PAN
ARCHIVES OF ACOUSTICS
Vol. **44**, No. 3, pp. 429–437 (2019)
DOI: 10.24425/aoa.2019.129259

Speech Enhancement Using Sliding Window Empirical Mode Decomposition and Hurst-based Technique

Selvaraj POOVARASAN*, Eswaran CHANDRA

Department of Computer Science, Bharathiar University
Coimbatore, India; e-mail: crcspeech@gmail.com
*Corresponding Author e-mail: pppoovarasan@gmail.com

(*received February 25, 2019; accepted April 4, 2019*)

The most challenging in speech enhancement technique is tracking non-stationary noises for long speech segments and low Signal-to-Noise Ratio (SNR). Different speech enhancement techniques have been proposed but, those techniques were inaccurate in tracking highly non-stationary noises. As a result, Empirical Mode Decomposition and Hurst-based (EMDH) approach is proposed to enhance the signals corrupted by non-stationary acoustic noises. Hurst exponent statistics was adopted for identifying and selecting the set of Intrinsic Mode Functions (IMF) that are most affected by the noise components. Moreover, the speech signal was reconstructed by considering the least corrupted IMF. Though it increases SNR, the time and resource consumption were high. Also, it requires a significant improvement under non-stationary noise scenario. Hence, in this article, EMDH approach is enhanced by using Sliding Window (SW) technique. In this SWEMDH approach, the computation of EMD is performed based on the small and sliding window along with the time axis. The sliding window depends on the signal frequency band. The possible discontinuities in IMF between windows are prevented by the total number of modes and the number of sifting iterations that should be set a priori. For each module, the number of sifting iterations is determined by decomposition of many signal windows by standard algorithm and calculating the average number of sifting steps for each module. Based on this approach, the time complexity is reduced significantly with suitable quality of decomposition. Finally, the experimental results show the considerable improvements in speech enhancement under non-stationary noise environments.

Keywords: Speech Enhancement; Empirical Mode Decomposition; Intrinsic Mode Functions; Hurst exponent; Sliding Window EMD.

1. Introduction

In recent years, the suppression of acoustic distortion in noisy speech signals has been mostly required to enhance the speech signals. Various speech enhancement techniques and algorithms have been proposed by many researchers to reduce the noise from the speech signals (Vishari *et.al.*, 2016; Kulkarni *et al.*, 2016). Typically, in real non-stationary environments, the major problem in speech enhancement is concerned with the estimation of the noise statistics precisely. The conventional estimators are based on Voice Activity Detectors (VAD) (Kasap, Arslan, 2013; Zhang *et al.*, 2014). After that, the power spectrum of the noise components is determined as a smoothed adaptation of its previous values obtained during the speech pauses. These processes offer a reasonable accuracy for stationary background noises but they cannot accurately estimate time-varying spectra. The complexity in tracking the non-stationary noises becomes more obvious for long speech segments and low Signal-to-Noise Ratio (SNR) (Hawaldar, Dixit, 2011; Mai *et al.*, 2015). Different power spectrum-based methods have been proposed to deal with such situations (Zhao *et al.*, 2014; Jin *et al.*, 2017b).

In the past researches, Time-Frequency-based (TF based) speech enhancement solutions (Soni *et al.*, 2018) were proposed based on the Empirical Mode Decomposition (EMD) (Mai *et al.*, 2015; Mert, Akan, 2014). Generally, the EMD is a nonlinear time-domain adaptive method to decompose the signals into a series of oscillatory Intrinsic Mode Functions (IMF) and a residual one (Mandic *et al.*, 2013; Zeileret *et al.*, 2010). It does not need a set of basic functions for appropriately analyzing the target signal. In addition, it does not restrict the stationary signals. To tackle the challenges in non-stationary noisy atmospheres, a novel EMD-based speech enhancement technique

(Zao *et al.*, 2014) was proposed in which the noise components of each IMF were identified and chosen by its Hurst exponent statistics. Here, the selection of IMF and the speech reconstruction were performed on the frame-by-frame basis by considering both quality and intelligibility objective measures. However, this technique consumes a lot of time and computer resources. Also, a significant improvement under Babble noise scenarios was not achieved effectively.

Hence in this article, Sliding Window EMDH (SWEMDH) is proposed to improve the EMDH approach. This approach is performed based on the calculation of EMD in a comparatively small window and sliding this window along with the time axis. Window size is depending on the signal's frequency band. The possible discontinuities in IMF between windows are prevented by the total number of modes and the number of sifting iterations that should be set *a priori*. The number of sifting steps should be tailored for each module. This parameter depends on the sampling frequency and on analyzed signal, its complexity and spectrum. The number of sifting iterations is determined by decomposition of many signal windows by a standard algorithm and calculating the average number of sifting steps for each module. Thus, the speech enhancement technique is improved efficiently.

The rest of the article is structured as follows: Sec. 2 presents the literature survey related to the speech enhancement techniques. Section 3 describes the proposed speech enhancement technique. Section 4 shows the experimental results of the proposed technique. Finally, Sec. 5 concludes the research work and presents the Future Enhancement.

2. Literature survey

A noise reduction algorithm (Taal et al., 2011) was proposed for the intelligibility prediction of time-frequency weighted noisy speech. A Short-Time Objective Intelligibility Measure (STOI) was proposed which has a strong monotonic relation with the intelligibility scores of various listening tests where noisy speech was processed by some type of TF-weighting. This model has a simple structure in the sense that it was based on only two free parameters. However, the performance was not effective.

A colored noise based multi-condition training technique (Zao, Coelho, 2011) was proposed for robust speaker identification in unknown noisy environments. In this technique, the colored noise samples generation was based on filtering a white Gaussian sequence. Gaussian Mixture Models (GMM) was applied for obtaining the speaker models by using the noisy speech signals with a single SNR. However, the identification accuracy was less precise.

The variational Bayesian algorithm (wa Maina, Walsh, 2011) was proposed for joint speech enhancement and speaker identification. This technique was constructed on the intuition that speaker dependent priors may operate better than priors that attempt for capturing global speech properties. An iterative algorithm was derived that exchanges information between the speech enhancement and speaker identification processes. However, the computational complexity of this algorithm was high.

A novel technique (Gerkmann, Hendriks, 2012) was proposed to estimate the noise power spectral density by means of an unbiased Minimum Mean-Square Error (MMSE) optimal estimation. In this technique, a VAD-based noise power estimator was used that neglects the bias compensation by a soft Speech Presence Probability (SPP) with fixed priors. By selecting fixed priors, decoupling of the noise power estimator was achieved such as the estimation of the speech power and the estimation of the clean speech. However, the processing time was high.

EMD-based Filtering (EMDF) of low-frequency noise (Chatlani, Soraghan, 2012) was proposed for speech enhancement. In this technique, an adaptive method was developed for selecting the IMF index to separate the noise components from the speech according to the second-order IMF statistics. Then, the low-frequency noise components were separated by a partial reconstruction from the IMF. Based on this technique, a residual noise was suppressed from the speech signals that were enhanced by the conventional optimally modified log-spectral amplitude approach that utilizes a minimum statistics-based noise estimate. However, a minor improvement was required with the non-stationary Babble noise.

The speech enhancement strategy (Khaldi *et al.*, 2014) was proposed based on time adaptive thresholding of IMF of the signal extracted by EMD. The denoised signal was reconstructed by the superposition of its adaptive thresholded IMFs. The adaptive thresholds were estimated by using the Teager-Kaiser energy operator (TKEO) of signal IMFs. It was used to identify the type of frame by expanding differences between speech and non-speech frames in each IMF. However, the parameters used for implementing a compromise between noise removal and speech distortion were required to optimize for further improvement.

Enhancement of speech dynamics for VAD (Dwijayanti *et al.*, 2018) was proposed by using Deep Neural Network (DNN). In this technique, the dynamics were highlighted by speech period candidates which are computed based on the heuristic rules for the patterns of the first and second derivatives of the input signals. Then, these candidates combined with the log power spectra were given as input to the DNN for obtaining VAD decisions. However, the performance of VAD was degraded while it eliminates the sub bands F0 and its neighbours.

A fast and robust VAD (Ghahabi *et al.*, 2018) was proposed for a real-time Automatic Speech Recognition (ASR) process. The major objective of this method was filtering the non-speech segments before processing the speech segments of the audio signal by the decoder. This method was a hybrid supervised or unsupervised model based on the zero-order Baum-Welch statistics obtained from a Universal Background Model (UBM). During testing, the Baum-Welch statistics of an unknown audio segment was compared with speech and non-speech VAD vectors. Finally, the decision was made based on the robust threshold. However, Equal Error Rate (EER) was high.

3. Proposed methodology

In this section, the proposed SWEMDH approach is explained in brief. The basic block diagram of the proposed approach is shown in Fig. 1. The enhancement of speech signals involves the following processes:

- Initially, noisy speech signals are collected from the database and the extrema are extracted i.e., the noisy speech signals are decomposed into a set of windowed IMFs by using SWEMDH technique.
- Once all windowed IMFs are obtained, the Hurst exponent is applied to select the most optimal IMF low-frequency noisy components.
- Finally, the noisy speech signals are reconstructed by using the selected windowed IMFs efficiently.

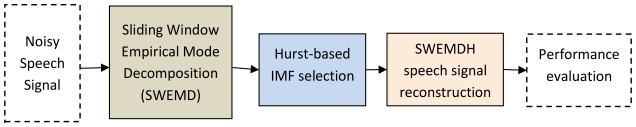

Fig. 1. Block diagram for proposed approach.

3.1. Sliding Window Empirical Mode Decomposition (SWEMD)

Initially, the extrema i.e., maxima and minima are extracted from the original signal $x(t)$. Then, the upper ($e_{\max}$) and lower ($e_{\min}$) envelopes are obtained by interpolating the local maxima and minima, respectively. The average between these envelopes is computed as:

$$m(t) = \frac{e_{\max}(t) + e_{\min}(t)}{2}. \tag{1}$$

The obtained average value is subtracted from the original signal to obtain imf as:

$$\mathrm{imf}_1(t) = x(t) - m(t). \tag{2}$$

Generally, this process is known as sifting process. The computed $\mathrm{imf}_1(t)$ is used as the input for next sifting process which is applied on the residual as:

$$\mathrm{imf}_1(t) \coloneqq \mathrm{imf}_1(t) - m(t). \tag{3}$$

This sifting process is iterated until $\mathrm{imf}_1(t)$ satisfies the conditions of imf signal. The original signal is reduced by the first mode while the sifting process is completed

$$\mathrm{IMF}_1(t) \coloneqq \mathrm{imf}_1(t), \tag{4}$$

$$r_1(t) = x(t) - \mathrm{IMF}(t). \tag{5}$$

The residue $r_1(t)$ is used as input for extracting the second IMF and this process is looped for extracting all IMF as follows:

$$r_i(t) = r_{i-1}(t) - \mathrm{imf}_i(t), \tag{6}$$

where i refers to the index of current mode. When residue $r_i(t)$ consists of less than three extrema or all its points are closely equal to zero, the decomposition process is completed. The original signal is obtained by sum of all IMF components and the residue as:

$$r_n + \sum_{i=1}^{n} \mathrm{imf}_i(t) = x(t), \tag{7}$$

where n refers to the number of all modes.

3.2. Termination criteria for sifting process

The following criteria are used to terminate the sifting process:

- The first one is that the number of extrema and the number of zero-crossings should vary at most by 1.
- The second one is that the mean between the upper and lower envelopes should equal to zero at each point of IMF.

In this proposed approach, the second criterion is used for terminating the sifting process. In accordance with the second criteria of IMF, the mean of its envelope is equivalent to zero at each point of IMF. Therefore, a termination criterion is used for the sifting process. In each iteration, the ratio of the mean value of the envelope of iterated mode and the amplitude of this envelope is verified

$$\tau(t) = \left| \frac{m(t)}{a(t)} \right|, \tag{8}$$

where

$$a(t) = \frac{e_{\max}(t) - e_{\min}(t)}{2}. \tag{9}$$

There are two thresholds used such as ϑ_1 and ϑ_2, where ϑ_1 is ensuring globally small fluctuations of the envelope mean around zero and ϑ_2 is locally allowing higher fluctuations. The sifting process is terminated if $\tau(t) < \vartheta_1$ is true for $(1-\varepsilon)$ part of the number of signal's points and if $\tau(t) < \vartheta_2$ is true for the remaining points. The typical values of these parameters are given as:

$$\varepsilon = 0.05; \qquad \vartheta_1 = 0.05; \qquad \vartheta_2 = 10 \cdot \vartheta_1. \tag{10}$$

These values of parameters tolerate a compromise between quality and speed of the decomposition process.

3.3. Hurst-based IMF selection

Once all IMF are obtained, the Hurst exponent is applied to decide which IMFs should be chosen for the speech signal reconstruction. Since those selected IMFs affect by the noise components. Once all IMF are obtained, the Hurst exponent is applied to select IMF for speech signal reconstruction. Consider the speech signal $x(t)$ with the normalized autocorrelation coefficient function $(\delta(k))$ as:

$$\delta(k) = \frac{E[(x(t)-\mu_x)(x(t+k)-\mu_x)]}{E[(x(t)-\mu_x)^2]}. \tag{11}$$

In equation (Soni *et al.*, 2018), μ_x refers the mean of $x(t)$ and k refers the time lag. For a fractional Gaussian noise, $\delta(k)$ is given as:

$$\delta(k) = \frac{1}{2}\left(|k-1|^{2H} - 2|k|^{2H} + |k+1|^{2H}\right), \tag{12}$$

where $0 \le H \le 1$ refers the Hurst exponent of $x(t)$. The value of H is defined by using autocorrelation coefficient function decaying rate whose asymptotic characteristic is given by,

$$\delta(k) \sim H(2H-1)k^{2(H-1)}, k \to \infty. \tag{13}$$

The Hurst exponent defines the time-dependence or scaling degree of $x(t)$ and is associated with its spectral characteristics. Within the entire range [0, 1], the power spectral density $S_x(f)$ is exposed to be proportional to f^{1-2H} when $f \to 0$ (Zhao *et al.*, 2014). For $H = \frac{1}{2}$, $S_x(f)$ is a constant over the entire frequency spectrum, where low frequencies are important in the case where $H > \frac{1}{2}$ and $H \to 1$. The Hurst exponent is estimated from non-overlapping frames of samples and it is used to enable the identification criteria for selecting the IMF low-frequency noise components. Figure 2 illustrates the first five IMFs obtained

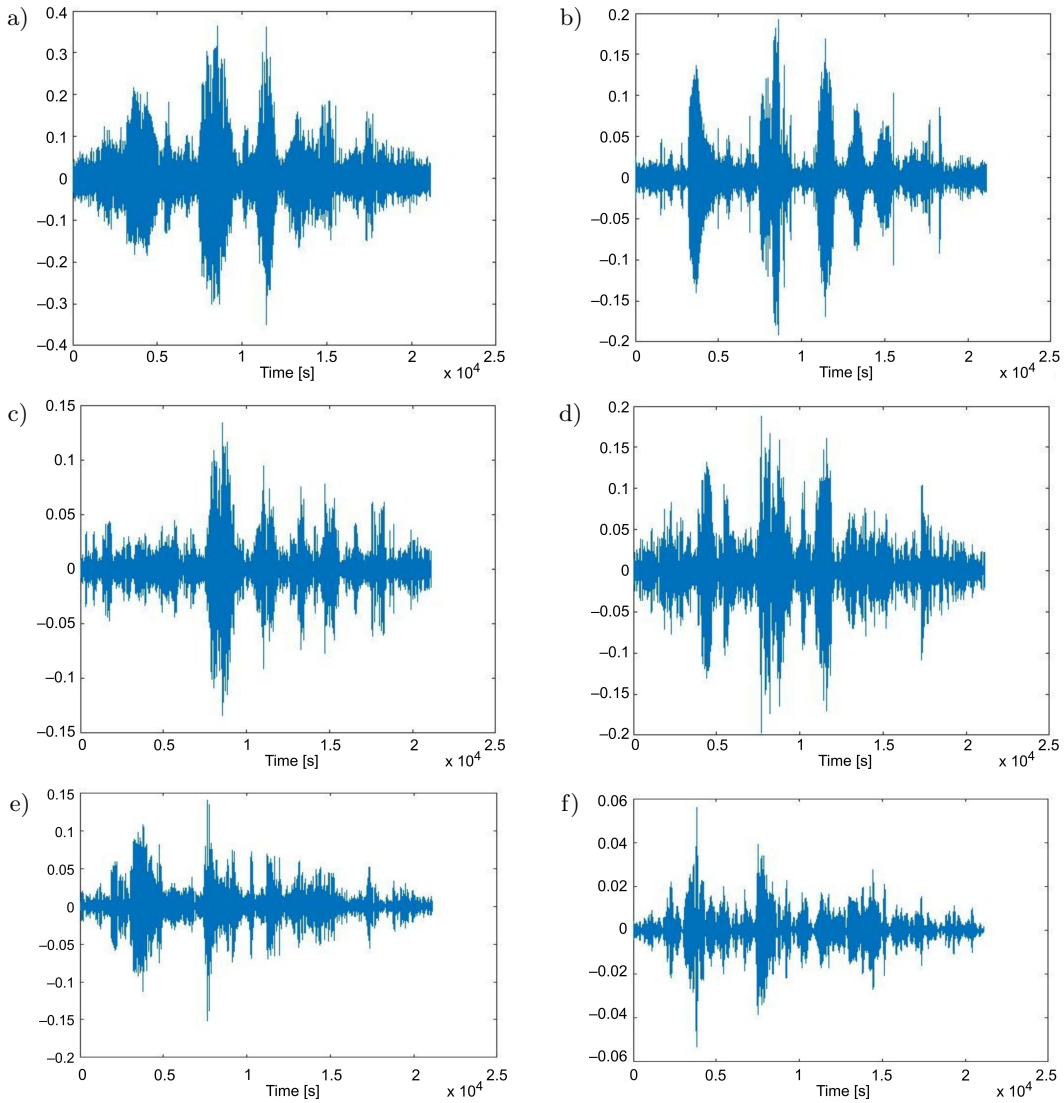

Fig. 2. The first five IMFs obtained from the decomposition of a speech signal segment: a) input speech signal, b)–f) IMF#1–IMF#5 , respectively.

from decomposing the sample input speech signal segment of 2500 ms collected from the NOISEX-92 database. It shows that the first IMF is composed of faster oscillations than the second one which in its turn has faster fluctuations than the third one and so on. It implies that, at each time interval, the SWEMD applies a high-frequency *versus* low-frequency partition between IMFs. Therefore, the first mode should present the high-frequency content of the signal. Also, the cut-off frequency between consecutive IMFs is time-varying and signal dependent.

3.4. SWEMDH speech signal reconstruction

The speech signal reconstruction is performed to validate the decomposition. Normally, the speech signal reconstruction defines the determination of an original speech signal from a sequence of equally spaced segments i.e., IMFs. It initiates with the decomposition of the input noisy speech into n modes by using (Mai *et al.*, 2015). After that, windowed IMF are obtained by separating each mode into Q non-overlapping short-time frames

$$\mathrm{w\_imf}_{i,q}(t) = \begin{cases} \mathrm{imf}_i\left(t + qT_d\right), & t \in [0, T_d], \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $q \in \{0, ..., Q-1\}$ refers the frame index and T_d refers the fixed time-duration of the frames. Then, the Hurst exponent is estimated to all the windowed IMF $(\mathrm{w\_imf}_{i,q}(t))$ to select the IMF low-frequency noise components for each frame index q. In the next step, for each frame, the index N_q of the last windowed IMF whose value of H is below a given threshold i.e., $H_q(N_q) < H_{th}$. If $\widehat{x}(t)$ is an enhanced speech signal, then each of it's $\widehat{x}_q(t)$ is reconstructed as follows:

$$\widehat{x}_q(t) = \sum_{m=1}^{N_q} \mathrm{w\_imf}_{i,q}(t). \tag{15}$$

Finally, $\widehat{x}(t)$ is given as follows:

$$\widehat{x}(t) = \sum_{q=0}^{Q-1} \widehat{x}_q(t - qT_d). \tag{16}$$

Thus, based on this proposed SWEMDH, the sudden changes in the power spectrum of non-stationary noises are avoided and the selection of IMF for entire speech signal is achieved efficiently.

4. Results and discussions

In this section, performance effectiveness of the proposed SWEMDH is evaluated and compared with the existing EMDH approaches by using MATLAB 2014a. In this experiment, a subset of 12 speakers including 7 male and 5 female is randomly chosen that provides a total of 420 speech data segments, 10 per speaker with sampling rate of 16 kHz and average time duration of 2 seconds. Also, acoustic noises such as Airport, Babble, Car, Exhibition, Restaurant, Station, Street and Train are used for corrupting the speech signals considering different SNR values like 0 dB, 5 dB, 10 dB and 15 dB. The noises are collected from the NOISEX-92 database. The following are the performance metrics used to evaluate the effectiveness of the proposed technique:

- **Signal-to-Noise Ratio (SNR):** It is defined as the fraction of the speech signal power to the corrupting noise power. It is computed as:

$$\mathrm{SNR\ [dB]} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right). \tag{17}$$

  In Eq. (17), P_{signal} is the average power of speech signal and P_{noise} is the average power of noise. It can be rewritten as:

$$\mathrm{SNR\ [dB]} = 20\log_{10}\left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right). \tag{18}$$

  In equation (Taal *et al.*, 2011), A_{signal} and A_{noise} are the Root Mean Square (RMS) amplitude of signal and noise, respectively.
- **Mean Square Error (MSE):** It represents the cumulative squared error between the reconstructed and original speech signal. The MSE is calculated as:

$$\mathrm{MSE} = \frac{1}{l}\sum_{i=1}^{n} e_i^2, \quad \text{where} \quad e = \widehat{x}(t) - x(t). \tag{19}$$

  In equation (Zao, Coelho, 2011), l refers the signal length and e refers the error between the original signal $x(t)$ and reconstructed signal $\widehat{x}(t)$.
- **Peak Signal-to-Noise Ratio (PSNR):** It is defined as the fraction of maximum possible signal power to the corrupting noise power. Generally, it is computed by using MSE as:

$$\mathrm{PSNR\ [dB]} = 10\log_{10}\frac{255^2}{\mathrm{MSE}}. \tag{20}$$

- **Mean Absolute Error (MAE):** It is defined as the absolute error between the reconstructed speech signal and original signal. It is computed as:

$$\mathrm{MAE} = \frac{1}{l}\sum_{i=1}^{n} e_i. \tag{21}$$

- **Perceptual Evaluation of Speech Quality (PESQ):** It can be applied to provide an end-to-end quality assessment for characterizing the listening quality as perceived by users.

$$\mathrm{PESQ} = \alpha_0 - \alpha_1 \cdot D - \alpha_2 \cdot A, \tag{22}$$

  where $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, and $\alpha_2 = 0.0309$.

Table 1. MSE comparison.

Noise	EMDH 0 dB	SWEMDH 0 dB	EMDH 5 dB	SWEMDH 5 dB	EMDH 10 dB	SWEMDH 10 dB	EMDH 15 dB	SWEMDH 15 dB
Airport	0.005775	0.001500	0.003425	0.000574	0.003703	0.000245	0.002460	0.000175
Babble	0.005829	0.001506	0.003546	0.000572	0.002682	0.002489	0.002428	0.000178
Car	0.005574	0.001521	0.003438	0.000567	0.002683	0.001901	0.002434	0.000179
Exhibition	0.004849	0.001501	0.002967	0.000576	0.002607	0.000226	0.004142	0.000305
Restaurant	0.006223	0.001478	0.003437	0.000568	0.002693	0.001444	0.002452	0.000176
Station	0.005827	0.001502	0.003167	0.000568	0.002629	0.001429	0.002411	0.000303
Street	0.004502	0.001495	0.003227	0.000569	0.002607	0.000398	0.002392	0.000308
Train	0.004846	0.001501	0.003074	0.000571	0.002567	0.001027	0.002388	0.000309

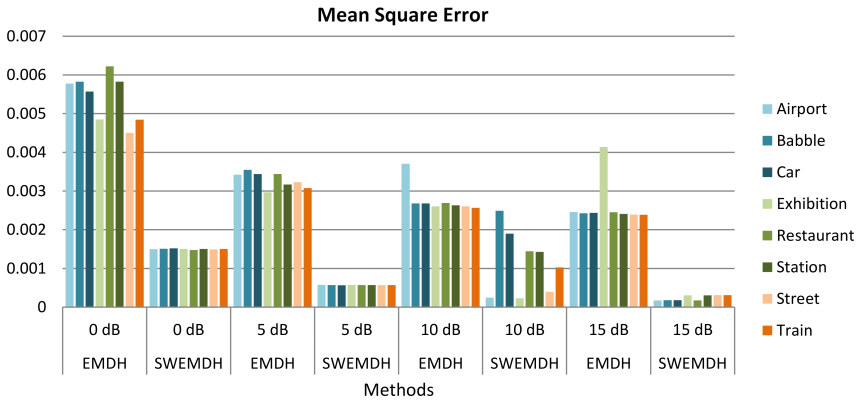

Fig. 3. MSE comparison.

Table 2. MAE comparison.

Noise	EMDH 0 dB	SWEMDH 0 dB	EMDH 5 dB	SWEMDH 5 dB	EMDH 10 dB	SWEMDH 10 dB	EMDH 15 dB	SWEMDH 15 dB
Airport	0.058149	0.030742	0.041596	0.018812	0.034674	0.010447	0.028804	0.010411
Babble	0.058241	0.030531	0.040286	0.017854	0.032865	0.039863	0.028596	0.010397
Car	0.056816	0.031154	0.041438	0.018981	0.031981	0.002595	0.028534	0.010331
Exhibition	0.052171	0.030548	0.036580	0.018083	0.031631	0.006500	0.037259	0.013411
Restaurant	0.056832	0.028862	0.039627	0.018559	0.032467	0.024235	0.028333	0.010201
Station	0.052369	0.028160	0.038677	0.019042	0.032034	0.024235	0.027558	0.013685
Street	0.049660	0.030168	0.038047	0.018010	0.030643	0.024285	0.027848	0.014016
Train	0.050743	0.029732	0.037712	0.018608	0.030818	0.024272	0.027511	0.013279

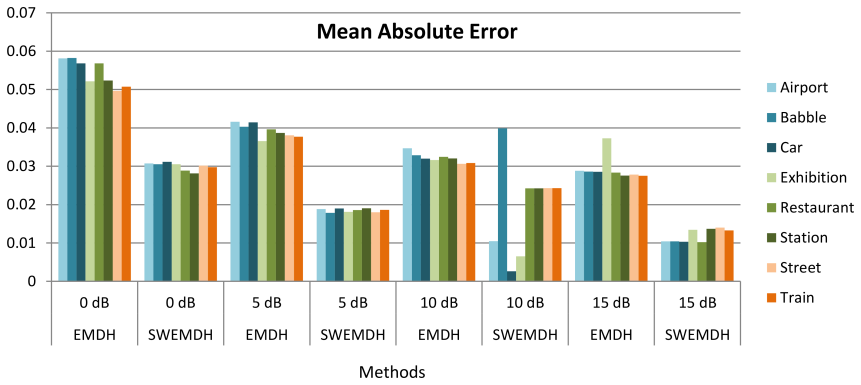

Fig. 4. MAE comparison.

The following Table 1 and Fig. 3 give the comparison results of MSE for both EMDH and SWEMDH using different acoustic noises that corrupt the speech signal during transmission.

The following Table 2 and Fig. 4 gives the comparison results of MAE for both EMDH and SWEMDH using different acoustic noises that corrupt the speech signal during transmission.

The following Table 3 and Fig. 5 gives the comparison results of SNR for both EMDH and SWEMDH using different acoustic noises that corrupt the speech signal during transmission.

The following Table 4 and Fig. 6 give the comparison results of PSNR for both EMDH and SWEMDH using different acoustic noises that corrupt the speech signal during transmission.

The following Table 5 and Fig. 7 give the comparison results of PESQ for both EMDH and SWEMDH using different acoustic noises that corrupt the speech signal during transmission.

Table 3. SNR comparison [dB].

Noise	EMDH 0 dB	SWEMDH 0 dB	EMDH 5 dB	SWEMDH 5 dB	EMDH 10 dB	SWEMDH 10 dB	EMDH 15 dB	SWEMDH 15 dB
Airport	3.659574	2.195915	3.334045	4.426647	3.302368	6.338962	3.066095	8.405252
Babble	3.605948	2.271026	3.453267	4.469029	3.123635	2.434380	3.018528	8.323599
Car	3.437381	2.204023	3.324013	4.502377	3.074241	1.192139	3.018716	8.306291
Exhibition	2.820614	2.272402	2.689272	4.427774	2.976287	5.929496	2.789780	8.534318
Restaurant	3.833914	2.408575	3.272254	4.545853	3.165528	4.058466	3.035596	8.396639
Station	3.628947	2.258953	2.941313	4.522969	3.003922	4.058466	2.994448	8.369053
Street	2.550461	2.237160	3.068816	4.466179	3.022369	7.058466	2.948060	8.295345
Train	2.812100	2.276606	2.851648	4.458775	2.901905	6.008366	2.955278	8.257501

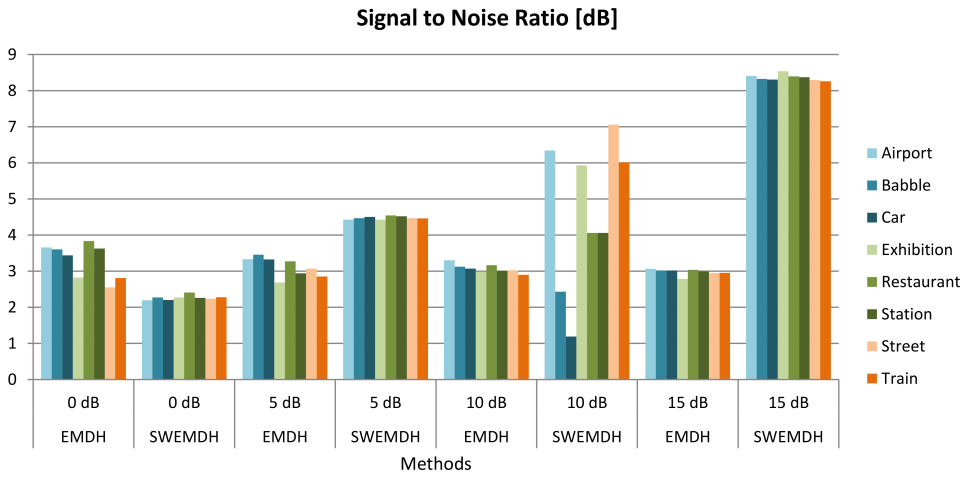

Fig. 5. SNR comparison.

Table 4. PSNR comparison [dB].

Noise	EMDH 0 dB	SWEMDH 0 dB	EMDH 5 dB	SWEMDH 5 dB	EMDH 10 dB	SWEMDH 10 dB	EMDH 15 dB	SWEMDH 15 dB
Airport	12.68316	18.53865	15.53044	23.29113	17.46393	25.73540	17.05465	28.52600
Babble	14.24706	20.12404	14.66708	22.58938	16.51313	24.83623	16.52640	27.86852
Car	13.46210	19.10351	15.31891	23.14530	16.24344	26.74035	16.67701	28.00202
Exhibition	14.91570	20.00872	16.27362	23.39067	16.46082	17.14640	14.22290	25.54700
Restaurant	13.38445	19.62694	15.19231	23.01042	16.19545	18.90165	16.82320	28.25544
Station	15.36787	21.25577	16.04403	23.50831	16.30616	26.90765	16.93355	27.09873
Street	15.32990	20.11752	15.48762	23.02262	16.82815	24.90065	16.81731	27.29751
Train	15.67256	20.76127	15.70213	23.01255	16.75199	25.90235	16.71708	27.32014

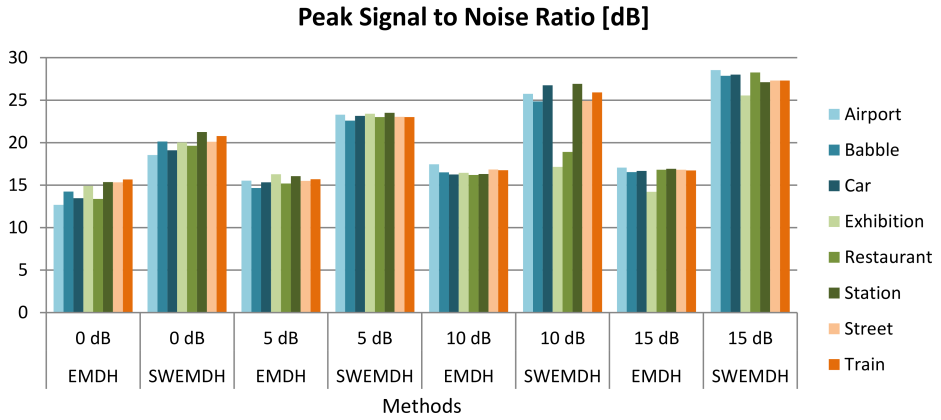

Fig. 6. PSNR comparison.

Table 5. PESQ comparison.

Noise	EMDH 0 dB	SWEMDH 0 dB	EMDH 5 dB	SWEMDH 5 dB	EMDH 10 dB	SWEMDH 10 dB	EMDH 15 dB	SWEMDH 15 dB
Airport	3.546681	3.772296	3.663083	3.671065	3.590500	3.927475	3.689458	4.023654
Babble	3.671733	3.849439	3.718452	3.822388	3.581719	3.943854	3.839045	3.867399
Car	3.334989	3.828398	3.765506	3.962600	3.716819	4.043255	3.598346	3.847243
Exhibition	3.658974	3.725694	3.629111	4.037867	3.446903	3.807430	3.681904	4.090796
Restaurant	3.023658	3.698754	3.763355	3.837838	3.882143	3.805933	3.881704	4.042269
Station	3.159864	3.897642	3.400217	3.837912	3.780931	3.888985	3.658974	4.042365
Street	3.451811	3.979580	3.473544	4.041857	3.867542	3.945786	3.708424	3.843790
Train	3.285802	3.800800	3.309595	3.873858	3.595872	4.025271	3.577204	4.006613

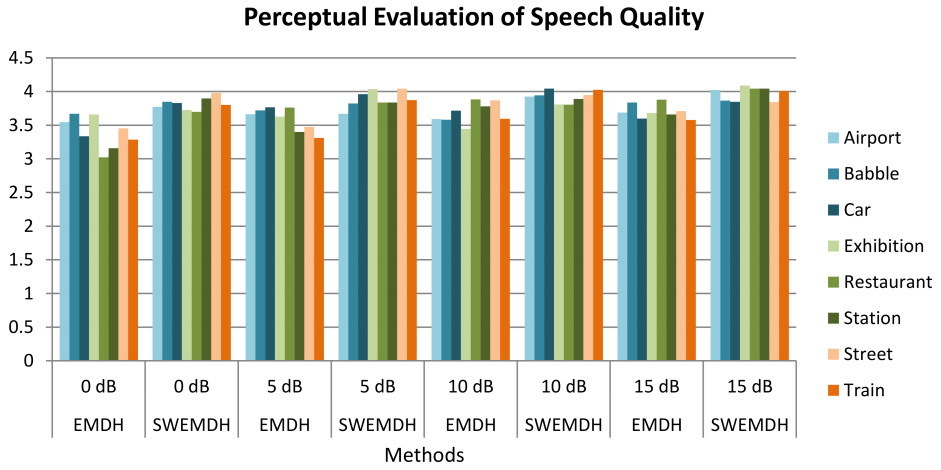

Fig. 7. PESQ comparison.

From this analysis, it is observed that SWEMDH approach achieves higher performance than the existing EMDH based speech enhancement. For example, consider the Babble noise environment with SNR is 15dB. For this case, the MSE of SWEMDH is 92.67% reduced differently than the EMDH technique. The MAE of SWEMDH is 63.64% less than the EMDH. Similarly, the PSNR value for the proposed SWEMDH technique is 68.63% increased than the existing technique. In addition, the PESQ of proposed technique is 0.74% higher than the existing EMDH technique. Thus, the proposed SWEMDH technique achieves high PSNR, SNR and PESQ with less MSE and MAE compared to the EMDH technique.

5. Conclusions

In this article, a Sliding Window-based EMDH approach (SWEMDH) is proposed to improve the speech enhancement under non-stationary acoustic noise envi-

ronments. In this proposed approach, the EMD computation is performed that estimates IMF according to the small and sliding window that depends on the signal's time frequency. To compute consecutive IMFs for each frame, the number of sifting iterations is determined by decomposition of many signal's windows by a standard algorithm and calculating the average number of sifting steps. After that, the Hurst exponent is applied on all IMFs to select the IMF low frequency components which are used to reconstruct the original speech signal. Thus, the time complexity of speech enhancement is reduced with an appropriate decomposition quality. Finally, the experimental results prove that the proposed SWEMDH approach has better performance than the existing EMDH in speech enhancement under non-stationary noise scenarios.

References

1. Chatlani N., Soraghan J.J. (2012), *EMD-based filtering (EMDF) of low-frequency noise for speech enhancement*, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 4, 1158–1166.
2. Dwijayanti S., Yamamori K., Miyoshi M. (2018), *Enhancement of speech dynamics for voice activity detection using DNN*, EURASIP Journal on Audio, Speech, and Music Processing, **2018**, 10, 15 pages.
3. Gerkmann T., Hendriks R.C. (2012), *Unbiased MMSE-based noise power estimation with low complexity and low tracking delay*, IEEE Transactions on Audio, Speech, and Language Processing, **20**, 4, 1383–1393.
4. Ghahabi O., Zhou W., Fischer V. (2018), *A robust voice activity detection for real-time automatic speech recognition*, [in:] Proceedings of ESSV 2018, Ulm, Germany.
5. Hawaldar S., Dixit M. (2011), *Speech enhancement for non-stationary noise environments*, Signal Image Processing, **2**, 4, 129–136.
6. Ji Y., Baek Y., Park Y.C. (2017a), *Robust noise power spectral density estimation for binaural speech enhancement in time-varying diffuse noise field*, EURASIP Journal on Audio, Speech, and Music Processing, **2017**, 1, 25.
7. Jin Y.G., Shin J.W., Kim N.S. (2017b), *Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement*, The Journal of the Acoustical Society of America, **141**, 3, EL228–EL233.
8. Kasap C., Arslan M.L. (2013), *A unified approach to speech enhancement and voice activity detection*, Turkish Journal of Electrical Engineering Computer Sciences, **21**, 2, 527–547.
9. Khaldi K., Boudraa A.O., Komaty A. (2014), *Speech enhancement using empirical mode decomposition and the Teager-Kaiser energy operator*, The Journal of the Acoustical Society of America, **135**, 1, 451–459.
10. Kulkarni D.S., Deshmukh R.R., Shrishrimal P.P. (2016), *A review of speech signal enhancement techniques*, International Journal of Computer Applications, **139**, 14, 23–26.
11. Mai V.K., Pastor D., Aïssa-El-Bey A., Le-Bidan R. (2015), *Robust estimation of on-stationary noise power spectrum for speech enhancement*, IEEE Transactions on Audio, Speech, and Language Processing, **23**, 4, 670–682.
12. Mandic D.P., Rehman N.U., Wu Z., Huang N.E. (2013), *Empirical mode decomposition-based time-frequency analysis of multivariate signals: the power of adaptive data analysis*, IEEE Signal Processing Magazine, **30**, 6, 74–86.
13. Mert A., Akan A. (2014), *Detrended fluctuation thresholding for empirical mode decomposition based denoising*, Digital Signal Processing, **32**, 48–56.
14. Soni M.H., Shah N., Patil H.A. (2018), *Time-frequency masking-based speech enhancement using generative adversarial network*, [in:] 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5039–5043.
15. Taal C.H., Hendriks R.C., Heusdens R., Jensen J. (2011), *An algorithm for intelligibility prediction of time-frequency weighted noisy speech*, IEEE Transactions on Audio, Speech, and Language Processing, **19**, 7, 2125–2136.
16. wa Maina C., Walsh J.M. (2011), *Joint speech enhancement and speaker identification using approximate Bayesian inference*, IEEE Transactions on Audio, Speech, and Language Processing, **19**, 6, 1517–1529.
17. Zao L., Coelho R. (2011), *Colored noise based multicondition training technique for robust speaker identification*, IEEE Signal Processing Letters, **18**, 11, 675–678.
18. Zao L., Coelho R., Flandrin P. (2014), *Speech enhancement with EMD and hurst-based mode selection*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, **22**, 5, 899–911.
19. Zeiler A., Faltermeier R., Keck I.R., Tomé A.M., Puntonet C.G., Lang E.W. (2010), *Empirical mode decomposition – an introduction*, [in:] 2010 IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 18–23 July, Barcelona, Spain.
20. Zhang Y., Tang Z.M., Li Y.P., Luo Y. (2014), *A hierarchical framework approach for voice activity detection and speech enhancement*, The Scientific World Journal, **2014**, Article ID 723643, 8 pages.
21. Zhao Y., Zhao X., Wang B. (2014), *A speech enhancement method based on sparse reconstruction of power spectral density*, Computers Electrical Engineering, **40**, 4, 1080–1089.