Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19744-1

Cross channel interaction based ECA-Net using gated recurrent convolutional network for speech enhancement

Manaswini Burra[1] · Sunny Dayal Vanambathina[2] · Venkata Adi Lakshmi A[3] · Loukya Ch[3] · Siva Kotiah N[3]

Received: 14 October 2022 / Revised: 3 May 2024 / Accepted: 19 June 2024

Abstract
Recently channel attention mechanism playing a major role in improving the performance of deep convolution neural networks. Even though there is an improvement in the performance, but there is an increase in complexity of the model network. It is difficult for CNN alone to correctly model the long-range dependencies of speech signals. The local receptive field of the convolution limits the model's ability to capture long-range dependencies across input sequences. Moreover, LSTMs are easily prone to the problem of overfitting, and it also requires a long time to train. To overcome the above drawbacks in the proposed model, dilated dense blocks and GRUs are introduced. Based on multi-task learning, we propose a gated convolutional recurrent network with efficient channel attention (GCRN-ECA) for complex spectral mapping, which amounts to a causal system for monaural speech enhancement. Each layer in encoder and decoder consists of dense block. Complex spectral mapping seeks to predict the real and imaginary spectrograms of clear speech based on those of noisy speech, thereby enhancing both the magnitude and phase responses of the speech. The advantage of dilated convolutions in dense block is the receptive field increases with increasing dilation rates, which are used to capture long-range speech contexts. And the dense connectivity provides a feature map with more precise target information by passing through multiple layers. To represent the correlation between neighboring noisy speech frames, a two Layer GRU is added in the bottleneck, which has the advantage of increased training speed because of its simpler architecture. GRU captures the long-range dependencies across input sequences. The advantage of GRU is that it is easier to modify and doesn't require memory units, which means it can train faster than LSTM. The ECA module can implement cross-channel interaction without dimensionality reduction. An appropriate cross-channel interaction can preserve performance while significantly decreasing model complexity. Our results reveal that the proposed GCRN-ECA outperforms existing baselines in terms of quality and intelligibility. The proposed model increases the average PESQ and STOI scores by 35.9% and 18.1%, respectively, for the Common voice dataset, and by 35.2% and 4%, respectively, for the VCTK dataset compared to noisy speech.

Keywords Deep learning · Speech enhancement · Attention · Convolutional neural networks · Efficient channel attention

Extended author information available on the last page of the article

Published online: 28 June 2024

1 Introduction

In everyday listening environments, background noise distorts the speech signals. These kinds of change, degrade the both quality and intelligibility of the speech. This is challenging in applications like automatic speech recognition and speech intelligibility. This is also a challenging task in case of monaural speech enhancement especially at very low signal to noise ratio (SNR) In the speech processing community, many studies have been done on monaural speech enhancement. Some of the Single channel speech enhancement techniques are statistical based approaches [92]. Non-negative matrix factorization [40, 49, 81, 82, 105], Deep Neural Networks [27, 36, 38, 39].

The advancement in speech and vision processing systems has enabled tremendous research and development in the areas of human-computer interactions [70], biometric applications [89, 90], security and surveillance [79], and most recently in computational behavioral analysis [13, 16, 43, 76, 104], Audio-Visual Speech Enhancement [59, 60], Speech Separation [1, 62] Automatic Speech Recognition [5], Human Listening and Live Captioning [15], Accents Identification [57], automatic speaker age estimation [2]. Emotions can alter the acoustic properties of speech, such as pitch, intensity, and duration. For example, speech produced in a fearful state might be higher in pitch and intensity compared to speech produced in a neutral state. Emotions can also impact how speech is perceived. Listeners are often able to accurately infer the emotional state of a speaker based on the acoustic cues present in their speech. In [73] a feature vector by minimum number of elements is proposed for recognizing emotional states of speech. The low complexity spectral enhancement methods are very suitable for hearing aids users [52]. The spectral subtraction technique, initially introduced by Boll [4], uses the assumption of uncorrelated speech and noise to remove noise in speech. This approach was further enhanced by Berouti et al [3]. to minimize the artifacts caused by noise reduction. These methods can be generalized to enhance quality by appropriately adjusting the parameters [42]. In line with this concept, Sim et al. [77] proposed a methodfor optimal parameter selection based on minimum mean squared error. Additionally, Hu and Yu [29] suggested an adaptive noise estimation method to improve quality.

Deep learning has revolutionized speech processing by autonomously extracting meaningful features from raw speech signals, eliminating the need for manual feature engineering. This advancement has led to significant improvements in speech processing performance, particularly in challenging scenarios with noise, various accents, and dialects [58]. It is commonly acknowledged that transcribing noisy speech using automatic speech recognition (ASR) systems trained on clean data results in notably reduced recognition accuracy. This challenge is further exacerbated when working with child speakers. Children's speech features, such as pitch and formant frequency, vary significantly with age, presenting a significant obstacle to accurate recognition. In [75], the authors explored methods to enhance the noise robustness of ASR systems, focusing on children's speech. They also proposed the incorporation of a foreground speech segmentation and enhancement module to improve noise robustness. A method for enhancing Dysarthric speech, designed specifically for individuals with cerebral palsy aged 40-60, was introduced in [50]. This method employed Kepstrum analysis and was assessed using dysarthric speech samples. The evaluation encompassed monosyllabic and bisyllabic samples, which exhibit distinct Consonant-Vowel and Consonant-Vowel-Consonant-Vowel patterns. The outcomes indicated notable changes in formants and energy levels in the processed speech signal.

The SEGAN [69] is an end-to-end SE model where only strided convolutions are used in the generator and discriminator. In this model also only ordinary convolution operations are

used. Even though the performance of the model is good but it suffers from computational complexity. Wave-U-Net [56] is a time domain SE model with basic U-NET [41] architecture with 1D ordinary convolution layers in the encoder and decoder with a 1D convolution as a bottleneck. The CNN alone cannot well model the long-range dependencies of speech signal. In all the existing models only, ordinary convolutional layers are used. The local receptive field of the convolution limits the model's ability to capture long-range dependencies across input sequences. In CRN [86] model to further enhance the performance of U-NET [41] the LSTMs are used in between encoder and decoder ofU-NET [41] to learn long term dependencies of speech signals. Even though the performance of model is better the LSTMs are easily prone to the problem of overfitting and it also requires a large time to train. LSTM requires 4 linear layers (MLP layer) per cell to run at each time step. Linear layers require large amounts of memory bandwidth to be computed. Speech enhancement performance is influenced by CNN's limited receptive field, which restricts its ability to extract long-range dependency of speech sequences.

The existing baseline models, such as the SEGAN [69], Wave-U-Net [56], U-NET [41], Masking [26], CRN [86], Self-attention [6], Autoencoder [66],Parallel RNN [51] are built using convolution layers only. It is difficult for CNN alone to correctly model the long-range dependencies of speech signals. The local receptive field of the convolution limits the model's ability to capture long-range dependencies across input sequences.

To deal with the long range dependency of speech, some models [51, 86] incorporated LSTMs in the bottleneck. Even though the performance of models [51, 61, 85, 108] is better, the LSTMs are easily prone to the problem of overfitting, and it also requires a long time to train. LSTM requires 4 linear layers (MLP layer) per cell to run at each time step. Linear layers require large amounts of memory bandwidth to be computed. The Self-attention model computes attention scores by comparing each element in the input sequence with every other element, resulting in a dense attention matrix. This computation becomes computationally expensive as the sequence length increases.

In recent years, speech enhancement has been thought of as supervised learning [94], based on the idea of time-frequency (T-F) Masking in computational auditory scene analysis (CASA). For supervised speech enhancement, it is important to choose the right training target [98]. On the one hand, training with a well defined target can improve speech quality and intelligibility. On the other hand, the training target should be something that can be supervised learning. In the T-F domain, a lot of training targets have been made, and most of them can be put into two groups. One set includes targets like the ideal ratio mask (IRM) [98], which characterize the time-frequency connections between noisy speech and clean speech. There are other goals that are based on mapping, such as the target magnitude spectrum (TMS) [25, 55] and the log-power spectrum (LPS) [103], which display the spectral characteristics of clean speech.

The magnitude spectrum of noisy speech is used to determine the majority of these training goals, which is obtained using a short-time Fourier transform (STFT). So, most speech enhancement algorithms only change the magnitude spectrogram and then use the noisy phase spectrogram to resynthesize the improved time-domain waveform. There are two reasons why we are unable to improve the phase spectrogram. First, it was found that the phase spectrogram does not have a clear structure, this makes it hard to figure out the phase spectrogram of clean speech [101]. Second, people thought that phase enhancement was not required to improve speech [95]. But newer research by Paliwal et al. [67] shows that a correct phase estimate can improve subjective and objective speech quality a lot, especially when the analytical window for phase spectrum calculation is set up correctly. Following that, several phase enhancement algorithms were developed for the separation of speech. Mowlaee et al. [72] used the mean

squared error (MSE) to figure out the phase spectra of two sources in a mixture. Krawczyk and Gerkmann [46] did phase enhancement on voiced frames but not on unvoiced frames. Kulmer et al. [47] calculated the phase of clean speech by disassembling the instantaneous noisy phase spectrum and temporal smoothing. T-F Masking can also take phase information into consideration. Wang and Wang [97] trained a deep neural network (DNN) to use the noisy phase and an inverse Fourier transform layer to directly rebuild the time-domain enahnced signal. The results suggest that combining training speech resynthesis with mask estimation improves perceived quality while keeping objective intelligibility. The phase-sensitive mask (PSM) is another way to do things [14].

The results conclude that the signal-to-distortion ratio (SDR) is greater when PSM estimation is utilized instead of just enhancing the magnitude spectrum. Williamson et al. [101] found that the phase spectrogram does not have spectrotemporal structure, but that both the real and imaginary parts of the clean speech spectrogram have clear structure and can be learned this way. So, they made the complex ideal ratio mask (cIRM), which can take noisy speech and reconstruct it sound like clean speech again. In their experiments, they employ a DNN to estimate both the imaginary and real spectra simultaneously. CIRM estimation is different from [46, 47, 72] in that it can improve both the phase spectrum and magnitude of noisy speech. The results show that complex ratio Masking (cRM) improves perceived quality more than IRM estimation, while improving objective intelligibility only slightly or not at all. Fu et al. [18] then used a convolutional neural network (CNN) to estimate the clean real and imaginary spectra from the noisy ones. The time-domain waveform is then made from the estimated real and imaginary spectra. They also trained a deep neural network (DNN) to turn noisy LPS characteristics into clean ones. Their results show that complex spectral mapping with a DNN does better than LPS spectral mapping in terms of STOI and PESQ.

In the last ten years, the use of CNNs and recurrent neural networks has greatly helped supervised speech enhancement (RNNs). RNNs with long-term short-term memory (LSTM) are used to improve speech in [99, 100]. Chen et al. [7] came up with an RNN with four hidden LSTM layers to deal with the problem of speaker generalization of noise-free models. They found that the RNN works well with untrained speakers and does better on STOI than a feedforward DNN. Furthermore, CNNs have also been employed to estimate masks and map spectral data [17, 23, 71]. In [71], Park et al. did spectral mapping with a convolutional encoder-decoder network (CED). The CED can remove noise just as well as a DNN or an RNN, but it has a lot fewer trainable parameters. Grais et al. [23] also came up with a similar architecture for encoders and decoders. We just made a gated residual network based on dilated convolutions that can use long-term contexts and has wide receptive fields [88]. Convolutional recurrent networks (CRNs) take the ability of CNNs to pull out features and the ability of RNNs to represent time and put them together. The CRN was made by Naithani et al. [61] by putting convolutional layers, recurrent layers, and fully connected layers on top of each other in that order. In [108], a CRN architecture like this one was made. To make CRN, CED and LSTMs and put them together, which is like a causal system [86]. Takahashi et al. [85] made a CRN with many low-scale convolutional and recurrent layers.

However, since spectrogram of speech and the complex targets are inherently complex valued, using complex networks could potentially lead to richer representations and more efficient modeling [20, 32]. This occurs because complex models adhere to the rules of complex multiplication, allowing them to simultaneously acquire the real as well as imaginary components based on previous knowledge. In prior research, the authors developed complex models utilizing convolutional recurrent architecture, yielding encouraging results [109, 110]. Lately, Transformer models [37, 68], particularly the Conformer architecture,

have significantly enhanced sequence modeling capabilities [24]. Unlike recurrent learning, the Conformer model utilizes self-attention [93] to capture overall dependencies within a sequence, while also considering local dependencies through convolutional layers. Therefore, it is highly desirable to extend the convolutional recurrent model into a Conformer-based model that utilizes full-complex networks for enhancing speech.

In [54, 65, 80, 84, 102], the authors proposed a DenseNet to reduce the number of dilated convolution layers to cover the large receptive area. With DenseNet, we aggregate the early and later layer features directly within a single convolution layer via dense skip connectivity. It is inefficient to have many parameters, especially for high-resolution data, especially when transforming local features into global ones. In [21], the authors proposed a network design that combines the advantages of DenseNet with the advantages of dilated convolution. The typical dilated convolutions were indeed employed, and dilation factors were computed based on layer depth, resulting in significant aliasing.

We just came up with a new CRN to do complex spectral mapping for monaural speech enhancement [87] in a preliminary study. This CRN was made using the architecture in [86]. In this study, we improve the CRN architecture [87] and look at how complex spectral mapping can be used to improve monoaural speech. First, each convolutional or deconvolutional layer is replaced with a gated linear unit (GLU) block [10] followed by dense layer and efficient channel attention [12]. Second, we add a linear layer on top of the last deconvolutional layer to guess the real and imaginary spectra.

The main objective of the proposed work is that to improve the quality and intelligibility of the degraded speech. The main advantage of GCRN is that it performance better than normal CNN approaches. The output of GCRN is given to Dense layer. The main advantage of dense layer is that it avoids vanishing gradient because the input of the given layer is not completely depend on the previous layer but also other previous several layers. Also, thinner (less number of channels) dense network outperform than wider dense network and hence the efficiency of the parameter network improves. The output of dense is given to ECA to improve information flow across layers by learning a dynamic representation without reducing the parametric space dimension. The motivation of the proposed work is that we want to improve the performance of the network and to improve the computational cost by keeping the same dimensionality. Efficient Channel Attention (ECA) module extracts the useful channels information by using a cross-channel interaction method without affecting the channel dimensions. In module testing, choosing an adaptable kernel size K for the ECA improved network performance significantly.

1. A gated convolutional recurrent network with efficient channel attention (GCRN-ECA) for complex spectral mapping is proposed, which amounts to a causal system for monaural speech enhancement. Each layer in encoder and decoder consists of dense block.
2. Convolutional Neural Networks (CNN) based techniques suffers from limited receptive field. To overcome this effect, Gated Convolutional Recurrent Neural Networks is proposed.
3. The advantage of dilated convolutions in the receptive field increases with increasing dilation rates, which are used to capture long-range speech contexts. And the dense connectivity provides a feature map with more precise target information by passing through multiple layers.
4. To represent the correlation between neighboring noisy speech frames, a two Layer GRU is added in the bottleneck of Wave-U-NET [56], which has the advantage of increased training speed because of its simpler architecture. GRU captures the long-range dependencies across input sequences.

5. The model is incorporated with a novel ECA network, which can improve information flow across layers by learning a dynamic representation without reducing the parametric space dimension. ECA chooses an adaptable kernel size (k) in model testing, which can improve accuracy and efficiency by allowing cross-channel interactions while preserving dimensions. ECA module can implement cross-channel interaction without dimensionality reduction.

The remainder of this paper is organized as follows. Section 2 discuss about monaural speech in the STFT domain. Section 3 explains description of the system. Section 4 discuss about experimental set up, Section 5 discusses the experiment outcomes. Section 6 concludes the paper.

2 Monaural speech enhancement in STFT domain

Monaural speech enhancement separates the speech s[t] from the noise n[t] in the background. A noisy mixture y can be modeled as

$$y[t] = n[t] + s[t] \tag{1}$$

where time sample index is t. Applying STFT on both sides will lead us to

$$Y_{m,f} = N_{m,f} + S_{m,f} \tag{2}$$

where N, Y, and S are the STFTs of n, y, and s, and f and m are the indices for the frequency bin and time frame respectively. In polar coordinates, Eq.(2) is written

$$\left|Y_{m,f}\right| e^{i\theta} = \left|N_{m,f}\right| e^{i\theta S(m,f)} + \left|S_{m,f}\right| e^{i\theta S(m,f)} \tag{3}$$

where θ and ||show the phase response and the magnitude response, respectively. In Clean Speech's target magnitude spectrum (TMS), the letter i stands for the "imaginary unit". The target magnitude spectrum (TMS) of clean speech is often used as a training target in most spectral mapping-based approaches [25, 55].

In the reconstruction process, the estimated magnitude $\left|Y_{m,f}\right|$ is combined with the noisy phase $\left|\hat{S}_{m,f}\right|$. The STFT of speech signal may also be represented in Cartesian coordinates, which offers a distinct perspective. So, Eq.(2) can be re-written as

$$Y_{m,f}^{(r)} + iY_{m,f}^{(i)} = \left(S_{m,f}^{(r)} + N_{m,f}^{(r)}\right) + i\left(S_{m,f}^{(i)} + N_{m,f}^{(i)}\right) \tag{4}$$

where superscripts (i) and (r) stand for the imaginary and real components, respectively. The cIRM [101] is as defined as

$$M = \frac{Y^{(r)}S^{(r)} + Y^{(i)}S^{(i)}}{\left(Y^{(r)}\right)^2 + \left(Y^{(i)}\right)^2} + i\frac{Y^{(r)}S^{(i)} - Y^{(i)}S^{(r)}}{\left(Y^{(r)}\right)^2 + \left(Y^{(i)}\right)^2} \tag{5}$$

The noisy spectrogram can be turned into the improved spectrogram by adding an estimate of the cIRM $\hat{M}$ is

$$S = \hat{M} \times Y \tag{6}$$

where the "X" multiplication complex operator. Signal Approximation [33] performs the Masking by reducing the difference between clean speech and estimated speech. The definition of the loss for cRM-based signal approximation (cRM-SA) is:

$$SA = |cRM \times Y - S|^2 \tag{7}$$

where (|||) is the complex modulus. Spectral mapping is learned from the real and imaginary spectra of noisy speech (Y(r) and Y(i)) to those of clean speech (S(r) and S(i)). The time-domain signal is obtained by combining the estimated real and imaginary spectra. Williamson et al. [101] claimed that it is not a good idea to use a DNN to try to predict the real and imaginary components of the STFT. We show that complex spectral mapping is always better than magnitude spectral mapping, complex ratio Masking, and complex ratio Masking-based signal approximation in terms of STOI and PESQ.

3 Description of the system

3.1 Convolutional recurrent network

A convolutional recurring network [86] was built, which is simply an encoder-decoder architecture with LSTMs between the encoder and the decoder. There are five convolutional layers in the encoder, while the decoder has also five deconvolutional layers. Two LSTM layers describe temporal dependencies between the encoder and decoder. The encoder-decoder structure is constructed symmetrically, with the number of kernels increasing in the encoder and decreasing in the decoder. A stride of 2 is used in both convolutional and deconvolutional layers with frequency dimension to aggregate the context along the frequency direction. In those other terms, the frequency dimensionality of the feature maps is halved in the encoder and doubled in the decoder, ensuring that the output has same form as the input. Also we used Skip Connections to connect each encoder layer's output to the input of the matching decoder layer.

3.2 Gated linear units

The way information moves through the network is controlled by "gating" mechanisms, which could make it possible to model more complex interactions. They were first developed for RNNs [28]. In a recent study on convolutional modelling of images, Van den Oord et al [64]. used an LSTM-style gating mechanism. This led to masked convolutions:

$$\begin{aligned} \mathbf{y} &= \tanh\left(\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1\right) \odot \sigma\left(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2\right) \\ &= \tanh\left(\mathbf{v}_1\right) \odot \sigma\left(\mathbf{v}_2\right) \end{aligned} \tag{8}$$

Let $v_1 = x * W_1 + b_1$ and $v_2 = x * W_2 + b_2$, where W's and b's denote kernels and biases, respectively, and σ denotes the sigmoid function. The symbols $*$ and $\circ$ represent convolution operation and element-wise multiplication, respectively.The gradient of the gating is

$$\begin{aligned} \nabla\left[\tanh\left(\mathbf{v}_1\right) \odot \sigma\left(\mathbf{v}_2\right)\right] = & \tanh'\left(\mathbf{v}_1\right) \nabla \mathbf{v}_1 \odot \sigma\left(\mathbf{v}_2\right) \\ & + \sigma'\left(\mathbf{v}_2\right) \nabla \mathbf{v}_2 \odot \tanh\left(\mathbf{v}_1\right) \end{aligned} \tag{9}$$

where $\tanh'\left(\mathbf{v}_1\right), \sigma'\left(\mathbf{v}_2\right) \in (0, 1)$ are both in the interval $(0, 1)$, and the prime symbol denotes differentiation. As the network depth increases, the gradient vanishes gradually because of the downscaling factors $\tanh'\left(\mathbf{v}_1\right)$ and $\sigma'\left(\mathbf{v}_2\right)$. To address this issue, Dauphin et al. [10] introduced Gated Linear Units (GLUs):

$$\begin{aligned} \mathbf{y} &= \left(\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1\right) \odot \sigma\left(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2\right) \\ &= \mathbf{v}_1 \odot \sigma\left(\mathbf{v}_2\right). \end{aligned} \tag{10}$$

The gradient of the GLUs

$$\nabla [\mathbf{v}_1 \odot \sigma (\mathbf{v}_2)] = \nabla \mathbf{v}_1 \odot \sigma (\mathbf{v}_2) + \sigma' (\mathbf{v}_2) \nabla \mathbf{v}_2 \odot \mathbf{v}_1 \tag{11}$$

includes a path $\mathbf{v}_1 \odot \sigma (\mathbf{v}_2)$ Without downscaling, which you can be treated as a multiplicative skip link that makes it easier for gradients to flow through layers. Figure 1(a) & (b) shows that a deconvolutional GLU block, called "DeconvGLU," is similar to a convolutional GLU block, except that it has deconvolutional layers instead of convolutional layers.

3.3 Dense block

The idea behind densely connected network is that feature reuse in which an output at a given layer is reused multiple times in the subsequent layers. i.e., the input to the given layer is not only the output of previous layer but also the outputs of previous several layers. This type of network has two advantages. First, it avoids vanishing gradient because the input of the given layer is not completely depend on the previous layer but also other previous several layers. Second, Thinner (less number of channels) dense network outperform than wider dense network and hence the efficiency of the parameter network improves. Finally, the dense connection can be defined as

$$\mathrm{y}^l = \mathrm{g}\left(\mathrm{y}^{l-1}, \mathrm{y}^{l-2}, \ldots . \mathrm{y}^{l-D}\right) \tag{12}$$

y_l denotes the output at the layer l, g is the function in the single layer, D represents depth of dense connections. The proposed network uses dense block after each layer in encoder and decoder. The dense block is shown in Fig. 2.

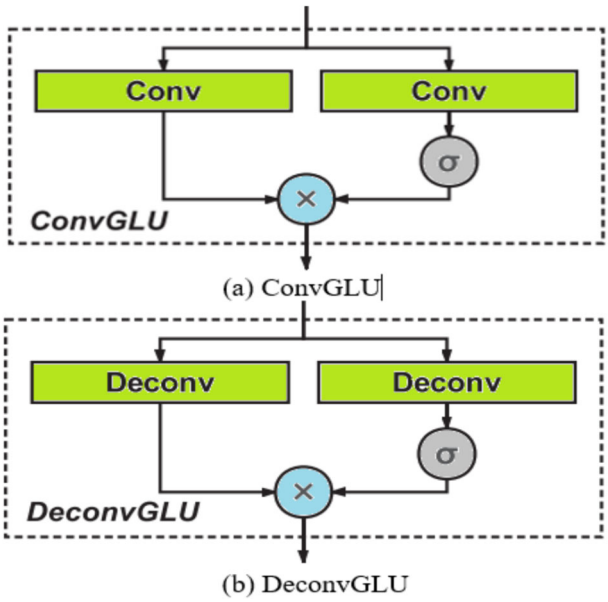

Fig. 1 Diagram of convolutional GLU and deconvolutional GLU

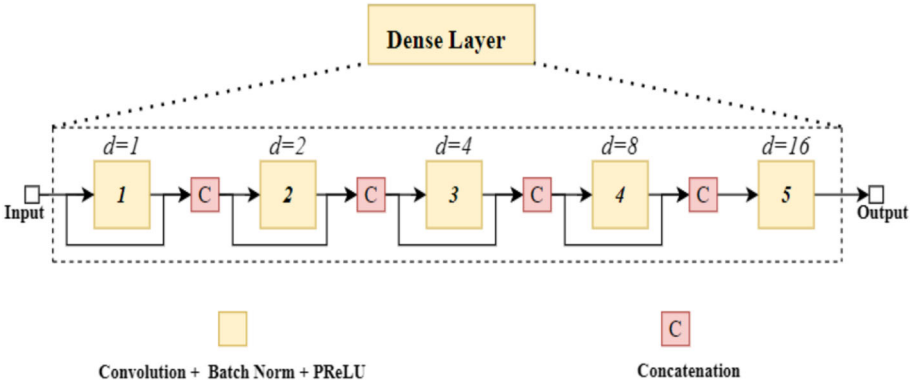

Fig. 2 Dense block

The dense block consists of five convolution layers followed by layer normalization and ReLu. The input of given layer is formed by the output of previous layer along with the outputs of several previous layers. The input channels increases linearly with the successive layers as C,2C,3C,4C,5C. The output of each convolution has C channels

3.4 Efficient channel attention module

The speech signal characteristics of convolutional neural networks are usually obtained by integrating spatial and channel dimensions on a local receptive field. Based on the significance of the channel, to improve the selected channel features and suppress channels that are less useful in the present work. In [96], we see the reviewed model of Squeeze-and-Excitation (SE) in [30]. In the Squeeze-and-Excitation module, global average pooling (GAP) for each channel is applied first separately based on the input features. In order to capture cross-channel interactions, it requires two fully connected non-linear layers. The dimensionality of channel is reduced by using this method, and reducing the dimensions has negative impact on network prediction. Hence, a one-dimensional convolution is employed on the fully connected layer of the SE module to increase the efficiency of channel attention. In this paper, the Efficient channel attention (ECA) module is proposed as a novel cross-channel interaction network without reducing the dimensionality. A significant improvement has been made to the overall network calculation speed as well as its prediction results as compared with the previous SE module. This module is structured as shown in Fig. 3. In the ECA model, the input channel is first sent through the global average pooling (GAP) layer and then employs a 1D convolution for local channel interaction. The 1D convolution kernel size is the same as the convolution kernel, and the kernel size parameter is used to calculate the coverage of cross-channel interaction. The 1D convolution layer by default sets the padding value equal to half of the kernel and it takes the integer part. Local cross-channel interaction is completed and then it is sent to the sigmoid function. The sigmoid function output is element-wise multiplied with the input channel and its product as an ECA module output. To determine the mapping relationship(ψ) between the number of channels and kernel size(k), the number of channels should be 2n. The eq. (2) gives the mapping relationship (ψ), where b and γ are set to 2 and 1:

$$k = \psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}} \tag{13}$$

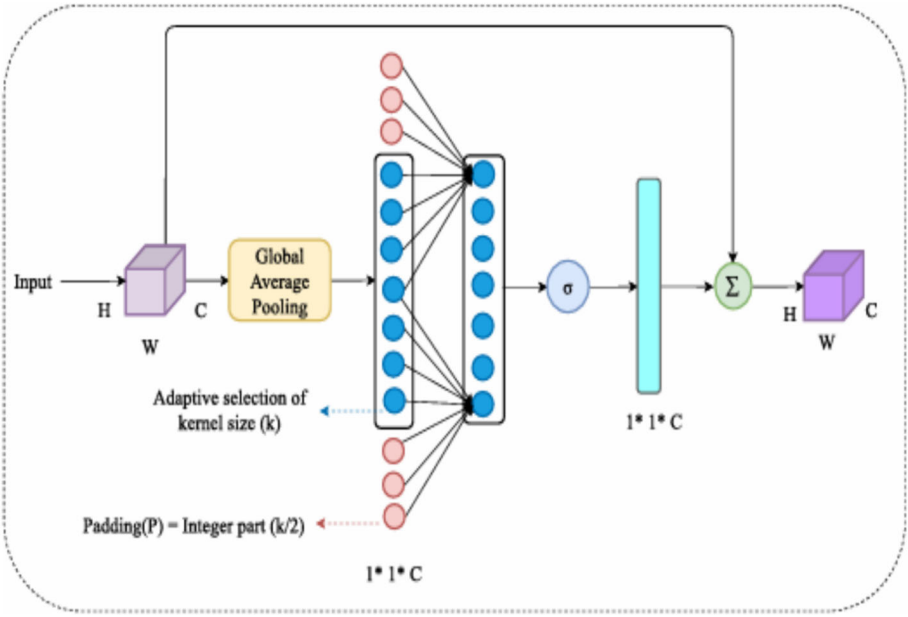

Fig. 3 Efficient channel attention mechanism

where $|\gamma|_{\text{odd}}$ represents the closet odd number of γ. It is possible to calculate the extent cross-channel interaction by selecting k. ECA improves accuracy and efficiency by allowing cross-channel interactions while preserving dimensions. Due to this, it attempts to add ECA modules in this paper.

3.5 Gated Recurrent Unit (GRU)

Combining the forget and input gates in LSTM into a single one, GRU is introduced with two gates rt and zt , named reset and update gates, respectively. GRU as a variation of LSTM is faster and computationally more efficient than LSTM, while in some cases, it yields even better performance on less training data. The extension of the GRU model in the given figure is displayed through multiple unified hidden layers. The module structure of GRU is repetitive, which is more straightforward than long and short-term memory because each recurrent neural network feature of the module is the same. It has only two doors, the updated door and the reset door, namely zt and rt in Fig. 4. The update gate is used to supervise the extent to which the knowledge of the previously hidden state is extended to the current state. The greater the value of the update gate, the more knowledge of the previous state is introduced. Therefore, if the reset gate is used to adjust the degree of knowledge transfer of the past state, the smaller the value of the reset gate, the more it will be transferred. Therefore, the capture of short-term dependence is usually in the cyclic activation of the reset gate, while the long-term dependence is in the activation of the update gate. A gate controller zt, controls the both the input gate and forget gate. When zt=1, the forget gate is closed and the input gate is open. When zt=0, the forget gate is open, and the input gate is closed. At each step, the previous (t-1) memory is saved, and the input of the time step is cleared. GRU uses tow gates

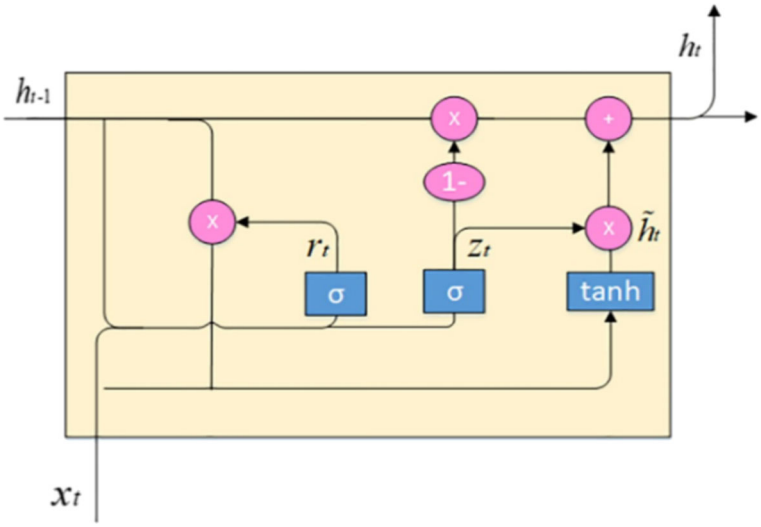

Fig. 4 Block diagram of GRU

instead of three gates like LSTM. So GRU reduces network complexity as well as improves the performance. each layer models the temporal dynamics of speech. The reset gate rt is used to determine how much of the previous memory information needs to be retained. The smaller rt is the lesser information from the previous state is written. The gate zt is used to control the extent of which the state information from the previous moment is brough into the current state. The larger zt is the more state information from previous moment is brought in.

3.6 Network architecture

This research tends to add to the CRN architecture outlined in [87] to perform complex spectral mapping. The resulting CRN includes GLUs, and gated convolutional recurrent network (GCRN), dense block and ECA.

Figure 5 shows our proposed design for the GCRN structure. As shown in [18], the real and imaginary spectrograms of noisy speech are treated as two different input channels. Figure 5 shows that the encoder and GRU modules are used to estimate both real and imaginary components, while real and imaginary spectrograms are approximated by two different decoder configurations. This design is based on multi-task learning [48, 107], which means that related prediction tasks are learned at the same time by sharing information across tasks. Estimating the real and imaginary parts is a part of spectral mapping that is connected to two other tasks [101]. We will assume that all signals are analysed at 16 kHz. Using a 20-ms Hamming window, a set of time frames are made in which each pair of frames overlaps by 50%. We use spectra with 161 dimensions, which is the same as a 320-point STFT (16 kHz X 20 ms). Remember that the number of feature maps in each decoder layer doubles when skip connections are used. We use a kernel size of 1X3, which will not affect the performance. After each convolutional or deconvolutional GLU block is followed by an exponential linear unit (ELU) [8] activation function and a batch normalization [34] operation.

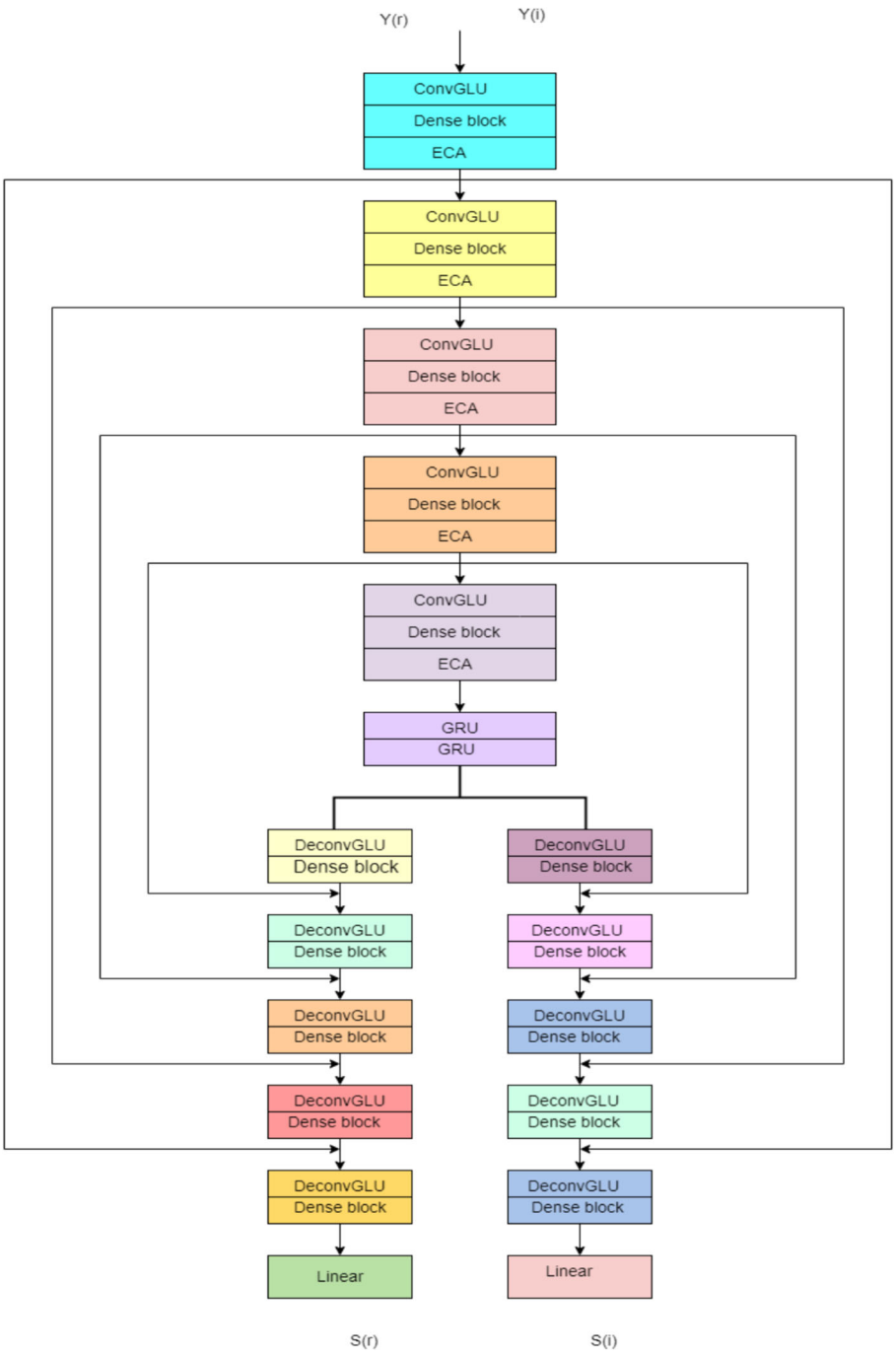

Fig. 5 Proposed GCRN-ECA structure

4 Experimental setup

The Voice Bank + DEMAND [91] dataset serves as the basis for both training and testing. It consists of 11,572 pairs of clean and noisy speech for training, along with 824 noisy clips designated for testing. We use the Common Voice corpus [9] to test our system, which is a publicly available voice dataset, powered by the voices of volunteer contributors around the world. People who want to build voice applications can use the dataset to train machine learning models. The data set contains 1653880(1.6 million) utterances from 84659 speakers. From the Common Voice, we select the English corpus and randomly choose 2000 utterances for the training set and 400 utterances for the validation set, respectively. The test set is also taken from Common Voice, which consists of 400 utterances. We built training and validation sets using different types of noise from Noizeus [63], consisting of white, pink, restaurant, and babble noises. Five SNR levels are used to test the noise mixture i.e.,-6dB, -3dB, 0dB, 3dB and 6dB. We used cross-validation for validation set.

Hyperparameters: There are two convolutional layers with 256 filters. We have the stride as 1 and 16 and the filter size is 11 and 32 for respective layer. The hidden size of the lstm is 1024 and there are two layers. The batch size is 32. Initial standard deviation is 0.02. A pretraining section is done of 10 epochs The learning rate starts at 1e-3 decays by 0.5 until it reaches a minimum of 1e-6 and the optimizer is the Adam Optimizer. Early stopping is implemented, otherwise it runs for a maximum of 200 epochs.

4.1 Experimental results and analysis

The short-time objective intelligibility (STOI) [83], the perceptual evaluation of speech quality (PESQ) [31] and the signal to noise ratio (SNR) is used as the objective metrics. The experiment results are compared with the existing techniques Wiener [74], SEGAN [69], Wave-U-Net [56], U-NET [41], Masking [26], CRN [86], Self-attention [6], Autoencoder [66], Parallel RNN [51].

Table 1 shows PESQ values for the existing techniques like wiener [74] , SEGAN [69], UNET, CRN, Self-attention, Auto encoder and Parallel RNN. In case of babble noise, the PESQ values at -3 dB and 0 dB are 1.83 and 2.08 respectively for SEGAN [69]. The PESQ values at -6dB and 3 dB are 1.85 and 2,41 respectively in case of U-NET [41]. For CRN method, at -3dB and 6dB the PESQ values are 2.24 and 3.02 respectively. For Autoencoder [66] method, the PESQ values at 0 dB and 3dB are 2.90 and 3.17 respectively. The proposed method yields PESQ values of 2.42, 2.94, 3.47 for the input test SNRs of -6dB, 0dB and 6dB respectively. The proposed method shows better results than other techniques. In case of street noise, Wave-U-NET [56] gives PESQ values of 1.70 and 2.11 respectivley at the input SNRs of -6 dB and 0 dB. The PESQ values at input SNR of -3dB and 3dB are 2.46 and 2.86 respectively for Self attention [6] method. For the proposed method the PESQ values of 3.11 and 3.27 for the input test SNRs of 3dB and 6dB.

Comparative performance of STOI is shown in Table 2. In case of street noise, SEGAN [69] gives STOI values of 64.4 and 75.6 respectively at the input SNRs of -6 dB and 0 dB. The STOI values at input SNR of -3dB and 3dB are 78.2 and 89.1 respectively for Self attention [6] method. For the proposed method the STOI values of 91.9 and 95.4 for the input test SNRs of 3dB and 6dB. In case of babble noise, the STOI values at -3 dB and 0 dB are 69.4 and 75.8 respectively for wiener [74] . The STOI values at -6dB and 3 dB are 67.2 and 72.2 respectively in case of U-NET [41]. For CRN [86] method, at -3dB and 6dB the STOI values are 77.9 and 92.1 respectively. For Autoencoder method, the STOI values at 0 dB and

Table 1 Comparative performance in terms of PESQ in the case of babble noise and street noise

Noise type	Babble noise						Street noise					
Test SNR	-6dB	-3dB	0dB	3dB	6dB	Avg.	-6dB	-3dB	0dB	3dB	6dB	Avg.
Noisy	1.51	1.70	1.89	2.09	2.30	1.898	1.47	1.65	1.81	1.98	2.20	1.822
Wiener [74]	1.57	1.78	1.97	2.17	2.39	1.976	1.53	1.72	1.90	2.10	2.29	1.908
SEGAN [69]	1.69	1.83	2.08	2.25	2.48	2.066	1.64	1.80	2.00	2.17	2.38	1.998
Wave-U-NET [56]	1.76	1.90	2.18	2.33	2.59	2.152	1.70	1.85	2.11	2.20	2.49	2.07
U-NET [41]	1.85	1.95	2.27	2.41	2.69	2.234	1.75	1.90	2.22	2.31	2.53	2.142
Masking [26]	1.99	2.12	2.50	2.63	2.89	2.426	1.80	2.07	2.42	2.52	2.63	2.288
CRN [86]	2.09	2.24	2.63	2.85	3.02	2.566	1.95	2.18	2.54	2.72	2.89	2.456
Self-attention [6]	2.29	2.57	2.85	3.10	3.29	2.82	2.15	2.46	2.76	2.86	3.07	2.66
Autoencoder [66]	2.37	2.68	2.90	3.17	3.37	2.898	2.23	2.57	2.80	3.01	3.12	2.746
Parallel RNN [51]	2.39	2.72	2.68	3.21	3.41	2.882	2.29	2.50	2.60	3.07	3.20	2.732
Proposed GCRN-ECA	2.42	2.76	2.94	3.27	3.47	2.972	2.36	2.67	2.81	3.11	3.27	2.844

Table 2 Comparative performance in terms of STOI in the case of babble noise and street noise

Noise type	Babble noise						Street Noise					
Test SNR	-6dB	-3dB	0dB	3dB	6dB	Avg.	-6dB	-3dB	0dB	3dB	6dB	Avg.
noisy	63.0	67.4	74.1	81.1	86.1	74.34	62.1	66.7	71.3	76.2	82.1	71.68
wiener [74]	64.1	69.4	75.8	82.3	87.3	75.78	63.5	68.3	73.2	79.3	83.3	73.52
SEGAN [69]	65.4	70.5	76.2	83.8	88.2	76.82	64.4	69.2	75.6	81.8	85.2	75.24
Wave-U-NET [56]	66.0	71.2	78.9	84.1	89.4	77.92	65.0	70.9	77.2	83.4	86.4	76.58
U-NET [41]	67.2	72.2	79.5	85.3	90.5	78.94	66.3	71.2	78.8	84.3	89.5	78.02
Masking [26]	69.3	75.1	82.1	86.9	91.6	81	68.1	73.5	82.1	85.9	90.2	79.96
CRN [86]	72.1	77.9	83.1	88.1	92.1	82.66	71.7	75.5	83.5	86.2	91.7	81.72
Self attention [6]	76.4	81.5	87.7	92.1	93.6	86.26	74.4	78.2	84.7	89.1	92.3	83.74
Autoencoder [66]	77.1	82.2	88.3	92.7	94.5	86.96	75.5	80.2	86.1	90.7	93.6	85.22
Parallel RNN [51]	77.4	83.0	89.4	92.1	95.9	87.56	76.2	82.0	88.7	91.1	94.2	86.44
Proposed GCRN-ECA	78.5	85.2	90.3	92.9	96.6	88.7	77.5	83.2	89.3	91.9	95.4	87.46

3dB are 88.3 and 92.7 respectively. The proposed method yields STOI values of 78.5, 90.3, 96.6 for the input test SNRs of -6dB, 0dB and 6dB respectively. The proposed method shows better results than other techniques.

Table 3 shows SNR values for the existing techniques like wiener [74], SEGAN [69], UNET, CRN, Self attention, Auto encoder and Parallel RNN. In case of babble noise, the SNR values at -3 dB and 0 dB are -2.18 and 1.19 respectively for wiener [74] . The SNR values at -6dB and 3 dB are -2.71 and 6.19 respectively in case of U-NET [41]. For CRN [86] method, at -3dB and 6dB the SNR values are 4.36 and 12.3 respectively. For Autoencoder method, the SNR values at 0 dB and 3dB are 10.2 and 13.3 respectively. The proposed method yields SNR values of 6.07, 11.98, 17.26 for the input test SNRs of -6dB, 0dB and 6dB respectively. The proposed method shows better results than other techniques. In case of street noise, SEGAN [69] gives SNR values of -4.12 and 1.17 respectively at the input SNRs of -6 dB and 0 dB. The SNR values at input SNR of -3dB and 3dB are 7.19 and 12.31 respectively for self-attention method. For the proposed method the SNR values of 13.54 and 16.56 for the input test SNRs of 3dB and 6dB.

5 Discussion on results

The SEGAN [69] is an end-to-end SE model where only strided convolutions are used in the generator and discriminator. In this model also only ordinary convolution operations are used. Even though the performance of the model is good but it suffers from computational complexity. The SEGAN [69] has the PESQ of 2.06& 1.99 on average in babble and street noise environments which is better than wiener [74] model. The SEGAN [69] has the STOI of 76.82 & 75.84 on average in babble and street noise environments which is better than wiener [74] model. The SEGAN [69] has the SNR of 1.72 & 0.926 on average in babble and street noise environments which is better than wiener [74] model. The reason for enhanced result is that the deep learning models can automatically learn relevant features directly from the input data. Moreover, the wiener [74] model is sensitive to noise type. The limitation of SEGAN is its computational complexity. The Wave-U-NET [56] is a time domain SE model with basic U-NET [41] architecture with 1D ordinary convolution layers in the encoder and decoder with a 1D convolution as a bottleneck. The Wave-U-NET [56] has the PESQ of 2.15 & 2.07 on average in babble and street noise environments which is better than wiener [74] model. The Wave-U-NET [56] has the STOI of 77.92 & 76.58 on average in babble and street noise environments which is better than wiener [74] model. The Wave-U-NET [56] has the SNR of 2.95 & 1.88 on average in babble and street noise environments which is better than wiener [74] model and SEGAN [69]. The performance is poor at low SNRs.

The U-NET [41] is a basic U-NET [41] model with encoder-decoder architecture. Even though the encoder extracts better features from noisy speech it is also necessary to deal with long term dependency of speech signal. The U-NET [41] has the PESQ of 2.23 & 2.24 on average in babble and street noise environments which is better than wiener [74] model. The U-NET [41] has the STOI of 78.94 & 78.02 on average in babble and street noise environments which is better than wiener [74] model. The U-NET [41] has the SNR of 3.40& 2.91 on average in babble and street noise environments. The CNN models such as SEGAN [69] , Wave-U-NET [56] and U-NET [41] alone cannot well model the long-range dependencies of speech signal. In all the existing models only, ordinary convolutional layers are used. The local receptive field of the convolution limits the model's ability to capture long-range dependencies across input sequences. In CRN [86] [55] model to further enhance

Table 3 Comparative performance in terms of SNR in the case of babble noise and street noise

Noise type	Babble noise						Street Noise					
Test SNR	-6dB	-3dB	0dB	3dB	6dB	Avg.	-6dB	-3dB	0dB	3dB	6dB	Avg.
Noisy	-5.92	-3.22	0.08	2.08	6.08	-0.18	-6.92	-4.32	-1.28	1.08	4.08	-1.472
wiener [74]	-4.13	-2.18	1.19	3.81	6.95	1.128	-5.31	-3.28	0.19	2.61	5.25	-0.108
SEGAN [69]	-3.61	-1.20	1.67	4.12	7.82	1.76	-4.12	-2.10	1.17	3.46	6.22	0.926
Wave-U-NET [56]	-2.01	0.57	2.28	5.20	8.71	2.95	-3.32	-1.27	2.00	4.22	7.81	1.888
U-NET [41]	-2.71	1.03	3.13	6.19	9.38	3.404	-2.25	0.03	2.73	5.79	8.28	2.916
Masking [26]	0.13	2.28	5.02	8.67	11.74	5.568	0.03	1.28	3.12	7.79	10.54	4.552
CRN[86]	1.90	4.36	6.59	9.91	12.3	7.012	1.20	3.36	5.39	8.41	11.7	6.012
Self attention [6]	4.35	7.02	8.22	12.08	14.29	9.192	3.52	5.12	6.42	10.08	13.91	7.81
Autoencoder [66]	5.56	8.09	10.2	13.3	15.4	10.51	4.16	7.19	8.29	12.31	14.67	9.324
Parallel RNN [51]	5.85	8.17	10.71	13.91	16.10	10.948	4.75	7.97	9.61	12.91	15.34	10.116
Proposed GCRN-ECA	6.07	9.03	11.98	14.04	17.26	11.676	5.87	8.12	10.26	13.54	16.56	10.87

the performance of U-NET [41] the LSTMs are used in between encoder and decoder ofU-NET [41] to learn long term dependencies of speech signals. The CRN [86] has the PESQ of 2.56 &2.45 on average in babble and street noise environments which is better than wiener [74] model. The CRN [86] has the STOI of 82.66& 81.72 on average in babble and street noise environments which is better than wiener [74] model. The CRN [86] has the SNR of 7.01& 6.01 on average in babble and street noise environments which is better than SEGAN [69], Wave-U-NET [56] and U-NET [41]. Even though the performance of model is better the LSTMs are easily prone to the problem of overfitting and it also requires a large time to train. LSTM requires 4 linear layers (MLP layer) per cell to run at each time step. Linear layers require large amounts of memory bandwidth to be computed. Speech enhancement performance is influenced by CNN's limited receptive field, which restricts its ability to extract long-range dependency of speech sequences. Autoencoder [66] model does not use any attention mechanism for better feature extraction. Later attention mechanisms are added to selectively focus on relevant features of the speech signal that are important for enhancement. The Self attention [6], has the PESQ of 2.82 & 2.66 on average in babble and street noise environments which is better than wiener [74] model. The Self attention [6], has the STOI of 86.26 & 83.74 on average in babble and street noise environments which is better than wiener [74] model. The Self attention [6], has the SNR of 9.91 &7.81 on average in babble and street noise environments which is better than U-NET [41] models and CRN models. The limitation of self-attention is it results in a dense attention matrix. This is computationally expensive as the sentence length increases.

The existing baseline models, such as the SEGAN [69], Wave-U-NET [56], U-NET [41], Masking [26], CRN , Self-attention [6], Autoencoder [66], Parallel RNN [51] are built using convolution layers only. It is difficult for CNN alone to correctly model the long-range dependencies of speech signals. The local receptive field of the convolution limits the model's ability to capture long-range dependencies across input sequences. To deal with the long range dependency of speech, some models [51, 86] incorporated LSTMs in the bottleneck. Even though the performance of models [51, 61, 85, 108] is better, the LSTMs are easily prone to the problem of overfitting, and it also requires a long time to train. LSTM requires 4 linear layers (MLP layer) per cell to run at each time step. Linear layers require large amounts of memory bandwidth to be computed. The Self-attention model computes attention scores by comparing each element in the input sequence with every other element, resulting in a dense attention matrix. This computation becomes computationally expensive as the sequence length increases.

To overcome the above drawbacks in the proposed model, dilated dense blocks and GRUs are introduced. First, the advantage of dilated convolutions in the receptive field increases with increasing dilation rates, which are used to capture long-range speech contexts. And the dense connectivity provides a feature map with more precise target information by passing through multiple layers. Second, to represent the correlation between neighboring noisy speech frames, a two Layer GRU is added in the bottleneck of U-NET [41], which has the advantage of increased training speed because of its simpler architecture. GRU captures the long-range dependencies across input sequences. The vanishing gradient problem is solved by GRUs using update gates and reset gates. The flow of information into and out of memory is controlled by the update and reset gates, respectively. The advantage of GRU is that it is easier to modify and doesn't require memory units, which means it can train faster than LSTM and also give performance results as fast as LSTM. Moreover, the ECA module can implement cross-channel interaction without dimensionality reduction. An appropriate cross-channel interaction can preserve performance while significantly decreasing model complexity. Hence, the performance of the model is enhanced compared to existing models,

Table 4 Subjective Assessment using VCTK dataset

Input feature	Model Mixture	PESQ 1.97	STOI 92	CSIG 3.35	CBAK 2.44	COVL 2.63	Parameters –
Time domain	SEGAN [69]	2.16	93	3.48	2.94	2.80	43.2M
Time domain	Wave-U-NET [56]	2.40	–	3.52	3.24	2.96	10.2M
Time domain	Attention-Wave-U-NET [22]	2.62	–	3.91	3.35	3.27	–
Gamma-tone spectral	MMSE-GAN [78]	2.53	93	3.80	3.12	3.14	0.79M
Magnitude	Metric GAN [19]	2.86	–	3.99	3.18	3.42	1.89M
STFT	MHSA-SPK [45]	2.99	–	4.15	3.42	3.53	–
Magnitude	MB-TCN [106]	2.94	93.64	4.21	3.41	3.59	1.66M
STFT	STFT-TCN [44]	2.89	–	4.24	3.40	3.56	–
Waveform	DEMUCS [11]	3.07	95	4.31	3.40	3.63	58M
Magnitude	5-stage SA-TCN [53]	3.02	94	4.29	3.50	3.67	9.91M
Time domain	SADNUNet [102]	2.82	–	4.18	3.47	3.51	2.63M
Time domain	Proposed model	3.04	95	4.20	3.51	3.69	3.82M

such as SEGAN [69] , Wave-U-NET [56] , U-NET [41], Masking [26], CRN [86] , Self attention [6], Autoencoder [66], Parallel RNN [51].

6 Subjective assessment using VCKT dataset

Subjective assessment using VCKT dataset are shown in Table 4. Subjective listening test methodology is designed by ITU in recommendation ITU-T P.835 [35]. This methodology was designed to evaluate the speech quality along three dimensions: signal distortion (CSIG), background distortion (CBAK) and overall quality (COVRL). This evaluation removes the uncertainty of listeners in listening tests by increased readability in terms of rating given to the enhanced speech on a five-point scale. The mean opinion score (MOS) for CSIG, CBAK and COVRL scales are described [31].

7 Conclusion

In this work we proposed a gated convolutional recurrent network with efficient channel attention (GCRN-ECA) for complex spectral mapping, which is a causal system for monaural speech enhancement. Each layer in encoder and decoder consists of dense block. The advantage of dilated convolutions present in dense block is the receptive field increases with increasing dilation rates, which are used to capture long-range speech contexts. And the dense connectivity provides a feature map with more precise target information by passing through multiple layers. The GRU captures the long-range dependencies across input sequences. The ECA module can implement cross-channel interaction without dimensionality reduction. An appropriate cross-channel interaction can preserve performance while significantly decreasing model complexity. Our results reveal that the proposed GCRN-ECA outperforms existing convolutional neural networks (CNN) and CRNs in terms of quality and intelligibility. The proposed method yields higher objective and subjective scores than existing techniques. The findings showed that the proposed model outperforms other competitive baseline methods in both PESQ and STOI metrics across the extensive VCTK and Common voice datasets.

Funding No Funding

Availability of data Commonvoice dataset and VCTK dataset.

Declarations

Conflict of interest No conflict of interest.

References

1. Aroudi A, Braun S (2020) Dbnet: doa-driven beamforming network for end-to-end farfield sound source separation. arXiv:2010.11566
2. Bastanfard A, Amirkhani D, Hasani M (2019) Increasing the accuracy of automatic speaker age estimation by using multiple ubms. In: 2019 5th conference on knowledge based engineering and innovation (KBEI), IEEE, pp 592–598

3. Berouti M, Schwartz R, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise. In: ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp 208–211
4. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoustics Speech Signal Process 27(2):113–120
5. Braun S, Gamper H (2022) Effect of noise suppression losses on speech distortion and asr performance. ICASSP 2022–2022 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 996–1000
6. Burra M, Yerva PKR, Eemani B, et al (2023) Densely connected dilated convolutions with time-frequency attention for speech enhancement. In: 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, pp 602–607
7. Chen J, Wang D (2017) Long short-term memory for speaker generalization in supervised speech separation. J Acoustical Soc America 141(6):4705–4714
8. Clevert DA, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289
9. Commonvoice (2017): https://commonvoice.mozilla.org/en
10. Dauphin YN, Fan A, Auli M, et al (2017) Language modeling with gated convolutional networks. In: International conference on machine learning, PMLR, pp 933–941
11. Defossez A, Synnaeve G, Adi Y (2020) Real time speech enhancement in the waveform domain. arXiv:2006.12847
12. Duan X, Sun Y, Wang J (2023) Eca-unet for coronary artery segmentation and three-dimensional reconstruction. Signal Image Video Process 17(3):783–789
13. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44(3):572–587
14. Erdogan H, Hershey JR, Watanabe S et al (2015) Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. 2015 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 708–712
15. Eskimez SE, Wang X, Tang M, et al (2021) Human listening and live captioning: multi-task training for speech enhancement. arXiv:2106.02896
16. Fayek HM, Lech M, Cavedon L (2017) Evaluating deep learning architectures for speech emotion recognition. Neural Netw 92:60–68
17. Fu SW, Tsao Y, Lu X, et al (2016) Snr-aware convolutional neural network modeling for speech enhancement. In: Interspeech, pp 3768–3772
18. Fu SW, Hu Ty, Tsao Y, et al (2017) Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP), IEEE, pp 1–6
19. Fu SW, Liao CF, Tsao Y, et al (2019) Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In: International Conference on Machine Learning, PMLR, pp 2031–2041
20. Fu Y, Liu Y, Li J et al (2022) Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation. ICASSP 2022–2022 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 7417–7421
21. Fuchs A, Priewald R, Pernkopf F (2019) Recurrent dilated densenets for a time-series segmentation task. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, pp 75–80
22. Giri R, Isik U, Krishnaswamy A (2019) Attention wave-u-net for speech enhancement. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, pp 249–253
23. Grais EM, Plumbley MD (2017) Single channel audio source separation using convolutional denoising autoencoders. In: 2017 IEEE global conference on signal and information processing (GlobalSIP), IEEE, pp 1265–1269
24. Gulati A, Qin J, Chiu CC, et al (2020) Conformer: Convolution-augmented transformer for speech recognition. arXiv:2005.08100
25. Han K, Wang Y, Wang D (2014) Learning spectral mapping for speech dereverberation. 2014 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 4628–4632
26. Hao X, Su X, Wen S et al (2020) Masking and inpainting: a two-stage speech enhancement approach for low snr and non-stationary noise. ICASSP 2020–2020 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 6959–6963
27. Harsh H, Indraganti A, Vanambathina SD, et al (2022) Convolutional gru networks based singing voice separation. In: 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), IEEE, pp 1–5

28. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
29. Hu H, Yu C (2007) Adaptive noise spectral estimation for spectral subtraction speech enhancement. IET Signal Process 1(3):156–163
30. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
31. Hu Y, Loizou PC (2007) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Language Process 16(1):229–238
32. Hu Y, Liu Y, Lv S, et al (2020) Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. arXiv:2008.00264
33. Huang PS, Kim M, Hasegawa-Johnson M et al (2014) Deep learning for monaural speech separation. 2014 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 1562–1566
34. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pmlr, pp 448–456
35. ITU-T P (2003) 835: subjective test methodology for evaluating speech communication systems that include noise suppression algorithms. ITU-T recommendation
36. Jannu C, Vanambathina SD (2023a) An attention based densely connected u-net with convolutional gru for speech enhancement. In: 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), IEEE, pp 1–5
37. Jannu C, Vanambathina SD (2023b) Convolutional transformer based local and global feature learning for speech enhancement. Int J Advan Comput Sci Appl 14(1)
38. Jannu C, Vanambathina SD (2023) Multi-stage progressive learning-based speech enhancement using time-frequency attentive squeezed temporal convolutional networks. Circuits Syst Signal Process 42(12):7467–7493
39. Jannu C, Vanambathina SD (2023d) An overview of speech enhancement based on deep learning techniques. Int J Image Graphics:2550001
40. Jannu C, Vanambathina SD (2023) Weibull and nakagami speech priors based regularized nmf with adaptive wiener filter for speech enhancement. Int J Speech Technol 26(1):197–209
41. Jansson A, Humphrey E, Montecchio N, et al (2017) Singing voice separation with deep u-net convolutional networks. ISMIR Conference
42. Kamath S, Loizou P, et al (2002) A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: ICASSP, Citeseer, pp 44164–44164
43. Kim Y, Lee H, Provost EM (2013) Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE, pp 3687–3691
44. Kishore V, Tiwari N, Paramasivam P (2020) Improved speech enhancement using tcn with multiple encoder-decoder layers. In: Interspeech, pp 4531–4535
45. Koizumi Y, Yatabe K, Delcroix M et al (2020) Speech enhancement using self-adaptation and multi-head self-attention. ICASSP 2020–2020 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 181–185
46. Krawczyk M, Gerkmann T (2014) Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Trans Audio Speech Language Process 22(12):1931–1940
47. Kulmer J, Mowlaee P (2014) Phase estimation in single channel speech enhancement using phase decomposition. IEEE Signal Process Lett 22(5):598–602
48. Kumar A, Daume III H (2012) Learning task grouping and overlap in multi-task learning. arXiv:1206.6417
49. Kumar K, Cruces S et al (2017) An iterative posterior nmf method for speech enhancement in the presence of additive gaussian noise. Neurocomputing 230:312–315
50. Lalitha V, Prema P, Mathew L (2010) A kepstrum based approach for enhancement of dysarthric speech. In: 2010 3rd International Congress on Image and Signal Processing, IEEE, pp 3474–3478
51. Le X, Lei T, Chen K et al (2022) Inference skipping for more efficient real-time speech enhancement with parallel rnns. IEEE/ACM Trans Audio Speech Language Process 30:2411–2421
52. Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. Proc IEEE 67(12):1586–1604
53. Lin J, van Wijngaarden AJdL, Wang KC et al (2021) Speech enhancement using multi-stage self-attentive temporal convolutional networks. IEEE/ACM Trans Audio Speech Language Process 29:3440–3450
54. Liu JY, Yang YH (2019) Dilated convolution with dilated gru for music source separation. arXiv:1906.01203
55. Lu X, Tsao Y, Matsuda S, et al (2013) Speech enhancement based on deep denoising autoencoder. In: Interspeech, pp 436–440

56. Macartney C, Weyde T (2018) Improved speech enhancement with the wave-u-net. arXiv:1811.11307
57. Mahdavi R, Bastanfard A, Amirkhani D (2020) Persian accents identification using modeling of speech articulatory features. 2020 25th international computer conference. Computer Society of Iran (CSICC), IEEE, pp 1–9
58. Mehrish A, Majumder N, Bharadwaj R, et al (2023) A review of deep learning techniques for speech processing. Inform Fusion:101869
59. Michelsanti D (2021) Audio-visual speech enhancement based on deep learning. Aalborg Universitet
60. Michelsanti D, Tan ZH, Zhang SX et al (2021) An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans Audio Speech Language Process 29:1368–1396
61. Naithani G, Barker T, Parascandolo G, et al (2017) Low latency sound source separation using convolutional recurrent neural networks. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, pp 71–75
62. Neri J, Braun S (2023) Towards real-time single-channel speech separation in noisy and reverberant environments. ICASSP 2023–2023 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 1–5
63. Noizeus (2007) https://ecs.utdallas.edu/loizou/speech/noizeus
64. Van den Oord A, Kalchbrenner N, Espeholt L, et al (2016) Conditional image generation with pixelcnn decoders. Advan Neural Inform Process Syst 29
65. Oord Avd, Dieleman S, Zen H, et al (2016) Wavenet: A generative model for raw audio. arXiv:1609.03499
66. Oostermeijer K, Du J, Wang Q et al (2021) Speech enhancement autoencoder with hierarchical latent structure. ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 671–675
67. Paliwal K, Wójcicki K, Shannon B (2011) The importance of phase in speech enhancement. Speech Commun 53(4):465–494
68. Parisae V, Bhavanam SN (2024) Adaptive attention mechanism for single channel speech enhancement. Multimed Tool Appl:1–26
69. Pascual S, Bonafonte A, Serra J (2017) Segan: Speech enhancement generative adversarial network. arXiv:1703.09452
70. Rautaray SS, Agrawal A (2015) Vision based hand gesture recognition for human computer interaction: a survey. Artif Intell Rev 43:1–54
71. Rim Park S, Lee J (2016) A fully convolutional neural network for speech enhancement. pp arXiv–1609
72. Saeidi R, Mowlaee P, Martin R (2012) Phase estimation for signal reconstruction in single-channel source separation. Interspeech
73. Savargiv M, Bastanfard A (2016) Real-time speech emotion recognition by minimum number of features. In: 2016 Artificial Intelligence and Robotics (IRANOPEN), IEEE, pp 72–76
74. Scalart P, et al (1996) Speech enhancement based on a priori signal to noise estimation. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, IEEE, pp 629–632
75. Shahnawazuddin S, Deepak K, Pradhan G et al (2017) Enhancing noise and pitch robustness of children's asr. 2017 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 5225–5229
76. Shriberg LD, Paul R, McSweeny JL, et al (2001) Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome. Journal of Speech, Language, and Hearing Research
77. Sim BL, Tong YC, Chang JS et al (1998) A parametric formulation of the generalized spectral subtraction method. IEEE Trans Speech Audio Process 6(4):328–337
78. Soni MH, Shah N, Patil HA (2018) Time-frequency masking-based speech enhancement using generative adversarial network. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5039–5043
79. Srivastava S, Bisht A, Narayan N (2017) Safety and security in smart cities using artificial intelligence–a review. 2017 7th International Conference on Cloud Computing. Data Science & Engineering-Confluence, IEEE, pp 130–133
80. Stoller D, Ewert S, Dixon S (2018) Wave-u-net: a multi-scale neural network for end-to-end audio source separation. arXiv:1806.03185
81. Sunnydayal V, Kumar TK (2016) Speech enhancement using β-divergence based nmf with update bases. 2016 International Conference on Microelectronics. Computing and Communications (MicroCom), IEEE, pp 1–6
82. Sunnydayal V et al (2017) Speech enhancement using posterior regularized nmf with bases update. Comput Electrical Eng 62:663–675

83. Taal CH, Hendriks RC, Heusdens R et al (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans Audio Speech Language Process 19(7):2125–2136
84. Takahashi N, Mitsufuji Y (2017) Multi-scale multi-band densenets for audio source separation. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, pp 21–25
85. Takahashi N, Goswami N, Mitsufuji Y (2018) Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In: 2018 16th International workshop on acoustic signal enhancement (IWAENC), IEEE, pp 106–110
86. Tan K, Wang D (2018) A convolutional recurrent neural network for real-time speech enhancement. In: Interspeech, pp 3229–3233
87. Tan K, Wang D (2019) Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. IEEE/ACM Trans Audio Speech Language Process 28:380–390
88. Tan K, Chen J, Wang D (2018) Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM Trans Audio Speech Language Process 27(1):189–198
89. Tompson JJ, Jain A, LeCun Y, et al (2014) Joint training of a convolutional network and a graphical model for human pose estimation. Advan Neural Inform Process Syst 27
90. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1653–1660
91. Valentini-Botinhao C, Wang X, Takaki S, et al (2016) Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In: SSW, pp 146–152
92. Vanambathina S, Kumar TK (2016) Speech enhancement by bayesian estimation of clean speech modeled as super gaussian given a priori knowledge of phase. Speech Commun 77:8–27
93. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advan Neural Inform Process Syst 30
94. Wang D, Brown GJ (2006) Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press
95. Wang D, Lim J (1982) The unimportance of phase in speech enhancement. IEEE Trans Acoustics Speech Signal Process 30(4):679–681
96. Wang Q, Wu B, Zhu P, et al (2020) Eca-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11534–11542
97. Wang Y, Wang D (2015) A deep neural network for time-domain signal reconstruction. 2015 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 4390–4394
98. Wang Y, Narayanan A, Wang D (2014) On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Language Process 22(12):1849–1858
99. Weninger F, Eyben F, Schuller B (2014a) Single-channel speech separation with memory-enhanced recurrent neural networks. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 3709–3713
100. Weninger F, Hershey JR, Le Roux J, et al (2014b) Discriminatively trained recurrent neural networks for single-channel speech separation. In: 2014 IEEE global conference on signal and information processing (GlobalSIP), IEEE, pp 577–581
101. Williamson DS, Wang Y, Wang D (2015) Complex ratio masking for monaural speech separation. IEEE/ACM Trans Audio Speech Language Process 24(3):483–492
102. Xiang X, Zhang X, Chen H (2021) A nested u-net with self-attention and dense connectivity for monaural speech enhancement. IEEE Signal Process Lett 29:105–109
103. Xu Y, Du J, Dai LR et al (2013) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68
104. Yang Y, Fairbairn C, Cohn JF (2012) Detecting depression severity from vocal prosody. IEEE Trans Affective Comput 4(2):142–150
105. Yechuri S, Vanabathina SD (2023) Genetic algorithm-based adaptive wiener gain for speech enhancement using an iterative posterior nmf. Int J Image Graph 23(06):2350054
106. Zhang Q, Nicolson A, Wang M, et al (2019) Monaural speech enhancement using a multi-branch temporal convolutional network. arXiv:1912.12023
107. Zhang Y, Yang Q (2018) An overview of multi-task learning. National Sci Rev 5(1):30–43
108. Zhao H, Zarar S, Tashev I et al (2018) Convolutional-recurrent neural networks for speech enhancement. 2018 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 2401–2405
109. Zhao S, Nguyen TH, Ma B (2021) Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses. ICASSP 2021–2021 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 6648–6652

110. Zhao S, Ma B, Watcharasupat KN et al (2022) Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement. ICASSP 2022–2022 IEEE International Conference on Acoustics. Speech and Signal Processing (ICASSP), IEEE, pp 9281–9285

Authors and Affiliations

Manaswini Burra[1] · Sunny Dayal Vanambathina[2] · Venkata Adi Lakshmi A[3] · Loukya Ch[3] · Siva Kotiah N[3]

✉ Sunny Dayal Vanambathina
sunny.dayal@vitap.ac.in

Manaswini Burra
manaswini.burra@gmail.com

Venkata Adi Lakshmi A
adilakshmiakurathi2901@gmail.com

Loukya Ch
chinthaloukya@gmail.com

Siva Kotiah N
s8008512139@gmail.com

1 Department of CSE, Potti Sriramulu Chalavadi Mallikarjuna Rao College of Engineering & Technology(Autonomous), Vijayawada 520001, Andhra Pradesh, India

2 School of Electronics Engineering, VIT-AP University, Beside AP Secretariat, Amaravati 522 237, Andhra Pradesh, India

3 Computer Science and Engineering Department, Lakireddy Balireddy College of Engineering, Myalavaram, Andhra Pradesh, India