Speech Enhancement Using EMD Based Adaptive Soft-Thresholding (EMD-ADT)

Md. Ekramul Hamid[1], Somlal Das[1], Keikichi Hirose[2] and Md. Khademul Islam Molla[2]

[1]*Dept. of Computer Science and Engineering, University of Rajshahi, Bangladesh*
[2]*Dept. of Information and Communication Eng., University of Tokyo, Tokyo, Japan*
ekram_hamid@yahoo.com

Abstract

This paper presents a novel algorithm of speech enhancement using data adaptive soft-thresolding technique. The noisy speech signal is decomposed into a finite set of band limited signals called intrinsic mode functions (IMFs) using empirical mode decomposition (EMD). Each IMF is divided into fixed length subframes. On the basis of noise contamination, the subframes are classified into two groups – noise dominant and speech dominant. Only the noise dominant subframes are thresholded for noise suppression. A data adaptive threshold function is computed for individual IMF on the basis of its variance. We propose a function for optimum adaptation factor for adaptive thresholding which was previously prepared by the least squares method using the estimated input signal to noise ratio (SNR) and calculated adaptation factor to obtain maximum output SNR. Moreover, good efficiency of the algorithm is achieved by an appropriate subframe processing. After noise suppression, all the IMFs (including the residue) are summed up to reconstruct the enhanced speech signal. The experimental results illustrate that the proposed algorithm show a noticeable efficiency compared to the recently developed speech denoising methods.

Keywords: *Adaptive thresolding, adaptation factor, empirical mode decomposition, speech enhancement*

1. Introduction

The background noise degrades the quality and intelligibility of the speech signals resulting in a severe drop in performance of speech related applications. There are different types of noise signals which affect the quality of the original speech. It may be a wide-band noise in the form of a white or colored noise, a periodic signal such as in hum noise, room reverberations etc. It is also possible that the speech signal may be simultaneously attacked by more than one noise source. The most common type of noise in time series analysis and signal processing is the white noise. Although this work is mainly concerned with white noise, the pink and the high frequency channel noise are also used in order to illustrate the performance of the proposed algorithm. Many of the existing speech enhancement algorithms suffer from the residual noise problem which is often referred to as the musical noise [1]-[5]. In a single channel speech enhancement method, the residual noise is a usual issue. The reported algorithms are mainly concerned with minimizing such effects. Fourier Transform (FT) and Wavelet Transform (WT) are dominating methods widely used in speech processing algorithms. However, both suffer to analyze non-stationary signals like speech. The FT is a powerful tool for stationary signals. Whereas, wavelet is relatively more suitable for non-stationary signal analysis; however, it depends on the basis wavelet. Therefore, a tool for analyzing non-stationary signal is highly desirable [6, 7].

In spectral domain, it is easier to remove the noise components from a frequency band where only noise is present. But in a frequency band where both speech and noise components are present, in such a case; it is difficult remove to the noise without degrading the speech signal. So the algorithm on noise suppression implemented in time domain is more appropriate for speech enhancement. Moreover, Thresholding is a widely used method in signal denoising [8]-[11]. The idea of thresholding is to determine an effective threshold value and to apply different subtraction on the basis of that threshold in the segmented regions. Hard thresholding sets any coefficient less than or equal to the threshold value to zero. The soft thresholding takes the risk of degrading the quality of the speech signal in order to remove the noise components [8, 9]. One of the major drawbacks of these kinds of processes is the degradation of the speech signal, especially with the signals of high signal-to-noise ratios (SNR). In order to minimize the degradation of the original speech components, a modified soft-thresholding strategy is proposed by Salahuddin in [8]. It is a powerful technique for removing noise components from the noisy signal while paying attention on the original speech. A speech enhancement method in discrete cosine transform (DCT) domain using hard and soft thresholding criteria is proposed by Hasan [9]. The author estimates the high frequency region of the DCT coefficients to obtain the critical threshold parameter. The results show that the method is more effective for a wide range of SNRs. But the unpleasant musical noise is introduced in most of the existing soft-thresholding algorithms which hampers the performance of speech enhancement.

In this paper, an adaptive thresholding algorithm is introduced on the basis of empirical mode decomposition (EMD) which is developed by Huang [12] to analyze non-linear and non-stationary signals. The EMD represents any signal into a finite set of AM-FM basis functions called intrinsic mode functions (IMFs). In our previous study, Hamid [6, 13] proposed speech enhancement algorithm with estimating the degree of noise (DON) related to the input SNR. Its main drawback is that DON is estimated on the basis of pitch period over the voiced frame only and the enhanced speech degraded in high SNRs. In this research, we estimate input SNR as well as DON by estimating SNRs of clean and noisy speech (both voiced and unvoiced) signals. The observed speech variance is calculated in subframe basis and sorted in ascending order, and then the noise variance is considered from the beginning parts of the sorted array. Moreover, it is found that each IMF has different noise and speech energy and hence the variances of speech and noise are changed for various IMFs. Then the proposed subframes in each IMF are classified either as noise dominant or speech dominant on the basis of noise contamination which also minimizes the misclassification problem of frames. Therefore, an adaptive threshold function is estimated for the individual IMF and only noise dominant subframes are thresholded to obtain higher degree of speech enhancement. It is experimentally observed that the better speech enhancement is achieved for an optimum adaptation factor. For that we derive a function which was previously prepared, using the least squares method, from the estimated input SNRs and values of adaptation factor to obtain maximum output SNRs, is used to compute the adaptive threshold [10]. The derived optimum value of adaptation factor improves the performance of proposed EMD-ADT method. The speech enhancement performances are illustrated using the proposed adaptive thresholding approach and other recent algorithms.

2. Adaptive Thresholding with EMD

Empirical mode decomposition (EMD) is a fully data-driven method decomposing any signal into sub-bands in time domain. Its basis functions named IMFs are estimated via an iterative procedure called sifting without any predefined basis in contrast to Fourier

and wavelet transform. The principle of this basis construction is based on the physical time scales that characterize the oscillations of the phenomena. These IMFs basically are acting like a filtering process from higher frequencies to lower frequencies but with self-adaptive time varying filters. They are of the same length as the original signal and preserve the frequency variations with time. Each IMF must satisfy two properties: (i) the number of extrema and the number of zero crossings are either equal or differ by one; (ii) the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. After completing EMD, any signal $x(t)$ can be represented as: $x(t) = \sum_{b}^{B} c_b(t) + \zeta_B(t)$, a decomposition of the data into B-empirical modes (IMFs) are achieved, where $c_b(t)$ is the bth IMF and $\zeta_B(t)$ is the final residue.

The completeness of EMD implies that the original signal can be reconstructed without any loss of data by simply adding up the IMFs up to the residue. Thus, the IMFs can be viewed as linear components of the original or source signal. The IMFs of a speech signal is shown in Figure 1.

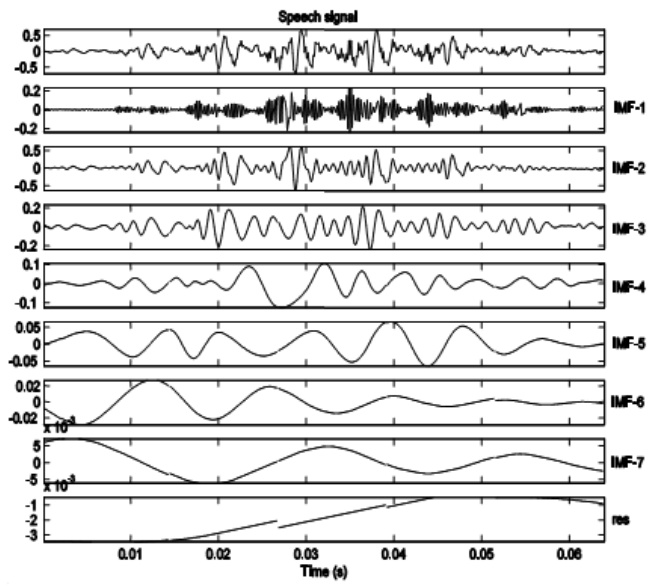

Figure 1. The IMFs of the Speech Signal obtained by EMD

2.1. Denoising by Soft-thresholding

Soft-thresholding strategy proposed by Salahuddin in [8] is a powerful technique of speech enhancement for a wide range of input SNRs. It thresholded only the noise-dominant frames and kept remain the same in case of the signal-dominant frames. The misclassification of frames is a major drawback that causes musical noise [7]. All the frames are processed with a unique noise variance estimated globally from the input speech. Many noise-dominant frames can be identified as signal-dominant due to the fluctuations in the noise variance of the frames when noise energy distribution is not uniform over the speech. This method is mainly appropriate for white noise which has a flat spectrum but not applicable to the color or real world noise with fluctuating spectra. As a result of misclassification of frames, the remaining noise components from both the noise and signal-dominant frames will generate the musical noise.

The drawbacks of traditional soft-thresholding algorithms are significantly reduced by the proposed EMD based adaptive thresholding technique. The frame classification criteria described in [8] is modified here. The soft thresholding is applied on each IMF. It is known that the thresholding function is dependent on the signal (speech) and noise variances of each IMF. The signal and noise variances are computed for individual IMF. Then soft thresholding technique is applied on each subframe of each IMF on the basis of computed variances. The threshold function is computed for individual IMF and hence such thresholding technique is termed as adaptive thresholding. We calculate the noise variance of speech from its silent part of the observed speech signals. For that, each IMFs is divided into frames of duration of 20 ms. The global noise variance $\sigma_{n,i}^2$ is calculated from the silent part of the ith IMF. In order to remove noise from the ith IMF, each frame is further subdivided into subframes of duration of 4ms. Then the subframes are classified as either a speech dominant or a noise dominant based on the noise variance $\sigma_{n,i}^2$ [10]. The proposed adaptive thresholding technique provides an effective boundary for the subframe classification. The soft thresholding is carried out on each subframe of each IMF adaptively. After properly suppression of noise using soft-thresholding, all the IMFs are summed up to get the enhanced speech signal.

2.2. Subframe Classification

The classification of the subframes plays an important role in the adaptive thresholding algorithm. The performance of this algorithm depends on the correct classification of the subframes. It makes the algorithm to be applicable for a wide range of SNRs. Due to the decomposition of noisy speech, the variance of the frames as well as subframes of each IMF will be more fluctuating than that of the noisy speech frames. Therefore, the separate noise variance of each IMF is effective for better denoising. It is required to define a sufficiently higher boundary value for the subframe classification to guarantee that all the noisy subframes are thresholded. A novel boundary relies on the idea that a subframe can be defined as a noise-dominant, if the noise power is higher than the power of the observed signal within that subframe. The boundary is set to the case where the noise and speech variances are equal. Hence, generally for any frame, we can write

$$\sigma^2 = \sigma_{s+n}^2 \tag{1}$$

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

where, σ_s^2 and σ_n^2 denote to the speech and noise variance of a frame. Since, speech and noise are independent, the covariance between the two will be zero and thus we have,

$$\sigma^2 = \sigma_s^2 + \sigma_n^2 . \tag{3}$$

To properly classify the subframes as speech dominant and noise dominant, the threshold point is selected at which the speech and noise variances are equal. Then the signal variance (at threshold point) can be written as:

$$\sigma^2 = 2\sigma_n^2 \tag{4}$$

Therefore, in case of equal noise and speech power and with the assumption of independency, the variance of a frame is equal to twice the noise variance of that frame. The classification condition of rth subframe of ith IMF defined as:

$$\varphi_i^{(r)} \geq 2\sigma_{n,i}^2 . \tag{5}$$

The average power of subframe, r of ith IMF is calculated as:

$$\varphi_i^{(r)} = \frac{1}{Q}\sum_{q=1}^{Q}\left|Y_{q,i}^{(r)}\right|^2 \tag{6}$$

where, Q is the sample length of the subframe (here Q=64 for 16kHz sampling frequency), $Y_{q,i}^{(r)}$ denotes the samples of rth subframe of the ith IMF, and $\sigma_{n,i}^2$ denotes the globally estimated noise variance of that IMF. Then the proposed classification condition for r^{th} subframe $s_i^{(r)}(t)$ of ith IMF is expressed as:

$$s_i^{(r)}(t) = \begin{cases} s_{i(s)}^{(r)}(t), & if \quad \varphi_i^{(r)} \geq 2\sigma_{n,i}^2 \\ s_{i(n)}^{(r)}(t), & otherwise \end{cases} \tag{7}$$

where $s_{i(s)}^{(r)}(t)$ and $s_{i(n)}^{(r)}(t)$ are the classified speech and noise dominant subframes, respectively. The speech dominant subframes are not thresholded. We express the adaptive thresholding for r^{th} subframe of ith IMF as:

$$\hat{Y}_{q,i}^{(r)} = \begin{cases} Y_{q,i}^{(r)}, & if \quad \varphi_i^{(r)} \geq 2\sigma_{n,i}^2 \\ \psi_{q,i}^{(r)}, & otherwise \end{cases} \tag{8}$$

where, $\psi_{q,i}^{(r)} = sign(Y_{q,i}^{(r)})[\max\{0,(|Y_{q,i}^{(r)}| - j\gamma_i)\}]$, $\hat{Y}_{q,i}^{(r)}$ and $Y_{q,i}^{(r)}$ denote the thresholded sample and q^{th} coefficient of rth subframe of ith IMF and the multiplication ($j\gamma_i$) is the adaptive threshold function while j being the sorted index-number of $|Y_{q,i}^{(r)}|$. The threshold factor γ_i is varied adaptively for individual IMF according to its variance. An estimated value of γ_i can be obtained as:

$$\gamma_i = \frac{\lambda\sigma_{n,i}}{\sqrt{\frac{1}{Q}\sum_{q=1}^{Q} q^2}} \tag{9}$$

where, $\sigma_{n,i}^2$ is the noise variance of the ith IMF and λ is the adaptation factor defined as:

$$\lambda = \frac{\sigma_{-n}}{\sigma_n} \tag{10}$$

where σ_n^2 is the globally estimated average noise power and σ_{-n}^2 is the average noise power added to a frame.

The Equation (10) is used to calculate the value of λ. In the experiment we use 5 different speech signals (from TIMIT database) of 10dB SNR degraded by white noise which is shown in Figure 2. It can be observed from Figure 2 that the value of λ varies in between 0.35 to 0.8 for all speech signals. Therefore, the value of λ is selected in this range experimentally and discussed later in detail

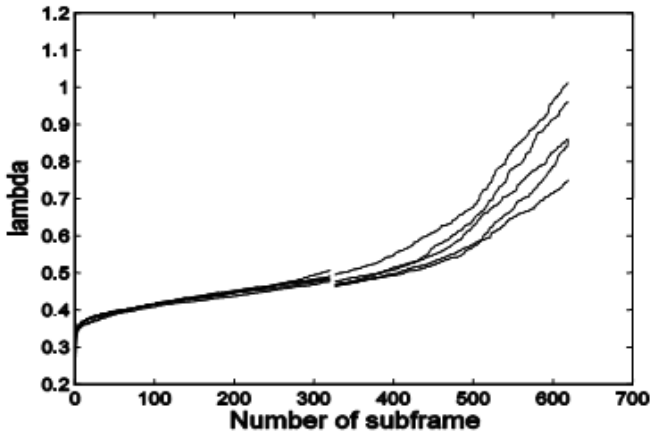

Figure 2. The Estimated Value of λ in Noise Dominant Subframes

2.3. Variance Estimation of IMFs

The variance of each IMF plays a vital role in the performance of the proposed EMD domain adaptive thresholding algorithm. In order to estimate the variance, the IMFs are divided into frames of each of 20 ms duration and the variance of each frame is stored in a variance array. The variance array is sorted in ascending order [7]. Since the silent parts will mostly have the lowest variance, the noise variance of the IMFs is selected from the beginning part of the sorted variance array. Figure 3 illustrates the variance of the frames for the first 8 IMFs of a noisy speech signal (10dB SNR). The differences in between the noise variance and the length of the silent parts of the IMFs are observed. It is clear that the noise signals are concentrated in the first 3 IMFs. The later IMFs contain mainly the speech signals, but also have significant amount of noise. An effective estimation of the noise variance of each IMF is obtained using this method. The noise components of all the IMFs are effectively removed using the proposed algorithm with the estimated variance of individual IMF.

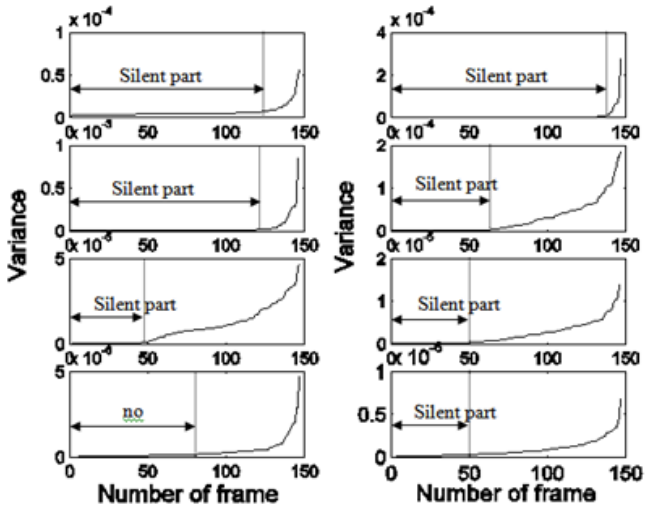

Figure 3. Sorted variance of 20 ms frames for the first 8 IMFs out of 14 (from left to right) of a noisy speech signal degraded by white noise at 10dB SNR.

2.4. Estimation of Optimum Adaptation Factor

The results of average output SNR using EMD based thresholding algorithm is used to estimate the optimum adaptation factor. In this experiment, the speech signals of English sentences uttered by 7 male and 7 female are randomly selected from TIMIT database. The output SNRs (corresponding to different input SNRs) for two different values of the adaptation factor λ are given in Table 1. It is observed that the higher value of λ is more effective at low input SNR and lower value for high input SNR. For this reason, we introduce an expression for the optimum value λ_{opt} of adaptation factor based on the input SNR of the given noisy speech. The derived optimum value of λ improves the performance of proposed EMD based method.

Table 1. Comparison of the overall output SNR of EMD based adaptive thresholding method for different values of adaptation factor. The speech is corrupted by white noise.

Input SNR (dB)	Output SNR in (dB)	
	λ=0.5	λ=0.8
0	5.98	7.51
5	10.58	11.05
10	14.56	14.46
15	18.46	18.15
20	22.51	22.14
25	26.79	26.47
30	31.29	30.97

Figure 4 illustrates the effect of λ on the output SNR corresponding to different input SNRs. Hence, optimum adaptation factor λ_{opt} is related to the estimated value of the input SNR. The estimation of input SNR of the noisy speech signal is explained below. It is observed from Fig. 4 that the maximum output SNR is achieved with a specific value of λ at different input SNRs and the results are listed in Table 2. In the experiment we have used male and female speech sentences degraded by while noise at different SNRs.

The SNR of noisy speech signal is calculated in the similar way of estimating the noise variance of the IMFs. The observed speech signal is segmented into frames of length 20ms and the variance of each frame is stored in a variance array in ascending order. The noise variance of the noisy speech is estimated from lower (silent) parts of the array. The input SNR can be estimated as:

$$SNR_{input} = 10\log_{10}\left(\frac{\sigma_{s+n}^2 - \sigma_n^{\ 2}}{\sigma_n^2}\right) = 10\log_{10}\left(\frac{\sigma_s^2}{\sigma_n^2}\right) \tag{11}$$

where σ_{s+n}^2, σ_s^2 and σ_n^2 are represent the observed, clean and noise variance, respectively. It is found that a specific value of adaptation factor corresponding to an input SNR produces the maximum output SNR. We introduce a formulation to compute the optimum value of λ for any given input SNR to achieve maximum speech enhancement as indicated in Fig. 4. The expression to calculate the optimum adaptation factor (λ_{opt}) is defined as

$$\lambda_{opt} = f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \tag{12}$$

to be fitted to the data points (x_i, y_i), i=1, 2, ….., d; where x_i and y_i are the input SNR and optimum value of λ (to obtain the maximum output SNR) respectively for the training

speech data and d (=9) represent the index of different SNRs as listed in Table 2. We experimentally found that there is a non-linear relation between the input SNR and adaptation factor, for that we choose a third degree polynomial to fit the non-linear data points with minimum stable coefficients. To obtain the coefficients, Eq. (12) can written as $Y=XA$ where $Y=[y_1, y_2, \ldots, y_d]^T$, $A=[a_0, a_1, a_2, a_3]^T$ and X is a matrix with d rows. The ith row of X can be defined as $X_i = [1 \quad x_i \quad x_i^2 \quad x_i^3]$. The matrix representation $Y=XA$ can also be written as $X^TY=X^TXA$ and hence the final expression to find the coefficient vector A is defined as:

$$A = (X^T X)^{-1} X^T Y \tag{13}$$

The Equation (13) is solved by using least square method to obtain the values of the coefficients $A=[a_0, a_1, a_2, a_3]^T$. Then the value of optimum adaptation factor λ_{opt} can easily be calculated using Equation (11). It is not necessary to use the only input SNRs listed in Table 2. The λ_{opt} can be computed for any given input SNR satisfying the least square fit method.

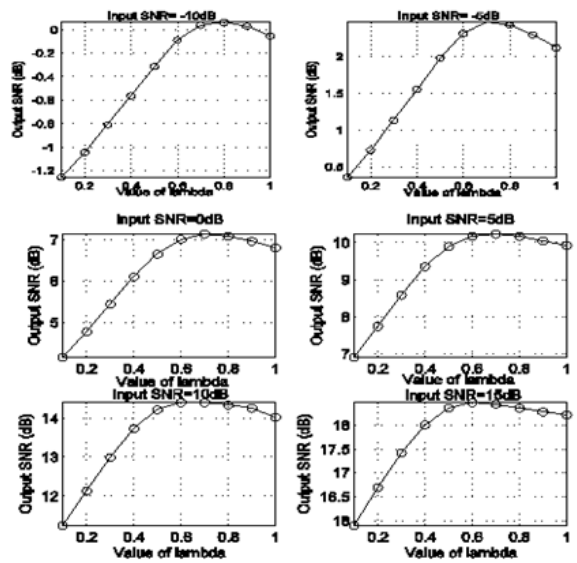

Figure 4. The Graphical Representation of Output SNR for Various Values of λ at input SNRs from –10dB to 15dB of step 5.

Table 2. The values of λ to obtain maximum output SNR for different input SNRs

Input SNR in (dB)	-10	-5	0	5	10	15	20	25	30
λ	0.80	0.73	0.71	0.7	0.62	0.60	0.54	0.51	0.5

3. Experimental Results and Discussions

The effectiveness of the proposed algorithm is tested using computer simulation with different 7 male and 7 female utterances (of English sentences) randomly selected from TIMIT database. The sampling frequency of all the speech signals is set to 16kHz. The white

noise is added to the clean speech to obtain the noisy speech signals at different noise levels. The simulation is performed over those noisy speech signals. The performance of the proposed method is presented in Figure 5. The spectrograms as well as the waveforms of the clean and noisy of 10 dB SNR are shown in Figure 5(a) and 5(b). The outputs of the proposed algorithm are illustrated in Figure 5(c) and 5(d) for λ=0.5 and λ=0.8 respectively.

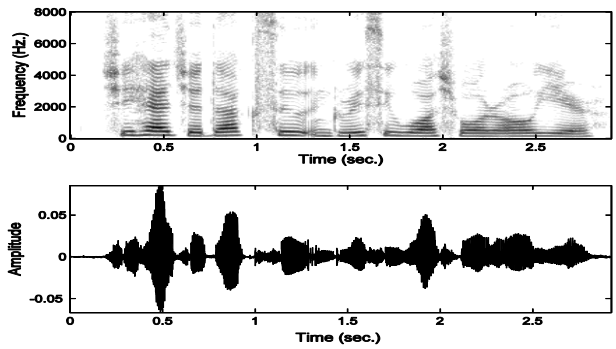

Figure 5(a). Spectrogram and Waveform of Clean Speech

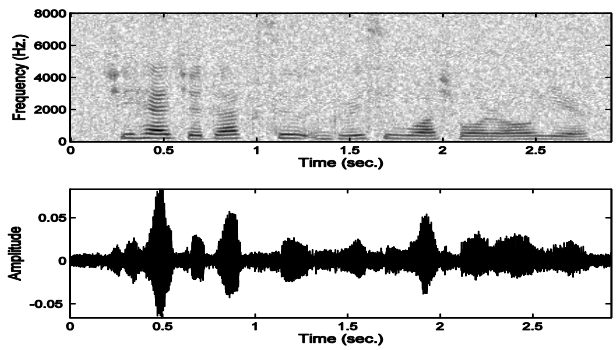

Figure 5(b). Spectrogram and Waveform of Noisy Speech (White at 10dB)

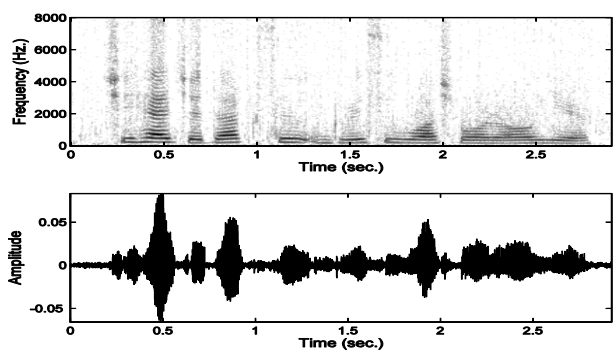

Figure 5(c). Spectrogram and Waveform of Enhanced Speech are obtained by EMD based Adaptive Thresholding with λ=0.5

It is observed in Figure 5(c) that a significant amount of noise is still remaining in the enhanced speech (with λ=0.5); whereas, some low energy speech components are degraded with λ=0.8 as shown in Figure 5(d). It is obvious that the choice of λ should be somewhere between 0.5 and 0.8 in order to have better performance. Hence we propose the optimum value of λ i.e. λ_{opt} and its result is illustrated in Figure 5(e). The Figure 5(f) shows the spectrogram and waveform of the enhanced speech by using DCT based soft

thresholding (SDCT) [11].

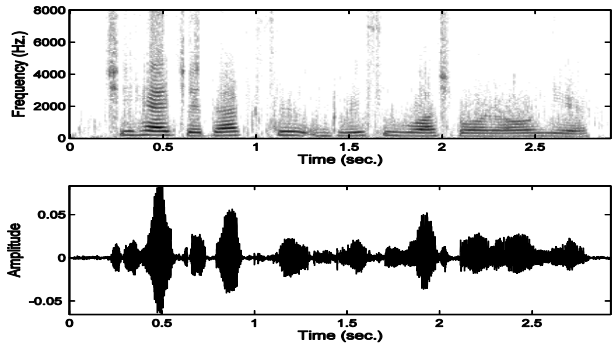

Figure 5(d). Spectrogram and Waveform of Enhanced Speech are obtained by EMD based Adaptive Thresholding with λ=0.8

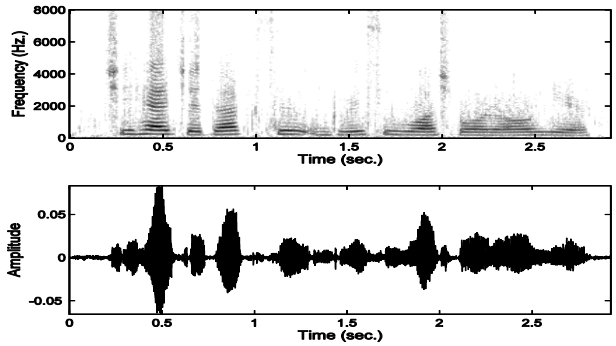

Figure 5(e). Spectrogram and Waveform of Enhanced Speech are obtained by the Proposed EMD-ADT Methods with λ_{opt}

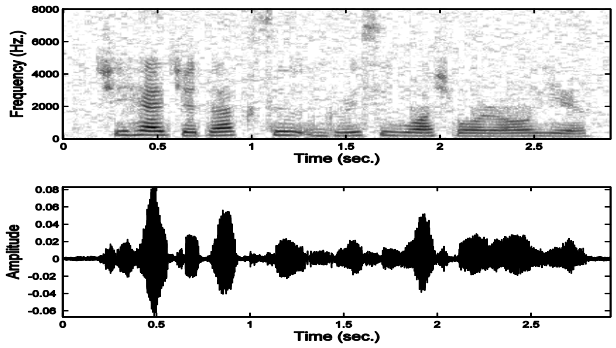

Figure 5(f). Spectrogram and Waveform of Enhanced Speech are obtained by SDCT [11] Method

The performance of the proposed EMD based adaptive thresholding technique is studied here with the optimal value of λ i.e. λ_{opt}. To prove the robustness of the proposed algorithm, pink and HF channel color noises collected from NOISEX database are also added to the clean speech to generate noisy speech signals of different SNRs. Then the experiment is carried out with these noisy signals to observe the efficiency of the algorithm in terms of overall output SNR, segmental SNR and the perceptual evaluation of speech quality (PESQ) [14]. The overall output SNR (for white noise) of the proposed

method is compared with the recently developed algorithms – two-stage speech enhancement (TSSE) method [6], hard and soft thresholding (HST) technique [9], DCT based soft thresholding (SDCT) [11] and combination of weighted noise subtraction and blind source separation (WNS+BSS) indicated as BSS [13] as illustrated in Table 3. In the case of HF channel and pink noise, the performance (in term of output SNR) is compared with WNS+BSS [13] and SDCT [11] methods as shown in Table 4. It is observed that for all types of noise, at higher SNRs (above 20dB), our previous method WNS+BSS has failed to avoid signal degradation. We can conclude from the outcomes of the Tables 1-3 and Fig. 5 that the proposed EMD-ADT results in a high speech enhancement score and clear sound without loss of speech content.

Table 3. The results of overall output SNR using different methods [6, 9, 11, 13] and a comparison with the proposed algorithm. Added noise is white at different SNRs.

Input	White noise				
	TSSE [6]	HST [9]	SDCT [11]	BSS [13]	Proposed (λ_{opt})
0dB	8.0	7.56	8.21	8.70	8.85
5dB	10.5	10.21	11.51	11.10	11.94
10dB	13.4	13.14	14.68	14.00	15.15
15dB	15.1	15.87	18.27	16.50	18.72
20dB	19.2	20.01	21.10	20.30	22.62
25dB	22.1	24.85	25.99	22.50	26.85
30dB	25.7	29.24	30.39	26.10	31.27

Table 4. The results of overall output SNR for various types of color noise at different input SNR of the EMD-ADT method and a compare with previous study WNS+BSS [13] (indicated as BSS) and SDCT [11].

Input SNR (dB)	HF channel noise			Pink noise		
	SDCT [11]	BSS [13]	Proposed (λ_{opt})	SDCT [11]	BSS [13]	Proposed (λ_{opt})
0	2.9	2.5	6.29	1.3	1.0	2.82
5	7.2	7.8	9.74	6.1	5.9	7.22
10	11.7	11.2	13.46	10.9	10.1	11.71
15	16.2	16.1	17.42	15.6	15.1	16.26
20	20.7	19.6	21.64	20.4	18.5	20.91
25	25.4	21.4	26.12	25.6	21.0	25.64
30	30.1	25.5	30.77	29.9	24.5	30.44

Although overall SNR is a good measure for quantifying performance, it has a little perceptual meaning. A better measure can be achieved by calculating average segmental SNR (ASEGSNR) over frames of short duration (20ms of frame length with 13.75ms overlapping is used here) of the speech signal that exhibits a high correlation to subjective results as compared to overall SNR [15]. Figure 6 shows the comparisons between the input and output ASEGSNR for different types of noises as a function of input SNR using the proposed algorithm with λ_{opt}. Figure 7 shows the speech enhancement performance of the proposed method and a comparison of that with SDCT in term of PESQ. The values 4 and 0 of PESQ measurement represent highest and lowest perceptual quality of the speech respectively.

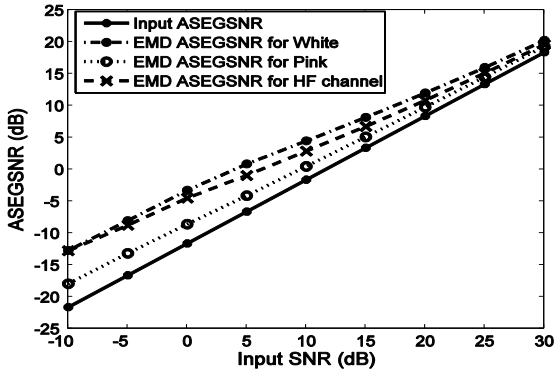

Figure 6. The Comparisons between the Output ASEGSNR and Input ASEGSNR for Different Types and Levels of Noises using EMD-ADT Method

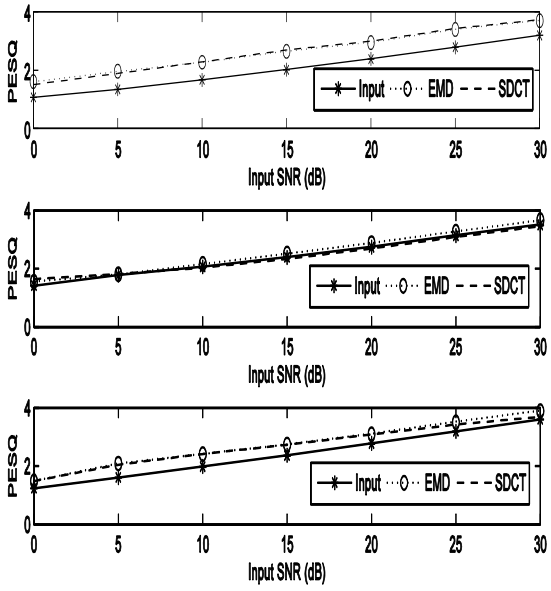

Figure 7. The Graphical Representations of Input PESQ and its Corresponding Output PESQ for a Speech Contaminated with white (Top), HF Channel (Middle) and Pink (Bottom) Noises at Different Levels by using Proposed EMD-ADT Method and SDCT [11]

The segmental SNR (SEGSNR) of individual frame (20ms of frame length with 13.75ms overlapping is used here)also highly correlated with the subjective quality of speech signal than the overall SNR [15]. Figure 8 shows the graphical illustration of input SEGSNR and its corresponding output SEGSNR obtained by applying the proposed algorithm. It is observed from Fig. 8 that the segmental output SNR is higher than that of the input SNR over all the frames of 0dB and 5dB noisy speech. Hence, the noise-dominant subframes are classified properly and the noise is removed from those subframes. With noisy speech of 10dB input SNR, the segmental SNR does not increase substantially at few frames (75 to 80) compared with the others. It is happened due to the misclassification of subframes i.e. the noise-dominant subframes are classified as signal-

dominant and kept without thresholding. Without considering those few frames, we can conclude that the classification of subframes as well as frames is performed effectively over the whole speech. Since the segmental SNR provides high correlation of subjective result, the proposed EMD based adaptive thresholding algorithm works well in this respect.

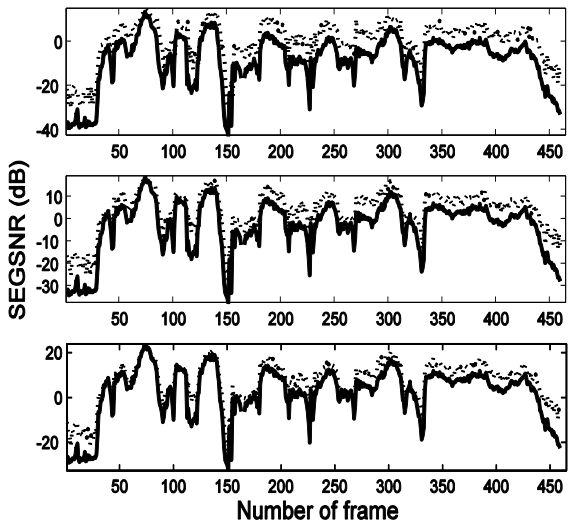

Figure 8. The Comparisons of Input SNR (solid line) and Segmental SNR (dotted line) at 0dB (Top), 5dB (Middle) and 10dB SNR (Bottom) of a Female Noisy (with white noise) Speech by using EMD-ADT Method

4. Conclusions

In this paper, a novel data adaptive algorithm is presented to effectively suppress the noise components in all frequency levels of noisy speech signal. The improvement of SNR of noise contaminated speech is achieved by removing noise using EMD based adaptive thresholding technique. An adaptation factor is introduced in the adaptive threshold function. The optimal value of adaptation factor is computed on the basis of estimated input SNR. The experimental result shows that the proposed speech enhancement algorithm works most efficiently for a wide range of input SNR. The performance of this algorithm (in terms of subjective measure, spectrogram and waveforms) is tested with the speech contaminated with white noise, pink noise and HF channel noise. However, the EMD based algorithm suffers from computational complexity and the empirical process takes long time and it is not suitable to apply for real time processing. The further research can be conducted to decrease the computational cost of EMD based methods.

References

[1] I. Cohen and B. Berdugo, “Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement”, IEEE Signal Processing Letters, vol. 9, no. 1, **(2002)** January, pp. 12-15.

[2] R. Martin, “Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors”, Proc. Int. Conf. Acoustics, Speech and Signal Processing, vol. I, **(2002)**, pp. 253–256.

[3] R. Martin, “Spectral Subtraction Based on Minimum Statistics”, Proc. EUSIPCO, **(1994)**, pp. 1182-1185.

[4] R. Martin, “Statistical Methods for the Enhancement of Noisy Speech”, Proc. IWAENC2003, **(2003)**, pp. 1-6.

[5] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands", Proc. EUROSPEECH, **(1995)**, pp. 1513-1516.

[6] M. E. Hamid and T. Fukabayashi, "A Two-Stage Method for Single-Channel Speech Enhancement", IEICE Trans. Fundamentals, vol. E89-A. 4, **(2006)** April, pp. 1058-1068.

[7] E. Deger, M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan, "Speech Enhancement using Soft Thresholding with DCT-EMD based Hybrid Algorithm", Proc. EUSIPCO, **(2007)** September.

[8] S. Salahuddin, et. al., "Soft thresholding for DCT speech enhancement", Electronics Letters, vol. 38, **(2002)**.

[9] M. K. Hasan, M. S. A. Salahuddin, M. R. Khan, "DCT speech enhancement with hard and soft thresholding criteria", Electronics Letters, vol. 38, no. 13, **(2002)**.

[10] S. Das, M. E. Hamid, K. Hirose and M. K. I. Molla, "Single-Channel Speech Enhancement by NWNS and EMD", Signal Processing: An International Journal (SPIJ), vol. 3, no. 5, **(2010)** December, pp. 279-291.

[11] D. L. Donoho, "De-noising by soft thresholding", IEEE Trans. Inf. Theory, vol. 41, **(1995)**, pp. 613-627.

[12] N. E. Huang et. al., "The empirical mode decomposition and Hilbert spectrum for non-linear and non-stationary time series analysis", Proc. Roy. Soc. London A, vol. 454, **(1998)**, pp. 903-995.

[13] M. E. Hamid, K. Ogawa and T. Fukabayashi, "Improved signal-channel noise reduction method of speech by blind source separation", Acoust. Sci. & Tech., vol. 28, no. 3, **(2007)**.

[14] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), a new method for speech quality assessment of telephone networks and codecs", Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing, vol. 2, **(2001)**, pp. 749-752.

[15] S. Quackenbush, T. Barnwell and M. Clements, "Objective Measures for Speech Quality Testing", Prentice-Hall, **(1988)**.

Authors

Md. Ekramul Hamid received his B.Sc and M.Sc degree from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh. After that he obtained the Masters of Computer Science from Pune University, India. He received his PhD degree from Shizuoka University, Japan. During 1997-2000, he was a lecturer in the Department of Computer Science and Engineering, Rajshahi University. Dr. Hamid was working as Assistant Professor at the King Khalid University, Abha, KSA from 2009 to 2011. He is currently working as an Associate Professor and Chairman in the Department of Computer Science and Engineering, Rajshahi University. His research interests include Digital Signal Processing, Analysis and synthesis of speech signal, Speech Enhancement and Image Processing.

Somlal Das received his B.Sc and M.Sc degree from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh. During 1998-2001, he was a lecturer in the Department of Computer Science and Engineering, Rajshahi University. He is currently working as an Associate Professor in the same Department. His research interests include Digital Signal Processing and Speech Enhancement.

Keikichi Hirose received the B.E. degree in electrical engineering in 1972, and the Ph.D. degree in electronic engineering in 1977, respectively, from the University of Tokyo, Tokyo, Japan. In 1977, he joined the University of Tokyo as a Lecturer in the Department of Electrical Engineering, and, in 1994, became a Professor in the Dept. of Electronic Engineering. From 1996, he was a Professor at the Graduate School of Engineering, Department of Information and Communication Engineering, University of Tokyo. On April 1, 1999, he moved to the University's Graduate School of Frontier Sciences (Department of Frontier Informatics), and again moved to Graduate School of Information Science and Technology (Department of Information and Communication Engineering) on October 1, 2004. From March 1987 until January 1988, he was a Visiting Scientist of the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, U.S.A. His research interests cover widely spoken language information processing. He led a project "Realization of advanced spoken language information processing from prosodic features," Scientific Research on Priority Areas, Grant in Aid on Scientific Research, Ministry of Education, Culture, Sports, Science and Technology, Japanese Government. He is a member of the Institute of Electrical and Electronics Engineers, the Acoustical Society of America, the International Speech Communication Association, the Institute of Electronics, Information and Communication Engineers, the Acoustical Society of Japan, and other professional organizations.

Md. Khademul Islam Molla received B.Sc. and M.Sc. degrees in Electronics and Computer Science from Shahjalal University of Science and Technology, Bangladesh in 1995 and 1997 respectively. He joined in the same Department as a lecturer in 1997. He obtained PhD degree from the Department of Frontier Informatics under the Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan in 2006. He is serving as an Associate Professor in the Dept. of Computer Science and Engineering of the University of Rajshahi, Bangladesh from August, 2006. He was a JSPS postdoctoral fellow in the Dept. of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan from Sep, 2006 to Sep, 2008. Dr. Molla was working as postdoctoral researcher in the field seismic signal processing at the University of Alberta, Edmonton, AB, Canada from Nov, 2010 to Oct, 2011. His research interest includes speech and audio signal processing, blind source separation, statistical and environmental signal processing, biomedical signal and image processing. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).