A Novel Approach to Speech Enhancement Based on Deep Neural Networks

Maryam SALEHI, Sattar MIRZAKUCHAKI
Department of Electrical Engineering, Iran University of Science & Technology, Tehran, Iran
m_kuchaki@iust.ac.ir

***Abstract*—Minimum mean-square error (MMSE) approaches have been shown to achieve state-of-the-art performance on the task of speech enhancement. However, MMSE approaches lack the ability to accurately estimate non-stationary noise sources. In this paper, a long short-term memory fully convolutional network (LSTM-FCN) is utilized to accurately estimate *a priori* signal-to-noise ratio (SNR) since the speech enhancement performance of an MMSE approach improves with the accuracy of the used *a priori* SNR estimator. The proposed MMSE approach makes no assumptions about the characteristics of the noise or the speech. MMSE approaches that utilize the LSTM-FCN estimator are evaluated using the mean opinion score of the perceptual evaluation of speech quality (PESQ) and the short-time objective intelligibility (STOI) measures of speech. The experimental investigation shows that the speech enhancement performance of an MMSE approach that utilizes LSTM-FCN estimator significantly increases**.

***Index Terms*—long short-term memory, machine learning, mean square error methods, recurrent neural networks, speech enhancement.**

I. INTRODUCTION

Speech processing applications, such as automatic speech recognition, mobile communications, health care, and voice activity detection, received a lot of attention over the past decade. Speech intelligibility and quality may be degraded by coloured or non-stationary background noises, with examples including airplane, car, factory, and street noises. Background noises can affect the performance of a speech processing system. However, speech processing robustness can be improved by speech enhancement techniques. A speech enhancement algorithm aims to improve the intelligibility and quality of noisy speech. These algorithms can be categorized as single-channel and multi-channel algorithms [1-3]. In this paper, single-channel speech enhancement algorithms are considered.

The objective of a single-channel speech enhancement algorithm is to recover the components of the clean speech from the noisy speech with improved perceptual quality and intelligibility. Most single-channel speech enhancement algorithms need an estimate of the noise power spectral density (PSD). Therefore, inaccurate estimation of the noise PSD can affect the performance of a speech enhancement algorithm. Spectral subtraction is commonly used for the enhancement of single-channel speech. In these methods, the noise PSD is estimated during speech absence and is subtracted from the noisy speech spectrum to estimate the clean speech [4-5]. Such methods are effective under fairly stationary noise conditions, but often fail to estimate coloured or non-stationary noise sources during active regions of the noisy speech signal.

To improve speech enhancement algorithms, several approaches have been proposed during the last decade. Among the most speech enhancement algorithms are those based on minimum mean square error (MMSE) [6-7]. In [6], the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator is proposed to optimally estimate the magnitude spectrum of the clean speech. A gain function is utilized to minimize the mean-square error between the clean and enhanced speech spectra. In [8], an optimum non-linear estimator based on an MMSE sense is proposed to minimize the background noises in different conditions of the speech signal. In [9], the short-time spectral amplitude (STSA) of the bioacoustics signal is used to identify the estimated clean speech signal in the MMSE sense. A modified version of the MMSE-STSA estimator used in [6] has been developed in [10]. In [10], a non-stationary noise tracking approach based on a log-spectral power MMSE estimator is proposed. In [11], the two-step noise reduction (TSNR) technique is utilized to enhance speech signals in noisy environments. The TSNR technique introduces harmonic distortion in the enhanced speech because of the unreliability of estimators for small signal-to-noise ratio (SNR). In [12], a nonlinear mapping technique is utilized to regenerate the degraded harmonics of the distorted signal. However, all these approaches have in common that they are a function of the *a priori* SNR estimate of noisy speech spectral component. Voice activity detection (VAD) is commonly utilized to update the *a priori* SNR estimate during speech absence [13]. This algorithm is effective in numerous applications but often causes detection errors mainly due to the loss of discrimination at low SNR levels. In [6], the *a priori* SNR is estimated by employing the decision-directed (DD) approach. The DD approach lacks the ability to accurately estimate non-stationary noise sources. In [14], the problem of maximum likelihood (ML) estimation of the SNR parameter is considered. To extend the maximum likelihood estimation method of the *a priori* SNR, Selective cepstro-temporal smoothing (SCTS) is performed in [15]. In [15], an estimate of the *a priori* SNR is obtained by adaptively smoothing its maximum likelihood estimate in the cepstral domain.

Recently, deep neural networks have been employed for speech enhancement systems. Deep learning techniques are able to establish very complex relationships and nonlinear feature interactions to solve a particular problem [16-19]. In [20], the convolutional neural network is used to segregate speech in background noise. This network is utilized to explain the features of the noisy speech signal. In [21], a

Digital Object Identifier 10.4316/AECE.2022.02009

boosted two-stage neural network with a multi-objective learning method has been applied to enhance the noisy speech signal. In the first stage, a single deep neural network is utilized to obtain multiple base predictions. Also, convolution layers are employed to concatenate the base predictions in the second stage [21]. In [22], a three-stage deep neural network by boosting contextual information is used for VAD. In [23], a stack of two deep neural networks is proposed for speech enhancement via a multi-objective learning and ensembling (MOLE) framework. Static noise is estimated using the first several frames of an utterance and thus fixed within that utterance. Also, the dynamic noise is calculated from the output of the first DNN. Unlike MMSE approaches, these speech enhancement algorithms do not require *a priori* SNR estimator. However, the performance of these speech enhancement systems is not always satisfactory due to non-stationary and coloured noises that make it difficult to estimate noise signals mathematically.

In [24], a long short-term memory (LSTM) network is developed when the source is impaired by bursty impulsive noise. The trained model estimate the noise PSD online and thus applies a linear minimum mean square error (LMMSE) approach to enhance speech signal in noisy environments. In [25], a deep learning approach to *a priori* SNR estimation is utilized to increase the performance of MMSE approaches to speech enhancement. This framework utilizes a residual long short-term memory (ResLSTM) recurrent neural network (RNN) to estimate the *a priori* SNR directly from the noisy speech magnitude spectrum of a given time-frame. In [26], an MMSE-based noise PSD estimator is proposed for the *a priori* SNR estimation. In [26], a temporal convolutional network (TCN) is used to estimate the *a priori* SNR. The bottlenecks of these MMSE-based estimators are a large number of hyperparameters and the complexity of training a recurrent architecture.

In this paper, an LSTM-FCN framework is utilized to accurately estimate *a priori* SNR. Compared with the literature in the field, the main contribution of this paper is as follows: 1) a speech enhancement algorithm is designed and developed with better results in comparison with other algorithms; 2) the framework outperforms other machine learning algorithms, while using fewer parameters and avoiding the complexity of training a recurrent architecture. The proposed LSTM-FCN framework is motivated by the following advantages [27-28]: 1) MMSE approaches lack the ability to accurately estimate *a priori* SNR in presence of non-stationary noise sources; 2) LSTM-FCN does not make assumptions about the characteristics of the noise or the speech; 3) LSTM-FCN framework shows no phase delay.

This paper is organized as follows. In section II, the signal model and notations is described. In section III, the speech enhancement based on MMSE approaches are reviewed. In section IV, an LSTM fully convolutional network (LSTM-FCN) framework is introduced and a novel speech enhancement algorithm is developed. In section V, to investigate the proposed framework, simulation results are performed in different scenarios and compared with DD approach [13], TSNR technique [11], HRNR technique [12], and ResLSTM framework [25]. Finally, the paper is summarized in the last section.

II. Signal Model and Notation

In this paper, the speech enhancement of the noisy speech signal is performed in the time-frequency domain by using the short-time Fourier transform (STFT). This framework consists of three stages: (1) the analysis stage, where the noisy speech signal is analyzed frame-wise using the STFT, (2) the compensation stage, where the noisy speech STFT is modified to enhance the noisy STFT, (3) the reconstruction stage, where inverse STFT procedure is used to construct enhance speech in time-domain. In the time-domain, the speech signal can be written as (1). In (1), n is discrete-time index, $y(n)$ is noisy speech, $x(n)$ is clean speech and $d(n)$ is uncorrelated additive noise.

$$y(n)=x(n)+d(n) \tag{1}$$

Let $Y(l,k)$, $X(l,k)$,and $D(l,k)$ denote the complex-valued STFT coefficients of the noisy speech, the clean speech, and the noise respectively, l denotes time-frame index, k denotes discrete frequency index, and having defined the above, the STFT coefficients of noisy speech signal error can be defined as:

$$Y(l,k)=X(l,k)+D(l,k) \tag{2}$$

It is assumed that $D(l,k)$ and $X(l,k)$, are statistically independent across time and frequency, follow conditional zero-mean Gaussian distributions, and satisfy (3) and (4) for each l and k respectively. In (3) and (4), $E\{\bullet\}$ is the expected value, $\lambda_x(l,k)$ is the spectral variance of clean speech, and $\lambda_d(l,k)$ is the spectral variance of the noise.

$$E\left\{|X(l,k)|^2\right\}=\lambda_x(l,k) \tag{3}$$

$$E\left\{|D(l,k)|^2\right\}=\lambda_d(l,k) \tag{4}$$

An MMSE approach to speech enhancement uses the *a priori* SNR to calculate a gain function. The gain function is utilized to the magnitude spectrum of the noisy speech, which constructs the enhanced speech magnitude spectrum. The *a priori* SNR and the *a posteriori* SNR of a noisy speech spectral component can be written as [6]:

$$\xi(l,k)=\frac{\lambda_x(l,k)}{\lambda_d(l,k)} \tag{5}$$

and

$$\gamma(l,k)=\frac{|Y(l,k)|^2}{\lambda_d(l,k)} \tag{6}$$

respectively. In (6), $|Y(l,k)|$ is the noisy speech magnitude spectrum. The *a priori* SNR and the *a posteriori* SNR must be estimated from the noisy speech since $\lambda_x(l,k)$ and $\lambda_d(l,k)$ are unobserved during speech enhancement. In this paper, an LSTM-FCN framework is utilized to accurately estimate the *a priori* SNR. Clearly, when training an LSTM-FCN framework to estimate the *a priori* SNR, the variances of the clean speech and noise speech are given. In the following, l and k are omitted from the notation unless otherwise indicated.

III. Speech Enhancement Based on MMSE Approaches

An MMSE approach utilized the noisy signal to compute a gain function for each time-frame and discrete frequency indexes. The gain function is used to enhance the noisy speech magnitude, $|Y(l,k)|$ spectrum. The minimum mean-

square error short-time spectral amplitude (MMSE-STSA) algorithm minimizes the mean-square error (MSE) between the clean and enhanced speech spectra [29-30]. The MMSE-STSA algorithm gain function is given by (7). In (7), $I_0(\bullet)$ and $I_1(\bullet)$ are modified Bessel functions of the zero and first kind, respectively, and $h(l,k)$ is given by (8).

$$G_{MMSE-STSA(l,k)} = \frac{\sqrt{\pi}}{2}\frac{\sqrt{h(l,k)}}{\gamma(l,k)}\exp\left(-\frac{h(l,k)}{2}\right)x \left(\left(1+h(l,k)I_0\left(\frac{h(l,k)}{2}\right)+h(l,k)I_1\left(\frac{h(l,k)}{2}\right)\right)\right) \quad (7)$$

$$h(l,k) = \frac{\xi(l,k)}{(1+\xi(l,k))}\gamma(l,k) \quad (8)$$

The minimum mean-square error log-spectral amplitude (MMSE-LSA) is used to minimize the MSE between the clean speech and enhanced speech log-magnitude spectra [31]. The MMSE-LSA algorithm gain function is given by

$$G_{MMSE-LSA} = \frac{\xi(l,k)}{(1+\xi(l,k))}\exp\left(\frac{1}{2}\int_{h(l,k)}^{+\infty}\frac{e^{-t}}{t}dt\right) \quad (9)$$

In the derivation of the classic Wiener Filter (WF), *a posteriori* SNR is assumed to be equal to "1+the *a priori* SNR". The WF algorithm is utilized to minimize the MSE between the clean and enhanced speech complex discrete Fourier transform (DFT) coefficients [7]. The WF algorithm gain function is given by

$$G_{WF} = \frac{\xi(l,k)}{(1+\xi(l,k))} \quad (10)$$

Considering all the above, the MMSE-based approach uses both the *a priori* SNR, ξ, and *a posteriori* SNR, γ, of the noisy speech to enhance the noisy speech magnitude spectrum. As the clean speech and noise are unobserved during speech enhancement, the *a priori* SNR must be estimated from the noisy speech. Therefore, an accurate the *a priori* SNR estimate, $\hat{\xi}$, affects the speech enhancement performance of an MMSE approach. In this paper, an LSTM-FCN framework is utilized to accurately estimate the *a priori* SNR.

IV. The LSTM-FCN Estimator

Temporal convolutions have proven to be an effective learning framework for speech enhancement problems. In this paper, an LSTM fully convolutional network (LSTM-FCN) framework is used to estimate the *a priori* SNR of the noisy signal for MMSE approaches, as shown in Fig. 1. In the proposed architecture, the input feature tensor is conveyed into a fully connected layer. The output of the fully connected layer is then passed into the LSTM block. The LSTM block is followed by a dropout. Simultaneously, the input feature vector is passed into a fully convolutional block in each step. The global average pooling layer is applied after the output of the fully convolutional block. Finally, the output of the LSTM block followed by the dropout is augmented by the output of the global average pooling layer. In Fig. 1, **FC** denotes a fully-connected layer, **OL** denotes the output layer, the dropout block is as [32], **FCN** denotes a fully convolutional network. The output layer is a fully-connected layer with sigmoidal units. Also, the FCN block is followed by a pooling layer. The pooling layer is utilized to reduce the number of parameters in the model before the *a priori* SNR estimation.

The pooling layer is a global average pooling layer proposed by [33]. This layer creates one feature map from the previous convolutional layer and takes the average of each feature map. An illustration of the global average pooling layer is shown in Fig. 2. Clearly, in the proposed approach, the output of the pooling layer and LSTM are concatenated.

Remark 1. Without the **FC** and the dropout layers, the performance of the LSTM block is significantly reduced due to the rapid overfitting of small short-sequence speech datasets and a failure to learn long-term dependencies in the long-sequence speech datasets. In the proposed LSTM-FCN framework, a high dropout rate of 80% was used after the LSTM to combat overfitting.

A. Fully Convolutional Block

Fully convolutional networks (FCNs) are generally utilized to extract features. In the proposed architecture, the FCN block contains of three temporal convolutional networks with filter sizes of 128, 256, and 128 respectively. A temporal convolutional network is a one-dimensional causal convolutional network [26]. An illustration of the FCN block is shown in Fig. 3. The filter for a temporal convolutional network is parameterized by tensor $W \in \Re^{N_O \times d \times N_i}$ and biases $b \in \Re^{N_i}$, where N_0 is the output dimension, d is the filter duration, and N_i is the input dimension [28]. Lea et al. [34] define the computation at t-th step as follows:

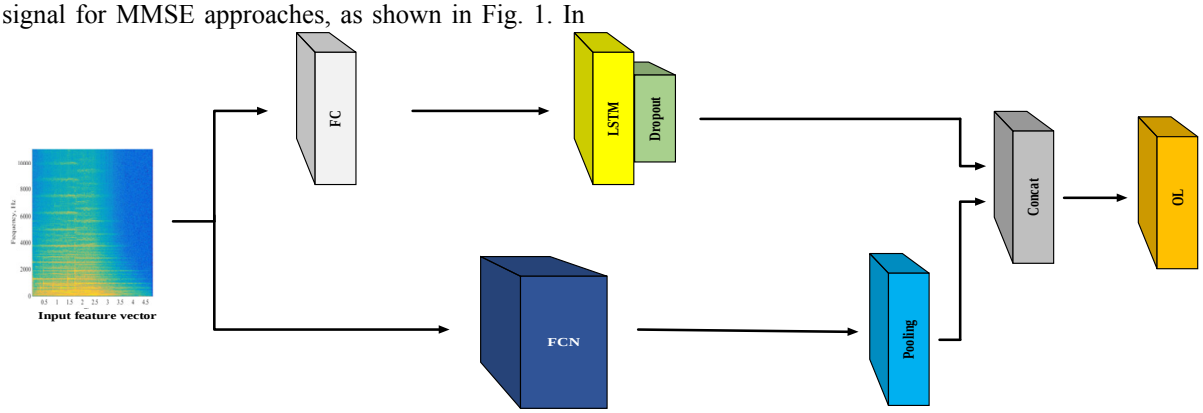


Figure 1. The overall architecture of the LSTM-FCN

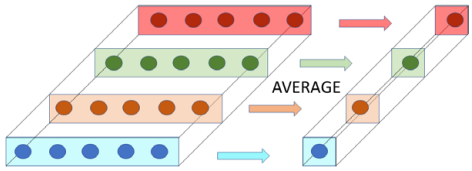


Figure 2. The global average pooling layer

$$E^{O}_{i,t} = b_i + \sum_{t'=1}^{d} \left\langle W_{i,t',.}, E^{I}_{.,t+d-t'} \right\rangle \tag{11}$$

In (11), $E^{I} \in \Re^{N_I \times (2m+1)}$ denotes the input feature tensor, and $E^{O} \in \Re^{N_O \times (2m+1)}$ denotes the output feature tensor.

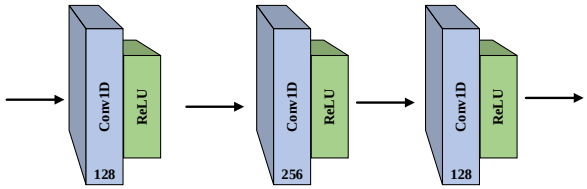


Figure 3. The Fully convolutional networks (FCNs) block

B. *Long Short-Term Memory Block*

The LSTM block contains three residual long short-term memory recurrent neural networks, (LSTM RNNs), as shown in Fig. 4. The LSTM RNNs are utilized to solve the vanishing gradient problem. The LSTM RNNs address the vanishing gradient problem commonly found in ordinary recurrent neural networks by incorporating gating functions into their state dynamics [35]. Each residual LSTM RNN contains an LSTM cell, as shown in Fig. 5.

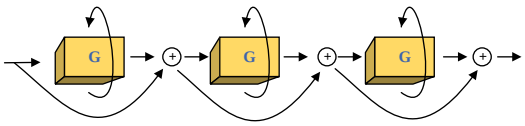


Figure 4. The residual LSTM block. It consists of three residual LSTM RNN, G

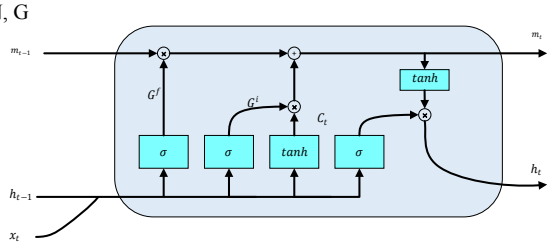


Figure 5. Basic architecture of the LSTM cell

Based on the architecture shown in Fig. 5, the computation at each time step is as of the following form [36]:

$$\begin{aligned} G^{i} &= \sigma(W^{i} h_{t-1} + B^{i} x_t) \\ G^{f} &= \sigma(W^{f} h_{t-1} + B^{f} x_t) \\ G^{o} &= \sigma(W^{o} h_{t-1} + B^{o} x_t) \\ C_t &= \tanh(W^{c} h_{t-1} + B^{c} x_t) \\ m_t &= G^{f} \circ m_{t-1} + G_i \circ C \\ h_t &= \tanh(G^{O} \circ m_t) \end{aligned} \tag{12}$$

In (12), G^i, G^f, G^0 and C_t are the activation vectors of the input, forget, output, and cell state respectively, W^i, W^f, W^0,and W^c are the weight matrices of the input, forget, output, and cell state gates respectively, B^i, B^f, B^0 and B^c are the projection matrices of the input, forget, output and cell state gates respectively,h_t denotes the hidden state vector of the LSTM cell,m_t denotes the memory vector,x_t denotes the input vector, σ(•) denotes a standard logistic sigmoid function, and the elementwise multiplication is represented by ∘ .

C. *Mathematical Model*

Let $X_i \in \Re^{l_0}$ denotes the input feature vector of length l_0 for *i*-th discrete-time index. In the proposed architecture, the input feature tensor can be defined as (13). In (13), $\widetilde{X}_t \in \Re^{l_0 \times (2m+1)}$ is the input feature tensor in a window whit *2m+1* frames for t-th step, and N_s denotes frame shift.

$$\widetilde{X}_t = \left\{ X_{(1+N_s)+i} \right\}_{i=-m}^{m}, \quad t = \{1,2,...,\} \tag{13}$$

Considering all the above, the proposed LSTM-FCN estimator can be written as:

$$\hat{\bar{\xi}}_t = softmax\left(\left[Y_{GPL}, Y_{DL} \right] \right) \tag{14}$$

In (14), $\hat{\bar{\xi}}_t$ is estimate of the cumulative distribution function (CDF) of $\xi_{dB}(l,k)$ (in dB) for *t*-th step, *softmax(•)* is softmax function, $Y_{GPL} \in \Re^{\tilde{N}_1}$ is output of the global pooling layer, and $Y_{DL} \in \Re^{\tilde{N}_1}$ is output of the dropout layer.

Remark 2. In this paper, the Adam optimizer [37] is used since it converges better than stochastic gradient descent (SGD) [38] and root mean squared propagation (RMSprop) [39] optimizers when using the LSTM-FCN framework. Also, a *priori* SNR was mapped to the interval [0-1] to improve the rate of convergence of the used Adam algorithm.

In [25], it can be seen that $\xi_{dB}(l,k)$ is distributed normally with mean μ_k and variance $\tilde{\sigma}^2$ for a given frequency component. Therefore, CDF of $\xi_{dB}(l,k)$ was used as the map in the proposed LSTM-FCN framework. The map is given by:

$$\bar{\xi}(l,k) = \frac{1}{2}\left[1 + erf\left(\frac{\xi_{dB}(l,k)}{\tilde{\sigma}_k \sqrt{2}} \right) \right] \tag{15}$$

In (15), $\bar{\xi}(l,k)$ is the cumulative distribution function of $\xi_{dB}(l,k)$, and *σ(•)* is the error function.

Remark 3. Since the loss function (cross-entropy) is an arbitrary unknown sequence of a convex function, the convergence of the proposed LSTM-FCN framework can be proved the same as [37].

V. Experimental Setup

All clean speech and noise signals are single-channel, with a sampling frequency of 16 kHz. The Hamming window function is utilized for spectral analysis and synthesis, with a frame-length of 32 ms (512 time-domain samples) and a frame-shift of 16 ms (256 time-domain samples). The *a priori* SNR estimate was calculated by using the single-sided noisy speech magnitude spectrum, which included both the Nyquist frequency component and the DC frequency component. To investigate the proposed LSTM-FCN framework, simulation results are performed in different scenarios and compared with the decision-directed (DD) approach [13], the two-step noise reduction (TSNR) technique [11], harmonic regeneration noise reduction (HRNR) technique [12], and ResLSTM framework [25]. In simulation results, the complexity and the design specifications of the ResLSTM framework is similar to that

of the LSTM-FCN framework. In the LSTM-FCN framework, the *a posteriori* SNR estimate, $\hat{\gamma}(l,k)$, is as (16) where $\hat{\xi}(l,k)$ is the *a priori* SNR estimate.

$$\hat{\gamma}(l,k) = 1 + \hat{\xi}(l,k) \tag{16}$$

A. Training Set

The training procedure is performed on a speech dataset that is produced from the TIMIT corpus [6], and train-clean-100 set from the Librispeech corpus [7]. To improve the performance of the speech enhancement approach processing methods in various noisy backgrounds, babble, airport, factory, car, F16, street, pink, white, Hall, destroyer engine, and HF channel noises are selected from the NOISEX-92 [8] and Aurora-4 [9] databases. Therefore, both real-world non-stationary (e.g., voice babble) and coloured noise sources (e.g., factory), were included in the noise training set.

B. Test Set

For the test set, real-world noise sources, including two non-stationary (babble and street) and two coloured (F16 and factory), were included in the test set. 20 clean speech signals were randomly selected from the TIMIT speech corpus [6] for each of the four noise signals. Finally, the test set is built by adding four real-world noise sources to the clean speech signals with various SNR: -5 to 15 dB, in 5 dB increments.

C. Validation Set

In order to validate the model obtained, 10% of the clean speech training set was used as a validation set. Also, to create the validation set, a random section of the noise signals was mixed with the clean speech at the following SNR levels: -5 to 15 dB, in 5 dB increments.

D. Training Strategy

The training target for a deep neural network within the LSTM-FCN framework is the mapped a *priori* SNR, as described in the previous section. Table I gives a formal algorithmic description of training strategy. The following strategy was utilized to train the LSTM-FCN:

- Cross-entropy as the loss function;
- 16 time-frames were used simultaneously for the LSTM training;
- The LSTM-FCN was trained via Adam optimizer [10], with an initial learning rate of 10^{-3} and a final learning rate of 10^{-4}. At each epoch, the learning rate is halved;
- The number of training epochs was kept constant at 200 epochs;
- The batch size of the noisy speech signals is set to 10;
- 90% of the clean speech training set was for LSTM-FCN training, and the remaining 10% was used as validation set;
- A clean speech and noise signals were randomly selected from speech and noise training datasets, respectively, to build a noisy speech signal. The noise signal was mixed with a randomly selected SNR between −5 and 10 dB.

E. Evaluation Metrics

In this paper, the LSTM-FCN framework was evaluated using the short-time objective intelligibility (STOI) [40], and the perceptual evaluation of speech quality (PESQ) [26]. STOI and PESQ measures were utilized to evaluate both the quality and intelligibility of the enhanced speech signal respectively. Average STOI and PESQ scores were computed over the test set.

TABLE I. A FORMAL ALGORITHMIC DESCRIPTION OF TRAINING STRATEGY

$Model_{weights}$ = *Initial_*$Model_{weights}$
while *(epoch < max_epoch || Covergence_Flag)*
train(model; Data; $initial_{lr}$*; batch_size)*
Update $Model_{weights}$
epoch = epoch +1;
$initial_{lr}$*=UpdateLearningRate(*$initial_{lr}$*, epoch)*
Covergence_Flag = ConvergenceFunction($Model_{weights}$*)*
end while

F. A priori SNR Estimation Spectral Distortion (SD) Levels

To evaluate the performance of the LSTM-FCN framework, real-work noise sources, including two non-stationary (babble and street) and two coloured (F16 and factory) at multiple SNR levels, were included in the test set, and the results are compared to DD approach [13], TSNR technique [11], HRNR technique [12], and ResLSTM framework [25]. The average frame-wise SD of the *a priori* SNR estimation as (17) is used to evaluate the accuracy of the *a priori* SNR estimators. In (17), N_F is the total number of frames, and N_l is the frame length in discrete-time samples.

$$SD = \sqrt{\frac{1}{N_F}\sum_{n=1}^{N_F}\left(\frac{1}{N_l+1}\sum_{k=0}^{N_l/2}\left(\xi_{dB}(n,k)-\hat{\xi}_{dB}(n,k)\right)^2\right)} \tag{17}$$

The *a priori* SNR estimation SD levels are shown in Table II. The LSTM-FCN framework outperformed all previous *a priori* SNR estimation methods (DD, TSNR, HRNR, and ResLSTM) with respect to SD levels for all noise types. The LSTM-FCN SD level averages around 15.2 while the DD, TSNR, HRNR, and ResLSTM frameworks give SD levels that average about 22.1, 19.6, 17.8, and 16.2 respectively. It can be seen that the average SD level of the LSTM-FCN framework is reduced by 31% for the DD approach, 22% for the TSNR technique, 14% for the HRNR technique, and 6% for the ResLSTM framework.

TABLE II. SD LEVELS FOR MMSE-STSA ESTIMATOR USING EACH OF THE A PRIORI SNR ESTIMATORS

Noise type	$\hat{\xi}(n,k)$	**SNR level(dB)**			
		-5	**0**	**5**	**10**
babble	DD	20.2	18.1	16.7	16.8
	TSNR	18.3	17.2	17.0	17.0
	HRNR	15.7	13.5	12.9	12.8
	ResLSTM	13.9	12.7	11.3	11.4
	LSTM-FCN	**12.8**	**11.9**	**11.1**	**11.0**
street	DD	21.7	20.1	18.2	18.2
	TSNR	21.5	19.8	18.1	18.0
	HRNR	17.4	16.8	15.9	15.7
	ResLSTM	14.6	13.3	13.1	13.1
	LSTM-FCN	**13.6**	**12.9**	**12.1**	**12.0**
F16	DD	22.6	22.3	20.9	21.1
	TSNR	22.5	21.7	21.1	21.0
	HRNR	21.7	20.6	20.1	20.1
	ResLSTM	20.6	20.1	18.6	18.3
	LSTM-FCN	**18.9**	**18.1**	**17.4**	**17.2**
factory	DD	25.8	24.3	22.6	22.5
	TSNR	24.7	23.4	21.9	21.4
	HRNR	22.3	21.6	20.8	20.8
	ResLSTM	20.9	20.1	19.8	19.9
	LSTM-FCN	**20.1**	**19.1**	**18.6**	**18.2**

G. MMSE-STSA Estimator

The STOI and PESQ scores for the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator using each of the *a priori* SNR estimators are shown in Table III and Table IV respectively. The STOI and PESQ evaluate the enhanced speech from two different perspectives, i.e., speech intelligibility and quality. The enhanced speech from the MMSE-STSA using the LSTM-FCM framework achieves the highest STOI and PESQ scores for both the real-world non-stationary and coloured noise sources. It can be seen that an average STOI score of the MMSE-STSA using the LSTM-FCN framework has improved by 15% for the DD approach, 11% for the TSNR technique, 10%, for the HRNR technique, and 5% for the ResLSTM framework. An average PESQ score of the MMSE-STSA using the LSTM-FCN framework has also improved by 10% for the DD approach, 9.7% for the TSNR technique, 9%, for the HRNR technique, and 4% for the ResLSTM framework.

TABLE III. THE STOI SCORE FOR MMSE-STSA ESTIMATOR USING EACH OF THE A PRIORI SNR ESTIMATORS

Noise type	$\hat{\xi}(n,k)$	SNR level(dB)			
		-5	**0**	**5**	**10**
babble	DD	0.6	0.75	0.8	0.9
	TSNR	0.63	0.8	0.85	0.92
	HRNR	0.62	0.82	0.85	0.93
	ResLSTM	0.67	0.84	0.9	0.96
	LSTM-FCN	**0.7**	**0.89**	**0.92**	**0.98**
street	DD	0.58	0.69	0.79	0.86
	TSNR	0.6	0.68	0.81	0.88
	HRNR	0.63	0.7	0.85	0.9
	ResLSTM	0.64	0.74	0.89	0.95
	LSTM-FCN	**0.66**	**0.78**	**0.91**	**0.98**
F16	DD	0.6	0.74	0.81	0.9
	TSNR	0.65	0.76	0.84	0.88
	HRNR	0.58	0.74	0.86	0.87
	ResLSTM	0.64	0.72	0.86	0.95
	LSTM-FCN	**0.69**	**0.79**	**0.91**	**0.96**
factory	DD	0.55	0.69	0.79	0.83
	TSNR	0.61	0.72	0.82	0.82
	HRNR	0.63	0.75	0.85	0.86
	ResLSTM	0.67	0.79	0.86	0.94
	LSTM-FCN	**0.69**	**0.82**	**0.9**	**0.96**

TABLE IV. THE PESQ SCORE FOR MMSE-STSA ESTIMATOR USING EACH OF THE A PRIORI SNR ESTIMATORS

Noise type	$\hat{\xi}(n,k)$	SNR level(dB)			
		-5	**0**	**5**	**10**
babble	DD	75.2	85.7	90.1	95.4
	TSNR	76.5	84.9	91.3	95.4
	HRNR	78.6	85.9	92.1	95.3
	ResLSTM	80.5	86.9	93.6	96.8
	LSTM-FCN	**84.4**	**90.6**	**94.6**	**97.2**
street	DD	73.8	78.5	85.6	90.1
	TSNR	72.6	77.9	84.2	91.1
	HRNR	72.5	76.6	85.9	91.1
	ResLSTM	76.3	81.1	86.3	93.6
	LSTM-FCN	**79.3**	**85.8**	**88.7**	**94.9**
F16	DD	70.3	77.2	83.6	87.4
	TSNR	71.0	79.5	84.1	88.2
	HRNR	71.6	78.2	84.3	89.1
	ResLSTM	75.4	82.6	89.4	91.9
	LSTM-FCN	**79.7**	**85.9**	**90.1**	**93.9**
factory	DD	67.2	74.9	80.6	87.6
	TSNR	66.9	75.1	80.9	86.7
	HRNR	69.3	78.2	81.6	87.9
	ResLSTM	74.9	81.4	83.6	90.5
	LSTM-FCN	**82.3**	**87.3**	**91.1**	**95.9**

It can be seen that there is a correlation between *a priori* SNR estimation accuracy and STOI and PESQ scores (speech enhancement performance).

H. MMSE-LSA Estimator

The STOI and PESQ scores for the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator using each of the *a priori* SNR estimators are shown in Table V and Table VI respectively. Simulation results show that the MMSE-LSA using the LSTM-FCM framework outperforms DD, TSNR, and HRNR algorithms.

It can be seen that an average STOI score of the MMSE-STSA using the LSTM-FCN framework has improved by 22% for the DD approach, 28% for the TSNR technique, 18%, for the HRNR technique, and 8% for the ResLSTM framework. An average PESQ score of the MMSE-STSA using the LSTM-FCN framework has also improved by 15% for the DD approach, 10% for the TSNR technique, 11%, for the HRNR technique, and 5% for the ResLSTM framework.

TABLE V. THE STOI SCORE FOR MMSE-LSA ESTIMATOR USING EACH OF THE A PRIORI SNR ESTIMATORS

Noise type	$\hat{\xi}(n,k)$	SNR level(dB)			
		-5	**0**	**5**	**10**
babble	DD	0.58	0.76	0.84	0.91
	TSNR	0.55	0.7	0.81	0.9
	HRNR	0.52	0.69	0.79	0.9
	ResLSTM	0.71	0.78	0.9	0.94
	LSTM-FCN	**0.75**	**0.85**	**0.93**	**0.97**
street	DD	0.6	0.68	0.77	0.89
	TSNR	0.62	0.7	0.79	0.86
	HRNR	0.62	0.69	0.81	0.91
	ResLSTM	0.65	0.75	0.87	0.94
	LSTM-FCN	**0.7**	**0.8**	**0.92**	**0.97**
F16	DD	0.6	0.71	0.8	0.9
	TSNR	0.64	0.73	0.82	0.89
	HRNR	0.59	0.73	0.84	0.91
	ResLSTM	0.66	0.78	0.86	0.93
	LSTM-FCN	**0.73**	**0.8**	**0.93**	**0.98**
factory	DD	0.57	0.7	0.79	0.87
	TSNR	0.63	0.69	0.78	0.86
	HRNR	0.62	0.72	0.82	0.9
	ResLSTM	0.68	0.79	0.85	0.92
	LSTM-FCN	**0.72**	**0.84**	**0.92**	**0.95**

TABLE VI. THE PESQ SCORE FOR MMSE-LSA ESTIMATOR USING EACH OF THE A PRIORI SNR ESTIMATORS

Noise type	$\hat{\xi}(n,k)$	SNR level(dB)			
		-5	**0**	**5**	**10**
babble	DD	74.2	81.4	89.7	96.1
	TSNR	75.1	84.5	88.9	95.9
	HRNR	76.2	85.2	90.1	95.7
	ResLSTM	79.6	88.1	92.6	96.7
	LSTM-FCN	**83.7**	**91**	**95.6**	**97.1**
street	DD	72.9	77.9	86.1	91.2
	TSNR	73.1	78.2	85.4	91.3
	HRNR	71.8	77.5	85.7	90.9
	ResLSTM	75.9	81.1	86.9	93.4
	LSTM-FCN	**78.6**	**86.2**	**89.2**	**94.5**
F16	DD	70.1	76.3	84.1	88.4
	TSNR	69.9	77.1	83.6	88.1
	HRNR	71.2	75.4	84.2	89.2
	ResLSTM	76.3	81.9	88.6	90.1
	LSTM-FCN	**80.1**	**86.3**	**91.1**	**92.8**
factory	DD	71.3	75.3	80.2	89.1
	TSNR	68.6	75.8	81.4	90.2
	HRNR	67.8	74.6	82	88.3
	ResLSTM	75.3	80.7	84.3	93.4
	LSTM-FCN	**84.1**	**88.4**	**92.5**	**96.3**

I. Enhanced Speech Spectrograms

To generate the noisy speech, babble noise at an SNR level of 5 dB is mixed with the clean speech. The noisy speech is then enhanced by each of the MMSE-STSA estimators. The clean and noisy speech magnitude spectrograms are shown in Fig. 6 (a) and (b), respectively. The enhanced speech magnitude spectrograms produced by the MMSE-STSA estimator using the DD approach, TSNR technique, HRNR technique, ResLSTM framework, and LSTM-FCN framework are shown in Fig. 7. It can be seen that the MMSE-STSA estimator using the LSTM-FCN framework produced enhanced speech with less residual noise than other estimators. The enhanced speech produced by the LSTM-FCN framework (Fig. 7 (e)) demonstrates significantly less musical noise and speech distortion than the ResLSTM framework (Fig. 7 (d)). To evaluate the performance of the proposed algorithm, same regions are highlighted in Fig. 7 (d) and (e). Clearly, it can be seen in Fig. 7 (d) that the ResLSTM framework heavily distorted these same regions. Also, the MMSE-STSA estimator using DD approach exhibits poor speech enhancement performance.

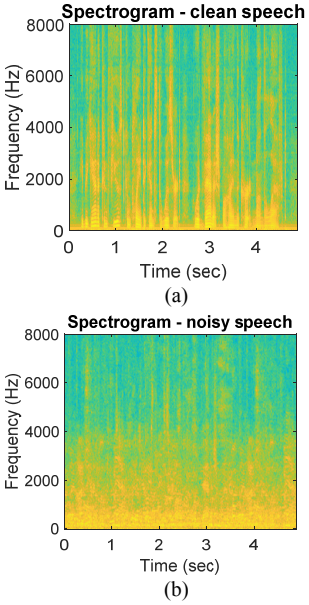
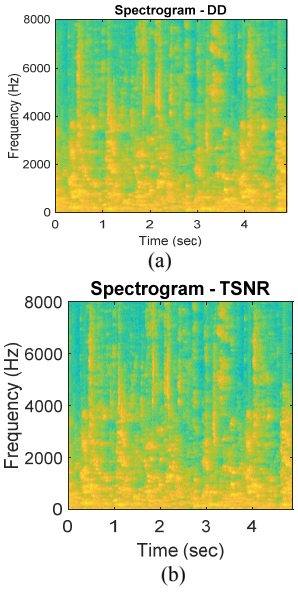


Figure 6. Measuring geometry relationship between target and sensors platform



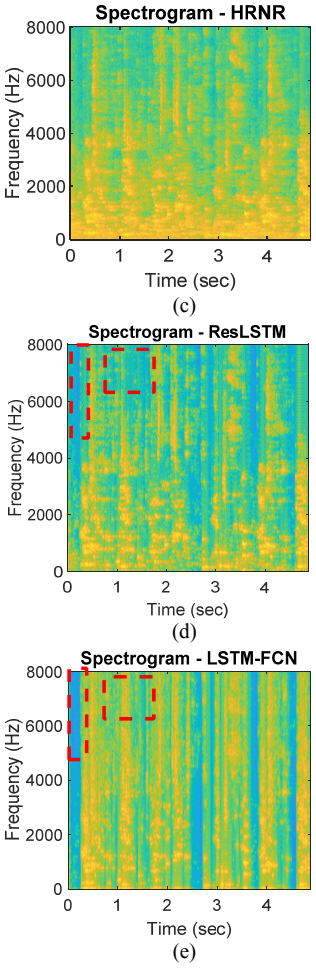


Figure 7. The spectrograms of the enhanced speech (a) DD approach, (b)TSNR technique, (c) HRNR technique, (d) ResLSTM framework, and (e) LSTM-FCN framework

VI. Conclusion

In this paper, a deep MMSE-based speech enhancement problem is investigated. The performance of an MMSE-based speech enhancement approach depends on the accuracy of the used *a priori* SNR estimator. An LSTM-FCN deep neural network is utilized to estimate the *a priori* SNR. The proposed approach makes no assumptions about the characteristics of the noise. Moreover, it can estimate sudden changes in the *a priori* SNR level. The proposed MMSE-based speech enhancement is evaluated for both real-world non-stationary and coloured noise sources. Experimental results indicate that the proposed speech enhancement approach has gained better performance than DD approach, TSNR technique, HRNR technique, and ResLSTM framework. In future work, we will research on how to optimize the LSTM-FCN model architecture and hyperparameters such as epochs, learning rate, optimizer, and loss function. In addition, if the *a priori* SNR changes during speech presence, this change can only be detected with a delay. This may be investigated in future work to achieve a better improvement in speech enhancement performance.

References

[1] S. K. Roy, A. Nicolson, K. K. Paliwal, "DeepLPC: A deep learning approach to augmented Kalman filter-based single-channel speech enhancement," IEEE Access, vol. 9, no. 4, pp. 64524-64538, 2021. doi:10.1109/ACCESS.2021.3075209

[2] Z. Q. Wang, P. Wang, D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 28, no. 5, pp. 1778 -1787, 2020. doi:10.1109/TASLP.2020.2998279

[3] S. Othman, A. Mohamed, A. Abouali, Z. Nossair, "Lossy compression using adaptive polynomial image encoding," Advances in Electrical and Computer Engineering, vol.21, no.1, pp.91-98, 2021. doi:10.4316/AECE.2021.01010

[4] T. G. Yadava, H. S. Jayanna, "Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing," International Journal of Speech Technology, vol. 22, no. 3, pp. 639-648, 2019. doi:10.1007/s10772-018-9506-9

[5] Y. Zhang, Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," Speech Communication, vol. 55, no. 4, pp. 509-522, 2013. doi:10.1016/j.specom.2012.09.005

[6] Y. Ephraim, D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. on Acoustics, Speech, and Signal processing, vol. 32, no. 6, pp. 1109-1121, 1984. doi:10.1109/TASSP.1984.1164453

[7] P. C. Loizou, Speech enhancement: Theory and practice. CRC press, 2007

[8] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussian, S. A. R. Al-Haddad, W. A. Jass, "Low-distortion MMSE speech enhancement estimator based on Laplacian Prior," IEEE Access, vol. 5, no. 4, pp. 9866-9881, 2017. doi:10.1109/ACCESS.2017.2699782

[9] A. Brown, S. Garg, J. Montgomery, "Automatic and efficient denoising of bioacoustics recordings using MMSE STSA," IEEE Access, vol. 6, no. 12, pp. 5010-5022, 2017. doi:10.1109/ACCESS.2017.2782778

[10] Q. Zhang, M. Wang, Y. Lu, M. Idrees, L. Zhang, "Fast nonstationary noise tracking based on log-spectral power MMSE estimator and temporal recursive averaging," IEEE Access, vol. 7, no. 6, pp. 80985-80999, 2019. doi:10.1109/ACCESS.2019.2923680

[11] C. Plapous, C. Marro, L. Mauuary, P. Scalart, "A two-step noise reduction technique," in IEEE International Conf. on Acoustics, Speech, and Signal Processing, Montreal, 2004, pp. 289-292. doi:10.1109/ICASSP.2004.1325979

[12] C. Plapous, C. Marro, P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2098-2108, 2006. doi:10.1109/TASL.2006.872621

[13] Y. G. Thimmaraja, B. Nagaraja, H. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC," International Journal of Speech Technology, vol. 24, no. 1, pp. 165-172, 2021. doi:10.1007/s10772-020-09786-9

[14] F. Bellili, R. Meftehi, S. Affes, A. Stephenne, "Maximum likelihood SNR estimation of linearly-modulated signals over time-varying flat-fading SIMO channels," IEEE Trans. on Signal Processing, vol. 63, no. 2, pp. 441-456, 2014. doi:10.1109/TSP.2014.2364017

[15] C. Breithaupt, T. Gerkmann, R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in IEEE International Conf. on Acoustics, Speech and Signal Processing, Las Vegas, 2008, pp. 4897-4900. doi:10.1109/ICASSP.2008.4518755

[16] V. Timcenko, S. Gajin, "Machine learning enhanced entropy-based network anomaly detection," Advances in Electrical and Computer Engineering, vol.21, no.4, pp.51-60, 2021. doi:10.4316/AECE.2021.04006

[17] A. Albu, R. E Precup, T. A Teban, "Results and challenges of artificial neural networks used for decision-making in medical applications," FACTA Universitatis Series: Mechanical Engineering, vol. 17, no. 3, pp. 285-308, 2019. doi:10.22190/FUME190327035A

[18] T. Zhang, F. Xu, T. Wu, "A software tool for spiking neural P systems," Romanian Journal of Information Science and Technology, vol. 23, no. 1, pp. 84-92, 2020

[19] E. L. Hedrea, R. E. Precup, R. C. Roman, E. M. Petriu, "Tensor product-based model transformation approach to tower crane systems modeling," Asian Journal of Control, vol. 23, no. 3, pp. 1313-1323, 2021. doi:10.1002/asjc.2494

[20] J. B. Awotunde, R. O. Ogundokun, F. E. Ayo, O. E. Matiluko, "Speech segregation in background noise based on deep learning," IEEE Access, vol. 8, no. 9, pp. 169568-169575, 2020. doi:10.1109/ACCESS.2020.3024077

[21] J. Kim, M. Hahn, "Speech enhancement using a two-stage network for an efficient boosting strategy," IEEE Signal Processing Letters, vol. 26, no. 5, pp. 770-774, 2019. doi:10.1109/LSP.2019.2905660

[22] X. L. Zhang, D. Wang, "Boosting contextual information for deep neural network based voice activity detection," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 24, no. 2, pp. 252-264, 2015. doi:10.1109/TASLP.2015.2505415

[23] Q. Wang, J. Du, L. R. Dai, C. H. Lee, "A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1185-1197, 2018. doi:10.1109/TASLP.2018.2817798

[24] I. Ahmed, S. Alam, J. Hossain, G. Kaddoum, "Deep learning for MMSE estimation of a Gaussian source in the presence of Bursty impulsive noise," IEEE Communications Letters, vol. 25, no. 4, pp. 1211-1215, 2020. doi:10.1109/LCOMM.2020.3045665

[25] A. Nicolson, K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," Speech Communication, vol. 111, no. 8, pp. 44-55, 2019. doi:10.1016/j.specom.2019.06.002

[26] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 28, no. 4, pp. 1404-1415, 2020. doi:10.1109/TASLP.2020.2987441

[27] F. Karim, S. Majumdar, H. Darabi, S. Chen, "LSTM fully convolutional networks for time series classification," IEEE Access, vol. 6, no. 12, pp. 1662-1669, 2017. doi:10.1109/ACCESS.2017.2779939

[28] F. Karim, S. Majumdar, H. Darabi, "Insights into LSTM fully convolutional networks for time series classification," IEEE Access, vol.7, no. 5, pp. 67718-67725, 2019. doi:10.1109/ACCESS.2019.2916828

[29] R. A. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in IEEE International Conf. on Acoustics, Speech and Signal Processing, Taipei, 2009, pp. 4421-4424. doi:10.1109/ICASSP.2009.4960610

[30] R. C. Hendriks, R. Heusdens, J. Jensen, "MMSE based noise PSD tracking with low complexity," in IEEE International Conf. on Acoustics, Speech and Signal Processing, Dallas, 2010, pp. 4266-4269. doi:10.1109/ICASSP.2010.5495680

[31] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 33, no. 2, pp. 443-445, 1985. doi:10.1109/TASSP.1985.1164550

[32] I. Goodfellow I, Y. Bengio, A. Courville, Deep learning. MIT press, pp. 257-267, 2016

[33] T. Y. Hsiao, Y. C. Chang, H. H. Chou, C. T. Lin, "Filter-based deep-compression with global average pooling for convolutional networks," Journal of Systems Architecture, vol. 95, no. 5, pp. 9-18, 2019. doi:10.1016/j.sysarc.2019.02.008

[34] C. LeaEmail, R. Vidal, A. Reiter, G. D. Hager, "Temporal Convolutional Networks: A unified approach to action segmentation," in European Conf. on Computer Vision, Amsterdam, 2016, pp. 47-54. doi:10.1007/978-3-319-49409-8_7

[35] P. Dhruv, S. Naskar, "Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review," Machine Learning and Information Processing, vol. 1101, no. 3, pp. 367-381, 2020. doi:10.1007/978-981-15-1884-3_34

[36] X. Wu, X. Shen, J. Zhang, Y. Zhang, "A wind energy prediction scheme combining cauchy variation and reverse learning strategy," Advances in Electrical and Computer Engineering, vol.21, no.4, pp.3-10, 2021, doi:10.4316/AECE.2021.04001

[37] A. Barakat, P. Bianchi, "Convergence and dynamical behavior of the ADAM algorithm for non-convex stochastic optimization," SIAM Journal on Optimization, vol. 31, no. 1, pp. 244-274, 2021. doi:10.1137/19M1263443

[38] P. Netrapalli, "Stochastic gradient descent and its variants in machine learning," Journal of the Indian Institute of Science, vol. 99, no. 2, pp. 201-213, 2019. doi:10.1007/s41745-019-0098-4

[39] R. V. K. Reddy, B. S. Rao, K. P. Raju, "Handwritten Hindi digits recognition using convolutional neural network with RMSprop optimization," in Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2018, pp. 45-51. doi:10.1109/ICCONS.2018.8662969

[40] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 7, pp. 2125-2136, 2011. doi:10.1109/TASL.2011.2114881