Hindawi Publishing Corporation
Advances in Acoustics and Vibration
Volume 2014, Article ID 765454, 11 pages
http://dx.doi.org/10.1155/2014/765454

Research Article

Subband DCT and EMD Based Hybrid Soft Thresholding for Speech Enhancement

Erhan Deger,[1] Md. Khademul Islam Molla,[1,2] Keikichi Hirose,[1] Nobuaki Minematsu,[3] and Md. Kamrul Hasan[4]

[1] *Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan*
[2] *Department of Computer Science and Engineering, The University of Rajshahi, Rajshahi 6205, Bangladesh*
[3] *Graduate School of Engineering, The University of Tokyo, Tokyo 113-8656, Japan*
[4] *Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh*

Correspondence should be addressed to Md. Khademul Islam Molla; molla@gavo.t.u-tokyo.ac.jp

Received 5 February 2014; Accepted 17 April 2014; Published 20 May 2014

Academic Editor: Rama B. Bhat

This paper presents a two-stage soft thresholding algorithm based on discrete cosine transform (DCT) and empirical mode decomposition (EMD). In the first stage, noisy speech is decomposed into eight frequency bands and a specific noise variance is calculated for each one. Based on this variance, each band is denoised using soft thresholding in DCT domain. The remaining noise is eliminated in the second stage through a time domain soft thresholding strategy adapted to the intrinsic mode functions (IMFs) derived by applying EMD on the signal obtained from the first stage processing. Significantly better SNR improvement and perceptual speech quality results for different noise types prove the superiority of the proposed algorithm over recently reported techniques.

1. Introduction

In many speech related systems, the desired signal is not available directly; rather it is mostly contaminated with some interference sources. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the post applications. Speech enhancement aims at improving the perceptual quality and intelligibility of such speech signals degraded in noisy environments, mainly through noise reduction algorithms [1]. Due to its significant importance in today's information technology, many methods have been developed for this purpose. A major problem in most algorithms is that the enhanced speech signal has distortions compared to the original one which results in loss of some speech details. The residual noise is another problem which affects the performance of the postprocessing systems.

Soft thresholding is a powerful technique used for removing the noise components by subtracting a constant value from the coefficients of the noisy speech signal obtained by the analyzing transformation. However, such type of direct subtraction results in a degradation of the speech components. Unlike the conventional constant noise-level subtraction rule [2, 3], a new soft thresholding strategy based on frequency frames was proposed in [4]. The later one is able to remove the noise components while giving significantly less damage to the speech signal. This enables even signals with high SNRs to be processed effectively. However due to the thresholding criteria, a noticeable amount of noise still remains in the enhanced signal. Another disadvantage is the lack of robustness of the algorithm to different noise types.

The empirical mode decomposition (EMD), recently pioneered by Huang et al. [5] as a new and powerful data analysis method for nonlinear and nonstationary signals, has made a novel and effective path for speech enhancement studies. Recent studies have shown that, with EMD, it is possible to successfully remove the noise components from the IMFs of the noisy speech. Since the extraction of the IMFs relies on

frequency characteristics, the IMFs with higher index contain lower frequency components. This property helps the noise and speech components to be roughly separated in terms of frequency and to dominate in different IMFs. Therefore, it will be even possible to identify and remove the noise parts that are embedded in the speech components.

In this paper, we propose a hybrid algorithm which will include a two-stage soft thresholding. In the first stage, a subband approach DCT domain soft thresholding is adapted to the noisy speech. The remaining noise in the enhanced speech looks like random tones and results in an irritating sound. Hence further denoising should be applied to get rid of this artifact. However, it is not an easy task to identify and remove these noise components without degrading the speech signal. Due to the frequency characteristics of the IMFs, further enhancement is achieved in the second stage through an EMD based soft thresholding strategy.

2. DCT Soft Thresholding

Transform domain speech enhancement methods commonly use amplitude subtraction based soft thresholding defined by [2, 3]

$$\widehat{X}_k = \begin{cases} \operatorname{sign}(X_k)\left(|X_k| - \sigma_v\right), & \text{if } |X_k| > \sigma_v, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where σ_v denotes the noise level, X_k is the kth coefficient of the noisy signal obtained by the analyzing transformation, and $\widehat{X}_k$ represents the corresponding thresholded coefficient. Since all the coefficients are thresholded by σ_v, the speech components are also degraded during this process. This degradation results in a loss in speech quality. Unlike the conventional constant noise-level subtraction rule in (1), a frame based soft thresholding strategy was proposed in [4]. The strategy depends on segmenting the signal into short time intervals and applying discrete cosine transform (DCT) on each frame. The DCT coefficients of each frame are divided into frequency bins which are categorized as either signal- or noise-dominant depending on their speech and noise energy distribution. Figure 1 shows an illustration of typical noise- and speech-dominant frequency bins. The problems of the conventional constant noise-level subtraction rules given in (1) can be well observed in this figure. For instance, it is apparent from Figure 1(a) that subtracting a constant value from the noisy speech coefficients in order to obtain the clean speech coefficients is inadequate. Furthermore, due to the second part of thresholding a significant amount of speech information may be lost, resulting in a source of musical noise. Therefore a linear thresholding is followed in noise-dominant frames. On the other hand, Figure 1(b) proves that soft thresholding is very inaccurate for signal-dominant frequency bins and will most probably degrade the speech components, therefore giving more damage than its contribution to the enhanced speech. Therefore, the signal-dominant frames should better be kept as they are in order not to degrade the high energy speech components. This enables even signals with high SNRs to be processed effectively.

The noisy speech is first segmented into 32 ms frames and a 512-point DCT is applied on each frame. The DCT coefficients of the frames are further divided into 8 frequency bins, each containing 64 DCT coefficients. As discussed before, for adaptive thresholding, each bin is categorized as either signal- or noise-dominant. The classification pertains to the average noise power associated with that particular bin. If the ith bin satisfies the following inequality:

$$\frac{1}{N}\sum_{k=1}^{N}\left|X_k^i\right|^2 \geq \sigma_n^2, \tag{2}$$

where σ_n^2 denotes the variance of the noise, X_k^i is the kth DCT coefficient of the ith frequency bin, and N (=64) is the number DCT coefficients of the bin; then the bin is characterized as signal-dominant, otherwise as noise-dominant. The signal-dominant bins are not thresholded, since it is highly possible to degrade the speech signal, especially for high SNRs. In the case of a noise-dominant frequency bin, the absolute values of the DCT coefficients are sorted in ascending order and a linear thresholding is applied:

$$\widehat{X}_k = \operatorname{sign}(X_k)\left[\max\left\{0, \left(|X_k| - \eta_j\right)\right\}\right], \tag{3}$$

where η_j is the linear threshold function obtained as

$$\eta_j = j\frac{\lambda\sigma_n N}{\sum_{k=1}^{N} k^2}, \tag{4}$$

where j is the index of sorted $|X_k|$. It is evident from (2) that, for the noise-dominant frequency bins, the average noise power added would be less than the average noise power estimated over the entire speech signal. Here, the added average noise power over any of these frequency bins is denoted as $\lambda\sigma_n$. To find a reasonable value for λ, three speech signals contaminated with white noise at 10 dB SNR are used. Using the categorization in (2) at each frequency bin, the noise dominants are identified and a value of λ is calculated by simply dividing the variance of that frequency bin by the overall noise variance. The sorted variation of λ is shown in Figure 2. It can be observed that the value of λ varies between 0.2 and 0.8 for all speech signals. Therefore, experimentally, the value of λ should be selected in this range.

3. Basics of EMD

The principle of EMD technique is to decompose any signal $s(t)$ into a set of band-limited functions $C_n(t)$, which are zero mean oscillating components, simply called the IMFs. Each IMF satisfies two basic conditions: (i) in the whole data set the number of extrema and the number of zero crossings must be the same or differ at most by one and (ii) at any point the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero [5]. The first condition is similar to the narrow-band requirement for a Gaussian process and the second condition is a local requirement induced from the global one and is necessary to ensure that the instantaneous frequency will

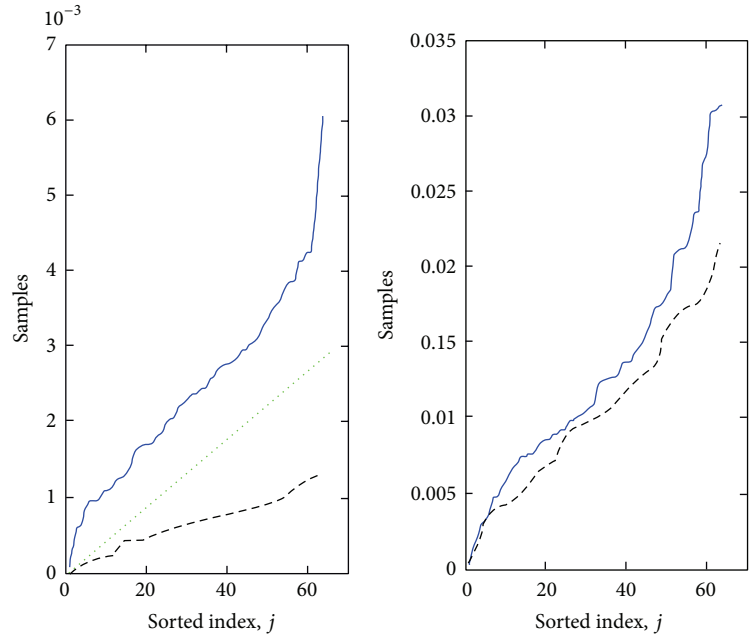

FIGURE 1: A typical (a) noise-dominant and (b) signal-dominant bin noisy frame (solid line), threshold (dotted line), and clean speech frame (dashed line).

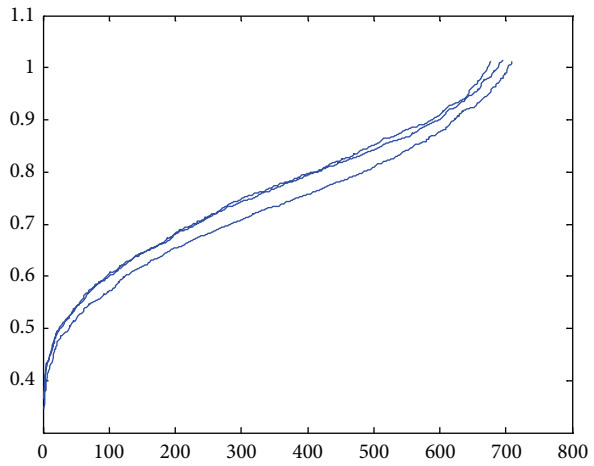

FIGURE 2: The calculated value of λ in noise-dominant frequency bins.

not have redundant fluctuations as induced by asymmetric waveforms. The name intrinsic mode function is adopted because it represents the oscillation mode in the data. With this definition, the IMF in each cycle, defined by the zero crossings, involves only one mode of oscillation; no complex riding waves are allowed [5]. IMF is not restricted to a narrow-band signal; it can be both amplitude and frequency modulated; in fact it can be nonstationary.

The idea of finding the IMFs relies on subtracting the highest oscillating components from the data with a step by step process, which is called the sifting process. Although a mathematical model has not been developed yet, different methods for computing EMD have been proposed after its introduction [6, 7]. The very first algorithm is called the sifting process. The sifting process is simple and elegant. It includes the following steps:

(1) identify the extrema (both maxima and minima of $s(t)$),

(2) generate the upper and lower envelopes ($u(t)$ and $l(t)$) by connecting the maxima and minima points by cubic spline interpolation,

(3) determine the local mean $\mu_1(t) = [u(t) + l(t)]/2$,

(4) since IMF should have zero local mean, subtract out $\mu_1(t)$ from $s(t)$ to get $h_1(t)$,

(5) check whether $h_1(t)$ is an IMF or not,

(6) if not, use $h_1(t)$ as the new data and repeat steps 1 to 6 until ending up with an IMF.

Once the first IMF $h_1(t)$ is derived, it is defined as $C_1(t) = h_1(t)$, which is the smallest temporal scale in $s(t)$. To compute the remaining IMFs, $C_1(t)$ is subtracted from the original data to get the residue signal $r_1(t)$: $r_1 = s(t) - C_1(t)$. The residue now contains the information about the components of longer periods. The sifting process will be continued until the final residue is a constant, a monotonic function, or a function with only one maximum and one minimum from which no more IMF can be derived [6]. The subsequent IMFs

and the residues are computed as

$$r_1(t) - C_2(t) = r_2(t), \ldots, r_{m-1}(t) - C_m(t) = r_m(t). \quad (5)$$

At the end of the decomposition, the data $s(t)$ will be represented as a sum of m IMF signals plus a residue signal,

$$s(t) = \sum_{i=1}^{m} C_i(t) + r_m(t). \quad (6)$$

A noisy speech signal and some selected IMF components are shown in Figure 3. It can be observed that higher order IMFs contain lower frequency oscillations than those of lower order IMFs. This is reasonable, since the sifting process is based on the idea of subtracting the component with the longest period from the data till an IMF is obtained. Therefore the first IMF will have the highest oscillating components: the components with the highest frequencies. Consequently, the higher the order of the IMF is, the lower its frequency content will be. However, the IMFs may have frequency overlaps but at any time instant the instantaneous frequencies represented by each IMF are different. This phenomenon can be well understood in Figure 4 which shows the instantaneous frequencies of the first 6 IMFs. Therefore EMD is not band pass filtering but is an effective decomposition of nonlinear and nonstationary signals in terms of their local frequency characteristics. The recent development of EMD focused on the use of ensemble EMD (EEMD) [8] and noise assisted multivariate EMD (MEMD) [9, 10] to implement the traditional univariate EMD (UEMD). The key advantage of the newly developed EMD methods is to achieve the accurate decomposition of the analyzing signal. The EEMD approach consists of sifting an ensemble of white noise-added signal and threatens the mean as the final true result. The effect of the added white noise is to provide a uniform reference frame in the time-frequency space; therefore, the added noise collates the portion of the signal of comparable scale in one IMF. A noise-assisted approach in conjunction with MEMD is also used for the computation of EMD, in order to produce localized frequency estimates at the accuracy level of instantaneous frequency [9]. The traditional EMD is prone to mode-mixing and is designed for univariate data. The noise assisted MEMD (NA-MEMD) approach utilizes the dyadic filter bank property of the MEMD providing the solution to the problem of standard EMD.

With these powerful characteristics, recent studies have shown that it is possible to successfully identify and remove a significant amount of the noise components from the IMFs of a noisy speech. Although all IMFs contain energy from both the original speech and the noise, the amount of the energy distribution is different. Since speech signals are mainly concentrated in the low and mid frequency bands, the high frequency noise components dominate the first IMFs. For instance, in case of white noise, most of the noise components are centered on the first three IMFs, while the speech signals dominate between the 3rd and 6th IMFs, as can be observed in Figure 3. Therefore, EMD makes it possible to some extent to separate the high frequency noise from the major speech components.

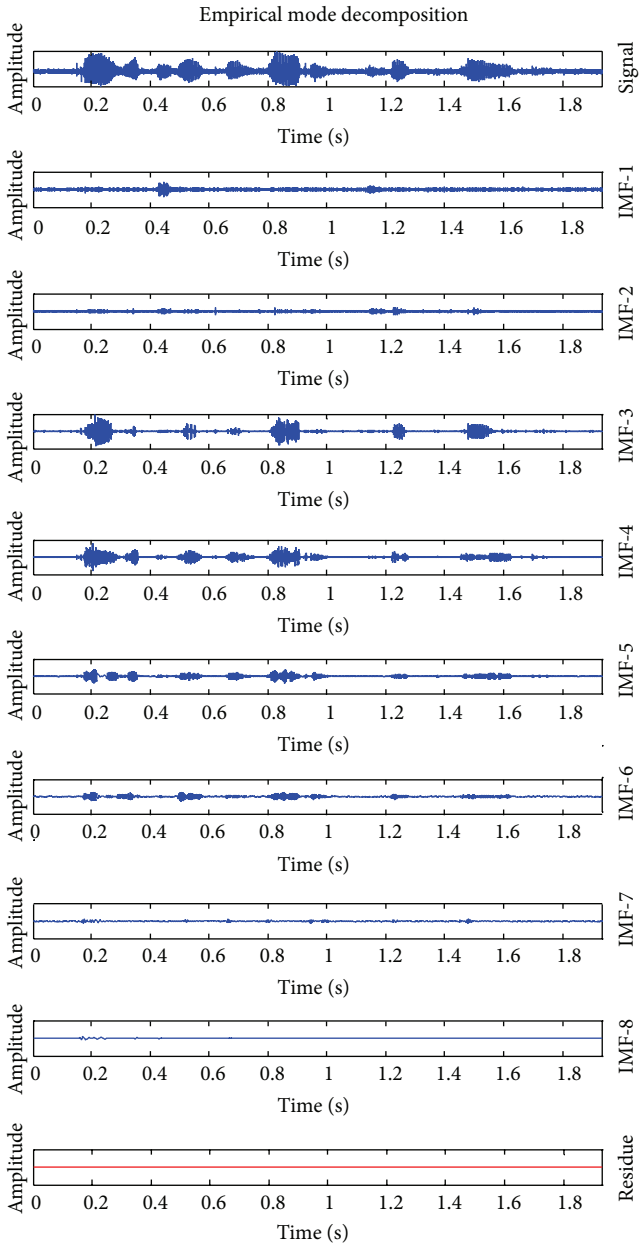

Figure 3: The illustration of EMD. A noisy speech signal at 10 dB SNR and its first 8 IMFs out of 14, plus a residue signal which can be observed to be close to a constant.

4. Proposed Hybrid Algorithm

The proposed hybrid algorithm is based on applying the frame based soft thresholding strategy [4] in two stages. The first stage includes the DCT domain soft thresholding with a subband approach in order to provide robustness to different noise types. The second stage of the algorithm consists of an EMD domain soft thresholding for further enhancement.

4.1. Subband DCT Soft Thresholding. The major problem in DCT soft thresholding algorithm given in [4] is that it is not robust to different noise types. Since all the frequency

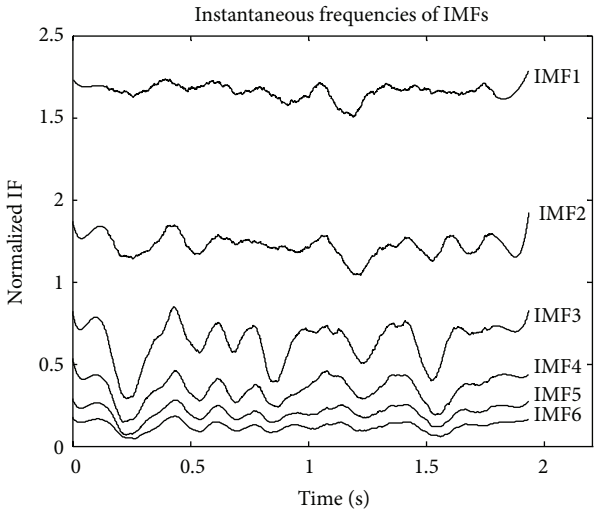

Figure 4: Instantaneous frequencies of the first 6 IMFs.

bins are processed with a unique noise variance estimated in the time domain, the algorithm is mainly applicable to white noise which has a flat spectrum. The method fails for other noise types that show different spectral distribution within the frequency bins. Therefore, it is important to have a subband approach where a specific noise variance is calculated for each frequency band. The index of the frequency bins represents the index of the subband. For instance, the first frequency subband consists of the first frequency bins of each frame. The variance of each subband is calculated through a minimum statistics approach from the frequency bins. With this subband approach, each band will have an effective bin categorization. Therefore, the algorithm will be robust to different noise types.

Apart from the subband approach, a novel strategy is introduced here for the bin categorization. The limit given in (2), which is set to noise variance, is not efficient to identify all the noise-dominant bins. Since the variance of the noisy bins will have fluctuations, there will be many noise-dominant bins which will be identified as signal-dominant. Therefore, the limit for bin categorization should have a larger value than the noise variance, in order to guarantee that all the noisy bins are thresholded. A novel limit relies on the idea that a bin can be defined as noise-dominant, if the noise power in that bin is higher than the speech power. Therefore, the limit should be set to the case where the noise and speech variances σ_n^2 and σ_s^2, respectively, are equal. The variance σ^2 of the noise contaminated speech for any frequency bin is represented as

$$\sigma^2 = \sigma_s^2 + \sigma_n^2 + 2\varphi(s,n), \tag{7}$$

where $\varphi(s,n)$ is the covariance term of signal and noise. If the signal and noise are independent, the covariance function gives zero; thus we have

$$\sigma^2 = \sigma_s^2 + \sigma_n^2. \tag{8}$$

For frame categorization (into signal- and noise-dominant frames), the threshold is considered with equal noise and speech power, and hence $\sigma^2 = 2\sigma_n^2$. Therefore, in case of equal noise and speech power, the variance of the bin is equal to $2\sigma_n^2$. The variance of a speech segment directly corresponds to its power. The equal variance of speech and noise exhibits the equilibrium contribution of speech noise power to the noisy speech frame. Hence such level of power is considered as the threshold for speech frame categorization. It is treated as the minimum power level of noise-free speech frame. Any frame with power higher than such threshold exhibits that the speech power is dominating. Otherwise, the noise power dominates the analyzing frame. That is why the limit for the categorization of the bins in (2) should be set to this value. With the proposed strategy, if

$$\frac{1}{N}\sum_{k=1}^{N}\left|x_k^i\right|^2 \geq 2\sigma_n^2, \tag{9}$$

where σ_n^2 denotes the variance of the noise for the ith subband and x_k^i is the k'th sample of the ith bin, then this bin is categorized as signal-dominant, otherwise as noise-dominant. Noise-dominant frequency bins are thresholded as in (3). The optimum value for λ is defined here.

4.2. Optimum Value of λ. The soft thresholding algorithm can further be improved by defining an optimum value for λ. As we discussed, it is better to have a higher λ for low SNRs and a lower value for high SNR input signals. This dependency of λ on the input SNR can be better observed in Figure 5, which shows the effect of λ on the SNR improvement results at different input SNRs. Therefore, the optimum value of λ can be related with an estimated value of the input SNR. The input SNR can be estimated as

$$\mathrm{SNR}_{\mathrm{input}} = 10\log\left(\frac{\sigma_s^2}{\sigma_n^2}\right), \tag{10}$$

where σ_s^2 denotes the variance of the speech signal and σ_n^2 denotes the variance of the noise signal within the whole noisy mixture. From the independency of the speech and noise, σ_s^2 is determined as $\sigma_s^2 = \sigma^2 - \sigma_n^2$. Extensive computer simulations are performed to determine the values of the parameters α_0 ($0.6 < \alpha_0 < 0.8$) and α_1 ($0.01 < \alpha_1 < 0.03$); hence the optimum value of λ is obtained as

$$\lambda_{\mathrm{opt}} = \alpha_0 - \alpha_1\left(\mathrm{SNR}_{\mathrm{input}}\right). \tag{11}$$

4.3. EMD Domain Soft Thresholding. A significant amount of the noise components is reduced in the first stage. However, there is still remaining noise from both the thresholded noise-dominant and unthresholded signal-dominant frequency bins. It is possible to extract a considerable amount of this residual noise in the second stage from the IMFs of the enhanced speech. Due to the frequency characteristics of EMD, the noise and speech signals mostly dominate in different IMFs. Mainly, the high frequency noise components centre in the first few ones. Therefore a noticeable amount of high frequency noise components that were in signal-dominant bins in the first stage can be identified from the first

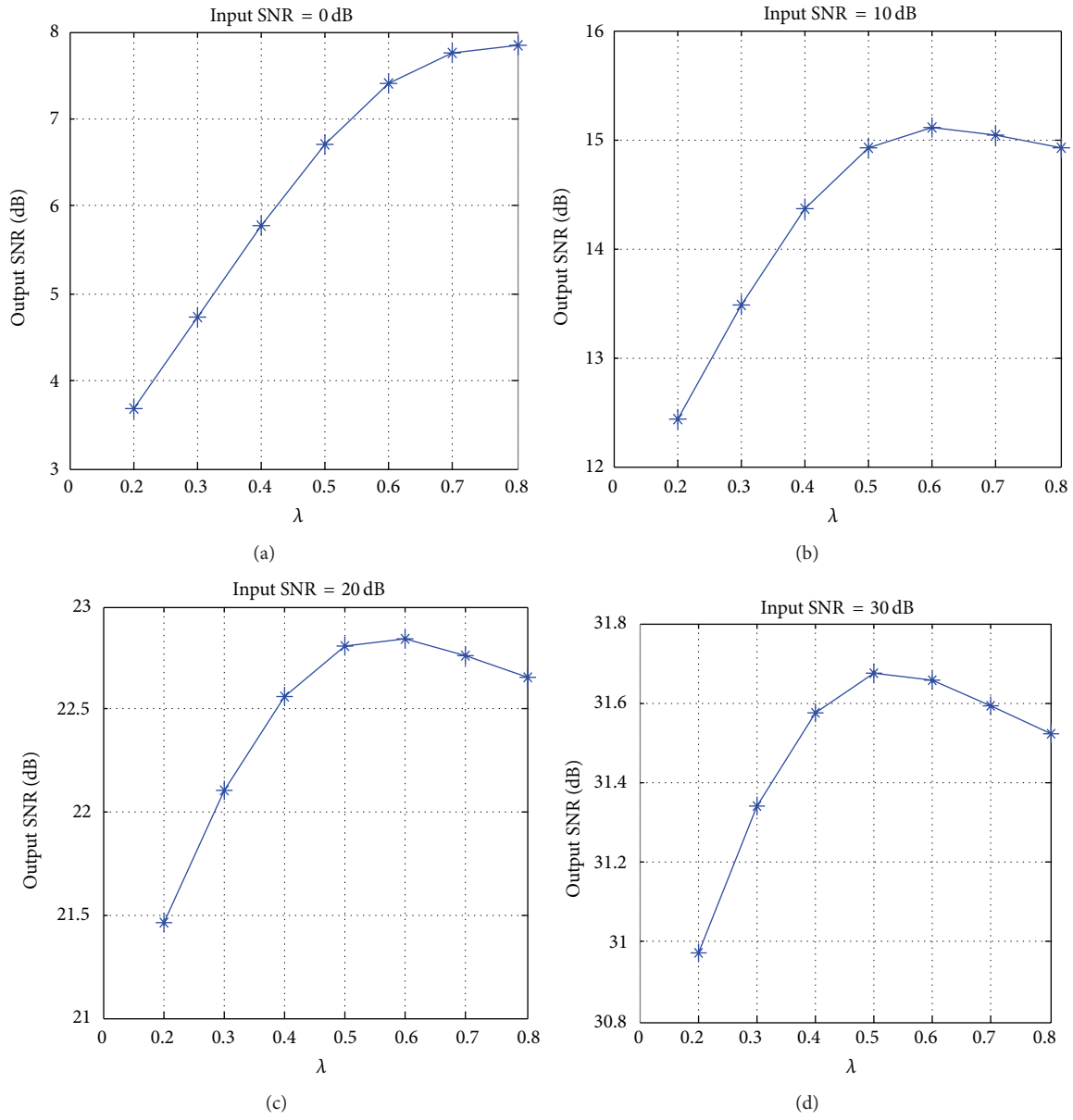

FIGURE 5: The effect of λ on the SNR improvement results in different input SNRs.

IMFs of the enhanced speech. Similarly, the lower frequency noise signals can be identified from the later IMFs.

The IMFs are in time domain and may have frequency overlaps. However, at any time instant, the instantaneous frequency represented by each IMF is different. That is why, although the IMFs are in time domain, they have spectral difference at time instances. Therefore, the DCT soft thresholding algorithm can be applied to the IMFs as given in [11]. First, the EMD is applied to the enhanced speech. The obtained IMFs are divided into 4 ms frames, thus each having 64 data for a 16 kHz sampling frequency. Due to the decomposition characteristics, the IMFs differ in terms of noise and speech energy distribution. Therefore the specific noise variance of each IMF is estimated from the speechless parts. As, in the DCT bin categorization case, the frames are characterized as either signal- or noise-dominant frames with the novel categorization limit given in (9). The noise-dominant frames are thresholded using (3), while the signal-dominant frames are not.

5. Experimental Results and Discussion

To illustrate the effectiveness of the EMD based hybrid algorithm, extensive computer simulations were conducted with 10 male and 10 female utterances sampled at 16 kHz, randomly selected from the TIMIT database. The clean speech samples were corrupted with weighted noise from the NOISEX database in order to obtain the noisy speech samples. To illustrate the robustness of the univariate EMD

Table 1: Comparison of the SNR, AvgSegSNR, and PESQ improvements of different denoising methods for a high range of SNR values (white noise).

A				
Input SNR (dB)	Output SNR (dB)			
	WP [3]	DCT [11]	Soft DCT [4]	$U_{\text{EMD}}(\lambda_{\text{opt}})$
0	4.86	6.44	6.42	**7.91**
5	8.92	10.03	10.07	**11.22**
10	12.52	13.61	13.95	**14.98**
15	15.64	17.26	18.05	**18.87**
25	20.95	24.92	26.78	**27.18**
30	23.27	28.86	31.32	**31.51**
B				
Input AvgSegSNR (dB)	Output AvgSegSNR (dB)			
	WP [3]	DCT [11]	Soft DCT [4]	$U_{\text{EMD}}(\lambda_{\text{opt}})$
−4.111	−1.933	−0.669	−0.317	**0.779**
−1.341	0.926	2.01	2.246	**3.166**
2.079	3.666	4.823	5.187	**6.078**
5.758	6.504	7.83	8.472	**9.294**
13.837	11.64	14.516	15.902	**16.394**
18.002	13.71	18.092	19.679	**19.998**
C				
Input SNR (dB)	PESQ			
	Input	WP [3]	DCT [11]	Soft DCT [4]	$U_{\text{EMD}}(\lambda_{\text{opt}})$
0	1.06	1.27	1.38	1.36	**1.74**
5	1.36	1.58	1.78	1.76	**2.07**
10	1.69	1.95	2.19	2.14	**2.39**
15	2.04	2.31	2.58	2.52	**2.71**
25	2.81	2.86	3.32	3.21	**3.32**
30	3.21	3.06	3.64	3.53	**3.66**

(U_{EMD}) scheme to different noise types, white, pink, and high frequency (HF) radio channel noise samples have been used. For evaluating the performance of the method, overall and average segmental SNR improvements as well as objective speech quality results were used. The quality of the enhanced signals has been measured with the perceptual evaluation of speech quality (PESQ).

Figures 6(a) and 6(b) show the spectrogram for the male clean speech "do not ask me to carry an oily rag like that" from the TIMIT database and the corresponding noisy speech corrupted with white noise at 10 dB SNR. The spectrogram of the enhanced speech after the first stage of the algorithm is illustrated in Figure 6(c). It can be observed that, with the first stage, there is a reasonable enhancement in the noisy speech signal. Although the noise components are effectively removed for a wide range of frequencies, the remaining noise in the enhanced speech can be observed. With the second stage, we could manage to efficiently remove the remaining noise. By this way, not only do we have a significant improvement in the SNR but we also get rid of the irritating residual noise. The spectrogram of the overall enhanced signal in Figure 6(d) illustrates the effectiveness of the proposed method. Figure 7 shows the corresponding waveforms.

Similar to the DCT soft thresholding, the algorithm can be applied for a wide range of SNRs. Since the signal-dominant frames are never thresholded, there is still significant improvement even in case of high SNRs where even the most proposed U_{EMD} based methods fail to hold on to the input SNR. The average results of the computer simulations for 10 male and 10 female utterances for a wide range of SNR values with a comparison of different denoising methods are listed in Table 1(A) for white noise. The superiority of the U_{EMD} scheme can be well observed in this table.

It can be observed that, for all SNR levels, the proposed U_{EMD} method gives significantly better results. Although SNR improvement is a good measure for quantifying performance, it has little perceptual meaning and is therefore not a good measure for speech quality [12]. Instead, the average segmental SNR (AvgSegSNR) is relatively a better measure.

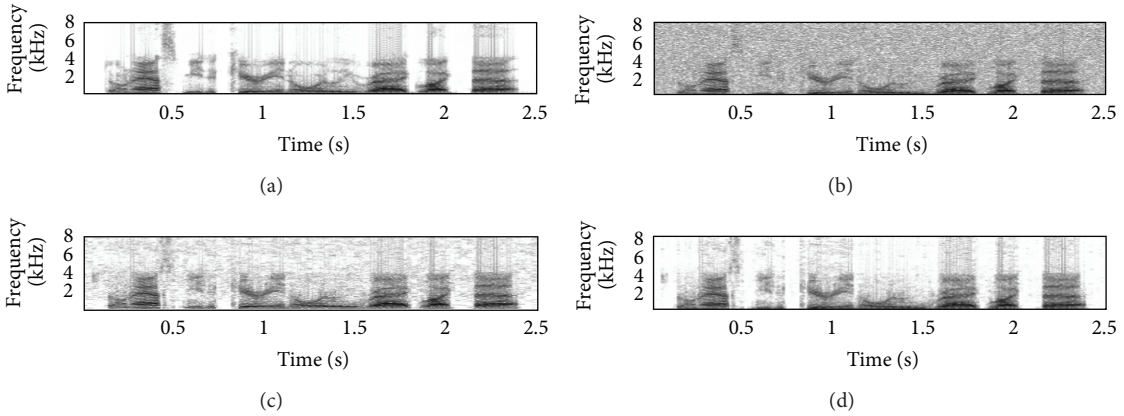

Figure 6: Spectrogram of (a) the clean speech, (b) the noisy speech corrupted with white noise at 10 dB SNR, (c) the recovered speech after soft thresholding with subband DCT, and (d) the overall recovered speech of the U_{EMD} based method.

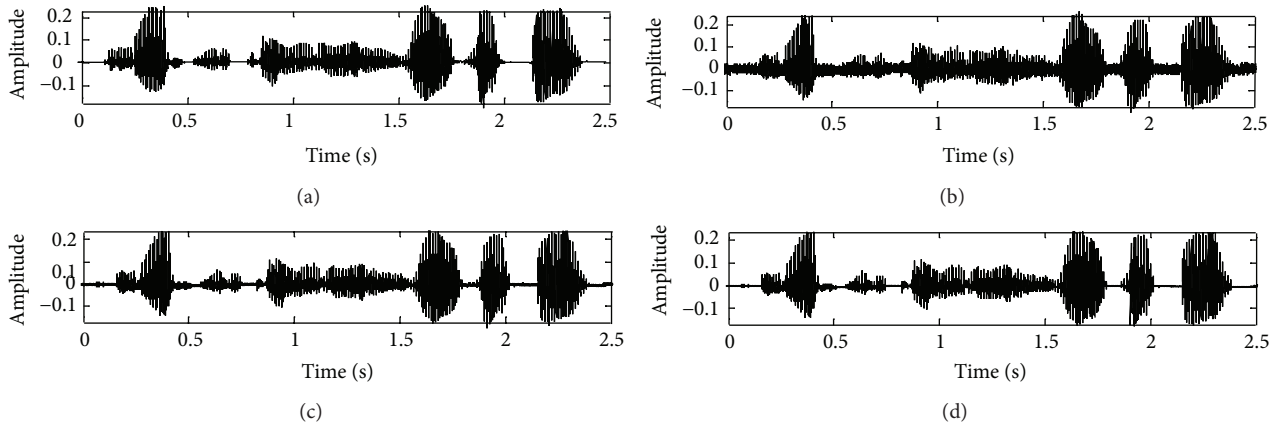

Figure 7: Waveform of (a) the clean speech, (b) the noisy speech corrupted with white noise at 10 dB SNR, (c) the recovered speech after soft thresholding with subband DCT, and (d) the overall recovered speech of the U_{EMD} method.

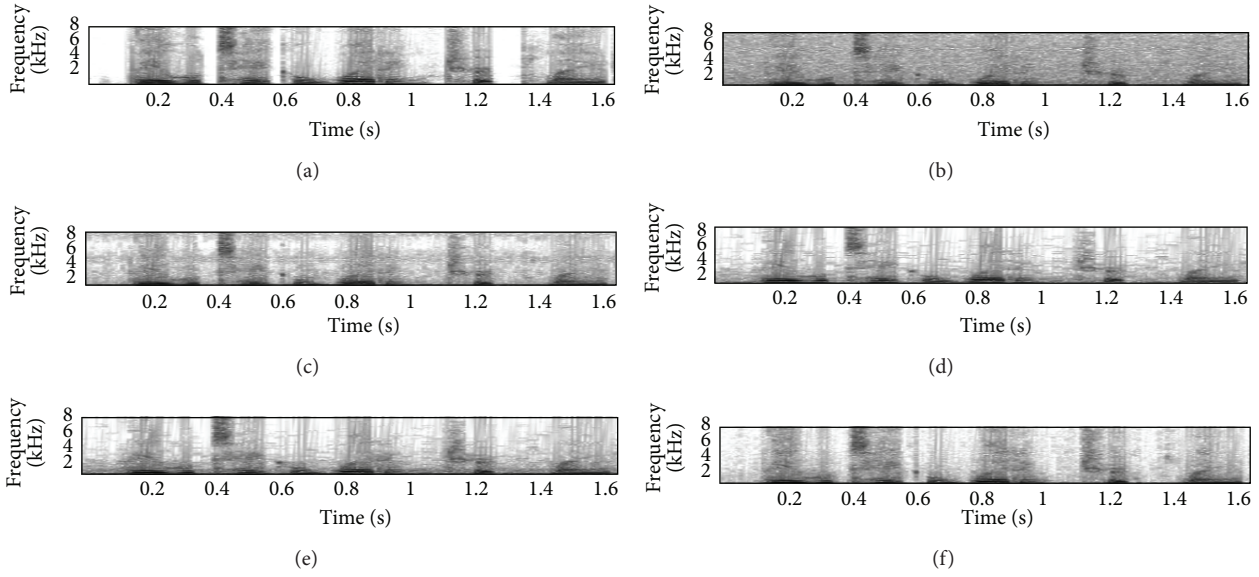

Figure 8: The spectrogram of (a) clean speech, (b) noisy mixture at 10 dB (pink noise), and enhanced speech with (c) wavelet packets thresholding [3], (d) DCT hard thresholding [11], (e) DCT soft thresholding, and (f) proposed U_{EMD} based hybrid method (λ_{opt}).

Table 2: Comparison of overall SNR, average segmental SNR (AvgSegSNR), and PESQ improvements of different denoising methods for pink and HF channel noise.

Input SNR (dB)	Output SNR (dB)					
	0	5	10	15	25	30
PINK						
WP [3]	2.57	7.19	11.66	15.81	22.69	25.20
DCT [11]	2.12	6.78	11.35	15.81	24.58	28.98
S. DCT [4]	1.41	5.98	10.73	15.51	25.24	30.13
U_{EMD}	**4.51**	**8.27**	**12.41**	**16.81**	**26.01**	**30.44**
HF						
WP [3]	1.96	6.72	11.63	16.45	24.24	26.47
DCT [11]	3.59	7.84	11.88	15.94	24.11	28.21
S. DCT [4]	0.94	5.38	10.08	14.92	24.70	29.61
U_{EMD}	**4.92**	**8.95**	**12.96**	**17.14**	**26.21**	**30.84**
In. AvgSegSNR (dB)	Output AvgSegSNR (dB)					
	−4.047	−1.124	2.256	5.959	14.059	18.188
PINK						
WP [3]	−2.983	0.017	3.196	6.373	12.354	14.904
DCT [11]	−3.149	−0.162	3.057	6.435	13.695	17.526
S. DCT [4]	−3.598	−0.649	2.704	6.328	14.292	18.341
U_{EMD}	**−1.594**	**0.927**	**3.538**	**7.074**	**15.088**	**18.834**
In. AvgSegSNR (dB)	Output AvgSegSNR (dB)					
	−4.162	−1.287	2.079	5.781	13.906	18.049
HF						
WP [3]	−3.574	−0.476	3.006	6.685	13.441	16.017
DCT [11]	−2.683	0.218	3.219	6.411	13.319	17.007
S. DCT [4]	−4.171	−1.349	1.948	5.599	13.603	17.725
U_{EMD}	**−1.234**	**1.526**	**4.416**	**7.671**	**15.342**	**19.239**
Input SNR (dB)	PESQ					
	0	5	10	15	25	30
PINK						
Input	1.33	1.68	2.06	2.43	3.22	3.61
WP [3]	1.64	2.04	2.38	2.66	3.15	3.32
DCT [11]	1.91	2.27	2.59	2.93	3.51	3.77
S. DCT [4]	1.85	2.17	2.51	2.84	3.50	3.79
U_{EMD}	**1.93**	**2.29**	**2.62**	**2.95**	**3.55**	**3.83**
HF						
Input	1.58	1.84	2.14	2.44	3.15	3.49
WP [3]	1.67	1.87	2.12	2.45	3.15	3.47
DCT [11]	1.60	1.83	2.13	2.46	3.11	3.37
S. DCT [4]	1.49	1.62	1.84	2.14	2.94	3.32
U_{EMD}	**1.61**	**1.96**	**2.32**	**2.66**	**3.34**	**3.65**

The results for the AvgSegSNR are listed in Table 1(B), which still proves the superiority of the U_{EMD} based algorithm in all SNRs. In order to have a better idea about the perceptual quality of the enhanced speech signals, PESQ has been used. Recently regarded as the best algorithm for estimation of the results of a subjective test, PESQ returns a score between −0.5 and 4.5, with higher scores indicating better quality. The results of the PESQ simulation results can be observed in Table 1(C). It can be observed that the U_{EMD} based algorithm is still more effective in terms of perceptual quality than the other methods.

In order to prove the robustness of the algorithm to different noise types, extensive computer simulations were conducted with pink and high frequency (HF) channel noise.

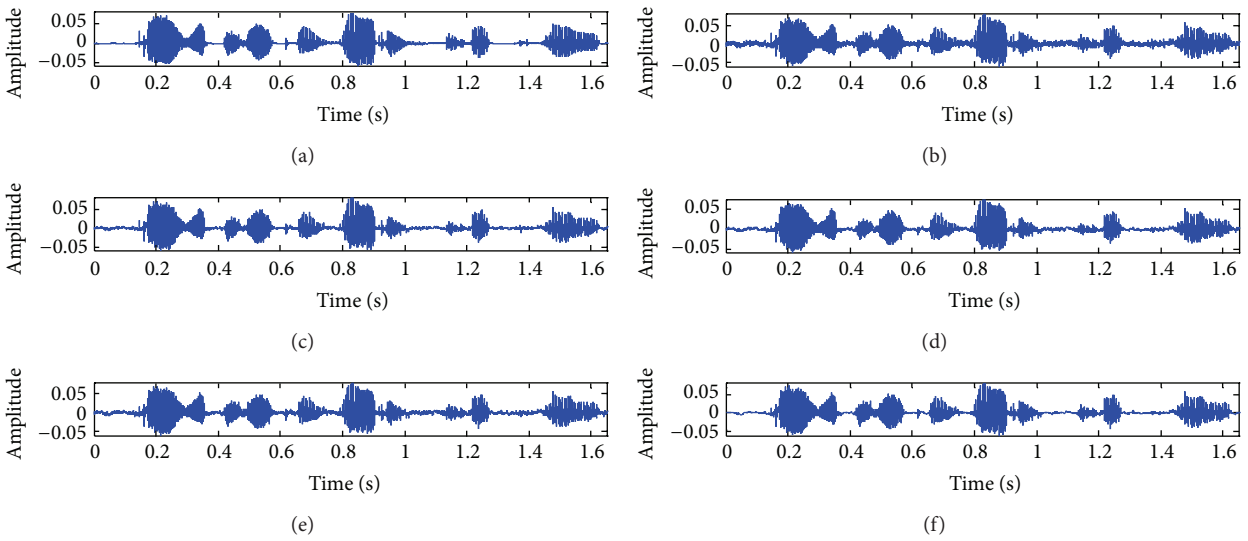

Figure 9: The waveform of (a) clean speech, (b) noisy mixture at 10 dB (pink noise), and enhanced speech with (c) wavelet packets thresholding [3], (d) DCT hard thresholding [11], (e) DCT soft thresholding, and (f) U_{EMD} based hybrid method (λ_{opt}).

The average results of computer simulations for 10 male and 10 female utterances for overall SNR, average segmental SNR, and PESQ results are listed in Table 2.

As discussed before, it can be seen that the DCT soft thresholding algorithm in [4] dramatically fails in such noise types that do not have flat spectral distribution in the frequency spectrum. Due to the subband variance approach adapted in the first stage, our proposed hybrid method is significantly robust to such noise types and highly superior to other methods. Moreover, since the signal-dominant subframes are never thresholded, the algorithm is always performing improvement in all SNR values. The EMD based soft thresholding in the second stage not only improves the SNR but also plays a critical role in removing the irritating musical noise, therefore extensively increasing the perceptual speech quality. Figures 8 and 9 show the spectrograms and waveforms of the clean speech, the noisy speech at 10 dB SNR contaminated with pink noise, and the enhanced speech signals for the female speech "they will take a wedding trip later."

The performance of U_{EMD} based speech enhancement is also compared with the methods in which the traditional EMD is computed using EEMD (E_{EMD}) [8] and MEMD (M_{EMD}) [9]. The comparative results for a wide range of SNRs obtained by three EMD methods for white noise are illustrated in Figure 10. Only the white noise is taken into consideration.

It is found that the EEMD based approach exhibits lower performance than that of the traditonal EMD for white noise, whereas a slight improvement is acheived with MEMD based implementation of standard EMD. One underlying consideration of having improved result using MEMD based approach is that the noise assisted MEMD fully uses the dyadic filter property of MEMD to implement traditional EMD. It does not suffer from the mod-mixing problem and hence the improvement of denoising results. The improvement of other EMDs (e.g., EEMD and MEMD) is more prominent in lower SNR, that is, highly noise contaminated speech signals.

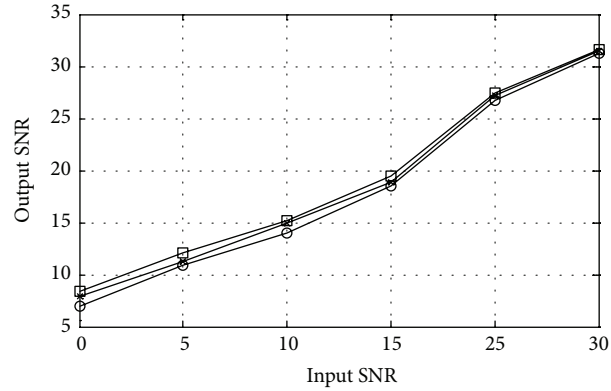

Figure 10: Performance comparison of speech enhancement using EMD based hybrid algorithm (for white noise). The EMD is implemented by univariate EMD (UEMD), enssemble EMD (EEMD), and multivariate EMD (MEMD).

6. Conclusions

In this paper, we presented a hybrid speech enhancement method based on DCT and EMD. In order to provide robustness to different noise types, a DCT soft thresholding strategy with a subband approach is proposed in the first stage of the algorithm. Furthermore, a novel limit for frame categorization was given in order to have a better identification of the noise components. In the second stage, we proposed an EMD domain soft thresholding strategy in order to remove the remaining noise components within the first stage enhanced signal.

One of the main advantages of the method is that it does not include any prior knowledge of the noise signal. Its robustness to different noise types is another significance of the method. The major drawback of the algorithm is its time cost. Since a mathematical representation is not yet given for EMD, the process takes long time. Therefore, the algorithm is not applicable to real time speech processing.

The algorithm can be further improved by adapting an optimum value calculation for the number of subbands. This can be achieved by analyzing the spectral distribution of the noise signal which can be obtained from the speechless parts of the noisy speech.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2000.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[3] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10–12, 2001.

[4] S. Salahuddin, S. Z. Al Islam, M. K. Hasan, and M. R. Khan, "Soft thresholding for DCT speech enhancement," *Electronics Letters*, vol. 38, no. 24, pp. 1605–1607, 2002.

[5] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and Hilbert spectrum for non-linear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, pp. 903–995, 1998.

[6] P. Flandrin, G. Rilling, and P. Gonçalvés, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[7] M. C. Ivan and G. B. Richard, "Empirical mode decomposition based frequency attributes," in *Proceedings of the 69th SEG Meeting*, Houston, Tex, USA, 1999.

[8] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[9] D. P. Madic, N. U. Rehman, Z. Wu, and N. E. Huang, "Empirical mode decomposition based time-frequency analysis of multivariate signals: the power of adaptive data analysis," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 74–86, 2013.

[10] N. U. Rehman, C. Park, N. E. Huang, and D. P. Mandic, "EMD via MEMD: multivariate noise-aided computation of standard EMD," *Advances in Adaptive Data Analysis*, vol. 5, no. 2, pp. 1–25, 2013.

[11] M. K. Hasan, M. S. A. Zilany, and M. R. Khan, "DCT speech enhancement with hard and soft thresholding criteria," *Electronics Letters*, vol. 38, no. 13, pp. 669–670, 2002.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, May 2001.