SPEECH ENHANCEMENT USING SOFT THRESHOLDING WITH DCT-EMD BASED HYBRID ALGORITHM

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

[1]Dept. of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan
[2]Dept. of Frontier Informatics, School of Frontier Sciences, 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 new speech enhancement method using soft thresholding with a Discrete Cosine Transform (DCT) and Empirical Mode Decomposition (EMD) based hybrid algorithm. Soft thresholding for DCT-enhancement is a powerful method for enhancing the noisy speech signal in a wide range of signal-to-noise ratios (SNRs). However, due to the thresholding criteria a significant amount of noise is left in the enhanced signal. EMD is applied here to remove the remaining noise components. Due to the frequency characteristics of the intrinsic mode functions (IMFs), the noise components are mainly centered in the lower order IMFs. Therefore, it is possible to successfully identify and remove the remaining noise. The experimental results show that the proposed hybrid method is significantly more effective in removing the noise components from the noisy speech signal; thus giving better results in output SNR and quality compared to recently reported techniques.

1. INTRODUCTION

Speech enhancement aims at suppressing noise and improving the perceptual quality and intelligibility of speech in speech-based human-machine interfaces [1]. Due to its significant importance in today's information technology, many methods have been developed for this purpose. Since speech signals are nonlinear and non-stationary in nature, the performance of related studies is significantly dependent on the analysis method. Although Fourier transform and wavelet analysis made great contributions, they suffer from many shortcomings in the case of nonlinear and non-stationary signals [2].

The empirical mode decomposition (EMD), recently been pioneered by Huang *et. al.* [2] as a new and powerful data analysis method for nonlinear and non-stationary signals has made a new and effective path for speech enhancement studies. Basically, EMD is a data-adaptive decomposition method with which any complicated data set can be decomposed into zero mean oscillating components, named intrinsic mode functions (IMFs). Such functions give sharp and meaningful identifications of instantaneous frequencies. Recent studies have shown that with EMD, it is possible to successfully remove the noise components from the IMFs of the noisy speech. It is mentioned in [3] that, in case of white noise, most of the noise components of a noisy speech signal are centered on the first three IMFs due to their frequency characteristics. Therefore EMD can be used for effectively identifying and removing these noise components.

Soft thresholding is a powerful technique used for removing the noise components by subtracting a constant value from the coefficients of the noisy 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. The results suggest that, although the method removes the noise components for a wide frequency range, a noticeable amount of noise still remains in the enhanced signal. The remaining noise 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 IMFs, EMD makes it possible to remove these remaining noise components effectively.

In this paper, we illustrate a hybrid method which will include a two-stage soft thresholding: (*i*) soft thresholding with DCT coefficients as a pre-process to remove the noise components for a wide range of frequencies, (*ii*) soft thresholding with IMFs to identify and remove the remaining noise components in the enhanced signal from the first stage.

2. EMPIRICAL MODE DECOMPOSITION

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 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:

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

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:

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

3. SOFT THRESHOLDING FOR DCT SPEECH ENHANCEMENT

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

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

where σ_v 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_v, 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 in high SNR environments 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 categorised 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:

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

then this sub-frame is characterized as a signal dominant subframe, 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 sub-frame, the absolute values of the DCT coefficients are first sorted in ascending order and then a linear thresholding is applied:

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

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:

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

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

4. PROPOSED HYBRID ALGORITHM

The proposed method is based on applying the soft thresholding algorithm in two stages. In the first stage, the soft thresholding for DCT enhancement algorithm is used as a pre-process. As discussed above, this algorithm is effective in removing the noise components for a wide range of SNR values. However, since the signal dominant sub-frames are not thresholded, the noise signals in these sub-frames are not removed. Moreover, a significant amount of the noise signals in the noise dominant sub-frames remain within the signal due to the subtraction rule. Therefore a significant amount of noise still exists in the enhanced signal, which results in an irritating sound. It is not an easy task to detect these noise components and to remove them without degrading the speech signal. The second stage of the proposed hybrid algorithm is effective in removing these remaining noise components.

In the second stage, we introduce a new method by adapting the soft thresholding algorithm to the intrinsic mode functions of the signal. Similarly, the recovered noisy signal from the first stage is first segmented into 32ms frames and each frame is decomposed into its IMFs with empirical mode decomposition. Since EMD decomposes the signal depending on its frequency content, most of the noise components are centered on the first three IMFs as reported in [3]. Therefore with this decomposition, we can have a powerful identification of the residual noise in the enhanced signal of the first stage. By this way, it will be possible to remove even the noise components within the signal dominant sub-frames that were not removed in DCT enhancement algorithm. However, which IMFs to be thresholded should be carefully defined. Extra attention should also be paid to the threshold values of each IMFs, because the signal has already been thresholded once in the first stage and the IMFs differ in terms of noise and speech content. In order to determine these points, recent studies and our experimental analysis gave us the following conclusions:

1) As reported in [3], in case of a noisy speech signal, the first IMF mainly consists of the noise components. However, this IMF also has a reasonable amount of speech signal which should be kept. Therefore, this IMF should be thresholded with a threshold vector that will keep the signal components.
2) The second IMF is still mainly noise, but has more speech signal components compared to the first IMF. Thus the threshold vector should be less compared to the first one.
3) A significant amount of the noise components have already been removed, but there are still major noise components in the third and fourth IMFs. Therefore, these IMFs should also be thresholded but the threshold values should be less compared to the first two IMFs.
4) Since the DCT speech enhancement has already been applied in the first stage and it is known that most of the noise signals are within the first three IMFs, the lower IMFs are mainly the speech signal. Further thresholding will mostly degrade the speech components. These IMFs should not be thresholded.

Similar to the soft thresholding for DCT, each IMF is further divided into 8 sub-frames, each having 64 samples. Depending on the average noise power as discussed in equation (4), similar to the DCT case, each sub-frame is characterized as either noise dominant or signal dominant. Signal dominant sub-frames are not thresholded. In case of a noise dominant sub-frame, the absolute values of the samples are sorted in ascending order and the following thresholding strategy is followed:

$$\hat{X}_k = sign(X_k)\left[max\left\{0,\left(|X_k| - \frac{mj}{4i}\right)\right\}\right] \qquad (7)$$

where threshold function mj is same as in equation (5) and i is the index of the IMF in concern. Therefore, $\frac{mj}{4i}$ is the weighted linear threshold function defined for the IMFs.

5. EXPERIMENTAL RESULTS

To illustrate the effectiveness of the proposed hybrid algorithm, extensive computer simulations were conducted with different 10 male and 10 female utterances, which were selected randomly from TIMIT database. In order to observe the performance for a wide range of SNRs, computer generated white noise sequences were added to the clean speech signal to obtain the noisy signals at different SNRs. The variance of the noise signal was estimated from the speechless parts of the noisy speech signal.

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 [9]. The reported algorithms usually result in a residual noise. Our proposed method is very effective in removing the noise components while significantly reducing this residual noise.

In the first stage, many simulations with DCT speech enhancement (λ=0.8) were conducted in order to get the recovered signal for a wide range of SNR values. As discussed before, with DCT speech enhancement, there is the residual noise problem which makes an irritating sound. Figure 1(a) shows the spectrogram of the female clean speech “she had your dark suit in greasy wash water all year” from TIMIT database. The corresponding noisy speech signal at 10dB SNR can be observed in Figure 1(b). To better understand the noise distribution of the enhanced signal in the DCT stage, the spectrogram of the recovered signal after the DCT enhancement can be seen in Figure 1(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 successfully removed for a wide range of frequencies, the remaining noise components in the enhanced signal can easily be observed. This noise signal is randomly distributed in all frequency ranges, thus looks like white noise. As illustrated in Figure 1(c), due to the thresholding criteria, the remaining noise components have less power compared to the noise signal in the real mixture. This explains why careful attention should be paid to the threshold values in the second stage. Applying the same linear threshold function as in the first stage, while removing the noise signal, will degrade the speech signal dramatically. Therefore, it is significantly important to define lower threshold values which will be enough to remove the noise signals in each IMFs. As discussed before, since each IMFs differ in terms of noise content, the linear threshold functions should also have different weights for each IMFs.

After extensive simulations, we have defined the threshold functions as in equation (7). The first four IMFs of the enhanced signal from the first stage, as in Figure 1(c), were thresholded with these defined linear threshold functions. With this second stage, we could manage to efficiently remove the noise components while successfully keeping the speech signals. By this way, we not only have a significant improvement in the SNR but also get rid of the irritating residual noise. The spectrogram of the overall recovered signal in Figure 1(d) illustrates the effectiveness of our proposed method. It can be observed that the spectrogram of the recovered signal is very close to that of the clean speech signal. The noise signals in the speechless parts are completely removed. The noise components in the speech signal are also significantly removed. The speech quality is very close to the clean signal, with significantly reduced residual noise. There is a significant increase in the SNR.

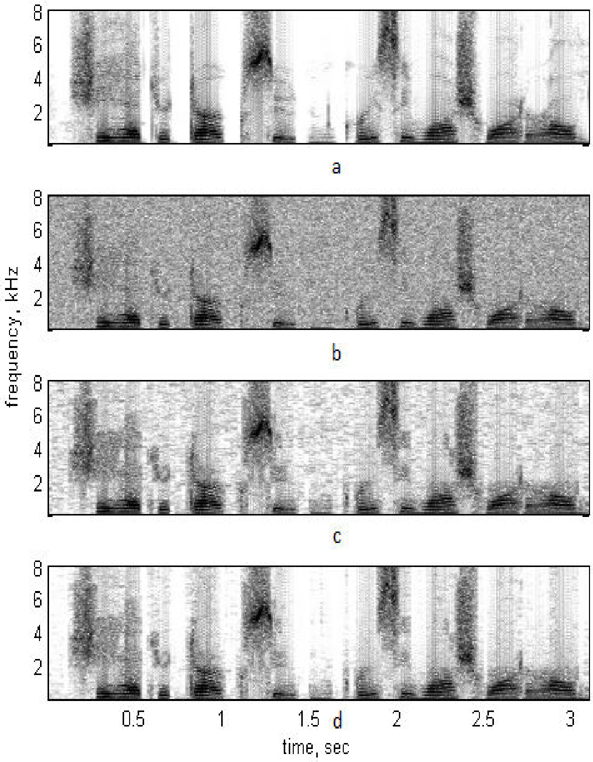

Figure 1 - Spectrogram of a) clean speech, b) noisy speech at 10dB SNR, c) the recovered speech after soft thresholding with DCT, and d) the overall recovered speech of the proposed method.

The power of the algorithm is not only limited with these results. 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. Even for very high SNRs, the algorithm at least keeps the input SNR. Therefore, the proposed algorithm prevents degradation from the given input SNR, which is another significant power of the method. Table 1 shows the effectiveness of the algorithm compared to previously reported results for a wide range of SNRs.

It can be observed that for all SNR levels, the proposed method gives better results. The effectiveness of the method can be better observed for high SNR values. The reason why we have such a result is simple. In the case of high SNRs, the noise power is significantly less compared to the audio signal. Therefore the variance of individual frames is dependent on the speech signal, which means that most of the frames are signal dominant and are not thresholded in the first stage.

Table 1 - Comparison of the SNR improvements of different denoising methods for a high range of SNR values.

Input SNR (dB)	Output SNR (dB)			
	WP[5]	DCT[10]	Soft DCT[6] ($\lambda = 0.8$)	Proposed ($\lambda = 0.8$)
0	4.86	6.24	7.31	8.45
5	8.86	10.01	10.81	11.82
10	12.36	13.61	14.42	15.51
15	15.45	17.38	18.34	19.33
25	20.82	25.09	26.56	27.59
30	23.16	29.14	30.56	31.93

By introducing the second stage with EMD, this problem is solved very effectively. Since the IMFs depend on the frequency content, the noise components dominate in the first few IMFs. Therefore, with this effective separation, these IMFs mainly include the noise dominant frames, which will be thresholded. By this way, a significant amount of the remaining noise components are removed.

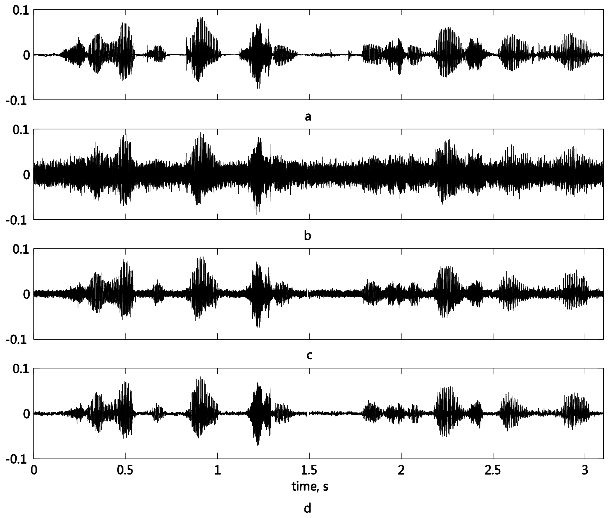

Figure 2 - Waveform of a) clean speech, b) noisy speech at 0dB SNR, c) the recovered speech after soft thresholding with DCT, and d) the overall recovered speech of the proposed method.

For very low SNR values, we still can observe the effectiveness of the proposed algorithm in removing the noise components. The reason why the results are close to the other results is due to the degradation of the speech signal. For instance, in case of 0dB, the noise dominant frames are significantly high. Therefore during thresholding, we not only remove the noise components but also degrade some speech signals. The power of the method in removing the noise components can be observed in Figure 2.

Considering that, at 0 dB 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. The spectrogram of the results for 0 dB SNR can be well observed in Figure 3. We can see the similarity between the spectral distribution of the clean speech and the recovered speech signals. The degradation of the signal is mainly in the low energy signal components. Therefore, the proposed method not only gives higher SNR but also a reasonably better speech due to significantly less noise components.

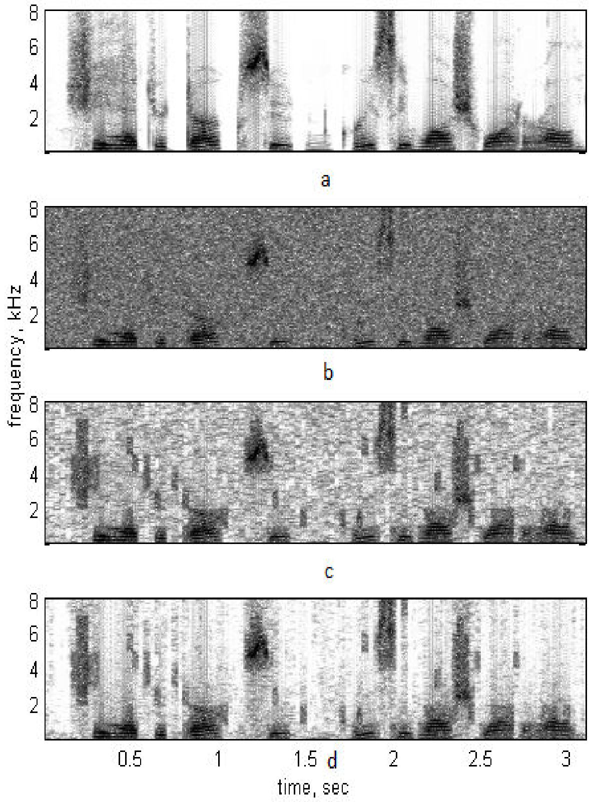

Figure 3 - Spectrogram of a) clean speech, b) noisy speech at 0dB SNR, c) the recovered speech after soft thresholding with DCT, and d) the overall recovered speech of the proposed method.

The hybrid algorithm has shown how two different methods can be combined in order to make a powerful algorithm that can remove most of the noise components within the signal. The results are very promising for further studies. Our future work will mainly include the following ideas:

i. The discrete noise left in the speech signals can be eliminated with spectral smoothing.
ii. The SNR can be estimated roughly for the noisy speech signal. For low SNR values, we can adjust another threshold criterion which will not degrade the speech signal.
iii. For low SNR case, since the third and fourth IMFs will have significant signal components, we can threshold the first two IMFs instead of the first four.

6. CONCLUSION

In this paper, we presented a new hybrid algorithm to effectively remove the noise components while paying significant attention on the speech signal. We have combined two powerful methods, soft thresholding and empirical mode decomposition in order to clean the noise signals in two stages.

The results have shown that the proposed method is very powerful compared to the recently proposed results for all SNR values. The main advantage of the algorithm is the effective removal of the noise components for a wide range of SNRs. We not only have better SNR but also a better speech quality with significantly reduced residual noise.

REFERENCES

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

[2] N. E. Huang *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] X. Zou, X. Li, and R. Zhang: "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] D. L. Donoho, "De-noising by soft thresholding," in *IEEE Trans. Inf. Theory,* vol. 41, pp. 613–627, 1995.

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

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

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

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

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

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