THE UNIVERSITY OF TOKYO

MASTER's THESIS

Noise-Thresholding with Empirical Mode Decomposition for Low Distortion Speech Enhancement

(EMD に基づく閾値操作による低歪み雑音除去)

Erhan Deger
(エルハン　デゲル)
February 2008

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
THE UNIVERSITY OF TOKYO

Contents

ABSTRACT

Degrading the quality and intelligibility of the speech signals, background noise is a severe problem in communication and other speech related systems. In order to get rid of this problem, it is important to enhance the original speech signal mainly through noise reduction. Speech enhancement is the term used to describe such algorithms and devices whose purpose is to improve some perceptual aspects of the speech for the human listener or to improve the speech signal so that it may be better exploited by other speech processing algorithms. Development and widespread deployment of digital communication systems during the last twenty years have brought increased attention to the role of speech enhancement in speech processing problems. For this purpose, this thesis presents novel speech enhancement methods based on applying some thresholding strategies in Empirical Mode Decomposition (EMD) domain.

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. The EMD, recently been pioneered by Huang *et. al.* as a new and powerful data analysis method for nonlinear and non-stationary signals has made a novel 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 identify the noise components from the IMFs of the noisy speech. For instance, 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.

Thresholding is a widely used process in noise reduction algorithms. The idea is to determine a threshold value and to apply different subtraction algorithms for the segmented regions. However, it is never easy to identify and remove the noise components while keeping the original speech components non-degraded. That is why; one of the major drawbacks of these kinds of processes is the degradation of the speech signal, especially in the process of noisy signals with high signal-to-noise ratios (SNR). In order to minimize the degradation of the original speech components, a modified soft-thresholding strategy that works on a frame basis is

adapted in this study. The IMFs of the noisy speech signal are denoised by applying the modified soft-thresholding strategy on the coefficients of each IMF. With the proposed strategy, most of the noise components are successfully removed while the speech components are mainly kept. This strategy enables even signals with high SNRs to be processed effectively.

It is never possible to remove all the noise components in a noise reduction method. The remaining noise parts may result in an irritating sound which is referred as the musical noise. That is why; most speech enhancement algorithms not only introduce speech distortion but also suffer from the musical noise artifact. The proposed EMD based algorithm is highly effective in noise removal and introduces a rather discrete noise than a continuous musical sound. In order to obtain even better results for speech quality and quantity, the method was further improved by introducing a Discrete Cosine Transform (DCT) based thresholding as a first stage. The two-stage algorithm gives efficient results, successfully improving the SNR of the noisy speech and removing most of the noise components. The thesis work mainly concentrates on white noise; however the method has been further improved with a sub-band approach so that it may be applied to different noise types.

Acknowledgements

First of all, I would like to express my sincerest gratitude and appreciation to my supervisor, Prof. Keikichi Hirose, for providing me the unique opportunity to work in the area of speech processing, for his expert guidance and mentorship, and for his encouragement and support at all levels. His great attention not only motivated me in my research but also made feel like home during my stay in Japan.

Next, I would like to thank Dr. Md. Khademul Islam Molla for his guidance, patience and valuable suggestions that shaped my research study and most importantly for his great friendship. My sincere thanks to Prof. Md. Kamrul Hasan for his directions and suggestions helping me find my research path and the ideas that lead to the effective solutions. I would also like to thank Prof. Nobuaki Minematsu for his fruitful suggestions during this master work.

I would like to give my sincere thanks to all my laboratory members for their welcoming, friendship and kind attention. Their smile and support always motivated me during this research and gave me the taste of a family in Japan.

My further thanks will go to the Japanese Government (Monbukagakusho), for funding my study and living expenses during my time in The University of Tokyo, and to the NEC C&C foundation for their financial support which helped me to attend the 2007 European Signal Processing Conference (EUSIPCO'07) held in Poznan, Poland.

At last but not the least, I would like to thank my family for their life-long love and support without boundaries, which have always been a source of motivation and happiness for me.

Chapter 1

Introduction

1.1. General Introduction

In many speech related systems, the desired signal is not available directly; rather it is mostly contaminated with some interference sources. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the 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, or it can take the form of fading noise. 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. That is why; this thesis is mainly concerned in this kind of noise. Pink and high frequency channel noise are also used in order to show the robustness of the proposed algorithms.

The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should be eliminated through speech enhancement algorithms. Speech enhancement aims at improving the perceptual quality and intelligibility of a speech signal in noisy environments, mainly through noise reduction algorithms. Such types of processes may be applied to a mobile radio communication system, a speech recognition system, a set of low quality recordings, or to improve the performance of aids for the hearing impaired. Figure 1.1 shows an illustration of the usage of speech enhancement. It can be observed that enhancement may also be applied directly to the clean speech signal in order to reduce the effect of the channel noise in communication systems.

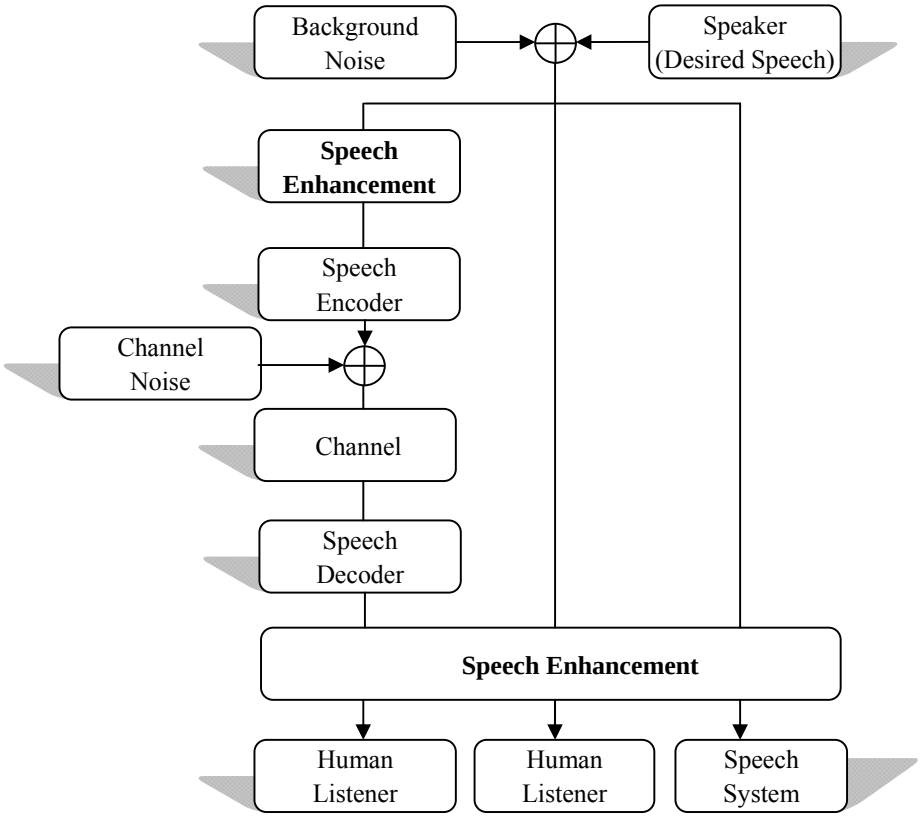


Figure 1.1: The application areas of speech enhancement.

Due to its significant importance in today's information technology, the topic is widely researched. The performance of such systems is mainly evaluated according to the quality and intelligibility. The quality of the enhanced signals refers to its clarity, distorted nature and the level of the residual noise in that signal. Most speech enhancement methods improve the quality of the signal however degrades its intelligibility, which refers to the understandability of the enhanced speech; the percentage of words that could be correctly identified by the listener. Human listeners can usually extract more information from the noisy signal than from the enhanced signal by careful listening.

Since quality and intelligibility require live listening sessions, they are both time consuming and expensive to measure. That is why; researchers mostly use some mathematical measures which are believed to be correlated with the quality and intelligibility of the enhanced speech. For this purpose, signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) tests are widely used to show the performance of the proposed algorithms in terms of their quality. For assessing the intelligibility of the enhanced signal, automatic speech recognition tests are commonly used.

1.2. Speech Enhancement Methods

There are several methods proposed for speech enhancement. The reported algorithms can be categorized into two main classes as parametric and non-parametric methods. Parametric approaches assume a model for the signal generation process. This model describes the predictable structures and the observable patterns in the process. Noise reduction is performed depending on this a-priori information. Since the enhancement is based on the parametric model, selection of the model is crucially important in those algorithms. Non-parametric approaches simply need an estimation of the noise spectrum. The noise spectrum can be estimated from the pause periods where the speaker is silent (single channel) or from a reference source (multi channel).

The speech enhancement methods can also be classified into single channel and multi channel approaches. In case of single channel, there is only one microphone and therefore only one noisy mixture which will just give spectral information. The noise spectrum has to be estimated from the pause period and the noise is regarded as stationary and uninterrupted. This makes the single channel speech enhancement challenging. That is why, the performance of the single channel techniques are limited. In case of multi channel, multiple microphones are available in the environment, leading more noisy mixtures which exhibit the advantage of incorporating both spatial and spectral information. However, since multi microphones will come at an increased cost and may not be always available, the single channel speech enhancement always attracts attention.

In this thesis, we propose non-parametric and single channel speech enhancement algorithms. Therefore, an introduction to some of the same types of noise reduction algorithms will be given.

1.2.1. Spectral Subtraction

Spectral subtraction is one of the earliest and most popular methods of reducing the effect of the background noise. The very first spectral subtraction technique for speech enhancement was given by Weiss et al, in 1974 (1). As discussed above, spectral subtraction is non-parametric method which requires only an estimate of the noise spectrum. In case of single channel, that estimation is done in the periods where the speaker is silent, which is referred as the pause periods. The main idea of the method is to estimate the spectrum of the noise signal and to subtract it from the mixture signal in order to obtain the original clean speech. Since the noise

spectrum is estimated from the pause periods and used for the whole data, spectral subtraction is suitable for stationary noises or very slowly varying noises so that the change in the noise power spectrum can be updated. Suppose that clean speech signal *x(m)* is corrupted with the noise *n(m);*

$$y(m)=x(m)+n(m) \tag{1.1}$$

Taking the Fourier transform of the signals will yield;

$$Y(e^{jw}) = X(e^{jw}) + N(e^{jw}) \tag{1.2}$$

Multiplying both sides with their complex conjugates will give;

$$|Y(e^{jw})|^2 = |X(e^{jw})|^2 + |N(e^{jw})|^2 + 2|X(e^{jw})||N(e^{jw})|cos(\Delta\theta) \tag{1.3}$$

Where $\Delta\theta$ is the phase difference between speech and noise:

$$\Delta\theta = \angle X(e^{jw}) - \angle N(e^{jw}) \tag{1.4}$$

Taking the expected value of both sides we get:

$$\begin{aligned} E\{|Y(e^{jw})|^2\} &= E\{|X(e^{jw})|^2\} + E\{|N(e^{jw})|^2\} + E\{2|X(e^{jw})||N(e^{jw})|cos(\Delta\theta)\}, \\ &= E\{|X(e^{jw})|^2\} + E\{|N(e^{jw})|^2\} + 2E\{|X(e^{jw})|\}E\{|N(e^{jw})|\}E\{cos(\Delta\theta)\} \end{aligned} \tag{1.5}$$

In deriving the last equation, two reasonable assumptions are being made:

1. Noise and speech magnitude spectrum values are independent of each other.
2. The phase of noise and speech are independent of each other and of their magnitude.

There have been many proposed methods based on spectral subtraction. It is possible to classify them in two main classes,

1.2.1.1. Power Spectral Subtraction

In power spectral subtraction, it is assumed that $E\{cos(\Delta\theta)\}$ in (6) is zero, yielding:

$$E\{|Y(e^{jw})|^2\} = E\{|X(e^{jw})|^2\} + E\{|N(e^{jw})|^2\} \tag{1.6}$$

The power spectrum of the noise is estimated from the speech inactive periods and assuming that the variations of noise spectrum are tolerable, it is subtracted from the noisy speech spectrum in order to obtain the enhanced speech.

$$|X(e^{jw})|^2 = |Y(e^{jw})|^2 - E\{|N(e^{jw})|^2\} \tag{1.7}$$

1.2.1.2. Magnitude Spectral Subtraction

In magnitude spectral subtraction it is assumed that E{$cos(\Delta\theta)$}=1, hence:

$$E\{|Y(e^{j\omega})|^2\} = E\{|X(e^{j\omega})|^2\} + E\{|N(e^{j\omega})|^2\} + 2E\{|X(e^{j\omega})|\}E\{|N(e^{j\omega})|\}$$
$$= (E\{|X(e^{j\omega})|\} + E\{|N(e^{j\omega})|\})^2$$

$$E\{|Y(e^{j\omega})|\} = E\{|X(e^{j\omega})|\} + E\{|N(e^{j\omega})|\} \quad (1.8)$$

The magnitude spectrum of the noise is averaged during speech inactive periods and with the same assumptions that the noise is stationary, the magnitude spectrum of speech is estimated by subtracting the average spectrum of noise from the magnitude spectrum of the mixture signal (2).

$$E\{|X(e^{j\omega})|\} = E\{|Y(e^{j\omega})|\} - E\{|N(e^{j\omega})|\} \quad (1.9)$$

Figure 1.2 shows an illustration of the magnitude spectral subtraction algorithms.

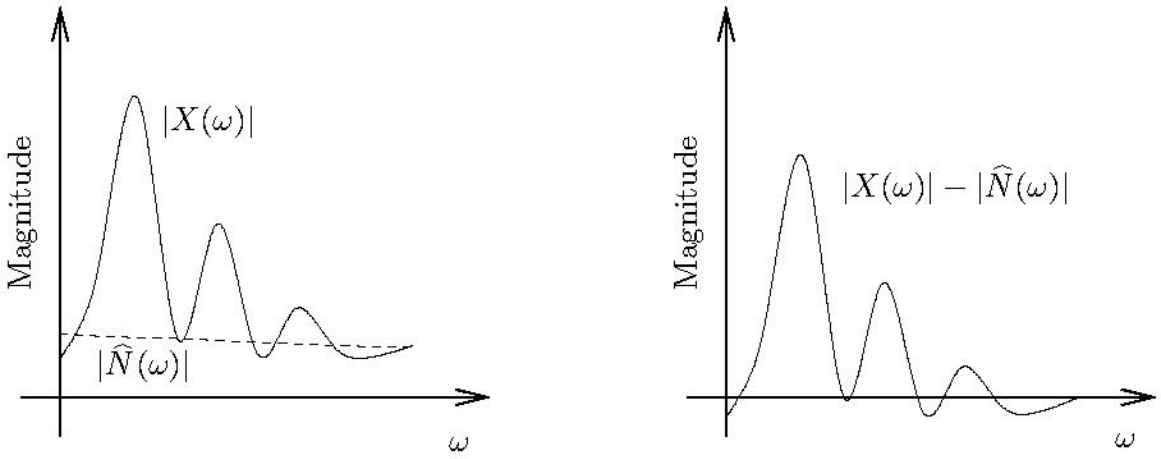


Figure 11.2: Spectral magnitude subtraction. The noise and noisy observation magnitude spectrum is illustrated to the left and the estimated signal magnitude spectrum is illustrated to the right.

1.2.1.3. Residual Noise Problem

A common problem for different kinds of speech enhancement algorithms is that as a result of the fluctuations of noise spectrum (whether power or magnitude) around its mean value, there is always some difference between the actual noise and its mean. Hence at the end of the spectral subtraction, some of the noise remains in the spectrum in the case that the value of noise is greater than its mean and some of the speech spectrum also is removed in the case that our

estimate of noise is greater than the actual value of noise. The latter produces negative values in spectrum as illustrated in Figure 1.2. These negative values are prevented or often corrected by the use of a full wave rectifier or a half wave rectifier with different techniques. However, the overall effect puts a noise in the output signal known as *residual noise*. The narrow band relatively long-lived portion of residual noise is sometimes referred to as *musical noise*. In order to better understand the musical noise artifact, the following description helps;

"To explain the nature of the musical noise, one must realize that peaks and valleys exist in the short—term power spectrum of white noise; their frequency locations for one frame are random and they vary randomly in frequency and amplitude from frame to frame. When we subtract the smoothed estimate of the noise spectrum from the actual noise spectrum, all spectral peaks are shifted down while the valleys (points lower than the estimate) are set to zero (minus infinity on a logarithmic scale). Thus, after subtraction there remain peaks in the noise spectrum. Of those remaining peaks, the wider ones are perceived as time varying broadband noise. The narrower peaks, which are relatively large spectral excursions because of the deep valleys that define them, are perceived as time varying tones which we refer to as musical noise (3)*."*

1.2.2. Wavelet Thresholding

Wavelet transform has been widely used in various fields of signal processing. Due to its advantage of using variable size time-windows for different frequency bands, wavelet transform is a powerful tool for modeling non-stationary signals like speech. Moreover, in case of single channel speech enhancement, generally the use of the sub-band processing can result in a better performance. Therefore, wavelet transform can provide an appropriate model of speech signal for denoising applications.

In case of a noisy speech signal, the energy of the clean speech is mostly concentrated in a small number of wavelet dimensions. On the other hand, the energy of the noise signal is spread over a large number of coefficients. That is why; the coefficients of a small number of dimensions where clean speech is present are relatively large compared to those of other dimensions. Hence, by defining a threshold value and setting smaller coefficients to zero, one can nearly optimally eliminate noise while preserving the important information of the original signal (4).

Suppose that the clean speech signal *x(m)* is corrupted by zero-mean, white Gaussian noise *n(m)* with variance σ^2, as in (1.1). By applying the discrete wavelet transform, we will have;

$$Wy=Wx+Wn \tag{1.10}$$

where W denotes the wavelet transformation. The wavelet of observation mixture signal is thresholded in order to obtain the original speech signal. The thresholding can be done in many ways. Different techniques have been proposed for this purpose. However there are two popular versions known as *hard* and *soft* thresholding.

1.2.2.1. Hard Thresholding

Let *Y* refer the coefficients of a wavelet dimension of the noisy mixture signal and *T* be the threshold value for the denoising strategy. Hard thresholding sets any coefficient whose absolute values is less than or equal to the threshold to zero;

$$\widehat{Y} = \begin{cases} Y & ,if\ |Y| > T \\ 0 & ,if\ |Y| < T \end{cases} \tag{1.11}$$

where $\hat{Y}$ denotes to the thresholded coefficients. The value of *T* is related with the estimated standard deviation of the noise signal σ and may change depending on the proposed algorithm. Donoho has suggested the following formula for its value;

$$T = \sigma\sqrt{2log\,(N)} \tag{1.12}$$

1.2.2.2. Soft Thresholding

Soft thresholding sets any coefficient with absolute value less than or equal to the threshold to zero and subtracts the threshold value from the other coefficients.

$$\hat{Y} = \begin{cases} sgn(y)(|Y| - |T|) & ,if\ |Y| > T \\ 0 & ,if\ |Y| < T \end{cases} \tag{1.13}$$

Those thresholding algorithms are widely used and have been also adapted in spectral subtraction methods, with Fourier transform analysis. It can be observed that the soft thresholding algorithm removes more noise components than the hard thresholding algorithm. However in soft thresholding the amount of the signal degradation is also higher. Therefore, the thresholding strategy should be selected depending on the subjective and objective perspectives.

1.3. Applications of Speech Enhancement

In the current information technology, there are many areas that speech enhancement is used in order to improve the performance of the system;

- *Robust Automatic Speech Recognition (RASR):* The accuracy of automation speech recognition degrades in the presence of background noise or other interfering sources. Noise reduction for speech signals has therefore critical importance as a pre-process of such types of systems, including human-computer interactions, robotics and audio driven systems, etc.

- *Telecommunication:* Background noise is a common problem which degrades the quality of the communication for the human listener. Speech enhancement may be applied to such systems in order to remove the unwanted noise sources. Another problem in telecommunication is the channel noise. By enhancing the speech signal before it goes into the channel, it is also possible to reduce the effect of the channel noise.

- *Digital Hearing Aids:* The digital hearing aid users often complain of difficulty in understanding speech in the presence of background noise. Therefore, speech enhancement is an important process to improve the speech perception in a noisy environment for the human listener.

1.4. Thesis Overview

As discussed in section 1.2., there are many speech enhancement methods including the parametric and non-parametric, single and multi-channel approaches. Since the single channel algorithms depend only one source, they are more challenging compared to the multi-channel algorithms. However, due to its low cost and simplicity of single microphone systems, single channel speech enhancement is significantly important and attracts attention of the researchers interested in this field of study. This thesis is focused on a single channel and non-parametric speech enhancement method that does not require any a priori knowledge of the noise signal.

1.4.1. Problem Statement

As discussed above, many of the reported speech enhancement algorithms suffer from the residual noise problem often referred as the musical noise. In single channel speech enhancement, the residual noise is an inevitable issue; therefore the algorithms are mainly concerned to minimize its effect. Fourier transform and wavelet analysis have dominated the speech processing algorithms. However they both suffer in the analysis of non-stationary signals. Fourier is powerful for periodic signals and easy to implement but not suitable for non-linear and non-stationary signals like speech. Wavelet is more suitable for non-stationary signal analysis, but once the basic wavelet is selected, one should follow it to analyze the whole data. Moreover, since the most commonly used Morlet wavelet is Fourier based, it also suffers from many shortcomings of the Fourier analysis. Therefore, an analysis that is highly applicable to non-linear and non-stationary signals is desired.

In a noisy mixture, it is easier to remove the noise components from a frequency band where speech is not present. At a frequency band where both speech and noise are present, it is very hard to identify and remove the noise components without degrading the speech signal. Hard thresholding prefers not to remove the noise signal in those bands, while soft thresholding takes the risk of degrading the quality of the speech signal in order to remove some of these noise components. Therefore in the spectral domain it is not easy to identify and remove the noise parts whose frequencies are same as that of the speech components at a time instant. Therefore an analysis which will roughly separate the noise and speech in the time domain is desired.

The empirical mode decomposition (EMD), recently been pioneered by Huang *et. al.* 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) (5). Such functions give sharp and meaningful identifications of instantaneous frequencies. The IMFs may have frequency overlaps, but at any time instant the instantaneous frequencies represented by each IMF will be different. Therefore EMD is not band pass filtering, but is an effective decomposition of non-linear and non-stationary signals in terms of their local frequency characteristics. With this powerful property, in case of a noisy speech signal, EMD makes it possible to successfully separate the noise components that are imbedded in speech signals. Each IMF will still have speech and noise components but the intensity and the distribution in time will be different. Therefore it is possible to have an effectual identification of the noise components. In this thesis, we are proposing EMD based thresholding algorithms for speech enhancement.

1.4.2. Outline of the Thesis

To increase readability, the chapters are briefly described in this section and thereby providing an overview of the contents of the thesis.

- ***Chapter 2*** includes detailed information about Empirical Mode Decomposition (EMD) and gives some data analysis results in order to show the efficiency and superiority of EMD to Fourier transform and Wavelet analysis in analyzing the non-stationary signals.

- ***Chapter 3*** includes a DCT based soft thresholding algorithm for speech enhancement. The soft thresholding strategy introduced here is effective in noise removal while paying attention to the original speech. Our proposed methods are mainly based on this soft thresholding strategy.

- ***Chapter 4*** describes the proposed EMD based soft thresholding algorithm for speech enhancement. The soft thresholding strategy given in Chapter 3 is adapted to the IMFs of

the noisy speech signal with some modifications. The method is very effective in noise removal. Extensive experimental results prove the superiority of the proposed algorithm to other recently reported techniques in terms of both SNR improvement and speech quality. The major drawback of the algorithm is that it is mainly applicable to white noise.

- ***Chapter 5*** includes a hybrid DCT and EMD based soft thresholding algorithm. The DCT soft thresholding is applied as a pre-process to the noisy speech signal. The remaining noise components in the enhanced speech are further removed from its IMFs through an EMD based soft thresholding. In order to provide robustness to different noise types, a sub-band approach for the DCT domain is further given as a modification to the algorithm.

- ***Chapter 6*** describes an EMD domain speech enhancement method based on joint hard and soft thresholding criteria.

- ***Chapter 7*** includes the conclusion of the thesis work.

1.5. Publications derived from this work

Peer-reviewed conference papers

[1] E. Deger; M. K. Islam Molla; K. Hirose; N. Minematsu and M. K. Hasan, "Hard and soft thresholding with EMD for speech enhancement", *Proc. of International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08),* Gold Coast, Australia; 6-8 March, 2008.

[2] E. Deger; M. K. Islam Molla; K. Hirose; N. Minematsu and M. K. Hasan, "Speech enhancement using soft thresholding with DCT-EMD based hybrid algorithm", *Proc. of European Signal Processing Conference 2007 (EUSIPCO'07),* Poznan, Poland; 3-7 September 2007.

[3] E. Deger; M. K. Islam Molla; K. Hirose; N. Minematsu, "EMD based soft thresholding for speech enhancement", *Proc. of EUROSPEECH'07*, Antwerp, Belgium; 27-31 August, 2007.

Technical Reports

[1] E. Deger; M. K. Islam Molla; K. Hirose; N. Minematsu, "DCT-EMD based hybrid soft-thresholding technique with a sub-band approach for speech enhancement", *IEICE Technical Report* (ISSN 0913-5685), Vol.107, No. 235, pp. 7-12, September 2007.

Chapter 2

Empirical Mode Decomposition

2.1. Introduction

Data analysis is an essential part in pure research and practical applications. Linear and stationary processes are easy to analyze, but the real world signals are mostly non-linear and non-stationary. Analysis of such time varying data is not an easy process. Fourier spectral analysis has provided a general and easy method for examining the global energy-frequency distributions. As a result, the term *spectrum* has become almost synonymous with the Fourier transform of the data. The spectrum gives us the frequencies that exist over the entire duration of the data set. However, the main idea of time-frequency analysis is to understand and describe where the frequency content of the data is changing in time.

The time-frequency (TF) representation, a two-dimensional function which indicates the energy content of a signal as a function of both time and frequency, is a powerful tool for time-varying signals. Therefore, TF representation provides temporal and spectral information simultaneously. There exists a numerous number of TF representation methods of time domain signals, such as short-time Fourier transform (STFT), wavelet transform, Wigner-Ville distribution, evolutionary spectrum, empirical orthogonal function expression. Inside those, STFT and wavelet have dominated the time-frequency analysis in signal processing.

The STFT represents the short time, snapshot like spectral representation which is nothing but a limited time window-width Fourier spectral analysis. It is simply obtained by sliding a selected size window along the time axis and applying Fourier transform in each segment. However,

since it relies on the traditional Fourier spectral analysis, one has to assume the data to be piecewise stationary. Therefore, in case of non-stationary signals, the STFT has limited usage.

Currently, the most famous time-frequency analysis method is wavelet transform. Wavelet transform expands the signal in terms of wavelet functions which are localized in both time and frequency. The most commonly used wavelet is Morlet, defined as Gaussian enveloped sine and cosine wave groups with 5.5 waves (6). The problem with Morlet wavelet is the leakage generated by the limited length of the basic wavelet function, which makes the quantitative definition of the energy-frequency-time distribution difficult. Once the basic wavelet is selected, one has to apply it for the whole data (7). Moreover, since Morlet wavelet is Fourier based, it also suffers from many shortcomings of Fourier spectral analysis.

The most recently introduced technique for analyzing non-linear and non-stationary signals is the Hilbert Huang Transform (HHT), which is a combination of Empirical Mode Decomposition (EMD), recently pioneered by Huang et.al. (5) and Hilbert transform (HT). The key ingredient in HHT is the EMD which decomposes the signal into many modes with different frequency characteristics, called the intrinsic mode functions (IMFs), and thus also alleviates the problem of sharp frequency change in the original signal. IMFs are free of riding waves; hence they give sharp identifications of the instantaneous frequencies. Therefore they are highly suitable for Hilbert transformation. Once these IMFs are obtained, HT is applied on each IMF in order to obtain the time-frequency representation. Since EMD is specifically introduced for nonlinear and non-stationary signals, it has attracted the attention of the researchers from many areas soon after its introduction and has been implemented in numerous kinds of data, often proving its efficiency and superiority.

In this chapter, we will give brief information to Empirical Mode Decomposition and in order to show its efficiency in terms of analyzing the non-linear and non-stationary signals, a comparison with Fourier and Wavelet analysis will be given in spectral domain.

2.2. Empirical Mode Decomposition

Empirical mode decomposition (EMD) was recently developed by Huang *et al.* to decompose any non-stationary and nonlinear signal into oscillating components obeying some basic properties, called intrinsic mode function (IMFs).

2.2.1. Intrinsic Mode Functions

The principle of EMD technique is to decompose any signal $s(t)$ into a set of band-limited functions $C_n(t)$, which are zero mean oscillating components, simply called the IMFs. Each IMF satisfies two basic conditions: (*i*) in the whole data set the number of extrema and the number of zero crossings must be 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 (5). The first condition is similar to the narrow-band requirement for a Gaussian process and the second condition is a local requirement induced from the global one, and is necessary to ensure that the instantaneous frequency will not have redundant fluctuations as induced by asymmetric waveforms. The name intrinsic mode function is adopted because it represents the oscillation mode in the data. With this definition, the IMF in each cycle, defined by the zero crossings, involves only one mode of oscillation, no complex riding waves are allowed (5). IMF is not restricted to a narrow-band signal; it can be both amplitude and frequency modulated, in fact it can be non-stationary. A typical IMF can be observed in Figure 2.1.

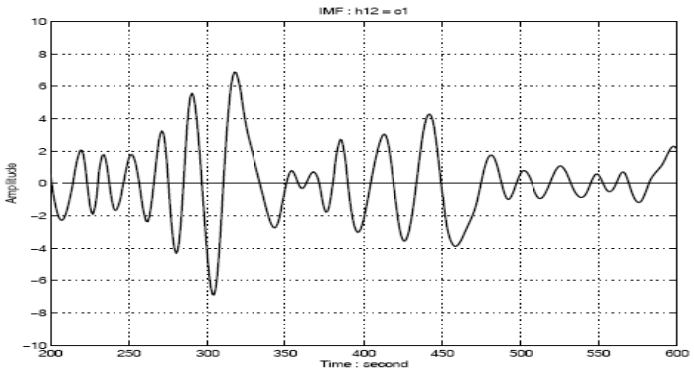


Figure 2.1: A typical IMF with the same numbers of zero crossings and extrema, and symmetry of the upper and lower envelopes with respect to zero.

The idea of finding the IMFs relies on subtracting the highest oscillating components from the data with a step by step process, which is called the sifting process.

2.2.2. Sifting Process

Although a mathematical model has not been developed yet, different methods for computing EMD have been proposed after its introduction (8, 9). The very first algorithm is called the sifting process. The sifting process is simple and elegant. It includes the following steps:

1. Identify the extrema (both maxima and minima of $s(t)$)
2. Generate the upper and lower envelopes ($u(t)$ and $l(t)$) by connecting the maxima and minima points by cubic spline interpolation
3. Determine the local mean $m_1(t)=[u(t)+l(t)]/2$
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 (8). The subsequent IMFs and the residues are computed as:

$$r_1(t) - C_2(t) = r_2(t), \cdots, r_{n-1}(t) - C_n(t) = r_n(t) \tag{2.1}$$

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

$$s(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \tag{2.2}$$

A noisy speech signal and some selected IMF components are shown in Figure 2.2. It can be observed that higher order IMFs contain lower frequency oscillations than that of lower order IMFs. This is reasonable, since sifting process is based on the idea of subtracting the component with the longest period from the data till an IMF is obtained. Therefore the first IMF will have the highest oscillating components; the components with the highest frequencies. Consequently,

the higher the order of the IMF, the lower its frequency content will be. However, the IMFs may have frequency overlaps but at any time instant the *instantaneous frequencies* represented by each IMF are different. This phenomenon can be well understood in Figure 2.3 which shows the instantaneous frequencies of the first 6 IMFs. Therefore EMD is not band pass filtering, but is an effective decomposition of non-linear and non-stationary signals in terms of their local frequency characteristics.

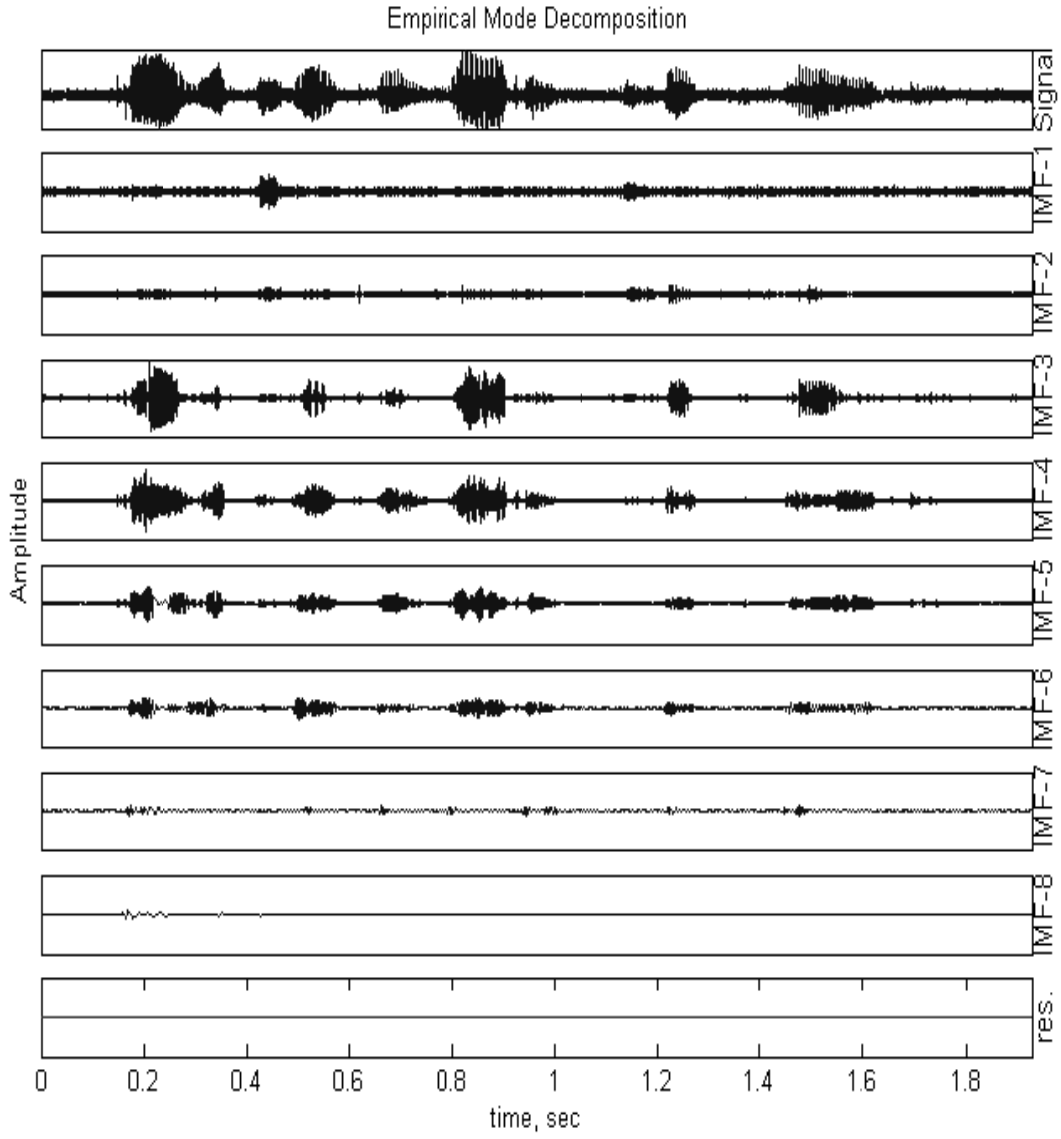


Figure 2.2: The illustration of EMD. A noisy speech signal at 10 dB SNR and its first 8 IMFs out of 14, plus a residue signal which can be observed to be close to a constant.

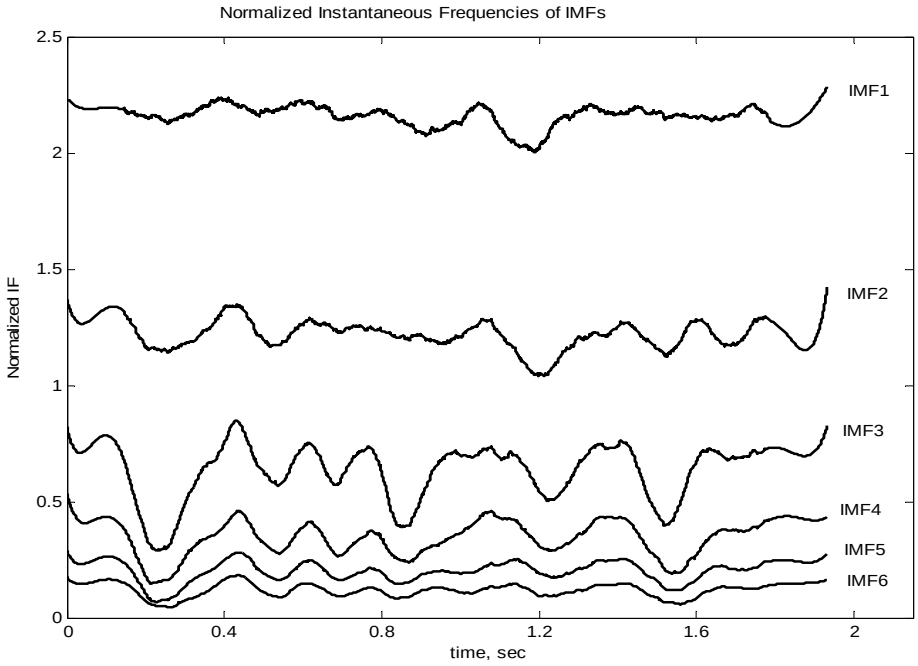


Figure 2.3: Instantaneous Frequencies of the IMFs.

2.2.3. Instantaneous Frequency

Instantaneous frequency (IF) represents signal's frequency at an instance and it is defined as the rate of change of the phase angle at the instant of the analytic (complex) version of the signal. Every IMF is a real valued signal. Analytic signal method (10) is used to compute the instantaneous frequency of the IMF components. The analytic signal corresponding to the m^{th} IMF $C_m(t)$ is defined as,

$$Z_m(t) = C_m(t) + j\hbar[C_m(t)] = \alpha_m(t)e^{j\theta_m(t)} \tag{2.3}$$

where $\hbar[.]$ refers to the Hilbert transform operator, $\alpha_m(t)$ and $\theta_m(t)$ are the instantaneous amplitude and phase respectively of the m^{th} IMF and j is the notation of complex term.

The Hilbert transform provides a phase-shift of $\pm\pi/2$ to all frequency components, whilst leaving the real parts unchanged (10). The analytic signal is advantageous in determining the instantaneous quantities such as energy, phase and frequency. Then the IF of the m^{th} IMF can easily be derived as,

$$\omega_m(t) = \frac{d\widehat{\theta}_m(t)}{dt} \tag{2.4}$$

where $\hat{\theta}_m(t)$ is the unwrapped version of the instantaneous phase θ_m(t). The concept of IF is physically meaningful only when applied to mono-component signals, which have been loosely defined as narrow band. To apply the concept of IF to arbitrary signals it is necessary to first decompose the signal into a series of mono-component contributions. Since EMD technique decomposes any time domain signal into a series of mono-component IMFs, the derivation of IF on each component provides the meaningful physical information.

2.2.4. Hilbert Spectrum

Hilbert Spectrum (HS) represents the distribution of the signal energy as a function of time and frequency. Having obtained the IMFs, to construct HS of the signal, the Hilbert transform is applied to each IMF and the instantaneous frequency is computed according to equation (2.4). After performing this calculation, the data can be expressed in the following form;

$$X(t) = \sum_{i=1}^{n} a_j\,(t) e^{\left(j \int \omega_j(t) dt\right)} \tag{2.5}$$

Here the residue is left out, since it is either a monotonic function or a constant. It can be observed that the Hilbert Huang Transform representation of the data in equation (2.5) gives both amplitude and the frequency as functions of time. This makes HHT highly efficient for analyzing non-stationary signals. Figure 2.4 illustrates the Hilbert Huang spectrum of the noisy mixture signal shown in Figure 2.2.

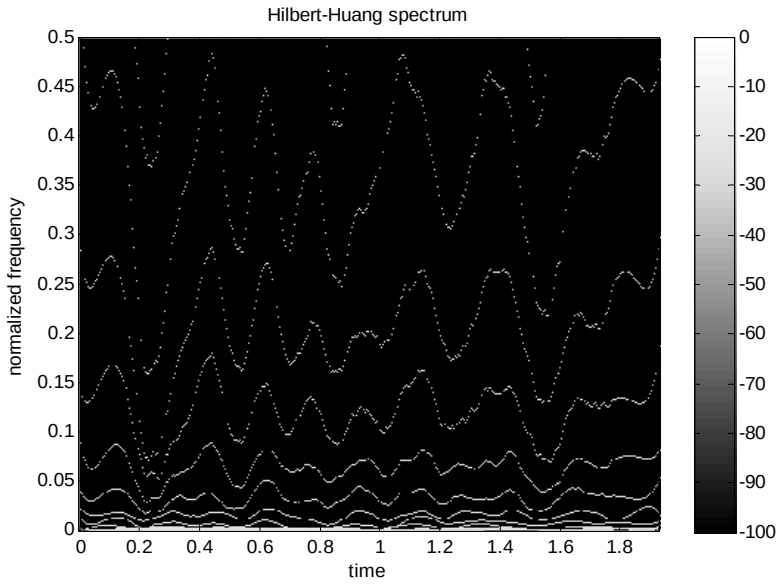


Figure 2.4: Hilbert Huang Spectrum of the noisy speech signal.

2.2.5. Completeness and Orthogonality

The original data can be easily reconstructed by summing up all the IMFs obtained in the decomposition as given in equation (2.2). When experimentally done, the difference between the reconstructed and original data is less than 5×10^{-15}, the roundoff error from the precision of the computer. Therefore EMD is a reversible representation which proves its *completeness.*

To measure the efficiency of EMD, the *orthogonality* of the decomposition should also be checked. Higher the orthogonality corresponds to less amount of information leakage between the IMFs. The orthogonality is satisfied in all practical sense, but is not guaranteed theoretically. By virtue of the decomposition, all the IMFs should be locally orthogonal to each other, for each element is obtained from the difference between the signal and its local mean through the maximal and minimal envelopes (5). Therefore;

$$\overline{\left(x(t) - \overline{x(t)}\right).\overline{x(t)}} = 0 \tag{2.6}$$

Nevertheless, since the mean is computed via the envelopes, hence is not true mean, equation (2.6) is not strictly true. Moreover, each successive IMF component is only part of the signal constituting $\overline{x(t)}$. Therefore leakage is unavoidable; however any leakage should be small. To check the orthogonality of IMFs, an overall index orthogonality, *IO*, is defined as;

$$IO = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{x^2(t)}\left(\sum_{l=1}^{M+1}\sum_{m=1}^{M+1} C_l(t) * C_m(t)\right) \tag{2.7}$$

where *l* and *m* refers to the indices of the IMFs. Since the residue signal is also included, the index goes to *M+1* instead of *M*, the number of the IMFs. For the decomposition to be orthogonal, *IO* should be zero. The theoretical value of the *IO*, given by Huang et.al., is less than 0.001, therefore very close to zero (5). The orthogonality therefore can be said to be satisfied. Figure 2.5 illustrates the index of orthogonality values between all possible pairs of IMF components of the noisy mixture signal shown in Figure 2.2. It can be observed that the maximum value of the indices of orthogonality is in the order of 9×10^{-3} and the overall index of orthogonality is 0.054 which is close to zero.

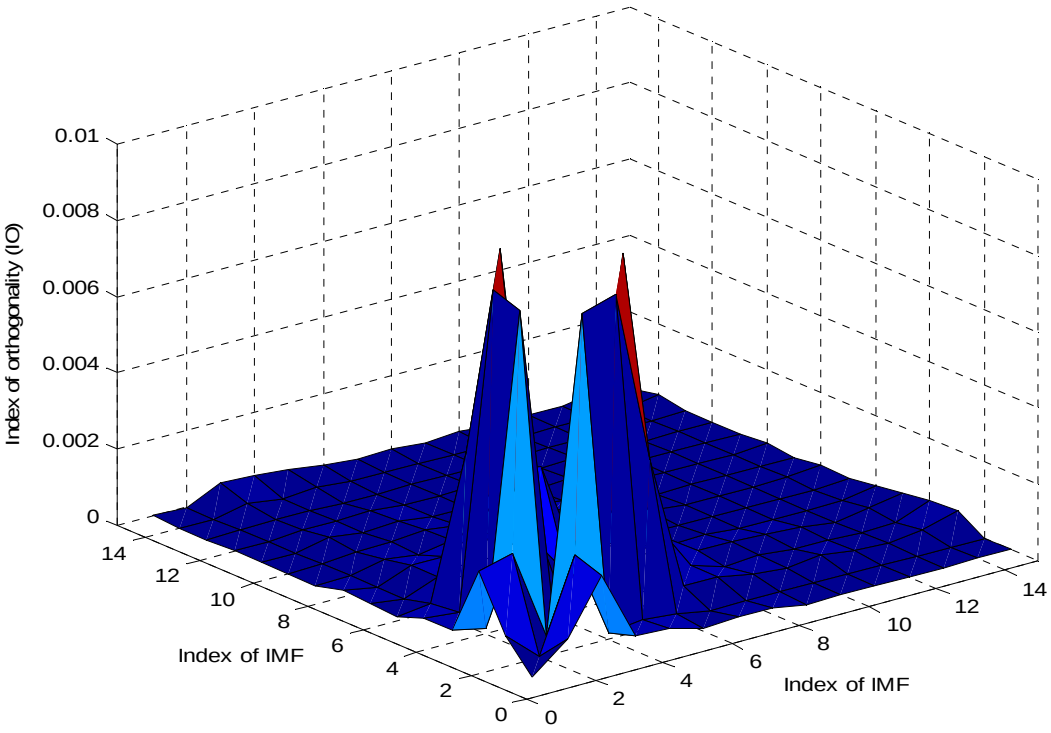


Figure 2.5: The values of indices of orthogonality between all possible pairs of IMFs.

2.3. Comparison of HHT, Fourier and Wavelet

It can be observed that the Hilbert Huang Transform representation of the data in equation (2.5) gives both amplitude and the frequency as functions of time. This makes the HHT highly efficient for analyzing non-stationary signals. The same data if expanded in Fourier representation would be;

$$X(t) = \sum_{i=1}^{n} a_j \, e^{j\omega_j t} \tag{2.8}$$

where both a_j and ω_j are constants. The contrast between the two equations is clear: the IMF represents a generalized Fourier expansion. The variable amplitude and the instantaneous frequency have not only greatly improved the efficiency of the expansion, but also enabled the expansion to accommodate non-stationary data. With IMF expansion, the amplitude and the frequency modulations are also clearly separated. Thus, we have broken through the restriction of the constant amplitude and fixed-frequency Fourier expansion, and arrived at a variable amplitude and frequency representation. This expression is numerical. If a function is more desired, an empirical polynomial expression can be easily derived from the IMFs (5).

In order to observe the superiority of the HHT to the STFT, an illustration of time-frequency analysis of a frequency modulated (FM) signal is given here. Figure 2.6 shows an FM signal with a sampling rate of 1 kHz. The STFT spectrum for the FM signal is given in Figure 2.7(a) with 256 point FFT and a Hamming window of 30ms with 60% overlap. It can be observed that with even such a stationary FM sinusoidal signal, the STFT has produced a band of energy with a remarkable amount of cross-spectral terms. On the other hand, the Hilbert Huang spectrum of the signal, depicted in Figure 2.7(b), gives extremely sharp identifications spectral components.

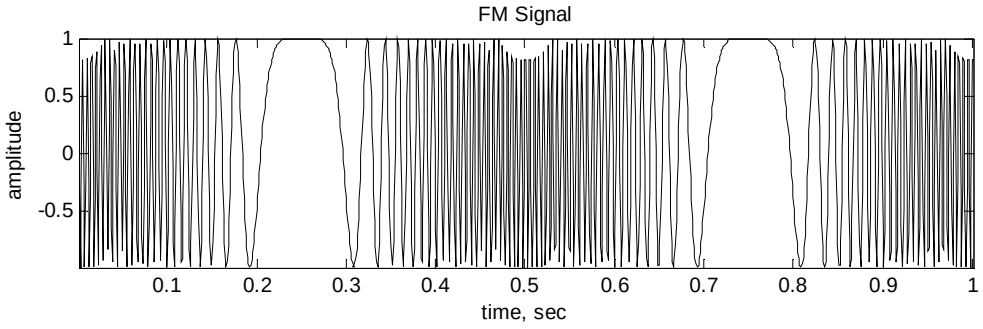


Figure 2.6: A frequency modulated signal carrying a 2 Hz sinusoid with a carrier frequency of 100 Hz at a sampling rate of 1 kHz.

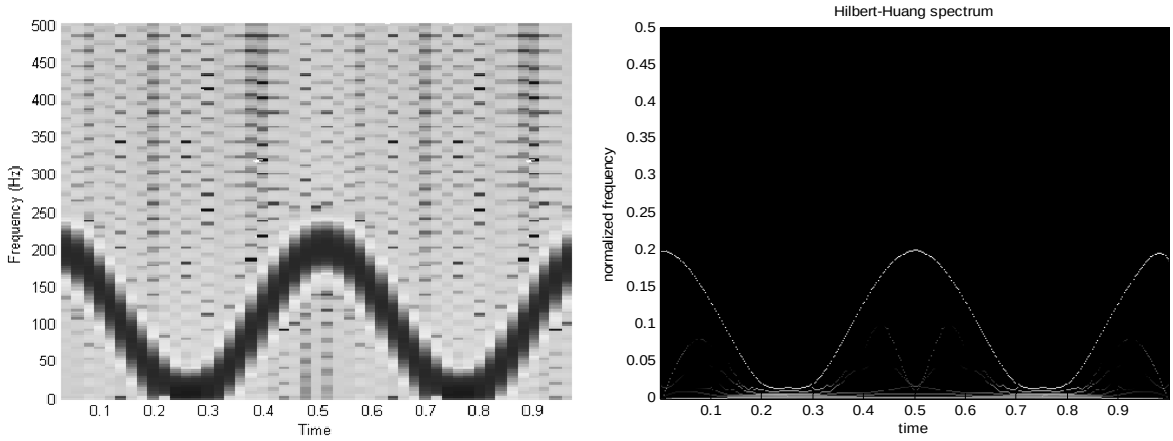
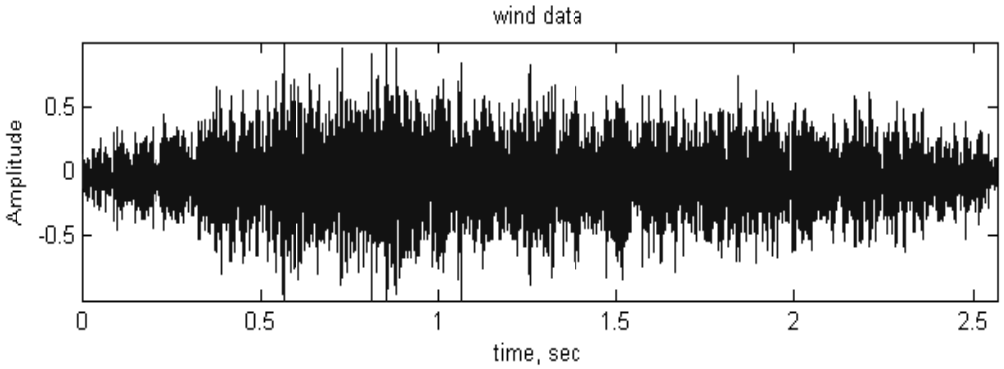


Figure 2.7: Comparison of HHT and STFT spectrum. a) STFT spectrum of the FM signal with, 256 point FFT (Hamming window 50ms with 60% overlap) b) Hilbert Huang Spectrum obtained by applying the Hilbert transform to the IMFs of the FM signal.

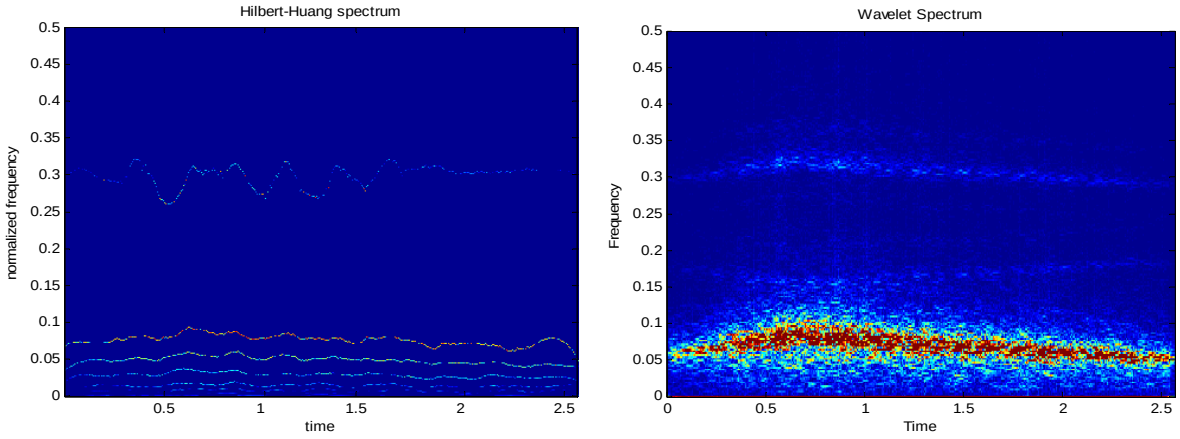
Wavelets have been widely used in analysis of the non-stationary data. Indeed, it is very useful for analyzing data with gradual frequency changes. As discussed before, the most commonly used Morlet wavelet is Fourier based, therefore it suffers many shortcomings of the Fourier spectral analysis. However the main problem with wavelet is its leakage generated by the limited length of the basic wavelet function, which makes the quantitative definition of the time-frequency-energy distribution difficult. Moreover, sometimes the interpretation of wavelet can also be counter-intuitive. For instance, one must analyze the result in the high frequency range in order to define a change occurring locally, since the basic wavelet is more localized in the higher frequencies. Therefore, if a local event occurs in low frequency range, in order to observe its effects, one will still have to look to the high frequency range. Another difficulty of the wavelet analysis is its non-adaptive structure, once the basic wavelet is selected, it must be used to analyze all the data, which is against the nature of the non-stationary signals. Despite these problems, Wavelets have been paid enormous attraction for the analysis of non-stationary data.

The Hilbert Huang Transform, as given in equation (2.5) represents the non-stationary data in terms of a small number of IMF components with well defined instantaneous frequencies. Being data-adaptive and local, it is highly effective for analyzing non-linear signals. As a result, HHT spectrum gives much sharper time-frequency-energy distribution than the Wavelet analysis. As an illustration to observe the superiority of HHT, a wind data and its HHT and wavelet spectrums can be observed in Figure 2.8. It can be observed that HHT gives very sharp

identification of the frequencies at all time instants. The wind energy is distributed in skeleton lines, representing each IMF. However, the spectrum obtained by wavelet is very blurred. Wind energy appears in smoothed contours with a rich energy distribution in the high harmonics.



a)



b) **c)**

Figure 2.8: Comparison of HHT and Wavelet spectrum. a) Wind data b) Hilbert Huang Spectrum obtained by applying the Hilbert transform to the IMFs of the FM signal. c) The Morlet wavelet spectrum for the wind data with 200 frequency cells.

The major drawback of EMD comes from its computational cost. As the name suggests, the IMFs are found through an empirical process not through a mathematical expression. That is why; EMD is not efficient for real time processes. However, many researchers are working to derive a mathematical expression and long steps have already been taken (11, 12).

2.4. Empirical Mode Decomposition for Speech Signals

Due to its efficiency in decomposing the non-stationary signals in terms of zero-mean oscillating components with well behaved instantaneous frequencies, EMD has been adapted to almost all kinds of data analysis soon after its introduction, always proving its efficiency (13). Therefore EMD has made a new and effective path for many signal processing research areas. Speech processing is one of these fields that EMD has successfully been applied (14, 15, 16, 17, 18).

As explained in Section 2.2., the idea of finding the IMFs relies on subtracting the highest oscillating components from the data, called the sifting process. Therefore the IMFs have different frequency characteristics; the upper the IMF, the higher its frequency content. The IMFs may have frequency overlaps but at any time instant the instantaneous frequencies represented by each IMF is different, the upper one having the higher frequency. With these powerful characteristics, recent studies have shown that it is possible to successfully identify and remove a significant amount of the noise components from the IMFs of a noisy speech. Although all IMFs contain energy from both the original speech and the noise, the amount of the energy distribution is different. Since speech signals are mainly concentrated in the low and mid frequency bands, the high frequency noise components dominate the first IMFs. For instance, in case of white noise, most of the noise components are centered on the first three IMFs, while the speech signals dominate between 3^{rd} and 6^{th} IMFs, as can be observed in Figure 2.2. Therefore, EMD makes it possible to at some extent separate the high frequency noise from the major speech components.

In this thesis we have shown that by applying a thresholding algorithm, it is possible to successfully eliminate the noise components from each IMF. Since we do not want to degrade the original speech while having an effective noise removal, a frame based soft thresholding strategy proposed by (19) has been adapted to the IMFs, with some modified criteria. Therefore before further going on the proposed algorithms, to make the flow of thesis more comprehensible, it would be appropriate to give brief information about the algorithm given in (19).

Chapter 3

DCT Soft Thresholding

3.1. Introduction

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. Transform domain speech enhancement methods commonly use amplitude subtraction based soft thresholding defined by (4, 20);

$$\hat{X}_k = \begin{cases} sign(X_k)(|X_k| - \sigma_v), & if|X_k| < \sigma_v \\ 0 & , otherwise \end{cases} \tag{3.1}$$

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. This degradation results in a loss in speech quality. Together with the residual noise problem discussed in Section 1.2, the enhanced speech may be even less desired than the noisy mixture.

Unlike the conventional constant noise-level subtraction rule in equation (3.1), a new soft thresholding strategy was proposed in (19). The later one is capable to remove the noise components while giving significantly less damage to the speech signal through a linear vector thresholding instead of a constant.

The strategy depends on segmenting the signal into short time intervals and applying Discrete Cosine Transform (DCT) on each frame. The DCT coefficients represent the whole frequency band within that interval. The signal is divided into frequency bands, therefore each time interval is represented by frequency bins. The frequency bins are categorized as either signal or noise dominant depending on its speech and noise energy distribution. Figure 3.1 shows an illustration of a typical noise and speech dominant frequency bins. The problems of the conventional contant noise level subtraction rules given in (3.1) can be well observed in this figure. For instance, it is apparent from Figure 3.1(a) that subtracting a constant value from the noisy speech coefficients in order to obtain the clean speech coefficients is inadequate. Furthermore, due to the second part of thresholding a significant amount of speech information may be lost, resulting as a source of musical noise. Therefore a linear thresholding is followed in noise dominant frames. On the other hand, Figure 3.1(b) proves that soft thresholding is very inaccurate for signal dominant frequency bins and will most probably degrade the speech components, therefore giving more damage than its contribution to the enhanced speech. Therefore, the signal dominant frames should better be kept as they are in order not to degrade the high energy speech components. This enables even signal with high SNRs to be processed effectively.

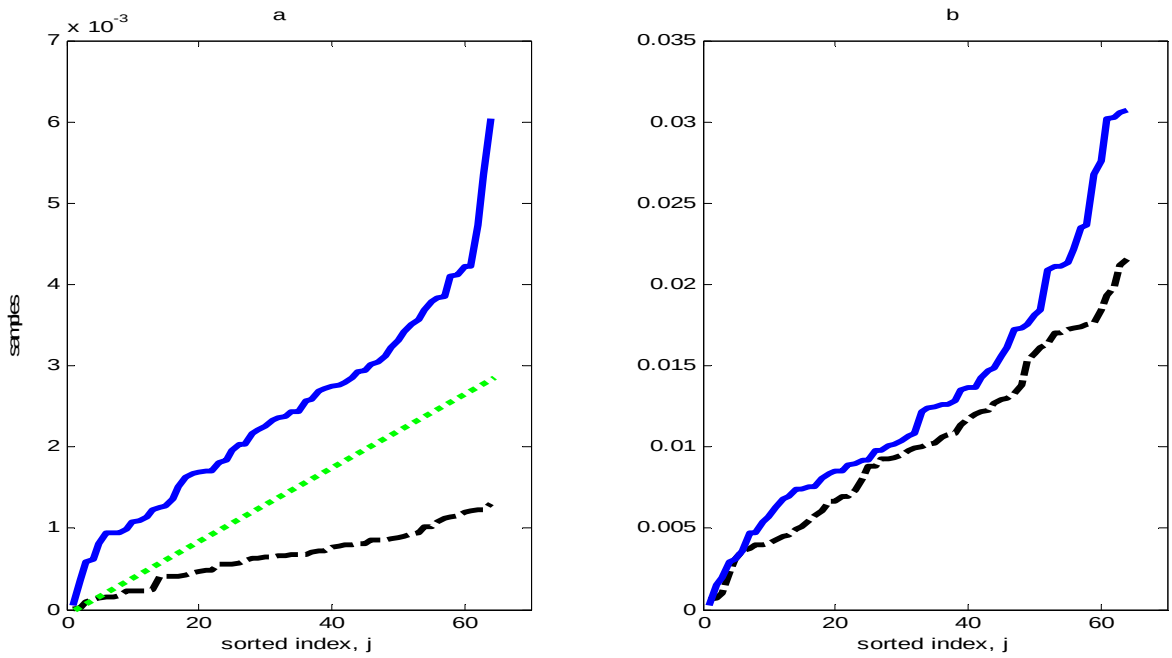


Figure 3.1: A typical; a) noise dominant bin, b) signal dominant bin.

3.2. Frequency Bin based DCT Soft Thresholding

The noisy speech is segmented into 32 ms frames and a 512 point DCT is applied on each frame. The DCT coefficients of the frames are further divided into 8 frequency bins, each containing 64 DCT coefficients. For adaptive thresholding, each bin is categorized as either signal or noise-dominant, as discussed in Section 3.1. The classification pertains to the average noise power associated with that particular bin. If for the *i*'th bin

$$\frac{1}{64}\sum_{k=1}^{64}\left|X_k^{(i)}\right|^2 \geq \sigma_n^2 \tag{3.2}$$

where σ_n^2 denotes the variance of the noise and $X_k^{(i)}$ is the *k*'th DCT coefficient of the *i*'th frequency bin, then this bin is characterized as signal dominant, otherwise as noise dominant. The signal dominant bins are not thresholded, since it is highly possible to degrade the speech signal, especially for high SNRs. In case of a noise dominant bin, absolute values of the DCT coefficients are sorted in ascending order and a linear thresholding is applied:

$$\hat{X}_k = sign(X_k)[max\{0, (|X_k| - mj)\}] \tag{3.3}$$

where the multiplication *mj* is a linear threshold function as can be observed in Figure 3.1(a) while *j* being the sorted index-number of $|X_k|$. An estimated value of *m* can be obtained as

$$m = \frac{\sigma_{-n}}{\sqrt[2]{\frac{1}{64}\sum_{k=1}^{64} k^2}} \tag{3.4}$$

where $\sigma_{-n} = \lambda\, \sigma_n$. It is evident from equation (3.2) that for the noise-dominant frequency bins, the average noise power added would be less than the average noise power estimated over the entire speech signal. Here, the added average noise power over any of these frequency bins is denoted as σ_{-n}. To find a reasonable value for $\lambda = (\sigma_{-n} / \sigma_n)$, three speech signals contaminated with white noise at 10dB SNR are used. Using the strategy in equation (3.3) at each frequency bin, the noise dominant ones are identified and the value of λ is calculated by simply dividing the variance of that frequency bin with the overall noise variance. The sorted variation of λ is shown in Figure 3.2. It can be observed that the value of λ vary between 0.2 and 0.8 for all of the speech signals.

Therefore, experimentally, the value of λ should be selected in this range. For signals with low SNRs, the higher λ gives better SNR improvement. On the other hand, for signals with high SNRs, a lower value results in better performance. The spectrograms of the clean, noisy and enhanced speech signals for λ=0.5 and λ=0.8 can be observed in Figure 3.3. It can be observed that with λ=0.8, the noise removal is more however the speech degradation is also higher, hence effecting the speech quality. Therefore, the choice of λ should depend on subjective and objective purposes.

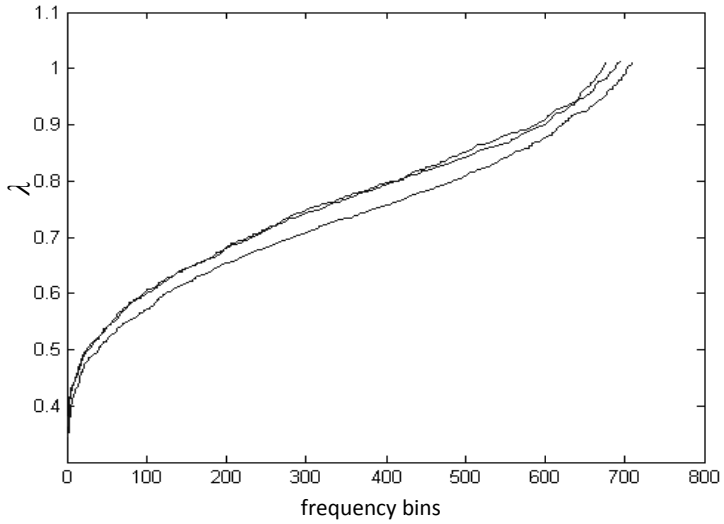


Figure 3.2: The calculated value of λ in the noise dominant frequency bins.

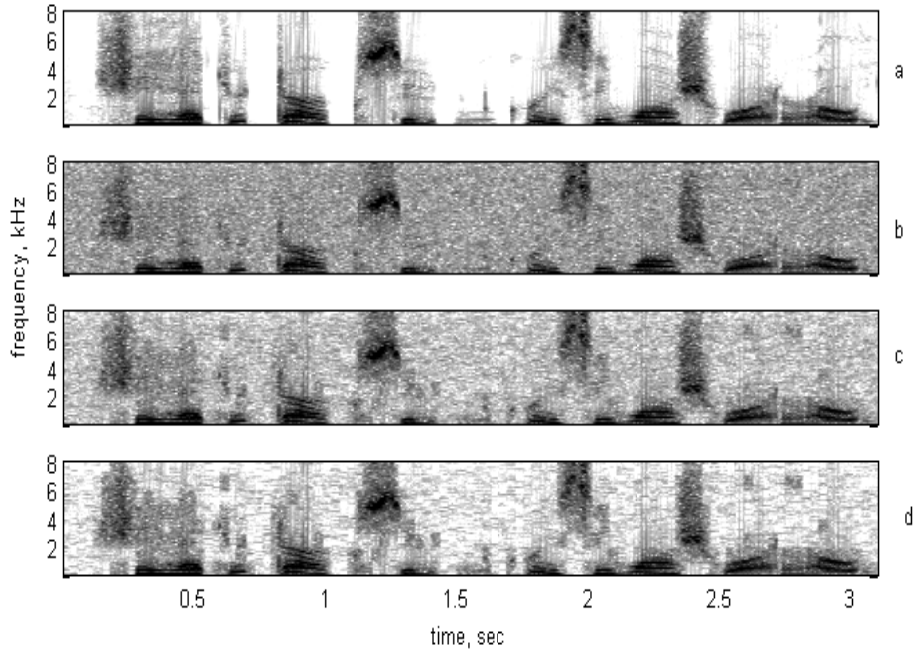


Figure 3.3: The spectrogram of a) clean, b) noisy and enhanced speeches c) for *λ=0.5*, d) for *λ=0.8*.

Chapter 4

EMD Based Soft Thresholding

4.1. Introduction

The empirical mode decomposition (EMD), recently been pioneered by (5) as a novel and powerful data analysis method for nonlinear and non-stationary signals has made a new and effective path for speech enhancement studies. As given in Chapter 2, EMD is a data-adaptive decomposition method with which any complex signal can be decomposed into zero mean oscillating components, the intrinsic mode functions (IMFs). Recent studies have shown that with EMD, it is possible to successfully remove the noise components from the IMFs of the noisy speech.

Soft thresholding strategy proposed by (19) is a powerful technique for removing the noise components from the noisy signal while paying attention on the original speech. Since the signal dominant frames are not thresholded, the algorithm enables even signals with high SNRs to be processed effectively, where most reported methods even fail to hold on to the input SNR. However, this is also a drawback, since it is not possible to efficiently remove the noise components that are embedded in the higher energy speech components. Since the categorization of the frequency bins depend on the global noise variance, many noise dominant bins can be identified as signal dominant due to the fluctuations in the noise variance of the frequency bins. As a result, the remaining noise components from both the noise and signal dominant frames will result in an irritating musical noise. In addition to this, another disadvantage of the method

comes from the categorization in spectral domain. Since the DCT coefficients of each frame are divided into 8 frequency bins thus forming 8 sub-bands, we may have a sharp increase between a thresholded noise dominant frequency bin and a non-thresholded signal dominant frequency bin in the spectral domain. This also results in an irritating effect in the musical noise. All these drawbacks can be significantly reduced with the proposed EMD based soft thresholding strategy.

In this chapter, we illustrate a novel speech enhancement method based on applying the soft thresholding algorithm with EMD. The proposed method is significantly effective in noise removal. Since the extraction of the IMFs relies on frequency characteristics, the IMFs with higher index contain lower frequency components. This property helps the noise and speech components to be roughly separated in terms of frequency and to dominate in different IMFs. For instance, in case of white noise, the noise components dominate in the first few IMFs, mainly in the first one and the speech components dominate in the later IMFs. Therefore, the noise parts that are embedded in speech signals can also be extracted and thresholded. In order to identify the noisy frames in an efficient way, the proposed method also includes a modification in the soft thresholding strategy and a specific approach for each IMF of the noisy speech.

4.2. EMD based Soft Thresholding

First of all, EMD is applied to the noisy speech in order to obtain the IMFs of the signal. In DCT soft thresholding given in Chapter 3, the signal is divided into 32 ms frames and further divided into 8 frequency bins with 64 data each. EMD is not a transformation from time domain to spectral domain; rather it is a decomposition of the time domain signals into IMFs, which are also time domain signals. Therefore, the obtained IMFs are divided into 4 ms frames, thus each having 64 data for a 16 kHz sampling frequency and instead of the name frequency bin, it is more appropriate to use the definition 'frame' to refer to the time frames.

Similar to the DCT case, these frames are characterized as either a signal dominant or a noise dominant frame. However for categorizing the frames, unlike the limit defined in (3.2), a novel strategy is introduced here. This new soft-thresholding strategy provides an effective limit for the 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 and the enhanced speech is obtained from the thresholded IMFs. A block diagram of the algorithm can be observed in Figure 4.1.

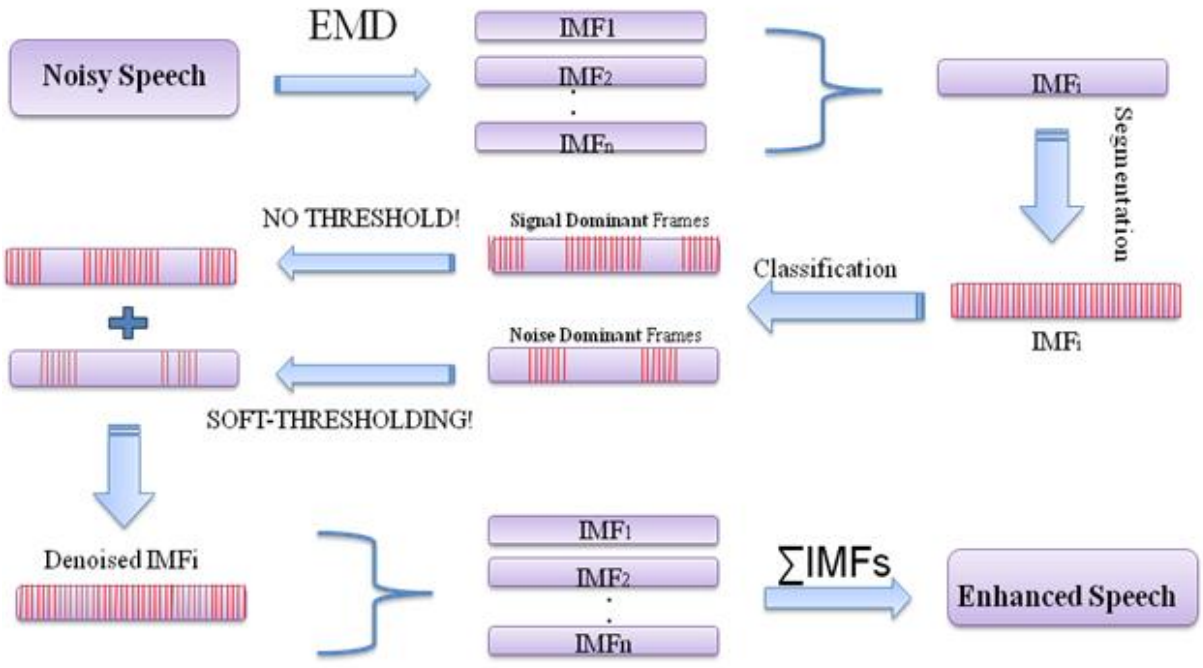


Figure 4.1: The block diagram of the proposed EMD soft thresholding method.

4.2.1. A Novel Soft Thresholding Strategy

The categorization of the 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 frames will have more fluctuations than that of the noisy speech frames. Therefore the noise variance of each IMF should be defined separately and the limit for frame categorization should have a larger value then the limit defined in (3.2), in order to guarantee that all the noisy frames are thresholded.

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

$$\sigma^2_{(frame)} = \sigma^2_{(sf+nf)} \tag{4.1}$$

$$\textit{thus}, \qquad \sigma^2_{(frame)} = \sigma^2_{sf} + \sigma^2_{nf} + 2 * Cov(s,n) \tag{4.2}$$

where $\sigma^2_{(frame)}$ denotes the noise variance of a frame, and σ^2_{sf} and σ^2_{nf} refer to the speech and noise variance within that frame, consecutively. In case of independence of speech and noise, the covariance between the two will be zero, thus we have;

$$\sigma^2_{(frame)} = \sigma^2_{sf} + \sigma^2_{nf} \tag{4.3}$$

For equal noise and speech power, we get;

$$\sigma^2_{(frame)} = \sigma^2_{sf} + \sigma^2_{nf} \quad \overset{\sigma^2_{sf}=\sigma^2_{nf}}{\Longrightarrow} \quad \sigma^2_{(frame)} = 2\sigma^2_{nf} \tag{4.4}$$

Therefore, in case of equal noise and speech power, with the assumption of independency, the variance of a frame is equal to twice the noise variance. That is why; the limit for the

categorization of frames should be set to two times of the globally estimated noise variance. With the proposed strategy, if for the *i*'th frame;

$$\frac{1}{64}\sum_{k=1}^{64}\left|X_{k,m}^{i}\right|^{2} \geq 2\sigma_{n,m}^{2} \tag{4.5}$$

where $X_{k,m}$ denotes the samples of the frames of the *m*'th IMF and $\sigma_{n,m}^{2}$denotes the estimated noise variance of this IMF; then this frame is defined as signal dominant, otherwise as a noise dominant frame. Signal dominant frames are not thresholded. In case of a noise dominant frame, absolute values of the frame samples are sorted in ascending order and a linear thresholding as in (3.3) is applied;

$$\hat{X}_k = sign(X_k)[max\{0, (|X_k| - mj)\}] \tag{4.6}$$

Here $\hat{X}_k$ refers to the thresholded samples and 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 as;

$$m = \frac{\sigma_{-n,m}}{\sqrt[2]{\frac{1}{64}\sum_{k=1}^{64} k^2}} \tag{4.7}$$

where $\sigma_{-n,m} = \lambda\sigma_{n,m}$. As discussed in Chapter 3, a reasonable value of λ is between 0.2 and 0.8.

4.2.2. Variance of the IMFs

The estimation of the variance of each IMF plays an important role in the performance of the proposed EMD domain soft thresholding algorithm. In order to estimate the variance, 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 the IMFs can be estimated from these speechless parts of the array. Figure 4.2 shows a plot of the variance of the frames for the first 6 IMFs of a noisy speech signal at 10dB. The differences in between the noise variance and the length of the speechless parts of the IMFs can be observed in Figure 4.2. It can be observed that 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 can have a very good estimation of the noise variance of each IMF. By using the IMF's specific noise variance, with the proposed soft thresholding algorithm, the noise components in all the IMFs can effectively be removed.

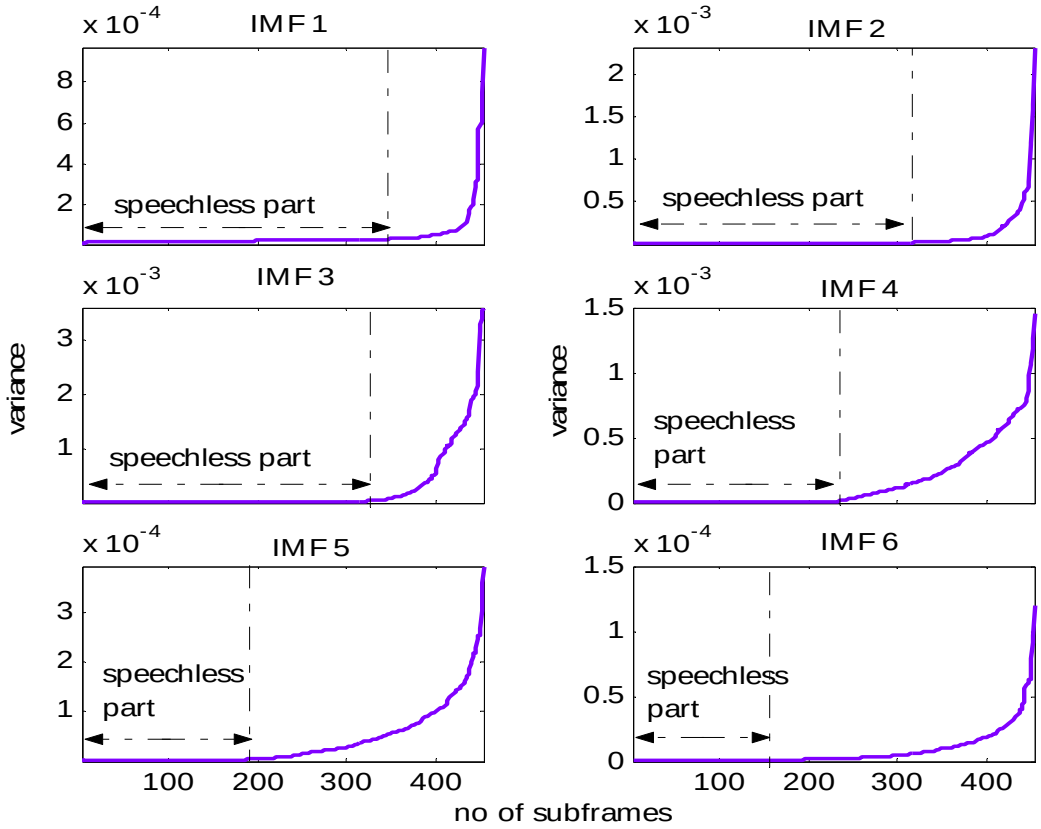


Figure 4.2: Sorted variance of 4ms frames for the first 6 IMFs of a noisy speech at 10dB SNR.

4.3. 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, weighted white noise samples from NOISEX database were added to the clean speech signal to obtain the noisy signals at different SNRs. White noise is considered here, since it is the most common type of noise in real world applications and it has been reported that this type of noise is more difficult to detect and remove than any other type (21). 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.

For a better understanding of the algorithm, Figure 4.3(a) shows the waveforms of the first 5 IMFs of the noisy speech signal "Don't ask me to carry an oily rag like that." from the TIMIT database. The corresponding denoised IMFs are illustrated in Figure 4.3(b).

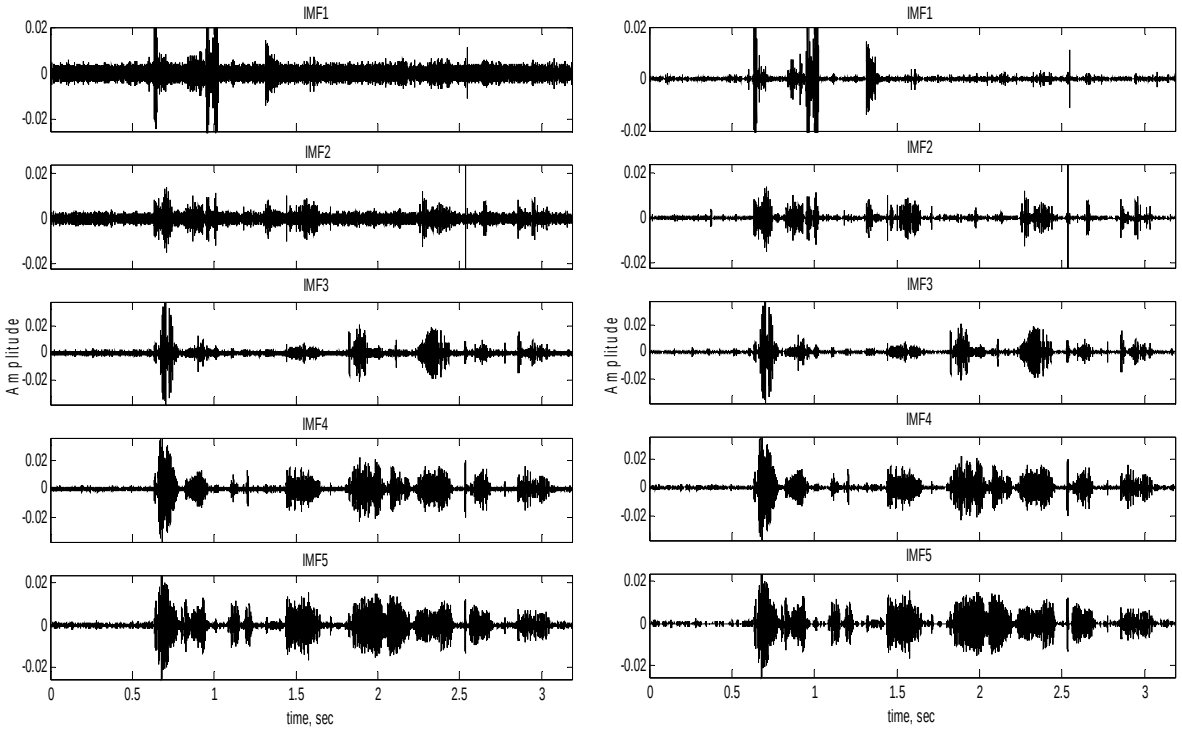


Figure 4.3: The waveforms of a) the first 5 IMFs of the noisy speech signal, b) the corresponding thresholded IMFs.

Figure 4.4 illustrates the spectrograms and the waveforms of the clean speech, noisy mixture at 10 dB SNR and recovered speech signals for different denoising algorithms and the proposed scheme. Here in order to observe its effect, the results are given for two values of λ (λ=0.5 and λ=0.8) both for DCT soft thresholding and proposed EMD based thresholding methods. For both values of λ, it can be observed that the proposed algorithm significantly performs better than the DCT soft thresholding algorithm. For λ=0.5, there is still significant remaining noise in the enhanced speech. On the other hand, for λ=0.8, although the noise components are effectively removed, there is some speech degradation in low energy speech components. Therefore, we can conclude that the choice of λ should be somewhere between these values in order to have a better performance. However, we will observe that the optimum value of λ differs depending on the SNR of the noisy mixture signal.

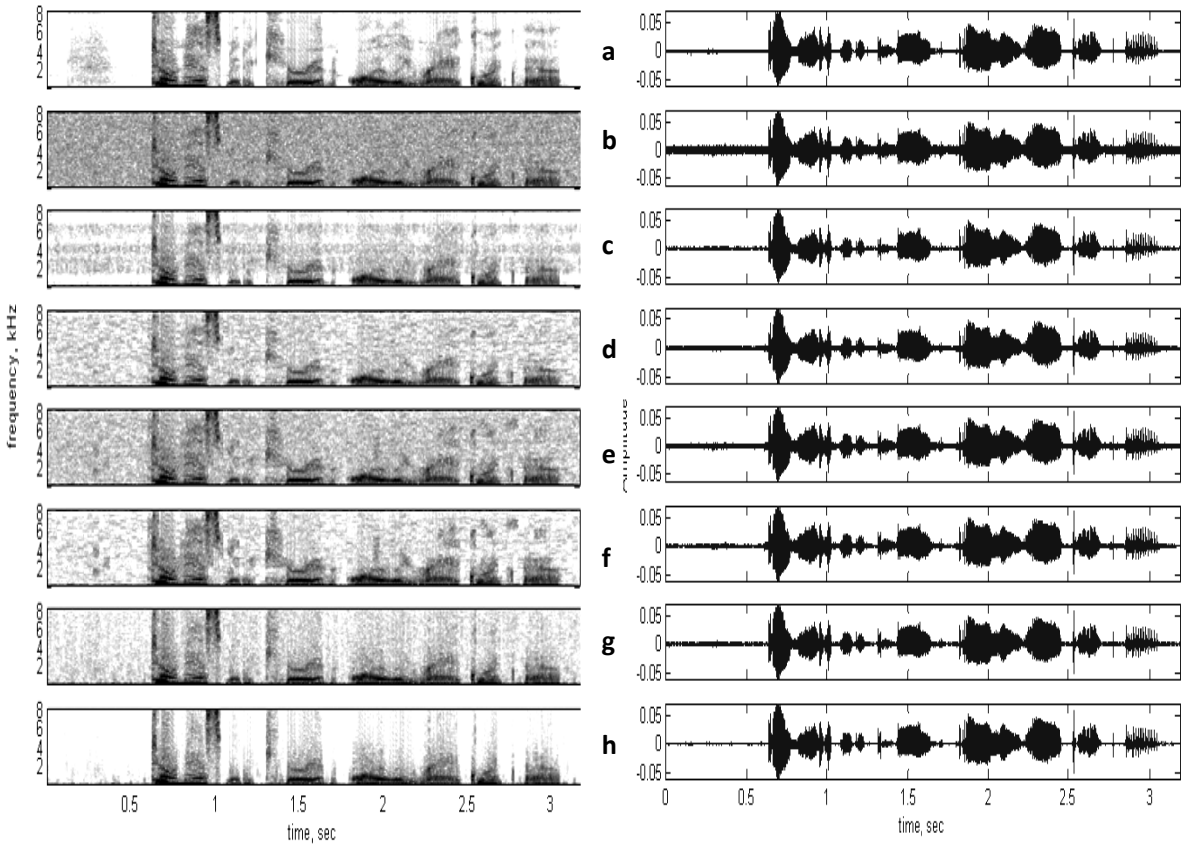


Figure 4.4: The spectrogram of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λ=0.5), f) DCT soft thresholding (λ=0.8), g) proposed method (λ=0.5), h) proposed method (λ=0.8)

The average SNR improvement results of the computer simulations for 10 male and 10 female utterances randomly selected from TIMIT database are listed in Table 4.1 for a wide range of SNR values. The superiority of the proposed scheme can be well observed in this table. The proposed algorithm is performing significantly better than other reported methods. 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 main reason is, for high SNRs, the noise power is significantly less compared to the speech power. Therefore it is much harder to identify and remove the noise components without degrading the speech signal. 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, they are effectively denoised.

In Table 4.1, the dependency of the optimum value of λ on the input SNR can be well observed. It is better to have a higher λ for low input SNR values, and to have a lower λ for high input SNRs. That is why, before giving further experimental results, an empirically defined optimum value for λ depending on the estimated input SNR is proposed here.

Table 4.1: *Comparison of the SNR improvements of different denoising methods.*

Input SNR (dB)	**Output SNR (dB)**					
	WP (20)	DCT (22)	Soft DCT (λ=0.5) (19)	**Proposed EMD (λ=0.5)**	Soft DCT (λ=0.8) (19)	**Proposed EMD (λ=0.8)**
0	4.86	5.69	5.33	**5.67**	6.49	**7.01**
5	8.86	9.76	9.67	**10.14**	10.04	**10.41**
10	12.36	13.74	13.75	**14.12**	13.45	**14.06**
15	15.45	17.65	17.93	**18.15**	17.56	**18.01**
25	20.82	25.53	26.35	**26.78**	26.03	**26.35**
30	23.16	29.52	30.56	**31.28**	30.28	**30.89**

4.4 Optimum value of λ

The soft-thresholding algorithm can be further improved by defining an optimum value for λ. As we discussed, it is better to have a higher value of λ for low SNRs and a lower value for high SNR input signals. This dependency of λ on the input SNR can be better observed in Figure 4.5, which shows the effect of λ on the SNR improvement results at different input SNRs. Therefore, the optimum value of λ can be related with an estimated value of the input SNR. An estimation of the SNR of the noisy speech signal can be obtained as explained below.

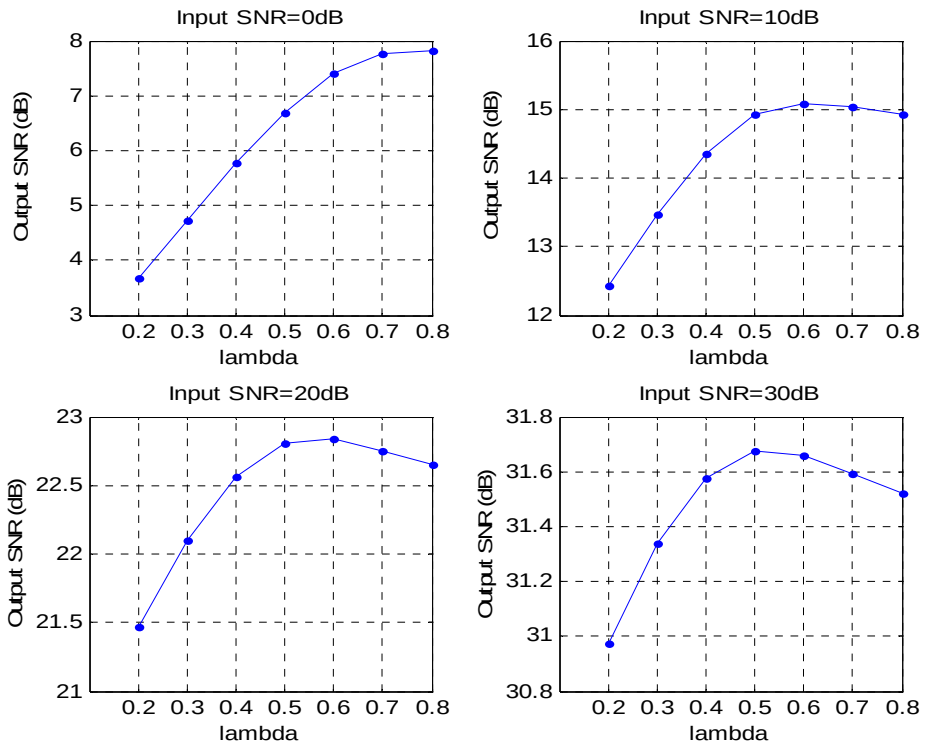


Figure 4.5: The effect of λ on the SNR improvement results at different input SNRs.

Similar to the noise variance estimation of the IMFs, the SNR estimation of the noisy speech signal –the input SNR- can be performed by segmenting the signal into frames. In order to achieve a reasonable estimation of the input SNR, the noisy speech signal is divided into 4 ms frames and the variance of each frame is stored in a variance array which is sorted in ascending order. The noise variance of the noisy speech, σ^2_{noise}, can be estimated from the speechless parts of this array.

Similar to (4.3), due to the independency of noise and speech, the variance of the noisy mixture ($\sigma^2_{mixture}$) is equal to the sum of the speech variance and noise variance;

$$\sigma^2_{mixture} = \sigma^2_{speech} + \sigma^2_{noise} \tag{4.8}$$

thus,

$$\sigma^2_{speech} = \sigma^2_{mixture} - \sigma^2_{noise} \tag{4.9}$$

The input SNR can be estimated by

$$SNR_{input} = 10\log\left(\frac{\sigma^2_{speech}}{\sigma^2_{noise}}\right) \tag{4.10}$$

Substituting (4.9) into (4.10) gives

$$SNR_{input} = 10\log\left(\frac{\sigma^2_{noisy} - \sigma^2_{noise}}{\sigma^2_{noise}}\right) \tag{4.11}$$

After extensive computer simulations, the optimum value of λ is obtained as

$$\lambda_{opt} = 0.7 - 0.01 * SNR_{input} \tag{4.12}$$

4.5. Experimental Results and Discussions

To illustrate the effectiveness of the proposed algorithm with the optimum value of λ 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 SNRs, weighted white noise samples from NOISEX database were added to the clean speech signal to obtain the noisy signals at different SNRs. For evaluating the performance of the method, overall and average segmental SNR improvements as well as objective speech quality results were used. The speech quality of the enhanced signals has been tested with the Perceptual Evaluation of Speech Quality (PESQ) algorithm provided by OPTICOM (23). PESQ provides accurate results for speech quality and is widely considered and used as the best algorithm as an estimation of a subjective test.

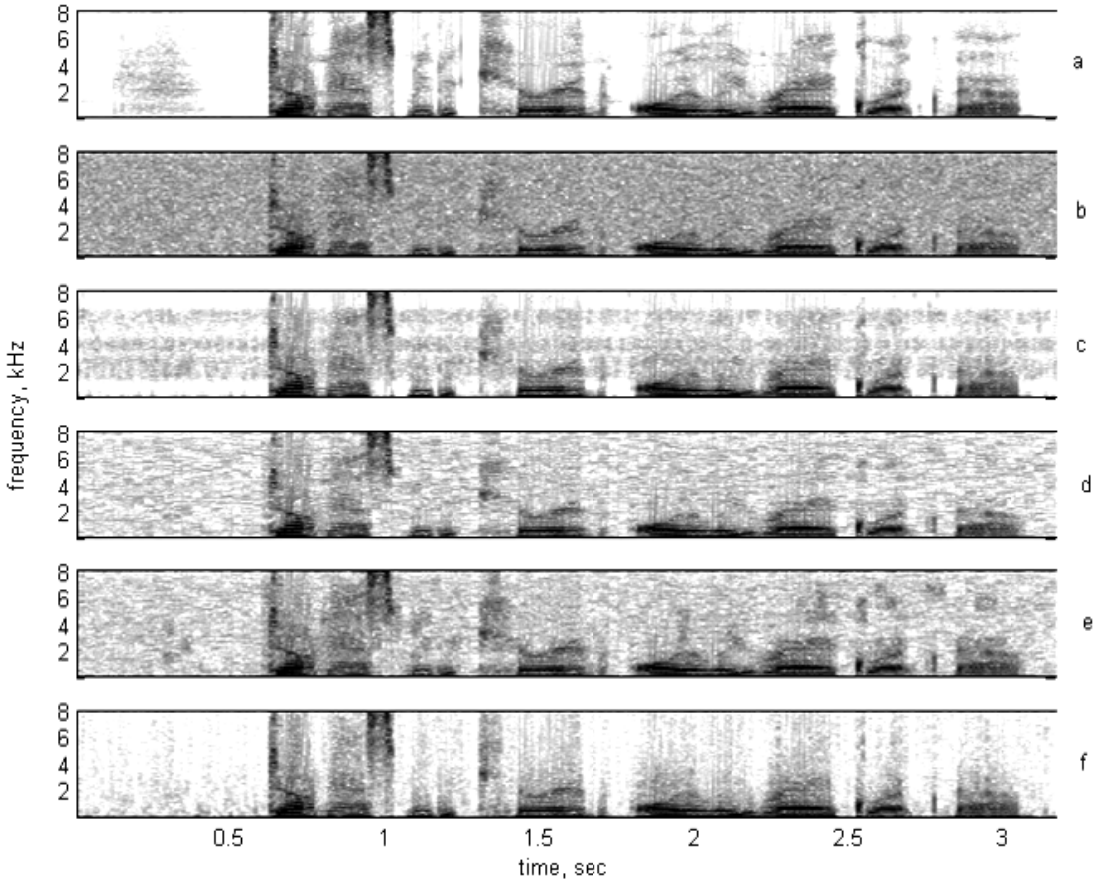


Figure 4.6: The spectrogram of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed method (λopt)

Figure 4.6 illustrates the spectrograms of the clean, noisy and recovered signals for the female speech signal 'Don't ask me to carry an oily rag like that.' from TIMIT database. It can be observed that the proposed algorithm performs significantly better than other reported methods. With the optimum value selection of λ, the noise signal is mostly removed with significantly low damage to the original speech. Figure 4.4(g)-(h) should be taken into account in order to better understand the effect of the optimum value of λ on the performance of the system. For λ=0.5, there is still remaining noise components as in Figure 4.4(g) and for λ=0.8, the noise signal is mostly removed however with a significant degradation to the original speech. Figure 4.6(f) shows the performance of the optimum value of λ which from equation (4.12) appears to be 0.6 for 10dB input SNR. Figure 4.7 shows the corresponding waveforms of the clean, noisy and enhanced speech signals.

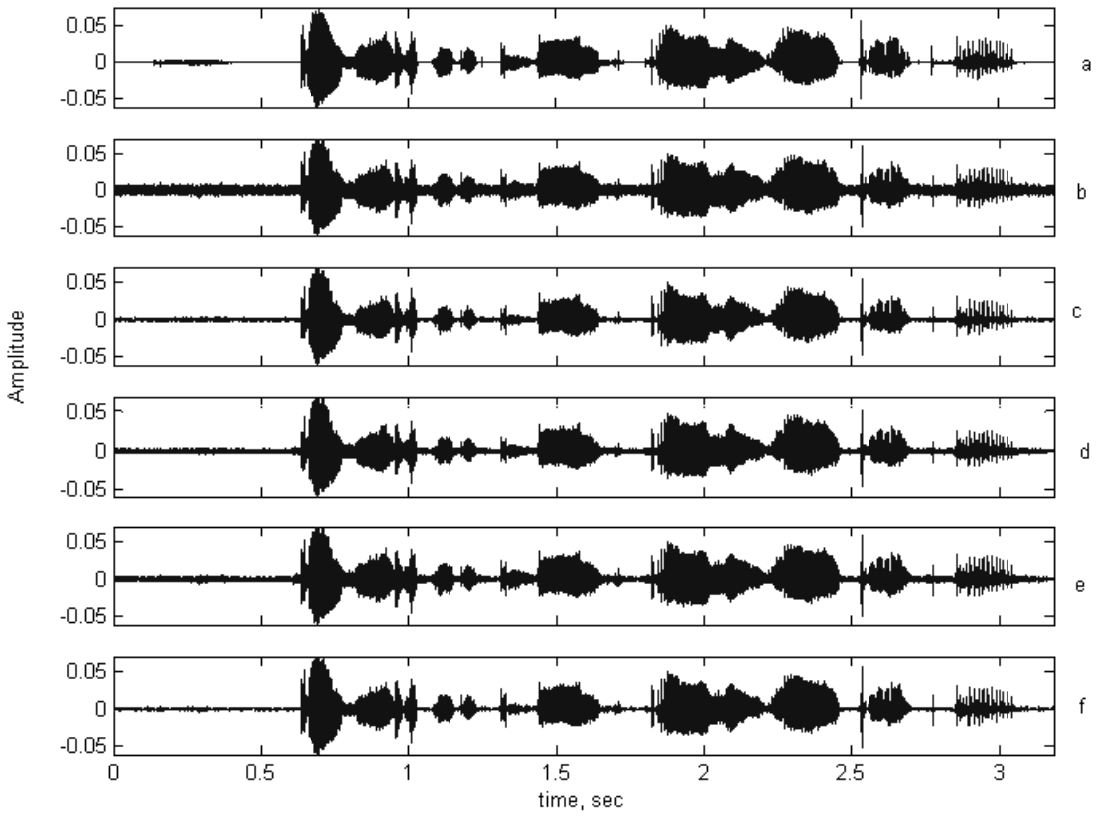


Figure 4.7: The waveform of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed method (λopt)

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 simulation for 10 male and 10 female utterances for a wide range of SNR values with a comparison of different denoising methods are listed in Table 4.2. The superiority of the proposed scheme can be well observed in this table.

Table 4.2: *Comparison of the overall SNR improvements..*

Input SNR (dB)	**Output SNR (dB)**			
	WP (20)	DCT (22)	Soft DCT (λopt) (19)	**Proposed EMD (λopt)**
0	4.86	5.69	6.55	**7.13**
5	8.86	9.76	10.15	**10.75**
10	12.36	13.74	13.93	**14.48**
15	15.45	17.65	18.03	**18.43**
25	20.82	25.53	26.43	**26.85**
30	23.16	29.52	30.68	**31.33**

It can be observed that for all SNR levels, the proposed method gives significantly better results. Although SNR improvement is a good measure for quantifying performance, it has little perceptual meaning and is therefore not a good measure for speech quality (24). Instead, the average segmental SNR (ASSNR) is relatively a better measure. The results for the ASSNR are listed in Table 4.3, which still proves the superiority of the proposed algorithm in all SNR values. It can be observed that for all SNR levels, the proposed algorithm performs better.

The main advantage of the algorithm comes with the effective decomposition introduced by EMD, which enables the higher frequency noise components to dominate the first IMFs helping the algorithm to perform better in noise detection. Moreover, the novel frame categorization limit helps the algorithm to perform significantly better than the DCT based soft-thresholding algorithm in (19). The reason is that during frame categorization in (19) where the categorization limit is set to noise variance as given in (3.2), there is a huge amount of noise components that

remain within the signal dominant frames. The novel limit given in (4.5) –which is set to twice the noise variance-, provides an efficient categorization of the frames enabling more noise components to be eliminated. Another advantage of the algorithm comes with the optimum value of λ calculation, which provides the optimum thresholding for each IMF.

Table 4.3: *Comparison of the average segmental SNR (ASSNR) improvements..*

Input ASSNR (dB)	**Output ASSNR (dB)**			
	WP (20)	DCT (22)	Soft DCT (λopt) (19)	**Proposed EMD (λopt)**
-4.111	-1.944	-0.636	-1.565	0.081
-1.341	0.797	1.747	1.039	2.512
2.079	3.623	4.687	4.254	5.529
5.758	6.427	7.852	7.712	8.767
13.837	11.723	14.799	15.307	15.877
18.002	13.780	18.382	19.222	19.542

In order to have a better idea about the perceptual quality of the enhanced speech signals, PESQ has been used. Recently regarded as the best algorithm for estimation of the results of a subjective test, PESQ returns a score between -0.5 and 4.5, with higher scores indicating better quality. The results of the PESQ simulations can be observed in Table 4-4. It can be observed that the proposed algorithm is still more effective in terms of perceptual quality than the other methods. However, it is surprising to note that the DCT algorithm with hard thresholding in (22), which thresholds all the DCT coefficients of the noisy speech, perform better in perceptual quality at high SNR values, for 25 and 30dB. The reason is that the subjective evaluation of speech signals at very high SNRs will not be as accurate as that of low SNRs. Since the signals are very close to the clean speech, the subject cannot distinguish the insignificant speech degradations. Therefore, the subject would prefer a speech signal with less noise components. Since (22) aims to remove the noise signal at all frequency levels, the resultant signal -although having more speech degradation- will have less noise components as a whole and is reasonable to be preferred. That's why for the perceptual quality, it is more reasonable to evaluate the PESQ values of the lower SNR values where the enhanced speech signals can be better evaluated and

where speech enhancement is more important to be applied. The superiority of the proposed scheme can be well observed in Table 4.4. The effectiveness of the results for speech quality in this table can be better understood when Table 4.2 is taken into consideration. For instance, the enhanced signal of 5dB signal has an average SNR of 10.75dB and an average PESQ of 2.01. This PESQ value is very close to that of 15dB input signal which has an average PESQ value of 2.06. The same discussion can be observed for other input SNRs.

Table 4.3: *Comparison of the PESQ improvements..*

Input SNR (dB)	**PESQ**				
	Input	WP (20)	DCT (22)	Soft DCT (λopt) (19)	**Proposed EMD (λopt)**
0	1.14	1.24	1.41	1.46	**1.59**
5	1.37	1.54	1.78	1.79	**2.01**
10	1.70	1.93	2.18	2.16	**2.31**
15	2.06	2.30	2.57	2.48	**2.58**
25	2.84	2.85	3.26	3.18	**3.20**
30	3.23	3.06	3.66	3.55	**3.57**

One of the major advantages of the method is that it does not include any a priori knowledge of the noise signal. However, due to the frame based thresholding which depends on the variance of the signal, the algorithm is mainly applicable for stationary noise. Moreover, since EMD decomposes the signal in terms of their frequency characteristics, the algorithm performs best for white noise case for which high frequency noise components dominate the first IMFs. Although the method performs improvement for colored noise types, the performance is the best for white noise case. Therefore, the algorithm is mainly applicable for white noise, which is the most common noise type. Another drawback of the algorithm is its time cost. Since a mathematical representation is not yet given for EMD, the process takes long time. Therefore, the algorithm is not applicable to real time speech processing. However, in order to reduce the computational cost, it is possible to divide the noisy speech signal into frames and to apply EMD in each frame instead of applying directly to the whole noisy speech.

Chapter 5

DCT-EMD Based Hybrid Soft Thresholding

5.1. Introduction

In this chapter, we introduce a novel speech enhancement method using soft thresholding with a Discrete Cosine Transform (DCT) and Empirical Mode Decomposition (EMD) based hybrid algorithm. As given in Chapter 3, soft thresholding for DCT-enhancement is a powerful method for enhancing the noisy speech signal in a wide range of SNRs. However, due to the thresholding criteria a significant amount of noise is left in the enhanced signal as can be observed in Figure 4.6(e), resulting in an irritating musical noise. EMD based soft thresholding is applied here to remove the remaining noise components. Due to the frequency characteristics of the EMD, in case of white noise, the remaining noise components are mainly centered in the lower order IMFs. Therefore, it is possible to successfully identify and remove these noise components from the first few IMFs. Since IMFs are time domain signals, the hybrid method provides a successful spectral and time domain thresholding. This two stage thresholdinag efficiently removes the noise components and significantly suppresses the musical noise problem. The degradation in the speech signal is also highly reduced.

5.2 DCT- EMD based Hybrid Soft Thresholding

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. In the second stage, we apply EMD soft thresholding algorithm to the enhanced signal of the first stage. 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, our experimental analysis gave us the following conclusions:

- In case of a noisy speech signal contaminated with white noise, 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.
- 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.
- 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 threshold values should be less compared to the first two IMFs.
- Since thresholding is already 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. Thresholding will mostly degrade the speech. These IMFs should not be thresholded.

First EMD is applied to the enhanced speech and the first four IMFs are divided into 4ms frames, each having 64 samples for 16 kHz sampling frequency. Depending on the average noise power as in equation (4.5), each frame is characterized as either noise or signal dominant. Signal dominant frames are not thresholded. In case of a noise dominant frame, the absolute values of the samples are sorted in ascending order and the following thresholding strategy is followed:

$$\hat{X}_k = sign(X_k)\left[max\left\{0, \left(|X_k| - \frac{mj}{4i}\right)\right\}\right] \tag{5.1}$$

where threshold function *mj* is same as in equation (3.4) and *i* is the index of the IMF in concern. Therefore, $\frac{mj}{4i}$ is the weighted linear threshold function defined for the IMFs. A block diagram of the proposed algorithm is illustrated in Figure 5.1.

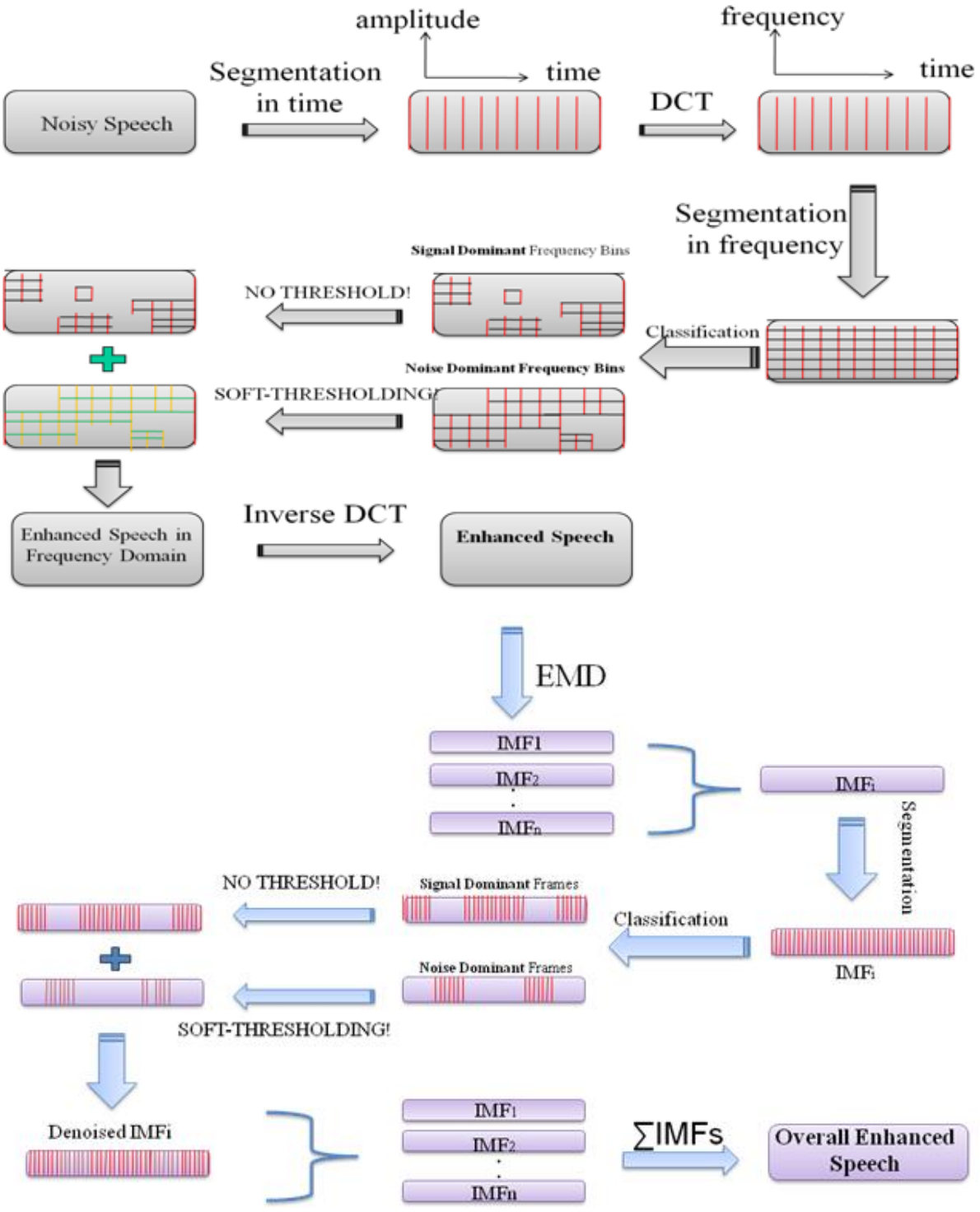


Figure 5.1: Block Diagram of the Proposed Hybrid DCT-EMD Soft Thresholding Algorithm.

5.3 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, weighted white noise samples from NOISEX database 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.

Figure 5.2 illustrates the spectrogram of the clean, noisy at 10dB SNR and enhanced speech signals for the female speech "she had your dark suit in greasy wash water all year" from TIMIT database. Figure 5.3 gives the corresponding waveforms. It can be observed that the proposed hybrid scheme is very effective in noise removal and extremely superior to other reported methods. Figure 5.4 illustrates the waveforms of the results for 0dB SNR noisy mixture.

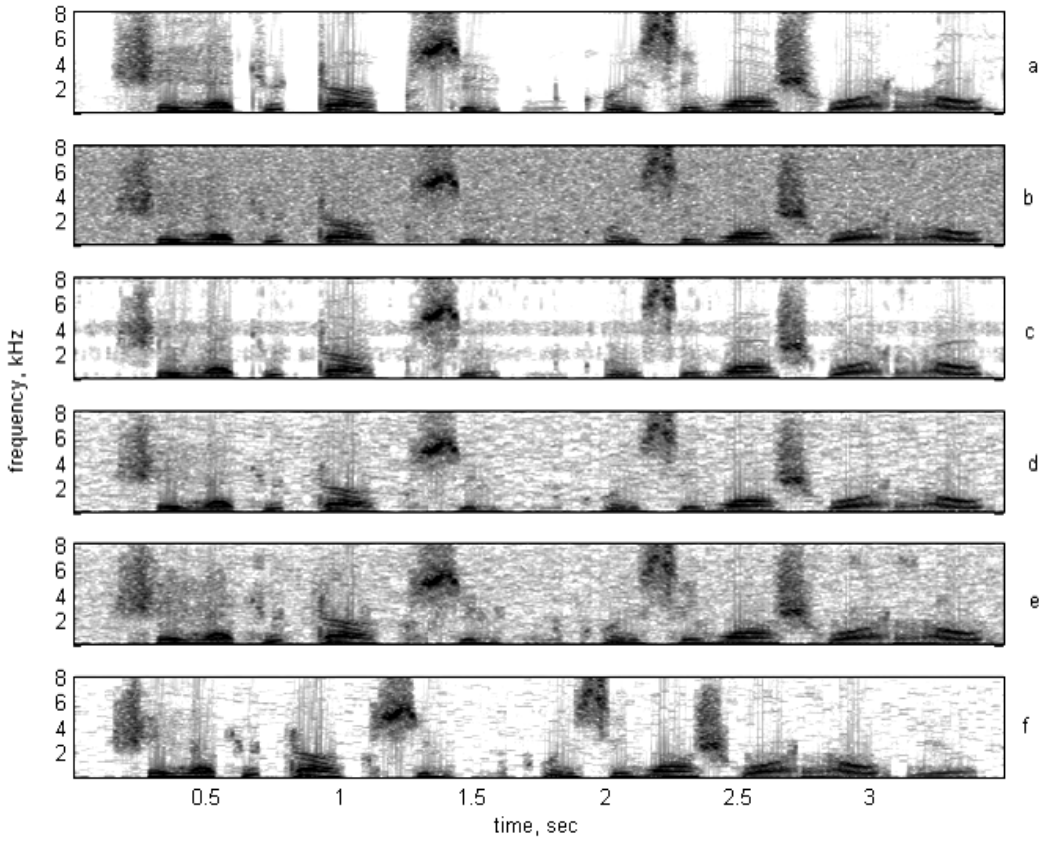


Figure 5.2: The spectrogram of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

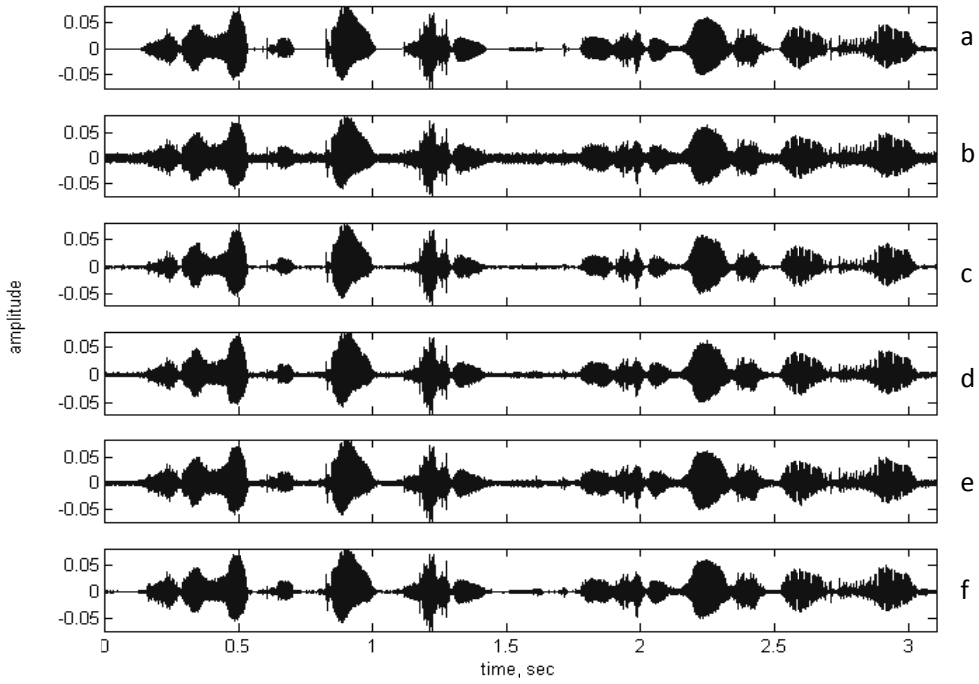


Figure 5.3: The waveform of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

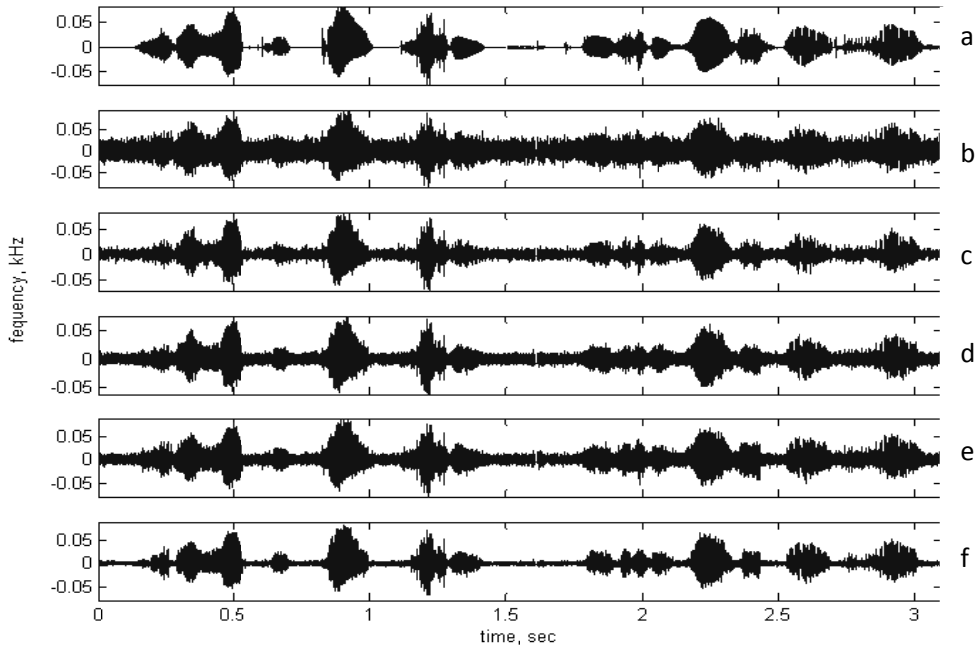


Figure 5.4: The waveform of a) the clean speech, b) noisy mixture at 0 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

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 a result of thresholding, 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 the first four IMFs. As discussed before, since the IMFs differ in terms of noise content, the linear threshold functions should also have different weights for each IMF. After extensive simulations, we have defined the threshold functions as in equation (5.1). The first four IMFs of the first stage enhanced signal 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 average results of the computer simulation for 10 male and 10 female utterances for a wide range of SNR values with a comparison of different denoising methods are listed in Table 5.1, which proves the superiority of the proposed hybrid method.

Table 5.1: *Comparison of the overall SNR improvements..*

Input SNR (dB)	**Output SNR (dB)**			
	WP (20)	DCT (22)	Soft DCT (λopt) (19)	**Proposed EMD (λopt)**
0	4.86	5.69	6.55	8.45
5	8.86	9.76	10.15	11.82
10	12.36	13.74	13.93	15.51
15	15.45	17.65	18.03	19.33
25	20.82	25.53	26.43	27.59
30	23.16	29.52	30.68	31.93

As in the EMD based soft thresholding algorithm, the main disadvantage of the proposed method is the computational cost due to the empirical calculation of the EMD. This is a major drawback of all EMD based algorithms and many researchers are working on EMD in order to derive a mathematical expression. Once a mathematical expression is given, EMD based algorithms will be applicable to real time processes.

Similar to the EMD based soft thresholding, another disadvantage in this algorithm is that it is not robust to different noise types. In the first stage, since all the frequency bins are processed with a unique noise variance estimated in time domain, the method is mainly applicable to white noise which has a flat spectrum. The method fails for other noise types that show different spectral distribution within the frequency bins. Therefore, it is important to have a sub-band approach where a specific noise variance is calculated for each frequency band. Here we propose a sub-band approach for the first stage in order to provide robustness to different noise types.

5.4 Sub-band DCT-EMD Hybrid Method

In order to make the hybrid algorithm robust to different noise types, some further modifications are described in this section. The DCT soft-thresholding, first stage of the hybrid method, is effective in removing the noise components while significantly keeping the original speech. However, since all frequency bins are processed with a unique noise variance estimated in time domain, the algorithm is mainly applicable to white noise which has a flat spectrum. The method is not robust to other noise types which show different spectral distribution within the frequency bins. For instance, in case of pink noise, the lower frequency bins will have higher noise variance. The estimated noise variance will be around the mean value of the whole spectrum; thus it will be less than the variance of the lower frequency bands. That's why, the lower DCT frequency bins will be always categorized as signal-dominant and will never be thresholded. Since most of the noise is in these bands, the algorithm will dramatically fail. Therefore, it is important to have a sub-band approach where a specific noise variance is calculated for each frequency band.

5.4.1. Sub-band Variance Approach for DCT Stage

Eight frequency sub-bands is adapted here. In order to find the noise variance of each sub-band, a frame by frame variance calculation is introduced. As discussed in Chapter 3, the 32 ms frames are divided into eight frequency bins. Therefore, each frequency bin represents a portion of a sub-band. For instance, the first frequency bin in each frame is representing the first sub-band of the whole signal which has the lowest frequency components. In order to find its noise variance, the variance of the first frequency bins of each frame are calculated and stored in an array in ascending order. Since the speechless parts will mostly have the lowest variance, its specific noise variance can be estimated from these parts of the array. The same procedure is followed for the other sub-bands. Figure 5.5 shows the sub-bands of a noisy speech signal corrupted with pink noise and Figure 5.6 shows the sorted variance array of each frequency band. The noise variance of each sub-band is estimated from their speechless parts as illustrated in Figure 5.6. Since the signal is corrupted with pink noise, it can be observed that each band has a different noise variance. The calculated noise variance of each band is used in the thresholding. With this sub-band approach, each band will have an effective bin categorization. Moreover, in order to provide better noise removal in both stages, for bin categorization, unlike the limit given in (3.2), the novel limit in (4.5) which is set to twice the noise variance is used.

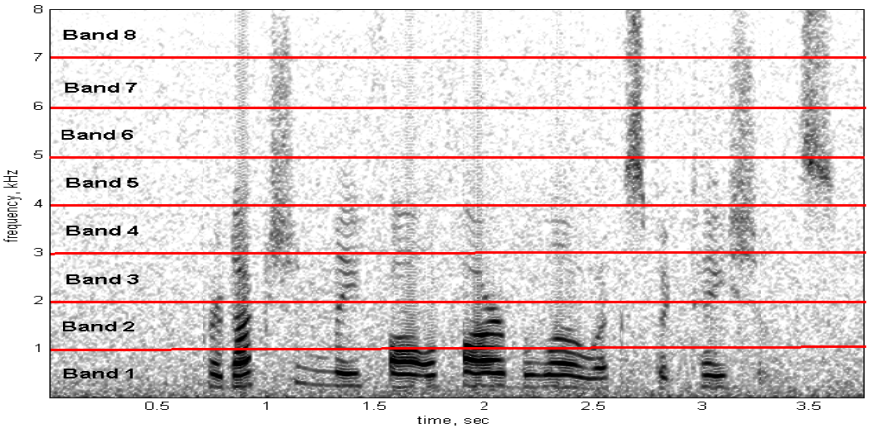


Figure 5.5: Eight frequency bands of a speech signal corrupted with pink noise (10 dB SNR).

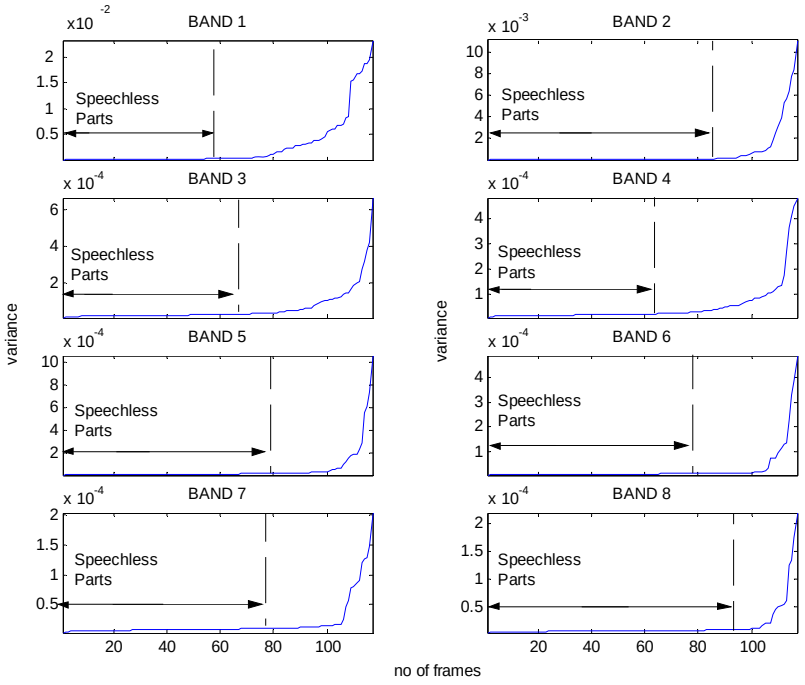


Fig. 5.6: The sorted variance of the bins of each frequency band of the noisy speech signal.

5.4.2. EMD Stage

In the second stage of the hybrid algorithm given in section 5.2, thresholding is applied only to the first four IMFs with the weighted threshold function in (5.1). This strategy is not applicable to the noise types whose energy is mainly at lower frequencies, such as pink noise. Such noise types will dominate in later IMFs. Therefore in order to provide robustness, all the IMFs are thresholded here. Instead of the weighted function, the variance of each IMF of the enhanced speech is calculated and thresholding is applied as in (4.6).

5.4.3. Experimental Results

To illustrate the effectiveness of the proposed hybrid algorithm, extensive computer simulations were conducted with 10 male and 10 female utterances sampled at 16 kHz, randomly selected from TIMIT database. The clean speech samples were corrupted with weighted noise from NOISEX database in order to obtain the noisy speech samples. To illustrate the robustness of the proposed scheme to different noise types; white, pink and high frequency (HF) radio channel noise samples have been used. For evaluating the performance of the method, overall and average segmental SNR improvements as well as objective speech quality results were used. The quality of the enhanced signals has been measured with the Perceptual Evaluation of Speech Quality (PESQ) (23).

Figure 5.7 shows the waveforms and spectrograms of the clean and noisy speech at 10dB SNR contaminated with white noise and the enhanced speech signals for the female speech “they will take a wedding trip later”. Figure 5.8 and Figure 5.9 show the spectrograms and waveforms for the pink noise and HF channel noise. The superiority and robustness of the proposed algorithm can be well observed in these figures.

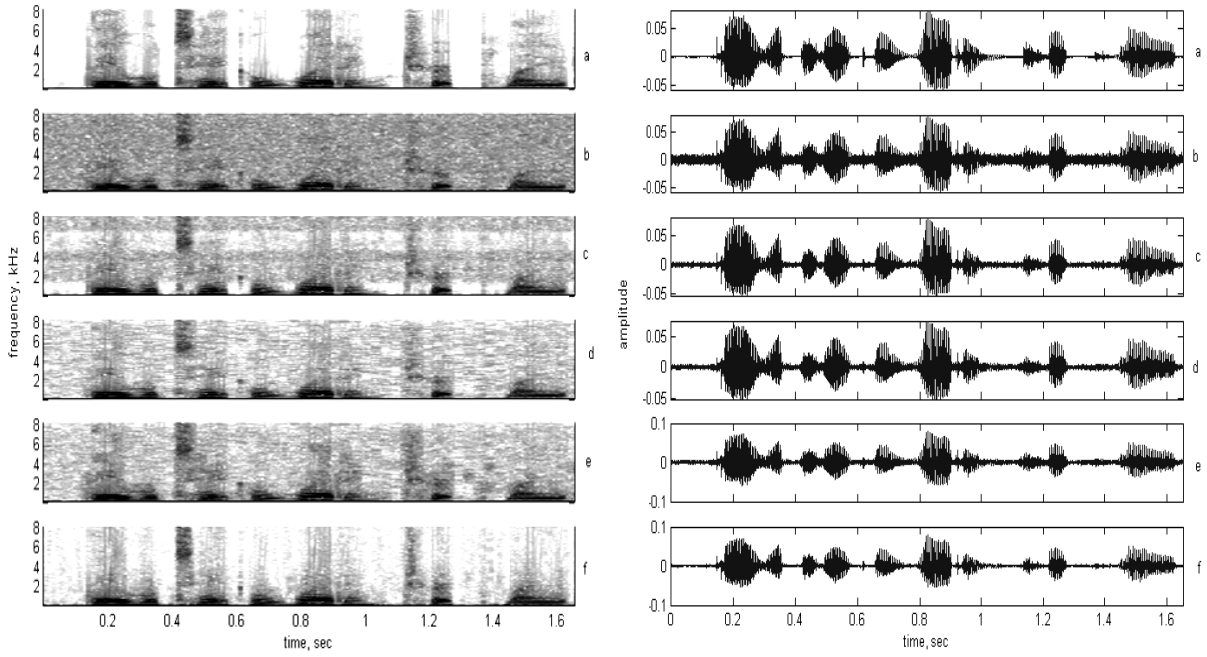


Figure 5.7: The spectrogram and waveform of a) the clean speech, b) noisy mixture at 10 dB (white noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

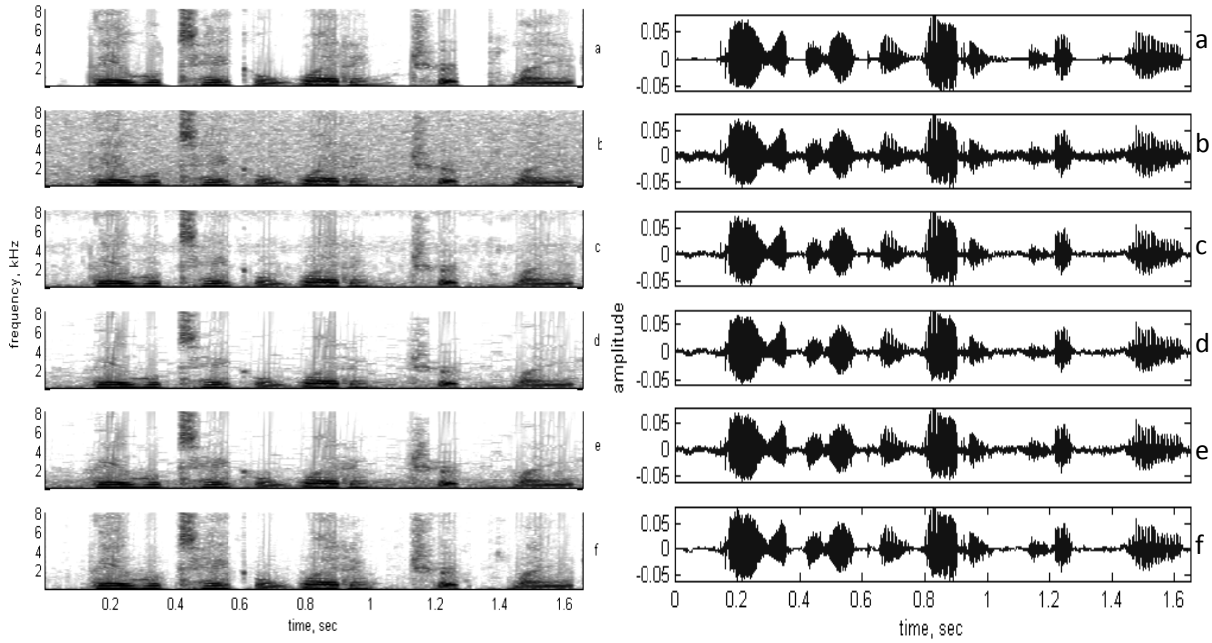


Figure 5.8: The waveform of a) the clean speech, b) noisy mixture at 10 dB (pink noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

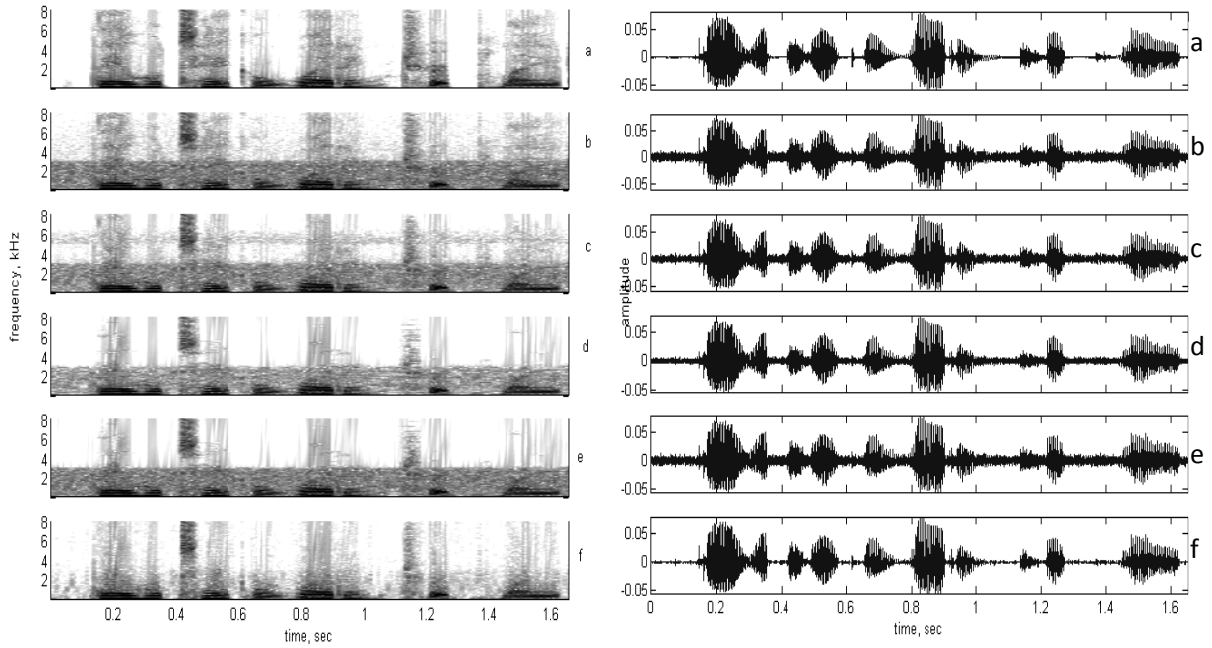


Figure 5.9: The waveform of a) the clean speech, b) noisy mixture at 10 dB (hfchannel noise), and enhanced speech with c) Wavelet packets thresholding (20), d) DCT hard thresholding (22), e) DCT soft thresholding (λopt), f) proposed hybrid method (λopt)

With the sub-band approach, the proposed algorithm is robust to different noise types and always performs significantly better than other reported techniques. The average results of the overall SNR improvement, average segmental SNR (ASSNR) and the perceptual evaluation of speech quality results (PESQ) for white noise are given in table 5.2.

Table 5.2: *Comparison of a) overall SNR, b) Average Seg. SNR (ASSNR) and c) PESQ improvements*

Input SNR (dB)	Output SNR (dB)			
	WP (20)	DCT (22)	Soft DCT (λopt)(19)	**Proposed EMD (λopt)**
0	4.86	5.69	6.55	**7.91**
5	8.86	9.76	10.15	**11.22**
10	12.36	13.74	13.93	**14.98**
15	15.45	17.65	18.03	**18.87**
25	20.82	25.53	26.43	**27.18**
30	23.16	29.52	30.68	**31.51**
Input ASSNR (dB)	**Output ASSNR (dB)**			
	WP (20)	DCT (22)	Soft DCT (λopt)(19)	**Proposed EMD (λopt)**
-4.111	-1.944	-0.636	-1.565	**0.779**
-1.341	0.797	1.747	1.039	**3.166**
2.079	3.623	4.687	4.254	**6.078**
5.758	6.427	7.852	7.712	**9.294**
13.837	11.723	14.799	15.307	**16.394**
18.002	13.780	18.382	19.222	**19.998**

Input SNR (dB)	PESQ				
	Input	WP (20)	DCT (22)	Soft DCT (λopt)(19)	**Proposed EMD (λopt)**
0	1.14	1.24	1.41	1.46	**1.74**
5	1.37	1.54	1.78	1.79	**2.07**
10	1.70	1.93	2.18	2.16	**2.39**
15	2.06	2.30	2.57	2.48	**2.71**
25	2.84	2.85	3.26	3.18	**3.32**
30	3.23	3.06	3.66	3.55	**3.66**

The same results for the pink and HF channel noise are given in Table 5.3. Both tables prove the superiority of our proposed hybrid algorithm.

Table 5.3: *Comparison of overall SNR, Average Segmental SNR (ASSNR) and PESQ improvements of different denoising methods for pink and hfchannel noise*

		Output SNR (dB)					
Input SNR (dB)		0	5	10	15	25	30
PINK	WP[3]	2.57	7.19	11.66	15.81	22.69	25.20
	DCT[8]	2.12	6.78	11.35	15.81	24.58	28.98
	S. DCT[4]	1.41	5.98	10.73	15.51	25.24	30.13
	Proposed	**4.51**	**8.27**	**12.41**	**16.81**	**26.01**	**30.44**
HF	WP[3]	1.96	6.72	11.63	16.45	24.24	26.47
	DCT[8]	3.59	7.84	11.88	15.94	24.11	28.21
	S. DCT[4]	0.94	5.38	10.08	14.92	24.70	29.61
	Proposed	**4.92**	**8.95**	**12.96**	**17.14**	**26.21**	**30.84**
		Output ASSNR (dB)					
In.ASSNR(dB)		-4.047	-1.124	2.256	5.959	14.059	18.188
PINK	WP[3]	-2.983	0.017	3.196	6.373	12.354	14.904
	DCT[8]	-3.149	-0.162	3.057	6.435	13.695	17.526
	S. DCT[4]	-3.598	-0.649	2.704	6.328	14.292	18.341
	Proposed	**-1.594**	**0.927**	**3.538**	**7.074**	**15.088**	**18.834**
In.ASSNR(dB)		-4.162	-1.287	2.079	5.781	13.906	18.049
HF	WP[3]	-3.574	-0.476	3.006	6.685	13.441	16.017
	DCT[8]	-2.683	0.218	3.219	6.411	13.319	17.007
	S. DCT[4]	-4.171	-1.349	1.948	5.599	13.603	17.725
	Proposed	**-1.234**	**1.526**	**4.416**	**7.671**	**15.342**	**19.239**
		PESQ					
Input SNR (dB)		0	5	10	15	25	30
PINK	Input	1.33	1.68	2.06	2.43	3.22	3.61
	WP[3]	1.64	2.04	2.38	2.66	3.15	3.32
	DCT[8]	1.91	2.27	2.59	2.93	3.51	3.77
	S. DCT[4]	1.85	2.17	2.51	2.84	3.50	3.79
	Proposed	**1.93**	**2.29**	**2.62**	**2.95**	**3.55**	**3.83**
HF	Input	1.58	1.84	2.14	2.44	3.15	3.49
	WP[3]	1.67	1.87	2.12	2.45	3.15	3.47
	DCT[8]	1.60	1.83	2.13	2.46	3.11	3.37
	S. DCT[4]	1.49	1.62	1.84	2.14	2.94	3.32
	Proposed	**1.61**	**1.96**	**2.32**	**2.66**	**3.34**	**3.65**

In this chapter, we presented a hybrid speech enhancement method based on DCT and EMD. In order to provide robustness to different noise types, a DCT soft-thresholding strategy with a sub-band approach is given for the first stage of the algorithm. The effectiveness of the proposed algorithm with the sub-band approach can be well observed. With this modification, the method can be applied to any stationary type of noise.

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

Chapter 6

Hard and Soft Thresholding with EMD

6.1. Introduction

Hard and soft thresholding are commonly used techniques in speech enhancement. A significant problem in such kinds of direct subtraction is the degradation of the speech signal. Therefore, it is important to have a thresholding algorithm that will minimize the speech degradation while removing a significant amount of the noise components. In the previous chapters, the given proposed algorithms include a frame based soft thresholding algorithm which does not threshold the frames that are identified as signal dominant.

In this chapter, a novel speech enhancement method based on applying a frame based joint hard and soft thresholding algorithm to the intrinsic mode functions (IMFs) of the noisy speech is given. With this strategy, hard thresholding is applied in the noise dominant frames. On the other hand, a soft thresholding method is applied to the signal dominant frames. The given strategy is effective in noise removal, however the degradation of the original speech signal is a major drawback.

6.2. Joint Hard and Soft Thresholding with EMD

The proposed method is based on applying a joint hard and soft thresholding algorithm to the IMFs of the noisy speech. First of all, EMD is applied to the noisy signal and the IMFs are obtained. Since each IMF has a different noise and speech energy distribution as can be observed, it is essential to find the specific noise variance of each IMF in order to apply an effective thresholding. This specific noise variance of each IMF is obtained from the speechless part as discussed in Chapter 4.

The IMFs are segmented into 4 ms frames, each frame having 64 coefficients for a 16 kHz sampling frequency. The frames of each IMF are categorized as either signal or noise dominant depending on its average energy content. If for the i'th frame of an IMF,

$$\frac{1}{64}\sum_{k=1}^{64}\left|X_k^{(i)}\right|^2 \geq \sigma_{n_m}^2 \tag{6.1}$$

where $\sigma_{n_m}^2$ denotes the noise variance of the *m*'th IMF, and $X_k^{(i)}$ is the *k*'th coefficient of the *i*'th frame of the *m*'th IMF, then this frame is categorized as signal dominant, otherwise as noise dominant. Hard thresholding is applied in noise dominant frames. Therefore all the coefficients in this frame are set to zero. In case of a signal dominant frame, the following soft thresholding strategy as given in (22) is applied;

$$\widehat{X_k} = \begin{cases} sign(X_k)\sqrt{\left[|X_k|^2 - \sigma_{n_m}^2\right]} & ,if\ |X_k| > \sigma_{n_m} \\ \frac{X_k\,|X_k|}{\sigma_{n_m}} & ,if\ |X_k| < \sigma_{n_m} \end{cases} \tag{6.2}$$

where $\widehat{X_k}$ denotes the thresholded coefficients. The enhanced speech is obtained by summing up the thresholded IMFs.

6.3. Experimental Results

To illustrate the effectiveness of the proposed algorithm, extensive computer simulations were conducted with 10 male and 10 female utterances sampled at 16 kHz, randomly selected from TIMIT database. The clean speech samples were corrupted with weighted white noise from NOISEX database in order to obtain the noisy speech samples in a wide range of SNR values. For evaluating the performance of the method, overall and average segmental SNR improvements were used.

Figure 6.1 illustrates the spectrograms of the clean, noisy and enhanced speech signals by different denoising algorithms for the female speech 'She had your dark suit in greasy was water all year'. Figure 6.2 shows the corresponding waveforms. It can be observed that the spectrogram of the enhanced speech signal by the proposed algorithm is very close to that of the clean speech signal and significantly better than those of the other methods. The noise components are significantly removed from the noisy speech.

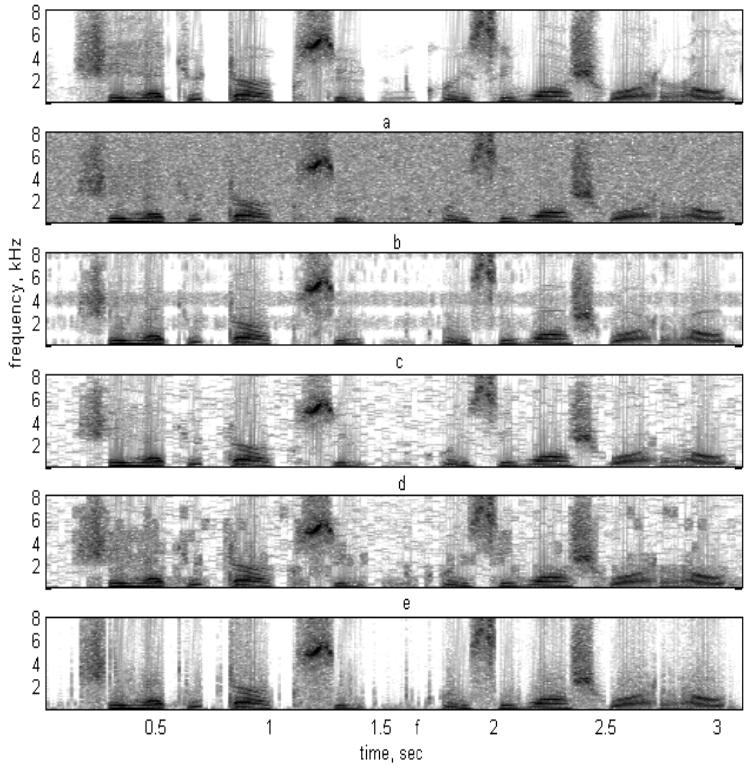


Figure 6.1: Spectrogram of a) clean speech, b) noisy speech at 10dB SNR(white noise), enhanced speech signals with c) wavelet thresholding, d)hard and soft thresholding with DCT, e) soft thresholding with DCT, and f) proposed algorithm.

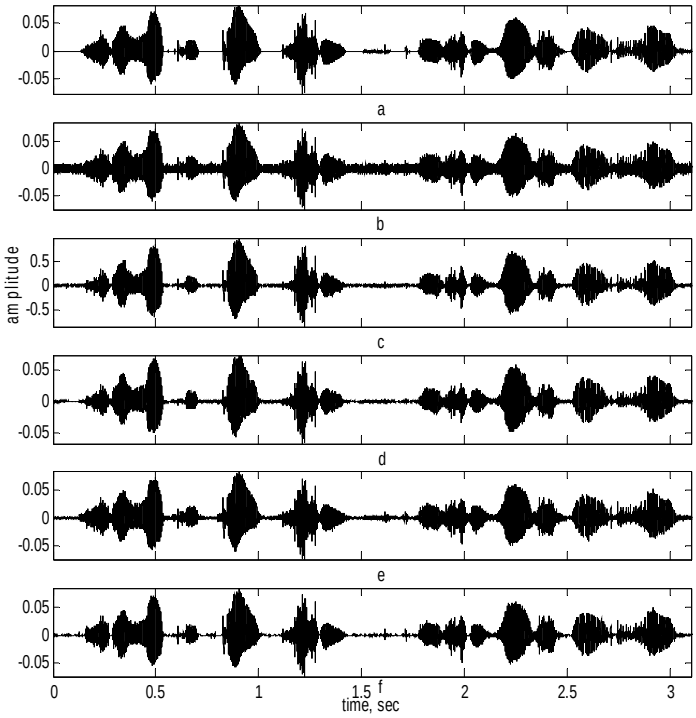


Figure 6.2: Waveform of a) clean speech, b) noisy speech at 10dB SNR(white noise), enhanced speech signals with c) wavelet thresholding, d)hard ans soft thresholding with DCT, e) soft thresholding with DCT, and f) proposed algorithm.

The average SNR improvement results of the computer simulations for 10 male and 10 female utterances are listed in Table 6.1. The superiority of the proposed scheme can be well observed in this table. It can be observed that for all SNR levels, the proposed method gives significantly better results. The results for the ASSNR are listed in Table 6.2, which still proves the superiority of the proposed algorithm in all SNRs.

Table 6.1: *Comparison of the overall SNR improvements*.

Input SNR (dB)	**Output SNR (dB)**			
	WP (20)	DCT (22)	Soft DCT (λopt) (19)	Proposed EMD
0	4.86	5.69	6.55	7.03
5	8.86	9.76	10.15	10.33
10	12.36	13.74	13.93	14.14
15	15.45	17.65	18.03	18.06
25	20.82	25.53	26.43	26.58
30	23.16	29.52	30.68	30.96

Table 6.2: *Comparison of the average segmental SNR (ASSNR) improvements*.

Input ASSNR (dB)	Output ASSNR (dB)			
	WP (20)	DCT (22)	Soft DCT (λopt) (19)	Proposed EMD
-4.111	-1.944	-0.636	-1.565	-0.001
-1.341	0.797	1.747	1.039	2.489
2.079	3.623	4.687	4.254	5.408
5.758	6.427	7.852	7.712	8.587
13.837	11.723	14.799	15.307	15.858
18.002	13.780	18.382	19.222	19.563

The proposed method is very effective in noise removal. However, since thresholding is also applied to the signal dominant frames, there is a reasonable degradation in the speech components which affects the speech quality. The soft thresholding strategy adapted to the signal dominant frames can be skipped, if better speech quality is desired.

Chapter 7

Conclusion

In this thesis, we have given EMD domain based thresholding algorithms for speech enhancement. Due to the intensity of the work given here, the conclusion will be given for each chapter;

Chapter 1: Speech enhancement aims at improving the perceptual quality and intelligibility of a noisy speech signal mainly through noise reduction. Speech enhancement may be applied to a mobile radio communication system, speech recognition system, robotics etc. Due to its importance in today's information technology, many methods have been developed for this purpose. The reported algorithms can be mainly classified as parametric and non-parametric. Parametric algorithms assume a model of the noise signal, whereas non-parametric approaches just need an estimation of the spectrum. The proposed algorithms in this thesis are non-parametric approach. Inside the parametric approaches, spectral subtraction and wavelet based thresholding have been paid great attention. Hard and soft thresholding are widely used thresholding strategies in those algorithms. The main problem in these methods is the residual noise which is generally referred as musical noise.

Chapter 2: Recently been pioneered by Huang et. al., Empirical Mode Decomposition (EMD) is a powerful data analysis method for non-linear and non-stationary signals. EMD decomposes such signals into zero mean oscillating components, referred as the intrinsic mode functions

(IMFs). IMFs give sharp and meaningful identifications of the instantaneous frequencies. This makes EMD highly efficient for non-stationary signal analysis and superior to Fourier and wavelet transforms. Soon after its introduction, EMD has been applied to a wide range of data analysis, always proving its efficiency.

Speech enhancement is one of these fields that EMD has been applied successfully. Since the extraction of IMFs relies on subtracting the highest oscillating components from the data with a step by step process, referred as the sifting process, the high frequency components dominate in the first IMFs. Therefore, the lower the index of the IMF, the higher its frequency content is. IMFs may have frequency overlaps, however at any time instant the instantaneous frequencies defined by each IMF is different, the lower order IMF having the higher instantaneous frequency. Therefore, although the IMFs are in time domain, they have spectral difference at time instances. Due to the frequency characteristics of the IMFs, the noise and speech components of a noisy speech dominate in different IMFs. A thresholding algorithm can be applied to the IMFs of the noisy speech to remove the noise components.

Chapter 3 A soft thresholding algorithm for DCT domain proposed by (19) was given here. The signal is divided into frequency bins in the spectral domain, and each bin is categorized as signal or noise dominant depending on the average noise power associated with that bin. The noise dominant bins are thresholded with a linear threshold vector instead of a constant value as in the traditional noise level subtraction rules. On the other hand, signal dominant bins are kept as they are. Since the signal dominant bins are never thresholded, the algorithm is applicable to a wide range of SNR values. The linear thresholding provides an effective noise removal, hence an effective increase in the SNR.

Chapter 4 An EMD domain soft thresholding method was proposed here. The soft thresholding strategy in Chapter 3 is adapted to the IMFs of the noisy speech with some modifications. Since the IMFs differ in terms of speech and noise energy, noise variance is calculated for each IMF. Each IMF is divided into time frames and each frame is categorized as signal or noise dominant frame similar to the DCT soft thresholding. The signal dominant bins are not thresholded, whereas the noise dominant bins are thresholded with a linear threshold function.

Due to the effective decomposition introduced by EMD and some modified criteria in the thresholding algorithm, the proposed method gives significantly better results than the DCT domain soft thresholding algorithm and other recently reported techniques. The major drawback of the algorithm is that it is mainly applicable to white noise, since white noise dominates in the first few IMFs and can be successfully suppressed with the proposed method. Another disadvantage of the algorithm comes form the computational cost of EMD. Since EMD does not have a mathematical expression, this is a major problem in all EMD based algorithms. Therefore the algorithm may not be used in real time processes. However, many researchers are working to derive a theoretical definition of EMD, which if achieved will let the EMD based algorithms be applicable to real time processes.

Chapter 5 A hybrid DCT and EMD based soft thresholding method was proposed here. The DCT soft thresholding strategy in Chapter 3 is used as a pre-process for noise reduction in spectral domain. The remaining noise components of the enhanced speech are denoised by EMD based soft thresholding in the second stage. In order to make the hybrid algorithm robust to different noise types, a sub-band variance approach is introduced for the DCT domain thresholding. With this sub-band approach, the algorithm is robust to different noise types and significantly performs better than other reported methods.

Chapter 6 A joint hard and soft thresholding criteria was adapted to the IMFs of the noisy speech signal. Similar to the EMD based thresholding, the IMFs of the noisy speech is divided into short time frames and each frame is categorized as signal or noise dominant. Hard thresholding is applied in noise dominant frames and a soft thresholding strategy is adapted in the signal dominant frames. Despite giving better results than other methods and performing significantly well in noise removal, the algorithm results in some speech degradation. The soft thresholding strategy in the signal dominant frames may be skipped if better speech quality is desired.

References

1. *Processing speech signals to attenuate interference.* **Weiss, M.R., Aschkanasy, E. and Parsons, T.** 1974. IEEE Symp Speech Recognition.

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

3. *Enhancement of speech corrupted by acoustic noise.* **Berouti, M., Schwartz, R. and Makhoul, J.** Washington : s.n., 1979. Proceedings of the IEEE ICASSP'79. pp. 208-211.

4. *Denoising by soft thresholding.* **Donoho, D. L.** 3, s.l. : IEEE Trans. on Information Theory, 1995, Vol. 41, pp. 613-627.

5. *The empirical mode decomposition and Hilbert spectrum for non-linear and non-stationary time series analysis.* **Huang, N. E., Shen, Z. and Long, S.R et al.** s.l. : Proc. Roy. Soc. London A, 1998, Vol. 454, pp. 903-995.

6. **Chan, Y. T.** *Wavelet Basics.* Boston, MA : Academic, 1995.

7. *The mechanism for frequency downshift in nonlinear wave evolution.* **Huang, N., Long, S. and Shen, Z.** s.l. : Advances in Applied Mechanics, April 1996, Vol. 32, pp. 59-117.

8. *Empirical mode decomposition as a filter bank.* **Flandrin, P., Rilling, G. and Goncalves, P.** 2, s.l. : IEEE Signal Processing Letters, 2004, Vol. 11, pp. 112-114.

9. *Empirical mode decomposition based frequency attributes.* **Ivan, M. C. and Richard, G. B.** Texas : Proceedings of the 69th SEG Meeting, 1999.

10. **Cooke, M.** *Modeling Auditory Processing and Organisation.* s.l. : Cambridge University Press, 1993.

11. *Enhanced empirical mode decomposition using a novel sifting-based interpolation points detection.* **Yannis, K. and Stephen, M.** s.l. : IEEE/SP Workshop on Statistical Signal Processing, 2007. pp. 725-729.

12. *Bivariate Empirical Mode Decomposition.* **Ricing, G., et al.** 12, s.l. : IEEE Signal Processing Letters, 2007, Vol. 14, pp. 936-939.

13. **Huang, N. E. and Nii, O. A. O.** *The Hilbert Huang Transform in Engineering.* s.l. : CRC Press, 2005. ISBN: 0-8493-3422-5.

14. *Empirical Mode Decomposition of Voiced Speech Signal.* **Bouzid, A. and Ellouze, N.** s.l. : IEEE, 2004. First Internation Symposium on Control, Communication and Signal Processing. pp. 603-606.

15. *Speech enhancement based on Hilbert-Huang transform.* **Liu, Z. F., Liao, Z. P. and Sang, E. F.** s.l. : IEEE, 2005. International Conference on Machine Learning and Cybernetics. pp. 4908-4912.

16. *Speech enhancement based on Hilbert-Huang transform theory.* **Zou, X., Li, X. and Zhang, R.** s.l. : IEEE Computer Society, 2006. First International Multi-Symposiums on Computer and Computational Sciences. pp. 208-213. ISBN: 0-7695-2581-4-01.

17. *Multi-band approach of audio source discrimination with empirical mode decomposition.* **Molla, M. K. I., Hirose, K. and Minematsu, N.** Lisbon : Proc. of EUROSPEECH, 4-8 September, 2005.

18. *Separation of mixed audio sources by decomposing Hilbert spectrum with modified EMD.* **Molla, M. K. I., Hirose, K. and Minematsu, N.** s.l. : IEICE Transaction on Fundamentals of Electronics, Communication and Computer Sciences, March, 2006.

19. *Soft thresholding for DCT speech enhancement.* **Salahuddin, S., et al.** 24, s.l. : Electronics Letters, 2002, Vol. 38.

20. *Wavelet speech enhancement based on the energy teager operator.* **Bahoura, M. and Rouat, J.** 1, s.l. : IEEE Signal Processing Letters, 2001, Vol. 8, pp. 10-12.

21. *Noisy speech enhancement using discrete cosine transform.* **Soon, I. Y., Koh, S.N. and Yeo, C. K.** s.l. : Speech Communication, 1998, Vol. 24, pp. 249-257.

22. *DCT speech enhancement with hard and soft thresholding criteria.* **Hasan, M. K., Zilany, M. S. A and Khan, M. R.** 13, s.l. : Electronics Letters, 2002, Vol. 38.

23. *Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs.* **Rix, A. W., et al.** s.l. : IEEE Proceedings, 2001, Vol. 2, pp. 749-752.

24. *Beta-Order MMSE Spectral Amplitude Estimation for Speech Enhancement.* **You, C. H., Koh, S.N. and Susanto, R.** 4, s.l. : IEEE Transactions on Speech and Audio Processing, 2005, Vol. 13.