SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 23, No. 2, June 2026, 305-320
UDC: 811.214.21:621.391.883.2 **DOI: https://doi.org/10.2298/SJEE2602305G**
Original scientific paper

Two-stage Adaptive Robust Noise Cancellation System for a Hindi Speech Database

Vijay Kumar Gupta[1], Mahesh Chandra[2], Alok Kumar[1], Subodh Kumar Yadav[1]

Abstract: This paper presents a two-stage adaptive noise cancellation system for noise cancellation of Hindi speech signals corrupted by babble and factory noise, under different input signal-to-noise ratio (SNR) levels. The corrupted signals with different SNR levels are passed through three different noise cancellation systems: an adaptive noise cancellation system with a fast block least mean squares algorithm (FBLMS) algorithm; a fully connected convolutional deep neural network (FCCDNN); and a two-stage system with an adaptive noise cancellation system consisting of the FBLMS algorithm followed by FCCDNN. The performance of each noise cancellation system is evaluated and compared. The parameters such as output SNR, perceptual evaluation of speech quality and short-time objective intelligibility are used for testing the intelligibility and quality of the recovered speech signals. The proposed system shows a maximum SNR improvement of 13.997 and 15.2965 dB at an input SNR of −5 dB for factory noise and babble noise, respectively, with significant increments in both of the evaluation metrics. The proposed system outperforms speech enhancement systems based on FCCDNN and FBLMS in terms of all performance parameters.

Keywords: Fully connected convolutional deep neural network, Perceptual evaluation of speech quality, Short-time objective intelligibility, Signal to noise ratio.

1 Introduction

Environmental noise is defined as an unwanted signal which when added to speech signal results in a reduction in the quality and intelligibility of speech

[1]Government Engineering College West Champaran, Bihar Engineering University, Bihar India
guptavk76@gmail.com, https://orcid.org/0000-0003-4261-9066
sayes2aloksahu@gmail.com, https://orcid.org/0009-0002-7018-5007
subodh.hit@gmail.com, https://orcid.org/0009-0001-6330-4242
[2]REVA University Bengaluru, India
shrotriya69@gmail.com, https://orcid.org/0000-0001-9892-5527

Colour versions of the one or more of the figures in this paper are available online at https://sjee.ftn.kg.ac.rs

signals. For someone using headphones or another hearing device while sitting in a restaurant, travelling on a train, or working in an office alongside many people who are talking, it is desirable to hear sound without distractions. Adaptive noise cancellation techniques have generally been used in these scenarios to reduce unwanted background noise; over the past few years, however, these techniques have mostly been replaced by deep-learning-based noise cancellation techniques. An adaptive noise cancellation system [1, 2] incorporates sophisticated adaptive algorithms such as least mean squares LMS and its variants, recursive least squares (RLS) and its variants, and frequency-domain adaptive algorithms. It continuously monitors the surrounding ambient sounds and then produces anti-noise output that is adjusted accordingly to reduce noise. The sound device captures the audio signals via built-in microphones in order to carry out real-time analysis and adjustments to optimise the performance of the noise cancellation system. Using this approach, the system can adapt itself to different noise environments and noise levels [1]. A block diagram of a system of this type is given in Fig. 1.

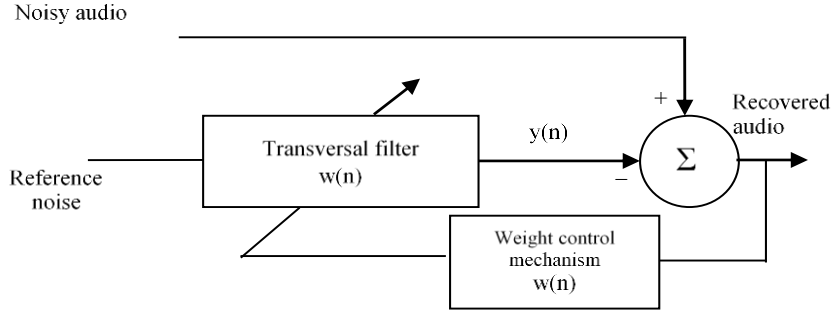


Fig. 1 – *Block diagram of an adaptive noise cancellation system.*

Compared to a time-domain adaptive filter, the numerical complexity of a frequency-domain adaptive filter (FDAF) [3, 4] is lower, as it uses an overlap-and-save implementation method. Their fast convergence and low computational complexity also make FDAF algorithms superior. Numerous methods have been developed by prior researchers for speech enhancement based on perceptual evaluation and speech quality [5, 6]. Deep learning [7 – 10] has become a popular term over the past few years. Recent research has focused on speech enhancement using algorithms of this type, which represent a subset of machine learning in which neural networks are used, with applications in regression, classification tasks and many others. A neural network is a model based on biological neuroscience, in which several artificial neurons are simulated using an input layer, several hidden layers and an output layer. The word 'deep' reflects the use of multiple hidden layers in the network, through which the data is transformed.

There are several variants, known as supervised, semi-supervised or unsupervised learning methods, which are used for different applications such as computer vision, speech enhancement and recognition, natural language processing, etc. Commonly used deep learning network architectures include feed-forward neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks [11]. In the scenario considered here, speech enhancement using deep neural networks has shown considerable success.

The temporal information associated with speech signals can be easily learned using a CNN; for example, a generative adversarial network is composed of two different neural networks that compete and learn together during training. The computational complexity of a CNN is significantly reduced compared with that of a DNN [7, 11]. This paper presents a fully connected deep neural network with a convolutional layer. A combination of an FBLMS-based FDAF using the overlap–save method and cascaded with a FCCDNN is also proposed for speech enhancement of a corrupted speech signal. FDAFs have shown good results for convergence, noise control and nonlinear modelling [12 – 14]. The FBLMS is a computationally efficient implementation of the block LMS (BLMS) algorithm, with a computational burden that is significantly lower than for the LMS algorithm, since the fast Fourier transform (FFT) is used to calculate both the block filtering output and the update terms in the frequency domain [19].

2 FBLMS-Algorithm-Based Speech Enhancement System

In FBLMS, block processing of the data samples in the frequency domain is applied to achieve significant reductions in the computational complexity of LMS adaptive filters. To perform a linear convolution between two sequences, one of finite length and another of infinite length, a sequence overlap–save sectioning method is generally used, in which the sequence of infinite length is partitioned into small blocks of a fixed length N. The adaptive weight w(n) of N element is appended by N zeros, to give a length of 2N, and a 2N-point FFT, i.e., W(K), is computed. In the infinite length input sequence, P(n) is changed to a length of 2N by concatenating the most recent N data block with the previous block of N samples. The 2N-point FFT of this sequence x(n) is computed, given by X(K). The block of output samples is then given by $Y(K) = X(K)W(K)$. A 2N-point inverse FFT (IFFT) of this Y(K) provides the output frame. The last N points of the output frame are taken as output, while the first N points are discarded. A block diagram of the overlap–save sectioning method is given in Fig. 2 [1].

A block diagram of a speech enhancement system based on the FBLMS algorithm is given in Fig. 3 [1], where x (n) is the reference noise signal, the desired signal d (n) is a noisy speech signal, and e (n) is the recovered speech signal. The noise in the noisy signal should be correlated with the reference noise. μ is the step size parameter.

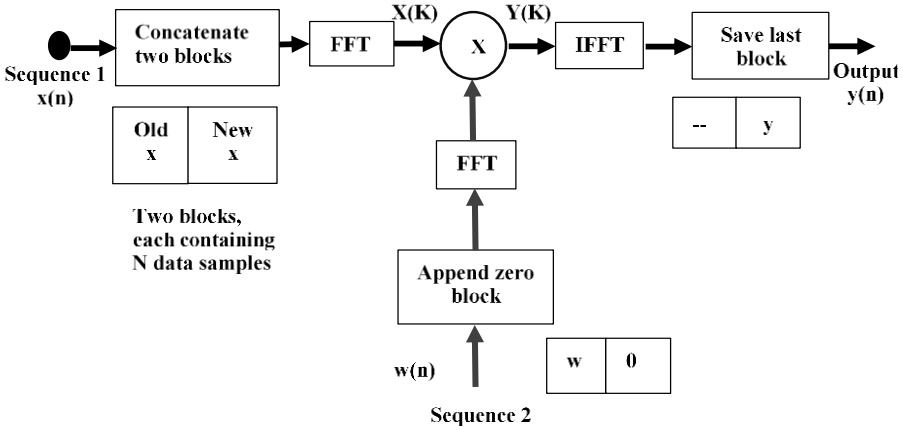


Fig. 2 – *Diagram illustrating the overlap–save sectioning method.*

For a block of L input data samples and filter length N, one block of the input vector, the desired output vector, the error and filter output at some instant n=k will be denoted as shown in (1) to (4).

$$\mathrm{x}(\mathrm{k}) = \left[\mathrm{x}(\mathrm{kL}-\mathrm{N}+1)\ \ \mathrm{x}(\mathrm{kL}-\mathrm{N}+2)\cdots\mathrm{x}(\mathrm{kL}+\mathrm{L}-1)\right]^{\mathrm{T}}, \tag{1}$$

$$\mathrm{d}(\mathrm{k}) = \left[\mathrm{d}(\mathrm{kL})\ \ \mathrm{d}(\mathrm{kL}+1)\cdots\mathrm{d}(\mathrm{kL}+\mathrm{L}-1)\right]^{\mathrm{T}}, \tag{2}$$

$$\mathrm{y}(\mathrm{k}) = \left[\mathrm{y}(\mathrm{kL})\ \ \mathrm{y}(\mathrm{kL}+1)\cdots\mathrm{y}(\mathrm{kL}+\mathrm{L}-1)\right]^{\mathrm{T}}, \tag{3}$$

$$\mathrm{e}(\mathrm{k}) = \mathrm{d}(\mathrm{k}) - \mathrm{y}(\mathrm{k}). \tag{4}$$

The frequency domain transforms (i.e. FFT) of the input and output vectors are represented by X(K) and Y(K). The weight updating process is based on (1) – (3). The output of the filter is given by (5):

$$\begin{bmatrix} -- \\ y(n) \end{bmatrix} = IFFT\left(FFT\left(\begin{bmatrix} w(n) \\ 0 \end{bmatrix} \right) \times FFT\left(x(n)\right) \right) \tag{5}$$

and the correlation vector $\hat{\nabla}$ between e(n) and x(n) is given by (6):

$$\begin{bmatrix} \hat{\nabla} \\ -- \end{bmatrix} = IFFT\left(FFT\left(\begin{bmatrix} 0 \\ e(n) \end{bmatrix} \right) \times \overline{FFT\left(x(n)\right)} \right). \tag{6}$$

The filter weight update process is then given by (7):

$$FFT\begin{bmatrix} w(n+1) \\ 0 \end{bmatrix} = FFT\left(\begin{bmatrix} w(n) \\ 0 \end{bmatrix} \right) \times \mu FFT\left(\begin{bmatrix} \hat{\nabla} \\ 0 \end{bmatrix} \right), \tag{7}$$

where '--' represents the first N points that are discarded and the last N points of the output frame are taken as output in (5). Similarly, the last N points are discarded in (6).

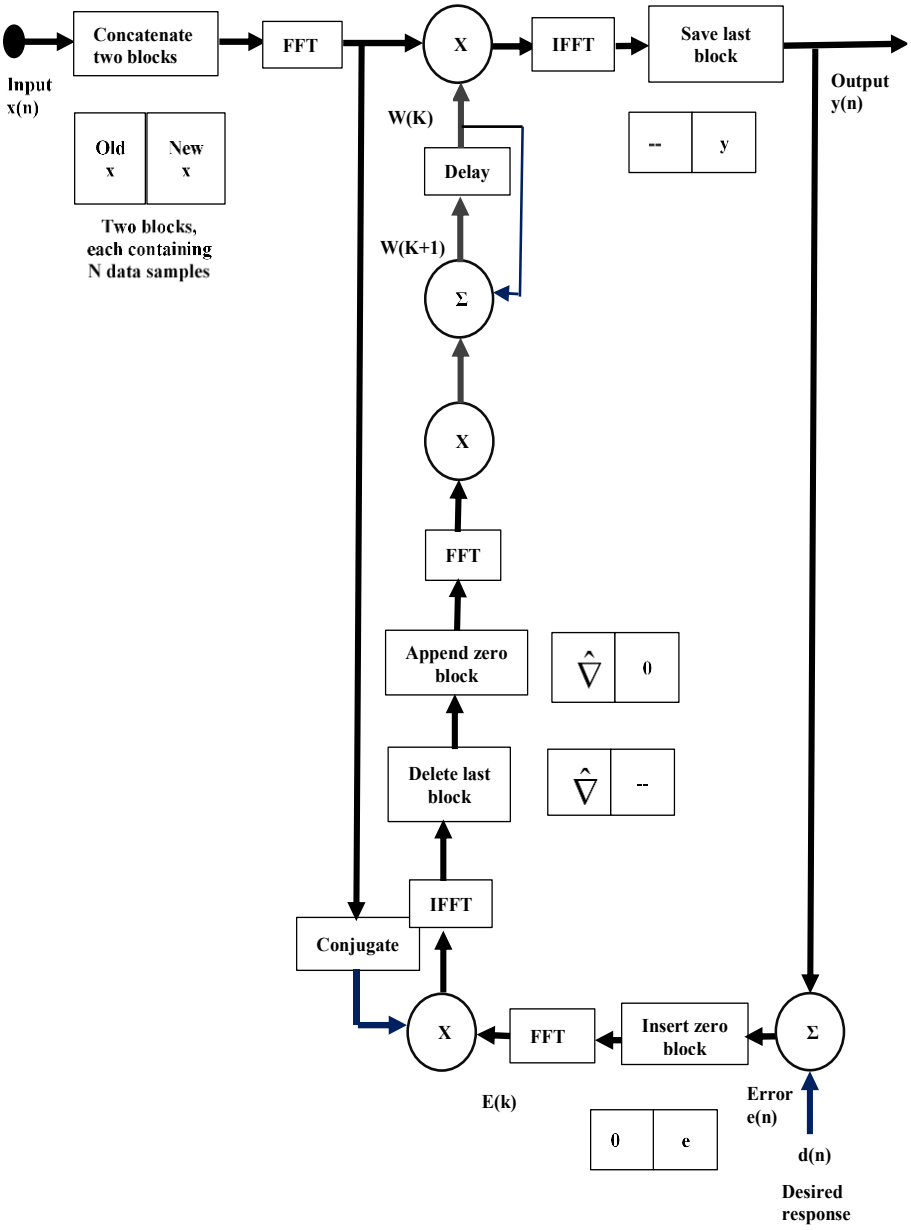


Fig. 3 – *Diagram of the fast block LMS algorithm.*

3 Proposed FCCDNN-Based Speech Enhancement System

The architecture of the FCCDNN model put forward in this paper is illustrated in Fig. 4, and consists of a fully connected layer followed by a convolution layer, ReLU and a max pooling layer. The final stage consists of a fully connected layer followed by a regression layer.

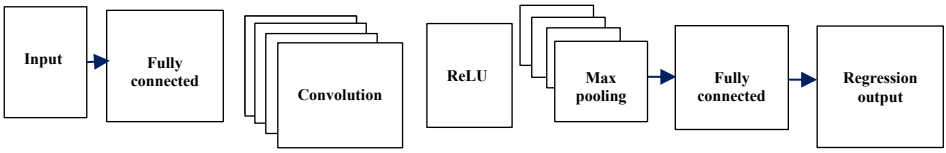


Fig. 4 – *Architecture of FCCDNN.*

The trained FCCDNN model is used for prediction of the magnitude spectrum of clean speech, using the magnitude spectrum of noisy speech. During the training process, a nonlinear relation between the noisy spectrum and clean spectrum, known as a mapping function, is developed to further minimise noise from unknown noisy speech signals, as shown in Figs. 5 and 6.

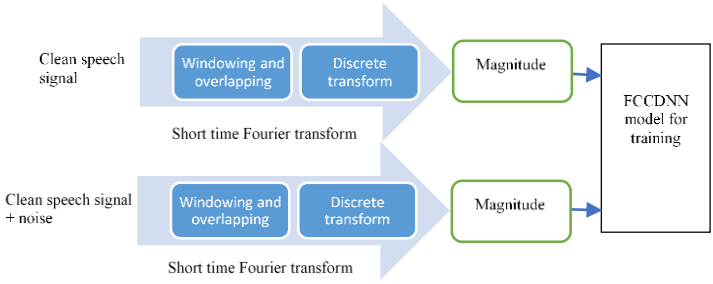


Fig. 5 – *Training of the FCCDNN model.*

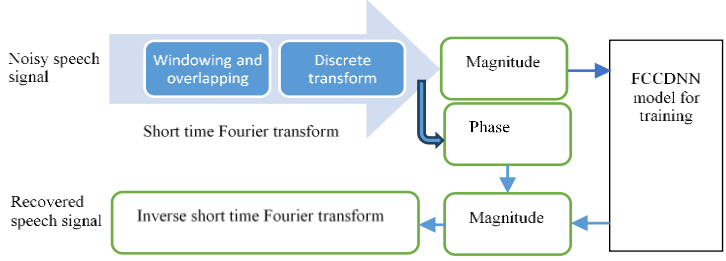


Fig. 6 – *Testing and validation of the trained FCCDNN.*

The proposed speech enhancement system consists of a training stage and a testing stage. The short time Fourier transform (STFT) is applied to both noisy

and clean signals to obtain their magnitude and phase spectra. The magnitude spectrum of noisy speech is used as the final predictor, whereas the magnitude spectrum of the clean speech signal is used as the final target. For the whole speech database predictors and targets are calculated and used to train DNN models. In the testing stage, the system performance is tested using the noisy test database, whereas in the validation stage, an unknown noisy speech signal is applied to the trained FCCDNN model.

First, the magnitude and phase spectra of the noisy speech signal are separated, and only the magnitude spectrum is passed to the trained DNN model. The trained FCCDNN model provides the magnitude spectrum of the recovered speech signal. Inverse STFT is then applied to the magnitude and phase spectra of the noisy speech signal to recover the speech signal, as shown in Fig. 6 [7,8]. A convolutional layer uses a convolution sum rather than relying solely on matrix multiplications, as in traditional neural networks. The neurons in a convolutional layer extract specific features. In this paper, a 2D convolutional layer is used in which sliding convolutional filters are applied to the 2D input to extract the features.

4 Two-Stage Proposed Speech Enhancement System Based on FBLMS Cascaded with FCCDNN

This research demonstrates the success of a two-stage speech enhancement system, and its performance is evaluated and compared as shown in Fig. 7.

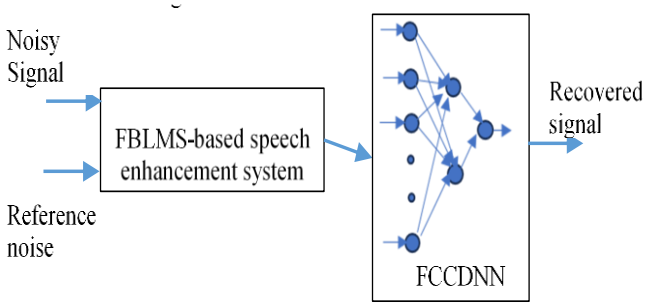


Fig. 7 − *Proposed two-stage speech enhancement system.*

In the first stage, the noisy signal is passed through an FBLMS-based speech enhancement system. In the second stage, the recovered signal is again passed through a trained FCCDNN-based speech enhancement system to remove the residual background noise. The FCCDNN system used in the second stage is trained on a speech database and then tested.

5 Performance Parameters

In general, the improvement in SNR is used to define the performance of a speech enhancement system, as it indicates that the quality of the speech is improved. In the last few years, deep neural networks have achieved substantial improvements in the quality and intelligibility of speech when used with supervised learning-based speech enhancement systems. However, recovered speech with a comparatively high improvement in SNR score compared to clean speech may not have high levels of quality and intelligibility; hence, objective functions based on human perception such as perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) are generally used to evaluate the quality and intelligibility of speech [6]. PESQ is the ITU-T standard representing the objective level of speech quality. Its value ranges from −0.5 to 4.5, with a higher value indicating better perceptual quality. STOI is a scalar quantity, with values ranging from −1 to 1; it measures the intelligibility of a recovered speech signal by comparing it with the clean speech signal, with a higher value corresponding to higher intelligibility [15].

6 Methodology and Experimental Setup

A suitable database of noisy signals should be created to test any speech enhancement system and to evaluate its performance. Noisy speech signals were therefore prepared for the Hindi speech database by combining them with factory and babble noise from the NOISEX-92 database [16]. Both the clean speech samples and noise samples had a 16 kHz sampling frequency. In this experimental setup, the noisy speech signal was passed through three different speech enhancement systems: an FBLMS-based system, an FCCDNN-based system, and a two-stage system. The noisy signal was first passed to the FBLMS-based speech enhancement system, with the output (recovered speech) then passed to the FCCDNN system to further remove the residual noise. The different layers used in FCCDNN system are given in **Table 1**. The clean speech signal dataset was taken from the "Common Voice" database of Hindi speech samples available from www.kaggle.com [18].

Table 1

Layers used in the proposed FCCDNN-based speech enhancement system.

Name	Dimensions and details
Image input	33×8×1 images with 'zero centre' normalisation
Fully connected	1024 fully connected layer
Convolution	512 1×1 convolutions with stride [1 1] and padding [0 0 0 0]
ReLU	ReLU
Max pooling	1×1 max pooling with stride [1 1] and padding [0 0 0 0]
Fully connected	33 fully connected layer
Regression output	Mean squared error

The training process of the FCCDNN model in terms of the root mean square error (RMSE) and the loss function is shown in Fig. 8. RMSE is useful for regression tasks where predictions of continuous values are required. It is also used with a loss function, which guides training and performance measurements. The training dataset consisted of total 1037 samples of Hindi speech, while the testing dataset consisted of 356 samples (i.e. 74% for training and 25% for testing). A total of 15 samples (1%) were used for validation. For STFT, a Hamming window of length 64 was used for framing. In the FCCDNN model, five layers were used, as shown in **Table 1**, and stochastic gradient descent with momentum was used for training, with the momentum set to 0.9000. The initial learning rate, maximum epochs, and mini-batch size were 0.01, 5 and 64, respectively. In FBLMS, the filter order and step size were taken as 20 and 0.1, respectively.

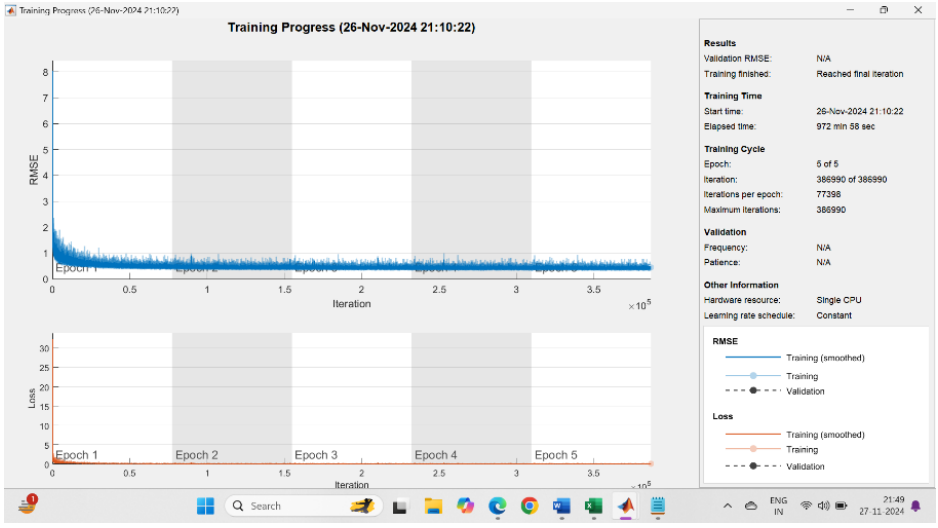
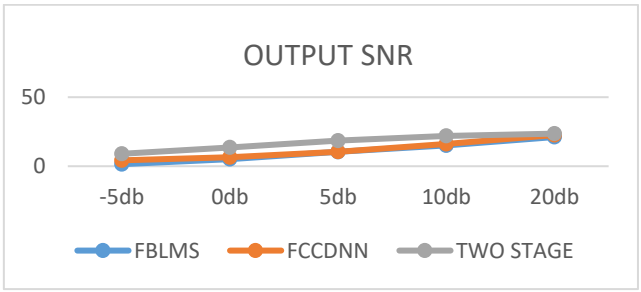


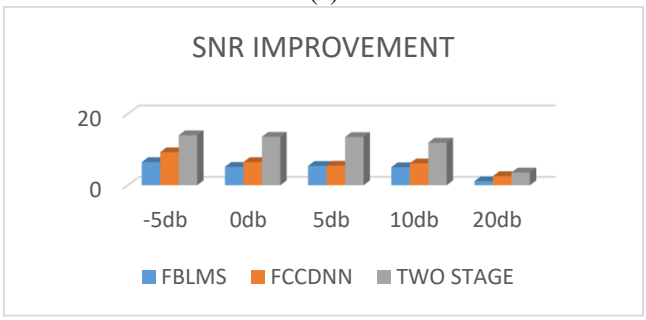
Fig. 8 – *Training process of the FCCDNN model.*

Table 2 presents a performance comparison of the quality and intelligibility of speech in terms of the output SNR, PESQ and STOI when all three speech enhancement systems are applied to speech signals corrupted with factory and babble noise at −5, 0, 5, 10 and 20 dB.

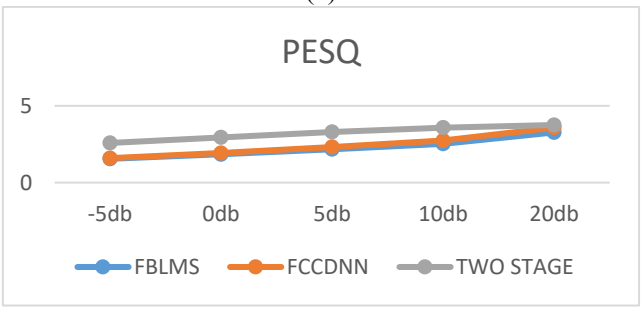
Fig. 9a presents a comparison of the output SNR for speech corrupted by factory noise at different SNR levels when it is passed through all three speech enhancement systems, while Fig. 9b enables a comparison of the improvements in SNR achieved by these systems. Figs. 9c and d present comparisons of the PESQ and STOI results, respectively, for the speech recovered from above systems. Fig. 10 presents the same results but for factory noise.



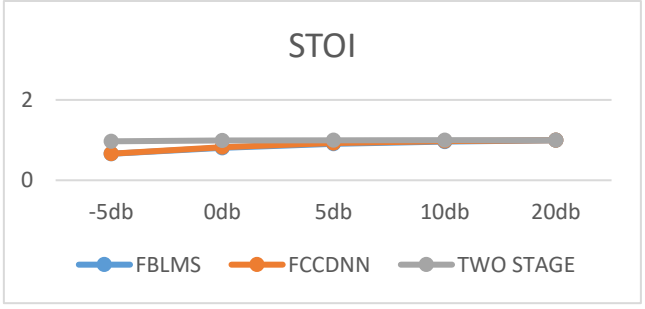
(a)



(b)



(c)



(d)

Fig. 9 – *Performance comparison of FBLMS, FCCDNN and two-stage speech enhancement systems for factory noise:* (a) *output SNR;* (b) *SNR improvement;* (c) *PESQ;* (d) *STOI.*

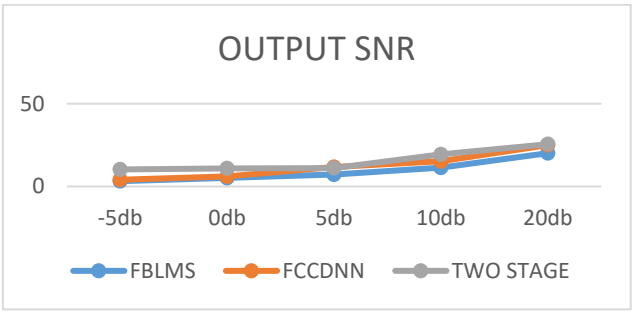
Figs. 9a and 10a show that the proposed system provides higher improvements in output SNR compared to the other two speech enhancement systems for both types of noise at all input SNR levels. From Fig. 9b, it can be seen that the proposed system provides a maximum improvement of 13.997 dB for speech signal corrupted by factory noise at −5 dB input SNR level, a value 4.7405 dB higher than for FCCDNN-based speech enhancement and 7.5152 dB than FBLMS-based speech enhancement.

Table 2

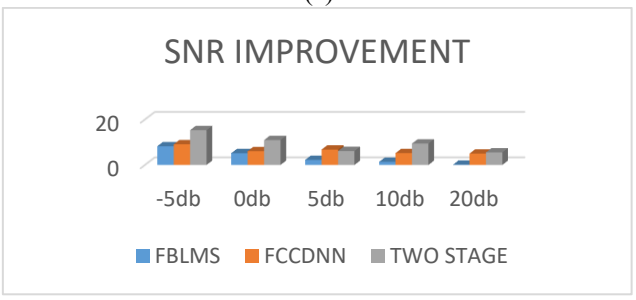
Performance comparison of speech quality and intelligibility in terms of output SNR, PESQ and STOI.

Noise	Input SNR (dB)	Output SNR			PSEQ		
		FBLMS	FCC DNN	Two-stage model	FBLMS	FCC DNN	Two-stage model
Factory	−5	1.4818	4.2565	8.997	1.5508	1.594	2.581
	0	5.1527	6.501	13.5866	1.8453	1.9275	2.9438
	5	10.4224	10.5012	18.5153	2.1703	2.3188	3.2962
	10	15.0551	16.1129	21.9253	2.5263	2.7408	3.5791
	20	21.1315	22.5694	23.578	3.2679	3.5814	3.7508
Babble	−5	3.2075	4.0738	10.2965	1.5465	1.5821	2.7562
	0	5.1738	6.0604	10.9363	1.8508	1.9098	2.8678
	5	7.221	11.7561	11.1202	2.2017	2.3143	2.9056
	10	11.3571	15.2045	19.3908	2.5727	2.7207	3.389
	20	20.112	24.9777	25.5249	3.4277	3.46	3.6144

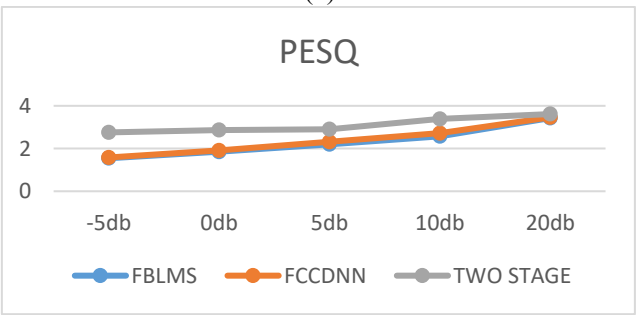
Noise	Input SNR (dB)	STOI		
		FBLMS	FCC DNN	Two-stage model
Factory	−5	0.6592	0.6611	0.9696
	0	0.8067	0.825	0.9891
	5	0.9083	0.9265	0.9955
	10	0.9614	0.9758	0.9971
	20	0.995	0.997	0.9977
Babble	−5	0.6413	0.6408	0.9697
	0	0.7765	0.786	0.9716
	5	0.885	0.901	0.9721
	10	0.949	0.9653	0.9947
	20	0.9941	0.9965	0.9978



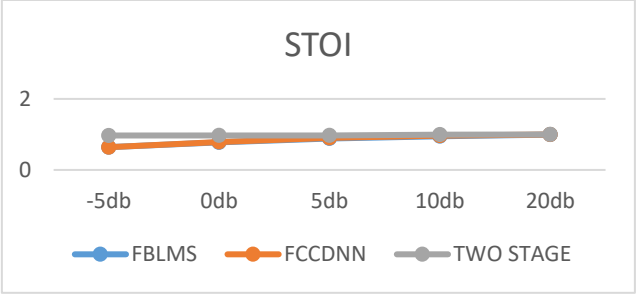
(a)



(b)



(c)



(d)

Fig. 10 – *Performance comparison of FBLMS, FCCDNN and two-stage speech enhancement systems under babble noise:* (a) *output SNR;* (b) *SNR improvement*; (c) *PESQ;* (d) *STOI.*

Similarly, it can be observed from Fig. 10b that the proposed system provides a maximum improvement of 15.2965 dB for a speech signal corrupted by babble noise at −5 dB input SNR level, a result that is 6.2227 dB higher than for FCCDNN-based speech enhancement and 7.089 dB higher than for FBLMS-based speech enhancement. It is also clear that the proposed system gives an improvement in SNR for speech signals corrupted by both types of noise for input at all SNR levels, except for babble noise at 5 dB, where the output is almost constant. This increase in SNR represents the quality improvement in the recovered signal.

In **Table 2**, the values of the PESQ and STIO parameters are also given for the recovered signal from both speech enhancement systems, to enable an objective evaluation of the quality and intelligibility of the speech. It can be observed that both parameters are improved for the proposed system compared to the FCCDNN-based speech enhancement system.

Figs. 9c and 10c represent the improvements in PESQ at all input SNR levels, whereas Figs. 9d and 10d represent the corresponding improvements in STOI. These increases indicate that the recovered speech from proposed system is better than from the FCCDNN-based speech enhancement system.

In addition to validating the proposed system based on reserved samples, it was also tested based on an independent speech database [17]. From this database, one test sample was selected from a range of different sentences in Hindi [17], namely “YAHA SAI LAGHBAG PANCH MEAL DAKSHIN PASCHIM MAI KATGHAR GAON HAI”, to assess the performance of the system. A noisy version of this sentence was prepared by adding factory and babble noise from the NOISEX-92 database [16] to this clean sentence at −5, 0, 5, 10 and 20 dB input SNR levels.

Fig. 11 shows the results for factory noise at 0 dB. Fig. 11a shows the clean speech signal and its spectrogram, Fig. 11b shows the noisy speech signal and its spectrogram, and Fig. 11c shows the output from the proposed speech enhancement system and its spectrogram.

Fig. 12 shows the corresponding results for babble noise at 0 dB: Fig. 12a shows the clean speech signal and its spectrogram, while Fig. 12b shows the noisy speech signal at 0 dB and its spectrogram, and Fig. 12c shows the output from the proposed speech enhancement system and its spectrogram. It can be observed that the output from the proposed system is very similar to the clean speech signals for both forms of noise at all SNR levels.

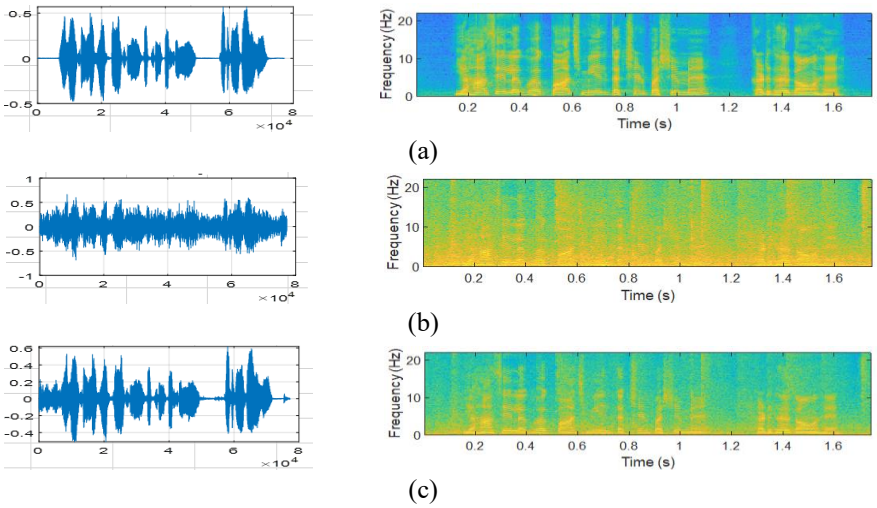


Fig. 11 – *Output for noisy signal corrupted by factory noise at* 0 dB *input SNR:*
(a) *clean speech signal and its spectrogram representation;*
(b) *noisy signal and its spectrogram representation;*
(c) *output from proposed system and its spectrogram representation.*

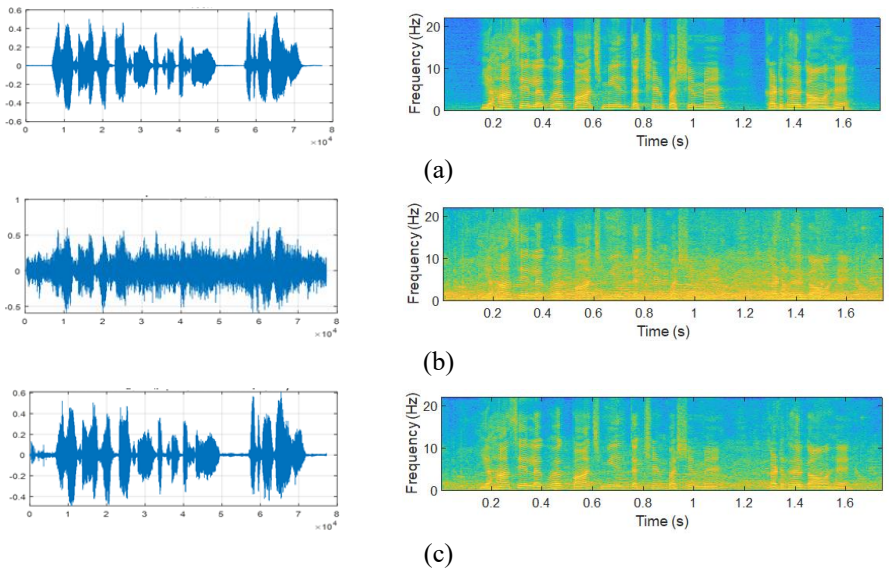


Fig. 12 – *Output for noisy signal corrupted by babble noise at* 0 dB *input SNR level:*
(a) *clean speech signal and its spectrogram representation;*
(b) *noisy signal and its spectrogram representation;*
(c) *output from proposed system and its spectrogram representation.*

7 Conclusion

In this paper, a two-stage system for speech enhancement based on a CNN architecture has been presented, consisting of a FBLMS network followed by an FCCDNN model. The quality and intelligibility of speech were evaluated in terms of the SNR, PESQ and STOI for a Hindi speech signal corrupted by babble and factory noise at −5, 0, 5, 10 and 20 dB input SNR levels. The performance of the proposed system was found to be better than FCCDNN- and FBLMS-based speech enhancement systems for both types of noise at all input SNR levels. Even at a −5 dB input SNR level, the proposed system showed maximum improvements in SNR of 13.997 and 15.2965 dB for factory noise and babble noise, respectively. The improvement is also represented in the spectrograms of the clean, noisy, and recovered speech signals for both types of noise at a 0 dB input SNR level.

8 References

[1] S. Haykin: Adaptive Filter Theory, 4th Edition, Pearson Education, Delhi, 2002.

[2] H. M. Lee, Y. Hua, Z. Wang, K. M. Lim, H. P. Lee: A Review of the Application of Active Noise Control Technologies on Windows: Challenges and Limitations, Applied Acoustics, Vol. 174, March 2021, p. 107753.

[3] M. Venkata Sudhakar, M. Prabhu Charan, G. Naga Pranai, L. Harika, P. Yamini: Audio Signal Noise Cancellation with Adaptive Filter Techniques, Materials Today: Proceedings, Vol. 80, Part 3, 2023, pp. 2956 – 2963.

[4] Kun Shi, Xiaoli Ma: A Frequency Domain Step-Size Control Method for LMS Algorithms, IEEE Signal Processing Letters, Vol. 17, No. 2, 2010, pp. 125 – 128.

[5] S.- W. Fu, C.- F. Liao, Y. Tsao: Learning with Learned Loss Function: Speech Enhancement with Quality-Net to Improve Perceptual Evaluation of Speech Quality, IEEE Signal Processing Letters, Vol. 27, May 2019, pp 26 – 30.

[6] X. Dong, D. S. Williamson: Towards Real-World Objective Speech Quality and Intelligibility Assessment Using Speech-Enhancement Residuals and Convolutional Long Short-Term Memory Networks, The Journal of the Acoustical Society of America, Vol. 148, No. 5, November 2020, pp. 3348 – 3359.

[7] G. E. Hinton, S. Osindero, Y.- W. The: A Fast Learning Algorithm for Deep Belief Nets, Neural Computation, Vol. 18, No. 7, July 2006, pp. 1527 – 1554.

[8] I. H. Sarker: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions, SN Computer Science, Vol. 2, No. 6, November 2021, p. 420.

[9] F. E. Wahab, Z. Ye, N. Saleem, R. Ullah: Compact Deep Neural Networks for Real-Time Speech Enhancement on Resource-Limited Devices, Speech Communication, Vol 156, January 2024, p. 103008.

[10] P. Kim: MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence, 1st Edition, Apress, Berkeley, Seoul, 2017.

[11] S. Haykin: Neural Networks and Learning Machines, 3rd Edition, Pearson, New York, Boston, San Francisco, 2010.

[12] F. Yang, J. Yang: Optimal Step-Size Control of the Partitioned Block Frequency-Domain Adaptive Filter, IEEE Transactions on Circuits and Systems II: Express Briefs, Vol. 65, No. 6, June 2018, pp. 814 – 818.

[13] F. Yang, G. Enzner, J. Yang: New Insights into Convergence Theory of Constrained Frequency-Domain Adaptive Filters, Circuits, Systems, and Signal Processing, Vol. 40, No. 4, April 2021, pp. 2076 – 2090.

[14] D. Comminiello, A. Nezamdoust, S. Scardapane, M. Scarpiniti, A. Hussain, A. Uncini: A New Class of Efficient Adaptive Filters for Online Nonlinear Modeling, IEEE Transactions on Systems, Man, and Cybernetics: Systems ,Vol 53, No. 3, March 2023, pp. 1384 – 1396.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra: Perceptual Evaluation of Speech Quality (PESQ)- A New Method for Speech Quality Assessment of Telephone Networks and Codecs, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, May 2001, pp. 749 – 752.

[16] A. Varga, H. J. M. Steeneken, D. Jones: The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition System, Reports of NATO Research Study Group (RSG.10), 1992.

[17] K. Samudravijaya, P. V. S. Rao, S. S. Agrawal: Hindi Speech Database, Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China, October 2000, p. 847.

[18] Kaggle: Common Voice, Available at:

https://www.kaggle.com/datasets/mozillaorg/common-voice

[19] J. Lu, X. Qiu, H. Zou: A Modified Frequency-Domain Block LMS Algorithm with Guaranteed Optimal Steady-State Performance, Signal Processing, Vol. 104, November 2014, pp. 27 – 32.