untrained noises. Table 4, illustrates how incorporating residual connections in the CRN-RCAG architecture affects performance. Adding residual connections is better than not adding them at all. While adding these connections improves the PESQ and STOI test scores, the best performance is achieved by including Conv (1 × 1) add-residual and Attention Gate Residual connections.

TABLE I: Performance Evaluation of CRN Speech Enhancement Models: A Comparative Analysis with STOI and PESQ Metrics.

Metric	Model	Noise	Babble Noise					Street Noise					Cafeteria Noise				
		SNR	-6	-3	0	3	6	-6	-3	0	3	6	-6	-3	0	3	6
STOI	CRN-WRC	noisy	54	60	68	73	80	58	66	72	73	83	56	62	67	71	77
		enh	72	81	84	86	88	74	80	85	87	91	75	82	85	87	89
	CRN-RCAG	enh	84	87	90	92	94	85	87	91	94	97	80	85	88	93	94
PESQ	CRN-WRC	noisy	1.4	1.4	1.8	1.9	2.1	1.6	1.6	1.7	1.9	2.0	1.5	1.6	1.7	1.9	2.3
		enh	2.2	2.6	2.7	2.8	2.9	2.3	2.6	2.7	2.9	3.0	2.4	2.5	2.7	3.0	3.0
	CRN-RCAG	enh	2.4	2.7	2.9	3.2	3.3	2.6	3.0	3.1	3.3	3.4	2.6	3.0	3.1	3.3	3.4



	Babble Noise	Street Noise	Cafeteria Noise
🟦 Noisy	66.860	70.440	66.580
🟧 CRN-WRC	82.080	83.380	83.560
🟨 CRN-RCAG	89.520	90.600	88.160

	Babble Noise	Street Noise	Cafeteria Noise
🟦 Noisy	1.706	1.750	1.810
🟧 CRN-WRC	2.634	2.722	2.716
🟨 CRN-RCAG	2.926	3.089	3.072

(a) STOI (b) PESQ

Fig. 5. PESQ and STOI Performance of CRN-WRC and CRN-RCAG Models Across Various Noise Conditions

TABLE II: Performance of Trained Noise Speech Enhancement under Various Noise Conditions.

Model	Metric	PESQ					STOI (in%)				
	SNR	-6	-3	0	3	6	-6	-3	0	3	6
CRN-WRC	noisy	1.39	1.48	1.63	1.86	2.18	53.7	61.7	66.3	72.1	78.3
	enh	2.26	2.48	2.62	2.88	3.01	72.6	79.3	82.7	87.1	88.9
CRN-RCAG	enh	2.57	2.91	3.17	3.38	3.40	83.7	89.5	92.1	94.6	95.3

TABLE III: Performance of Untrained Noise Speech Enhancement under Various Noise Conditions.

Model	Noise	Restaurant				Street			
	Metric	PESQ		STOI		PESQ		STOI	
	SNR	-6	-3	-6	-3	-6	-3	-6	-3
CRN-WRC	noisy	1.12	1.32	52.4	59.7	1.34	1.39	56.7	64.2
	enh	1.83	2.16	67.8	72.1	2.13	2.26	67.5	73.9
CRN-RCAG	enh	2.19	2.47	72.6	78.3	2.31	2.48	77.1	79.3

TABLE IV: Analysing the Impact of Residual Connections on Model Performance

Residual Types	STOI	PESQ
No Residual	78	2.34
Add Residual	83	2.49
Conv Residual	87	2.73
Residual-Attn	89	3.01

V. CONCLUSION

In this paper, we present a novel approach to speech enhancement tailored for resource-constrained devices operating in noisy environments. By leveraging supervised learning techniques and innovative model architectures, we have successfully developed lightweight yet efficient models capable of significantly improving speech quality and intelligibility. Our experiments demonstrate the superiority of the CRN-RCAG model over the CRN-WRC model, emphasizing the importance of incorporating residual connections and attention gates for optimal performance. The results highlight the effectiveness of the proposed models in enhancing speech signals across various noise conditions and signal-to-noise ratio levels. Furthermore, the proposed models exhibit promising performance even when tested with untrained noise types, showcasing their robustness in real-world scenarios. Overall, this research bridges the gap between speech enhancement and resource constraints, offering a practical solution for clear communication in noisy environments.

Future work will focus on further optimizing and extending the proposed model architecture to enhance its applicability across a wider range of scenarios and devices. This includes: Exploring advanced techniques for model compression and acceleration to ensure even greater efficiency on resource-limited devices.Expanding the dataset to include a broader range of noise types and languages to enhance the model's generalizability and effectiveness in different linguistic contexts. Collaborating with industry partners to implement and test the proposed models in practical applications, such as mobile communication systems and hearing aids, to facilitate seamless adoption and impact.By pursuing these future directions, we aim to further advance the state of speech enhancement technology, making high-quality, intelligible speech accessible in various challenging environments and on a wide array of devices.

ACKNOWLEDGMENT

First and foremost, I would like to express my heartfelt gratitude and appreciation to my supervisor, Dr. Fazal Muhammad, and my Co-supervisor Dr. Fazal-e-Wahab, for guiding me the academic research, supporting and encouraging me when I encountered difficulties, teaching me the theoretical knowledge and providing me valuable suggestions throughout my entire research. I also would like to extend my deep gratefulness to my friends for their nice help and comforts throughout my life during my master study.

REFERENCES

[1] Kheddar, Hamza, et al. "Deep transfer learning for automatic speech recognition: Towards better generalization." Knowledge-Based Systems 277 (2023): 110851.
[2] Kwak, Chanbeom, and Woojae Han. "Towards size of scene in auditory scene analysis: A systematic review." Journal of Audiology & Otology 24.1 (2020): 1.
[3] Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." IEEE/ACM transactions on audio, speech, and language processing 26.10 (2018): 1702-1726.
[4] Nossier, Soha A., et al. "Mapping and masking targets comparison using different deep learning based speech enhancement architectures." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
[5] Ye, Zhongfu, Nasir Saleem, and Hamza Ali. "Efficient gated convolutional recurrent neural networks for real-time speech enhancement." (2023).
[6] Hsieh, Tsun-An, et al. "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement." IEEE Signal Processing Letters 27 (2020): 2149-2153.
[7] Wang, Kai. Novel Deep Learning Approaches for Single-Channel Speech Enhancement. Diss. Concordia University, 2022.
[8] Haar, Lynn Vonder, Timothy Elvira, and Omar Ochoa. "An analysis of explainability methods for convolutional neural networks." Engineering Applications of Artificial Intelligence 117 (2023): 105606.
[9] Le, Xiaohuai, et al. "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement." arXiv preprint arXiv:2107.05429 (2021).
[10] Marcu, David C., and Cristian Grava. "The impact of activation functions on training and performance of a deep neural network." 2021 16th International Conference on Engineering of Modern Electric Systems (EMES). IEEE, 2021.
[11] Ye, Zhongfu, Nasir Saleem, and Hamza Ali. "Efficient gated convolutional recurrent neural networks for real-time speech enhancement." (2023).
[12] Ketkar, Nikhil, et al. "Convolutional neural networks." Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch (2021): 197-242.
[13] Hewamalage, Hansika, Christoph Bergmeir, and Kasun Bandara. "Recurrent neural networks for time series forecasting: Current status and future directions." International Journal of Forecasting 37.1 (2021): 388-427.
[14] Wahab, Fazal E., et al. "Compact deep neural networks for real-time speech enhancement on resource-limited devices." Speech Communication 156 (2024): 103008.
[15] Galvez, Daniel, et al. "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage." arXiv preprint arXiv:2111.09344 (2021).

Hamza Ali was born in Peshawar City, KPK, Pakistan in 1998. He received B.Sc. degree in Electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2021. He is currently working toward the M.Sc. degree with University of Engineering and Technology, Mardan. His current research interests are broad areas of mobile networking, signal processing, and Machine learning.

MUHAMMAD ISMAIL holds a Bachelor's degree in Electrical Engineering from the University of Engineering and Technology, Peshawar (2021), and a Master's degree from the University of Engineering and Technology, Mardan (2023). He received the "Ihsaas Undergraduate Scholarship" and the "Stori da Pakhtunkhwa," awarded by the HEC Pakistan. His research interests encompass vehicular ad hoc networks, intelligent transportation systems, and machine learning (ML).