EMD based Soft-Thresholding for Speech Enhancement

Erhan Deger[1], M. K. Islam Molla[1], Keikichi Hirose[1], Nobuaki Minematsu[2] and M. Kamrul Hasan[3]

[1]Dept. of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan
[2]Dept. of Frontier Informatics, The University of Tokyo, Tokyo, Japan
[3]Dept. of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
E-mail: [1,2]{erhan, molla, hirose, mine}@gavo.t.u-tokyo.ac.jp, [3]khasan@eee.buet.ac.bd

Abstract

This paper introduces a novel speech enhancement method based on Empirical Mode Decomposition (EMD) and soft-thresholding algorithms. A modified soft thresholding strategy is adapted to the intrinsic mode functions (IMF) of the noisy speech. Due to the characteristics of EMD, each obtained IMF of the noisy signal will have different noise and speech energy distribution, thus will have a different noise variance. Based on this specific noise variance, by applying the proposed thresholding algorithm to each IMF separately, it is possible to effectively extract the existing noise components. The experimental results suggest that the proposed method is significantly more effective in removing the noise components from the noisy speech signal compared to recently reported techniques. The significantly better SNR improvement and the speech quality prove the superiority of the proposed algorithm.

Index Terms: speech enhancement, empirical mode decomposition, soft-thresholding

1. Introduction

Speech enhancement aims at improving the perceptual quality and intelligibility of a speech signal through noise removal, while paying excessive attention on the original speech components [1]. The process has significant importance in speech processing systems operating in noisy environments. Due to its importance in today's information technology, the topic is widely researched and many methods have been developed for this purpose. Since speech signals are nonlinear and non-stationary in nature, the performance of related studies is dependent on the analysis method. Fourier transform and wavelet analysis made great contributions. However, for nonlinear and non-stationary signals, these analysis methods suffer from many shortcomings [2].

A new nonlinear technique, the empirical mode decomposition (EMD), has recently been pioneered by Huang *et. al.* [2] for analyzing the nonlinear and non-stationary signals. This powerful data analysis method, often proving its efficiency, has made a new and effective path in speech enhancement studies as well as in many other research areas. The purpose of the method is to adaptively represent the nonlinear and non-stationary signals as sums of zero-mean oscillating components, named the intrinsic mode functions (IMFs). The idea of finding the IMFs relies on subtracting the highest oscillating components from the data with a step by step process. Therefore the IMFs have different frequency characteristics; the upper the IMF, the higher its frequency content. With this powerful characteristic, 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. As mentioned in [3], in case of white noise, most of the noise components of a noisy speech signal are centered on the first three IMFs. Therefore, EMD makes it possible to at some extent separate the high frequency noise from the major speech components. A thresholding algorithm can be applied to each IMF depending on its specific noise level to eliminate the noise components while keeping the original speech signal.

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 [4, 5], a new soft thresholding strategy was proposed in [6]. The later one is capable 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, it is not possible to efficiently remove the noise components that are embedded in the higher energy speech components. Due to the frequency characteristics of IMFs, EMD makes it possible to also separate these noise components effectively. With a modification on this soft thresholding algorithm, we can successfully denoise the IMFs of the noisy speech signal.

In this paper, we illustrate a novel speech enhancement method based on applying the soft thresholding algorithm with EMD. The proposed method includes a modification of the soft thresholding strategy and a specific approach for each IMF of the noisy speech.

2. 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 the 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 same or differ at most by one, (*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 [2]. 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 not have redundant fluctuations as induced by asymmetric waveforms. Although a mathematical model has not been developed yet, different methods for computing EMD have been proposed after its introduction [7, 8]. The very first algorithm, called as the sifting process, is adopted here to find the IMFs of the data. 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 $m_1(t)=[u(t)+l(t)]/2$
4. Since IMF should have zero local mean, subtract out $m_1(t)$ 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 $r_1(t)$: $r_1(t) = s(t) - C_1(t)$. 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 maxima and one minima from which no more IMF can be derived [7]. The subsequent IMFs and the residues are computed as:

$$r_1(t) - C_2(t) = r_2(t), \cdots, r_{n-1}(t) - n(t) = r_n(t) \qquad (1)$$

At the end of the decomposition, the data $s(t)$ will be represented as a sum of n IMF signals plus a residue signal, which is generally a constant or a monotonic trend:

$$s(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \qquad (2)$$

3. DCT Soft-Thresholding

Transform domain speech enhancement methods commonly use amplitude subtraction based soft thresholding defined by [4, 5]:

$$\hat{X}_k = \begin{cases} sign(X_k)(|X_k| - \sigma_n), & if |X_k| < \sigma_n \\ 0 & , otherwise \end{cases} \qquad (3)$$

where σ_n denotes the standard deviation of the noise, X_k is the k'th coefficient of the noisy signal obtained by the analyzing transformation and $\hat{X}_k$ represents the corresponding thresholded coefficient. Since all the coefficients are thresholded by σ_n, the speech components are also degraded during this process. Giving effective results in the case of low SNR, this method cannot be applied for high SNR values, where components are mostly the speech signals.

As reported in [6], soft thresholding technique for DCT speech enhancement is effective in denoising the noisy speech signal for a wide range of SNR values. The main advantage of the technique comes from the new soft thresholding strategy which enables even signals with high SNRs to be enhanced.

The noisy signal is segmented into 32 ms frames and a 512 point DCT is applied to each frame separately. The DCT coefficients of each frame are further divided into 8 sub-frames each containing 64 DCT coefficients. For adaptive thresholding, each sub-frame is categorized as either signal-dominant or noise-dominant. The classification pertains to the average noise power associated with that particular sub-frame. If for the i'th sub-frame:

$$\frac{1}{64}\sum_{k=1}^{64} \left|X_k^i\right|^2 \geq \sigma_n^2 \qquad (4)$$

then this sub-frame is characterized as a signal dominant sub-frame, otherwise a noise dominant one. In case of a signal dominant sub-frame, the coefficients are not thresholded, since it is highly possible to degrade the speech signal, especially for high SNRs. In the case of a noise dominant subframe, the absolute values of the DCT coefficients are first sorted in ascending order and then a linear thresholding is applied:

$$\hat{X}_k = sign(X_k)[max\{0, (|X_k| - mj)\}] \qquad (5)$$

where the multiplication mj is the linear threshold function while j being the sorted index-number of X_k. An estimated value of m can be obtained by:

$$m = \frac{\lambda \sigma_n}{\frac{1}{64}\sum_{k=1}^{64} k^2} \qquad (6)$$

A reasonable value for λ is between 0.2 and 0.8 [6].

4. EMD Soft-Thresholding Algorithm

The proposed method is based on applying the soft thresholding algorithm in (5) to the IMFs of the noisy speech. First, EMD is applied to the noisy speech. The obtained IMFs are divided into 4 ms sub-frames, thus each having 64 data for a 16 kHz sampling frequency. Similar to the DCT case, these sub-frames are further characterized as either a signal dominant or a noise dominant sub-frame. However for categorizing the sub-frames, unlike the limit defined in (4), a novel strategy is introduced here. This new soft-thresholding strategy provides an effective limit for the sub-frame categorization. Moreover, the noise variance used in thresholding is estimated separately for each IMF. This new strategy is applied to the IMFs of the noisy speech signal.

4.1. A Novel Soft-Thresholding Strategy

The categorization of the sub-frames is one of the key points of the soft thresholding algorithm. The main purpose in this categorization is to make it possible to eliminate the noise signals without degrading the original speech components. This makes the soft thresholding algorithm to be applicable for a wide range of SNR values. However, applying this algorithm directly to the IMFs of the noisy speech signal will fail for two reasons. First, IMFs will have different noise and speech energy distribution, which suggests that each IMF will have a different noise and speech variance. Second, due to the decomposition, the variance of the IMF sub-frames will have more fluctuations than that of the noisy speech sub-frames. Therefore, the noise variance of each IMF should be defined separately and the limit for sub-frame categorization should have a larger value then the limit defined in (4), in order to guarantee that all the noisy sub-frames are thresholded.

A novel limit relies on the idea that a sub-frame can be defined as a noise dominant sub-frame, if the noise power is higher than the speech power. Therefore, the limit should be set to the case where the noise and speech variance (σ_n^2 and σ_n^2) are same. For any sub-frame;

$$\sigma_{(sub)}^2 = \sigma_{(s+n)}^2 \qquad (7)$$

$$thus, \quad \sigma_{(sub)}^2 = \sigma_s^2 + \sigma_n^2 + 2 * Cov(s,n) \qquad (8)$$

where $\sigma_{(sub)}^2$ denotes the noise variance of the sub-frame and Cov is the covariance function. In case of independence, the covariance function gives zero, thus we have;

$$\sigma^2_{(sub)} = \sigma_s^2 + \sigma_n^2 \quad (9)$$

For equal noise and speech power, we get;

$$\sigma^2_{(sub)} = \sigma_s^2 + \sigma_n^2 \overset{\sigma_s^2=\sigma_n^2}{\Longrightarrow} \sigma^2_{sub} = 2\sigma_n^2 \quad (10)$$

Therefore, in the case of equal noise and speech power, the variance of a sub-frame is equal to $2\sigma_n^2$. That is why; the limit for the categorization of sub-frames in (4) should be set to this value. With the proposed strategy, if for the *i*'th sub-frame;

$$\frac{1}{64}\sum_{k=1}^{64}\left|X_k^i\right|^2 \geq 2\sigma_n^2 \quad (11)$$

then this sub-frame is defined as signal dominant, otherwise as a noise dominant sub-frame. In case of a noise dominant sub-frame, the IMFs are thresholded as in (5), where the noise variance in (6) is calculated separately for each IMF.

4.2. Variance of the IMFs

The estimation of the variance of each IMF plays an important role in the performance of the EMD soft thresholding algorithm. The calculation is achieved by a data-adaptive, efficient algorithm. The IMFs are divided into 4 ms frames and the variance of each frame is stored in a variance array. The variance array is sorted in ascending order. Since the speechless parts will mostly have the lowest variance, the noise variance of each IMF can be estimated from the speechless parts of its variance array. Figure 1 shows a plot of the variance of the sub-frames in ascending order for the first 6 IMFs of a noisy speech signal at 10dB SNR.

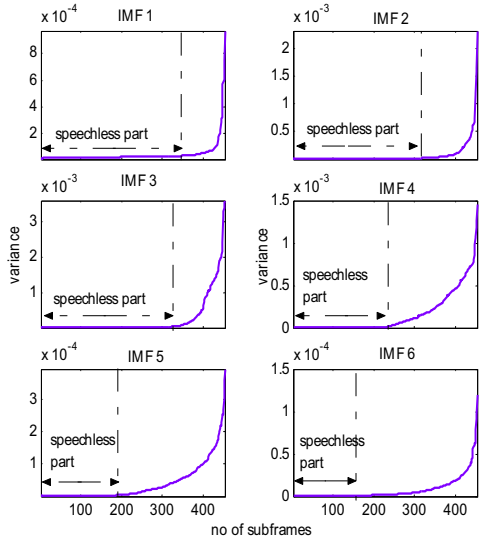

Figure 1: *Sorted noise variance of 4ms sub-frames for the first 6 IMFs of a noisy speech signal at 10dB SNR.*

The difference between the noise variance and the length of the speechless parts of the IMFs can be observed in Figure 1. As mentioned in [3], the noise signals are concentrated in the first 3 IMFs. The later IMFs are mainly the speech signals, but also have significant amount of noise. With this method, we have a very good estimation of the noise variance of each IMF. By this way, with the proposed soft thresholding algorithm, the noise components in all the IMFs can effectively be removed.

5. Experimental Results

To illustrate the effectiveness of the proposed algorithm, extensive computer simulations were conducted with different 10 male and 10 female utterances, which were randomly selected from TIMIT database. In order to observe the performance for a wide range of input SNRs, computer generated random white noise sequences were added to the clean speech signal to obtain the noisy signals at different SNRs. White noise is considered here, since it has been reported that this type of noise is more difficult to detect and remove than any other type [11]. The reported algorithms usually results in a residual noise. Our proposed method is very effective in removing the noise components while significantly reducing this residual noise.

Figure 2(c) illustrates the spectrogram of the clean, noisy and recovered signals for the female speech "Don't ask me to carry an oily rag like that." It can be observed that the spectrogram of the enhanced speech signal is very close to that of the clean speech signal. The noise components are significantly removed from the noisy speech. The enhanced speech has a high speech quality with significantly reduced residual noise. There is a significant increase in the SNR.

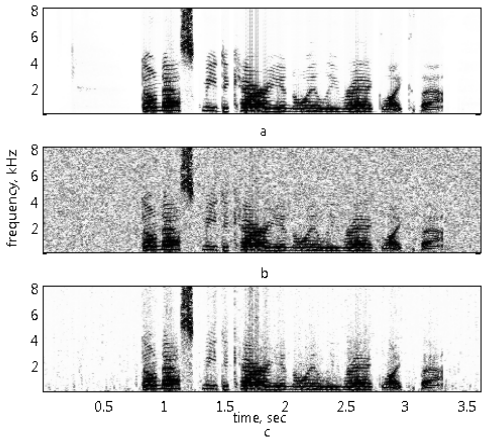

Figure 2: *Spectrogram of a) clean speech, b) noisy speech at 10dB SNR, c) enhanced speech with EMD soft thresholding.*

The power of the algorithm is not only limited with these results. Similar to the DCT soft thresholding case, the algorithm can be applied for a wide range of SNR values, basically for any value. Since the signal dominant frames are never thresholded, there is still a significant improvement even in case of high SNR values where most proposed methods even fail to hold on to the input SNR. The average results of the computer simulations of 10 male and 10 female utterances for different denoising methods in a wide range of input SNR values are listed in Table 1 (λ=0.5).

Table 1. *Comparison of the SNR improvements of different denoising methods for a wide range of SNR values.*

Input SNR (dB)	Output SNR (dB)			
	WP [5]	DCT [10]	Soft DCT [6] (λ=0.5)	Proposed EMD (λ=0.5)
0	4.86	5.69	5.33	5.67
5	8.86	9.76	9.67	10.14
10	12.36	13.74	13.75	14.12
15	15.45	17.86	17.93	18.15
25	20.82	26.02	26.35	26.78
30	23.16	30.25	30.56	31.28

The superiority of the proposed scheme can be well observed in the SNR improvement table. For all SNR levels, the proposed method gives better results apart from the 0dB case where the results are competitive. The effectiveness of the method can be better observed for high SNR values. The reason is, for high SNRs, the noise power is significantly less compared to the speech power. Therefore it is much harder to identify the noise components than it is for low SNRs. By introducing the EMD, this problem is solved very effectively. Since the IMFs depend on the frequency content, the high frequency noise components embedded in the speech signal are effectively separated from the speech components. As we discussed, these high frequency noise components dominate the first few IMFs. Therefore, these IMFs mainly include the noise dominant sub-frames and with the proposed soft thresholding algorithm, these IMFs are effectively denoised.

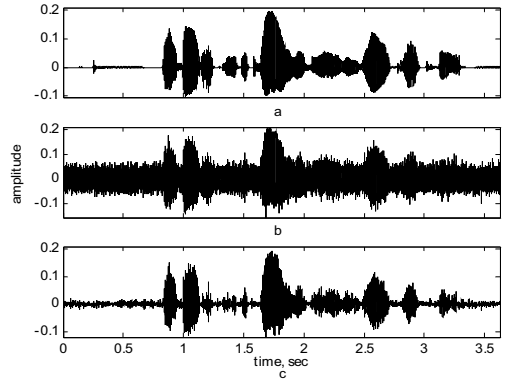

Figure 3: *Waveform of a) clean speech, b) noisy speech at 0dB SNR, c) enhanced speech with EMD soft thresholding.*

For very low SNR values, the effectiveness of the proposed algorithm can still be observed. For 0dB case, the reason why the results are close to the other results is due to the degradation of the speech signal. For such a low SNR, the noise dominant frames are significantly high. Therefore during thresholding, not only the noise components are removed but also some low energy speech components. The power of the method in removing the noise components at very low SNR can be observed in Figure 3, which shows the waveforms of clean, noisy and enhanced speech at 0dB SNR. Considering that, at 0dB SNR, it is not an easy task to remove the noise components without degrading the speech signal, it can be concluded that the proposed method is very promising in terms of noise removal even for such a low SNR.

6. Discussion

As the experimental results suggest, the proposed method is very powerful in terms of noise removal for a wide range of SNR values. The main advantage of the method comes from the characteristics of the IMFs and the new soft thresholding strategy. Unlike the soft thresholding criteria in [6], the new soft thresholding strategy assures that all the noise dominant sub-frames are thresholded. Moreover, processing each IMF separately depending on its noise-speech energy distribution provides a much better elimination of the noise components. If we apply the estimated noise variance of the noisy speech signal for all the IMFs, the thresholding would dramatically degrade the speech signal. Therefore the introduction of the variance calculation of each IMF has significant importance in the effectiveness of the algorithm.

The algorithm may further be improved by modifying the value of λ and the limit for noise categorization. Such a modification can be based on the SNR of the noisy speech signal. For instance, the remaining noise in Figure 3 suggests that it is better to have a higher λ for very low SNR inputs. However, it is also important not to degrade the speech signal. Therefore, the optimum value can be related with the input SNR. Similar approach can also be adapted for the sub-frame categorization limit. By this way, we can have a better elimination depending on the input signal. An estimation of the SNR of the noisy speech signal can be achieved by the noise variance estimation algorithm given in 4.2.

7. Conclusion

In this paper, we presented a novel speech enhancement method based on adapting a modified soft thresholding algorithm to the IMFs of the noisy speech signal. We have shown that the proposed method effectively removes the noise components while paying significant attention on the speech signal. The main advantage of the algorithm is the effective removal of the noise components for a wide range of SNRs.

Due to the successful thresholding algorithm and the advantage of EMD, the proposed method is significantly more effective in removing the noise components from the noisy speech signal compared to recently reported techniques for a wide range of input SNRs. Specifically, for high input SNRs, the algorithm is performing better than the previous methods. We not only have better SNR but also a fine speech quality with significantly reduced residual noise.

8. References

[1] Deller, J. R., Proakis, J. G. and Hansen, J. H. L., Discrete-time processing of speech signals, IEEE Press, New York, 2000.

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

[3] Zou, X., Li, X., and Zhang R., "Speech Enhancement Based on Hilbert-Huang Transform Theory," in First International Multi-Symposiums on Computer and Computational Sciences, vol. 1, pp. 208–213, 2006.

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

[5] Bahoura M., and Rouat, J., "Wavelet speech enhancement based on the teager energy operator," IEEE Signal Process. Lett., vol. 8, pp. 10–12, 2001.

[6] Salahuddin S., Al Islam, S. Z., Hasan, Md. K., and Khan, M.R., "Soft thresholding for DCT speech enhancement," Electronics Letters, vol. 38, pp. 1605– 1607, 2002.

[7] Flandrin, P., Rilling, G., and Goncalves, P., "Empirical mode decomposition as a filter bank," IEEE Signal Processing Letters, vol 11(2), pp. 112-114, 2004.

[8] Ivan, M. C., and Richard, G. B., "Empirical mode decomposition based frequency attributes," Proceedings of the 69th SEG Meeting, Texas, USA, 1999.

[9] Wu, B. Z., and Huang, N. E., "A study of the characteristics of white noise using the empirical mode decomposition method", Proc. Roy. Soc. Lond. A (460), pp.1597-1611, 2004.

[10] Hasan, M. K., Zilany, M. S. A., and Khan, M.R., "DCT speech enhancement with new hard and soft thresholding criteria," Electron. Lett., vol. 38, (13), pp. 669-670, 2002.

[11] Soon, I. Y., Koh, S. N., and Yeo, C. K., "Noisy speech enhancement using discrete cosine transform," Speech Communication, vol. 24, pp. 249-257, 1998.