Multimedia Tools and Applications (2023) 82:3647–3675
https://doi.org/10.1007/s11042-022-13302-3

Speech enhancement using long short term memory with trained speech features and adaptive wiener filter

Anil Garg [1]

Received: 23 June 2021 / Revised: 28 January 2022 / Accepted: 30 May 2022 /
Published online: 14 July 2022

Abstract
Speech enhancement is the process of enhancing the clarity and intelligibility of speech signals that have been degraded due to background noise. With the assistance of deep learning, a novel speech signal enhancement model is introduced in this research. The proposed model is divided into two phases: (i) Training (ii) Testing. In the training phase, the noise spectrum and signal spectrum are estimated via a Non-negative Matrix Factorization (NMF) from the noisy input signal. Then, Empirical Mean Decomposition (EMD) features are extracted from the Wiener filter. The de-noised signal is acquired from EMD, the bark frequency is evaluated and the Fractional Delta AMS features are extracted. The key contribution of this study is the use of the Long Short Term Memory (LSTM) model to properly estimate the tuning factor η of the Wiener filter for all input signals. The LSTM was trained by the extracted features (EMD) via a modified wiener filter for decomposing the spectral signal and the output of EMD is the denoised enhanced speech signal. A comparative evaluation is carried out between the proposed and existing models in terms of error measures.

Keywords Speech processing · Speech enhancement · Empirical mean decomposition · Empirical mean curve decomposition · Wiener filter

1 Introduction

Speech is typically distorted in real-world environments by both room resonances and background noises [10]. The goal of speech enhancement is to remove a specific amount of noise from a noisy speech signal while retaining the speech component and reducing speech distortion as much as possible [29]. Speech augmentation is required in a variety of applications, including mobile communication and speech recognition [46]. Speech is one of the most

✉ Anil Garg
anilgarg0778@gmail.com

[1] ECE Department, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar Deemed To Be University, Mullana, Ambala, Haryana 134007, India

common ways for humans to share information [11]. Speech is the most critical mode for interaction in today's technological society. Speech is indeed a tool that allows us to communicate with each other [34]. In recent days, Covid-19 is also detected from the speech signal [26]. To find out presence or absence of Covid-19 the speech signal is used [5]. The goal of single channel speech enhancement is to improve the quality and intelligibility of speech that has been corrupted by environmental noises, which degrades many real-world applications such as speech recognition, hearing aids, and speech telephony [32]. Speech enhancement is a common technique for improving speech quality [8]. These disruptions diminish speech quality and intelligibility, particularly whenever the Signal-to-Noise Ratio (SNR) is low. Binaural speech enhancement strategies are of particular interest for assistive listening devices, such as hearing aids or headsets, where the end user expects both high speech quality and speech clarity [38]. The noise corrupted signal is improved by using spatial or temporal modifications [20]. The speech signal is regarded as the quickest and most natural way to communicate with humans [12].Understanding distorted speech can be complicated for both Normal Hearing (NH) and Hearing Impaired (HI) listeners. Many voice-related applications, such as Automatic Speech Recognition (ASR) and Speaker Identification (SID) seem to perform poorly in presence of noise [39]. Therefore, speech enhancement is essential.

There's been a lot of study towards improving speech in noisy environments [2]. Speech recognition in background noise appears to be difficult for people with hearing loss [3]. The goal of speech enhancement algorithms is to remove additive background noise from a noisy speech signal in order to improve its quality or intelligibility [9]. In general, speech enhancement refers to the processing of noisy speech signals in order to improve signal perception via better decoding by systems or humans [14]. Noises such as airport noise, train noise, and street noise frequently distort speech signals. These noises have a negative impact on the quality of the speech signal, especially in voice communication, automatic speech recognition, and speaker identification [15]. Feature selection is an important step in improving a system for recognising emotions in speech [2]. Speech recognition in background noise appears to be difficult for people with hearing loss [3]. Background noise is the primary source of speech degradation, particularly in hands-free scenarios [22]. The field of speech enhancement (SE) is concerned with the enhancement of speech signals that have been degraded by noise [16]. Speech enhancement in non-stationary noise environments is a difficult area of study [13]. Speech enhancement aims to improve the clarity and intelligibility of noisy speech [28]. The major intention behind the speech enhancement is to suppress the noise and to boost up the SNR of noisy speech signals in challenging environments. The most renowned techniques like spectral subtraction, Minimum Mean Square Error (MMSE), Log MMSE, OM-LSA, Wiener filtering, etc. are being more commonly preferred for Speech Enhancement [19]. Speech enhancement (SE) is the problem of estimating clean-speech signals from noisy single-channel or multiple-channel audio recordings [27]. Speech enhancement techniques have been studied for several decades with a variety of promising applications, such as telecommunications and hearing aid systems, to mitigate the harmful effects of background noise and interference [33]. Acoustically added background noise to speech can degrade the performance of digital voice processors used for applications such as speech compression, recognition, transcription and authentication [7]. The key aim of these speech enhancement methods is to enhance the Speech SNR. Techniques have been introduced regarding boosting up the speech quality and compacting speech bandwidth by suppressing the additive background noise [6]. Deep Neural Networks (DNN) are mostly deployed in speech enhancement [41]. This method generally produces a measure of time-frequency mask that was employed to evaluate the clean speech spectrum [21]. Optimal mask generation was also introduced in the traditional method [44]. But this masking strategy generally provides residual and musical noise in the enhanced speech. Kalman Filtering (KF) based speech enhancement is introduced in [17]; here the Linear Prediction Coefficients (LPCs) are calculated using a DNN. Although under nonstationary noise settings, the noise covariance is computed during speech gaps, which is ineffective. In addition, a deep audio-visual speech enhancement is suggested [35], but this approach might break in low SNR values.

Recently, Generative Adversarial Network (GAN) based speech enhancement was utilized to overcome the traditional difficulties. Especially, Speech Enhancement GAN (SEGAN) [30], conditional GAN (cGAN) [23], Wasserstein GAN (WGAN) [18], and Relativistic Standard GAN (RSGAN) [1] techniques were introduced. Despite the success of GAN-based speech improvement techniques, two major difficulties were present that was training instability and a lack of consideration for varied speech characteristics [42]. Therefore, researchers have been making a significant contribution to this field for decades. However, the accuracy and intelligibility of the outcomes weren't always adequate.

Thus, to overcome the existing issues, an LSTM with trained speech features and an adaptive Wiener filter is introduced in this work. The major contribution of this research is listed below:

- For decomposing the speech spectral signal, a modified wiener filter is introduced.
- In addition, the LSTM model is introduced to properly estimate the tuning factor of the Wiener filter for all input signals.
- In a testing phase, the LSTM model has been trained by the extracted features (EMD) via a modified wiener filter.

The rest of this paper is organized as: Section 2 addresses the literature works undergone in speech enhancement. Section 3 tells about the proposed speech enhancement model: an architectural description. In addition, Section 4 depicts about the processing steps of proposed speech enhancement model. The results acquired with the proposed work are discussed in Section 5. This paper is concluded in Section 6.

1.1 Problem statement

Most studies have shown that reducing signal noise without distorting speech is a difficult challenge, which is one of the main reasons why perfect enhancement systems aren't available. In this research, we focus the issues such as lower robustness, not suitable for complex noise conditions [37], more residual noise existing, lower SNR [45], reduction in speech intelligibility [43], lower denoise effect, lower PESQ [40], and low consideration of speech quality measures [10].

Compared to the existing models, the proposed work introduces a wiener filter-assisted deep learning LSTM model. The LSTM model estimates the tuning factor of the Wiener filter with the aid of extracted features to obtain the de-noised speech signal. For simulation, the proposed model considers the speech quality measures such as SDR, PESQ, SNR, RMSE, CORR, ESTOI, and STOI. Moreover, the proposed model attains higher SNR, PESQ, robustness, and also the proposed model is well suited for complex noisy environments.

2 Literature review

In 2017, Zou et al. [46] introduced two speech amplification frameworks with super gaussian speech modeling. Under the assumption that the Discrete Cosine Transform (DCT) coefficients of clean speech were modelled by Laplacian or a Gamma distribution and the DCT coefficients of the noise were Gaussian distributed, the clean speech components were calculated using the MMSE estimator. Then, underneath the condition of speech presence ambiguity, MMSE estimators were retrieved. The correct estimators of speech statistical parameters were indeed recommended. A modern decision-directed approach has been used to approximate the speech Laplacian element. According to the simulation data, the suggested algorithm generates very little residual disruption and has higher speech efficiency than Gaussian-based speech amplification algorithms.

In 2020, Zhang et al. [37] developed an LSTM-Convolutional-BLSTM Encoder-Decoder (LCLED) for enhancing the speech signal. Transpose convolution and skip connection were both included in the LCLED. Besides that, a priori SNR has been used as a learning objective of LCLED to achieve a higher level of enhanced speech. Post-processing is done using the MMSE method. The findings indicate that the suggested LCLED increases the accuracy and intelligibility of enhanced speech. Furthermore, the running time of the LCLED model was 130 sec.

In 2021, Khattak et al. [29] proposed a "phase compensated perceptually weighted -order Bayesian estimator" to modify both magnitude and phase spectra to improve noisy speech. They have changed the step of noisy speech spectra alone in the proposed methodology. Second, they have manipulated the magnitude-spectra using a perceptually motivated-order Bayesian estimator. Further, to obtain a stronger gain function, the estimator combines the benefits of the perceptually-weighted and -order spectral amplitude estimators. To recreate the interference attenuated speech signals, the compensated phase spectra, and approximate magnitude spectra were merged. Using the NOIZEUS and AURORA repositories, the proposed speech amplification strategy was tested for various noise ranges (0 dB to +10 dB) in terms of quantitative accuracy and intelligibility tests. In both non-stationary and stationary noisy settings, the proposed improvement approach significantly enhances productivity and ensures intelligibility.

In 2020, Tan et al. [45] introduced a Fully Convolutional Neural Network (FCNN) "to achieve end-to-end speech enhancement. The encoder and decoder, as well as an extra Convolutional-Based Short-Time Fourier Transform (CSTFT) layer and CISTFT layer, were applied to simulate forward as well as inverse STFT operations, respectively. Since the fundamental phonetic information of speech is presented more clearly by Time-Frequency (T-F) representations, these layers seek to incorporate frequency-domain knowledge into the proposed model. In addition, the Temporal Convolutional Module (TCM), which would be successful for processing the long-term correlations of speech signals, was indeed integrated amongst encoder and decoder. According to the experimental findings, the suggested paradigm consistently outperforms other competitive speech amplification models.

In 2020, Zhu et al. [43] used the "Deep Neural Network (DNN)-augmented colored-noise Kalman filter" to develop a novel speech enhancement system. The authors have modelled the noise as well as clean speech signal in the form of an Autoregressive (AR) process. The multi-objective DNN was trained via LPCs to map the Line Spectrum Frequencies (LSF) from the noisy acoustic features. The denoising was done to the noisy speech by applying the 'colored-noise Kalman filter with DNN estimated parameters". Finally, residual noise in the Kalman-filtered speech was removed using a post-subtraction procedure. The proposed work has achieved the best estimation accuracy for street noise and produced better outcomes in unseen noise.

In 2021, Wei et al. [40] have proposed a Constant Q Transform (CQT) intending to enhance the resolution of the lower frequency speech signals. The NMF/ Sparse NMF (SNMF) algorithm has been used in the backend. At low SNR, PESQ, and STOI the experimental results demonstrate that the proposed approach outperforms the Short-Time Fourier Transform (STFT) baseline in terms of enhancement ability.

In 2020, Zhou et al. [32] have suggested a modified bark spectral distortion loss mechanism, which can be thought of as an auditory perception-based MSE to replace the traditional MSE in DNN-based speech amplification approaches to increase objective perceptual efficiency even further. When compared to DNN-based methods using the traditional MSE criteria, experiments demonstrated that the proposed method can boost speech enhancement efficiency, particularly in terms of objective perceptual quality in all experimental settings.

In 2021, Chen et al. [10] have proposed a multi-objective-based multi-channel speech amplification approach. For dealing with noise and reverberation, the proposed work used the Bidirectional Long Short-Term Memory (BiLSTM) network. To the BiLSTM network, the Log-Power Spectra (LPS) of noisy speech was given as an input for each channel of the microphone array to predict the LPS and Ideal Ratio Mask (IRM) of clean speech. The intermediary LPS including IRM obtained features from both channels was further treated as a single LPS using a fusion layer. Moreover, among the clean speech LPS and the fused single-channel, the interaction taking place was learned via a DNN. Experimental findings showed the suggested speech amplification method’s viability and adaptability.

The advantages, as well as the challenges of the existing literature works discussed in the literature section, are manifested in Table 1.

3 Proposed speech enhancement model: An architectural description

The proposed speech enhancement model’s design is shown in Fig. 1, with the overall mechanism divided into “two main phases (i) Training Phase (ii) Testing Phase”.

The proposed model will be constructed by following three major phases: (a) noise spectrum and signal spectrum estimation, (b) feature extraction, (c) speech enhancement. In the training phase, the noisy signal $W(t)$ (“airport noise, exhibition noise, restaurant noise, station noise, and Street noise”) is incorporated into the clear speech signal $S(t)$. The formulated noisy speech signal is shown in Eq. (1)

$$R(t) = S(t) + W(t) \tag{1}$$

Then, for this $R(t)$, the NMF-based spectrum is estimated to find the noise spectrum $Spe_N(n)$ and signal spectrum $Spe_S(n)$ respectively. The obtained spectrum (noise and signal) is given as input to the statistical wiener filter, from which the filtered signal $F(n)$ is generated. Since the tuning factor η plays a key role in the Wiener filter, it has to be determined for each signal and is trained in the LSTM algorithm. These, filtered signals $F(n)$ are subjected to EMD, from which the denoised signal can be obtained. Then, from the denoised signal acquired via EMD, the bark frequency $b(n)$ is evaluated. Then, from the computed bark frequency, the fractional delta AMS-based features are extracted. Subsequently, with these extracted features f^{FD-AMS},

Table 1 Review on Speech Enhancement models

Author [Citation]	Adopted Technique	Advantages	Drawbacks
Zou et al. [46]	DCT	✓ Better speech quality ✓ Lower MMSE	• Higher computational complexity
Zhang et al. [37]	LSTM-convolutional-BLSTM encoder-decoder network	✓ Reduces the model complexity and training time ✓ Improves the quality and the intelligibility of enhanced speech	• Lower robustness • Not applicable for complex noise condition
Khattak et al. [29]	Phase compensated perceptually weighted-order Bayesian estimator	✓ Applicable to noise sources with fast-changing Spectrum	• No consideration on the constant Or slowly varying frequency information
Tan et al. [45]	FLGCNN	✓ Higher STOI	• Lower SNR • More residual noise exists
Zhu et al. [43]	DNN-Augmented Colored-Noise Kalman Filter	✓ Lessen the voice mumbling ✓ Remove the residual noise ✓ In both seen and unseen noise conditions, has a strong generalization capability.	• There is no balance between residual noise and voice distortion. • Reduction in speech intelligibility
Wei et al. [40]	Nonnegative matrix factorization	✓ Increases the low-frequency resolution of the speech ✓ Improved SNR ✓ Focused mainly on low-frequency signals	• Lower denoise effect • Lower PESQ
Zhou et al. [32]	DNN	✓ Improved objective perceptual quality ✓ Achieves a better PESQ score	• Reduced noise suppression
Chen et al. [10]	BiLSTM network	✓ Achieves good robustness against reverberations as well as distortion	• Low consideration of speech quality measures

the LSTM algorithm (a deep learning model) is trained. The LSTM provides the suited tuning factor η^{tuned} for the entire input signal in modified Wiener filter.

This η^{tuned} is fed as input to EMD via a modified wiener filter for decomposing the spectral signal, and the output of EMD is the denoised signal.

4 Processing steps of proposed speech enhancement model

This is the initial step, where the noise spectrum $Spe_N(n)$ and signal spectrum $Spe_S(n)$ are extracted from the noisy signal $R(t)$. The NMF model has higher physical significance and is simpler to easy than the traditional matrix decomposition algorithm; this is the reason behind the utilization of the NMF in this research work. We can get a priori information in voice applications by using train data with NMF instead of the clean signal.

To improve the speech signal, the noisy signal and the speech signal in the time-frequency (γ, p) domain is computed using STFT as defined in Eq. (2). In Eq. (3), the clear speech STFT $S(p, \gamma)$ the distorted speech STFT $R(p, \gamma)$, and the noise signal STFT $W(p, \gamma)$ are included in the p^{th} frequency bin of the γ frame. Eq. (3) shows the statistical formula for the "noisy speech's magnitude spectrum" approximation, which is the most often, used assumption for NMF-based speech and audio signal processing.

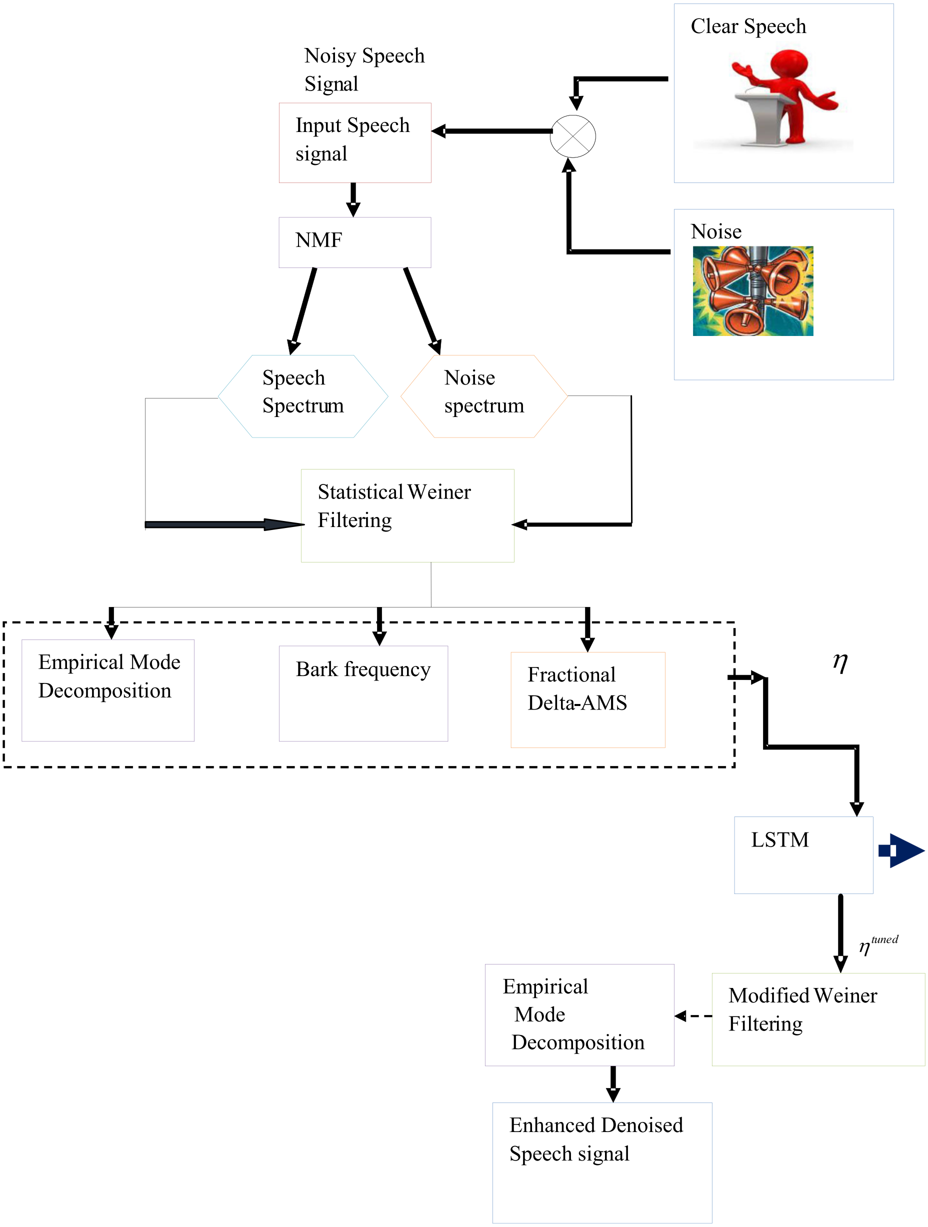


Fig. 1 Schematic overview of the proposed work

$$R(p, \gamma) = S(p, \gamma) + W(p, \gamma) \tag{2}$$

$$|R(p, \gamma)| = |S(p, \gamma) + W(p, \gamma)| \tag{3}$$

Eq. (5) shows the magnitude spectrum matrices of the various signals and $j_{p,\ \gamma}$ shows the magnitude spectral value corresponding to γ frame for the p^{th} bin. The frequency bin count is denoted by H, and the time frames are denoted by I.

$$J = \left[j_{p,\gamma}\right] \in N_+^{H\times I} \tag{4}$$

Eq. (4) is separately used in the training stage for the training data $J_S \in N_+^{H\times I_S}$ and $J_W \in N_+^{H\times I_W}$, and the results are the basis matrices in terms of clear speech $F_S = \left[r_{HI}^{S}\right] \in N_+^{H\times L_S}$ and noise $F_W = \left[r_{HI}^{W}\right] \in N_+^{H\times L_W}$, respectively. Moreover, L denotes the total number of base vectors. In Eq. (5), T' denotes the transpose of a $H \times I$ matrix ζ, whose entities are equal to one. In the enhancement stage, the basis matrices are fixed as $F_S = [F_S\ F_W] \in N_+^{H\times(L_S+L_W)}$. The activation matrix $E_{\widehat{T}} = \left[E_S^{T'} E_W^{T'}\right]^{T'} \in N_+^{(L_S+L_W)}\ \times I_{\widehat{T}}$ corresponding to the noisy speech is estimated from $J_{\widehat{T}} \in N_+^{H\times I_{\widehat{T}}}$ by means of employing the NMF activation update. Furthermore, the clear speech spectrum is evaluated from the speech signal only after obtaining the activation matrix as per Eq. (6), with the help of the Wiener Filter (WF). In Eq. (6), the approximate Positive Semi-Definite (PSD) matrices corresponding to simple speech are denoted by $V'_S = [V'_S(p, \gamma)]$, while the measured PSD matrices corresponding to noisy speech are denoted by $V'_W = [V'_W(p,\gamma)] \in N_+^{H\times I_{\widehat{T}}}$. The next solution is obtained by temporal smoothing the time, as seen in Eq. (7) and (8). Moreover, Eq. (7) and (8), demonstrate the temporal smoothing factor of speech ω_T and noise ω_W, respectively.

$$\begin{aligned} F &\leftarrow F \otimes \frac{(J/F.E)E}{\zeta E}, \\ E &\leftarrow E \otimes \frac{F(J/F.E)}{F^{T'}\zeta} \end{aligned} \tag{5}$$

$$Q = \frac{V'_S}{V'_S + V'_W} \otimes \widehat{T} \tag{6}$$

$$V'_S(p,\gamma) = \rho_S V'_S(p,\gamma-1) + (1-\rho_T)\left(\left[F_S E_S\right]_{p\gamma}\right)^2 \tag{7}$$

$$V'_W(p,\gamma) = \rho_W V'_W(p,\gamma-1) + (1-\rho_W)\left(\left[F_W E_W\right]_{p\gamma}\right)^2 \tag{8}$$

The signal spectrum $\boldsymbol{Spec}_S$ and the noise spectrum $\boldsymbol{Spec}_N$ is obtained as the outcomes. The Wiener filtering method is used to filter the received noise spectrum $\boldsymbol{Spec}_N$ and signal spectrum $\boldsymbol{Spec}_S$.

4.1 Wiener filter

“The Wiener filter’s purpose is to compute a statistical estimate of an unknown signal by taking a similar signal as an input and filtering it to create the estimate as an output”. The Wiener filter is being used on a wide scale in signal amplification techniques [36]. The Wiener filter is premised on the idea of estimating the clean signal from the distorted noise signal. The major goal of the Wiener filter is to diminish the noise from the corrupted signal. Thus, the approximation is done by reducing the MSE between the target signal and the noise distorted signal.

The estimated noise spectrum $\boldsymbol{Spec}_N$ and signal spectrum $\boldsymbol{Spec}_S$ are fed as input to statistical wiener filtering, in which $\boldsymbol{Spec}_N$ and $\boldsymbol{Spec}_S$ are filtered. The solution to this frequency-domain optimization problem is given by the filter transfer function shown in Eq. (9). The signal spectrum $\boldsymbol{Spec}_S$ and the noise spectrum $\boldsymbol{Spec}_N$ are treated as uncorrelated and stationary signals to arrive at this equation. Moreover, $\boldsymbol{Spec}_S$ has a power spectral density of $pdf_S(\omega)$, while $\boldsymbol{Spec}_N$ has a power spectral density of $pdf_W(\omega)$. Moreover, Eq. (10) shows the statistical formula for SNR, and Eq. (11) shows how the SNR formula can be used in the filter conversion function. Moreover. $G_W(\omega)$ represents the approximate signal magnitude range.

$$F(\omega) = \frac{pdf_S(\omega)}{pdf_S(\omega) + pdf_W(\omega)} \tag{9}$$

$$SNR = \frac{pdf_S(\omega)}{G_W(\omega)} \tag{10}$$

$$F(\omega) = \left[1 + \frac{1}{SNR}\right]^{-1} \tag{11}$$

At the end of filtration, the filtered signal $F(n)$ is generated. Then, from these filtered signals, the features like the EMD, bark frequency, and delta AMS are extracted.

4.2 Empirical mode curve decomposition

The EMD features are extracted from $F(n)$. Huang proposed EMD as an adaptive strategy in which a limited number of Intrinsic Mode Functions (IMF) were applied to reflect complex data. IMFs $y_e(n)$ and residue $q(n)$ are decomposed from the data set $F(n)$. Eq. (12) describes the logical formula that corresponds to this decomposition.

$$y(n) = \sum_e y_e(n) + y(n) \tag{12}$$

The steps are given below:

Step 1: Initialization,
Step 2: The d^{th} IMF is removed using the measures below.

(a) Let $k_0(n) \coloneqq q_{d-1}(n)$ and $m \coloneqq 1$ in the equation.

(b) The entire $k_{m\ -\ 1}(n)$'s local maxima and minima are established.
(c) Using cubic splines interpolation, the envelope $UB_{m\ -\ 1}(n)$ is defined by the maxima and $LB_{m\ -\ 1}(n)$ for $k_{m\ -\ 1}(n)$ by the minima.
(d) he mean $z_{m\ -\ 1}(n)$ for both envelopes belonging to $k_{m\ -\ 1}(n)$ is calculated as $z_{m-1}(n) = \frac{1}{2}(UB_{m-1}(n) - LB_{m-1}(n))$. Low-frequency local pattern is the name given to this moving mean. Furthermore, the assessment of high-frequency local detail takes place through the shifting method.
(e) $k_m(n) ≔ k_{m\ -\ 1}(n) - z_{m\ -\ 1}(n)$ is used to shape the m^{th} dimension.

- If $k_m(n)$ does not meet any of the IMF conditions, the shifting process is continued by increasing $mm + 1$ at step (b).
- If $k_m(n)$ satisfies all of the IMF conditions, then set $y_d(n) ≔ k_m(n)$ and $q_d(n) ≔ q_{d\ -\ 1}(n) - y_d(n)$

Step 3: If $q_d(n)$ represents a residuum, the shifting process may be stopped; otherwise, resume the shifting process by increasing d, $d + 1$ and starting from step 1.

Furthermore, the EMD algorithm immediately achieves the completeness of the decomposition process as $y(n) = \sum_{d=1}^{v} y_d + q$, which represents an identity. Since equivalent frequencies can be used by neighbouring IMFs at different time points, the locally orthogonal IMFs provided by the EMD algorithm do not guarantee global orthogonality. As a result of this, the bark frequency $b(u)$ is obtained.

4.3 Fractional DeltaAMS feature

The spectrum amplitudes of $b(n)$ are the AMS characteristics. The delta features are introduced by the minute variations in the frequency and time domains, and let $f(tim, \boldsymbol{freq})$ be the AMS function vector to $b(n)$. The determined feature vector is also known as Eq. (13) – Eq. (15).

$$f(tim, \boldsymbol{freq}) = \left[f(tim, \boldsymbol{freq}), \Delta f_T\left(tim, \boldsymbol{freq}\right), \Delta a_k\left(tim, \boldsymbol{freq}\right)\right] \tag{13}$$

$$\begin{gathered}\Delta f_{Tim}(tim, \boldsymbol{freq}) = f(tim, \boldsymbol{freq}) - f(tim-1, \boldsymbol{freq}); \\ \text{where} tim = 2, \ldots, Tim\end{gathered} \tag{14}$$

$$\Delta f_{Tim}(tim = 1, \boldsymbol{freq}) = f(tim = 2, \boldsymbol{freq}) - f(tim = 1, \boldsymbol{freq}) \tag{15}$$

The measured delta function vector across frequency and time is denoted by $\Delta f_{Tim}(tim, \boldsymbol{freq})$. The fractional calculus is used to obtain the most important delta-AMS features in the delta-AMS features. Incorporating the fractional calculus improves the convergence speed while reducing computing load. As a result, Eq. (15) can be rewritten as

$$\Delta f(tim, \boldsymbol{freq}) = f(tim-1, \boldsymbol{freq}) \cong E^{\sigma}[\Delta f(tim, \boldsymbol{freq})] \tag{16}$$

Here, t and s represents the windows length and count of frames, respectively and $E^{\sigma}[\Delta f(tim, \boldsymbol{freq})]$ is fractional calculus. The incorporation of $E^{\sigma}[\Delta f(tim, \boldsymbol{freq})]$ into delta-AMS features plays a key role in enhancing the speech signal. The formulated FD-AMS is denoted as per Eq. (17) and Eq. (18), respectively. The extracted fractional delta-AMS features are denoted as f^{FD-AMS}.

$$\Delta f_{Tim}(tim,\boldsymbol{freq}) = E^{\sigma}[\Delta f(tim,\boldsymbol{freq})] \tag{17}$$

$$\begin{aligned}\Delta f_{Tim}(tim,\boldsymbol{freq}) = \Delta f(tim,\boldsymbol{freq}) - \frac{1}{2}\Delta f(tim-1,\boldsymbol{freq}) - \\ \frac{1}{6}(1-\sigma)E^{\sigma}[\Delta f(tim-2,\boldsymbol{freq})] - \\ \frac{1}{24}\sigma(1-\sigma)(2-\sigma)[\Delta f(tim-3,\boldsymbol{freq})]\end{aligned} \tag{18}$$

The LSTM network is trained with the extracted features f^{FD-AMS}.

4.4 LSTM network

For speech enhancement, the extracted features f^{FD-AMS} are subjected to LSTM. A list of repeating LSTM cells has been used in the LSTM setup. Each LSTM cell is made up of three multiplicative units, which represent the "forget gate, input gate, and output gate" [31]. These units enable LSTM memory cells to store and transfer data for longer periods of time. Let the variables M and C denotes the hidden and cell states, respectively. The operation is performed by the benchmark LSTM cell while generating outputs η^{tuned}. The formulation of LSTM is given below:

$$I_t = \sigma(J_I X_t + K_I M_{t-1} + B_I) \tag{19}$$

$$F_t = \sigma(J_F X_t + K_F M_{t-1} + B_F) \tag{20}$$

$$O_t = \sigma(J_O X_t + K_O M_{t-1} + B_O) \tag{21}$$

$$C_t = F_t C_{t-1} + I_t G_t \tag{22}$$

$$G_t = tanh(J_G X_t + K_G M_{t-1} + B_G) \tag{23}$$

$$M_t = O_t tanh(C_t) \tag{24}$$

Here, I_t, F_t, and O_t are the input, forget, and output gates at a time t. The weights which map the hidden layer input to the input, forget as well as output gates are represented as J_I, J_F, and J_O. The weight matrices which map the hidden layer output to gates are denoted by K_I, K_F, and K_O. B_I, B_F, $\mathrm{B_O}$ and B_G are the bias vectors. The sigmoid function σ is used to represent the gate activation function. Furthermore, the cell outcome and layer outcome is denoted by G_t and M_t respectively. The architecture of the LSTM is shown in Fig. 2.

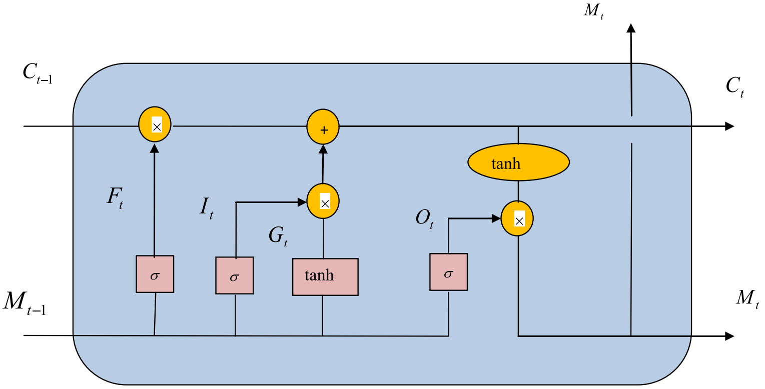


Fig. 2 The architecture of LSTM

4.5 Modified wiener filtering

The importance of the tuning ratio η^{tuned} has been well established in this research work. Based on the $b(u)$(bark frequency) of the NMF-based filtered EMD signal, the estimated tuning ratio of the Wiener filter is fine-tuned by LSTM. Mathematically, $b(u)$ can be expressed by Eq. (25).

$$b(u) = 13arctan(0.76u) + 3.5arctan\left[(0.33u)^2\right] \tag{25}$$

For tuning η in a much precise manner, we have introduced a new modified wiener filter model, which overcomes the drawbacks of the existing wiener filtering model. The existing wiener filter couldn't estimate the power spectra efficiently; it is challenging for the existing wiener filtering to acquire the perfect restoration for the random nature of the noise. Moreover, the existing wiener filter is comparatively slow to apply since they require working in the frequency domain. Interestingly, our new modified wiener filter overcomes all these drawbacks. The newly developed modified wiener filter can be formulated as per Eq. (26).

$$H(\omega) = \frac{R(\omega)}{R(\omega) + En/\left(E_y - En\right).\alpha.R(\omega)} \tag{26}$$

Here, En denotes the noise-free speech, E_y points to the energy of the noise speech energy, and $R(\omega)$ is the noisy speech signal. In addition, α represents the noise suppression factor.

The properly estimated tuning ratio η^{tuned} acquired from LSTM is fed as input to the Wiener filter, instead of the constant η. The outcomes of the Modified Wiener filter are the filtered signal $\overline{F_u(t)}$. Again, $\overline{F_u(t)}$ is decomposed using EMD and the result is the enhanced denoised signal $\overline{\overline{S(t)}}$. The training library is built using the established $b(u)$ and tuning ratio η^{tuned} as inputs during the training phase. The testing procedure is described as an online procedure, while the training procedure is described as an offline procedure. In the offline phase, the required tuning factor for various noises is identified, and the LSTM is trained with this information.

In the training process, the training library is constructed by giving the known $b(u)$ and tuning ratio η^{tuned} as inputs. The testing process is said to be the online process, while the training process is an offline process. The appropriate tuning factor for diverse noises is identified in the offline process and with this, the LSTM is trained. The tuning element is associated with the qualified network in the online mechanism, where the real improvement process takes place.

5 Results and discussion

MATLAB was used to implement the recently adopted speech amplification model. The data collection for the current study work was obtained from [4]. The five noise categories in this database, namely "airport noise, exhibition noise, restaurant noise, station noise, and street noise," are added to speech signals at differing SNR levels (0 dB, 5 dB, 10 dB, and 15 dB, respectively) to measure the efficacy of the suggested work for speech enhancement. Memory bandwidth is the amount of memory that can be used to process files in a second The total memory of the system utilized is 12GB in which the memory bandwidth of the proposed method is 1.6GB. The time step for each sequence is given in Table 2. Here time step represents the number of samples.

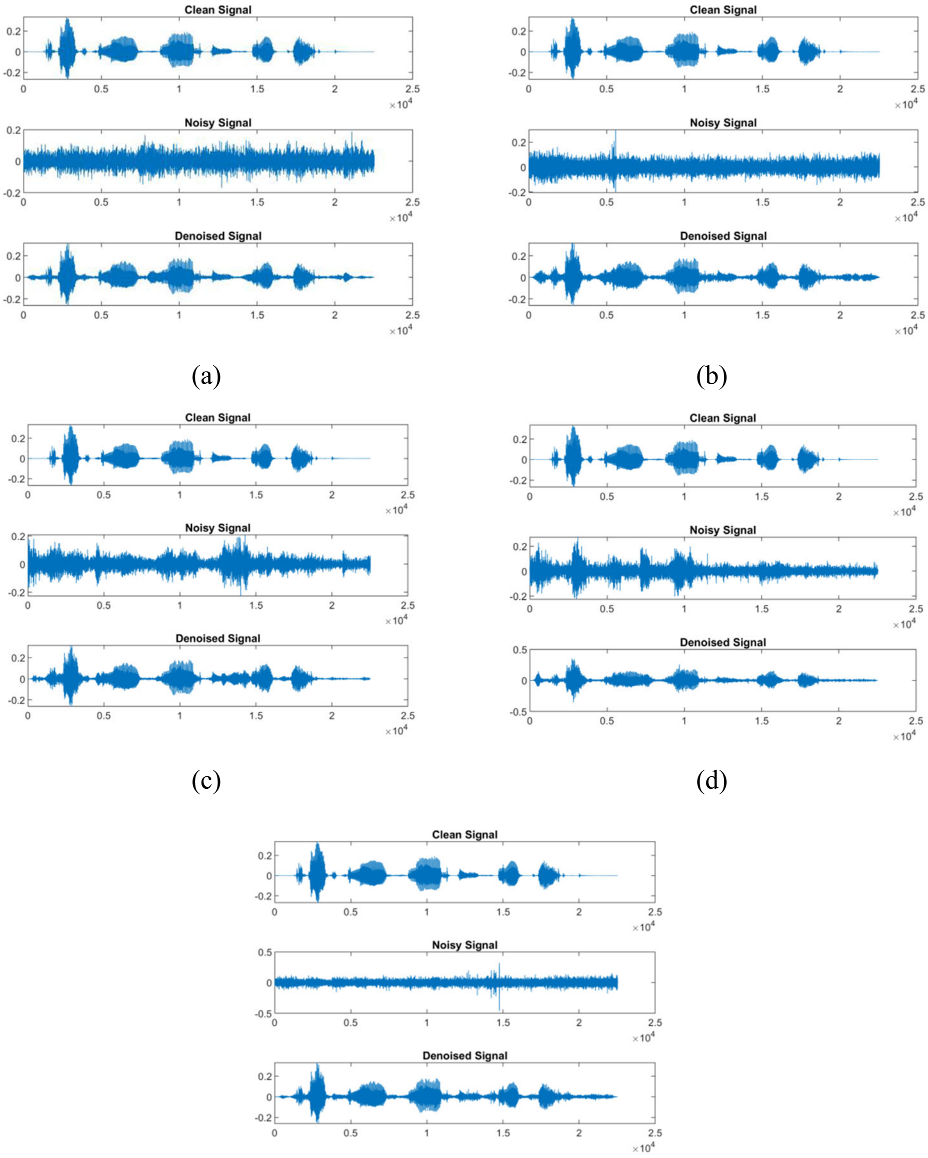
Figure 3 depicts the spectrums of the clean signal, noised signal (mixture of clean and noise signal), and de-noised signal for the airport, exhibition hall, restaurant, train station, and street, respectively. Signal-to-Distortion Ratio (SDR), Perceptual Evaluation Of Speech Quality (PESQ), Signal-To-Noise Ratio (SNR), Root-Mean-Square Error (RMSE), Correlation (CORR), Extended STOI (ESTOI), Short-Time Objective Intelligibility (STOI), and Cumulative Squared Euclidean Distance (CSED) are all used to analyze the performance of the proposed work. The comparative evaluation is made between the proposed and the existing models like multi-features+ DCNN based speech enhancement [15], Diminished Empirical Mean Curve Decomposition (D-EMCD) [14], Neural Network (NN) + auto correlation, Spectral Subtraction [7], Optimal Modified Minimum Mean Square Error Log-spectral Amplitude (OMLSA) [6], Two-step Noise Reduction (TSNR) [24], Harmonic Regeneration Noise Reduction (HRNR) [25], and Regularized Nonnegative Matrix Factorization (RNMF) [9], respectively.

5.1 Influence on airport noise under varying SNR

- In order to validate our proposed work as a significant one even under varying noise conditions, we have added the airport noise signal $W^{air}(t)$ onto the clear speech signal $S(t)$. The formulated noisy speech signal $R^{air}(t)$ is validated for varying SNR levels $R(t) = S(t) + W(t)$. At 0 dB, 5 dB, and 10 dB, 15 dB, respectively we have added the $W^{air}(t)$ onto the $S(t)$**.** Now, the formulated airport noisy signal $R^{air}(t)$ is validated over the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF in terms of SDR, PESQ,

Table 2 Time step for sequence

Sequence	Time steps
1	22,529
2	22,529
3	22,529
4	22,529
5	22,529



(a)

(b)

(c)

(d)

(e)

Fig. 3 The spectrum of the clean signal, noised signal (mixture of clean and noise signal), and de-noised signal for (**a**)airport, (**b**)exhibition hall, (**c**) restaurant, (**d**) train-station, and (**e**) street noise

SNR, RMSE, CORR, ESTOI, STOI, respectively. The obtained results are tabulated in Tables 3, 4, 5 and 6, which corresponds to different SNR rates of 0 dB, 5 dB, 10 dB, and 15 dB, respectively. When looking at the results, it's clear that the proposed work delivered the best results, with higher SDR, PESQ, CORR, ESTOI, STOI, and SNR, as well as a lower RMSE. Initially, when $R^{air}(t)$ is added at 0 dB, the proposed work seems to have achieved the highest value as 6.89, which is the best score compared to multi-

Table 3 Performance evaluation of proposed model over existing for Airport Noise at varying SNR = 0 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	5.98	2.07	32.47	0.02	0.88	0.46	0.75	4472.1
D-EMCD [14]	4.83	1.95	5.18	0.02	0.83	0.51	0.72	1870.3
NN + auto correlation	−43.19	0.54	−0.01	0.04	−0.000003	0.0008	1	7169.1
Spectral Subtraction [7]	−8.80	0.55	−0.21	0.04	−0.003	0.14	0.39	4923.3
OMLSA [6]	−23.11	1.29	−22.20	0.56	0.07	0.36	0.58	4931
TSNR [24]	−7.41	1.41	−1.004	0.05	0.01	0.34	0.55	3410.6
HRNR [25]	−7.42	1.38	−0.88	0.05	0.01	0.32	0.56	3352.9
RNMF [9]	5.82	1.93	2.78	0.03	0.80	0.47	0.68	0
Proposed	**6.89**	**2.17**	**35.20**	**0.02**	**0.89**	**0.51**	**0.73**	**1758.1**

features+ DCNN based Speech Enhancement = 5.98, D-EMCD = 4.83, NN + auto correlation = −43.19, spectral subtraction = −8.80, OMLSA = -23.11, TSNR = −7.41, HRNR = −7.42, and RNMF = 5.82. In addition, PESQ of the proposed work is 2.17 at SNR = 0 dB, which is 4.9%, 10.43%, 75.06%, 74.7%, 40.78%, 40.78%, 34.95%, 36.325% and 11.09% better than the existing models like like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. When, $R^{air}(t)$ is added at SNR = 5 dB, the proposed work is PESQ of the proposed work is 2.57, which is the maximal

Table 4 Performance evaluation of proposed model over existing for Airport Noise at varying SNR = 5 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	10.55	2.41	36.05	0.02	0.95	0.61	0.84	4676.9
D-EMCD [14]	9.26	2.36	8.23	0.02	0.92	0.64	0.81	1606.9
NN + auto correlation	−43.25	0.53	−0.004	0.04	−0.000004	0.0006	1	7952.9
Spectral Subtraction [7]	−7.70	0.82	−0.18	0.04	−0.005	0.19	0.47	4530.9
OMLSA [6]	−22.66	1.32	−22.19	0.55	0.08	0.48	0.67	4754.4
TSNR [24]	−6.77	1.86	−0.89	0.05	0.02	0.46	0.66	3215.7
HRNR [25]	−6.81	1.85	−0.82	0.05	0.02	0.44	0.66	3161.4
RNMF [9]	8.23	2.31	3.67	0.03	0.87	0.61	0.76	2999
Proposed	**10.87**	**2.57**	**38.29**	**0.01**	**0.95**	**0.71**	**0.84**	**1471.6**

Table 5 Performance evaluation of proposed model over existing for Airport Noise at varying SNR = 10 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	13.69	2.69	10.57	0.01	0.97	0.75	0.89	4604.2
D-EMCD [14]	12.45	2.69	10.54	0.01	0.96	0.77	0.88	1392.5
NN + auto correlation	−45.51	0.45	−0.001	0.04	0.00003	−0.00006	1	9049.1
Spectral Subtraction [7]	−7.46	0.99	−0.21	0.04	−0.002	0.25	0.52	4275.7
OMLSA [6]	−22.42	1.47	−22.19	0.56	0.08	0.59	0.74	4621.7
TSNR [24]	−6.69	2.28	−0.91	0.05	0.02	0.59	0.75	3032.3
HRNR [25]	−6.73	2.31	−0.85	0.05	0.02	0.59	0.75	3016.9
RNMF [9]	9.36	2.53	4.53	0.03	0.89	0.71	0.82	2761.1
Proposed	**13.19**	**2.63**	**40.68**	**0.009**	**0.972**	**0.79**	**0.91**	**1291.9**

Table 6 Performance evaluation of proposed model over existing for Airport Noise at varying SNR = 15 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	15.2	3.06	10.01	0.01	0.97	0.84	0.93	4851.6
D-EMCD [14]	14.21	2.97	11.91	0.01	0.97	0.84	0.91	1303.6
NN + auto correlation	−47.09	0.64	−0.0004	0.04	0.00002	0.00006	1	10,307
Spectral Subtraction [7]	−7.49	1.07	−0.24	0.04	−0.004	0.27	0.54	4136.2
OMLSA [6]	−22.26	1.67	−22.19	0.56	0.08	0.66	0.77	4562.3
TSNR [24]	−6.75	2.75	−0.92	0.05	0.02	0.69	0.80	2940.4
HRNR [25]	−6.78	2.79	−0.88	0.05	0.02	0.68	0.79	2929
RNMF [9]	9.95	2.72	5.15	0.02	0.90	0.75	0.84	0
Proposed	**16.48**	**2.98**	**42.97**	**0.007**	**0.98**	**0.89**	**0.95**	**1018.8**

value when compared to multi-features+ DCNN based Speech Enhancement = 2.41, D-EMCD = 2.36, NN + auto correlation = 0.53, spectral subtraction = 0.82, OMLSA = 1.32, TSNR = 1.86, HRNR = 1.85, and RNMF = 2.31. In addition, the proposed work has achieved the maximal ESTOI as 0.71, which is 13.83%, 8.6%, 99.9%, 71.9%, 31.4%,31.4%, 34.8%, 37.65%, and 13.5% better than the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively at SNR = 5 dB. In addition, when $R^{air}(t)$ is added at 10 dB, the proposed work seems to have attained the favorable outcome as shown in Table 4. The CORR of the proposed work at SNR = 10 dB is 0.97, which is 0.18%, 0.14%, 99.9%, 92%, 91.85%, 91.85%, 97.6%, 97.7% and 8.4% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. Moreover, on observing the outcomes from Table 5, the proposed work had achieved the least RMSE as 0.007, which is 50.2%, 34.1%, 82.8%, 83.2%, 98.6%, 98.6%, 84.5%, 84.4%, and 69.2% better than the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. The change is made in the wiener filter. Furthermore, the LSTM model is used to accurately estimate the tuning factor of the Wiener filter for all input signals. During the testing phase, the extracted features (EMD) were used to train the LSTM model using a modified Wiener filter. Thus, the proposed work had enhanced the quality of the speech signal even under the airport environment.

5.2 Influence on exhibition hall noise under varying SNR

The noise created from the exhibition hall $W^{hall}(t)$ is added to $S(t)$ at varying SNR rates. The formulated noisy signal $R(t) = S(t) + W^{hall}(t)$ is evaluated in terms of SDR, PESQ, SNR, RMSE, CORR, ESTOI, STOI, respectively. The result acquired is tabulated in Tables 7, 8, 9 and 10, corresponding to varying SNR rates: 0 dB, 5 dB, and 10 dB, 15 dB, respectively. When adding $W^{hall}(t)$=0 dB to $S(t)$, the proposed work has achieved the highest SNR as 34.27, which is better than existing models like the existing models like multi-features+ DCNN based Speech Enhancement = 5.24, D-EMCD = 5.16, NN + auto correlation = −0.006, spectral

Table 7 Performance evaluation of proposed model over existing for exhibition Hall Noise at varying SNR = 0 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	5.36	1.73	5.24	0.02	0.86	0.52	0.69	4607.1
D-EMCD [14]	5.26	1.88	5.16	0.02	0.83	0.56	0.73	2406.2
NN + auto correlation	−41.66	0.43	−0.006	0.04	−0.00003	−0.00009	1	6885.2
Spectral Subtraction [7]	−9.78	0.43	−0.31	0.04	−0.006	0.15	0.39	4998.4
OMLSA [6]	−23.21	1.15	−22.20	0.56	0.07	0.43	0.61	5203.8
TSNR [24]	−7.32	1.24	−0.99	0.05	0.01	0.43	0.61	3570.9
HRNR [25]	−7.35	1.21	−0.87	0.05	0.01	0.42	0.62	3523.7
RNMF [9]	6.22	1.85	2.81	0.03	0.79	0.53	0.69	4127
Proposed	**5.84**	**1.81**	**34.27**	**0.02**	**0.87**	**0.51**	**0.72**	**2392.1**

subtraction = −0.31, OMLSA = −22.20, TSNR = −0.99, HRNR = −0.87, and RNMF = 2.81. In addition, the RMSE of the proposed work is 18.7%, 17.7%, 53.5%, 55.18%, 96.4%, 96.4%, 58.5%, 57.9% and 36.3% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. In addition, when

Table 8 Performance evaluation of proposed model over existing for exhibition Hall Noise at varying SNR = 5 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	10.60	1.98	8.41	0.02	0.95	0.51	0.76	4181.5
D-EMCD [14]	9.37	2.25	8.15	0.02	0.92	0.68	0.82	2048.5
NN + auto correlation	−46.41	0.60	−0.002	0.04	0.00001	0.0002	1	7163.5
Spectral Subtraction [7]	−8.64	0.72	−0.24	0.04	−0.004	0.21	0.46	4556
OMLSA [14]	−22.64	1.34	−22.19	0.56	0.08	0.53	0.69	4956.9
TSNR [24]	−6.90	1.79	−0.96	0.05	0.02	0.54	0.70	3218.2
HRNR [25]	−6.94	1.79	−0.89	0.05	0.02	0.54	0.71	3179
RNMF [9]	8.69	2.19	3.84	0.03	0.86	0.64	0.77	3873.4
Proposed	**9.11**	**2.13**	**36.81**	**0.01**	**0.93**	**0.56**	**0.77**	**2000.4**

Table 9 Performance evaluation of proposed model over existing for exhibition Hall Noise at varying SNR = 10 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	13.76	2.49	12.95	0.01	0.97	0.65	0.87	4376.1
D-EMCD [14]	12.55	2.58	10.38	0.01	0.96	0.77	0.88	1812.9
NN + auto correlation	−47.84	0.39	−0.0009	0.04	−0.00001	0.0009	1	8363.9
Spectral Subtraction [7]	−7.67	0.92	−0.22	0.04	−0.006	0.25	0.51	4502.9
OMLSA [6]	−22.39	1.49	−22.19	0.56	0.08	0.61	0.74	4761.9
TSNR [24]	−6.69	2.25	−0.93	0.05	0.02	0.62	0.76	3168.5
HRNR [25]	−6.73	2.26	−0.88	0.05	0.02	0.62	0.76	3094.6
RNMF [9]	10.26	2.48	4.81	0.02	0.89	0.72	0.83	0
Proposed	**14.65**	**2.66**	**41.59**	**0.009**	**0.97**	**0.73**	**0.86**	**1543.1**

Table 10 Performance evaluation of proposed model over existing for exhibition Hall Noise at varying SNR = 15 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	16.23	2.82	13.85	0.01	0.98	0.79	0.93	4601
D-EMCD [14]	15.03	2.88	12.16	0.01	0.97	0.85	0.92	1666.9
NN + auto correlation	−47.88	0.45	−0.0004	0.04	0.00002	0.0005	1	9176.9
Spectral Subtraction [7]	−7.52	1.03	−0.24	0.04	−0.005	0.27	0.54	4272.4
OMLSA [6]	−22.26	1.62	−22.19	0.56	0.08	0.67	0.77	4637.6
TSNR [24]	−6.73	2.65	−0.93	0.05	0.020	0.69	0.80	3036.8
HRNR [25]	−6.76	2.67	−0.89	0.05	0.02	0.69	0.80	2974.6
RNMF [9]	11.13	2.72	5.59	0.02	0.91	0.78	0.86	3443.7
Proposed	**17.27**	**2.90**	**43.01**	**0.007**	**0.98**	**0.81**	**0.91**	**1567.6**

the $W^{hall}(t)$ is added to $S(t)$ at 5 dB, the SNR of the proposed work is 36.81, which is 77.16%, 77.85%, 99.9%, 99.3%, 39.7%, 39.7%, 97.3%, 97.52%, and 89.5% better than the existing models multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. In addition, the CORR of the proposed work is 2.1%, 0.4%, 99.9%, 99.56%, 91.8%, 91.8%, 97.9%, 97.9%, and 6.9% better than the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. Moreover, when the $W^{hall}(t)$ is added at 10 dB to $S(t)$, the $R(t)$ acquired from the proposed work in terms of PESQ is 2.663, which is better than existing models like multi-features+ DCNN based Speech Enhancement = 2.49, D-EMCD = 2.58, NN + auto correlation = 0.39, spectral subtraction = 0.92, OMLSA = 1.49, TSNR = 2.25, HRNR = 2.26, and RNMF = 2.48, respectively. Moreover, when $W^{hall}(t)$ is applied at 15 dB to $S(t)$, the proposed speech signal is 0.3%, 0.7%, 84.3%, 64.4%, 44.2%, 44.2%, 8.6%, 8.15%, 8.15% and 6.3% better than the existing work in terms of PESQ. Moreover, in case of SNR for $W^{hall}(t)$ applied at 15 dB, the proposed work is 67.8%, 71.7%, 99.9%, 99.43%, 48.3%, 48.3%, 97.8%, 97.9%, and 87% better than the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. The final result is indeed a speech-enhanced signal with negligible noise. The wiener filter has undergone changes. Moreover, for all input signals, the LSTM model is used to accurately estimate the Wiener filter tuning factor. The extracted features (EMD) were used to train the LSTM model using a modified Wiener filter during the testing phase. Therefore, from the evaluation, it is clear that the proposed work is applicable even under the exhibition hall.

5.3 Influence on restaurant noise under varying SNR

The restaurant noise $W^{rest}(t)$ is added to $S(t)$ at varying SNR rates. The formulated noisy signal $R(t) = S(t) + W^{rest}(t)$ is evaluated in terms of SDR, PESQ, SNR, RMSE, CORR, ESTOI, STOI, respectively. The result acquired is tabulated in Tables 11, 12, 13 and 14, corresponding to varying SNR rates: 0 dB, 5 dB, and 10 dB, 15 dB, respectively. While adding $W^{rest}(t)$=0 dB, the SNR of the proposed work is 33.90, which is better than the existing models like multi-features+ DCNN based Speech Enhancement =5.99, D-EMCD = 4.55, NN + auto correlation = −0.009, spectral subtraction = −0.17, OMLSA = −22.21, TSNR = −1.05, HRNR =

Table 11 Performance evaluation of proposed model over existing for restaurant noise at varying SNR = 0 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	5.29	2.05	5.99	0.03	0.86	0.56	0.78	4416
D-EMCD [14]	3.76	1.93	4.55	0.03	0.81	0.51	0.69	1964
NN + auto correlation	−44.85	0.47	−0.009	0.04	−0.00001	0.0001	1	6974.2
Spectral Subtraction [7]	−9.14	0.55	−0.17	0.043	−0.003	0.13	0.37	4852.8
OMLSA [6]	−23.03	1.37	−22.21	0.56	0.07	0.37	0.58	4977.1
TSNR [24]	−7.69	1.37	−1.05	0.05	0.01	0.34	0.55	3560.1
HRNR [25]	−7.71	1.35	−0.92	0.05	0.005	0.32	0.55	3538
RNMF [9]	5.19	1.85	2.58	0.03	0.78	0.49	0.66	3378.8
Proposed	**5.37**	**2.22**	**33.90**	**0.02**	**0.86**	**0.55**	**0.73**	**1765.8**

−0.92, and RNMF = 2.58, respectievly. In addition, for $W^{rest}(t)$=0 dB, the RMSE of the proposed work is 0.02, which is 26.4%, 20.6%, 52.3%, 53.4%, 96.3%, 96.3%, 57.9%, 57.3% and 36.5% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. Moreover, when adding $W^{rest}(t)$ at 5 dB, the CORR of the proposed work is 0.93016, which is better than models like multi-features+ DCNN based Speech Enhancement = 0.91, D-EMCD = 0.92, NN + auto correlation = −0.000008, spectral subtraction = −0.005, OMLSA = 0.08, TSNR = 0.02, HRNR = 0.02, and RNMF =

Table 12 Performance evaluation of proposed model over existing for restaurant noise at varying SNR = 5 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	8.92	2.35	7.89	0.01	0.91	0.62	0.85	4117.3
D-EMCD [14]	8.69	2.27	7.87	0.02	0.92	0.65	0.81	1729.6
NN + auto correlation	−41.91	0.55	−0.003	0.04	−0.000008	−0.0002	1	8276.5
Spectral Subtraction [7]	−7.90	0.80	−0.19	0.04	−0.005	0.19	0.46	4413.9
OMLSA [6]	−22.56	1.31	−22.19	0.56	0.08	0.49	0.67	4773.8
TSNR [24]	−7.06	1.84	−0.98	0.05	0.02	0.49	0.68	3129.5
HRNR [25]	−7.09	1.85	−0.91	0.05	0.02	0.48	0.68	3100.8
RNMF [9]	7.96	2.19	3.55	0.03	0.86	0.62	0.76	0
Proposed	**8.96**	**2.38**	**36.87**	**0.01**	**0.93**	**0.65**	**0.81**	**1598.8**

Table 13 Performance evaluation of proposed model over existing for restaurant noise at varying SNR = 10 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	11.72	2.68	10.87	0.01	0.96	0.78	0.93	4177
D-EMCD [14]	11.89	2.66	10.09	0.01	0.95	0.77	0.87	1532.7
NN + auto correlation	−44.57	0.48	−0.0009	0.04	−0.000007	0.0004	1	9166.9
Spectral Subtraction [7]	−7.57	0.96	−0.22	0.04	−0.004	0.25	0.52	4297.5
OMLSA [6]	−22.38	1.49	−22.19	0.56	0.08	0.59	0.74	4651
TSNR [24]	−6.73	2.24	−0.91	0.05	0.02	0.59	0.76	3098.7
HRNR [25]	−6.76	2.26	−0.85	0.05	0.02	0.58	0.75	3050.4
RNMF [9]	9.31	2.45	4.37	0.03	0.88	0.71	0.81	2974.8
Proposed	**13.94**	**2.77**	**40.97**	**0.009**	**0.97**	**0.79**	**0.90**	**1273.3**

Table 14 Performance evaluation of proposed model over existing for restaurant noise at varying SNR = 15 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	13.98	2.96	38.67	0.01	0.98	0.81	0.97	4826.8
D-EMCD [14]	13.82	2.91	11.50	0.01	0.97	0.84	0.91	1375.6
NN + auto correlation	−45.91	0.49	−0.0003	0.04	0.00003	0.002	1	10,500
Spectral Subtraction [7]	−7.44	1.05	−0.24	0.04	−0.005	0.28	0.54	4145.3
OMLSA [6]	−22.26	1.59	−22.19	0.56	0.08	0.66	0.77	4573.3
TSNR [24]	−6.74	2.62	−0.92	0.05	0.02	0.68	0.80	2944.5
HRNR [25]	−6.77	2.66	−0.88	0.05	0.02	0.68	0.79	2923.7
RNMF [9]	9.77	2.62	4.94	0.02	0.89	0.77	0.84	2747.8
Proposed	**15.44**	**2.88**	**41.32**	**0.009**	**0.98**	**0.85**	**0.93**	**1255.8**

0.86, respectievly. In addition, the PESQ of the proposed work is 0.12%, 4.7%, 77.01%, 66.4%, 45.1%, 45.1%, 22.7%, 22.49% and 8.1% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. Moreover, when the $W^{rest}(t)$=10 dB is added to the input clear speech signal, the processed outcomes from proposed model has generated a higher speech quality signal's. here, when $W^{rest}(t)$=10 dB, the RMSE of the proposed work is 33.6%, 31.9%, 78.3%, 78.8%, 98.3%, 98.3%, 80.5%, 80.37% and 64.6% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. In addition, ESTOI of the proposed work is 0.79, which is 0.15%, 3.9%, 99.9%, 68.9%, 25.3%, 25.3%, 25.79%, 26.8% and 11.455 better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. Moreover, when the $W^{rest}(t)$=15 dB is added to the clear signal, the processed outcomes from the proposed work in terms of SNR is 6.4%, 72.1%, 99.9%, 99.4%, 46.2%, 46.28%, 97.7%, 97.86% and 88.04% better than the existing models like the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectievly. The modification is made to the wiener filter. The improved wiener filter is used rather than using conventional one. Thus, the betterment of the proposed work has been proved over the other existng models.

5.4 Influence on railway station noise under varying SNR

To the clear speech signal $S(t)$, the railway station noise $W^{rail}(t)$ is added under varying SNR rates, and the outcomes acquired after de-noising is evaluated in terms of SDR, PESQ, SNR, RMSE, CORR, ESTOI, STOI, respectively. The results acquired are tabulated in Tables 15, 16, 17 and 18 corresponding to varying SNR rates: 0 dB, 5 dB, and 10 dB, 15 dB respectively. On applying the $W^{rail}(t)$ at SNR = 0 dB, the SNR of the proposed work is 35.03, which is the highest value and it is 83.26%, 84.2%, 99.9%, 99.9%, 36.6%, 36.6%, 97.6%, 97.9%, and 92.65% better than the existing models like multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. In addition, $W^{rail}(t)$ is applied to $S(t)$ at 5 dB, the proposed had achieved the least RMSE as 0.01, which is the least value when compared to multi-features+ DCNN

Table 15 Performance evaluation of proposed model over existing for station noise at varying SNR = 0 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	6.18	2.10	5.74	0.03	0.87	0.45	0.75	4140.4
D-EMCD [14]	6.11	2.01	5.52	0.03	0.86	0.47	0.69	2117.4
NN + auto correlation	−43.04	0.52	−0.009	0.04	−0.00001	0.0004	1	7277.9
Spectral Subtraction [7]	−8.78	0.59	−0.17	0.04	−0.007	0.13	0.41	5326.7
OMLSA [6]	−23.61	1.14	−22.20	0.56	0.06	0.33	0.55	5066.9
TSNR [24]	−6.96	1.41	−0.83	0.05	0.01	0.32	0.55	3502.5
HRNR [25]	−7.02	1.31	−0.722	0.05	0.01	0.28	0.55	3433.7
RNMF [9]	6.67	1.99	2.57	0.03	0.82	0.45	0.67	0
Proposed	**7.40**	**2.26**	**35.03**	**0.02**	**0.89**	**0.53**	**0.74**	**1904.9**

based Speech Enhancement = 0.02, D-EMCD = 0.02, NN + auto correlation = 0.04, spectral subtraction = 0.04, OMLSA = 0.56, TSNR = 0.05, HRNR = 0.05, and RNMF = 0.03. In addition, the RMSE of the proposed work after the application of $W^{rail}(t)$ at SNR = 10 dB is 0.0083884, which is 32.6%, 35.3%, 80.4%, 80.9%, 98.4%, 98.49%, 82.4%, 82.2% and 67.6% better than the existing models multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. In addition, while applying the $W^{rail}(t)$ at SNR = 15 dB, the proposed work has

Table 16 Performance evaluation of proposed model over existing for station noise at varying SNR = 5 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	11.42	2.34	9.92	0.02	0.96	0.54	0.83	4485.2
D-EMCD [14]	9.90	2.45	8.35	0.02	0.93	0.65	0.81	1770.9
NN + auto correlation	−43.05	0.58	−0.003	0.04	0.00001	0.0012	1	7636.7
Spectral Subtraction [7]	−7.59	0.86	−0.18	0.04	−0.002	0.19	0.48	4981.5
OMLSA [6]	−22.88	1.37	−22.19	0.56	0.074	0.47	0.66	4827.3
TSNR [24]	−6.69	1.96	−0.88	0.05	0.02	0.46	0.69	3293.8
HRNR [25]	−6.75	1.89	−0.80	0.05	0.02	0.42	0.66	3211.3
RNMF [9]	8.34	2.33	3.39	0.03	0.87	0.60	0.77	3494.5
Proposed	**11.98**	**2.57**	**39.25**	**0.01**	**0.96**	**0.66**	**0.82**	**1537.3**

Table 17 Performance evaluation of proposed model over existing for station noise at varying SNR = 10 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	13.45	2.46	11.68	0.01	0.97	0.77	0.92	4316
D-EMCD [14]	12.53	2.73	10.52	0.01	0.96	0.76	0.87	1534.6
NN + auto correlation	−48.81	0.53	−0.001	0.04	0.00003	−0.001	1	8473.8
Spectral Subtraction [7]	−7.49	0.99	−0.21	0.04	−0.005	0.24	0.52	4366.4
OMLSA [6]	−22.47	1.49	−22.19	0.56	0.08	0.59	0.73	4668.8
TSNR [24]	−6.67	2.35	−0.91	0.05	0.02	0.60	0.75	3066.5
HRNR [25]	−6.71	2.34	−0.85	0.05	0.02	0.59	0.75	3048.8
RNMF [9]	9.52	2.55	4.47	0.03	0.89	0.70	0.82	0
Proposed	**14.81**	**2.798**	**41.61**	**0.008**	**0.98**	**0.77**	**0.89**	**1380.5**

Table 18 Performance evaluation of proposed model over existing for station noise at varying SNR = 15 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	14.83	2.63	9.48	0.02	0.97	0.69	0.88	4909.3
D-EMCD [14]	14.04	3.01	11.6	0.01	0.97	0.84	0.91	1437.1
NN + auto correlation	−45.75	0.55	−0.0004	0.04	−0.000009	−0.0009	1	10,002
Spectral Subtraction [7]	−7.41	1.07	−0.23	0.04	−0.003	0.26	0.54	4170.2
OMLSA [6]	−22.29	1.64	−22.19	0.55	0.08	0.66	0.77	4586.9
TSNR [24]	−6.70	2.81	−0.93	0.05	0.02	0.68	0.79	2963.7
HRNR [25]	−6.73	2.79	−0.88	0.05	0.02	0.67	0.79	2947.7
RNMF [9]	9.96	2.71	5.05	0.02	0.90	0.76	0.84	2838.8
Proposed	**15.02**	**2.79**	**40.21**	**0.01**	**0.97**	**0.85**	**0.93**	**1312.3**

achieved the highest SNR value as 40.21, while the existing models had recorded the SNR value as multi-features+ DCNN based Speech Enhancement = 9.49, D-EMCD = 11.6, NN + auto correlation = −0.0004, spectral subtraction = −0.23, OMLSA = −22.19, TSNR = −0.93, HRNR = −0.87, and RNMF = 5.05. Furthermore, after analyzing the proposed work with $W^{rail}(t)$=15 dB, the proposed work seems to have obtained the best results. During the testing phase, the extracted features (EMD) were used to train the LSTM model using a modified Wiener filter. As a result of the evaluation, it is clear that the proposed study is effective in improving the speech signal even when station noise is present.

5.5 Influence on street noise under varying SNR

The street noise $W^{street}(t)$ is added to the clear signal $S(t)$ at varying, SNR, as well as the results obtained after de-noising, are measured in terms of SDR, PESQ, SNR, RMSE, CORR, ESTOI, and STOI. The obtained results are tabulated in Tables 19, 20, 21 and 22, which correspond to different SNR rates of 0 dB, 5 dB, 10 dB, and 15 dB, respectively. From the acquired outcomes, the RMSE of the proposed work is found to be lower even under every variation in the application of $W^{street}(t)$ rate. At SNR = 0 dB, 10 dB, 15 dB, 20 dB, the proposed work had achieved the least value RMSE as 0.02, 0.01, 0.009 and 0.006 respectively. Moreover, on analyzing the other outcomes, the proposed work had recorded the highest SDR, PESQ, SNR, CORR, ESTOI, and STOI, which are said to be the appropriate values for speech enhancement. In addition on adding $W^{street}(t)$ at 5 dB, the SNR of the proposed work is 39.26, which is the better value when compared to existing models like multi-features+ DCNN based Speech

Table 19 Performance evaluation of proposed model over existing for street noise at varying SNR = 0 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	6.61	1.88	6.38	0.02	0.89	0.45	0.70	4097.1
D-EMCD [14]	6.50	2.05	5.83	0.02	0.86	0.51	0.71	2275.8
NN + auto correlation	−29.07	0.39	−0.01	0.04	−0.000001	−0.003	1	6669.8
Spectral Subtraction [7]	−8.79	0.69	−0.16	0.04	−0.01	0.11	0.39	7117.7
OMLSA [6]	−23.38	1.29	−22.20	0.48	0.07	0.33	0.58	5198.1
TSNR [24]	−6.36	1.23	−0.88	0.04	−0.06	0.24	0.49	4264.4
HRNR [25]	−6.37	1.11	−0.77	0.04	−0.06	0.20	0.50	4264.5
RNMF [9]	7.20	2.09	2.75	0.03	0.83	0.49	0.682	4175
Proposed	**6.82**	**1.95**	**35.06**	**0.02**	**0.86**	**0.45**	**0.71**	**2330.5**

Table 20 Performance evaluation of proposed model over existing for station noise at varying SNR = 5 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	12.29	2.47	8.91	0.01	0.96	0.59	0.80	4628.4
D-EMCD [14]	10.49	2.41	8.65	0.02	0.93	0.64	0.79	1964.3
NN + auto correlation	−41.19	0.35	−0.0032	0.04	0.00002	−0.003	1	7176.4
Spectral Subtraction [7]	−8.47	0.84	−0.18	0.04	−0.009	0.21	0.47	5217.1
OMLSA [6]	−21.99	1.50	−22.19	0.48	0.09	0.49	0.67	4858.9
TSNR [24]	−7.06	1.88	−0.89	0.05	−0.01	0.44	0.67	3119.2
HRNR [25]	−7.09	1.89	−0.81	0.05	−0.01	0.42	0.68	3073.4
RNMF [9]	8.62	2.37	3.49	0.03	0.86	0.60	0.76	3934
Proposed	**11.85**	**2.54**	**39.26**	**0.01**	**0.96**	**0.66**	**0.83**	**1738.8**

Enhancement = 8.91, D-EMCD = 8.65, NN + auto correlation = −0.003, spectral subtraction = −0.18, OMLSA = −22.19, TSNR = −0.89, HRNR = −0.81, and RNMF = 3.49. Moreover, when $W^{street}(t)$=10 dB is applied to the clean speech signal, the outcome from the proposed work in terms of RMSE = 0.009, which is better than the existing models like multi-features+ DCNN based Speech Enhancement = 0.01, D-EMCD = 0.01, NN + auto correlation = 0.04, spectral subtraction = 0.04, OMLSA = 0.48, TSNR = 0.05, HRNR = 0.05, and RNMF = 0.02. Moreover, for all input signals, the LSTM model is used to accurately estimate the Wiener filter tuning factor. The extracted features (EMD) were used to train the LSTM model using a modified Wiener filter during the testing phase. Therefore from the evaluation, it is clear that the proposed work is much significant for enhancing the speed signal.

5.6 Statistical Anaysis

Table 23 shows the statistical analysis of the proposed model over existing methods. On considering the SDR measure, the STD value of the proposed model is 4.23%, 10.20%, 5.42%, 82.84%, 86.82%, 92.29%, 92.29%,and 58.46% better than the existing multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF models. In RMSE measure, the best value of the proposed model is 40%, 40%, 85%, 85%, 98.75%, 85%, 85%, and 70% superior to existing models multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. Further considering the SNR measure, the best value of the proposed model is 33.90, which is the better

Table 21 Performance evaluation of proposed model over existing for station noise at varying SNR = 10 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	13.77	2.71	9.97	0.01	0.97	0.77	0.89	4742.3
D-EMCD [14]	13.45	2.77	11.07	0.01	0.96	0.76	0.87	1766.5
NN + auto correlation	−39.2	0.42	−0.0018	0.04	−0.00003	0.0002	1	8980.5
Spectral Subtraction [7]	−8.35	1.02	−0.22	0.04	−0.006	0.23	0.50	4747.7
OMLSA [6]	−21.74	1.73	−22.19	0.48	0.09	0.64	0.79	4659.5
TSNR [24]	−6.68	2.25	−0.93	0.05	−0.02	0.57	0.77	3031.1
HRNR [25]	−6.72	2.29	−0.87	0.05	−0.02	0.56	0.78	3009.6
RNMF [9]	10.05	2.65	4.86	0.02	0.90	0.70	0.82	0
Proposed	**14.74**	**2.83**	**41.76**	**0.009**	**0.98**	**0.78**	**0.89**	**1422.5**

Table 22 Performance evaluation of proposed model over existing for station noise at varying SNR = 15 dB

Approach	SDR	PESQ	SNR	RMSE	CORR	ESTOI	STOI	CSED
multi-features+ DCNN based Speech Enhancement [15]	17.98	2.93	13.53	0.01	0.99	0.86	0.95	4567.2
D-EMCD [14]	14.98	2.98	12.20	0.01	0.97	0.82	0.89	1597.3
NN + auto correlation	−43.03	0.50	−0.00043	0.041	0.00008	0.003	1	9153.8
Spectral Subtraction [7]	−8.46	1.10	−0.24	0.04	−0.009	0.27	0.54	4621.3
OMLSA [6]	−21.55	1.85	−22.19	0.48	0.09	0.73	0.83	4601.2
TSNR [24]	−6.64	2.59	−0.94	0.05	−0.02	0.65	0.82	2978.2
HRNR [25]	−6.66	2.57	−0.89	0.05	−0.02	0.65	0.82	2931.2
RNMF [9]	10.56	2.75	5.44	0.02	0.91	0.75	0.83	3281.5
Proposed	**17.99**	**3.21**	**44.84**	**0.0061**	**0.99**	**0.87**	**0.93**	**1278.4**

value when compared to existing models like multi-features+ DCNN based Speech Enhancement = 5.24, D-EMCD = 4.55, NN + auto correlation = −0.01, spectral subtraction = −0.31, OMLSA = −22.21, TSNR = −1.05, HRNR = −0.92, and RNMF = 2.57 respectively. Likewise, other measures also show a better performance. Therefore, from the analysis, the proposed model is proven to be a suitable model for speech enhancement.

5.7 Discussions

The major goal of this study is to enhance the speech signals with various noise sources. The results section evaluated the proposed model with different noise sources such as "airport noise, exhibition noise, restaurant noise, station noise, and street noise". The various noise sources are analyzed under different SNR values in terms of speech quality measures. By utilizing the modified wiener filter and extracted features assisted LSTM model, the denoised speech signal is obtained. Compared to the existing models, the proposed method achieves higher SDR, PESQ, CORR, ESTOI, STOI, and SNR, as well as lower RMSE values. Moreover, the proposed model overcomes the drawbacks such as reduction in speech intelligibility [43], lower PESQ [40], lower robustness [37], not being suitable for complex noise environments [37], lower speech quality [10], and low SNR [45] [29]. However, this proposed method lacks at some noise sources and it did not determine the spectral magnitude and spectral phase estimation.

5.8 Practical implication

The main potential applications of the proposed model are given below:

- Hearing aids
- Automatic speech recognition
- Mobile communications
- Video captioning for teleconferences
- Voice over Internet protocol
- Hand-free communications

This research provides better outcomes and it suits many potential application fields.

Table 23 Statistical analysis of proposed model over existing methods

Approach	multi-features + DCNN based Speech Enhancement [15]	D-EMCD[14]	NN + auto correlation	Spectral Subtraction [7]	OMLSA[6]	TSNR [24]	HRNR [25]	RNMF [9]	Proposed
SDR									
Best	5.29	3.76	−48.81	−9.78	−23.61	−7.69	−7.71	5.19	**5.37**
Worst	17.98	15.03	−29.07	−7.41	−21.55	−6.36	−6.37	11.13	**17.99**
Mean	11.39	10.46	−43.66	−8.12	−22.55	−6.86	−6.89	8.64	**11.93**
Median	12.01	11.19	−43.91	−7.80	−22.44	−6.73	−6.76	8.99	**12.59**
STD	3.85	3.61	4.25	0.69	0.53	0.31	0.31	1.67	**4.02**
PESQ									
Best	1.73	1.88	0.35	0.43	1.14	1.23	1.11	1.84	**1.81**
Worst	3.06	3.01	0.64	1.10	1.85	2.81	2.79	2.75	**3.21**
Mean	2.44	2.48	0.49	0.85	1.46	2.04	2.03	2.37	**2.55**
Median	2.46	2.51	0.50	0.89	1.48	2.10	2.08	2.41	**2.60**
STD	0.38	0.38	0.08	0.20	0.19	0.52	0.55	0.30	**0.37**
SNR									
Best	5.24	4.55	−0.01	−0.31	−22.21	−1.05	−0.92	2.57	**33.90**
Worst	38.67	12.20	−0.0003	−0.16	−22.19	−0.83	−0.72	5.59	**44.84**
Mean	13.43	8.97	−0.003	−0.21	−22.19	−0.93	−0.86	4.03	**39.15**
Median	9.99	9.37	−0.002	−0.22	−22.19	−0.92	−0.88	4.11	**39.73**
STD	9.97	2.59	0.002	0.03	0.003	0.05	0.05	1.01	**3.28**
RMSE									
Best	0.01	0.01	0.04	0.04	0.48	0.04	0.04	0.02	**0.006**
Worst	0.03	0.03	0.04	0.04	0.56	0.05	0.05	0.03	**0.02**
Mean	0.02	0.02	0.04	0.04	0.54	0.05	0.05	0.03	**0.01**
Median	0.02	0.01	0.04	0.04	0.56	0.05	0.05	0.03	**0.01**
STD	0.005	0.005	0.0007	0.0009	0.03	0.0009	0.0008	0.003	**0.005**
CORR									
Best	0.86	0.81	−0.00003	−0.01	0.06	−0.06	−0.06	0.78	**0.86**
Worst	0.99	0.97	0.00008	−0.002	0.09	0.02	0.02	0.91	**0.99**
Mean	0.94	0.92	0.000007	−0.005	0.08	0.009	0.009	0.87	**0.95**
Median	0.96	0.94	−0.000002	−0.005	0.08	0.02	0.02	0.88	**0.96**
STD	0.04	0.05	0.00003	0.003	0.007	0.02	0.02	0.04	**0.04**
ESTOI									
Best	0.45	0.47	−0.003	0.11	0.33	0.24	0.20	0.45	**0.45**

Table 23 (continued)

Approach	multi-features + DCNN based Speech Enhancement [15]	D-EMCD[14]	NN + auto correlation	Spectral Subtraction [7]	OMLSA[6]	TSNR [24]	HRNR [25]	RNMF [9]	Proposed
Worst	0.86	0.84	0.003	0.27	0.73	0.69	0.69	0.78	**0.89**
Mean	0.65	0.69	0.0001	0.21	0.53	0.52	0.51	0.64	**0.69**
Median	0.64	0.72	0.0002	0.22	0.56	0.56	0.55	0.67	**0.72**
STD	0.14	0.13	0.001	0.05	0.12	0.14	0.14	0.11	**0.14**
STOI									
Best	0.69	0.69	1	0.37	0.55	0.49	0.50	0.67	**0.71**
Worst	0.95	0.91	1	0.54	0.83	0.82	0.82	0.86	**0.95**
Mean	0.84	0.83	1	0.48	0.69	0.69	0.69	0.77	**0.84**
Median	0.86	0.85	1	0.49	0.71	0.73	0.73	0.79	**0.85**
STD	0.08	0.08	0	0.058	0.08	0.10	0.09	0.07	**0.08**
CSED									
Best	4097.1	1303.6	6669.8	4136.2	4562.3	2940.4	2923.7	0	**1018.8**
Worst	4909.3	2406.2	10,500	7117.7	5203.8	4264.4	4264.5	4175	**2392.1**
Mean	4489.7	1758.7	8317.8	4722.7	4793.6	3227.3	3188.3	2201.5	**1592.1**
Median	4526.2	1748	8320.2	4543.4	4758.1	3124.3	3084	2906.8	**1540.2**
STD	255.52	303.52	1190.4	667.18	203.09	315.57	317.72	1706.7	**357.88**

6 Conclusion

In this modern world, there is a need to improve the speech signal, where the target speech signal is disturbed by different noise sources. This research considered the various noise problems for speech enhancement that is similar to real-world situations and many noise sources which simultaneously diminish the quality and intelligibility of the speech. In this work, a novel speech signal enhancement model was introduced with the assistance of a deep learning model. The main contribution of this research lies in the proper estimation of the tuning factor η of the Wiener filter for all input signals. The training of η was done using the *LSTM* model. The experimental outcomes at various input SNRs have verified the supremacy of the proposed model with respect to SDR, PESQ, SNR, RMSE, CORR, ESTOI, STOI, respectively. Especially, for airport noise, the PESQ of the proposed work is 2.17 at SNR = 0 dB, which is 4.9%, 10.43%, 75.06%, 74.7%, 40.78%, 40.78%, 34.95%, 36.325%, and 11.09% better than the existing multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF models. Additionally, in RMSE measure, the best value of the proposed model is 40%, 40%, 85%, 85%, 98.75%, 85%, 85%, and 70% superior to existing models multi-features+ DCNN based Speech Enhancement, D-EMCD, NN + auto correlation, spectral subtraction, OMLSA, TSNR, HRNR, and RNMF, respectively. Thus, the superiority of the proposed model has been proven with complex-noise environments. In the future, we consider the issue and develop the speech enhancement model with advanced GAN.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Jolicoeur-Martineau A (2018) The relativistic discriminator: a key element missing from standard GAN. arXiv preprint arXiv:1807.00734
2. Anita JS, Abinaya JS (2019) Impact of supervised classifier on speech emotion recognition. Multimedia Res 2(1):9–16
3. Arul VH, Sivakumar VG, Marimuthu R, Chakraborty B (2019) An approach for speech enhancement using deep convolutional neural network. Multimedia Res 2(1):37–44
4. NOIZEUS: https://ecs.utdallas.edu/loizou/speech/noizeus/ (Access Date: 2021-05-06)
5. Bekë K, Elezaj E, Millaku B, Dreshaj A, Hung NT (2021) The impact of COVID-19 (SARS-CoV-2) in tourism industry: evidence of Kosovo during Q1, Q2 and Q3 period of 2020. J Sustain Financ Invest:1–12
6. Cohen I, Berdugo B (2001) Speech enhancement for non-stationary noise environments. Sig Proc 81(11): 2403–2418
7. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Sig Proc 27(2):113–120
8. Chai L, Du J, Liu Q-F, Lee C-H (2021) A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 29:106–117
9. Chung H, Plourde E, Champagne B (2017) Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement. Speech Comm 87:18–30
10. Cuiv X, Chen Z, Yin F (2021) Multi-objective based multi-channel speech enhancement with BiLSTM network. Appl Acoust
11. Daniel M, Tan Z-H, Zhang S-X, Xu Y, Yu M, Yu D, Jensen J (2021) An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Trans Audio Speech Lang Process

12. Darekar RV, Dhande AP (2019) Emotion recognition from speech signals using DCNN with hybrid GA-GWO algorithm. Multimedia Res 2(4):12–22
13. Dionelis N, Brookes M (2018) Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering. IEEE/ACM trans Audio Speech Lang Process 26(5):937–950
14. Garg A (2020) Enhancement of speech signal using diminished empirical mean curve decomposition-based adaptive wiener filtering. *in comm*.
15. Garg A (2020) Deep convolutional neural network based speech signal enhancement using extensive speech features. *in comm*.
16. Gelderblom FB, Tronstad TV, Viggen EM (2019) Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement. IEEE/ACM trans Audio Speech Lang Process 27(3): 583–594
17. Hongjiang Y, Ouyang Z, Zhu WP, Champagne B, Ji Y (2019) A deep neural network based Kalman filter for time domain speech enhancement. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS), pp 1–5
18. Ishaan G, et al (2017) "Improved training of wasserstein gans." Advances in neural information processing systems vol 30
19. Kolbæk M, Tan Z, Jensen J (2019) On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement. IEEE/ACM Trans Audio speech Lang Process 27(2):283–295
20. Lavanya T, Nagarajan T, Vijayalakshmi P (2020) Multi-level Single-Channel speech enhancement using a unified framework for estimating magnitude and phase spectra. IEEE/ACM Trans Audio Speech Lang Process 28:1315–1327
21. Nicolson A, Paliwal KK (2018) Bidirectional long-short term memory network-based estimation of reliable spectral component locations. In: INTERSPEECH 1606-1610.
22. Pfeifenberger L, Zöhrer M, Pernkopf F (2019) Eigenvector-based speech mask estimation for Multi-Channel speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 27(12):2162–2172
23. Phillip I, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conf on comp vision and pattern recog* 1125–1134
24. Plapous C, Marro C, Mauuary L, Scalart P (2004) A two-step noise reduction technique. *2004 IEEE Int Conf Acoust, Speech, and Signal Process* 1:289–292
25. Plapous C, Marro C, Scalart P (2006) Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans ASLP 14:2098–2108
26. Reddy BG, Ofori M, Liu J, Ambati LS (2020) Early public outlook on the coronavirus disease (COVID-19): a soc med study
27. Sadeghi M, Leglaive S, Alameda-Pineda X, Girin L, Horaud R (2020) Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Trans Audio Speech Lang Process 28:1788–1800
28. Saleem N, Khattak MI, Al-Hasan M, Qazi AB (2020) On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks. IEEE Access 8:160581–160595
29. Saleem N, Khattak MI, Ochani MK (2021) Perceptually weighted β-order spectral amplitude Bayesian estimator for phase compensated speech enhancement. Appl Acoust 178:108007
30. Santiago P, Bonafonte A, Serra J (2017) SEGAN: Speech enhancement generative adversarial network *arXiv preprint arXiv*: 1703.09452
31. Sepp H, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
32. Shu X, Zhou Y, Liu H, Truong TK (2020) A human auditory perception loss function using modified bark spectral distortion for speech enhancement. Neural Process Lett 51(3):2945–2957
33. Sun X, Gao Z-F, Lu Z-Y, Li J, Yan Y (2020) A model compression method with matrix product operators for speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 28:2837–2847
34. Tayseer M, Adeel A, Hussain A (2018) A survey on techniques for enhancing speech. Int J Comput Appl 179(17):1–14
35. Triantafyllos A, Chung JS, Zisserman A (2018) The conversation: deep audio-visual speech enhancement. *arXiv preprint arXiv*:1804.04121
36. Venkateswarlu S, China K, Prasad S, Reddy AS (2011) Improve Speech Enhancement Using Weiner Filtering. Global J Comput Sci Technol
37. Wang Z, Zhang T, Ding B (2020) LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement. Appl Acoust 172:107647
38. Wood SUN, Stahl JKW, Mowlaee P (2019) Binaural codebook-based speech enhancement with atomic speech presence probability. IEEE/ACM Trans Audio Speech Lang Process 27(12):2150–2161
39. Xiang Y, Bao C (2020) A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network. IEEE/ACM trans Audio speech Lang Process 28:1826–1838

40. Xu L, Wei Z, Zaidi SFA, Ren B, Yang J (2021) Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain. *Appl Acoust* 174:107732
41. Yong X, Du J, Dai L-R, Lee C-H (2013) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68
42. Yong KH, Yoon JW, Cheon SJ, Kang WH, Kim NS (2021) A multi-resolution approach to GAN-based speech enhancement. Appl Sci 11(2):721
43. Yu H, Zhu W-P, Champagne B (2020) Speech enhancement using a DNN-augmented colored-noise Kalman filter. *Speech Comm* 125:142–151
44. Wang Y, Narayanan A, Wang D (2014) On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 22(12):1849–1858
45. Zhu Y, Xu X, Ye Z (2020) FLGCNN: a novel fully convolutional neural network for end-to-end monaural speech enhancement with utterance-based objective functions. Appl Acoust 170:107511
46. Zou X, Zhang X (2007) Speech enhancement using an MMSE short time DCT coefficients estimator with supergaussian speech modeling. J Electron (China) 24(3):332–337