Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.028090
Article

Tech Science Press

Real-Time Speech Enhancement Based on Convolutional Recurrent Neural Network

S. Girirajan and A. Pandian*

Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Engineering, Kattankulathur, Tamil Nadu, India
*Corresponding Author: A. Pandian. Email: pandiana@srmist.edu.in
Received: 02 February 2022; Accepted: 01 April 2022

Abstract: Speech enhancement is the task of taking a noisy speech input and producing an enhanced speech output. In recent years, the need for speech enhancement has been increased due to challenges that occurred in various applications such as hearing aids, Automatic Speech Recognition (ASR), and mobile speech communication systems. Most of the Speech Enhancement research work has been carried out for English, Chinese, and other European languages. Only a few research works involve speech enhancement in Indian regional Languages. In this paper, we propose a two-fold architecture to perform speech enhancement for Tamil speech signal based on convolutional recurrent neural network (CRN) that addresses the speech enhancement in a real-time single channel or track of sound created by the speaker. In the first stage mask based long short-term memory (LSTM) is used for noise suppression along with loss function and in the second stage, Convolutional Encoder-Decoder (CED) is used for speech restoration. The proposed model is evaluated on various speaker and noisy environments like Babble noise, car noise, and white Gaussian noise. The proposed CRN model improves speech quality by 0.1 points when compared with the LSTM base model and also CRN requires fewer parameters for training. The performance of the proposed model is outstanding even in low Signal to Noise Ratio (SNR).

Keywords: Speech enhancement; convolutional encoder-decoder; long short-term memory; noise suppression; speech restoration

1 Introduction

Speech enhancement plays an important role in speech processing applications and voice communication by separating speech and non-speech noise [1]. In recent years, research interest in speech enhancement has been increased consistently to address the challenges that have occurred in robust speech recognition, mobile speech communication, and hiring aids. Most of the speech enhancement algorithms work in the short-time Fourier transform (STFT) domain along with the weighting rule. To compute the weighting rule, the signal-to-noise ratio (SNR) with noise power is used. Various algorithms are available to estimate the SNR [2] and noise power [3], in which a frequently used technique is assuming noise to be more stationary than speech. This technique does not provide good

performance since this assumption cannot handle highly non-stationary noise types like babble noise or restaurant noise. Mask-learning [4] and feature-learning [5] are two major approaches available for enhancing the single-channel speech over the data that is collected through the stereo. Mask-learning outperforms when compared with feature-learning due to its fastness and dynamic range. The masking process removes the additive noise by assuming the scale of the masked signal is the same as the clean speech signal. Mask learning is carried out by two approximations, the distance between the target mask and the learning mask is minimised directly by mask approximation (MA) [6]. Similarly, the distance between the target signal and the distorted signal was minimized by signal approximation (SA) [7].

In [8], the researcher proposed a Smart Larynx (SL) device for vocal cord affected patients to increase the clarity of their Tamil speech. The SL device contains a smartphone with an inbuilt vibrator application along with a frequency ranging between 250 to 450 Hz. The Radial Dilation Wavelet Transformation (RDWT) algorithm is used for processing the Tamil speech signal. This proposed work achieves nearly 85% accuracy in obtaining clean speech signals from noisy environments.

Deep learning algorithms show better performance when compared with machine learning and other regular methodologies in most fields, including automatic speech recognition, image processing, computer vision, and speech enhancement [9] and researchers used phase spectrum approximation (PSA) loss to differentiate the noisy and clean speech signals. PSA outperformed when compared with MSA. Most of the models used for speech enhancement belong to the feedforward Deep Neural Network (DNN). The performance of speech enhancement was increased by using deep learning since no assumption on the stationarity of noise and loss function was made in the training phase along with the topology of the neural network. Deep learning usually predicts the label from each frame, but such a method does not hold on to long-term contexts. In the paper [10,11], the researcher suggested using sequence-to-sequence mapping for speech separation to strengthen the long-term context. In [12], the researcher shows better performance by using a large number of different noise types for multi-condition training that directly maps the noisy features with their corresponding clean speech features. In [13], the researcher concludes that speech separation can be done efficiently by estimating time-frequency (T-F) based on an ideal ratio mask (IRM) rather than estimating the clean spectrogram directly. Mask spectrum approximation (MSA) loss in the domain of the speech spectrum In [14], the researcher used Long Short-Term Memory (LSTM) that contains temporal dynamics, which shows better performance in speech separation tasks. By using LSTM, temporal dependencies are taken into account and more focused on the target speaker, which provides better speaker generalization. Similarly, the Convolutional Neural Network (CNN) model is also widely used in speech enhancement tasks with an increase in performance. Due to its encoder-decoder architecture, CNN is a popular model in the fields of image processing and computer vision due to its encoder-decoder architecture. But this CNN model uses the pooling layer for compressing the feature dimensions alone. The encoder compresses the features and the decoder will decompress the features by using up-sampling layers [15]. With the help of skip connection, high-resolution information is conserved by adding the same number of layers of the encoder to the decoder in a task like speech enhancement. That learns the clean speech spectrum by mapping it with the noisy speech spectrum.

Each network topology will have its own advantages over speech enhancement. To leverage the advantage of this different topology is to combine it into the multi-stage identical model. With such a formulation, CNN and Recurrent Neural Network (RNN) have been combined for noise and speaker-independent speech enhancement. In [16,17], the researchers proposed Recurrent Convolutional Encoder-Decoder (R-CED) that incorporates repeated convolution layers with ReLU as an activation layer to reduce the noise signal. In [18], a multi-stage feedforward DNN model is used for separation and enhancing the separated signal that is obtained from the music source. In [19], the researcher proposed a speaker-independent model that contains LSTM with 4 hidden layers to handle the noise. The proposed

model shows better performance for untrained speakers than DNN and also performs well on short-time objective intelligibility. To distinguish between noise and clean speech signals, a model that can handle long-term temporal dependencies is required.LSTM is widely used in image restoration since it can handle long-term temporal dependency. With its success in image restoration, LSTM is also used in speech enhancement. In [20], researchers estimated the clean speech along with the noise linear prediction coefficient using the Deep Neural network-based LSTM model. In this work, residual background noise is reduced by applying post-processing. LSTM perform well in speech enhancement task but it is highly difficult to implement due to its complex network. To overcome the above drawbacks of LSTM, two models were proposed recently for Speech Enhancement namely Gated Recurrent Unit (GRU) and Single Gated Unit (SGU) [21,22]. The GRU and SGU are easy to implement when compared with LSTM but the performance of SGU and GRU is not up to the level in speech enhancement application. In [23], the researcher introduces a new methodology to reduce the risk of over-pruning by evaluating the pruning and fine-tuning in each iteration using short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) metrics.

We designed a new CRN architecture for real-time speech improvement based on motivation from existing work [24]. When compared to speech enhancement created with the LSTM model alone, the suggested model integrates the CNN encoder-decoder gives superior performance. To assess the noise power spectral density, researchers used a model that included deep minimum mean-square error and a ResNet Temporal Convolutional network [25]. Using a Temporal Convolutional Neural Network (TCNN), a noisy speech input was directly translated to a clean speech signal. The researcher presented a methodology for compressing the RNN using pruning and integer quantization [26]. This decreases the size of the RNN by 38% while decreasing the SNR.

The rest of the paper is organized as follows: in Section 2, the baseline model and the proposed model is described in detail. In Section 3, experimental setup and evaluation results are presented followed by that in Section 4, conclusion and future work are given.

2 Traditional and Deep Neural Network Based Architecture for Speech Enhancement

The signal model for estimating the clean speech signal from the noisy signal is given in Eq. (1).

$$y(n) = s_{cs}(n) + d_{ns}(n) \tag{1}$$

Where, $s_{cs}(n)$ denotes the clean speech signal, $d_{ns}(n)$ denotes the noisy speech signal with discrete-time sample index n and y(n) denotes the noisy microphone signal. Similarly, the window function is applied frame-wise to compute the STFT representation by using k-point Discrete Fourier Transform (DFT) as shown below in Eq. (2).

$$y_l(m) = s_{(cs)l}(m) + d_{(ns)l}(m) \tag{2}$$

where, l denotes the frame length and m denotes the frequency bin index. Mostly frame and frequency bin-wise gain functions are used in traditional and DNN approaches to compute the clean speech as shown in below Eq. (3)

$$\hat{s}_{(cs)l}(m) = Q_l(m).\, y_l(m) \tag{3}$$

where, the gain function is denoted by $Q_l(m)$.

2.1 Traditional Approaches

In the traditional approach, gain function depends on prior and posterior signal-to-noise ratio (SNR) as shown in Eq. (4). In [27,28], the researcher used minimum mean-square error log spectrum amplitude with

the decision-directed approach for estimating the prior SNR along with its posterior SNR obtained by using minimum statistics (MS) for estimating the noise power.

$$Q_l(m) = q(\xi_l(m).\ \gamma_l(m)) \tag{4}$$

where, $\xi_l(m)$ denotes the prior and $\gamma_l(m)$ denotes the posterior SNR.

2.2 Deep Learning Approaches

In deep learning-based approaches, initially, neural networks are trained by mapping the input feature vector with the output feature vector to perform the speech enhancement task with the help of the Activation function present in each neuron and parameters that are required for training the network topology. In the case of the recurrent neural network, a hidden state will be given an additional input h_{l-1} to compute the temporal context in each frame.

$$u_l = f(x_l, h_{l-1}; \Theta) \tag{5}$$

Gain function is often used for separating the clean speech from the noisy signal as a T-F mask. The masks can be improvised by reducing the MA loss function as shown in Eq. (6). The Mask Approximation loss function doesn't directly maximize the goal of decreasing the variance and clear speech spectrums. Estimating the loss between each frequency bin is depends on the clean speech signal and noisy signal. Based on this mask value can be estimated up to a certain range alone, to overcome this mask can be estimated by using masked spectrum approximation loss function as shown in Eq. (7)

$$P_l^{MA}(\Theta) = \frac{1}{m}\sum\nolimits_{m\epsilon \mathrm{M}} \left(\hat{Q}_l(m) - Q_l^{ideal}(m)\right)^2 \tag{6}$$

$$P_l^{MSA}(\Theta) = \frac{1}{m}\sum\nolimits_{m\epsilon \mathrm{M}} \left(\hat{Q}_l(m).|y_l(m)| - \left|s_{(cs)l}(m)\right|\right)^2 \tag{7}$$

In (6) and (7), the loss function is computed based on spectral magnitude by using noisy speech. Similarly, loss function can be computed based on clean speech using the ideal complex mask. Those two-loss function has been combined to form a complex cMSA loss function as shown in Eq. (8), which provides better performance over speech enhancement.

$$P_l^{cMSA}(\Theta) = \frac{1}{m}\left(\sum_{m=0}^{m/2}\left(\hat{Q}_l^R(m).real|y_l(m)| - img\left|s_{(cs)l}(m)\right|\right)^2 + \sum_{m=1}^{\frac{m}{2}-1}\left(\hat{Q}_l^I(m).real|y_l(m)| - img\left|s_{(cs)l}(m)\right|\right)^2\right) \tag{8}$$

Using Eq. (9) enhanced signal is calculated by applying cMSA loss function in neural network

$$\hat{s}_{(cs)l}(m) = Q_l^R(m).real\{y_l(m)\} + a\hat{Q}_l^I(m).img\{y_l(m)\} \tag{9}$$

Using Eq. (10) real and imaginary parts of the clean speech spectrum can be estimated directly by applying cSA.

$$P_l^{cSA}(\Theta) = \frac{1}{m}\left(\sum_{m=0}^{m/2}\left(\hat{s}_l^R(m) - real|s_l(m)|\right)^2 + \sum_{m=1}^{\frac{m}{2}-1}\left(\hat{s}_l^I(m) - img\left|s_{(cs)l}(m)\right|\right)^2\right) \tag{10}$$

2.3 Convolutional Recurrent Neural Network

In the proposed model, speech denoising and restoration are carried out in two different stages. In the first, speech denoising is performed by training the LSTM with cMSA loss function. In the second stage, a Convolutional encoder-decoder is used for restoring clean speech obtained from the denoising stage. CED

performs well in restoring the slightly corrupted structured signal. In stage two direct spectral mapping is done since cSA loss function is used to train the CED network. The reason for using cSA loss function is restoring will be done more efficiently in missing T-F regions when compared with the mask-based approach. Similarly mapping the output of the same domain as the input is possible in the CED network.

3 System Descriptions

The proposed model is based on two-fold architecture, in the first stage, raw speech with noise will be given as an input for noise suppression. Normalized feature vectors are obtained during feature extraction by using mean and variance normalization (MVN) [29]. Similarly to extract the current frame, the context of previous and next frames are concatenated during feature extraction. Separate real value masks, as well as real and imaginary noisy speech spectrum, are estimated using input features and parameters that are obtained during the training phase of the noise suppression network. These mask values are applied in Eq. (8) to obtain the clean speech spectrum. Fig. 1 shows the architecture of the proposed model.

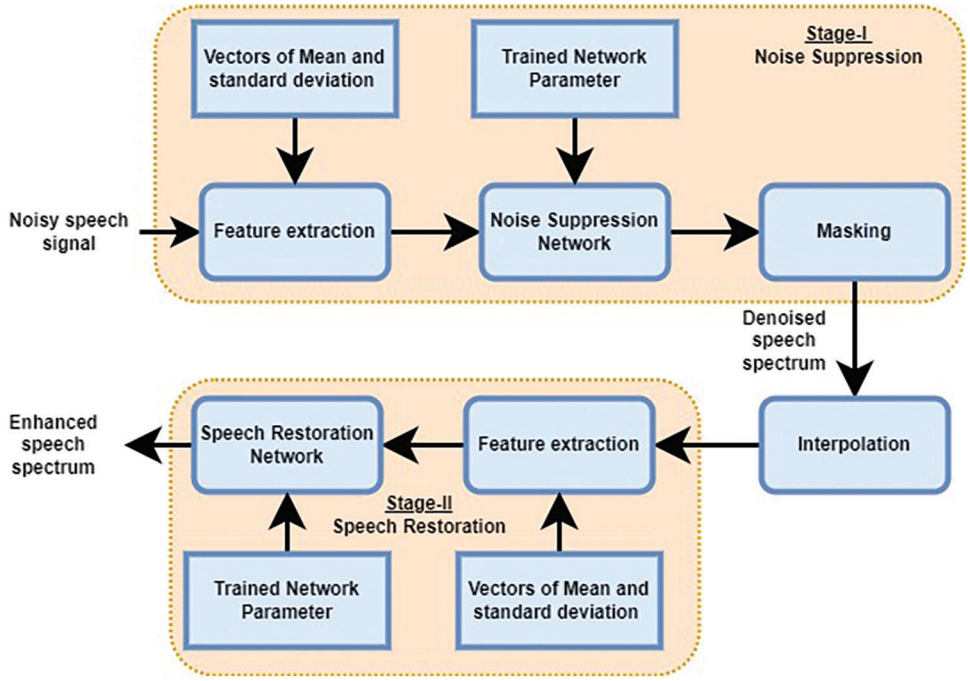

Figure 1: Proposed two-fold architecture for speech suppression and restoration

In between the Noise suppression and Speech restoration stage, the frequency resolution of the speech signal is increased for mapping CED network with high-resolution spectrum to obtain the clean speech signal by using interpolation. Interpolation controls the spectral leakage from negative frequency to estimate the significant errors when the signal contains small cycles. Such interpolation works by applying Inverse Discrete Fourier Transform (IDFT) followed by zero padding in the time domain and respective modification is made to the frequency domain as well.

After interpolation de-noised speech signal is processed to the speech restoration stage. This stage also contains feature extraction using MVN normalization along with standard deviation. Later extracted features are directly mapped with enhanced speech spectrum along with trained parameters. Finally enhanced time-domain signals are reconstructed by using IDFT and windowing [30].

3.1 LSTM Based Noise Suppression

In the First stage noise suppression is carried out using recurrent neural network type LSTM as shown in Fig. 2. LSTM performs well in the long-term context that helps in tracking the target speaker. LSTM [31] is a type of RNN that contains memory cells that shows a successful performance in temporal modeling in the field of the acoustic model over Automatic speech recognition. All RNNs have a chain of reiterating neural network modules. Likewise, LSTM has an equivalent structure, yet the repeating module has a substitute structure. There are four interfacing in an excellent way rather than having a singular neural system layer as shown in Fig. 3.

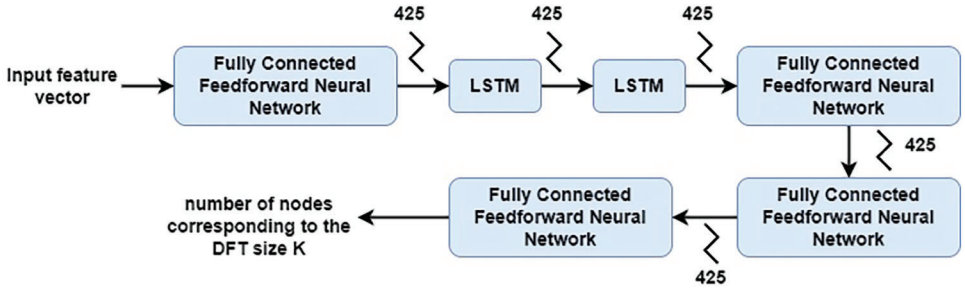

Figure 2: LSTM based noise suppression Network

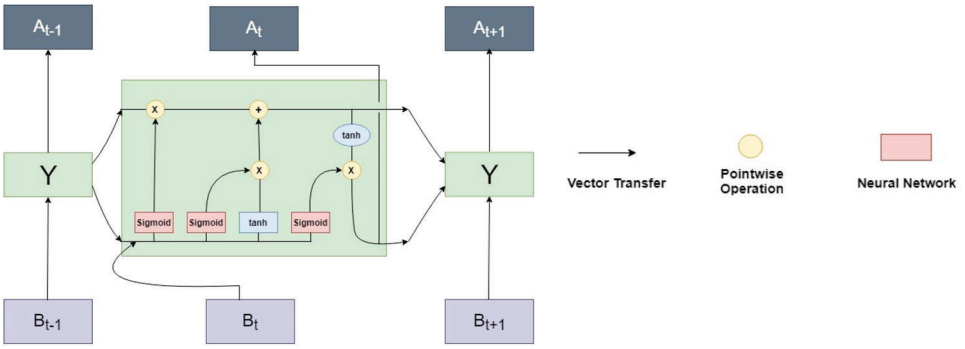

Figure 3: LSTM with four interacting layers

Steps involved in LSTM:

Step 1: Sigmoid layer in forgot gate layer chooses what data the cell state has to discard.

$$f_t = \sigma\left(W_f \cdot [A_{t-1}\ ,\ B_t]\ +\ x_f\right) \tag{11}$$

Step 2: Combine also gets fed into the input layer. This layer decides what data from the candidate should be added to the new cell state as shown in Eq. (9).

$$i_t = \sigma(W_i \cdot [A_{t-1}\ ,\ B_t]\ +\ x_i) \tag{12}$$

$$\tilde{C}_t = tanh\left(W_C \cdot \left[A_{(t-1)}, B_t\right] + x_C\right) \tag{13}$$

$$C_t = f_t * C_{t-1}\ +\ i_t * \tilde{C}_t \tag{14}$$

$$O_t = \sigma(W_0 * (A_{t-1}\ , B_t) + x_0) \tag{15}$$

$$A_t = O_t * \tanh(C_t) \tag{16}$$

A_{t-1} represents the output of the previous cell (or) LSTM, B_t input at that particular time, C_{t-1}, C_t, $\tilde{C}_t$ represents old cell state, new cell state and new candidate value, f_t forgot gate state, O_t output gate, it input gate, σ sigmoid function, x bias for the respective gate, W weight for the respective gate

In stage 1 for noise suppression, feature vectors that are extracted after MVN normalization are given as the input. The size of the feature vector is based on the previous and next frame size with a constant value 1.

$$C = (N^- + N^+ + 1).\left(\frac{m}{2} + 1\right) \tag{17}$$

where, N^- and N^+ denotes the previous and next frame features that can be extracted to create a current frame. In some situations, the next frame can also be considered as zero.

Initially, noise suppression consists of one fully connected feed-forward neural network that helps to identify the efficient features then forward it to the LSTM layers. Finally, three Feed Forward layers are attached with LSTM to estimate the T-F mask. Each feed-forward layer consists of 425 nodes with the rectified linear unit (ReLU) as an activation function.

3.2 Speech Restoration Using CED

In the proposed work, CED is used for speech restoration with the architecture shown in Fig. 4. Each layer is represented with frequency, frame axis and feature maps sizes. The output that is obtained from speech denoising is given as the input for speech restoration stage. Normalized feature are obtained by using separated feature maps for the real and imaginary parts.

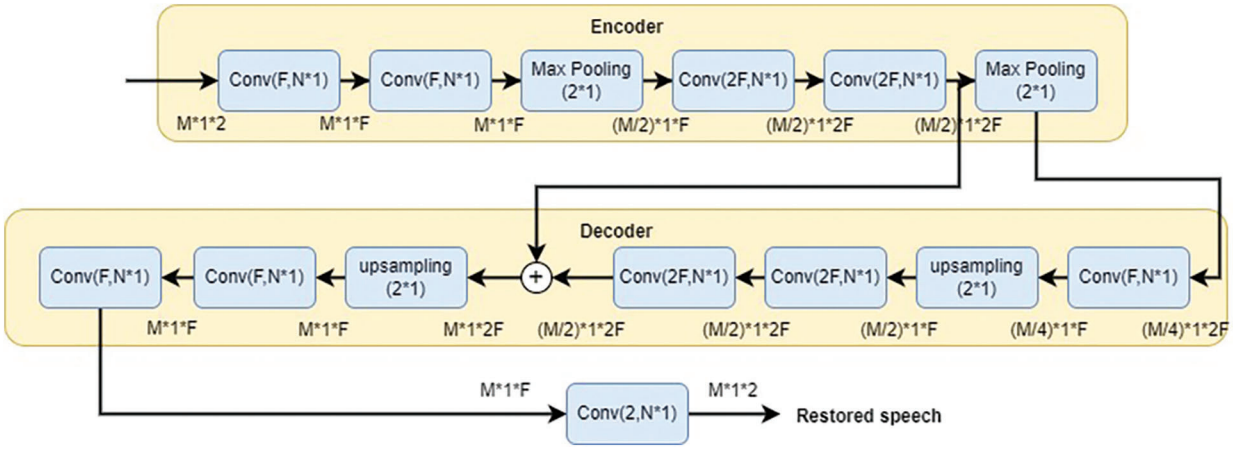

Figure 4: CED Architecture for speech restoration

The feature maps that are obtained should match the real and imaginary parts of the CED kernel in the first network layer. Frequency axis size is obtained by applying zero padding and multiplying by 4 to make the total dimension factor as 4 for the encoder. Similarly, decoder features will also be reconstructed to match the factor of the encoder [32–35]. After replacing the noise that is obtained from the speech de-noising stage, clean speech is estimated. Later without applying any algorithm delay or any other information in the future frame, speech restoration is generated. In speech restoration, the convolutional layer is denoted *Conv* with filter kernel F, frequency axis size N, and constant 1 is added. On the other hand, transposed convolutional layers are denoted by $Conv^T$. Both Convolutional layer and transpose convolutional layers

use *ReLu* as an Activation function with zero padding to map the input feature vector. The Convolutional and maximum pooling layers are used to reduce the feature size concerning the frequency axis in the encoder part. Similarly, Transpose Convolutional and upsampling layers are used to increase the frequency axis in the decoder part [36,37].

3.3 Data Preprocessing and Training the Models

3.3.1 Data Set

Audacity 3.0.2 is used to record the Tamil speech signal from various native Tamil speakers. Audacity is open-source software. It is a basic audio editor which can trim, copy, record, and manipulate sounds. It can be used to adjust the speed and pitch of the audio, or add an equaliser to it. For experimental purposes, a total of 83 speakers (42 male and 41 female) from different age groups were selected. For recording purposes, the WO Mic client interface was used to connect with Audacity.

In total, 7138 utterances of speech signals were collected. Each speaker recorded 100 samples of Tamil sentences. From the entire data set, 60% of recorded samples are used for training, 20% of samples are used for development, and 20% of samples are used for testing purposes. Overlapping between the speakers is avoided. The parameters that are considered for recording speech signals are listed in Tab. 1 shown below.

Table 1: Parameters considered for recording speech signal

Parameters	values
Sampling frequency	16 kHz
Coding technique	PCM
Recording mode	Mono
Bit rate	16 bits/s

The samples that are collected from various speakers are mixed with common model-independent noises that are downloaded from https://www.sound-ideas.com. For testing purposes, babble and cafeteria noises were used that are downloaded from https://www.auditec.com. The entire dataset contains 500 hrs of the mixed speech signal. Random utterances were selected from the clean speech and mixed with random noise based on the signal-to-noise ratio (SNR). Figs. 5 and 6 show the Speech signal waveform in Audacity. After mixing the noise with the recorded sample SNR value is set between 0 to 15 db with an equal interval of around 5 db. Speech to noise ratio is obtained by using effects in Audacity. Figs. 7 and 8 show the various noises that we are used in experiments. A separate window is used to obtain the noisy data by adding noise data with collected speech samples. The frame length is set to 256 and frameshift is set to 128 along with time-domain signals to compute the input feature and target for the proposed model. The size of DFT is set to 256 for training and evaluation.

3.3.2 Training LSTM and CNN Based CED for Noise Suppression and Restoration

Back-propagation through time (BPTT) is used for training the LSTM model for noise suppression along with cMSA loss function and Adam optimizer. The parameters like batch size and learning rate are set to 25 and 0.001 respectively for Adam optimizer. Overfitting is avoided by setting up the weight decay as 0.0002. Speech utterance is set to fixed-length size say 100 for training in BPTT. Speech utterances that are less than 100 sizes are zero-padded to match the size. In each epoch, the loss is calculated and if there is no change in the loss for the continuous three epochs then the particular loss is considered as least and the learning rate will be updated accordingly based on the development loss. During the experiment, we reached a learning rate up to 0.0001 which is considered as a minimum.

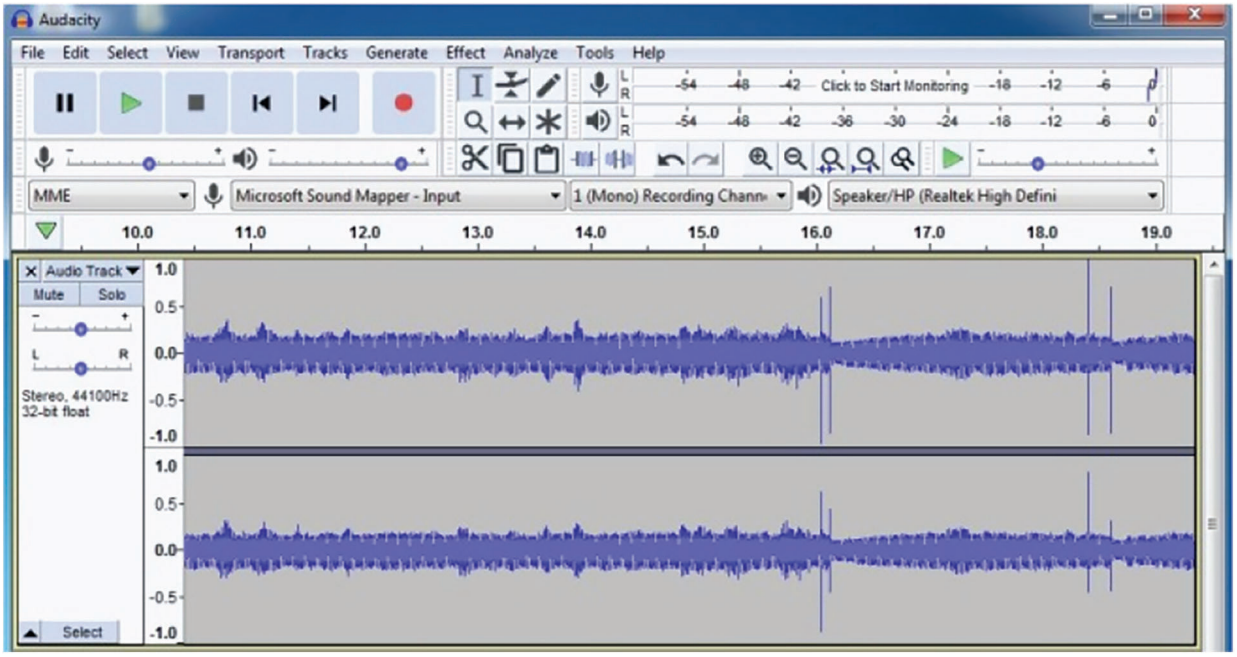

Figure 5: Speech signal in audacity

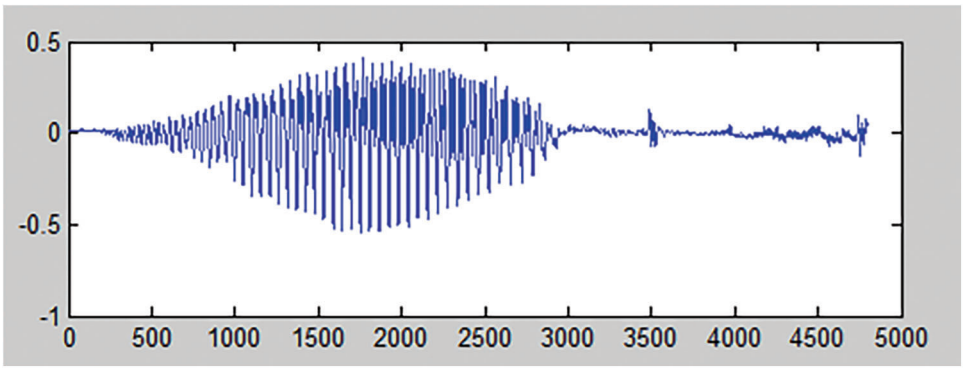

Figure 6: Speech signal

In Speech restoration, back-propagation is used for training the CED network along with cSA as a loss function. The Parameters like batch size and learning rate are set to the same values that are used in LSTM training for Adam optimizer. In each epoch, the loss is calculated and if there is no change in the loss for continuous two epochs then the particular loss is considered as least and the learning rate will be updated accordingly based on the development loss. In the proposed work, 88 filter kernels and 24 as the size of frequency axis are used for training the CED network.

4 Experimental Results

In the proposed work, we used PESQ and STOI as evaluation metrics and compared them with the traditional models. Based on the experimental result, the proposed deep learning model can attain 0.88 for unprocessed speech signal as an STOI value. On the other hand, traditional models can able attain the average of 0.75 as the STOI value. Similarly, the PESQ value also increased considerably while

using Deep learning models. Tabs. 2 and 3 show the STOI percentage for processed and unprocessed noise speech signals based on the trained and untrained speaker. Tabs. 4 and 5 show the PESQ for processed and unprocessed noise speech signals based on the trained and untrained speaker. Fig. 9 depicts the comparison of the model based on the evaluation metrics. In the low Speech to Noise ratio condition, the deep learning model performs well in stage 1 noise suppression. LSTM based on cMSA loss function shows outstanding performance when compared with other traditional approaches like LSTM-MSA and LSTM-IRM. Since LSTM-cMSA processes the information by making a clean speech spectrum into two separate parts like real and imaginary. In stage 2 speech restorations, the CED network is used along with cSA and perform well in high SNR condition as well. The proposed two-fold architecture noise suppression is implemented using LSTM-cMSA and speech restoration carried out using CED-cSA. The performance of the proposed model shows improvement even in low SNR conditions. In the case of -5 db, the proposed model increases 2% over STOI and 0.1 over PESQ. Similarly, when speech signal of untrained speaker mixed with unknown noise then proposed model shows nearly 18.12% increase in STOI and 0.54 increases in PESQ at -5db

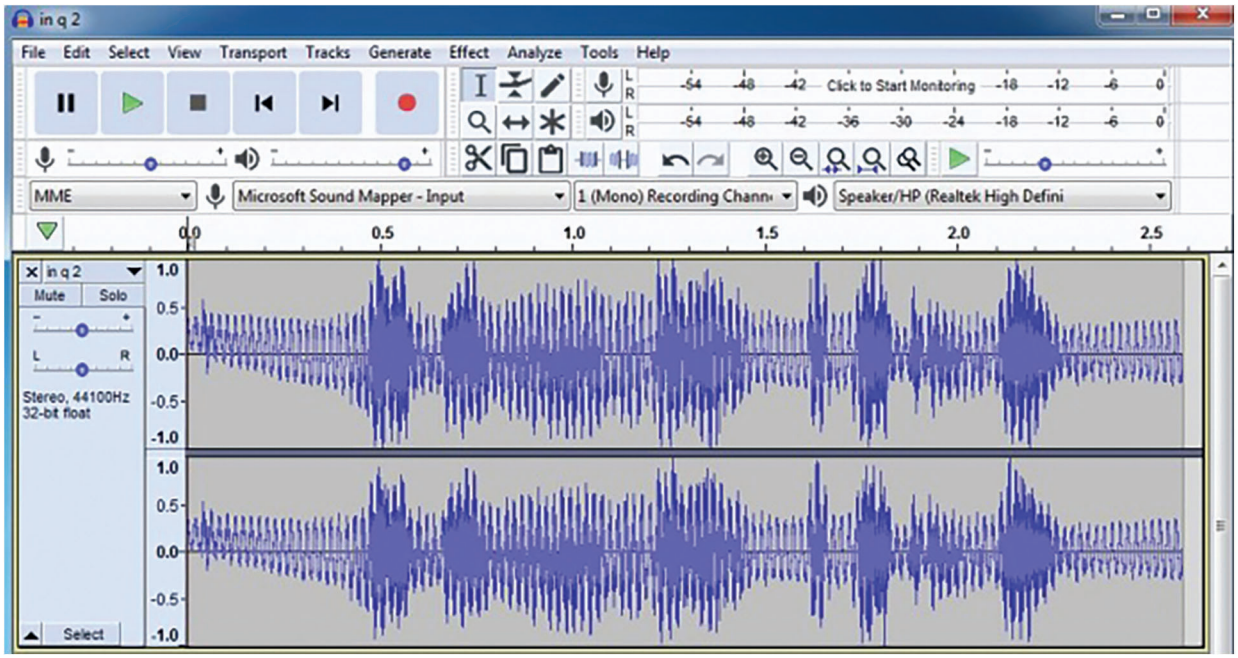

Figure 7: Fan noise signal in audacity

Features such as time taken for processing a single frame, several parameters are considered to analyze the computational complexity of the proposed model. Intel Core i7 9th Gen. Hexa Core, 2.6 GHz Clock Speed machine is used for measuring the time complexity frame. For the proposed model by using a 16-millisecond frameshift, the average time frame processing is calculated as 10.2 milliseconds. In stage 1 for noise suppression, LSTM along with up-sampling and pooling layers were used to increase the real-time factor up to 1.91 as shown in Tab. 6.

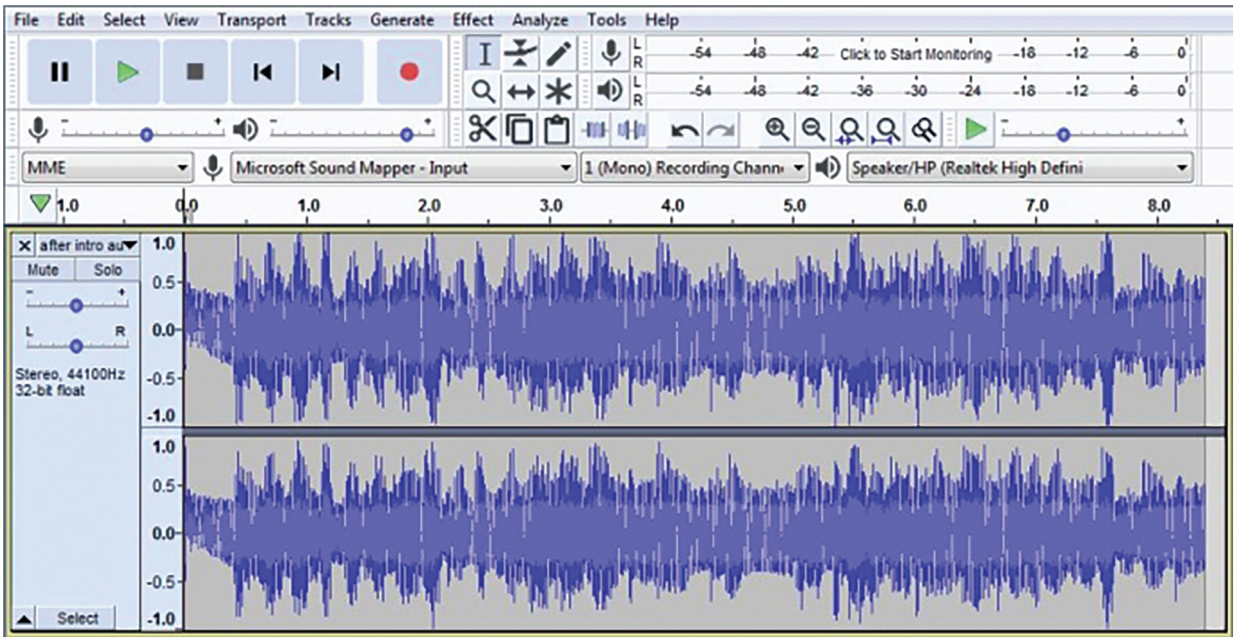

Figure 8: Wind noise signal in audacity

Table 2: STOI evaluation based on trained speaker over different model

Metric	STOI in percentage					
Test SNR	-5 db			-2 db		
Noise Type	Babble	Cafeteria	Average	Babble	Cafeteria	Average
Noisy	60.38	58.88	59.63	67.73	66.67	67.2
LSTM-IRM	78.72	75.8	77.26	84.05	82.86	83.455
LSTM-cMSA	78.88	75.62	77.25	85.23	82.73	83.98
CED-cSA	81.14	77.55	79.345	86.91	84.16	85.535
LSTM-cMSA + CED-cSA	83.78	80.76	82.27	89.55	87.37	88.46

Table 3: STOI evaluation based on untrained speaker over different model

Metric	STOI in percentage					
Test SNR	-5 db			-2 db		
Noise Type	Babble	Cafeteria	Average	Babble	Cafeteria	Average
Noisy	59.85	58.85	59.35	66.7	58.37	62.535
LSTM-IRM	76.52	75.11	75.815	83.9	74.63	79.265
LSTM-cMSA	76.86	74.96	75.91	84.12	74.48	79.3
CED-cSA	79.29	76.52	77.905	85.63	76.04	80.835
LSTM-cMSA + CED-cSA	82.403	78.38	80.3915	87.86	77.1	82.48

Table 4: PESQ evaluation based on trained speaker over different model

Metric	PESQ					
Test SNR	-5 db			-2 db		
Noise Type	Babble	Cafeteria	Average	Babble	Cafeteria	Average
Noisy	1.94	1.71	1.825	2.1	1.92	2.01
LSTM-IRM	2.37	2.23	2.3	2.67	2.52	2.595
LSTM-cMSA	2.37	2.22	2.295	2.65	2.5	2.575
CED-cSA	2.48	2.31	2.395	2.75	2.6	2.675
LSTM-cMSA + CED-cSA	2.77	2.42	2.595	2.87	2.72	2.795

Table 5: PESQ evaluation based on untrained speaker over different model

Metric	PESQ					
Test SNR	-5 db			-2 db		
Noise Type	Babble	Cafeteria	Average	Babble	Cafeteria	Average
Noisy	1.64	1.58	1.61	1.76	1.72	1.74
LSTM-IRM	2.02	2.08	2.05	2.33	2.33	2.33
LSTM-cMSA	2.02	2.07	2.045	2.32	2.33	2.325
CED-cSA	2.12	2.14	2.13	2.41	2.4	2.405
LSTM-cMSA + CED-cSA	2.27	2.27	2.27	2.55	2.53	2.54

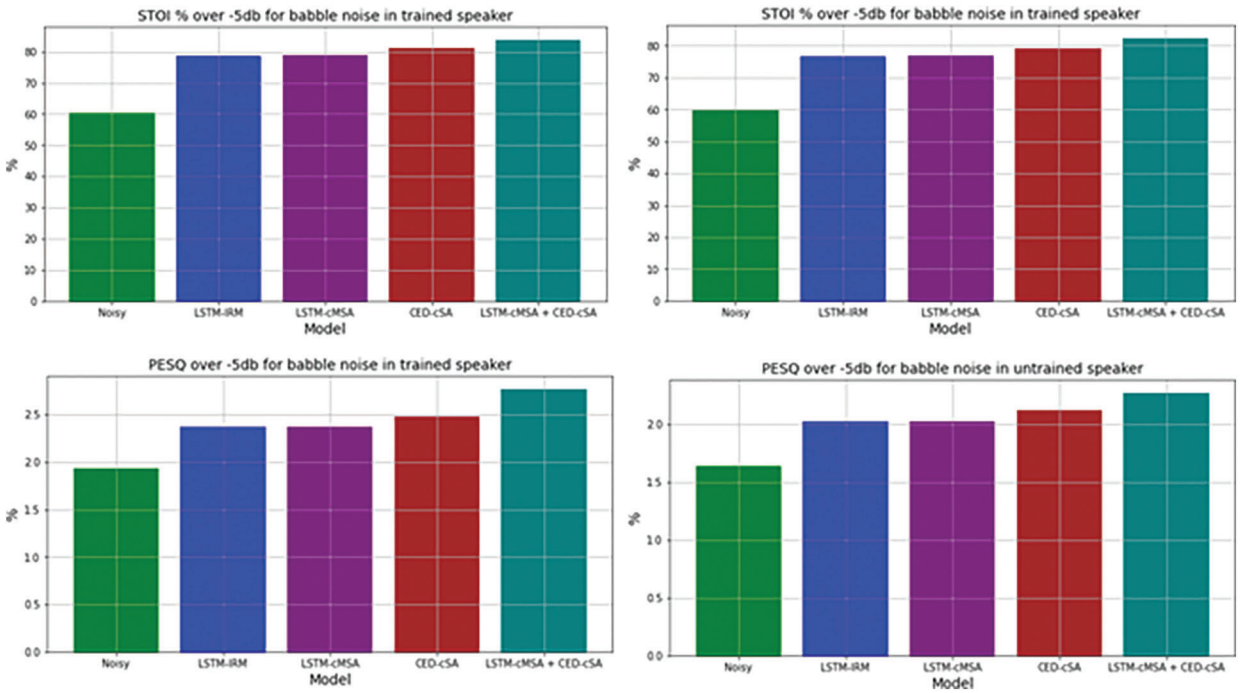

Figure 9: STOI and PESQ comparison for trained and untrained speaker over -5 db

Table 6: Computational complexity comparison based on no. of parameters, no. of multiplication, and real-time factor

Model	No. of parameters	No. of multiplication	Real-time factor
	10^6	10^6	
LSTM-cMSA	8.6	623	1.91
CED-cSA	6.9	501.2	1.20
LSTM-cMSA + CED-cSA	6.9	358.3	0.79

5 Conclusion

In this study, we have proposed a twofold architecture for speech enhancement using Recurrent and convolutional neural networks. In the first stage, LSTM is used for speech denoising by using cMSA as a loss function. Later in the second stage, the speech signal obtained from stage 1 is processed for speech restoration using CED with cSA as a loss function. The proposed model performs well in the speech denoising stage up to 5db in signal-to-noise ratio. On other hand in speech restoration using CED has shown very less improvement. Combining these two stages, the proposed model shows nearly 2% improvement in STOI and 0.1 Mean Opinion Score (MOS) points in PESQ. In addition, we found that the proposed model can able to reduce the computational complexity. Since proposed Convolutional Recurrent Network can able to perform well with fever number parameters. In recent years Automatic Speech Recognition are widely used in various real-world applications, we believe that the proposed model can be used in preprocessing stage to increase the accuracy of the ASR system.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] Y. Zhao, D. L. Wang, B. Xu and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1598–1607, 2020.

[3] S. Elshamy, N. Madhu, W. Tirry and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, 2017.

[4] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks," in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, Vancouver, Canada, pp. 7092–7096, 2013.

[5] L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen *et al.,* "Recurrent neural networks for noise reduction in robust ASR," in *Proc. Interspeech*, Portland, USA, pp. 22–25, 2012.

[6] F. Weninger, J. R. Hershey, J. Le Roux and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, Atlanda, USA, pp. 577–581, 2014.

[7] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, pp. 708–712, 2015.

[8] P. Malathi, G. Suresh, M. Moorthi and N. Shanker, "Speech Enhancement via smart larynx of variable frequency for laryngectomee patient for tamil language syllables using RADWT algorithm," *Circuits, Systems, and Signal Processing*, vol. 38, no. 9, pp. 1–15, 2019.

[9] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[10] Z. Zhao, H. Liu and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.

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

[12] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[13] Y. Wang, A. Narayanan and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[14] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[15] X. Mao, C. Shen and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. of Neural Information Processing Systems*, Barcelona, pp. 2802–2810, 2016.

[16] A. Karthik and J. L. MazherIqbal, "Efficient speech enhancement using recurrent convolution encoder and decoder," *Wireless Personal Communications*, vol. 119, no. 3, pp. 1959–1973, 2021.

[17] B. Fernandes and K. Mannepalli, "An analysis of emotional speech recognition for tamil language using deep learning gate recurrent unit," *Pertanika Journal of Science & Technology*, vol. 29, no. 3, pp. 1937–1961, 2021.

[18] B. Tolooshams, R. Giri, A. H. Song, U. Isik and A. Krishnaswamy, "Channel-attention dense U-Net for multichannel speech enhancement," 2001. [Online]. Available: https://arxiv.org/pdf/2001.11542.pdf.

[19] H. Noh, S. Hong and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, pp. 1520–1528, 2015.

[20] H. Yu, W. P. Zhu and B. Champagne, "Speech enhancement using a DNNaugmented colored-noise Kalman filter," *Speech Communication*, vol. 125, no. 2, pp. 142–151, 2020.

[21] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *2017 IEEE 60th Int. Midwest Symp. on Circuits and Systems (MWSCAS)*, IEEE, Medford, United States, pp. 1597–1600, 2017.

[22] X. Cui, Z. Chen and F. Yin, "Speech enhancement based on simple recurrent unit network," *Applied Acoustics*, vol. 157, no. 2, pp. 107019, 2020.

[23] P. Molchanov, A. Mallya, S. Tyree, I. Frosio and J. Kautz, "Importance estimation for neural network pruning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 11264–11274, 2019.

[24] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Transaction Audio, Speech, Language Process*, vol. 28, pp. 1404–1415, 2020.

[25] A. Pandey and D. L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. on Acoustics, Speech, & Signal Processing*, Brighton, UK, pp. 6875–6879, 2019.

[26] I. Fedorov, M. Stamenovic, C. Jensen, L. C. Yang, A. Mandell *et al.,* "TinyLSTMs: Efficient neural speech enhancement for hearing aids," in *ISCA, INTERSPEECH*, Shanghai, China, pp. 4054–4058, 2020.

[27] M. Kolbæk, Z. H. Tan and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[28] E. M. Grais, G. Roma, A. J. R. Simpson and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 9, pp. 1773–1783, 2017.

[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[30] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Advances in Signal Processing*, vol. 7, no. 7, pp. 1110–1126, 2005.

[31] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 2092–2102, 2019.

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

[33] X. R. Zhang, W. F. Zhang, W. Sun, X. M. Sun and S. K. Jha, "A robust 3-D medical watermarking based on wavelet transform for data protection," *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1043–1056, 2022.

[34] X. R. Zhang, X. Sun, X. M. Sun, W. Sun and S. K. Jha, "Robust reversible audio watermarking scheme for telemedicine and privacy protection," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 3035–3050, 2022.

[35] Mustaqeem and S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, no. 4, pp. 107101, 2021.

[36] Mustaqeem and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications: An International Journal*, vol. 167, no. C2021.

[37] S. Girirajan and A. Pandian, "Acoustic model with hybrid Deep Bidirectional Single Gated Unit (DBSGU) for low resource speech recognition," *Multimedia Tools Application*, vol. 18, no. 2, pp. 183, 2022.