Pattern Analysis and Applications (2020) 23:179–198
https://doi.org/10.1007/s10044-018-00768-x

THEORETICAL ADVANCES

# Enhancement of speech signal using diminished empirical mean curve decomposition-based adaptive Wiener filtering

Anil Garg[1] · O. P. Sahu[1]

Received: 28 July 2017 / Accepted: 26 December 2018 / Published online: 14 January 2019

**Abstract**

During the last few decades, speech signal enhancement has been one of the wide-spreading research topics. Numerous algorithms are being proposed to enhance the perceptibility and the quality of speech signal. These algorithms are often formulated to recover the clear signal from the signals that are ruined by noise. Usually, short-time Fourier transform and wavelet transform are widely used to process the speech signal. This paper attempts to overcome the regular drawbacks of the speech enhancement algorithms. As the frequency domain has good noise-removing ability, the short-time Fourier domain is also aimed to enhance the speech. Additionally, this paper introduces a decomposition model, named diminished empirical mean curve decomposition, to adaptively tune the Wiener filtering process and to accomplish effective speech enhancement. The performances of the proposed method and the conventional methods are compared, and it is observed that the proposed method is superior to the conventional methods.

**Keywords** Speech signal · Enhancement · STFT · D-EMCD · Wiener filtering

## 1 Introduction

Generally, speech enhancement implies the processing of noisy speech signals, so as to improve the signal perception through better decoding by systems or human beings [1–3]. A number of speech enhancement procedures are being formulated to recover the performance of a system, when the input given is a noise-ruined speech signal. Still, it is a tedious process to retain the denoised signal by reducing the noise. Hence, some limitations may be attained in the performance, compromising noise reduction and speech distortion [4–6]. Moreover, there are two categories of distorting speech signal based on medium to high SNR and low SNR. Under the first category, the objective is reducing the noise level to produce the natural signal. In contrast, in the second category, the objective is dropping the noise level, while preserving the intelligibility. Generally, the major factor that causes degradation in the speech's intelligibility and quality is the background noise. Further, the noise can be stationary or non-stationary and it is assumed as additive and uncorrelated with the speech signal [7]. More commonly, the entire speech enhancement approaches are intended at suppressing the background noise and they rely on one way or the other on the assessment of background noise. If the background noise gets modified at a rate that is much slower than the speech, that is, if the noise is more stationary than the speech, it is simple to assess the noise during the pauses in the speech.

More particularly, the speech enhancement approaches are broadly categorized as the temporal processing method and the spectral processing method. In case of the temporal processing method, the degraded speech is processed in time domain. On the contrary, the processing is achieved in the frequency domain for the spectral processing methods [8]. Spectral subtraction is one of the oldest procedures, which was proposed for reducing the background noise, and it is popular for its easy implementation and minimal complexity. The process of this technique is reducing or subtracting the average magnitude of the noise spectrum from the noisy speech spectrum. However, the estimation of the average magnitude of noise spectrum is carried out from the frames of speech absence. Mostly, in case of the stationary noise condition, initial frames are chosen for estimation. But, for the non-stationary noise condition, the noise estimation is formulated, whenever

✉ Anil Garg
anilgarg0778@gmail.com

[1] National Institute of Technology, Kurukshetra, Haryana, India

the characteristics of noise are changed. Therefore, the spectral subtraction algorithm becomes inefficient for corrupted speech with non-stationary noise [8–11].

Effective reduction of noise in the noisy speech signal allows the efficiency of speech-related applications to be improved [12, 13]. Various algorithms have been introduced currently to enhance the perceptibility and the quality of the speech signal. Those compensation methods are broadly classified into two, which include the multi-channel algorithms and the single-channel algorithms [14]. In most applications, the users are bound to the single-channel algorithm, since only one input channel is available. The statistical model-based techniques [15, 16] and spectral subtractions [17–19] are few modern single-channel algorithms, which usually use short-time Fourier transform (STFT) for processing the speech signal. The performance efficiency of the speech signal is improved, in case of the presence of little preservative noise. But, it is reversed, in case of the additive noise. Recently, wavelet transform (WT) is widely focused, when compared to STFT, because it uses large-sized windows at low frequency and small-sized windows at high frequency. It is different from STFT, since STFT uses the function of fixed window size. The variable size windows of WT result in low resolution and high resolution for high-frequency band and low-frequency band, respectively [20]. Thus, for all the frequency bands of speech signal, the time–frequency domain resolutions are highly improved. Mostly, high quantity of noisy speech is available in the real-time scenario [21–24]. Hence, the sub-band division methods effectively enhance the performance by making a better estimation of noise. Furthermore, WT works beneficially to build the approximation-based model from the estimated speech signal, even under adverse conditions [25].

*Contribution* In [26], a single-channel supervised speech enhancement algorithm on the basis of regularized NMF is implemented. In addition, a priori magnitude spectral distributions are modeled by the Gaussian mixtures. The work focuses on speech enhancement in the STFT domain. As the frequency domain is known for its noise removal ability, the adoption of the short-time Fourier domain further enhances the speech. A decomposition model, called the D-EMCD, is introduced here to remove the undesired signal. Further, the Wiener filtering process is adopted to accomplish speech enhancement and this paper is the improved version of [26]. This paper claims the following contributions in the speech enhancement method:

- An adaptive tuning factor is proposed to enhance the operation of Wiener filtering
- D-EMCD, which is a variant of EMCD, is proposed to decompose the signal, through which the tuning factor is defined.
- A sophisticated procedure is proposed to define the enhancement process, and so, the speech is enhanced under different noise conditions.

The proposed technique first estimates the noise spectrum and identifies the clean speech spectrum for achieving the unity tuning factor using Wiener filtering. The resultant signal is decomposed using D-EMCD. The bark frequency of the decomposed signal is determined, and then, it is used in the network. The network, in turn, predicts the tuning ratio. By using this tuning ratio, the second-stage Wiener filtering is carried out on the actual noisy signal. Subsequently, the resultant signal is decomposed for extracting the enhanced speech.

The rest of the paper is organized as follows: Sect. 2 reviews the literature work, and Sect. 3 describes the proposed speech enhancement algorithm. Moreover, Sect. 4 discusses the results and Sect. 5 concludes the paper.

## 2 Literature review

In 2017, Pejman et al. [27] have proposed an amplitude and phase estimator (ijMAP) and iterative joint maximum a posteriori (MAP) that assume a non-uniform phase distribution. The experimental outcomes proved the efficiency of the proposed method in improving both the phase and the amplitude of noise. The results were also justified using the instrumental measures like speech intelligibility, perceived quality and phase assessment error. Additionally, the approach enabled joint improvement in the perceived excellence. The speech intelligibility and the phase-blind joint MAP estimator exhibited comparable performance with the complex MMSE estimator.

In 2017, Sonay and Mohammad [28] have presented a novel unsupervised speech improvement method, projecting both the speech spectrogram and its temporal gradient as sparse. The sparse assumption was true because of the quasi-harmonic nature of the speech signals. In the approach, speech improvement was made by decreasing the suitable objective function, which was composed of a data fidelity term and a sparsity-imposing regularization term. Further, alternating direction scheme of multipliers (ADSM) was modified to determine the proposed methodology and a well-organized iterative procedure was established for carrying out the speech enhancement. Later, wide experiments showed that the proposed method outperformed the other competing schemes, in relation to varied performance assessment metrics.

In 2017, Hanwook et al. [26] have introduced a speech enhancement algorithm, named as the single-channel supervised speech enhancement algorithm. It was formulated on the basis of regularized nonnegative matrix factorization

(RNMF). The regularization in the NMF cost functions considered the log-likelihood functions of the spectrum of both the clean and the noisy speech signals, on the basis of the Gaussian mixture models. With the use of projected regularization as a priori information in the enhancement stage, the algebraic possessions of both the clean speech and the noise signals were exploited. The masking model of the human auditory system was also combined to improve the speech quality. Investigational upshots of source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ) and segmental signal-to-noise ratio (SNR) showed that their proposed speech enhancement algorithms offered improved performance in speech enhancement than the other benchmark algorithms.

In 2016, Ruwei et al. [29] have adopted a new filtering process, called improved least mean square adaptive filtering (ILMSAF). It was a speech enhancement algorithm with deep neural network (DNN) as well as noise classification. An adaptive coefficient of the filter's parameters was presented into the existing least mean square adaptive filtering algorithm (LMSAF). Initially, the authors have assessed the adaptive coefficient of the filter parameters using the deep belief network (DBN). Later, the enhanced speech was obtained by ILMSAF. Additionally, they presented a new classification method that was based on DNN to make the existing method as appropriate for several types of noise environments. In accordance with the consequence of noise classification, the ILMSAF model was nominated in the improvement process. The test results gave efficient results for the proposed model, under ITU-TG.160. Their method attained significant developments, in correspondence with varied subjective and objective quality measures of speech.

In 2016, Yanping et al. [30] have proposed a new procedure for the reduction of storage space and running time by utilizing low-rank estimate in a copying kernel Hilbert space, with tiny presentation loss in the enhanced speech. They also examined the root-mean-square error that was bound among the improved vectors, which were got by the approximation kernel matrix and the full kernel matrix. Further, it was observed that the method improved the speed of computation of the algorithm with the estimated presentation, while comparing with the full kernel matrix.

In 2016, Yang et al. [31] have developed the extension of gamma tone filter bank for speech enhancement by eliminating both the belongings of reverberation and noise through reinstating the appropriate amplitude and phase. Impartial and personal trials were carried out under numerous noisy reverberant circumstances to assess the delay efficiency of the proposed system. The signal-to-error ratio (SER), correlation, PESQ and SNR loss were also utilized in the objective assessments. The normalized mean preference score and the correctness in modified rhyme test (MRT) were utilized in the subjective evaluations. The results of all the estimations exposed that the proposed arrangement could effectively recover the quality and the intelligibility of speech signals under noisy reverberant situations.

In 2016, Sun et al. [32] have introduced a deep autoencoder (DAE) to represent the residual part, which was obtained by subtracting the valued fresh speech spectrum from the noisy speech spectrum. The enhanced speech signal was, therefore, found by transforming the valued clear speech spectrum back into the time domain. The overhead proposed method was known as separable deep autoencoder (SDAE). The under-determined nature of the above optimization problem was given, and the clear speech reconstruction was confined in the convex hull spanned by a pre-trained speech dictionary. New learning algorithms were investigated to value the nonnegativity of the parameters in the SDAE. Investigational results on TIMIT with 20 noise types, at various noise levels, demonstrated the dominance of the proposed technique over the conventional baselines.

In 2016, Chazan et al. [33] have presented a single-microphone speech enhancement algorithm. A hybrid approach was proposed by merging the generative mixture of Gaussians (MoG) model and the discriminative deep neural network (DNN). The proposed algorithm was executed in two phases, the training and the testing phases. First, the noise-free speech log power spectral density (PSD) was modeled as a MoG, representing the phoneme-based diversity in the speech signal. A DNN was then trained with the phoneme-labeled database of the clean speech signals for phoneme classification with mel-frequency cepstral coefficients (MFCC) as the input features. In the test phase, a noisy utterance of an untrained speech was processed. Lastly, they analyzed the contribution of all the components of the proposed process, indicating their combined importance.

In 2016, Wang et al. [34] have introduced the DWPT and NMF. Briefly, the DWPT remained chiefly practical in splitting a time-domain speech signal into a series of sub-band signals, without the introduction of any distortion. Then, they used NMF to emphasize the speech component for each sub-band. At last, the improved sub-band signals were combined through the inverse DWPT to rebuild a noise-abridged signal in the time domain. Further, they evaluated the proposed DWPT-NMF-based speech enhancement technique on the Mandarin hearing in noise test (MHINT) task. Investigational grades showed that this new way acted very well in encouraging the speech excellence and lucidity and outperformed the conservative STFT-NMF (Table 1).

## 3 Proposed speech enhancement algorithm

The architecture of the proposed speech enhancement algorithm is demonstrated in Fig. 1.

**Table 1** Features and challenges of speech enhancement processes

| References | Adopted methodology | Features | Challenges |
|---|---|---|---|
| Pejman et al. [27] | Maximum a posteriori | More accurate phase-aware amplitude estimate<br>Less speech distortion and more noise reduction | Prone to certain errors<br>Less consistent statistical model fitting |
| Sonay and Mohammad [28] | Iterative algorithm | Advanced SDR and PESQ<br>Higher SNR and lower LLR<br>Lower SD and higher short-time objective intelligibility (STOI) | Bi-level characteristics of the algorithm introduced computational complexity<br>Each phase of an iteration was rigid with no overlaps |
| Hanwook et al. [26] | Weiner filter | Enhanced speech quality<br>Less effect on noise | Occurrence of blurred results<br>Spatially invariant |
| Ruwei et al. [29] | Deep belief network | Enhancement of subjective and objective quality of speech<br>Suited for low SNR environments | Increased the complexity of testing or running time |
| Yanping et al. [30] | Kernel Hilbert space | High computational speed<br>Reduced the storage space | Quantified the function complexity |
| Yang et al. [31] | Kalman filtering | Improved the quality and lucidity of speech signals<br>Good performance under noisy reverberant conditions | High computational complexity<br>Not realistic in real-time situations |
| Sun et al. [32] | Deep learning neural network | Good performance under different noise conditions<br>Minimized the total reconstruction error | Required a large amount of data<br>Computationally expensive to train<br>Lack of theoretical foundation |
| Chazan et al. [33] | Neural network | Preserved the speech smoothness<br>Fast adaptation to noise<br>Required less formal statistical training | Greater computational burden<br>Proneness to overfitting |
| Wang et al. [34] | Discrete wavelet transform | Behaved very well in promoting the speech quality and intelligibility | Computationally intensive<br>The discrete wavelet transform was less efficient and natural |

Step 1: Let $S(n)$ be the clear signal. When the noise $N$ is added to the clear signal, it becomes a noisy signal $\bar{S}(n)$, which is given as the input to the NMF process. This results in two spectrums, namely noise spectrum $N^{\mathrm{s}}$ and signal spectrum $\bar{N}^{\mathrm{s}}$.

Step 2: The resultant spectrums, $N^s$ and $\bar{N}^{\mathrm{s}}$, are then filtered under Wiener filtering process, and this results in the filtered signal $\bar{S}_{\mathrm{f}}(n)$.

Step 3: $\bar{S}_{\mathrm{f}}(n)$ is then decomposed under the D-EMCD process, and this results in the bark frequency $b'\left(f'\right)$, which is utilized to train the NN classifier.

Step 4: The resultant 'tuned $\eta$' from the NN classifier and the spectrums, i.e., [$N^{\mathrm{s}}$ and $\bar{N}^{\mathrm{s}}$], are given as the inputs to the adaptive Wiener filtering process to filter the input signal $\bar{S}(n)$, resulting in the filtered signal $\overline{\overline{S_{\mathrm{f}}(n)}}$.

Step 5: Finally, the resultant $\overline{\overline{S_{\mathrm{f}}}}(n)$ is again decomposed using the D-EMCD process, and the decomposed signal is produced as the denoised signal $\overline{S_{\mathrm{D}}}(n)$.

The description of the adopted processes is as follows:

*NMF* The NMF is a dimensionality-lessening tool that decomposes the input signal into spectrums with nonnegative element constraints. The resultant spectrums are the noise spectrum and the signal spectrum.

*Wiener filter* The Wiener filter is a filter, which grants the assessment of the target random process with linear time-invariant (LTI) filtering of the additional noise. This filter reduces the mean square error among the assessed random process and the desired process.

*D-EMCD* The EMCD—empirical mean curve decomposition—decomposes a signal by smoothening its peaks. First, the maximal and the minimal points from the signals are extracted. Then, they are interpolated and the average is taken. The average signal is subtracted from the original signal to find the residue. The residue and the average signals are checked for their similarity with the original signal. Till it is smoothened, the process is repeated. In D-EMCD, the iteration is diminished and so, the first average signal is used for the further steps because the speech signal requires no loss on maxima and minima.

The D-EMCD is a signal decomposition process that has the similar process of EMCD. The only difference is that the D-EMCD does the decomposition without any iteration, whereas the existing EMCD is an iterative process.

*NN* This is a machine learning approach inspired by the brain's performance. The NN organization is associated with the learning algorithm, which is used for training purposes.

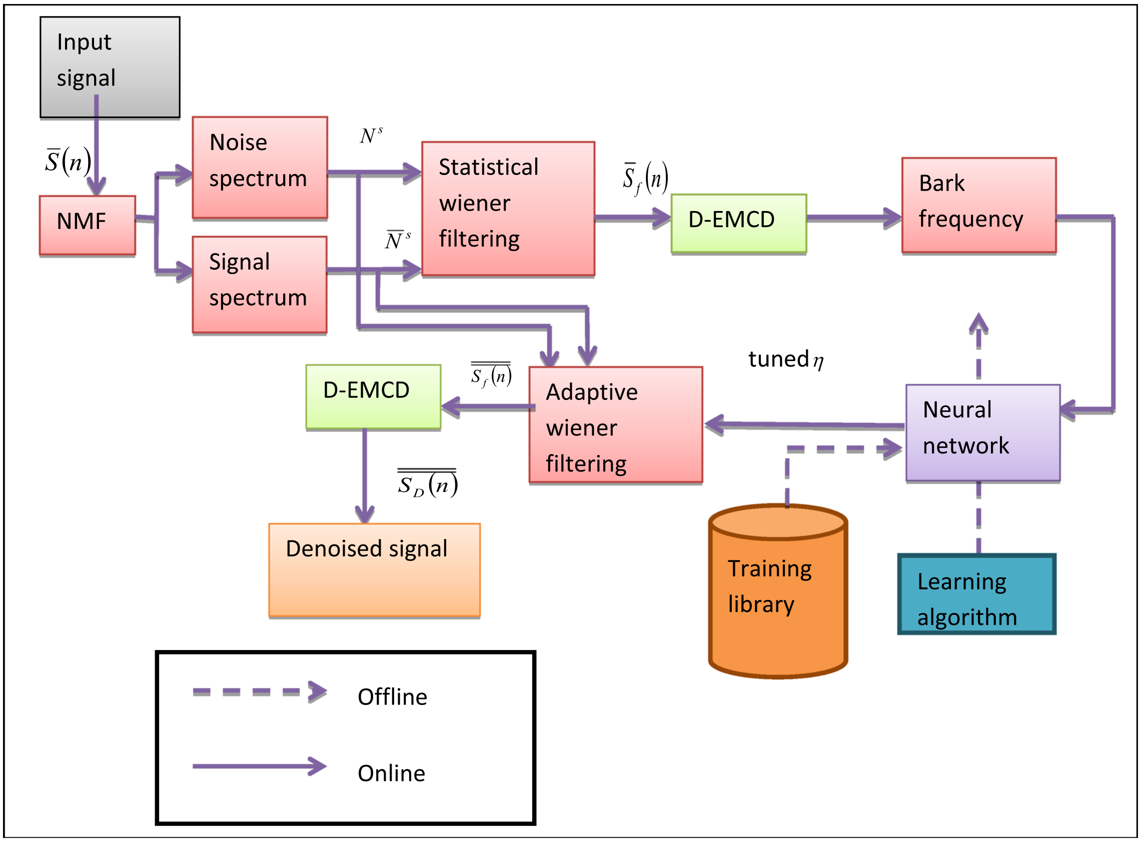


**Fig. 1** Proposed architecture for speech enhancement

*Adaptive Wiener filter* This is a filter that processes with the concept of Wiener filter, but the fact is that the filter process also incurs tuned $\eta$, which is the output of NN.

*Training library* The training library of NN is constructed by giving the known inputs (bark frequency) and its target $\eta$. With the knowledge of this, the unknown values are formulated.

*Offline and online process* The training process is considered as the offline process, and the testing process is termed as the online process, in which the testing is carried out on the trained system. Offline process means identifying appropriate tuning factor for different noise variances and training the neural network [35]. Online process implies the actual enhancement process, where the trained network is used for determining the tuning factor.

*Learning algorithm* In this work, the NN approach is trained using the Levenberg–Marquardt algorithm [36].

### 3.1 Noise estimation using STFT-minimum statistics

The minima are tracked from the noisy signal by the noise power spectral density estimator, which is based on minimum statistics.

$$W(\alpha,b) = \lambda(\alpha,b)W(\alpha-1,b) + (1-\lambda(\alpha,b))|W(\alpha,b)|^2 \tag{1}$$

where $W(\alpha,b)$ denotes the STFT coefficient of the frame $\alpha$, $b$ represents the frequency bin and $\lambda(\alpha,b)$ denotes the frequency- and time-dependent smoothing parameters. A bias compensation factor is applied to observe the mean power. Moreover, $F_{\min}$ represents the bias compensation factor, which defines the function of the length of minimum search interval and $\mathrm{var}\{W(\alpha,b)\}$ denotes the variance estimator of the smoothened power spectral density. The variance of $W(\alpha,b)$ is estimated, while fixing the search interval length for the algorithm. The variance estimator for frequency bin $b$ at $\alpha$ frame is defined as:

$$\hat{\mathrm{var}}\{W(\alpha,b)\} = \overline{W^2}(\alpha,b) - \bar{W}^2(\alpha,b) \tag{2}$$

where $\bar{W}(\alpha,b)$ and $\overline{W^2}(\alpha,b)$ denote the mean smoothened periodograms and a first-order recursive average of smoothened periodograms, respectively.

In this paper, we describe about the short-time Fourier transform (STFT)-based noise estimation. Figure 2 illustrates the noise power spectrum of the actual signal, the noise estimated signal by FFT and the noise estimated

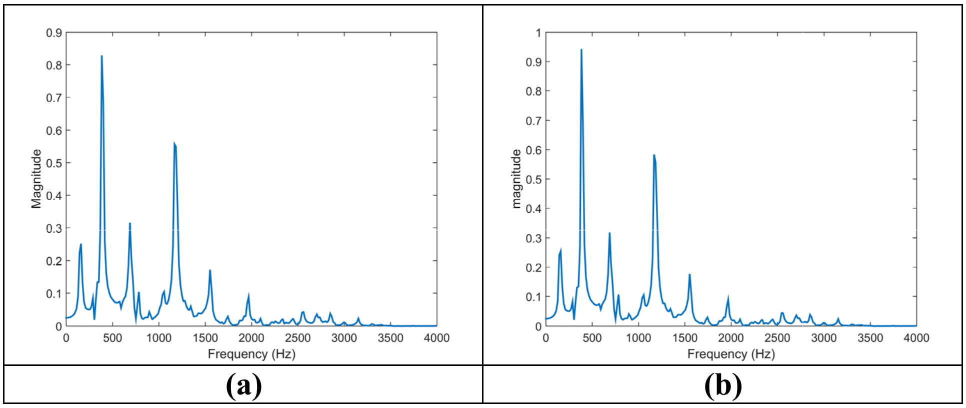


**Fig. 2** Noise power spectrum **a** estimated by FFT and **b** estimated by STFT—minimum statistics

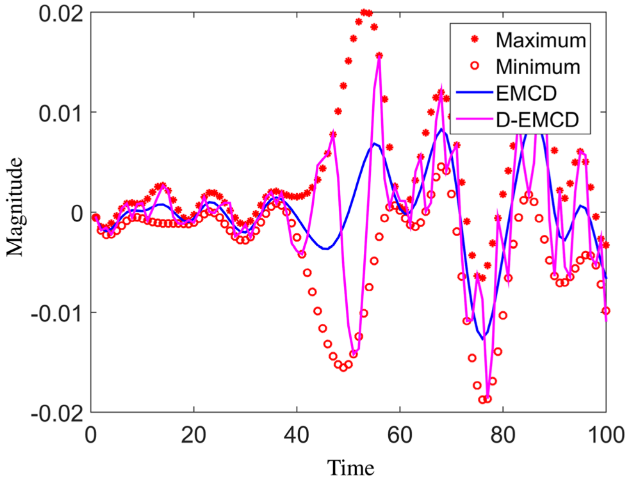


**Fig. 3** EMCD versus D-EMCD: information-preserving characteristics of D-EMCD

signal by STFT. Basically, STFT is used to determine the phase content and the sine wave frequency of a signal that changes over time. Practically, we can say that the longer time signals are divided into equal shorter length segments and the Fourier transform is applied separately on each segment. Moreover, the STFT can also be interpreted as a filtering operation. More particularly, there are two properties which satisfy the estimation strategy (i.e., shift invariance property that is based on magnitude and the properties of linear time–frequency distribution).

In Fig. 2, the power spectrum of the noisy speech, which is obtained from FFT and STFT-minimum statistics, is presented. There is a significant difference between them that the magnitude of the frequency component is presented well. Figure 3 actually describes that the D-EMCD decomposed signal correlates with the actual signal, but neglects the undesired spikes and surges. However, the EMCD signal loses huge information from the actual signal. For the reference, the maximum and the minimum peaks are also presented.

## 3.2 Adaptive Wiener tuning ratio

The role of tuning ratio is highly substantiated in [26]. This paper proposes neural network (NN) to estimate the tuning ratio, based on the bark frequency $b'\left(f'\right)$ of the NMF-based filtered D-EMCD signal, i.e., $\bar{S}_{\mathrm{D}}(n)$. The logical expression of the mapping function from $f'$ frequency to the bark frequency is given as:

$$b'\left(f'\right) = 13 \arctan\left(0.76f'\right) + 3.5 \arctan\left[\left(0.33f'\right)^2\right] \tag{3}$$

where $f'$ is the frequency of the $\bar{S}_{\mathrm{D}}(n)$. The basis function $a'_j$ is formulated, as defined in Eq. (4):

$$a'_j = \left(W^I_j b'\left(f'\right)\right) + W^0_j; \quad j = 1, \dots N_{\mathrm{h}'} \tag{4}$$

where $W^{\mathrm{I}}$ represents the weight between the input and the $j$th hidden neuron,$N_{\mathrm{h}'}$ denotes the number of hidden neurons in the NN network and $W^0_j$ is the weight of the $j$th bias neuron. Consequently, the activation function $\hat{a}'_j$ is formulated for limiting the amplitude, as represented in Eq. (5).

$$\hat{a}'_j = \phi\left(a_j\right) = \frac{1}{1 + \exp\left(-a'_j\right)} \tag{5}$$

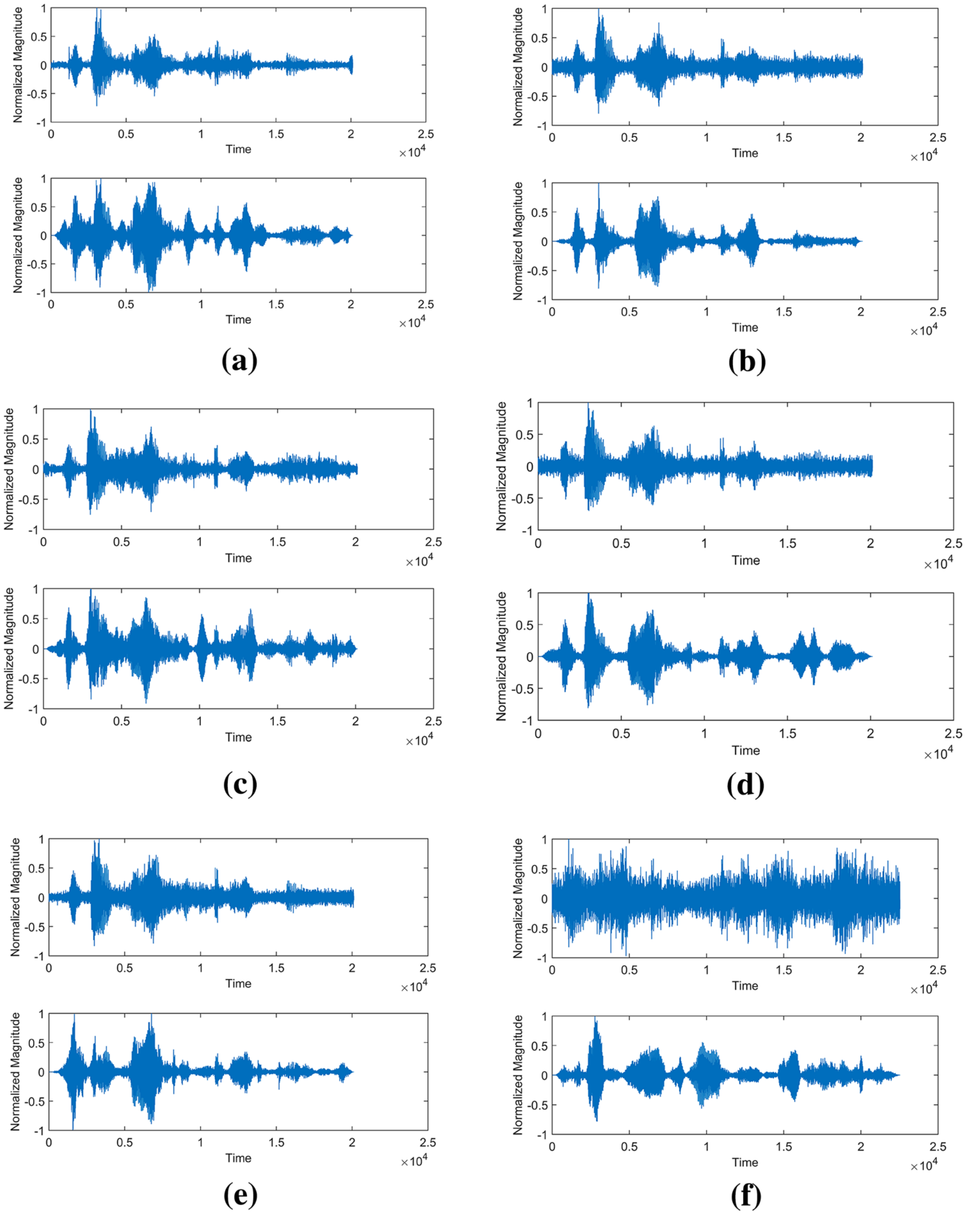
The network output $\eta$ is defined as given in Eq. (6), where $W^{\mathrm{H}}$ denotes the weight between the $j$th hidden neuron and the output neuron and $W^{\mathrm{H}_0}$ represents the weight of the bias.

$$\eta = \sum_{j=1}^{N_{\mathrm{h}'}} W^{\mathrm{H}}_j \hat{a}'_j + W^{\mathrm{H}_0} \tag{6}$$

## 3.3 Spectrum estimation using NMF

For speech signal enhancement, the noisy signal $\bar{S}(n)$ is voiced in time–frequency $(\alpha, b)$ domain through STFT, as given in Eq. (7).

$$\bar{S}(b, \alpha) = S(b, \alpha) + N(b, \alpha) \tag{7}$$



(a)

(b)

(c)

(d)

(e)

(f)

**Fig. 4** Temporal analysis of denoising performance: noisy and denoised signal of various noise types: **a** airport noise, **b** exhibition noise, **c** restaurant noise, **d** station noise, **e** street noise and **f** babble noise

where $S(b,\alpha)$,$\bar{S}(b,\alpha)$,$N(b,\alpha)$ present the STFT of the clear speech, noisy speech and noise, respectively, for the $b$th frequency bin of $\alpha$ frame. The approximation of the noisy speech's magnitude spectrum is defined as $|\bar{S}(b,\alpha)| = |S(b,\alpha) + |N(b,\alpha)|$. This is the widely used assumption in the processing of NMF-based speech and audio signals.

The magnitude spectrum matrices of varied signals are denoted as:

$$V' = \left[v'_{b\alpha}\right] \in R_{+}^{B\times T} \tag{8}$$

where $v'_{b\alpha}$ represents the magnitude spectral value for the $b$th bin of $\alpha$ frame, whereas $B$ and $T$ denote the number of frequency bins and time frames, respectively.

Generally, NMF-based speech enhancement processes are comprised of two stages, namely the training stage and the enhancement stage. In the training stage, Eq. (9) is separately applied to the training data $V'_{\mathrm{S}} \in R_{+}^{B\times T_{\mathrm{S}}}$ and $V'_{\mathrm{N}} \in R_{+}^{B\times T_{\mathrm{N}}}$, and this results in the basis matrices of both clear speech and noise, $W'_{\mathrm{S}} = \left[w'^{\mathrm{S}}_{Bm'}\right] \in R_{+}^{B\times M'_{\mathrm{S}}}$ and $W'_{\mathrm{N}} = \left[w'^{\mathrm{N}}_{Bm'}\right] \in R_{+}^{B\times M'_{\mathrm{N}}}$, respectively. Here, $M'$ represents the number of basis vectors:

$$\begin{aligned} W' &\leftarrow W' \otimes \frac{(V'/W'H')H'}{\Psi H'} \\ H' &\leftarrow H' \otimes \frac{W'(V'/W'H')}{W^{T'}\Psi} \end{aligned} \tag{9}$$

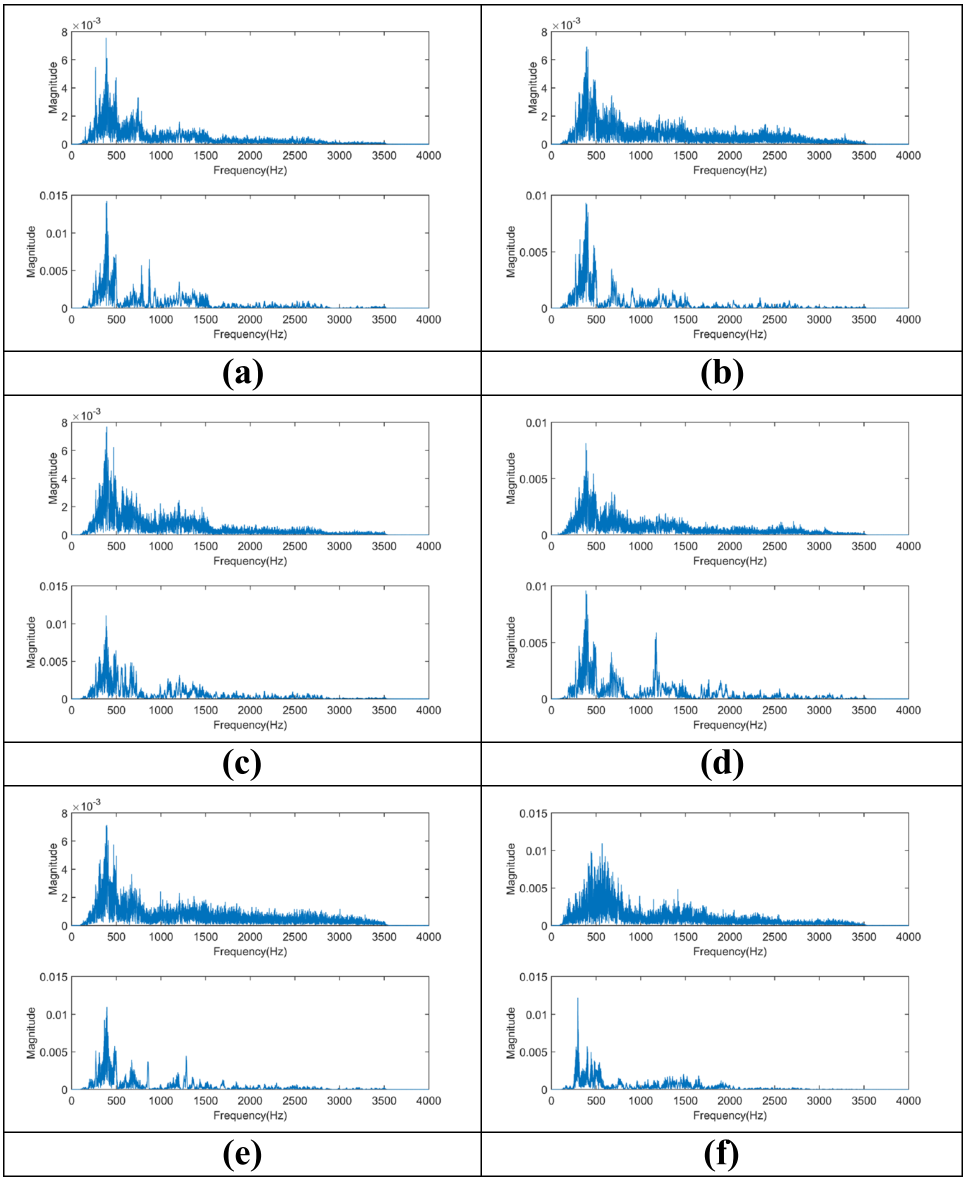


**Fig. 5** Spectral analysis of denoising performance: noisy and denoised signal of various noise types: **a** airport noise, **b** exhibition noise, **c** restaurant noise, **d** station noise, **e** street noise and **f** babble noise

where $\Psi$ is a $B \times T$ matrix with entries equal to one and $T'$ represents the matrix transpose.

In the enhancement stage, the basis matrices are fixed as $W'_{\hat{S}} = \left[W'_{S} W'_{N}\right] \in R_{+}^{B\times(M'_{S}+M'_{N})}$ and the estimation of activation matrix $H'_{\bar{S}} = \left[H'^{T'}_{S} H'^{T'}_{N}\right]^{T'} \in R_{+}^{(M'_{S}+M'_{N})\times T_{\bar{S}}}$ of noisy speech is done by applying the NMF activation update on $V'_{\bar{S}} \in R_{+}^{B\times T_{\bar{S}}}$. After getting the activation matrix of the speech signal, the estimation of clear speech spectrum is done with the aid of the Wiener filter (WF), as given in Eq. (10):

$$S' = \frac{P'_{S}}{P'_{S} + P'_{N}} \otimes \bar{S} \tag{10}$$

where $P'_{S} = \left[P'_{S}(b,\alpha)\right]$ and $P'_{N} = \left[P'_{N}(b,\alpha)\right] \in R_{+}^{B\times T_{\bar{S}}}$ represent the estimated power spectral density (PSD) matrices of the clear speech and the noise, respectively. The latter is obtained through the temporal smoothing of periodograms, as defined in Eqs. (11) and (12), respectively:

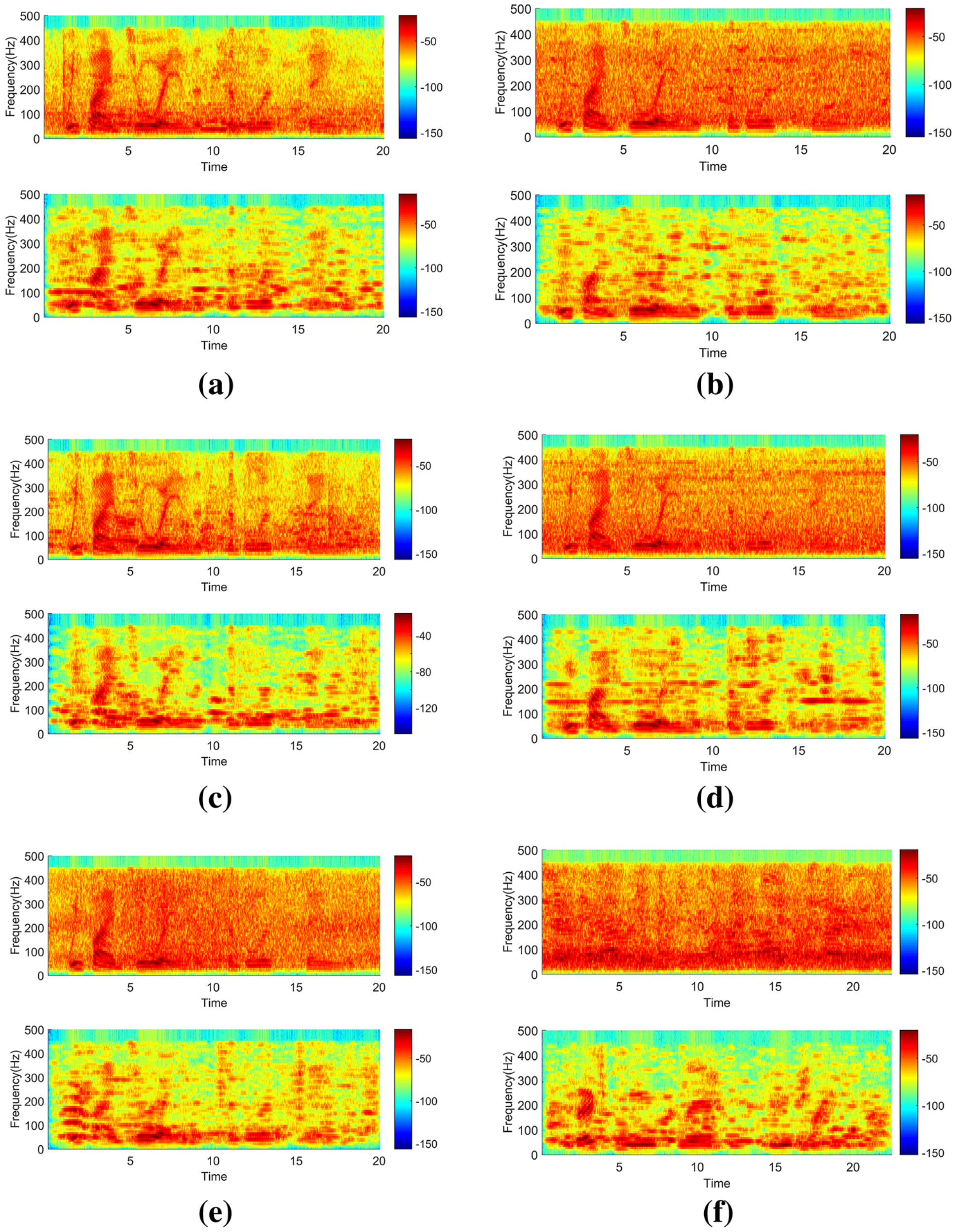
$$P'_{S}(b,\alpha) = \tau_{S} P'_{S}(b,\alpha-1) + \left(1-\tau_{S}\right)\left(\left[W'_{S}H'_{S}\right]_{b\alpha}\right)^{2} \tag{11}$$

$$P'_{N}(b,\alpha) = \tau_{N} P'_{N}(b,\alpha-1) + \left(1-\tau_{N}\right)\left(\left[W'_{N}H'_{N}\right]_{b\alpha}\right)^{2} \tag{12}$$

where $\tau_{S}$ and $\tau_{N}$ denote the temporal smoothing factors for speech and noise, respectively.

### 3.4 D-EMCD-based Wiener filtering

The Wiener filtering process is based on the proposed D-EMCD decomposition process. It is an iterative



**(a)** **(b)** **(c)** **(d)** **(e)** **(f)**

**Fig. 6** Time–frequency analysis of denoising performance: noisy and denoised signal of various noise types: **a** airport noise, **b** exhibition noise, **c** restaurant noise, **d** station noise, **e** street noise and **f** babble noise

decomposition process, and the initial step is the extraction of both the minima and the maxima. Figure 3 illustrates the information-preserving characteristics of D-EMCD. Let $\bar{S}^{\max}(n) : \{(P_i, S(P_i)), i = 1, \dots N_{\max}\}$ be the maxima signal of $\bar{S}(n)$ with $N$-element signal, where $P_i$ denotes the time index and $N_{\max}$ represents the number of maxima. Let the minima signal of actual signal $\bar{S}(n)$ be $\bar{S}^{\min}(n) : \{(Q_i, S(Q_i)), i = 1, \dots N_{\min}\}$, where $Q_i$ is the time index and $N_{\min}$ denotes the number of minima.

Furthermore, B-spline interpolation is used to interpolate both the maxima and the minima signal and they are defined below:

$$\bar{S}^{I-\max}(n) = B\{(P_i, y(P_i)), \bar{S}^{\max}(n)\}; \quad n = 1, \dots N \tag{13}$$

$$\bar{S}^{I-\min}(n) = B\{(Q_i, y(Q_i)), \bar{S}^{\min}(n)\} \quad n = 1, \dots N \tag{14}$$

$\widehat{\delta}_k(n)$ is defined as follows:

$$\widehat{\delta}_k(n) = \min|\bar{S}_k(n) - \bar{S}_k^{I-\max}(n)| \tag{15}$$

Moreover, the minimum and the maximum of $\widehat{\delta}(n)$ are also defined as given in Eqs. (16) and (17). Similarly,$\widehat{S}_k(n)$ and $\bar{S}_k^{\max-\min}$ are also represented as shown in Eqs. (18) and

**Table 2** Airport noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Airport noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | −10.4180 | 0.8024 | 26.8415 | 2.2741 | 0.1806 | 0.02439 | 0.5310 | 17,265.0321 |
| ILMSAF [29] | −43.1867 | 0.542019 | 28.2433 | 0.04294 | −0.0000027 | 0.000795 | 1 | 7169.128 |
| Vuvuzela [37] | −8.80113 | 0.549925 | 28.1619 | 0.043961 | −0.00318 | 0.137855 | 0.3948 | 4923.255 |
| OMLSA [38] | −23.1074 | 1.287055 | 5.5178 | 0.555896 | 0.069244 | 0.359471 | 0.578349 | 4930.989 |
| TSNR [39] | −7.40889 | 1.413942 | 27.0352 | 0.04818 | 0.011206 | 0.343843 | 0.553239 | 3410.615 |
| HRNR [40] | −7.42242 | 1.38407 | 27.1862 | 0.047539 | 0.011357 | 0.315499 | 0.557241 | 3352.881 |
| RNMF [26] | 5.824237 | 1.93443 | 30.9863 | 0.031407 | 0.803629 | 0.476359 | 0.675044 | 0 |
| Prop | 4.830108 | **1.946733** | **34.0685** | **0.023947** | **0.833765** | **0.509891** | **0.715761** | **1870.349** |
| *Airport noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | −10.5173 | 0.8024 | 27.1485 | 2.2607 | 0.1776 | 0.02332 | 0.5303 | 17,225.476 |
| ILMSAF [29] | −43.245 | 0.534249 | 28.2452 | 0.042914 | −0.0000043 | 0.000566 | 1 | 7952.943 |
| Vuvuzela [37] | −7.70199 | 0.818841 | 28.1042 | 0.04383 | −0.00492 | 0.197686 | 0.466444 | 4530.937 |
| OMLSA [38] | −22.6598 | 1.319643 | 5.5211 | 0.555559 | 0.075779 | 0.483455 | 0.671198 | 4754.392 |
| TSNR [39] | −6.77054 | 1.858997 | 26.9794 | 0.047592 | 0.018492 | 0.459935 | 0.659213 | 3215.675 |
| HRNR [40] | −6.81255 | 1.847263 | 27.0671 | 0.047161 | 0.018756 | 0.4399 | 0.664296 | 3161.433 |
| RNMF [26] | 8.227952 | 2.313639 | 32.1356 | 0.028369 | 0.867131 | 0.610264 | 0.764791 | 2999.043 |
| Prop | **9.257039** | **2.361946** | **36.9993** | **0.016827** | **0.924949** | **0.644831** | **0.80974** | **1606.895** |
| *Airport noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | −10.3620 | 0.7985 | 26.9106 | 2.2302 | 0.1826 | 0.0255 | 0.5291 | 17,126.674 |
| ILMSAF [29] | −45.513 | 0.451885 | 28.2478 | 0.042901 | 0.0000328 | −0.000062 | 1 | 9049.059 |
| Vuvuzela [37] | −7.45629 | 0.986987 | 28.0929 | 0.043998 | −0.00225 | 0.250451 | 0.522101 | 4275.661 |
| OMLSA [38] | −22.4185 | 1.46885 | 5.52273 | 0.555438 | 0.079178 | 0.594816 | 0.738666 | 4621.661 |
| TSNR [39] | −6.69568 | 2.280759 | 26.9317 | 0.047647 | 0.021737 | 0.594022 | 0.750943 | 3032.308 |
| HRNR [40] | −6.7336 | 2.30853 | 26.9818 | 0.047317 | 0.021952 | 0.58603 | 0.754258 | 3016.927 |
| RNMF [26] | 9.362473 | 2.534341 | 33.3167 | 0.02575 | 0.889264 | 0.705761 | 0.817231 | 2761.073 |
| Prop | **12.45125** | **2.690384** | **40.5341** | **0.013011** | **0.957877** | **0.771515** | **0.878077** | **1392.467** |
| *Airport noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | −10.3510 | 0.79992 | 26.5577 | 2.20896 | 0.1833 | 0.02876 | 0.52820 | 17,057.300 |
| ILMSAF [29] | −47.0965 | 0.637249 | 28.2483 | 0.042896 | 0.000017 | 0.000061 | 1 | 10,306.93 |
| Vuvuzela [37] | −7.49411 | 1.071891 | 28.0770 | 0.044129 | −0.00439 | 0.271151 | 0.53806 | 4136.205 |
| OMLSA [38] | −22.2604 | 1.665314 | 5.52358 | 0.55539 | 0.081069 | 0.662475 | 0.773733 | 4562.333 |
| TSNR [39] | −6.751 | 2.753715 | 26.9760 | 0.047724 | 0.02114 | 0.688655 | 0.802001 | 2940.364 |
| HRNR [40] | −6.77787 | 2.796623 | 27.01398 | 0.047493 | 0.021105 | 0.679954 | 0.798071 | 2928.979 |
| RNMF [26] | 9.953762 | 2.722153 | 33.71583 | 0.024022 | 0.90362 | 0.755009 | 0.83849 | 0 |
| Prop | **14.20475** | **2.9743** | **42.5818** | **0.011197** | **0.96908** | **0.836507** | **0.90994** | **1303.649** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

(19), respectively. The resultant signal from the filtering process is the denoised signal.

$$\bar{\delta}_k^{\max}(n) = |\bar{S}_k(n) - \bar{S}_k^{I-\max}(n)| \tag{16}$$

$$\bar{\delta}_k^{\min}(n) = |\bar{S}_k(n) - \bar{S}_k^{I-\min}(n)| \tag{17}$$

$$\widehat{S}_k(n) = \begin{cases} \bar{S}_k(n); & \text{if } \widehat{\delta}_k(n) > \delta_T \\ \bar{S}_k^{\max-\min}(n); & \text{otherwise} \end{cases} \tag{18}$$

$$\bar{S}_k^{\max-\min}(n) = \left\{ \begin{array}{ll} \bar{S}_k^{I-\max}(n); & \text{if } \bar{\delta}_k^{\max}(n) < \bar{\delta}_k^{\min}(n) \\ \bar{S}_k^{I-\min}(n); & \text{otherwise} \end{array} \right\} \tag{19}$$

## 4 Results and discussion

### 4.1 Dataset and experiments

The speech signal enhancement experimentation is conducted using MATLAB 2015a. The database including the speech signals is downloaded from the URL http://ecs.utdallas.edu/loizou/speech/noizeus/. The LRA [30] and ILMSAF [29] are the public databases, whereas the Vuvuzela [37], OMLSA [38], TSNR [39], HRNR [40] and RNMF [26] are the private databases. The experimentation is carried out

**Table 3** Exhibition noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Exhibition noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | –10.3437 | 0.8099 | 26.7835 | 2.2973 | 0.1841 | 0.02654 | 0.5324 | 17,315.887 |
| ILMSAF [29] | –41.6579 | 0.425503 | 28.2423 | 0.042925 | –0.000029 | –0.000091 | 1 | 6885.197 |
| Vuvuzela [37] | –9.7762 | 0.433939 | 28.1086 | 0.044531 | –0.0056 | 0.153763 | 0.397682 | 4998.421 |
| OMLSA [38] | –23.2058 | 1.153355 | 5.51202 | 0.555904 | 0.068713 | 0.430344 | 0.607923 | 5203.767 |
| TSNR [39] | –7.3184 | 1.243383 | 26.8617 | 0.048168 | 0.01497 | 0.432529 | 0.612161 | 3570.904 |
| HRNR [40] | –7.35267 | 1.21477 | 26.9480 | 0.047504 | 0.014563 | 0.421413 | 0.616776 | 3523.669 |
| RNMF [26] | 6.22385 | 1.853652 | 31.41849 | 0.031333 | 0.794168 | 0.528068 | 0.691038 | 4126.974 |
| Prop | **5.259774** | **1.884774** | **32.856** | **0.024258** | **0.831848** | **0.556889** | **0.727769** | **2406.215** |
| *Exhibition noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | –10.4051 | 0.7938 | 26.6512 | 2.19833 | 0.1818 | 0.02693 | 0.5263 | 17,014.009 |
| ILMSAF [29] | –46.4092 | 0.604895 | 28.2467 | 0.042905 | 0.0000145 | 0.000222 | 1 | 7163.531 |
| Vuvuzela [37] | –8.63867 | 0.715995 | 28.0074 | 0.044112 | –0.00404 | 0.206777 | 0.463506 | 4555.987 |
| OMLSA [38] | –22.6404 | 1.344596 | 5.52152 | 0.555568 | 0.075976 | 0.530204 | 0.685579 | 4956.877 |
| TSNR [39] | –6.90206 | 1.789926 | 26.9364 | 0.047941 | 0.019204 | 0.542492 | 0.701599 | 3218.206 |
| HRNR [40] | –6.93887 | 1.795809 | 26.9788 | 0.047558 | 0.018696 | 0.544629 | 0.709576 | 3178.957 |
| RNMF [26] | 8.688553 | 2.192665 | 32.5005 | 0.027868 | 0.86214 | 0.63504 | 0.772684 | 3873.355 |
| Prop | **9.365253** | **2.245781** | **37.4415** | **0.017025** | **0.92303** | **0.676195** | **0.819646** | **2048.487** |
| *Exhibition noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | –10.3808 | 0.79589 | 26.6526 | 2.2156 | 0.1828 | 0.0251 | 0.5276 | 17,090.679 |
| ILMSAF [29] | –47.8363 | 0.394362 | 28.2481 | 0.042898 | –0.000014 | 0.000995 | 1 | 8363.914 |
| Vuvuzela [37] | –7.66609 | 0.921928 | 28.07225 | 0.044039 | –0.0057 | 0.245008 | 0.506356 | 4502.93 |
| OMLSA [38] | –22.3901 | 1.491871 | 5.52280 | 0.555434 | 0.079361 | 0.606982 | 0.737645 | 4761.913 |
| TSNR [39] | –6.69004 | 2.250223 | 26.9636 | 0.047792 | 0.020464 | 0.619061 | 0.760259 | 3168.486 |
| HRNR [40] | –6.73292 | 2.262972 | 26.9889 | 0.047516 | 0.02038 | 0.619909 | 0.764679 | 3094.55 |
| RNMF [26] | 10.2548 | 2.480936 | 33.7476 | 0.024995 | 0.889274 | 0.719885 | 0.825086 | 0 |
| Prop | **12.54993** | **2.578956** | **40.61842** | **0.013242** | **0.955032** | **0.770816** | **0.876053** | **1812.882** |
| *Exhibition noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | –10.3828 | 0.7962 | 26.67419 | 2.2242 | 0.1817 | 0.0258 | 0.5285 | 17,121.4615 |
| ILMSAF [29] | –47.882 | 0.453097 | 28.2485 | 0.042897 | 0.000023 | 0.000552 | 1 | 9176.933 |
| Vuvuzela [37] | –7.51761 | 1.03181 | 28.0735 | 0.044154 | –0.00476 | 0.271344 | 0.536531 | 4272.38 |
| OMLSA [38] | –22.2591 | 1.617731 | 5.5235 | 0.555388 | 0.081044 | 0.672514 | 0.7742 | 4637.631 |
| TSNR [39] | –6.72701 | 2.652039 | 26.9625 | 0.047795 | 0.020428 | 0.688293 | 0.801937 | 3036.825 |
| HRNR [40] | –6.75953 | 2.666548 | 26.9868 | 0.047597 | 0.020457 | 0.687408 | 0.802052 | 2974.568 |
| RNMF [26] | 11.13144 | 2.718365 | 34.4738 | 0.022931 | 0.905955 | 0.784248 | 0.857782 | 3443.665 |
| Prop | **15.02555** | **2.880454** | **43.7957** | **0.010931** | **0.970195** | **0.84481** | **0.915692** | **1666.868** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

on about 30 speech signals. The number of hidden units is 10. Six noise types, namely airport noise, exhibition noise, restaurant noise, station noise, street noise and babble noise, are added to the speech signals. In addition, the investigation is carried out with different SNR dB levels, which include 0 dB, 5 dB, 10 dB and 15 dB.

The speech data are subjected to NMF decomposition, which estimates the signal spectrum as well as the noise spectrum of similar length. The decomposition is performed at different noise levels, so that diverse decomposition effect can be obtained via NMF. The Wiener filtering is applied on the decomposed signal of dimension $513 \times 86$, followed by D-EMCD. The resultant signal is subjected to feature extraction using bark frequency, and hence, the training library is constructed in the dimension of 1x30. The training data are obtained for different speech qualities, and the respective

**Table 4** Restaurant noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Restaurant noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | −10.411 | 0.8014 | 26.8545 | 2.272 | 0.1806 | 0.02190 | 0.5306 | 17,267.3618 |
| ILMSAF [29] | −44.8462 | 0.472523 | 28.2273 | 0.042936 | −0.000012 | 0.000107 | 1 | 6974.248 |
| Vuvuzela [37] | −9.14382 | 0.551319 | 28.1195 | 0.043784 | −0.00264 | 0.134361 | 0.370427 | 4852.803 |
| OMLSA [38] | −23.0271 | 1.367484 | 5.509799 | 0.556031 | 0.070911 | 0.372362 | 0.581629 | 4977.075 |
| TSNR [39] | −7.69256 | 1.374338 | 26.90630 | 0.048442 | 0.00607 | 0.337645 | 0.544972 | 3560.089 |
| eSHRNR [40] | −7.70968 | 1.350633 | 27.030 | 0.047764 | 0.005206 | 0.318896 | 0.553293 | 3537.96 |
| RNMF [26] | **5.194129** | 1.849504 | 30.97360 | 0.032143 | 0.783563 | 0.487437 | 0.661691 | 3378.811 |
| Prop | 3.76254 | **1.92735** | **32.02981** | **0.025676** | **0.809011** | **0.509527** | **0.699553** | **1964.05** |
| *Restaurant noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | −10.4433 | 0.79964 | 26.6245 | 2.2156 | 0.1796 | 0.0257 | 0.5276 | 17,096.7325 |
| ILMSAF [29] | −41.9114 | 0.547673 | 28.24716 | 0.042908 | −0.0000081 | −0.00016 | 1 | 8276.524 |
| Vuvuzela [37] | −7.90071 | 0.800465 | 28.10527 | 0.043901 | −0.00466 | 0.196542 | 0.458998 | 4413.882 |
| OMLSA [38] | −22.5559 | 1.306641 | 5.520834 | 0.555606 | 0.077011 | 0.486875 | 0.671337 | 4773.834 |
| TSNR [39] | −7.05857 | 1.841138 | 26.93318 | 0.048077 | 0.016885 | 0.488268 | 0.675045 | 3129.462 |
| HRNR [40] | −7.08685 | 1.84633 | 26.99694 | 0.047649 | 0.016611 | 0.477135 | 0.679056 | 3100.808 |
| RNMF [26] | 7.956189 | 2.187008 | 31.806648 | 0.028782 | 0.855729 | 0.617525 | 0.764193 | 0 |
| Prop | **8.698332** | **2.268126** | **36.07842** | **0.017556** | **0.915883** | **0.65133** | **0.809497** | **1729.632** |
| *Restaurant noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | −10.4429 | 0.7995 | 26.5665 | 2.20698 | 0.17990 | 0.0247 | 0.5279 | 17,066.5855 |
| ILMSAF [29] | −44.5729 | 0.48295 | 28.24830 | 0.042899 | −0.000007 | 0.0004 | 1 | 9166.868 |
| Vuvuzela [37] | −7.57176 | 0.964362 | 28.09070 | 0.044012 | −0.00406 | 0.24772 | 0.517621 | 4297.485 |
| OMLSA [38] | −22.3793 | 1.486634 | 5.52273 | 0.555447 | 0.079563 | 0.594998 | 0.736039 | 4650.964 |
| TSNR [39] | −6.72913 | 2.241367 | 26.98552 | 0.047656 | 0.022435 | 0.591289 | 0.751859 | 3098.7 |
| HRNR [40] | −6.75711 | 2.264289 | 27.031751 | 0.047326 | 0.022294 | 0.583314 | 0.752648 | 3050.421 |
| RNMF [26] | 9.310854 | 2.451338 | 32.75409 | 0.026238 | 0.88414 | 0.705571 | 0.811554 | 2974.847 |
| Prop | **11.8937** | **2.659263** | **39.28692** | **0.013656** | **0.953861** | **0.765353** | **0.873405** | **1532.681** |
| *Restaurant noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | −10.4461 | 0.7942 | 26.6383 | 2.2139 | 0.1803 | 0.02634 | 0.52721 | 17,083.1356 |
| ILMSAF [29] | −45.9102 | 0.499356 | 28.2485 | 0.042896 | 0.0000344 | 0.001954 | 1 | 10,499.8 |
| Vuvuzela [37] | −7.44252 | 1.052113 | 28.0778 | 0.044116 | −0.00527 | 0.277397 | 0.539848 | 4145.347 |
| OMLSA [38] | −22.2584 | 1.592738 | 5.52345 | 0.555395 | 0.081084 | 0.662682 | 0.775123 | 4573.293 |
| TSNR [39] | −6.7375 | 2.624678 | 26.9696 | 0.047729 | 0.02078 | 0.681692 | 0.801552 | 2944.466 |
| HRNR [40] | −6.76414 | 2.659293 | 26.9976 | 0.047516 | 0.020522 | 0.675716 | 0.799488 | 2923.681 |
| RNMF [26] | 9.769551 | 2.619633 | 33.08000 | 0.024631 | 0.897358 | 0.765587 | 0.836966 | 2747.786 |
| Prop | **13.81539** | **2.907064** | **42.65285** | **0.01171** | **0.967353** | **0.84411** | **0.912251** | **1375.572** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

tuning ratio of the Wiener filter is set as the target for the respective noise intensities. The training is performed using the Levenberg–Marquardt training algorithm. Given a corrupted test speech, the noise intensity is estimated and it is followed by the estimation of tuning ratio. Based on the estimated tuning ratio, the Wiener filtering is applied to enhance the corrupted speech signal.

## 4.2 Qualitative analysis

The quality of the selected speech signals is studied in this section. Moreover, the analysis such as temporal analysis, spectral analysis and time–frequency analysis for denoising performance is also observed for six noise types, namely airport noise, exhibition noise, restaurant noise, station noise, street noise and babble noise, which are added to the speech signals. Figure 4a–f illustrates the temporal analysis

**Table 5** Station noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Station noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | – 10.4298 | 0.8029 | 26.8864 | 2.2302 | 0.18068 | 0.0242 | 0.5289 | 17,134.7076 |
| ILMSAF [29] | – 43.0395 | 0.516516 | 28.2350 | 0.042933 | – 0.000018 | 0.000365 | 1 | 7277.851 |
| Vuvuzela [37] | – 8.78239 | 0.599902 | 28.16731 | 0.043788 | – 0.00688 | 0.12826 | 0.405335 | 5326.739 |
| OMLSA [38] | – 23.609 | 1.136879 | 5.516967 | 0.555847 | 0.064141 | 0.331076 | 0.554855 | 5066.86 |
| TSNR [39] | – 6.96222 | 1.407002 | 27.28262 | 0.047271 | 0.015467 | 0.3204 | 0.547674 | 3502.536 |
| HRNR [40] | – 7.02212 | 1.308554 | 27.41390 | 0.046677 | 0.015591 | 0.277251 | 0.546833 | 3433.749 |
| RNMF [26] | **6.671867** | 1.992645 | 30.79072 | 0.032156 | 0.817177 | 0.452036 | 0.665938 | 0 |
| Prop | 6.114214 | **2.014777** | **33.91242** | **0.023215** | **0.858347** | **0.474452** | **0.697443** | **2117.406** |
| *Station noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | – 10.3866 | 0.79468 | 26.5786 | 2.2127 | 0.1828 | 0.0259 | 0.5275 | 17,062.6255 |
| ILMSAF [29] | – 43.0508 | 0.579121 | 28.2460 | 0.04291 | 0.0000135 | 0.001277 | 1 | 7636.656 |
| Vuvuzela [37] | – 7.59672 | 0.861347 | 28.1401 | 0.043798 | – 0.0025 | 0.194482 | 0.48034 | 4981.494 |
| OMLSA [38] | – 22.8765 | 1.371561 | 5.52194 | 0.555557 | 0.073565 | 0.465456 | 0.656106 | 4827.331 |
| TSNR [39] | – 6.69946 | 1.959304 | 27.07713 | 0.047543 | 0.021207 | 0.456918 | 0.658742 | 3293.792 |
| HRNR [40] | – 6.74983 | 1.887786 | 27.1635 | 0.047094 | 0.021118 | 0.424697 | 0.65947 | 3211.321 |
| RNMF [26] | 8.343678 | 2.334204 | 31.82070 | 0.029265 | 0.865977 | 0.601457 | 0.765671 | 3494.473 |
| Prop | **9.903879** | **2.446756** | **38.29393** | **0.016576** | **0.931048** | **0.645581** | **0.810262** | **1770.943** |
| *Station noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | – 10.4142 | 0.8069 | 26.8453 | 2.2576 | 0.1809 | 0.0246 | 0.5308 | 17,237.252 |
| ILMSAF [29] | – 48.8073 | 0.531978 | 28.24765 | 0.0429 | 0.0000319 | – 0.00112 | 1 | 8473.772 |
| Vuvuzela [37] | – 7.48921 | 0.99167 | 28.09044 | 0.043997 | – 0.00472 | 0.239015 | 0.516205 | 4366.422 |
| OMLSA [38] | – 22.4691 | 1.495268 | 5.52299 | 0.555429 | 0.07843 | 0.587385 | 0.731867 | 4668.752 |
| TSNR [39] | – 6.67067 | 2.347521 | 26.95910 | 0.047692 | 0.020162 | 0.600952 | 0.752588 | 3066.506 |
| HRNR [40] | – 6.71314 | 2.338328 | 27.01854 | 0.047359 | 0.020229 | 0.589183 | 0.754907 | 3048.834 |
| RNMF [26] | 9.524805 | 2.546329 | 32.89604 | 0.02591 | 0.894694 | 0.701516 | 0.81523 | 0 |
| Prop | **12.53183** | **2.725255** | **40.88445** | **0.012982** | **0.959115** | **0.759757** | **0.870494** | **1534.63** |
| *Station noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | – 10.3658 | 0.79406 | 26.70513 | 2.2224 | 0.1835 | 0.02792 | 0.52893 | 17,097.2573 |
| ILMSAF [29] | – 45.754 | 0.549845 | 28.2482 | 0.042896 | – 0.0000096 | – 0.00099 | 1 | 10,002.46 |
| Vuvuzela [37] | – 7.41393 | 1.069357 | 28.07887 | 0.044107 | – 0.0033 | 0.262964 | 0.536423 | 4170.249 |
| OMLSA [38] | – 22.289 | 1.636335 | 5.523403 | 0.555392 | 0.080706 | 0.660804 | 0.771335 | 4586.93 |
| TSNR [39] | – 6.70273 | 2.807841 | 26.9491 | 0.047756 | 0.021589 | 0.678983 | 0.799672 | 2963.738 |
| HRNR [40] | – 6.73479 | 2.787415 | 26.98597 | 0.047498 | 0.021414 | 0.666929 | 0.795567 | 2947.717 |
| RNMF [26] | 9.960434 | 2.706672 | 33.70012 | 0.024279 | 0.902648 | 0.760993 | 0.842388 | 2838.799 |
| Prop | **14.03563** | **3.006517** | **41.28067** | **0.011652** | **0.966707** | **0.835726** | **0.910228** | **1437.081** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

of the denoising performance, in which the noisy and the denoised signals of various noise types and the performance of the proposed methodology are proved for their efficiency in denoising the signal are shown in Figs. 7, 8, 9, 10, 11, 12. Figure 5 illustrates the spectral analysis of the denoising performance for various noise types, which include airport noise, exhibition noise, restaurant noise, station noise, street noise and babble noise. Here, the noisy and the denoised signals are shown and it is found that the performance rate of the proposed method is high by ultimate reduction of the noise from the noisy signal. Further, Fig. 6 illustrates the time–frequency analysis of the denoising performance, in which the noisy and the denoised signals of various noise types are shown. The superior noise-removing ability of the proposed method is precisely understood from this figure.

### 4.3 Quantitative analysis

The proposed speech enhancement algorithm is compared to the state-of-the-art methods like low-rank approximation

**Table 6** Street noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Street noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | – 10.3508 | 0.8069 | 26.8003 | 2.2545 | 0.1830 | 0.02434 | 0.5299 | 17,206.0324 |
| ILMSAF [29] | – 29.0713 | 0.386485 | 28.24517 | 0.041198 | – 0.0000012 | – 0.00303 | 1 | 6669.757 |
| Vuvuzela [37] | – 8.79086 | 0.693556 | 28.16884 | 0.041617 | – 0.01414 | 0.111887 | 0.385911 | 7117.692 |
| OMLSA [38] | – 23.3815 | 1.297848 | 5.518987 | 0.481817 | 0.071972 | 0.32705 | 0.576932 | 5198.124 |
| TSNR [39] | – 6.35935 | 1.226439 | 27.29515 | 0.044753 | – 0.05652 | 0.23802 | 0.495741 | 4264.394 |
| eSHRNR [40] | – 6.37398 | 1.113361 | 27.40063 | 0.044295 | – 0.05983 | 0.200118 | 0.502984 | 4264.492 |
| RNMF [26] | **7.202237** | 2.085 | 31.48721 | 0.031483 | 0.826486 | 0.496823 | 0.682608 | 4175.865 |
| Prop | 6.505504 | **2.04939** | **35.09585** | **0.022436** | **0.863687** | **0.509153** | **0.71288** | **2275.803** |
| *Street noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | – 10.4028 | 0.7997 | 26.6633 | 2.2208 | 0.18210 | 0.02640 | 0.5280 | 17,090.374 |
| ILMSAF [29] | – 41.1891 | 0.347932 | 28.2475 | 0.041166 | 0.0000203 | – 0.00299 | 1 | 7176.363 |
| Vuvuzela [37] | – 8.46669 | 0.835758 | 28.1020 | 0.041951 | – 0.00917 | 0.212747 | 0.466567 | 5217.142 |
| OMLSA [38] | – 21.993 | 1.502347 | 5.52068 | 0.481596 | 0.08801 | 0.494314 | 0.674252 | 4858.855 |
| TSNR [39] | – 7.06186 | 1.881626 | 26.9478 | 0.046547 | – 0.01488 | 0.44168 | 0.668866 | 3119.174 |
| HRNR [40] | – 7.0982 | 1.897935 | 27.00894 | 0.046256 | – 0.01304 | 0.420817 | 0.684409 | 3073.434 |
| RNMF [26] | 8.622569 | 2.370913 | 32.41061 | 0.028934 | 0.864937 | 0.60159 | 0.757201 | 3934.027 |
| Prop | **10.48717** | **2.414182** | **37.92920** | **0.016037** | **0.934853** | **0.635419** | **0.79669** | **1964.252** |
| *Street noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | – 10.3086 | 0.8105 | 26.69456 | 2.2415 | 0.1847 | 0.03025 | 0.5315 | 17,153.806 |
| ILMSAF [29] | – 39.1995 | 0.416146 | 28.2473 | 0.04116 | – 0.000025 | 0.000222 | 1 | 8980.504 |
| Vuvuzela [37] | – 8.34898 | 1.020207 | 28.09422 | 0.041953 | – 0.00577 | 0.233092 | 0.502643 | 4747.651 |
| OMLSA [38] | – 21.7346 | 1.7324 | 5.522853 | 0.481429 | 0.092094 | 0.635414 | 0.789033 | 4659.472 |
| TSNR [39] | – 6.68444 | 2.24976 | 27.03556 | 0.046285 | – 0.0188 | 0.571822 | 0.771147 | 3031.058 |
| HRNR [40] | – 6.71648 | 2.286572 | 27.09301 | 0.046066 | – 0.01835 | 0.56431 | 0.776018 | 3009.596 |
| RNMF [26] | 10.0503 | 2.651448 | 33.17159 | 0.02482 | 0.902525 | 0.700434 | 0.818501 | 0 |
| Prop | **13.45219** | **2.769417** | **40.28013** | **0.012185** | **0.964429** | **0.756369** | **0.871417** | **1766.508** |
| *Street noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | – 10.4688 | 0.7959 | 26.67656 | 2.2268 | 0.17930 | 0.02352 | 0.5283 | 17,122.543 |
| ILMSAF [29] | – 43.0316 | 0.504134 | 28.24852 | 0.041157 | 0.0000757 | 0.002941 | 1 | 9153.841 |
| Vuvuzela [37] | – 8.45853 | 1.103083 | 28.06584 | 0.04207 | – 0.00851 | 0.273837 | 0.535553 | 4621.293 |
| OMLSA [38] | – 21.553 | 1.854615 | 5.523514 | 0.481373 | 0.093912 | 0.727654 | 0.83191 | 4601.213 |
| TSNR [39] | – 6.64125 | 2.593716 | 26.95361 | 0.046455 | – 0.02291 | 0.652852 | 0.818702 | 2978.182 |
| HRNR [40] | – 6.66462 | 2.573333 | 26.98725 | 0.046301 | – 0.02283 | 0.645521 | 0.819777 | 2931.21 |
| RNMF [26] | 10.5605 | 2.754843 | 34.25164 | 0.023246 | 0.907395 | 0.750891 | 0.834912 | 3281.504 |
| Prop | **14.97509** | **2.977879** | **43.37776** | **0.010818** | **0.971795** | **0.821248** | **0.898608** | **1597.327** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

(LRA) [30], ILMSAF [29], Vuvuzela [37], optimal modified minimum mean square error log-spectral amplitude (OMLSA) [38], two-step noise reduction (TSNR) [39], harmonic regeneration noise reduction (HRNR) [40] and regularized nonnegative matrix factorization RNMF [26]. The quality of the input speech signals is studied with different measures like PESQ, SNR, root-mean-square error (RMSE), correlation, STOI, extended STOI (ESTOI), SDR and cumulative squared Euclidean distance (CSED). Further, the investigation is proceeded with different SNR dB levels such as 0 dB, 5 dB, 10 dB and 15 dB. Table 2 shows the performance investigation of the proposed method against the existing methods for the airport noise at various dB levels. Similarly, Tables 3, 4, 5, 6 and 7 show the performance investigation of the proposed method, for exhibition noise, restaurant noise, station noise, street noise and babble noise, at different dB levels, respectively. From Table 2, while comparing the conventional methods, the proposed method leads position for airport noise at SNR = 5 dB with 9.26 SDR, 2.36 PESQ, 8.23 SNR, 0.017 RMSE, 0.92 correlation, 0.64 ESTOI, 0.81 STOI

**Table 7** Babble noise at different intensity levels

| Methods | SDR | PESQ | SNR | RMSE | Correlation | ESTOI | STOI | CSED |
|---|---|---|---|---|---|---|---|---|
| *Babble noise at SNR 0 dB* | | | | | | | | |
| LRA [30] | −29.3206 | 1.2492 | 10.0762 | 0.6066 | 0.0119 | 0.3452 | 0.5686 | 6700.345 |
| ILMSAF [29] | −44.5177 | 0.350636 | 3.705896 | 0.034231 | 0.000071 | 0.000541 | 1 | 6302.352 |
| Vuvuzela [37] | −10.538 | 0.291167 | 29.31127 | 0.034416 | 0.017666 | 0.123171 | 0.268919 | 5652.945 |
| OMLSA [38] | −28.5134 | 1.01435 | 29.26486 | 0.598757 | 0.043366 | 0.286956 | 0.465407 | 5825.889 |
| TSNR [39] | −4.03642 | 1.174523 | 4.454985 | 0.039327 | −0.15688 | 0.291666 | 0.51614 | 4372.283 |
| HRNR [40] | −4.04488 | 1.097431 | 28.10611 | 0.038533 | −0.15801 | 0.246216 | 0.508015 | 4083.372 |
| RNMF [26] | **6.906372** | 1.722543 | 28.28328 | 0.026349 | 0.76644 | 0.319191 | 0.574315 | 3844.686 |
| Prop | 6.20913 | **1.902121** | **31.64638** | **0.02233** | **0.818016** | **0.366412** | **0.595346** | **3072.769** |
| *Babble noise at SNR 5 dB* | | | | | | | | |
| LRA [30] | −28.2115 | 1.538 | 10.0059 | 0.606 | 0.0250 | 0.5807 | 0.7842 | 6257.3161 |
| ILMSAF [29] | −47.873 | 0.489815 | 5.52662 | 0.034219 | −0.0000024 | 0.000364 | 1 | 7012.314 |
| Vuvuzela [37] | −7.38911 | 0.612969 | 29.31455 | 0.035107 | −0.00749 | 0.207169 | 0.40774 | 4042.433 |
| OMLSA [38] | −27.1359 | 1.18378 | 29.0921 | 0.598583 | 0.052134 | 0.478167 | 0.647879 | 5622 |
| TSNR [39] | −5.37797 | 1.771651 | 4.45750 | 0.04101 | −0.13546 | 0.513049 | 0.681794 | 3176.579 |
| HRNR [40] | −5.50824 | 1.875257 | 27.74213 | 0.040561 | −0.13215 | 0.515815 | 0.705963 | 3140.937 |
| RNMF [26] | 8.91804 | 2.012063 | 27.83780 | 0.022675 | 0.839892 | 0.548873 | 0.726988 | 3704.015 |
| Prop | **9.446117** | **2.132058** | **32.83511** | **0.018107** | **0.889465** | **0.586192** | **0.745457** | **2728.421** |
| *Babble noise at SNR 10 dB* | | | | | | | | |
| LRA [30] | −27.9044 | 1.7605 | 10.0211 | 0.6066 | 0.02603 | 0.65721 | 0.8603 | 6123.426 |
| ILMSAF [29] | −39.8411 | 0.455403 | 10.2810 | 0.034216 | −0.00017 | 0.000155 | 1 | 8728.255 |
| Vuvuzela [37] | −11.4098 | 0.699444 | 29.3157 | 0.034655 | 0.002222 | 0.19033 | 0.361669 | 4409.774 |
| OMLSA [38] | −26.8735 | 1.327771 | 29.20465 | 0.59855 | 0.054035 | 0.547575 | 0.713626 | 5502.45 |
| TSNR [39] | −5.15497 | 2.164079 | 4.457990 | 0.040664 | −0.14 | 0.54818 | 0.733701 | 3104.883 |
| HRNR [40] | −5.28808 | 2.260716 | 27.81573 | 0.04033 | −0.13742 | 0.54411 | 0.750874 | 3057.119 |
| RNMF [26] | 11.04823 | 2.480265 | 27.8873 | 0.020863 | 0.874637 | 0.635716 | 0.805794 | 3624.103 |
| Prop | **13.20129** | **2.647154** | **33.35560** | **0.010474** | **0.961372** | **0.705626** | **0.857672** | **1806.428** |
| *Babble noise at SNR 15 dB* | | | | | | | | |
| LRA [30] | −27.7785 | 1.8918 | 10.0199 | 0.6066 | 0.02623 | 0.7474 | 0.9175 | 6098.077 |
| ILMSAF [29] | −39.8866 | 0.431535 | 11.20955 | 0.034213 | −0.0000058 | 0.000603 | 1 | 9848.481 |
| Vuvuzela [37] | −11.0845 | 0.789693 | 29.31616 | 0.034699 | 0.001369 | 0.20085 | 0.409755 | 4444.612 |
| OMLSA [38] | −26.7349 | 1.381456 | 29.19357 | 0.598537 | 0.055002 | 0.615176 | 0.756076 | 5480.147 |
| TSNR [39] | −5.21817 | 2.55182 | 4.45817 | 0.040696 | −0.14384 | 0.661409 | 0.807145 | 3089.656 |
| HRNR [40] | −5.29458 | 2.595153 | 27.80901 | 0.04048 | −0.14214 | 0.652382 | 0.809342 | 3013.952 |
| RNMF [26] | 10.57615 | 2.54707 | 27.85519 | 0.020398 | 0.870586 | 0.744149 | 0.839639 | 2960.952 |
| Prop | **14.2091** | **2.760114** | **33.92268** | **0.009413** | **0.968986** | **0.794852** | **0.892859** | **1502.005** |

Bold values indicated the best resulted value. Mostly the proposed model has shown superior value. But, in some cases other models too show best results when compared to that of proposed model

**Table 8** Computational time for denoising a speech signal

| Methods | Time in seconds |
|---|---|
| LRA [30] | 2.8346 |
| ILMSAF [29] | 1.2425 |
| Vuvuzela [37] | 0.28791 |
| OMLSA [38] | 1.0881 |
| TSNR [39] | 0.61697 |
| HRNR [40] | 0.51269 |
| RNMF [26] | 2.0361 |
| Prop | 2.7934 |

and 1606.895 CSED. Table 3 shows the proposed method, for the case of exhibition noise, with high SDR, PSEQ, SNR, correlation, ESTOI and STOI of 5.26, 1.88, 5.16, 0.83, 0.557 and 0.73, at 0 dB results. Subsequently, the RMSE and the CSED values of the proposed method are found to reduce gradually as 0.024 and 2406.215, respectively.

In the same way, for the other noise types such as restaurant noise, station noise, street noise and babble noise at varied dB levels, the proposed method outperforms, with respect to the performance rate. Further, it is observed that the measures like PESQ, SNR, correlation, STOI, ESTOI

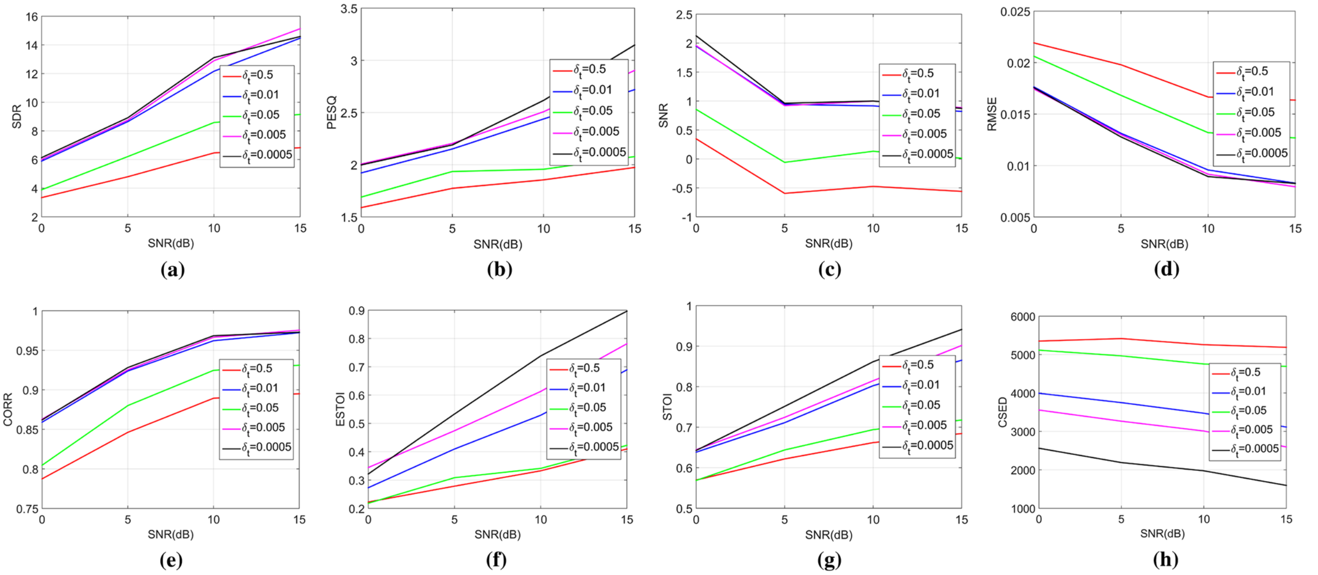


**Fig. 7** Performance analysis for mitigating the airport noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

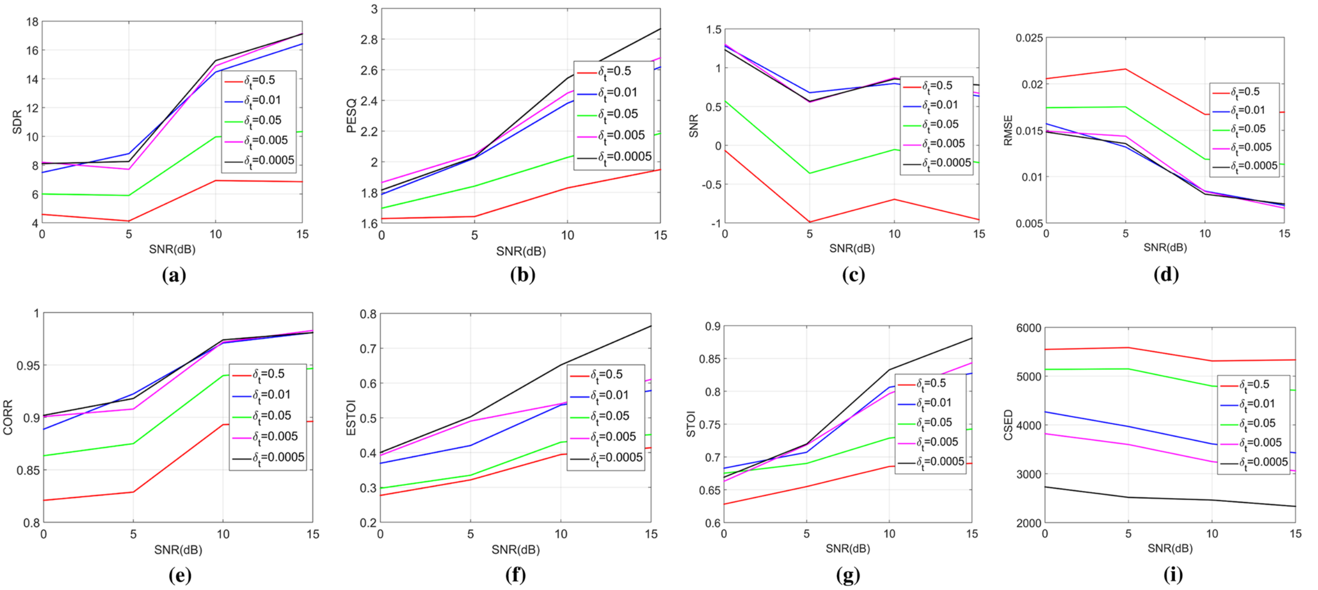


**Fig. 8** Performance analysis for mitigating the exhibition noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

and SDR of the proposed method are abundantly increased, whereas the existing methods showed poor performance with low values. Similarly, the measures like RMSE and CSED of the proposed method are decreased. But, the existing methods show increased values of RMSE and CSED, leading to the performance excellence of the proposed method. Apart from this, Table 8 demonstrates the computational time required for denoising the speech signal by the proposed methodology and the other existing methods. During comparison, it is observed that the proposed method requires 2.7934 s to denoise a speech signal. Even though the computational time is higher, the proposed method dominates all the existing methods in terms of speech enhancement.

### 4.4 Impact of D-EMCD thresholding

In this paper, the threshold value of the D-EMCD is fixed as $0.5e^{-4}$. The analysis is performed by varying the

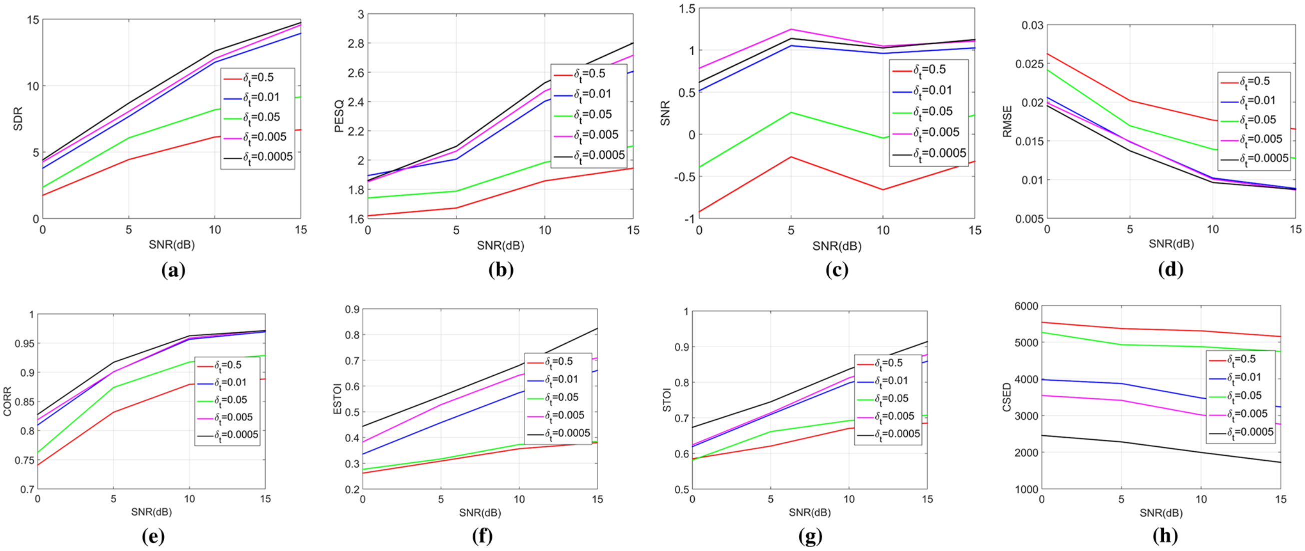


**Fig. 9** Performance analysis for mitigating the restaurant noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

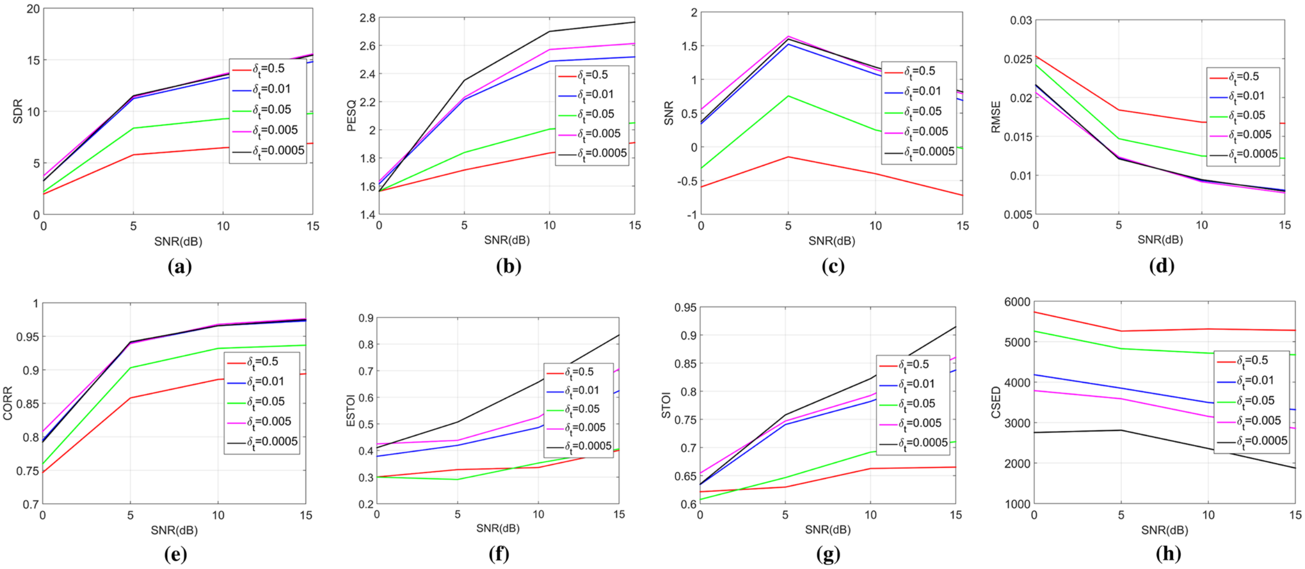


**Fig. 10** Performance analysis for mitigating the station noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

threshold $\delta_t$ values as 0.5, 0.01, 0.05, 0.005 and 0.0005, for all noise types with varied dB levels like 0, 5, 10, 15. Figure 7 illustrates the performance of varied measures like (a) SDR, (b) PESQ, (c) SNR, (d) RMSE, (e) correlation, (f) ESTOI, (g) STOI, (h) CSED of airport noise with different threshold values. As the value of threshold decreases, the performance of SDR, PESQ, SNR, correlation, ESTOI and STOI increases, in a sense that the mentioned measures exhibited a drastic improvement with the threshold $\delta_t$ value of 0.0005. Similarly, the measures like RMSE and CSED gradually decrease at the same $\delta_t$ value. The same analysis is observed for all the noise types like exhibition noise, restaurant noise, station noise, street noise and babble noise. Figures 8, 9, 10, 11 and 12 demonstrate the analysis of power spectrum estimation for the denoised signal of all the six noise types. These figures clearly characterize the frequency content of the denoised signals with varied threshold $\delta_t$ values such as 0.5, 0.01, 0.05, 0.005 and 0.0005. Threshold decides the

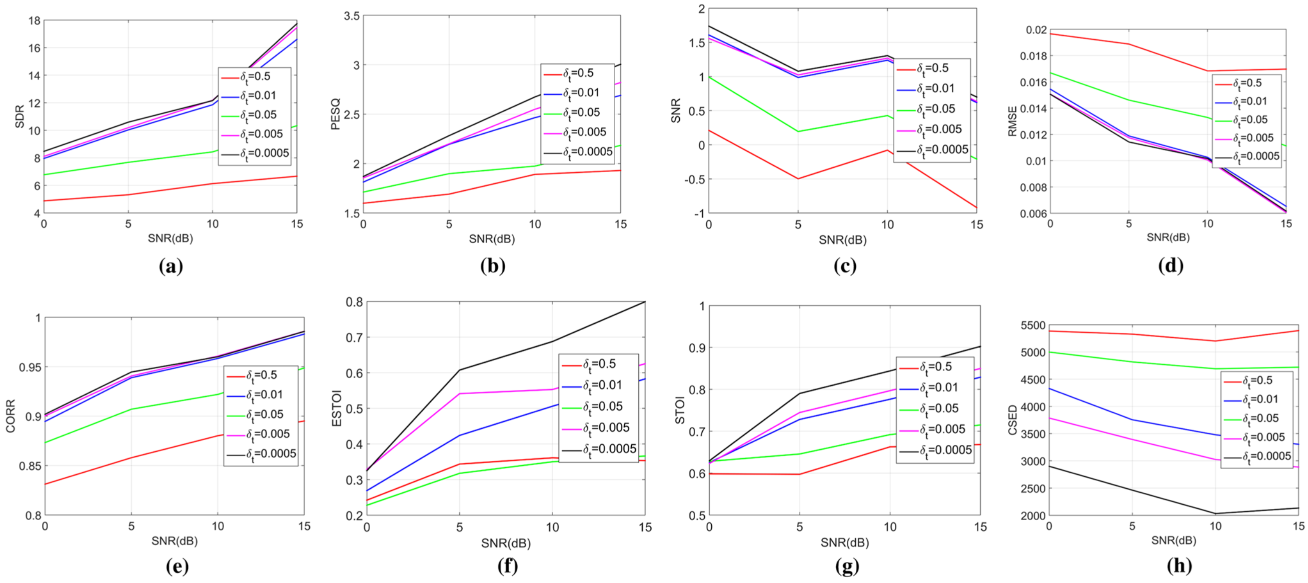


**Fig. 11** Performance analysis for mitigating the street noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

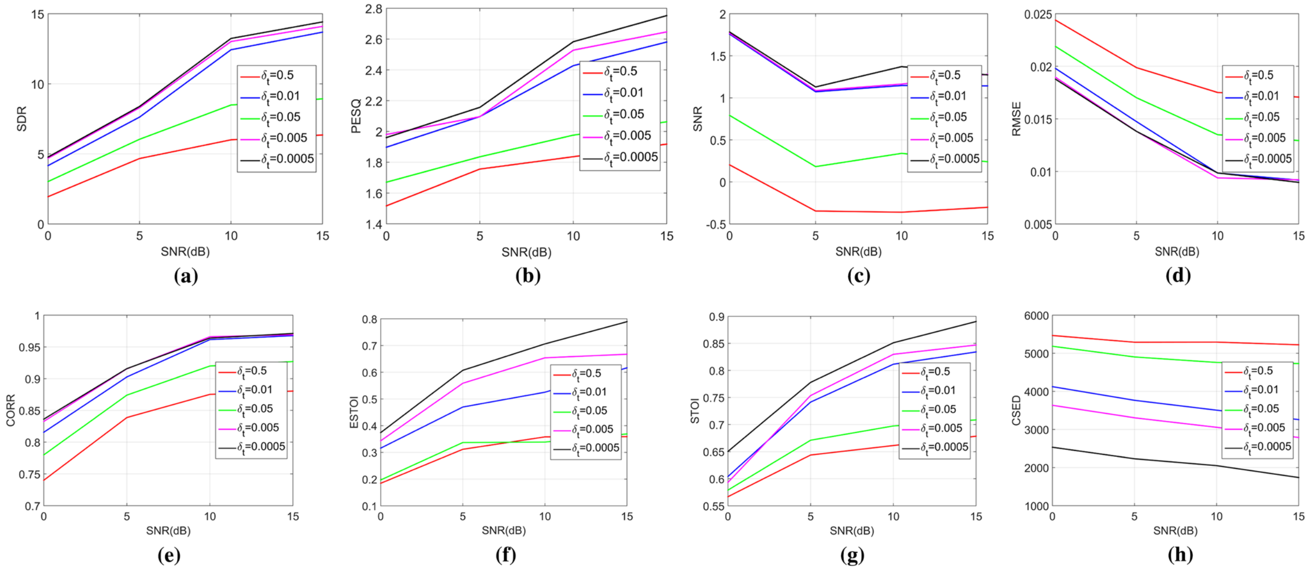


**Fig. 12** Performance analysis for mitigating the babble noise (with varying threshold): **a** SDR, **b** PESQ, **c** SNR, **d** RMSE, **e** correlation, **f** ESTOI, **g** STOI and **h** CSED

quality of D-EMCD, and it can be set based on trial and error (Fig. 13).

The impact of the threshold on the D-EMCD performance is high, but the relationship between the threshold and the performance remains unknown. Hence, the analysis is performed by varying the threshold. The results have revealed that minimum threshold leads to improved performance.

## 5 Conclusion

In this paper, a speech enhancement algorithm using short-time Fourier domain has been presented to overcome the regular drawbacks of the conventional speech enhancement algorithms. Further, a decomposition model, named diminished empirical mean curve decomposition (D-EMCD), has also been introduced to remove the undesired signals. Further,

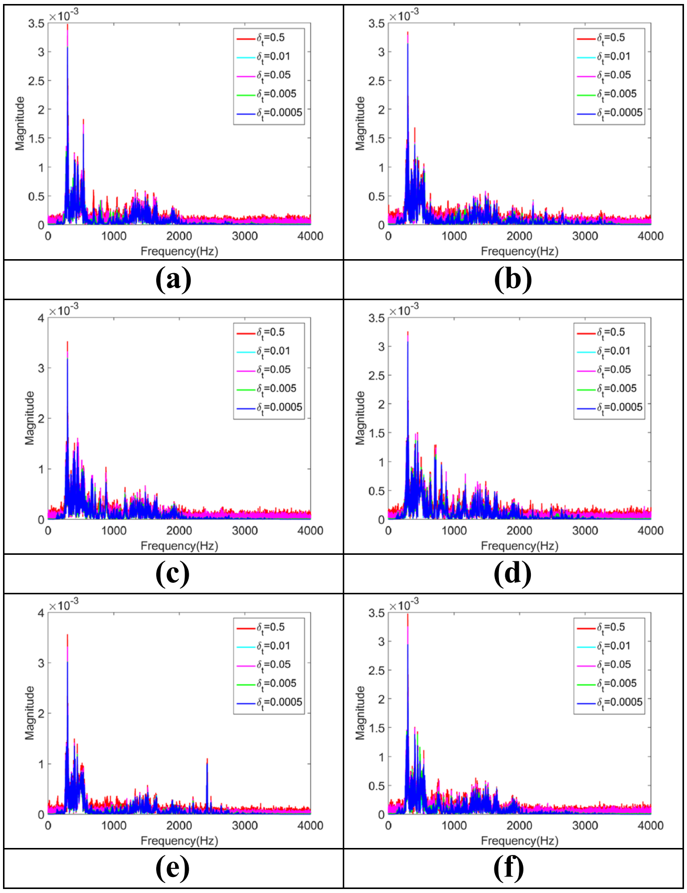


**Fig. 13** Power spectrum of denoised speech with varying threshold, in case of **a** airport noise, **b** exhibition noise, **c** restaurant noise, **d** station noise, **e** street noise and **f** babble noise

the Wiener filtering process has been adopted to accomplish an effective speech enhancement. The proposed methodology has been developed in MATLAB, and the performance of the proposed method has been analyzed with various measures. Moreover, the proposed method has been compared with the existing methods for proving its superiority.

## References

1. Moore AH, Peso Parada P, Naylor PA (2016) Speech enhancement for robust automatic speech recognition: evaluation using a baseline system and instrumental measures. Comput Speech Lang 86:85–96
2. Zao L, Coelho R, Flandrin P (2014) Speech enhancement with EMD and hurst-based mode selection. IEEE/ACM Trans Audio Speech Lang Process 22(5):899–911
3. Xu Y, Du J, Dai LR, Lee CH (2015) A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 23(1):7–19
4. Aroudi A, Veisi H, Sameti H (2015) Hidden Markov model-based speech enhancement using multivariate Laplace and Gaussian distributions. IET Signal Process 9(2):177–185
5. Baby D, Virtanen T, Gemmeke JF, Van Hamme H (2015) Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition. IEEE/ACM Trans Audio Speech Lang Process 23(11):1788–1799
6. Chen Z, Hohmann V (2015) Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation. IEEE/ACM Trans Audio Speech Lang Process 23(11):1904–1916
7. Deng F, Bao C, Kleijn WB (2015) Sparse hidden Markov models for speech enhancement in non-stationary noise environments. IEEE/ACM Trans Audio Speech Lang Process 23(11):1973–1987
8. Vihari S, Murthy AS, Soni P, Naik DC (2016) Comparison of speech enhancement algorithms. Procedia Comput Sci 89:666–676
9. Doi H, Toda T, Nakamura K, Saruwatari H, Shikano K (2014) Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. IEEE/ACM Trans Audio Speech Lang Process 22(1):172–183
10. Gerkmann T, Krawczyk-Becker M, Le Roux J (2015) Phase processing for single-channel speech enhancement: history and recent advances. IEEE Signal Process Mag 32(2):55–66
11. Islam MT, Shahnaz C, Zhu WP, Ahmad MO (2015) Speech enhancement based on student t modeling of teager energy operated perceptual wavelet packet coefficients and a custom thresholding function. IEEE/ACM Trans Audio Speech Lang Process 23(11):1800–1811
12. Jin YG, Shin JW, Kim NS (2014) Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain. IEEE Signal Process Lett 21(3):352–355
13. Kim SM, Kim HK (2014) Direction-of-arrival based SNR estimation for dual-microphone speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 22(12):2207–2217

14. Ghanbari Y, Karami-Mollaei MR (2006) A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets. Speech Commun 48(8):927–940
15. Ephraim Y, Malah D (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 33(2):443–445
16. Cohen I (2004) Speech enhancement using a noncausal a priori SNR estimator. IEEE Signal Process Lett 11(9):725–728
17. Berouti M, Schwartz R, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise. In: IEEE international conference on acoustics, speech, and signal processing, ICASSP '79, pp 208–211
18. Kamath S, Loizou P (2002) A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP). IEEE, Orlando, p IV-4164
19. Lu Y, Loizou PC (2008) A geometric approach to spectral subtraction. Speech Commun 50(6):453–466
20. Ayat S, Manzuri-Shalmani MT, Dianat R (2006) An improved wavelet-based speech enhancement by using speech signal features. Comput Electr Eng 32(6):411–425
21. Balaji GN, Subashini TS, Chidambaram N (2015) Detection of heart muscle damage from automated analysis of echocardiogram video. IETE J Res 61(3):236–243
22. Sunil Kumar BS, Manjunath AS, Christopher S (2018) Improved entropy encoding for high efficient video coding standard. Alexandria Eng J 57(1):1–9
23. Wagh AM, Todmal SR (2015) Eyelids, eyelashes detection algorithm and Hough transform method for noise removal in iris recognition. Int J Comput Appl 112(3):28–31
24. Sreedharan NPN, Ganesan B, Raveendran R, Sarala P, Dennis B, Rajakumar BR (2018) Grey Wolf optimisation-based feature selection and classification for facial emotion recognition. IET Biom 7(5):490–499
25. Bhowmick A, Chandra M (2017) Speech enhancement using voiced speech probability based wavelet decomposition. Comput Electr Eng 62:706–718
26. Chung H, Plourde E, Champagne B (2017) Regularized nonnegative matrix factorization with Gaussian mixtures and masking model for speech enhancement. Speech Commun 87:18–30
27. Mowlaee P, Stahl J, Kulmer J (2017) Iterative joint MAP single-channel speech enhancement given non-uniform phase prior. Speech Commun 86:85–96
28. Kammi S, Karami-Mollaei MR (2017) Noisy speech enhancement with sparsity regularization. Speech Commun 87:58–69
29. Li R, Liu Y, Shi Y, Dong L, Cui W (2016) ILMSAF based speech enhancement with DNN and noise classification. Speech Commun 85:53–70
30. Zhao Y, Qiu RC, Zhao X, Wang B (2016) Speech enhancement method based on low-rank approximation in a reproducing kernel Hilbert space. Appl Acoust 112:79–83
31. Liu Y, Nower N, Morita S, Unoki M (2016) Speech enhancement of instantaneous amplitude and phase for applications in noisy reverberant environments. Speech Commun 84:1–14
32. Sun M, Zhang X, Van Hamme H, Zheng TF (2016) Unseen noise estimation using separable deep auto encoder for speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 24(1):93–104
33. Chazan SE, Goldberger J, Gannot S (2016) A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier. IEEE/ACM Trans Audio Speech Lang Process 24(12):2516–2530
34. Wang SS et al (2016) Wavelet speech enhancement based on nonnegative matrix factorization. IEEE Signal Process Lett 23(8):1101–1105
35. Bhatnagar K, Gupta S (2017) Extending the neural model to study the impact of effective area of optical fiber on laser intensity. Int J Intell Eng Syst 10(4):274–283
36. Muaidi H (2014) Levenberg–Marquardt learning neural network for part-of-speech tagging of arabic sentences. Wseas Trans Comput 13:300–309
37. Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Signal Process 27(2):113–120
38. Cohen I, Berdugo B (2001) Speech enhancement for non-stationary noise environments. Signal Process 81(11):2403–2418
39. Plapous C, Marro C, Mauuary L, Scalart P (2004) A two-step noise reduction technique. In: 2004 IEEE international conference on acoustics, speech, and signal processing, vol 1, pp I-289–I292
40. Plapous C, Marro C, Scalart P (2006) Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans ASLP 14(6):2098–2108

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.