Noisy Speech Enhancement Using Soft Thresholding on Selected Intrinsic Mode Functions

Hadhami Issaoui *issaouihadhami@yahoo.fr*
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère B.P. 37, 1002 Tunis, Tunisia

Aïcha Bouzid *bouzidacha@yahoo.fr*
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère B.P. 37, 1002 Tunis, Tunisia

Noureddine Ellouze *N.Ellouze@enit.rnu.tn*
Electrical Engineering Department
National School of Engineers of Tunis
Le Belvédère B.P. 37, 1002 Tunis, Tunisia

Abstract

In this paper, a new speech enhancement method is introduced. It is essentially based on the Empirical Mode Decomposition technique (EMD) and a soft thresholding approach applied on selected modes. The proposed method is a fully data driven approach. First the noisy speech signal is decomposed adaptively into intrinsic oscillatory components called Intrinsic Mode Functions (IMFs) by using a time decomposition called sifting process. Second, selected IMFs are soft thresholded and added to the remaining IMFs with the residue to reconstitute the enhanced speech signal. The proposed approach is evaluated using speech signals from NOISEUS database corrupted with additive white Gaussian noise. Our algorithm is compared to other state of the art algorithms.

Keywords: Empirical Mode Decomposition, Speech Enhancement, Soft Thresholding, Mode-Selection.

1. INTRODUCTION

Speech enhancement is a challenging task aiming to suppress noise and to improve the perceptual quality and intelligibility of the speech signal through the noise removal. In the literature, various speech enhancement algorithms have been proposed to improve the performances of modern communications devices, particularly, in the case of additive white Gaussian noise [1, 2, 3, 4, 5].

In fact, linear methods such as the Weiner filtering are the most used because of their implementation simplicity [1]. However, these methods are not sufficiently effective for transient or pulse signals.

The spectral subtraction method introduced in [2] remains an interesting choice in reducing the additive noise. Despite its capability of removing the background noise, this method introduces additional artifacts called musical noise [3].

In recent years, a non-linear approach based on wavelet transform has been proposed. The main idea is to threshold the wavelet coefficients by keeping only those which are supposed to correspond to the signal [4, 5, 6]. This method has shown a good agreement. However, a

drawback of the wavelet approach is that the analyzed functions are predetermined in advance and it is not often optimal to describe the signal non stationarity.

In the last decade, a new non linear technique, termed empirical mode decomposition (EMD), has been introduced by N. E. Huang et al. [7] for adaptively representing non stationary signals. The most important characteristic is that the basis functions are directly derived from the speech signal itself. Thus the EMD allows the decomposition of a signal into a finite sum of components, called Intrinsic Mode Functions (IMFs).

In this paper, we will present a new speech enhancement approach based essentially on the Empirical Mode Decomposition technique (EMD) and a soft thresholding approach applied on selected modes. The basic idea is to fully reconstruct the signal with all IMFs by thresholding only the first IMFs (low order components) and keeping unthresholded the last components.

2. EMPIRICAL MODE DECOMPOSITION

2.1 Principle

The principle of the EMD technique is to decompose a given signal x(t) into series of oscillating components called Intrinsic Mode Functions (IMFs) via an iterative procedure called sifting process, each one with a distinct time scale. The decomposition is based on the local time scale of x(t), and yields adaptive basis functions.

By mean of the EMD, the signal x(t) is decomposed into fast oscillations superposed to slow oscillations. Thus, each IMF contains locally lower frequency oscillations than the one that was extracted just before.

An IMF must fulfill the two following conditions:
C1- In the whole data series, the number of local extrema and the number of zero crossings must be the same or differ at most by one.
C2- At any point, the mean value of the local maxima envelope and the local minima envelope is zero [7].

2.2 Algorithm

To determine the IMFs, denoted $imf_i(t)$, the sifting process can be summarized as follows:

1. Initialize: $r_0(t) = x(t)$, $i = 1$
2. Extract the i^{th} IMF:
   a. Initialize: $h_0(t) = r_i(t)$, $j = 1$,
   b. Identify the extrema (both maxima and minima) of the signal, $h_{j-1}(t)$,
   c. Interpolate the local maxima and the local minima by a cubic spline to form upper and lower envelopes of $h_{j-1}(t)$
   d. Compute the local mean, $m_{j-1}(t)$, by averaging the envelopes,
   e. $h_j(t) = h_{j-1}(t) - m_{j-1}(t)$,

      if the stopping criterion is satisfied then set $imf_i(t) = h_j(t)$

      else go to (b) with $j = j + 1$
3. $r_i(t) = r_{i-1}(t) - imf_i(t)$,
4. if $r_i(t)$ still has at least two extrema then go to (2) with $i = i + 1$

else the decomposition is finished and $r_i(t)$ is the residue.

At the end of the algorithm, the decomposition of $x(t)$ is given by:

$$x(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t) \quad \textbf{(1)}$$

Where n is the mode number and $r_n(t)$ is the residue of the decomposition.

3. MODE SELECTION APPROACH

In the literature many authors have proposed approaches for signal enhancement using EMD technique based on excluding the first IMFs.

EMD extracts, sequentially and intrinsically, the energy in the signal starting from small scales (high frequency modes) towards the larger ones (low-frequency modes). The selection method is based on the assumption that the first IMFs (high-frequency modes) are mostly dominated by noise and are not representative for information specific to the original signal. Thus, the enhanced signal is reconstructed only by a few IMFs in which pure signal mostly predominates. In fact, there will be a mode, $IMF_{js}(t)$, from which the energy distribution of the original signal is greater than the noise. The idea is to separate signal from noise. The basic of this approach is to set to zero the first $j_S - 1$ IMFs [9]. As a result, the signal is partially reconstructed from the remaining IMFs.

Let $x(t)$ be the clean speech signal, $x_n(t)$ the noisy speech signal and $n(t)$ the noise (additive white gaussian noise). $x_n(t)$ is given as follows:

$$x_n(t) = x(t) + n(t) \quad \textbf{(2)}$$

The aim of this section is to find an approximation $\tilde{x}(t)$ of the original signal $x(t)$ that minimizes the mean square error (MSE) defined by [10]:

$$MSE(x, \tilde{x}) \overset{\Delta}{=} \frac{1}{N} \sum_{i=1}^{N} \left[x(t_i) - \tilde{x}(t_i) \right]^2 \quad \textbf{(3)}$$

Where $x = [x(t_1), x(t_2) ... x(t_N)]^t$, $\tilde{x} = [\tilde{x}(t_1), \tilde{x}(t_2) ... \tilde{x}(t_N)]^t$ and N is the signal length.

After decomposing the signal $x_n(t)$ through the EMD algorithm, $\tilde{x}(t)$ is reconstructed using $(n - j_S + 1)$IMF indexed from j_S to n as follows:

$$\tilde{x}_{j_s}(t) = \sum_{j=j_s}^{n} imf_j(t) + r_n(t), \; j_S \in \{2, ..., n\} \quad \textbf{(4)}$$

Since, the original signal $x(t)$ is unknown; the MSE cannot explicitly be calculated. That's why a distortion measure, termed consecutive MSE (CMSE) that does not require any knowledge of $x(t)$ [9] is used. The CMSE is defined as:

$$CMSE(\tilde{x}_k, \tilde{x}_{k+1}) \overset{\Delta}{=} \frac{1}{N} \sum_{i=1}^{N} \left[\tilde{x}_k(t_i) - \tilde{x}_{k+1}(t_i) \right]^2, \quad k \in \{1, ..., n-1\} \quad \textbf{(5)}$$

$$\overset{\Delta}{=} \frac{1}{N} \sum_{i=1}^{N} \left[imf_k(t_i) \right]^2 \quad \textbf{(6)}$$

According to [10], the CMSE is reduced to the energy of the k^{th} IMF. It is also the classical empirical variance estimate of the IMF. Finally j_s is computed as:

$$j_S = \underset{1\le k\le n-1}{\arg\min} \left[\mathrm{CMSE}(\tilde{x}_k - \tilde{x}_{k+1})\right] \quad (7)$$

Where $\tilde{x}_k$ and $\tilde{x}_{k+1}$ are signals that are respectively reconstructed starting from the IMFs that are indexed by K and $(k+1)$.

By using the CMSE criterion, the IMF order corresponding to the first significant change in the energy distribution is identified.

4. SOFT THRESHOLDING

Many speech enhancement methods use amplitude subtraction based soft thresholding approach [5]:

$$\tilde{X} = \begin{cases} \mathrm{sign}(X)\left(|X| - \tau\right) & \text{if} \quad X > \tau \\ 0 & \text{if} \quad |X| \le \tau \end{cases} \quad (8)$$

Where X is the coefficient of the noisy speech signal $x_n(t)$ (as given in equation 2) obtained by the analyzing transformation, $\tilde{X}$ is the denoised version of X and τ is the threshold parameter. According to Donoho and Johnstone in [5], a universal threshold τ is given by:

$$\tau = \tilde{\sigma}\sqrt{2.\log_e(N)} \quad (9)$$

Where N is the number of samples and $\tilde{\sigma}$ represents the noise level estimation. The expression of $\tilde{\sigma}$ is:

$$\tilde{\sigma} = \mathrm{MAD}/0.6745 \quad (10)$$

Here MAD represents the absolute median deviation of X.

5. PROPOSED HYBRID APPROACH

Many speech enhancement algorithms excluding the first IMFs issued from the EMD technique are revealed not efficient.

In this work, we propose a hybrid approach for speech enhancement. We don't eliminate the first IMFs but we consider them after operating a soft thresholding. The enhanced signal is constituted by the thresholded IMFs [11, 1], the remaining ones and the residue. This approach permits us to preserve the signal components in the first IMFs. .
The proposed method follows four steps:

1. Decomposing a given noisy speech signal $x_n(t)$ into series of IMFs by EMD technique [7],
2. Applying on the obtained IMFs the Mode-Selection criteria to find the index j_S which minimizes the mean square error (MSE) [8],
3. Enhancing the first $(j_S - 1)$ IMFs by the soft thresholding algorithm [9] to obtain the denoised $imf_i(t)$ versions $\tilde{f}_i(t)$, $i = 1.....j_S - 1$,
4. Reconstructing the enhanced following signal as follows

$$\tilde{x}(t) = \sum_{i=1}^{j_s - 1} \tilde{f}_i(t) + \sum_{i=j_s}^{n} imf_i(t) + r_n(t) \quad \textbf{(11)}$$

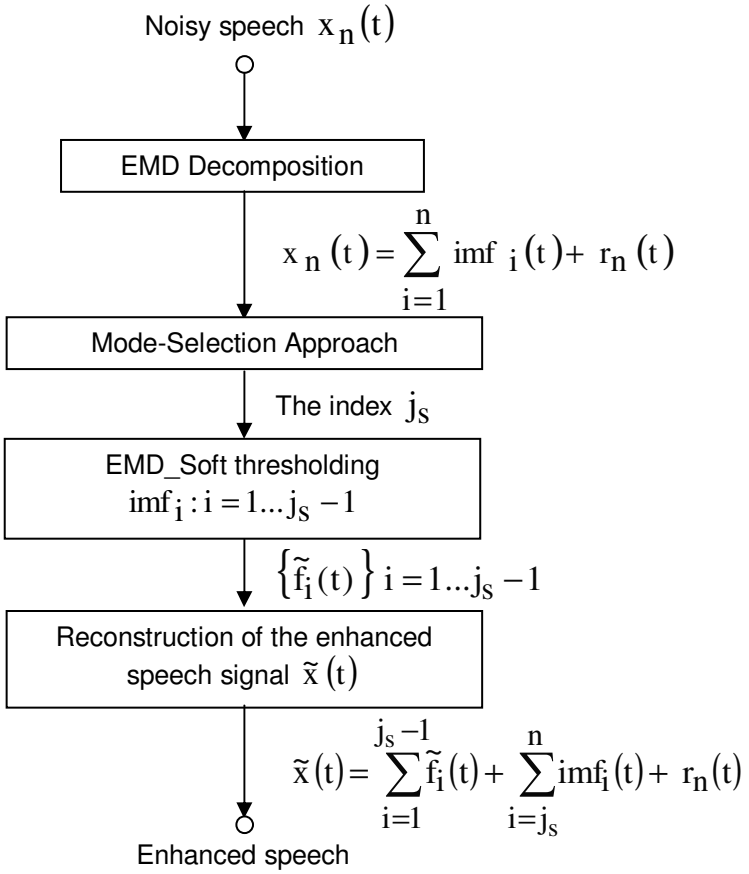

FIGURE 1: Overall block diagram of the proposed method for speech enhancement.

6. RESULTS

In order to illustrate the effectiveness of our proposed method, a total of ten sentences (5 male and 5 female speakers) taken from the NOISEUS database are used in our evaluation. The analysis is conducted by adding to the clean speech signal a white Gaussian noise with various SNR levels -5, 0, +5 and +10 dB. To operate an objective performance evaluation of our speech enhancement approach, both output SNR and Weighted Spectral Slope (WSS) distance proposed in [12], are computed.

Figure 2 illustrates the original clean speech signal taken from the NOISEUS database and pronounced by the speaker sp03 followed by the same speech signal corrupted by an additive white Gaussian noise with an input SNR of 5dB. The last signal shows the enhanced speech signal using our approach. It can be observed that the noise is reduced in the enhanced speech signal and has a shape very close to the corresponding clean speech.

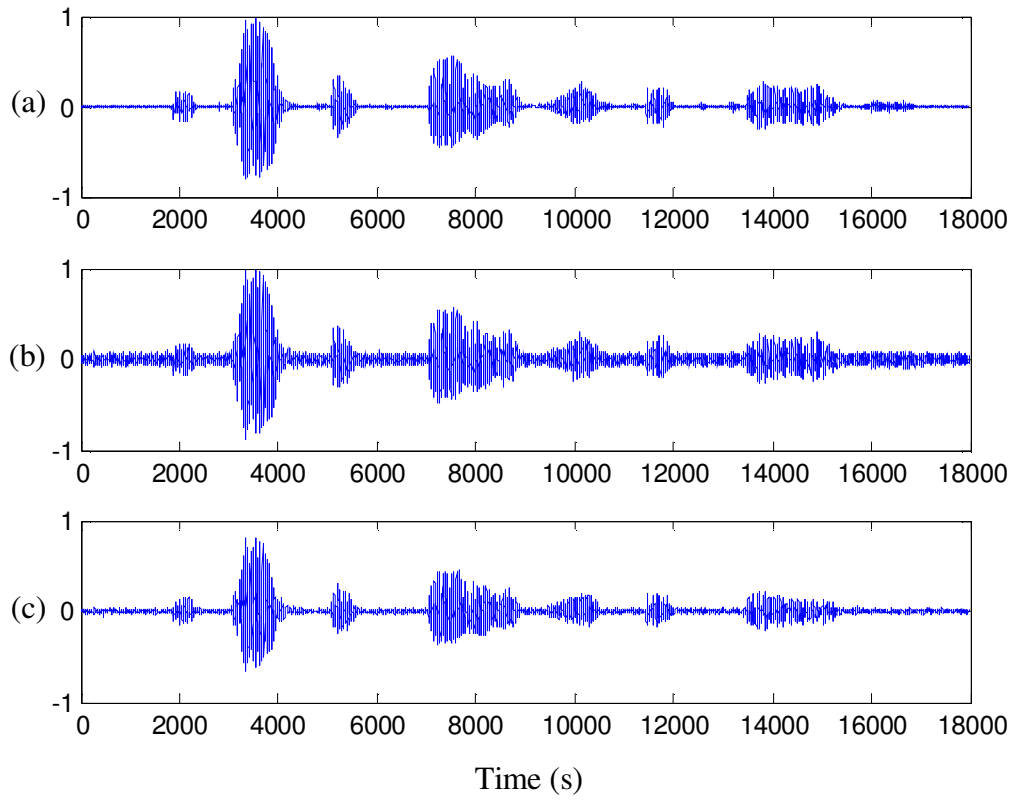

FIGURE 2: Waveform of the clean speech (a), the noisy speech at SNR of 5dB (b) and the enhanced speech signal with our proposed method (c).

In our evaluation, we compare our proposed method to two approaches using EMD soft thresholding of all IMFs and the elimination of the first IMFs [10]. We use two criterions:

- The output SNR of the enhanced speech signal.

Input SNR (dB)	Output SNR (dB)		
	First IMFs elimination	EMD_soft	Proposed
-5	0,290	**0,583**	0,387
0	1,090	1,211	**1,474**
5	3,683	4,379	**6,094**
10	8,449	8,719	**13,058**

TABLE 1: Comparison of the output SNR levels for various denoising methods.

- The Weighted Spectral Slope (WSS) measure:

The measure is based on the auditory model in which 36 overlapping filters of progressive larger bandwidth are used to estimate the smoothed short-time speech spectrum [12]. The measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. A per-frame measure in decibels is found as:

$$d_{wss}(j) = K_{SPL}(K - \hat{K}) + \sum_{k=1}^{36} w_a(k)\left[x(k) - \hat{x}(k)\right]^2 \quad (12)$$

Where $(K - \hat{K})$ is the difference between overall sound pressure level of the original and processed utterances. K_{SPL} is a parameter which can be varied to increase the overall performances.

	d_{WSS}		
Input SNR (dB)	**First IMFs elimination**	**EMD_soft**	**Proposed**
-5	105,157	122,557	**104,997**
0	**75,924**	124,531	79,438
5	66,167	104,320	**60,394**
10	46,235	72,143	**35,974**

TABLE 2: Comparison of the WSS measure for various denoising methods.

Referring to tables 1 and 2, one can clearly notice the following interpretations:

- Table 1 depicts the SNR of the enhanced speech signal compared to the SNR at the input. The proposed approach improves the speech quality by reducing the noise and performing better than the other methods at SNRs of 0, +5 and +10 dB.

- Table 2 shows the WSS evaluation criteria for our approach and two others. Our approach gives the less WSS distance for almost all the SNR levels showing its convenience for speech enhancement.

- The Mode Selection approach proposed by Boudraa in [10] is effective especially for very noisy signals. For this reason, the increase of the input SNR level leads to lower values of output SNR. This decrease is logical because on one hand this approach eliminates the first IMFs and on the other hand, for high values of input SNR, we tend toward the original signal. This causes the degradation of the original signal, and hence the interest of our approach whose principle is to keep all IMFs. As shown by Cexus and Boudraa in [11], the EMD soft thresholding performs almost better than the soft thresholding using Wavelet transform, what justifies the enhancement of the first IMFs by EMD soft thresholding applied in our work.

7. CONCLUSION

In this paper, we propose a new approach based on EMD technique for speech enhancement. It consists of four essential steps:

- The first step concerns the empirical mode decomposition of the noisy speech signal.
- The second step concerns the index mode selection j_s using an energy criterion.
- The third step concerns the soft thresholding of the first j_s -1 IMFs.
- And the forth step concerns the signal reconstruction by adding the thresholded IMFs, the remaining IMFs and the residue.

This approach shows efficiency when compared to other approaches based also on EMD technique.

8. REFERENCES

[1] A.O. Boudraa and J.C. Cexus, "Denoising via empirical mode decomposition," in Proc. IEEE ISCCSP, Marrakech Morocco, 2006.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-27 pp. 113-120, 1979

[3] A. Sumithra M G , B. Thanuskodi K, C. Anitha M R, "Modified Time Adaptive Wavelet Based Approach for Enhancing Speech from Adverse Noisy Environments," DSP Journal, Volume 9, Issue 1, p.p. 33-40, June, 2009

[4] D.L. Donoho, "De-noising by soft-thresholding," IEEE Trans. on Information Theory, 41(3):613–627, 1995.

[5] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," Biometrica, vol. 81, pp. 425–455, 1994.

[6] S. Mallat, " Une Exploration Des Signaux En Ondelettes," Ellipses, Paris, France, 2000.

[7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shin, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proc. R. Soc. Lond. A, Math. Phys. Sci., Mar. 1998, vol. 454, no. 1971, pp. 903–995.

[8] E. Derger, Md. K. I. Molla, M. Hirose, N. Minematsu, and Md. K. Hasan, "Speech Enhancement using Soft Thresholding with DCT-EMD Based Hybrid algorithm," Proc. EUSIPCO-2007, Poznań, POLAND, 2007, pp. 75–79.

[9] J.C. Cexus, "Analyse des signaux non-stationnaires par Transformation de Huang, Opérateur de Teager-Kaiser, et Transformation de Huang-Teager (THT)," Thesis, Université de Rennes 1, 2005.

[10] A. O. Boudraa, and J.C. Cexus, "EMD-Based Signal Filtering," IEEE Trans. On Instrumentation and measurement, vol. 56, no. 6, pp. 2196–2202, December 2007.

[11] A.O. Boudraa, J.C. Cexus, and Z. Saidi, "EMD-based signal noise reduction," Int. J. Sig. Process., vol. 1, no. 1, pp. 33–37, 2004, ISSN: 1304-4494.

[12] D. H. Klatt, "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step," Proc. IEEE ICASSP'82, May, 1982, Vol. 2, pp. 1278-1281.