Cascaded speech enhancement system using deep learning method

Kavitha A[1], Mahesh Chandra[2], Vijay Kumar Gupta[3]
[1,2]School of Electronics and Communication Engineering, REVA University, Bangalore, India
[3]Department of Electronics and Communication Engineering, Government Engineering College West Champaran Bihar, Bihar, India

Article Info

Article history:

Received Oct 12, 2025
Revised Dec 5, 2025
Accepted Jan 15, 2026

Keywords:

Fully connected deep neural network
Least mean square algorithm
SGDM optimizer
Speech enhancement
Wavelet domain adaptive filter

ABSTRACT

Here, a two-stage cascaded noise minimization from noisy speech is proposed for noise cancellation from highly corrupted speech signals. In the first stage, corrupted speech is passed through speech enhancement system based on wavelet domain adaptive filter using least mean square algorithm (WDAF-LMS) and performance is evaluated for noisy signal corrupted by babble noise, car noise and machine gun noises. Then this output is given to second stage for further improvement. This is fully connected deep neural network using stochastic gradient descent with momentum optimizer (FCDNN-SGDM) used to improve the quality of speech signal. The system is tested for highly corrupted noisy speech signals where noise signal power level is equal to or more than clean signal power. Input signal-to-noise ratio (SNR) level is taken as 0 dB and -5 to -13 dB. The proposed system improved the quality and intelligibility of speech at all SNR levels for all three noises.

Corresponding Author:

Vijay Kumar Gupta
Department of Electronics and Communication Engineering, Government Engineering College West Champaran Bihar
Railway Station, Opposite Kumarbagh, Kumarbagh, Bihar 845450, India
Email: guptavk76@gmail.com

1. INTRODUCTION

In voice communication systems, speech enhancement is required for improving the quality and intelligibility of speech signals which are captured by these devices under the presence of background or some specific noise [1] along with clean Hindi speech database [2]. The goal of these speech enhancement systems is to make noisy speech more clear and pleasant in listening through audio devices after application of noise minimization techniques. The perceptual quality and intelligibility of speech signal should be retained. One of the important applications is remote work meetings, *i.e.*, teleconferencing or video conferencing where clear voice from audio devices is an essential requirement for effective communication as well as for user satisfaction.

Traditionally, accurate noise estimation-based methods such as spectral subtraction and Wiener filtering. were used for denoising. The adaptive noise cancellation systems use popular adaptive algorithms such as least mean square (LMS), recursive least square (RLS) algorithms and their variants in time domain. LMS algorithm is mostly used due to its simplicity and convergence guaranty in stationary environments, but it may take long time to converge. Whereas RLS algorithm shows Faster convergence rates, but it requires large computational resources, often it is too large for real-time implementation. RLS algorithm performs better for non-stationary environments. Affine projection algorithm and its variants are also used for noise cancellation since it has fast convergence like recursive RLS and low complexity like LMS algorithm. The transform domain adaptive algorithms offer superior performance over conventional adaptive algorithms. These algorithms show lower computational complexity and better convergence as compared to time domain algorithms. Implementation of noise cancellation system proved more effectiveness for applications having impulse response of long duration. Due to highly correlated input signals, the convergence of LMS algorithm [3], [4] in time domain significantly degrades. Because of orthogonal property, transform domain algorithms provide faster convergence speed with low computational complexity. This is since transform domain adaptive filters [5], [6] uses decorrelation properties of some well-known signal transforms, such as the discrete Fourier transform (DFT), discrete cosine transform (DCT) and wavelet transform (WT) domain [7], [8]. WT domain LMS Newton adaptive filtering algorithm [9] has proved better than other domain algorithms for both first order and second order autoregressive (AR) process. In this paper the analysis of stability, misadjustment, and convergence performance has been done. The coefficients belonging to certain sub-bands of WT based LMS [10] are dynamically selected for the update based on largest decrement of the mean-square deviation. It resulted in a fast convergence speed and a low steady-state error as compared to simple WT based LMS. Other researchers have also proved that WT domain filters [11], [12] have also proved their superiority for applications in speech noise reduction and acoustic echo cancellation.

In the recent past years, many deep learning network-based speech enhancement systems have shown their superiority over the traditional speech denoising methods. In deep learning models, databases are used for complex mappings from noise to clean speech directly. In review paper on speech enhancement [13], reviewers have the opinion that convolutional neural networks (CNN) are better for speech enhancement as CNN is more effective in learning temporal information of speech signal. Many deep neural network models such as fully connected neural networks [14], deep denoising autoencoder, CNN, LSTM have been effectively used for speech denoising even in diverse noisy conditions [15], [16]. Deep learning methods in different applications and experimental set ups have been implemented [17], [18] for normal speech enhancement and speech enhancement in cocktail parties. LMS algorithm in wavelet domain [19] is used to minimize noise from signals. Both traditional methods and deep learning approaches [20] are compared for speech enhancement and deep learning approaches have proved better than traditional methods. Deep learning approaches have also proved better for speech enhancement [21], improvement in intelligibility of speech [22] and in combination with discrete transforms [23]. Cascaded CNN are used for speech emotion recognition in noisy conditions [24]. Two stage speech enhancement by using optimum values of magnitude and phase have been very effective in noise minimization or noise reduction [25]. An improved RLS algorithm is used to denoise electrocardiogram (ECG) signals for four types of real noises from MIT-BIT dataset [26]. Here use of a systolic architecture enabled faster processing and better noise reduction as compared to RLS. The proposed algorithm [26] outperformed the conventional RLS algorithm with and without systolic architecture in terms of convergence speed, signal-to-noise ratio (SNR), and mean squared error (MSE).

2. PROPOSED METHOD

The schematic of LMS adaptive filter-based speech enhancement system is given in Figure 1. The corrupted speech $x(n)$ corrupted by noise $N(n)$ and reference noise $p(n)$ is given as input to system. The $N(n)$ and $p(n)$ should be correlated. In LMS algorithm, the cost function which is least mean square value of error signal $e(n)$ is minimized having multiple iterations *i.e.*

Since $N(n)$ and $y(n)$ are correlated and signal power is constant, minimizing the cost function will minimize the noise term in (1). In LMS algorithm, the steepest descent algorithm is used for updating the filter. The weight update equation for LMS algorithm is given by (2).

$$\begin{aligned}\text{Cost function} &= min\{E[e^2(n)]\} = min\{E[(x(n)-y(n))^2]\}\\ &= min\{E[(s(n)+N(n)-y(n))^2]\}\\ &= min\{E[(s(n))^2]+E[(N(n)-y(n))^2]+2E[s(n)(N(n)-y(n))]\}\end{aligned} \tag{1}$$

$$w(n+1) = w(n) + \mu e(n)x(n) \tag{2}$$

The step size μ. play an important role in convergence of LMS algorithm. Convergence time will be high if step size is small or vice versa [6].

2.1. Wavelet domain adaptive filter based on LMS algorithm (WDAF-LMS)

In this algorithm, the input signal $x(n)$ is applied to N^{th} order wavelet domain adaptive LMS filter $i.e.$

$$x_N = [x(n), x(n-1), x(n-2), \ldots\ldots\ldots x(n+N-1) \tag{3}$$

Then this input signal is transformed to wavelet domain signal $X(n, f)$ as given in (4).

$$X(n, f) = WT\{x_N\} \tag{4}$$

where $X(n, f)$ is level 3 approximation coefficient of DWT of $x(n)$. The three level wavelet transform decomposition of $x(n)$ is shown in Figure 2 [27]. This wavelet domain signal $X(n, f)$ is used in WDAF system shown in Figure 3. The output $y(n)$ is given by (5).

$$y(n) = \sum_{n=0}^{N-1} X(n, f)\, w_n \tag{5}$$

This output is compared with the desired signal $d(n)$ and error signal $e(n)$ as given by (6).

$$e(n) = d(n) - y(n) \tag{6}$$

The weight of WDAF is changed according to error signal $e(n)$ through multiple iterations. The weight update equation is given by (2).

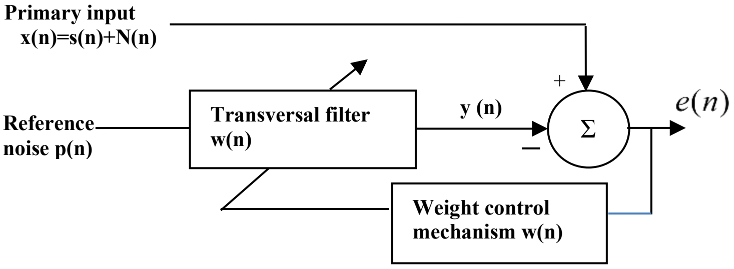


Figure 1. Block diagram of adaptive filter-based speech enhancement system

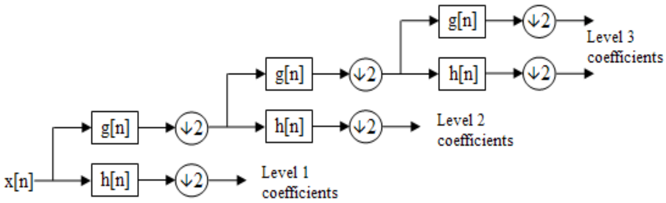


Figure 2. The three level wavelet transform decomposition of $x(n)$ [27]

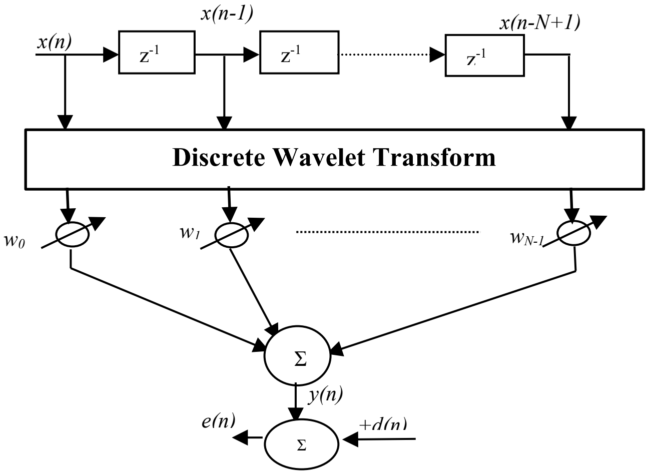


Figure 3. A wavelet domain adaptive filter [19]

2.2. A speech enhancement system based on fully connected deep neural network using stochastic gradient descent with momentum optimizer

A fully connected deep neural network (FCDNN) is a subset of artificial neural network where neurons of one layer are connected to every neuron of the next layer. These networks are widely used in various tasks like regression, classification, and pattern recognition as they can learn complex mappings from inputs and outputs. The choice of activation function is very important in determining the information to be transmitted from one layer to the next. Thus, it controls information exchange and learning from data for a neural network model. In this paper, the output of fully connected layer is passed through a rectified linear unit (ReLU) activation function defined by (7). A regression layer is used in neural networks for regression tasks. It computes the half-mean-squared-error loss for regression tasks. For speech enhancement problems, a regression layer must follow the final fully connected layer. DNN model is trained using stochastic gradient descent with momentum (SGDM) optimizer, progressively improving its prediction over multiple epochs.

$$f(x) = \max(0, x) \tag{7}$$

As shown in Figure 4, fully connected deep neural network using stochastic gradient descent with momentum optimizer (FCDNN-SGDM) is first trained from magnitude spectrum of noisy speech and clean speech to develop a nonlinear relation between the noisy spectrum and clean spectrum. The short time Fourier transform (STFT) is applied to both speeches to obtain its magnitude spectrum and phase spectrum. The magnitude spectrum of noisy speech is used as final predictor. Whereas the magnitude spectrum of clean speech signal is used as final target. For the whole speech database predictors and targets are calculated and used to train DNN models. In the testing stage performance is tested using the noisy test database. In validation stage, the unknown noisy speech signal is applied to trained FCCDNN model. First magnitude spectrum and phase spectrum are separated, and noisy speech signal is calculated by STFT. Only Its magnitude spectrum is given to trained DNN model. The trained FCCDNN model provides the magnitude spectrum of recovered speech signal. Then inverse STFT is applied to this magnitude spectrum and the phase spectrum of noisy speech signal to recover speech signal as shown in Figure 5.

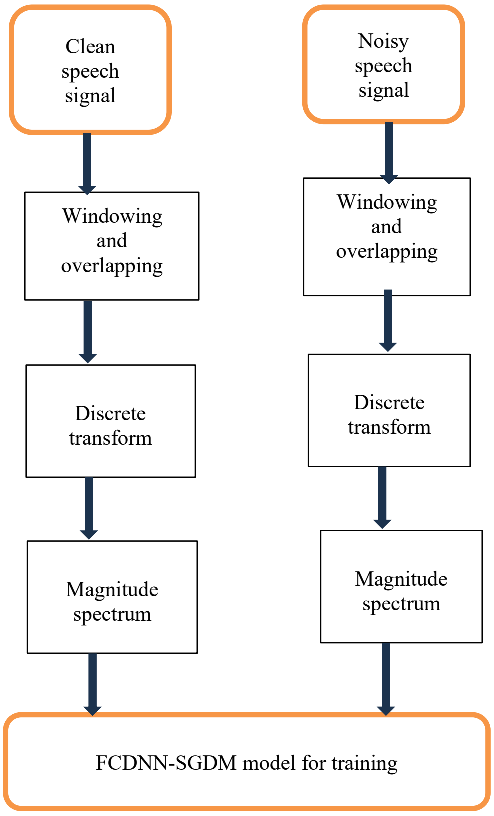


Figure 4. Training stage of FCDNN-SGDM model

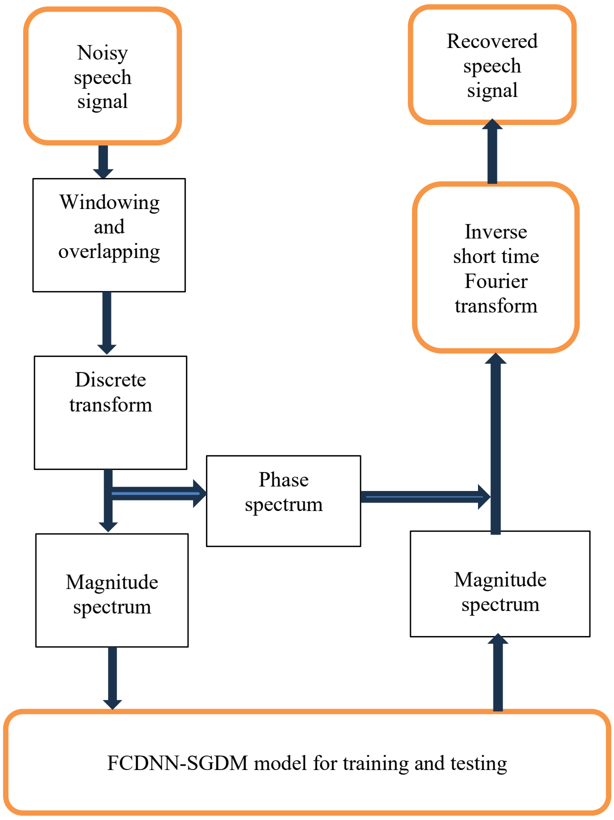


Figure 5. Testing and validation stage of trained FCDNN-SGDM model

3. EXPERIMENTAL SETUP

In the experimental setup shown in Figure 6, a two-stage cascaded speech enhancement method is proposed where first stage is WDAF-LMS and second stage is FCDNN-SGDM. In the second stage, the recovered signal is again passed through trained FCDNN-SGDM based speech enhancement system to further remove the residual background noise. For evaluation of a speech enhancement system, a proper database is required. Here clean sentence database is taken from Hindi speech database [2]. The noisy version of clean sentences is prepared by adding babble noise, car noise and machine gun noises from NOISEX-92 database [1] to this clean sentence at 0 dB down to –13 dB.

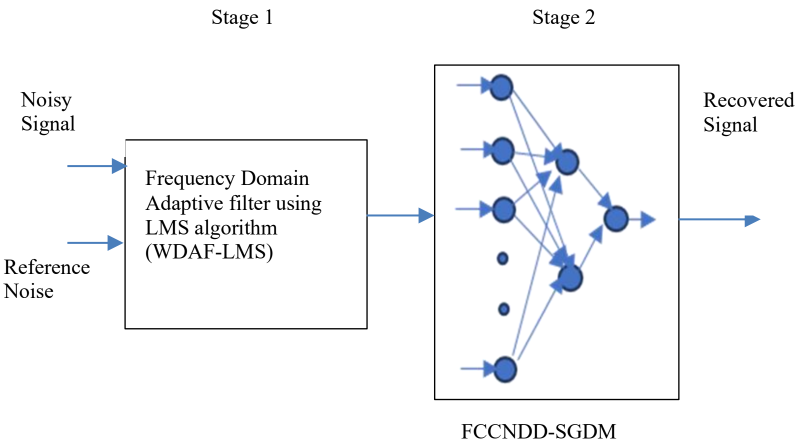


Figure 6. A two-stage speech enhancement system

In WDAF-LMS based speech enhancement system, the filter order is taken as 20 and step size μ is taken as 0.008. Here 'DB2' wavelet is used for DWT for reference noise signal taken as input to transversal filter For FCCDNN-SGDM, total 419 speech samples are taken for training, testing and validation. Speech database [2] of 21 speakers is taken for training, testing and validation. Where 20 sentences are taken from each speaker. Training dataset consists of total 314 Hindi speech samples and testing dataset consists of 84 Hindi speech samples. 75% and 20% of the database are used for training and testing respectively. Also, 21 samples *i.e.* 5% samples are used for validation of the results. For STFT, "Hanning window" of length 64 is used for framing. Five layers are used for training the model. SGDM is used for Training options with momentum equals to 0.9000, learn rate drop factor is .0001, maximum epochs are 20, mini batch size is 64.

The performance parameters used in experimental setup for evaluation of system are SNR improvement, perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI). PESQ value ranges from-0.5 to 4.5 and the higher PESQ means better perceptual quality. STOI value is a scalar quantity, and ranges from-1 to 1. It measures the intelligibility of the recovered speech signal by comparing it with the clean speech signal. A higher value for STOI corresponds to a higher intelligibility of speech signal.

4. RESULTS AND DISCUSSION

It is observed from Table 1 to Table 3 that the proposed system provides improvement in speech quality and intelligibly in terms of SNR improvement, PESQ and STOI at all input SNR level for all three noises. In this experimental setup, the corrupted signal was tested at input SNR levels ranging from 0 dB down to –13 dB. Even up to -13 dB input SNR level, the system provides subsequent improvements.

Table 1, Figure 7, and Figure 8 show the enhancement in noisy speech signal corrupted by babble noise at 0 dB, -5 dB, -6 dB, -7 dB, -8 dB, -9 dB, -10 dB, -11 dB, -12 dB, -13 dB input SNR levels in terms of SNR, PESQ and STOI, when it is passed through the system, with the specific metrics detailed in Figure 7(a) through Figure 7(d). Table 1 shows the comparisons of output at both stages. Where stage-1 is WDAF-LMS and stage-2 is FCCDNN-SGDM. The WDAF-LMS achieved a maximum SNR improvement of 15.238 dB at –10 dB input SNR while the subsequent FCCDNN-SGDM stage further enhanced the performance, achieving the overall maximum SNR improvement of 17.061 dB at –8 dB input SNR. Even at -13 dB input SNR level, the proposed algorithm provides subsequent improvement with improvements in

PESQ and STOI values which shows that at -13 dB input SNR level, speech quality and intelligibility is retained in denoised signal. Figure 8 illustrates the waveform comparison specifically at -5 dB input SNR, displaying the clean speech in Figure 8(a), the noisy speech in Figure 8(b), the denoised signal after stage-1 in Figure 8(c), and the denoised signal after stage-2 in Figure 8(d). As observed in these waveforms, the signal recovered after stage-2 is closer to original signal as compared to the signal recovered after stage-1.

Table 2, Figure 9, and Figure 10 show the enhancement in noisy speech signal corrupted by car noise at 0 dB, -5 dB, -6 dB, -7 dB, -8 dB, -9 dB, -10 dB, -11 dB, -12 dB, -13 dB input SNR levels in terms of SNR, PESQ and STOI, when it is passed through the system, with the specific metrics detailed in Figure 9(a) through Figure 9(d). Table 2 shows the comparisons of output at both stages. The WDAF-LMS shows maximum SNR improvement of 21.2406 dB at -13 dB input SNR level while the FCCDNN-SGDM further improves in SNR to 22.1705 dB at -13 dB input SNR level. It is observed that at -13 dB input SNR level, the proposed algorithm provides subsequent improvement with improvements in PESQ and STOI values. Which shows that at -13 dB input SNR level speech quality and intelligibility is retained in denoised signal. Figure 10 illustrates the waveform comparison specifically at -5 dB input SNR, displaying the clean speech in Figure 10(a), the noisy speech in Figure 10(b), the denoised signal after stage-1 in Figure 10(c), and the denoised signal after stage-2 in Figure 10(d). As observed in these waveforms, the signal recovered after stage-2 is closer to original signal as compared to the signal recovered after stage-1.

Table 1. Performance comparison of speech enhancement system at stage-1 and stage-2 (final stage) for babble noise

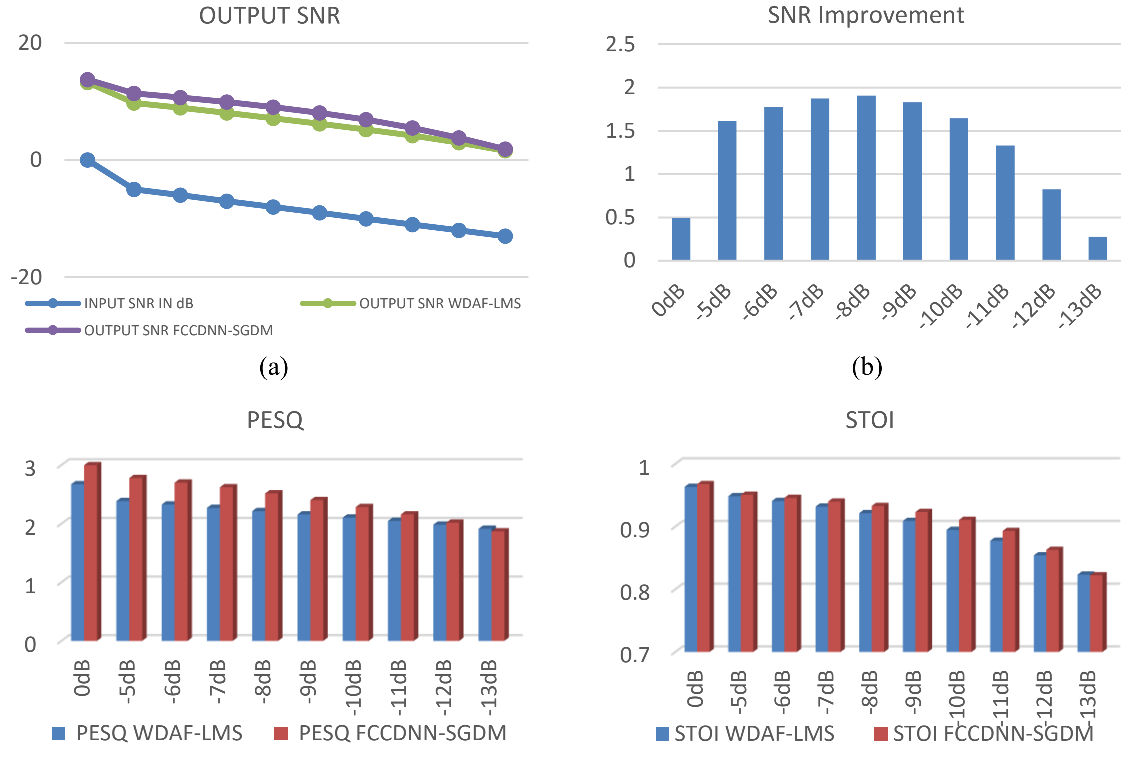
Input SNR	SNR improvement		PESQ		STOI	
	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM
0 dB	13.2457	13.7421	2.6716	2.9963	0.9634	0.9678
-5 dB	14.7478	16.3641	2.3847	2.7784	0.9486	0.9508
-6 dB	14.9184	16.6935	2.3271	2.7002	0.9409	0.9458
-7 dB	15.0552	16.9283	2.2691	2.62	0.932	0.9399
-8 dB	15.1552	17.061	2.2136	2.5161	0.9214	0.9329
-9 dB	15.2201	17.0488	2.1575	2.4017	0.909	0.9233
-10 dB	15.238	16.8831	2.1032	2.2838	0.8947	0.9108
-11 dB	15.1705	16.5017	2.0475	2.1573	0.8774	0.893
-12 dB	14.9654	15.7885	1.9823	2.0169	0.8542	0.863
-13 dB	14.5853	14.8621	1.9124	1.868	0.8234	0.8223

Table 2. Performance comparison of speech enhancement system at stage-1 and stage-2 (final stage) for car noise

INPUT SNR	SNR improvement		PESQ		STOI	
	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM
0 dB	9.0590	10.1693	3.4558	3.4654	0.9844	0.9876
-5 dB	13.9972	15.0926	3.4075	3.418	0.984	0.9875
-6 dB	14.9741	16.0659	3.3906	3.4109	0.9838	0.9875
-7 dB	15.9454	17.0304	3.3716	3.4023	0.9837	0.9874
-8 dB	16.9095	17.9966	3.348	3.3891	0.9835	0.9873
-9 dB	17.8641	18.9297	3.3221	3.371	0.9832	0.9872
-10 dB	18.8023	19.8659	3.2817	3.3463	0.9828	0.987
-11 dB	19.7088	20.7625	3.2261	3.3069	0.9822	0.9867
-12 dB	20.5409	21.5627	3.1212	3.2098	0.9807	0.9856
-13 dB	21.2406	22.1705	2.9694	3.0219	0.9774	0.9824

Table 3. Performance comparison of speech enhancement system at stage-1 and stage-2 (final stage) for machine gun noise

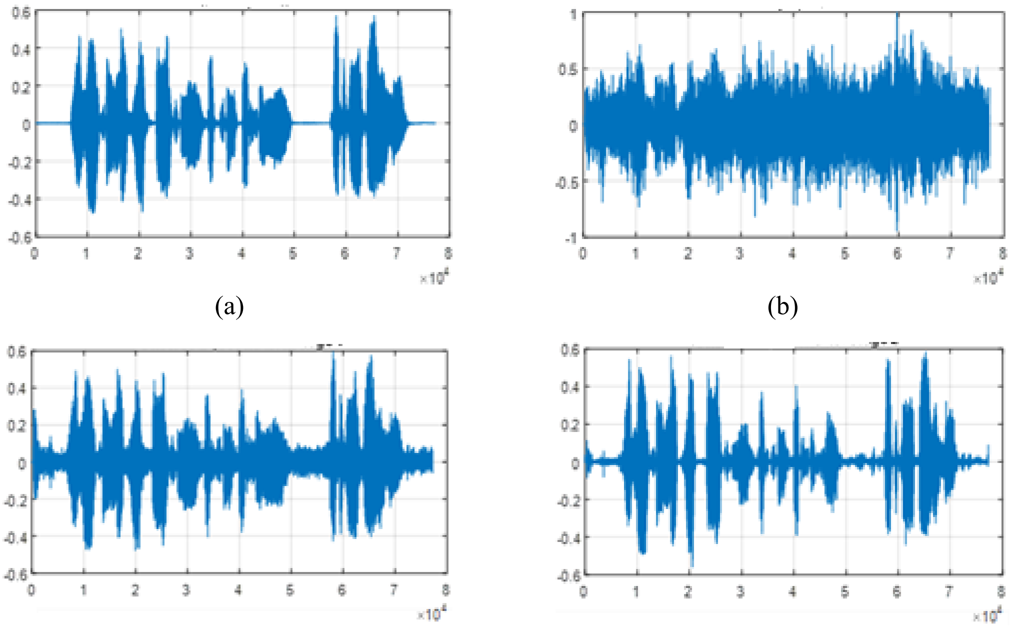
INPUT SNR	SNR improvement		PESQ		STOI	
	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM	WDAF-LMS	FCCDNN-SGDM
0 dB	11.1084	17.1918	3.1031	3.1687	0.9879	0.9884
-5 dB	11.1881	18.2533	2.692	2.7393	0.9649	0.9727
-6 dB	11.1312	17.9718	2.5228	2.6046	0.9533	0.9653
-7 dB	10.9965	17.7721	2.3785	2.5031	0.9397	0.9579
-8 dB	10.8086	17.5841	2.237	2.2968	0.9224	0.9494
-9 dB	10.5918	17.4199	2.128	2.3085	0.9022	0.9395
-10 dB	10.3505	17.3517	2.0198	2.2277	0.8775	0.9277
-11 dB	10.1153	17.3198	1.9421	2.1502	0.8517	0.915
-12 dB	9.9283	17.3148	1.8716	2.0815	0.8265	0.902
-13 dB	9.7934	17.2867	1.8241	2.0019	0.8028	0.8873



(a) (b)

(c) (d)

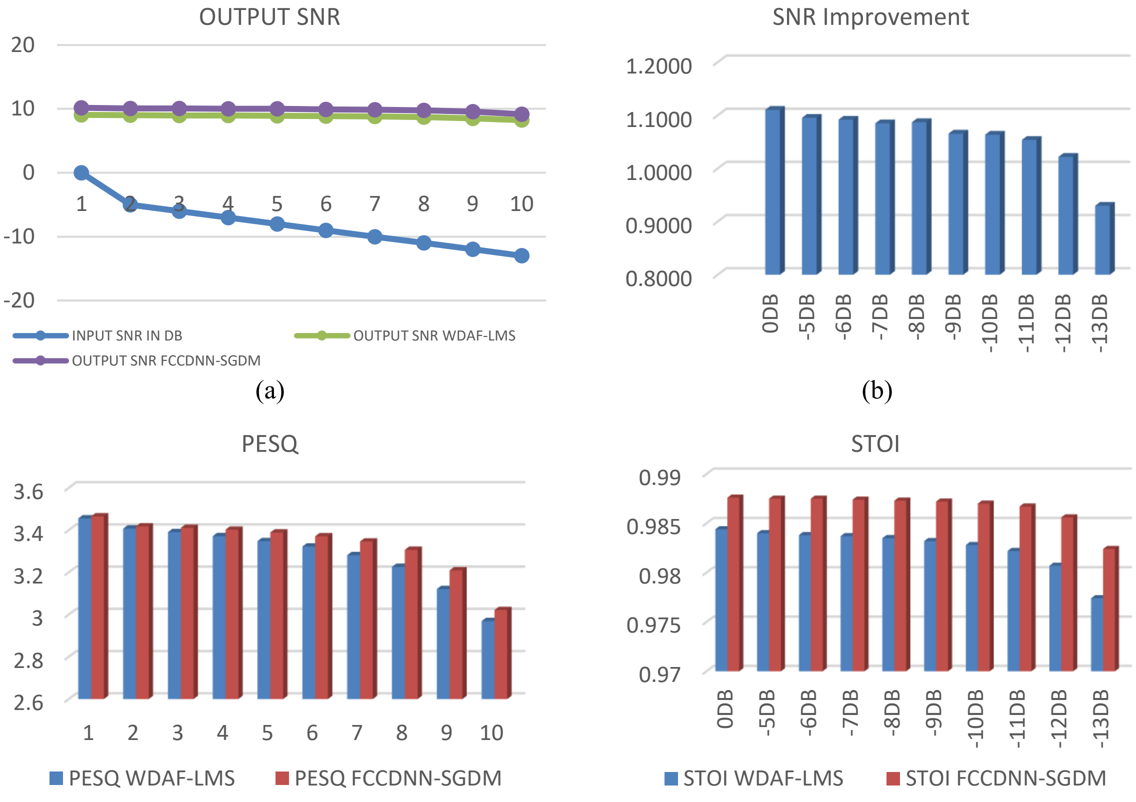
Figure 7. Performance comparison of speech enhancement system for babble noise at different input SNR levels: (a) output SNR for babble noise at all input SNR level, (b) SNR improvement for FCCDNN-SGDM as compared to WDAF-LMS, (c) PESQ improvement for FCCDNN-SGDM as compared to WDAF-LMS, and (d) STOI improvement for FCCDNN-SGDM as compared to WDAF-LMS for babble noise at all Input SNR level



(a) (b)

(c) (d)

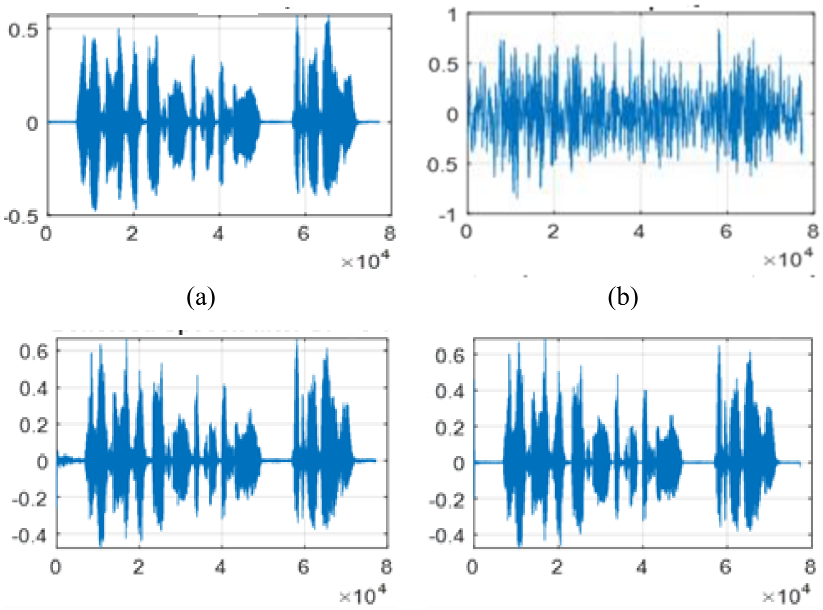
Figure 8. Waveform comparison of speech signals corrupted by babble noise: (a) clean speech, (b) noisy speech, (c) denoised speech signal after stage-1, and (d) denoised speech signal after stage-2 at -5 dB input snr for babble noise



(a) (b)

(c) (d)

Figure 9. Performance comparison of speech enhancement system for car noise at different input SNR levels: (a) output SNR for car noise at all input SNR level, (b) SNR improvement for FCCDNN-SGDM as compared to WDAF-LMS, (c) PESQ improvement for FCCDNN-SGDM as compared to WDAF-LMS for, and (d) STOI improvement for FCCDNN-SGDM as compared to WDAF-LMS for car noise at all input SNR level



(a) (b)

(c) (d)

Figure 10. Waveform comparison of speech signals corrupted by car noise: (a) clean speech, (b) noisy speech, (c) denoised speech signal after stage-1, (d) and denoised speech signal after stage-2 at -5 dB input SNR for car noise

Table 3 and Figure 11 show the enhancement in noisy speech signal corrupted by machine gun noise at 0 dB, -5 dB, -6 dB, -7 dB, -8 dB, -9 dB, -10 dB, -11 dB, -12 dB, -13 dB input SNR Levels in terms of SNR, PESQ and STOI, when it is passed through the system, with the specific metrics detailed in Figure 11(a) through Figure 11(d). Table 3 shows the comparisons of output at both stages. The WDAF-LMS shows maximum SNR improvement of 11.1881 dB at -5 dB input SNR level while FCCDNN-SGDM shows further SNR improvement of 18.2533 dB at -5 dB input SNR level. Even at -13 dB input SNR level, the proposed algorithm provides subsequent improvement with improvements in PESQ and STOI values. Which shows that at -13 dB input SNR level speech quality and intelligibility is retained in denoised signal.

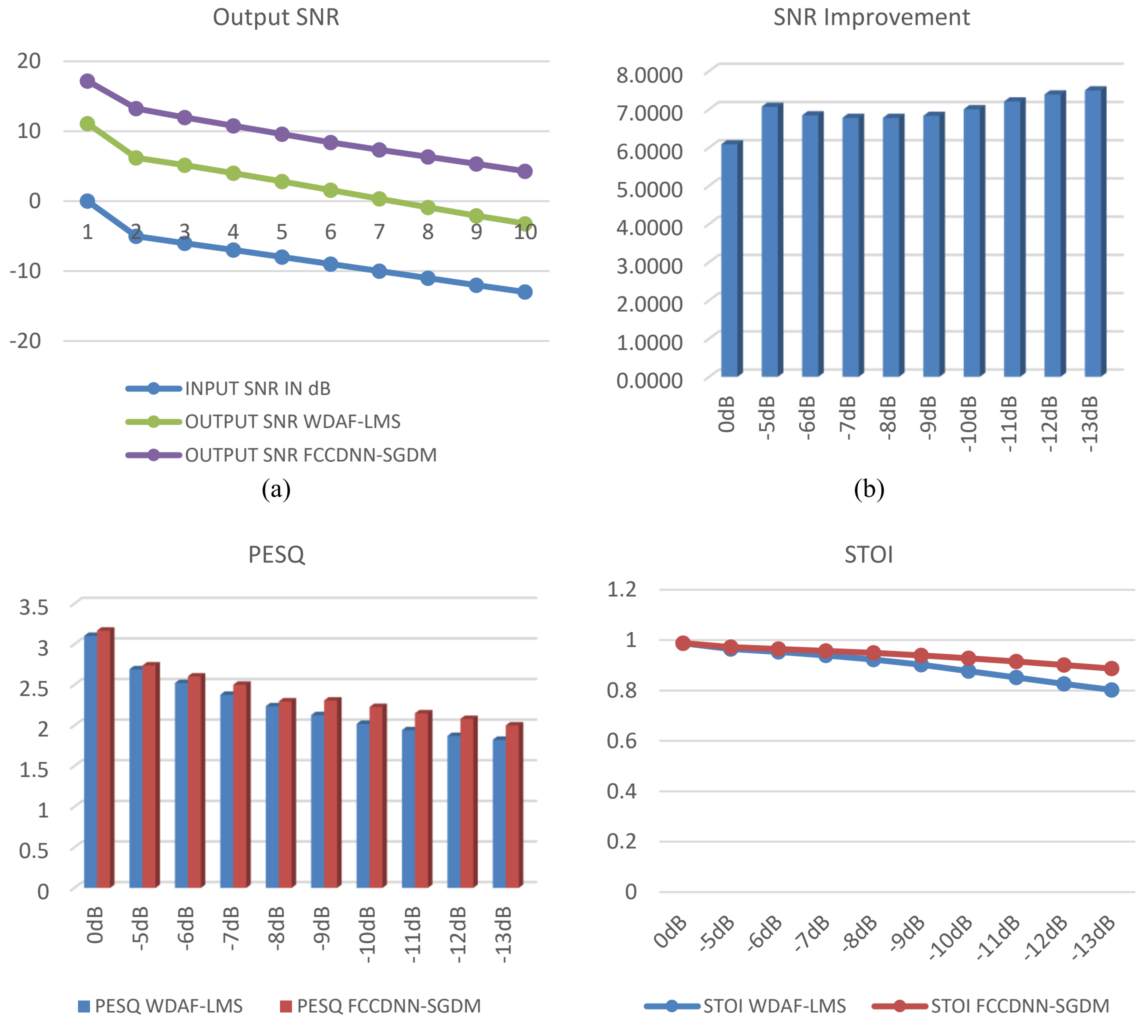
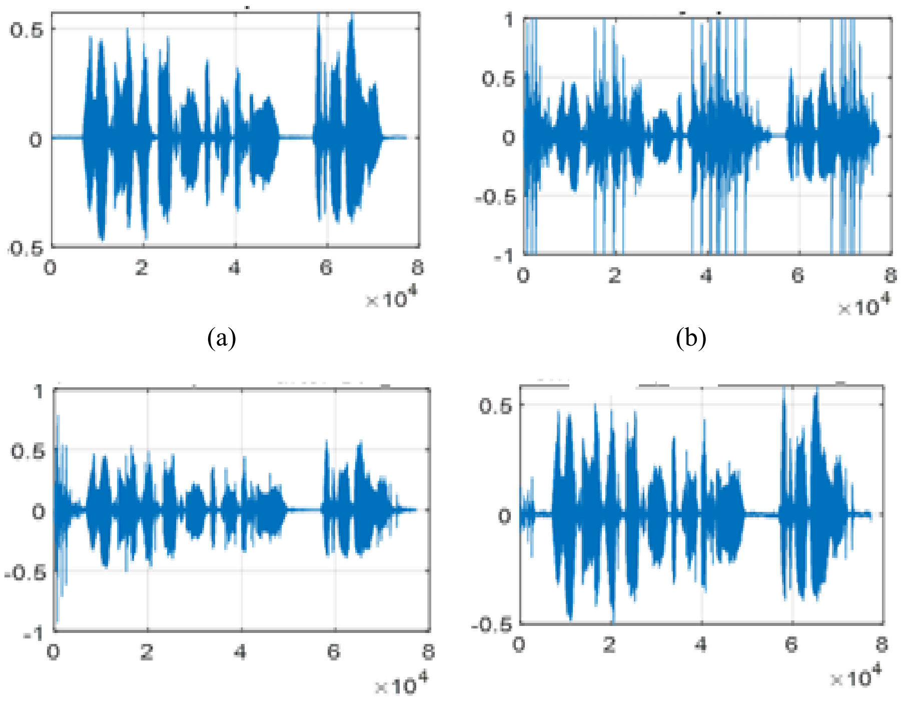


Figure 11. Performance comparison of speech enhancement system for machine gun noise at different input SNR levels: (a) Output SNR for machine gun noise at all input SNR level, (b) SNR improvement for FCCDNN-SGDM as compared to WDAF-LMS, (c) PESQ improvement for FCCDNN-SGDM as compared to WDAF-LMS, and (d) STOI improvement for FCCDNN-SGDM as compared to WDAF-LMS for machine gun noise at all input SNR level

Figure 12 shows denoised output signal of stage-1 and stage-2 at -5 dB input SNR level and the quality of signals received after stage-2 is better than that of stage-1. Displaying the clean speech in Figure 12(a), the noisy speech in Figure 12(b), the denoised signal after stage-1 in Figure 12(c), and the denoised signal after stage-2 in Figure 12(d).



(a) (b)

(c) (d)

Figure 12. Waveform comparison of speech signals corrupted by machine gun noise at -5 dB input SNR: (a) clean speech, (b) noisy speech, (c) denoised speech signal after stage-1, and (d) denoised speech signal after stage- 2 at -5 dB input SNR for machine gun noise

5. CONCLUSION

Here, a two-stage cascaded speech enhancement method is proposed for noise cancellation from highly corrupted speech signals. Where stage-1 is WDAF-LMS based system and stage-2 is FCCDNN-SGDM based system. The result is presented for speech corrupted by babble noise, car noise and machine gun noises at 0 dB and -5 dB to -13 dB input SNR levels and the results are compared at both stages *i.e.* stage-1 and stage-2. Speech quality and intelligibly is compared in terms of SNR improvement, PESQ and STOI at all input SNR levels for all three noises. It is proved that the proposed system shows improvement in SNR, PESQ, and STOI up to -13 dB input SNR level corrupted speech signal.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

First author one has contributed to conceptualization, methodology and formal analysis while second author has prepared the original draft of the paper. Third author is involved in review and editing of the paper.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kavitha A	✓	✓			✓									
Mahesh Chandra									✓					
Vijay Kumar Gupta										✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] A. Varga, “The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” *ical Report, DRA Speech Research Unit*, 1992.

[2] K. Samudravijaya, P. V. S. Rao, and S. S. Agrawal, “Hindi speech database,” in *Proceedings 6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. 456–459. doi: 10.21437/ICSLP.2000-847.

[3] S. S. Haykin, *Adaptive filter theory*, 4th ed. N.J: Prentice Hall, 2002.

[4] J. E. Greenberg, “Modified LMS algorithms for speech processing with an adaptive noise canceller,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 338–351, Jul. 1998, doi: 10.1109/89.701363.

[5] S. Zhao, Z. Man, S. Khoo, and H. Ren Wu, “Stability and convergence analysis of transform-domain LMS adaptive filters with second-order autoregressive process,” *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 119–130, Jan. 2009, doi: 10.1109/TSP.2008.2007618.

[6] S. Attallah, “The wavelet transform-domain LMS adaptive filter with partial subband-coefficient updating,” *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 1, pp. 8–12, Jan. 2006, doi: 10.1109/TCSII.2005.855042.

[7] M. S. Esfand Abadi, H. Mesgarani, and S. M. Khademiyan, “The variable step-size wavelet transform-domain LMS adaptive filter algorithm,” *Scientia Iranica*, vol. 27, no. 3, pp. 1398–1412, Aug. 2018, doi: 10.24200/sci.2018.20827.

[8] F. Beaufays, “Transform-domain adaptive filters: an analytical approach,” *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 422–431, 1995, doi: 10.1109/78.348125.

[9] T. Lutfor, M. Alam, and S. Sarker, “Analysis of wavelet transform-domain LMS-Newton adaptive filtering algorithms with second-order autoregressive process,” *IEEE Trans. Signal Process [J]*, vol. 3, no. 3, pp. 54–70, 2013, doi: 10.5923/j.ajsp.20130303.03.

[10] M. Shams Esfand Abadi, H. Mesgarani, and S. M. Khademiyan, “The wavelet transform-domain LMS adaptive filter employing dynamic selection of subband-coefficients,” *Digital Signal Processing*, vol. 69, pp. 94–105, Oct. 2017, doi: 10.1016/j.dsp.2017.05.012.

[11] E. Ozen and N. Ozkurt, “Speech noise reduction with wavelet transform domain adaptive filters,” in *2021 Global Congress on Electrical Engineering (GC-ElecEng)*, Dec. 2021, pp. 15–20. doi: 10.1109/GC-ElecEng52322.2021.9788190.

[12] J. Raghuwanshi, A. Mishra, and N. Singh, “The wavelet transform-domain adaptive filter for nonlinear acoustic echo cancellation,” *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 25853–25871, Sep. 2020, doi: 10.1007/s11042-020-09218-5.

[13] A. R. Yuliani, M. F. Amri, E. Suryawati, A. Ramdan, and H. F. Pardede, “Speech enhancement using deep learning methods: a review,” *Jurnal Elektronika dan Telekomunikasi*, vol. 21, no. 1, p. 19, Aug. 2021, doi: 10.14203/jet.v21.19-26.

[14] S. Haykin, *Neural networks and learning machines*. London: Pearson Education India, 2009.

[15] F. E. Wahab, Z. Ye, N. Saleem, and R. Ullah, “Compact deep neural networks for real-time speech enhancement on resource-limited devices,” *Speech Communication*, vol. 156, p. 103008, Jan. 2024, doi: 10.1016/j.specom.2023.103008.

[16] S. Kantamaneni, A. Charles, and T. R. Babu, “Speech enhancement with noise estimation and filtration using deep learning models,” *Theoretical Computer Science*, vol. 941, pp. 14–28, Jan. 2023, doi: 10.1016/j.tcs.2022.08.017.

[17] L. Yue and Q. Ji, “Speech enhancement based on the combination of deep learning and wavelet algorithm,” in *International Symposium on Automatic Control and Emerging Technologies*, 2024, pp. 178–188. doi: 10.1007/978-981-97-0126-1_16.

[18] T. Fischer, M. Caversaccio, and W. Wimmer, “Speech signal enhancement in cocktail party scenarios by deep learning based virtual sensing of head-mounted microphones,” *Hearing Research*, vol. 408, p. 108294, Sep. 2021, doi: 10.1016/j.heares.2021.108294.

[19] P. Goel, S. Rai, M. Chandra, and V. K. Gupta, “Analysis of LMS algorithm in wavelet domain,” in *Conference on Advances in Communication and Control Systems (CAC2S 2013)*, 2013, pp. 734–738.

[20] Y. Wang, “Research progress in speech enhancement technology,” in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Jul. 2020, pp. 222–226. doi: 10.1109/CVIDL51233.2020.00-97.

[21] J. Llombart, D. Ribas, A. Miguel, L. Vicente, A. Ortega, and E. Lleida, “Progressive loss functions for speech enhancement with deep neural networks,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 1, Dec. 2021, doi: 10.1186/s13636-020-00191-3.

[22] N. Y.-H. Wang *et al.*, “Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 184–195, 2021, doi: 10.1109/TNSRE.2020.3042655.

[23] W. A. Jassim and N. Harte, “Comparison of discrete transforms for deep-neural-networks-based speech enhancement,” *IET Signal Processing*, vol. 16, no. 4, pp. 438–448, Jun. 2022, doi: 10.1049/sil2.12109.

[24] Y. Nam and C. Lee, “Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions,” *Sensors*, vol. 21, no. 13, p. 4399, Jun. 2021, doi: 10.3390/s21134399.

[25] C. Qing, L. He, X. Fu, and H. Lin, “Two-stage cascaded speech enhancement by exploiting magnitude and phase optimization,” *Circuits, Systems, and Signal Processing*, vol. 44, no. 9, pp. 7048–7069, Sep. 2025, doi: 10.1007/s00034-025-03154-1.

[26] H. H. Thannoon and I. A. Hashim, “Efficient enhanced recursive least square algorithm adaptive filtering scheme for artifacts removal in ECG signals,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 6, p. 100318, Dec. 2023, doi: 10.1016/j.prime.2023.100318.

[27] K. P. Soman, *Insight into wavelets: from theory to practice*. Delhi: PHI Learning Pvt. Ltd., 2010.

BIOGRAPHIES OF AUTHORS

Kavitha A is currently pursuing Ph.D. from REVA University, Bengaluru, India. She is working as assistant professor at East-West College of Engineering Bengaluru, India. She can be contacted at kavithaln217@gmail.com.

Mahesh Chandra holds a Ph.D. in audio signal processing from AMU, Aligarh. He is F-IETE, SM-IEEE, M-IEI and LM-ISTE. Dr. Mahesh Chandra, presently working as professor at School of Electronics and Communication Engineering REVA University, Bengaluru has also served as professor and dean students welfare at Birla Institute of Technology, Mesra Ranchi. His areas of interest are audio signal processing, human computer interaction, machine learning and internet of things. He has delivered several keynotes talks and chaired many sessions at National/International levels. He can be contacted at email: shrotriya69@gmail.com.

Vijay Kumar Gupta holds a Ph.D. in adaptive signal processing from BIT Mesra Ranchi. He is M-IEEE, and LM-ISTE. He is presently working as associate professor at department of electronics and communication engineering, Government Engineering College West Champaran Bihar and director in-charge at the same college. His areas of interest are adaptive signal processing, machine learning and deep learning. He can be contacted at guptavk76@gmail.com.