Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement

Fazal-E-Wahab[1], Zhongfu Ye[1], Nasir Saleem[2], Hamza Ali[3]

[1] National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026 Anhui (China)
[2] Department of Electrical Engineering, Faculty of Engineering & Technology, Gomal University, D.I. Khan, KPK (Pakistan)
[3] Department of Electrical Engineering, University of Engineering & Technology, Mardan, KPK (Pakistan)

Received 5 March 2022 | Accepted 27 December 2022 | Early Access 31 May 2023

Abstract

Deep learning (DL) networks have grown into powerful alternatives for speech enhancement and have achieved excellent results by improving speech quality, intelligibility, and background noise suppression. Due to high computational load, most of the DL models for speech enhancement are difficult to implement for real-time processing. It is challenging to formulate resource efficient and compact networks. In order to address this problem, we propose a resource efficient convolutional recurrent network to learn the complex ratio mask for real-time speech enhancement. Convolutional encoder-decoder and gated recurrent units (GRUs) are integrated into the Convolutional recurrent network architecture, thereby formulating a causal system appropriate for real-time speech processing. Parallel GRU grouping and efficient skipped connection techniques are engaged to achieve a compact network. In the proposed network, the causal encoder-decoder is composed of five convolutional (Conv2D) and deconvolutional (Deconv2D) layers. Leaky linear rectified unit (ReLU) is applied to all layers apart from the output layer where softplus activation to confine the network output to positive is utilized. Furthermore, batch normalization is adopted after every convolution (or deconvolution) and prior to activation. In the proposed network, different noise types and speakers can be used in training and testing. With the LibriSpeech dataset, the experiments show that the proposed real-time approach leads to improved objective perceptual quality and intelligibility with much fewer trainable parameters than existing LSTM and GRU models. The proposed model obtained an average of 83.53% STOI scores and 2.52 PESQ scores, respectively. The quality and intelligibility are improved by 31.61% and 17.18% respectively over noisy speech.

Keywords

Convolutional Recurrent Networks, Deep Learning, GRU, Intelligibility, LSTM, Speech Enhancement.

DOI: 10.9781/ijimai.2023.05.007

I. Introduction

SPEECH enhancement (SE) aims to suppress background noise signals from the target speech, which include non-speech noise, competing speech, and room reverberations [1]. SE is used as a front-end in various real-world applications such as robust ASR systems and mobile phone communications where real-time processing is required. In such applications, SE is required to perform with little computational complexity and provide near-instantaneous outputs. The aim of this study is to focus on single-microphone speech enhancement, operating in real-time systems. For listeners using digital hearing aids, a delay of 3 milliseconds is perceptible, whereas delays longer than 10 msec are intolerable [2]. Speech enhancement techniques have made significant progress during the last several decades. Speech enhancement techniques may be divided into two categories depending on the quantity of microphones used, that is, single-channel based and multi-channel. The high-availability and low-cost single-channel approaches nevertheless have significant research significance, even if the extra spatial information of the microphone array may assist in reducing the direction-related noise interference. As a result, the goal of this work is to concentrate on real-time, single-microphone speech enhancement. Delays of 3 milliseconds or less are noticeable to listeners, whereas those of 10 milliseconds or more are unpleasant [2]. In these situations, a causal SE system is often necessary to prevent delays. A Causal SE system is often a requisite in such applications to avoid delays.

In recent years, SE has been formulated as a supervised learning problem where a deep neural network (DNN) learns a mapping function from the noisy features to a time-frequency mask [3]. The ideal binary mask, which categorises time-frequency units into the speech-dominant and noise-dominant, was the first training-target used in supervised speech enhancement [3]. More recent training-targets include ideal ratio mask (IRM) [4]-[5] complex ratio mask [6], mapping-based targets related to the magnitude or power spectra of the target speech [7]-[8], ideal amplitude mask [9]. For supervised SE, both noise and speaker generalizations are vital. An easy but useful approach to cope with noise generalization is to train a network with

* Corresponding author.
E-mail address: fwmarwat@mail.ustc.edu.cn

Please cite this article in press as:
F. E. Wahab, 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), http://dx.doi.org/10.9781/ijimai.2023.05.007

large noise types [10]. Likewise, a large number of speakers can be used in a training set to deal with speaker generalization. However, in the presence of several training speakers, a feedforward DNN is inept at tracking a target-speaker [10]-[11].

The training-targets in the time-frequency domain are mainly divided into two classes: the masking-based and mapping-based targets. The masking-based targets describe a time-frequency relationship between the clean speech and background noise signals, whereas the mapping-based targets correspond to the spectral representations of clean speech. In the masking-based class, ideal binary mask (IBM) [12], ideal ratio mask (IRM) [4] and spectral magnitude mask (SMM) [4] only use the magnitude between clean speech and mixture speech, overlooking the phase spectrum. Alternatively, the phase-sensitive mask (PSM) [13] incorporated the phase information and showed the importance of phase spectrum estimation. Afterward, complex ratio mask (cRM) [6] can be used to recover speech efficiently by improving both real and imaginary parts of the clean and noisy speech spectrograms simultaneously. Recently, [14] proposed a convolutional recurrent network using one encoder and two decoders to estimate the real and imaginary spectrograms (complex spectral mapping) of the noisy speech concurrently. It is important that the complex ratio mask and complex spectral mapping obtain the complete information of a speech signal to accomplish the best SE performance. The convolutional encoder-decoder (CED) and GRU produce a convolutional recurrent neural network (CRN), which is used to develop the SE in this article. CED is a strong tool for extracting temporal and spatial patterns from raw data. A causal system that is suitable for real-time speech processing is created by integrating a convolutional encoder-decoder and GRUs into the convolutional recurrent network architecture. In comparison to typical RNNs, GRU has the capacity to learn long-term temporal dependencies in speech signals with a far smaller number of trainable parameters. The contributions of this study are summarized as:

- A causal SE system which is appropriate for real-time speech processing is created by integrating a convolutional encode-decoder and GRUs into the convolutional recurrent network architecture.
- For an appropriate shape of the inputs required by GRUs, the proposed model has grouped the fully connected recurrent neural networks into disconnected parallel recurrent neural networks, where the forward information flow remains the same.
- By adding the skipped connections, to avoid gradient decay, which connect the output of the encoder to input of decoder output doubles the feature Maps, results in increasing the model complexity. Therefore, in the proposed model, add-skipped connection between conv-deconv layers having (1×1) kernel size is proposed which improves network performance at negligible complexity.

The remaining of the paper is organized as follows. Related studies on the research are presented in Section II. Description of the proposed model is given in Section III. The experimental setup and results are given in Section IV. Finally, the paper is concluded in Section V.

II. Related Studies

Generally, a DNN individually predicts labels for all time frames using small context windows and cannot control long-term context windows that are essential for target speaker tracking. Recently, studies [11], [15] suggest that it is better to formulate SE as a sequence-to-sequence process to control the long-term context windows. Recurrent Neural Networks [16] and Convolutional Neural Networks have been employed with such a formulation where training and testing with different noise types and speakers can be carried out. A four-layer LSTM model for speaker generalization is proposed in [15]. The results showed that the LSTM model generalized better to untrained speakers and considerably outperformed a DNN-based model in terms of speech intelligibility. Recently, a dilated convolution-based gated residual network is developed in [17]-[18] which demonstrated better generalization potential for untrained speakers at various SNRs when compared to the LSTM by [10]. However, the gated residual network requires future information for spectral masking or spectral mapping. Thus, it is not suitable for real-time SE. Motivated by recent studies [19]-[20] on convolutional recurrent networks; we designed a compact and efficient architecture for real-time speech enhancement. The first convolutional encode-decoder architecture has been introduced for SE by [21]. A redundant convolutional encode-decoder [22] was proposed, based on the convolution repetitions, batch normalization, and a ReLU activation layer. Moreover, to facilitate network optimization, skip connections are used. In this study, a skips-based convolutional encoder-decoder and the parallel GRUs are integrated into a convolutional recurrent architecture to estimate the complex ratio mask (cRM). We observed that the proposed architecture provided improved speech quality and intelligibility as compared to the GRU and LSTM with fewer trainable parameters. A deep residual GRU-based model to enhance noisy speech was proposed [23] which performed better as compared to SOTA for speech enhancement and recognition tasks. The study in [24] presented a joint structure to solve single-channel speech enhancement in the complex-domain. The RBM in [25] is extended for spectral masking and noisy speech enhancement. The acoustic features in traditional RBMs are extracted layerwise, where feature compression results in a loss of information during training. To address this problem of retaining information in raw speech, RBMs are extended for acoustic feature representation and speech enhancement. Acoustic features and regularized sparse features are combined to train DNNs for better speech enhancement [26]. Using short context windows, FNN model [27] independently predicts labels for all time frames. The CNN [28]-[29] may learn local features involved in the training data, in contrast to the FNN [8]-[9] which can fully use the previous knowledge of speech. The long-term contexts of speech signals cannot be leveraged by either the FNN or the CNN model, however. In order to regulate the long-term context windows, it has recently been recommended by the research [30]-[31], that it is preferable to design SE as a sequence-to-sequence process. The RNN model [32] can deal with long-term contexts in a sequence-based way, but often needs very complex hand-crafted features like MFCC. Convolutional recurrent networks (CRN) were first used for speech improvement by combining the CNN and RNN [31], [33]. Convolutional and recurrent neural networks [34] have been used in a formulation that enables training and testing with a variety of speakers and noise sources. Due to high computational load, most of the DL models for speech enhancement are difficult to implement for real-time processing. It is challenging to formulate resource efficient and compact networks.

III. Proposed System Description

The convolutional encoder-decoder (CED) and GRU produce a convolutional recurrent neural network (CRN), which is used to develop the SE in this article. CED is a strong tool for extracting temporal and spatial patterns from raw data. A causal system that is suitable for real-time speech processing is created by integrating a convolutional encoder-decoder and GRUs into the convolutional recurrent network architecture. In comparison to typical RNNs, GRU has the capacity to learn long-term temporal dependencies in speech signals with a far smaller number of trainable parameters. CED and GRU are explained in the following subsections.

A. Causal Convolution-Based Encoder-Decoder

The encoder in the causal convolution-based encoder-decoder framework is made up of stacked convolutional and pooling layers that extract high-level features from raw input data. Fundamentally similar structure as the encoder but in the reverse order, the decoder maps low-level features at the encoder output to full input feature size. This symmetric structure of the encoder-decoder ensures the shape of inputs and outputs. We imposed causal convolutions on the encoder-decoder framework to design a real-time SE system. Fig. 1. illustrates the causal convolutions with time-dimension. We treat the inputs as the sequence of the feature vectors, whereas the outputs are independent of the future sequence of the feature vectors. With such causal convolutions, the architecture leads to a causal encoder-decoder framework. The causal deconvolution can easily be applied to the decoder.

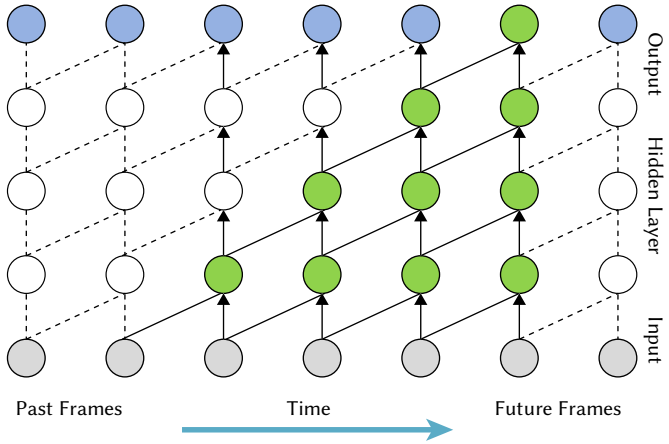

Fig. 1. An example of causal convolutions. The convolution output does not depend on future inputs.

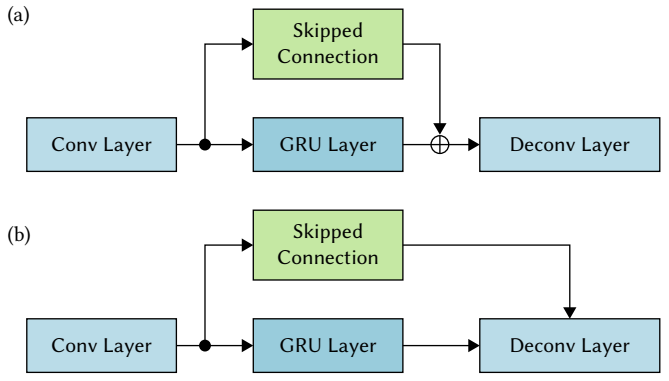

Fig. 2. Skip connections (a) add-skipped connections, (b) doubling the decoder inputs.

In the proposed network, the causal encoder-decoder is composed of five convolutional (Con2D) and deconvolutional (Decon2D) layers. Leaky linear rectified unit (ReLU) [35]-[36] is applied to all layers apart from the output layer where softplus activation (which can confine the network output to always be positive) [37] is utilized. Leaky ReLU has shown fast convergence and better generalization. Furthermore, batch normalization is adopted after every convolution (or deconvolution) and prior to activation. The kernel number is increased steadily in the encoder, whereas it is decreased steadily in the decoder, such that symmetric kernel numbers are adopted. So as to leverage large contexts, a stride of 2 is used along the frequency direction for all the convolutional (or deconvolutional) layers, whereas the time dimension of the features remains the same. To get a better flow of gradients and information all through the network, skip connections are utilized which connect the encoder outputs to the decoder inputs, as depicted in Fig. 2(a). In a recent study [17], the skip connections have been adopted by connecting the output of the encoder to the input of the decoder, as depicted in Fig. 2(b), which doubles the number of input channels to the decoder, resulting in increasing the complexity.

B. Temporal Modeling Using Parallel GRUs

Leveraging the long context is important to track a target speaker. The GRU [26] is the newer type of recurrent neural network that includes memory cells and is successful in temporal modeling. To integrate the temporal dynamics of the speech signals, we inserted a parallel GRU layer between the convolutional-encoder and the convolutional-decoder. Equations (1)-(4) describe the GRU network.

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

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

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

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

Where $z_t, r_t, \tilde{h}_l, h_l$ represent update gate, reset gate, intermediate memory, and output respectively whereas W_z, W_r, W_h, b_z, b_r and b_h are the model parameters that are learned during training. For a suitable shape of the inputs required by GRUs, the approach proposed by [17] has been adopted to group the wide fully connected recurrent neural networks into P disengaged parallel recurrent neural networks. But, noted that the forward information flow remains the same. The parallel GRUs are denoted by P, where P = 1 indicates that the last convolutional encoder output is flattened to a single vector and fed to a single GRU, whereas P > 1 indicates that the encoder output is reshaped to P vectors of the same length, fed through P disconnected GRUs, and reshaped again to the number of decoder channels. Another practical advantage is the possible parallel execution of the disconnected RNNs. It is important to note that the insertion of the GRUs does not impact the system's causality.

C. Network Architecture

In this paper, we used 161-dimensional STFT magnitude spectrum of noisy speech as the input features and a complex ratio mask as the training target. The proposed convolutional recurrent network is illustrated in Fig. 3, where inputs to the network are encoded into a high-dimension latent space, and the sequences of latent features are subsequently modelled by the GRU layer. Next, the output sequences of the GRU layer are transformed back by the decoder into their original input shape. The proposed convolutional recurrent network uses CNNs for feature extraction and RNNs for temporal modeling, thus combining two powerful topologies with improved results. A representation of the architecture is given in Table 1. The input and output layers' sizes are specified as FeatureMaps (FM), TimeSteps (TS), and FrequencyChannels (FCh), respectively, while the hyperparameters along the layer are specified as KernelSize (KZ), Strides (S), and OutChannels (OCh). In all the convolution and the deconvolution layers, a zero-padding in the time direction is applied, but no padding is involved in the frequency direction. For causal convolutions, a (2×3) kernel size is used, where (2×3) indicates (time×frequency). Note that by adding the skipped connections, which connect the output of the encoder to the input of the decoder output, doubles the feature maps, thus increasing the network complexity. By adding an add-skipped connection between the conv-deconv with (1×1) kernel size, it improves network performance at negligible added complexity, as shown in Fig. 4. We denoted the proposed network as CGCRN.

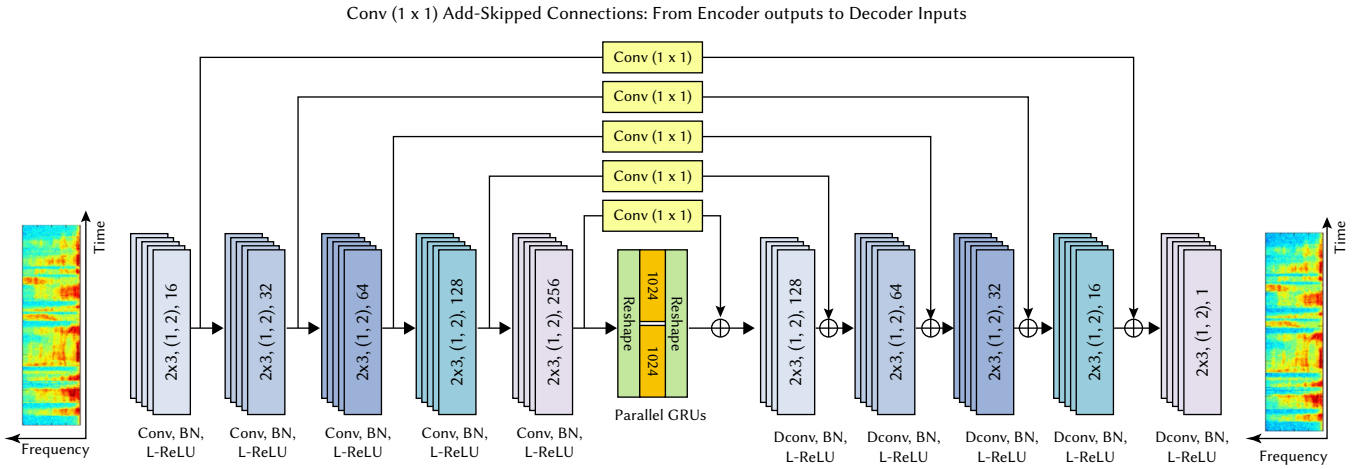

Fig. 3. Network architecture of our proposed CGCRN.

TABLE I. Network Architecture, Where *T* Denotes the Time Frames in the STFT Magnitude Spectrum, Here P = 2 in P-GRU Layer, Epochs= 100, and Learning Rate Is 0.0001

Layer	Input Size	Hyperparameters	Output Size
	FM × TS × FCh	KZ, S, OCh	FM × TS × FCh
Reshape-1	T × 161	---	1 × T × 161
Conv-1	1 × T × 161	2 × 3, (1, 2), 16	16 × T × 80
Conv-2	16 × T × 80	2 × 3, (1, 2), 32	32 × T × 39
Conv-3	32 × T × 39	2 × 3, (1, 2), 64	64 × T × 19
Conv-4	64 × T × 19	2 × 3, (1, 2), 128	128 × T × 9
Conv-5	128 × T × 9	2 × 3, (1, 2), 256	256 × T × 4
Reshape-2	256 × T × 4	---	T × 2048
P-GRU	T × 2048	2048	T × 1024
Reshape-3	T × 1024	---	256 × T × 4
Deconv-1	512 × T × 4	2 × 3, (1, 2), 128	128 × T × 9
Deconv-2	256 × T × 9	2 × 3, (1, 2), 64	64 × T × 19
Deconv-3	128 × T × 19	2 × 3, (1, 2), 32	32 × T × 39
Deconv-4	64 × T × 39	2 × 3, (1, 2), 16	16 × T × 80
Deconv-5	32 × T × 80	2 × 3, (1, 2), 1	1 × T × 4
Reshape-4	1 × T × 161	---	T × 161

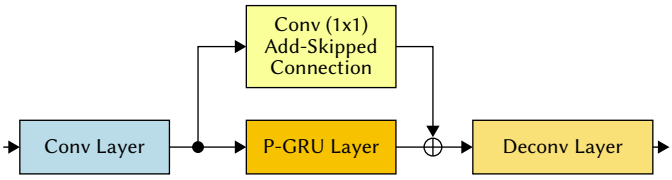

Fig. 4. (1 × 1) convolutions in the add-skipped connections.

D. LSTM and GRU Baselines

In the experiments, causal LSTM and GRU baselines were selected for comparison purposes. In the causal LSTM and GRU models, a context feature window of 11 frames, composed of 10 past speech frames and 1 current speech frame is used to estimate one frame of the target speech. A concatenated long vector of 11 frames of feature vectors is used as input to the network at all-time steps. We used the same network architectures for LSTM and GRU [11 161, 1024, 1024, 1024, 1024, 1024, 1024, 1024] units from the input to the output layer. No future information is used by baselines, which makes them causal speech enhancement systems.

IV. Experimental Setup

In the experiments, we evaluated SE networks on the LibriSpeech dataset [38] (derived from the read audiobooks, LibriVox project) including 0.25 Million utterances from 2.1k speakers. We have used the LibriClean version of LibriSpeech which contains 104014 clean utterances (about 360 hours) belonging to 921 different speakers. But to evaluate the networks used in this study, we randomly selected 5000 speech utterances from 40 speakers. Among these speakers, 2 male and 2 female speakers are used as untrained speakers, whereas the remaining 36 speakers are used to train the networks. In order to train noise-independent networks, we have used 60 noise types from the Perception and Neurodynamics Laboratory (http://web.cse.ohio-state.edu/pnl/data.html) and Laboratory for Recognition and Organization of Speech and Audio (https://www.ee.columbia.edu/~dpwe/sounds/) for network training. For testing purpose, we used three challenging noise types (multi-talker babble, street, and cafeteria). We created a training set by randomly selecting utterances with an indiscriminate cut from the 60 training noise types at SNRs selected from -5dB, -3dB, -1dB, 0dB, and 2dB. During testing, we used three SNRs for the test set, that is, -5dB, -2dB, and 2dB. In order to examine the speaker generalization, the models are tested with two test sets for all noise types (i.e., multi-talker babble, street, and cafeteria) using trained and untrained speakers, respectively. First test set is composed of 120 mixtures created from 30 × 4 utterances of 5 trained speakers, whereas the second test set is composed of 120 mixtures created from 30 × 4

TABLE II. Networks Comparison in Three Test-Noises in Terms of the STOI (in %)

Noise Types	Babble				Street				Cafeteria			
Input SNR	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg
Noisy Speech	58.95	66.30	75.55	66.93	58.30	66.20	75.08	66.52	57.40	65.19	74.21	65.60
LSTM	77.29	82.62	84.96	81.62	75.21	82.62	84.11	80.64	74.32	81.38	83.07	79.59
GRU	77.45	83.21	85.01	81.89	75.30	82.05	85.01	80.78	74.14	81.25	83.22	79.53
CRN	79.71	85.48	86.88	84.02	77.12	84.44	87.23	82.93	76.07	82.68	85.10	81.28
CNN-GRU	78.11	84.31	85.82	82.74	76.21	83.01	85.66	81.62	75.04	82.31	84.14	80.49
FCNN	70.22	75.21	80.34	75.26	71.44	75.57	81.02	76.00	70.34	76.70	80.87	75.97
CGCRN	**80.47**	**86.29**	**87.74**	**84.83**	**77.86**	**85.23**	**88.07**	**83.72**	**76.80**	**83.43**	**85.92**	**82.05**

TABLE III. Networks Comparison in Three Test-Noises in Terms of the PESQ

Noise Types	Babble				Street				Cafeteria			
Input SNR	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg	-5 dB	-2 dB	0 dB	Avg
Noisy Speech	1.63	1.79	1.86	1.76	1.58	1.71	1.84	1.71	1.52	1.70	1.82	1.68
LSTM	2.06	2.36	2.53	2.32	2.03	2.31	2.48	2.27	2.04	2.30	2.47	2.27
GRU	2.07	2.36	2.54	2.32	2.05	2.27	2.47	2.26	2.03	2.31	2.48	2.27
CRN	2.17	2.44	2.62	2.41	2.14	2.40	2.60	2.38	2.12	2.38	2.59	2.36
CNN-GRU	1.95	2.26	2.53	2.25	1.98	2.31	2.55	2.28	1.94	2.25	2.50	2.23
FCNN	1.81	2.15	2.44	2.13	1.88	2.21	2.51	2.20	1.85	2.20	2.48	2.18
CGCRN	**2.29**	**2.59**	**2.79**	**2.56**	**2.25**	**2.53**	**2.76**	**2.52**	**2.21**	**2.49**	**2.72**	**2.47**

utterances of 4 untrained speakers. Speech utterances and noise types are sampled at 16 kHz. The networks are optimized with the Adam optimizer [39]-[40]. We fixed the learning rate to 10-4 and the mean squared error (MSE) served as a loss function. The networks are trained with minibatch size of 16 and the number of epochs is fixed to 80. Inside all minibatches, the training samples are zero padded such that to contain the equal number of time steps.

The experiments use two widely used objective metrics to quantify the proposed speech enhancement, including the STOI (Short-Time Objective Intelligibility), the PESQ (Perceptual Evaluation of Speech Quality). Intelligibility and quality of the enhanced speech signals are determined by STOI and PESQ, respectively. PESQ [41], an ITU-T P.862 recommendation, scores the perceptual speech quality from -0.5 to 4.5. Similarly, STOI [42] assesses speech intelligibility with output values ranging from 0 to 100.

V. Results and Discussions

Two performance metrics are used in the experiments. Perceptual evaluation of speech quality (PESQ) [41] measures the speech quality whereas the short-term objective speech intelligibility (STOI) [42] measures the speech intelligibility, respectively. A high value for both the metrics indicates a better performance. We also included the Convolutional recurrent network (CRN) proposed by [17], CNN-GRU [43], and fully connected CNN (FCNN) [44] as SOTA for comparison. The proposed CGCRN network is the extension of CRN. Table II-III presents STOI and PESQ test scores of noisy speech and speech processed by different networks across all the noise types and input SNRs. The best performance is highlighted with boldface numbers. As indicated by Table II-III, the LSTM and GRU networks yielded almost similar STOI and PESQ scores which suggests that the noisy speech can effectively be enhanced by using GRU networks with less trainable parameters. The results indicated that replacing the LSTM layers by a parallel GRU layer in the CRN significantly improved the performance with fewer trainable parameters and network complexity. The CRN outperformed both the LSTM and GRU networks. On the other hand, the proposed CGCRN consistently outperformed the LSTM, GRU, and CRN in both the metrics. For example, the average STOI test scores are improved from 66.93% to 84.83% with CGCRN (ΔSTOI = 17.90%) in babble noise type. Here Δ indicates the improvements in metrics. Also, the average STOI test scores are improved from 65.60% to 82.05% with CGCRN (ΔSTOI = 16.45%) in cafeteria noise type. On average, 3.09% STOI gain is achieved when the CGCRN when compared to the LSTM network. Moreover, 0.79% improvement in STOI test scores is achieved against the CRN. In addition, the average PESQ test scores are improved from 1.76 to 2.56 with the proposed CGCRN (ΔPESQ = 0.80 equivalent to 31.25%) in babble noise type. Similarly, the average PESQ test scores are improved from 1.71 to 2.52 with the proposed CGCRN (ΔPESQ = 0.81 equivalent to 32.14%) in street noise type.

Table IV-V presents the speaker generalization potential of the neural networks used in this study. It can be observed from Tables that the CGCRN and CRN showed better generalization to untrained speakers. In the most challenging noisy cases, where the utterances from the untrained speakers are mixed with three untrained noise types at -5dB and -2dB, the proposed CGCRN improved the average STOI by 15.72% and the PESQ by 0.70 (28.29%) over the noisy speech. The CGCRN improved the PESQ by 9.16 % over the GRU in trained speakers whereas by 9.91% in untrained speakers, respectively. Thereby indicates that the proposed models can success-fully be implemented in untrained situations.

The batch normalization in convolution operations accelerated the training and improved the performance. We observed a faster convergence and less MSEs with the CGCRN as compared the LSTM and GRU networks. Importantly the fewer trainable parameters are the key significance of the CGCRN, as illustrated in Fig. 5. In addition, the causal convolution operations captured the local patterns in the magnitude spectra exclusive of the future in-formation. In contrast, the GRU and LSTM networks deal all input frames as flattened feature vectors, thereby lack ample control over the time-frequency structures in the magnitude spectra. The parallel GRUs layer models the long-term temporal dependencies in a compressed space which is vital to speaker classification in the speaker-independent SE. As shown in experiments, replacing the LSTM layers in CRN with a single parallel GRUs layer yielded a considerable performance gain and enormous computational savings. A single GRU layer reduces 25% of trainable parameters (network complexity) as com-pared to single LSTM layer.

TABLE IV. Speaker Generalization of the Networks in Terms of STOI (in %)

Speaker Types	Trained Speakers				Untrained Speakers			
Noise Type	Babble	Street	Cafeteria	Avg	Babble	Street	Cafeteria	Avg
LSTM	81.62	80.64	79.59	80.61	79.32	78.22	77.01	78.18
GRU	81.89	80.78	79.53	80.73	79.57	78.41	77.19	78.39
CGCRN	**84.83**	**83.72**	**82.05**	**83.53**	**83.66**	**82.33**	**80.23**	**82.07**

TABLE V. Speaker Generalization of the Networks in Terms of PESQ

Speaker Types	Trained Speakers				Untrained Speakers			
Noise Types	Babble	Street	Cafeteria	Avg	Babble	Street	Cafeteria	Avg
LSTM	2.32	2.27	2.27	2.29	2.22	2.19	2.18	2.20
GRU	2.32	2.26	2.27	2.28	2.21	2.17	2.17	2.18
CGCRN	**2.56**	**2.51**	**2.47**	**2.51**	**2.45**	**2.43**	**2.38**	**2.42**

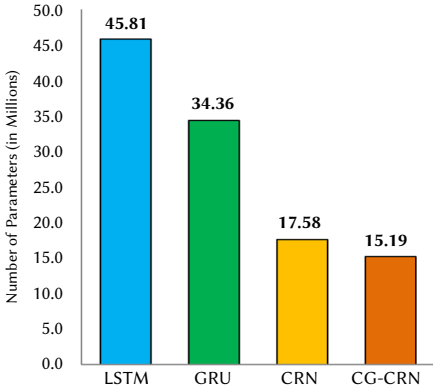

Fig. 5. Parameter efficiency comparison of different models. We compare the number of trainable parameters in different models.

Table VI shows impact of add-skipped connections in the CGCRN architecture. Adding the skipped connections is superior to no skipped connections. Although add-skips improved the PESQ and STOI test scores, but a better performance is achieved by inserting Conv (1 × 1) add-skipped connections. In order to visualize the spectrotemporal characteristics, the spectrograms are presented in Fig. 6. which belongs to the clean speech, noisy speech, and speech processed by the LSTM, GRU, and CGCRN with the cRM as the training-target. It is evident that few speech parts are missing in the spectrograms (highlighted with boxes) of speech enhanced by LSTM and GRU. In contrast, the speech enhanced by CGCRN demonstrates comparable spectrotemporal patterns to the clean speech and less distortion can also be noticed.

TABLE VI. Effects of Skipped Connections

Skip Types	STOI	PESQ
No Skips	79.21	2.34
Add Skips	81.33	2.40
Conv Skips	**83.45**	**2.49**

The proposed speech enhancement CGCRN performed better at all input SNRs in terms of speech intelligibility and quality. However, to confirm the success at SNRs, one-way analysis-of-variance (ANOVA) statistical analyses are conducted at -5dB, -2dB and 0dB. The statistical tests are performed at 95% confidence interval. Differences between test results are believed statistically important if the probability (Pvalue) is less than 0.05 (P<0.05) and Fvalue is higher than the critical value of FDistribution (Fvalue>FCritical). Table VII shows the statistical tests in terms of speech intelligibility at 95% confidence interval with FCritical is 3.09. It is clear that Pvalues of the proposed model are less than 0.05 and the values of FCritical are higher than 3.09, which indicates that the intelligibility results of the proposed model are statistically significant. Similarly, Table VIII shows the statistical tests in terms of speech intelligibility at 95% confidence interval with FCritical is 3.09. For illustration at adverse noise levels (-5dB), the CGCRN against the noisy speech (CGCRN ⟶ Noisy), we achieved [F (2, 100) = 39.5, p < 0.001] for STOI and [F (2, 100 = 32.2), p < 0.001] for PESQ, respectively. Also, against the CRN (CGCRN ⟶ CRN), we achieved [F (2, 100) = 21.1, p < 0.001] for STOI and [F (2, 100) = 24.1, p < 0.001] for PESQ. Moreover, against LSTM (CGCRN ⟶ LSTM), we achieved [F (2, 100) = 22.2, p < 0.001] for STOI and [F (2, 100) = 18.3, p < 0.015] for PESQ. The ANOVA results at low SNRs indicate that the proposed model achieved better results statistically over the competing deep learning models.

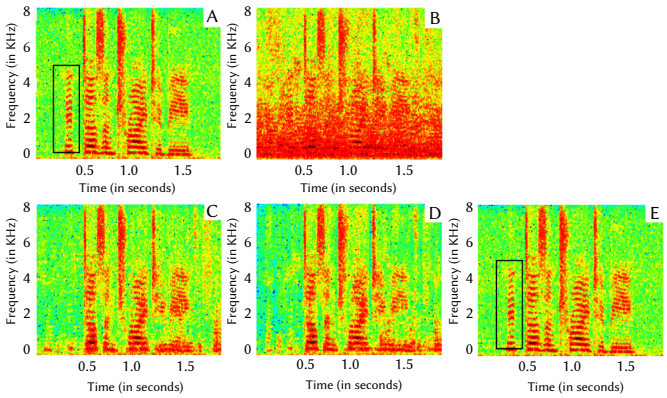

Fig. 6. Spectrotemporal characteristics of the speech processed by different networks.

VI. Summary and Conclusions

In this paper we propose resource efficient Convolutional recurrent network to learn the complex ratio mask for real-time speech enhancement. Convolutional encode-decoder and gated recurrent unit are integrated into the Convolutional recurrent network architecture thereby formulated a causal system, which is suitable for the real-time speech processing. Parallel GRU grouping and efficient skipped connections techniques are used to achieve compact network. Different noise types and speakers are used in training and testing to observe the speaker and noise generalization. With LibriSpeech dataset, the experiments showed that the proposed real-time approach led to improve perceptual speech quality and intelligibility with much fewer trainable parameters than existing LSTM and GRU models. The quality and intelligibility are improved by 31.61% and 17.18% over noisy speech. CGCRN proves comparable spectrotemporal patterns to the clean speech and less distortion can also be noticed. We showed gains on the speech quality and intelligibility with less computational complexity by more effective skip connections and a parallel GRUs

TABLE VII. Statistical Analysis of Average Intelligibility at 95% Confidence Interval With $F_{Critical}$ Is 3.09 and $P_{Critical}$ Is 0.05

ANOVA SE Models	STOI					
	-5dB		-2dB		0dB	
	P_{Value}	F_{Value}	P_{Value}	F_{Value}	P_{Value}	F_{Value}
CGCRN ⟶ Noisy	<0.001	39.5	<0.001	21.4	<0.001	19.1
CGCRN ⟶ LSTM	<0.001	22.2	<0.002	31.2	<0.001	26.1
CGCRN ⟶ GRU	<0.002	29.4	<0.005	28.6	<0.001	25.3
CGCRN ⟶ CRN	<0.001	21.1	<0.001	19.1	<0.002	18.3

TABLE VIII. Statistical Analysis of Average Quality at 95% Confidence Interval With $F_{Critica}$ Is 3.09 and $P_{Critical}$ Is 0.05

ANOVA SE Models	PESQ					
	-5dB		-2dB		0dB	
	P_{Value}	F_{Value}	P_{Value}	F_{Value}	P_{Value}	F_{Value}
CGCRN ⟶ Noisy	<0.001	32.2	<0.001	37.4	<0.001	22.2
CGCRN ⟶ LSTM	<0.015	18.3	<0.020	28.3	<0.020	27.3
CGCRN ⟶ GRU	<0.001	30.2	<0.001	24.5	<0.001	23.2
CGCRN ⟶ CRN	<0.001	24.1	<0.001	22.1	<0.001	20.5

structure. The proposed model used fewer parameters and causal operations; therefore, suitable for real-time speech enhancement. The ANOVA statistical analysis confirmed that the intelligibility and quality results are statistically significant. The average STOI test scores are improved from 66.93% to 84.83% with CGCRN (ΔSTOI = 17.90%) in babble noise type. Here Δ indicates the improvements in metrics. Also, the average STOI test scores are improved from 65.60% to 82.05% with CGCRN (ΔSTOI = 16.45%) in cafeteria noise type. On average, 3.09% STOI gain is achieved when the CGCRN when compared to the LSTM network. Moreover, 0.79% improvement in STOI test scores is achieved against the CRN. In addition, the average PESQ test scores are improved from 1.76 to 2.56 with the proposed CGCRN (ΔPESQ = 0.80 equivalent to 31.25%) in babble noise type. at adverse noise levels (-5dB), the CGCRN against the noisy speech (CGCRN ⟶ Noisy), we achieved [F (2, 100) = 39.5, p < 0.001] for STOI and [F (2, 100 = 32.2), p < 0.001] for PESQ, respectively. Also, against the CRN (CGCRN ⟶ CRN), we achieved [F (2, 100) = 21.1, p < 0.001] for STOI and [F (2, 100) = 24.1, p < 0.001] for PESQ. Moreover, against LSTM (CGCRN ⟶ LSTM), we achieved [F (2, 100) = 22.2, p < 0.001] for STOI and [F (2, 100) = 18.3, p < 0.015] for PESQ.

Speech perception quality also depends on the phase. However, since phase lacks spectrotemporal structure, it seems to be impossible to correctly estimate phase spectra using masking-based supervised learning, like in this proposed model. The complex spectral mapping, which concurrently improves the magnitude and phase responses of noisy speech, tries to estimate the real and imaginary spectrograms of clean speech from those of noisy speech. In future work, we would be devoted to proposing more flexible, scalable, and phase included CRNs for real-time speech enhancement, trained on large datasets and tested on the real recordings.

References

[1] Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702-1726.

[2] Agnew, J., & Thornton, J. M. (2000). Just noticeable and objectionable group delays in digital hearing aids. *Journal of the American Academy of Audiology*, *11*(06), 330-336.

[3] Wang, Y., & Wang, D. (2013). Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(7), 1381-1390.

[4] Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *22*(12), 1849-1858.

[5] Saleem, N., & Khattak, M. I. (2020). Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(1), 84-91.

[6] Williamson, D. S., Wang, Y., & Wang, D. (2015). Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *24*(3), 483-492.

[7] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7-19.

[8] Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2013). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, *21*(1), 65-68.

[9] Saleem, N., & Khattak, M. I. (2020). Multi-scale decomposition based supervised single channel deep speech enhancement. *Applied Soft Computing*, *95*, 106666.

[10] Chen, J., & Wang, D. (2017). Long short-term memory for speaker generalization in supervised speech separation. *The Journal of the Acoustical Society of America*, *141*(6), 4705-4714.

[11] Kolbæk, M., Tan, Z. H., & Jensen, J. (2016). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 153-167.

[12] Li, Y., & Wang, D. (2009). On the optimality of ideal binary time–frequency masks. *Speech Communication*, *51*(3), 230-239.

[13] Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015, April). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 708-712). IEEE.

[14] Tan, K., & Wang, D. (2019). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 380-390.

[15] Chen, J., Wang, Y., Yoho, S. E., Wang, D., & Healy, E. W. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *The Journal of the Acoustical Society of America*, *139*(5), 2604-2612.

[16] Saleem, N., Khattak, M. I., Al-Hasan, M. A., & Jan, A. (2021). Multi-objective long-short term memory recurrent neural networks for speech enhancement. *Journal of Ambient Intelligence and Humanized Computing*, *12*(10), 9037-9052.

[17] Tan, K., Chen, J., & Wang, D. (2018, April). Gated residual networks with dilated convolutions for supervised speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 21-25). IEEE.

[18] Tan, K., & Wang, D. (2018, September). A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech* (Vol. 2018, pp. 3229-3233).

[19] Zhang, Z., Sun, Z., Liu, J., Chen, J., Huo, Z., & Zhang, X. (2016). Deep recurrent convolutional neural network: Improving performance for

speech recognition. *arXiv preprint arXiv:1611.07174*.

[20] Naithani, G., Barker, T., Parascandolo, G., Bramsl, L., Pontoppidan, N. H., & Virtanen, T. (2017, October). Low latency sound source separation using convolutional recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 71-75). IEEE.

[21] Park, S. R., & Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

[22] Jha, S., Dey, A., Kumar, R., & Kumar-Solanki, V. (2019). A Novel Approach on Visual Question Answering by Parameter Prediction using Faster Region Based Convolutional Neural Network. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(5), 30-38.

[23] Saleem, N., Gao, J., Khattak, M. I., Rauf, H. T., Kadry, S., & Shafi, M. (2022). Deepresgru: residual gated recurrent neural network-augmented kalman filtering for speech enhancement and recognition. *Knowledge-Based Systems*, *238*, 107914.

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

[25] Khattak, M. I., Saleem, N., Nawaz, A., Almani, A. A., Umer, F., & Verdú, E. (2022). ERBM-SE: Extended Restricted Boltzmann Machine for Multi-Objective Single-Channel Speech Enhancement. *International Journal of Interactive Multimedia & Artificial Intelligence*, *7*(4).

[26] Khattak, M. I., Saleem, N., Gao, J., Verdu, E., & Fuente, J. P. (2022). Regularized sparse features for noisy speech enhancement using deep neural networks. *Computers and Electrical Engineering*, *100*, 107887.

[27] Gao, T., Du, J., Dai, L. R., & Lee, C. H. (2016, September). SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement. In *Interspeech* (pp. 3713-3717).

[28] Laishram, A., & Thongam, K. (2022). Automatic Classification of Oral Pathologies Using Orthopantomogram Radiography Images Based on Convolutional Neural Network. *International Journal of Interactive Multimedia & Artificial Intelligence*, *7*(4).

[29] Kounovsky, T., & Malek, J. (2017, May). Single channel speech enhancement using convolutional neural network. In *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)* (pp. 1-5). IEEE.

[30] Grais, E. M., & Plumbley, M. D. (2017, November). Single channel audio source separation using convolutional denoising autoencoders. In *2017 IEEE global conference on signal and information processing (GlobalSIP)* (pp. 1265-1269). IEEE.

[31] Zhao, H., Zarar, S., Tashev, I., & Lee, C. H. (2018, April). Convolutional-recurrent neural networks for speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2401-2405). IEEE.

[32] Gao, T., Du, J., Dai, L. R., & Lee, C. H. (2018, April). Densely connected progressive learning for lstm-based speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5054-5058). IEEE.

[33] Tan, K., & Wang, D. (2018, September). A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech* (Vol. 2018, pp. 3229-3233).

[34] Pirhosseinloo, S., & Brumberg, J. S. (2019, November). Dilated convolutional recurrent neural network for monaural speech enhancement. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers* (pp. 158-162). IEEE.

[35] Manju, N., Harish, B. S., & Nagadarshan, N. (2020). Multilayer Feedforward Neural Network for Internet traffic classification. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(1), 117-123.

[36] Dubey, A. K., & Jain, V. (2019). Comparative study of convolution neural network's relu and leaky-relu activation functions. In *Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018* (pp. 873-880). Springer Singapore.

[37] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[38] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206-5210). IEEE.

[39] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[40] Alvarez, A. A., & Gómez, F. (2021). Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks. *International Journal of Interactive Multimedia & Artificial Intelligence*, *6*(6).

[41] Beerends, J. G., Hekstra, A. P., Rix, A. W., & Hollier, M. P. (2002). Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model. *Journal of the Audio Engineering Society*, *50*(10), 765-778.

[42] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 4214-4217). IEEE.

[43] Hasannezhad, M., Ouyang, Z., Zhu, W. P., & Champagne, B. (2020, December). An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 764-768). IEEE.

[44] Ouyang, Z., Yu, H., Zhu, W. P., & Champagne, B. (2019, May). A fully convolutional neural network for complex spectrogram processing in speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5756-5760). IEEE.

Fazal-e-Wahab

Fazal-e-Wahab received a B.E degree in Electronics engineering from the Dawood University of Engineering and Technology, Karachi, Pakistan, in 2009, and the M.Sc. degree in electrical engineering from CECOS University, Peshawar, in 2015. Since 2012, he has been involved in teaching and research with Department of Electrical Engineering, UET Peshawar. Currently, he is pursuing a Ph.D. degree from the University of Science and Technology of China, Hefei, China in Signal and Information Processing. His current research interests are Speech Enhancement, Speech Denoising and Machine learning applications.

Zhongfu Ye

Zhongfu Ye received the BE and MS degrees from the Hefei University of Technology, Hefei, China, in 1982 and 1986, respectively, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 1995. He is currently a Professor of the University of Science and Technology of China. His current research interests are in statistical and array signal processing and image processing.

Nasir Saleem

Nasir Saleem received B.Sc. Telecom Engineering degree from University of Engineering and Technology, Peshawar, Pakistan in 2008, M.S. Electrical Engineering degree from CECOS University, Peshawar, Pakistan in 2012; and Ph.D. degree in Electrical Engineering, specialization in Digital speech processing and Deep Learning from University of Engineering and Technology, Peshawar, Pakistan in 2021. From 2008 to 2012, he was a lecturer at Institute of Engineering Technology, 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, Gomal University. Human-Machine Interaction, Speech Enhancement, and Machine Learning Applications are the areas he is currently researching.

Hamza Ali

Hamza Ali 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, Speech Enhancement and Machine learning.