SPEECH PROCESSING USING THE EMPIRICAL MODE DECOMPOSITION AND THE HILBERT TRANSFORM

Ke Gong

A Thesis

in

The Department

of

Computer Science

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Computer Science at
Concordia University

Montreal, Québec, Canada

September 2004

Your file Votre référence
ISBN: 0-612-94739-4
Our file Notre référence
ISBN: 0-612-94739-4

ACKNOWLEDGMENTS

I wish to express my sincere appreciation and thanks to my advisor, Dr. T.D. Bui, for his continued guidance and unwavering support during the course of this thesis. I am forever grateful to him for giving me an opportunity to finish my study as well as the freedom to explore the areas of research opened up to my curiosity.

A special thanks to my wife, Zhen Li, for her love, encouragement and support when it was most required. I am also grateful to all my other family and friends for their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

EMD: Empirical Mode Decomposition

HT: Hilbert Transform

IMF: Intrinsic Mode Function

DS: Degree of stationarity

DDS: Degree of Statistic Stationarity

IE: Instantaneous Energy density level

EGG: Electrogastrogram

MFCC: Mel Frequency Cepstral Coefficient

GMM: Gaussian Mixture Model

FFT: Fast Fourier Transform

LPC: Linear Predictive Coding

IDFT: Inverse Discrete Fourier Transform

DFT: Discrete Fourier Transform

DCT: Discrete Cosine Transform

FIR: Finite Impulse Response Filter

LMMD: Local Mean Mode Decomposition

HT: Hilbert Transform

Chapter 1

INTRODUCTION

Most of the signals in practice are time-domain signals in their raw format. When we plot the signal, the x-axis usually is time (independent variable) and the y-axis is for the amplitude (dependent variable). But, in many cases, the most important information is hidden in the frequency domain. Historically, there are methods to obtain the frequency content from the raw signal. In these methods, Fourier analysis and Wavelet analysis are the most famous ones. Very recently, Huang et al (1998) [1] introduced a new tool called Empirical Mode Decomposition (EMD) associated with the Hilbert transform to perform comprehensive analysis of nonlinear and non-stationary data.

In this thesis we discuss the EMD and the Hilbert transform and propose to use the new method to solve some problems in speech processing.

1.1. The Empirical Mode Decomposition and The Hilbert Transform

In the real world whether from physical measurements or numerical modeling, most signals are nonlinear and non-stationary. Facing such data, we have limited options to use in the analysis. Historically, Fourier spectral analysis has provided a general method for examining the global energy-frequency distributions. Although the Fourier transform is valid under extremely general conditions, there are some crucial restrictions of the Fourier spectral analysis: the system must be linear, and the data must be strictly periodic or stationary,

otherwise, the resulting spectrum will make little physical sense. In 1998 Huang et al [1] proposed a new method to analyze nonlinear and non-stationary signals in time and frequency, which only uses the signal itself and describes the frequency along the characteristic functions obtained from the signal. This method requires two steps. In the first step the signal is decomposed into a finite and often small number of "Intrinsic Mode Functions" that admit well-behaved Hilbert transforms. This decomposition method so-called Empirical Mode Decomposition is an iterative and adaptive process that uses only the signal itself. In the second step with the Hilbert transform, the "Intrinsic Mode Functions" yield instantaneous frequencies as functions of time that give sharp identifications of imbedded structures. The final presentation of the results usually is an energy-frequency-time distribution, designated as the Hilbert spectrum. In this method, the main conceptual innovations are the introduction of "Intrinsic Mode Functions" based on local properties of the signal, which makes the instantaneous frequency meaningful; and the introduction of the instantaneous frequencies for complicated data sets, which eliminate the need for spurious harmonics to represent nonlinear and non-stationary signals. The new method would be ideal for nonlinear and non-stationary data analysis.

1.2. Text-Independent Speaker Identification

Speech signal is typical nonlinear and non-stationary data. Speech processing is a diverse field with many applications. In this thesis, enlightened by the successful applications, we try to test the EMD algorithm and the Hilbert transform in some speech processing; it includes pitch detection, formant detection and text-independent speaker identification. Here we must

emphasize that in this thesis the purpose of the speaker identification system is just for testing and comparison of the new method, not for improvement of the performance of the speaker identification system itself. The reasons why we choose a speaker recognition system to test the EMD and the Hilbert transform include: firstly speaker recognition system is a complete and independent practical application. The recognition rate usually can directly indicate the performance of the new algorithm. Second, the experimental requirement of speaker recognition is not too high, especially for a text-independent speaker recognition system. In this thesis we select a subset of speech sentences from TIMIT speech database [21] and all the development work are performed in MATLAB environment.

Speaker recognition is classified into two specific types: speaker identification and speaker verification. Speaker identification is to determine which speech in a known group of speeches best matches the speaker, whereas the speaker verification is to determine if the speaker is who he or she claims to be. In speaker recognition, speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). Text-independent means that the identification procedure should work for any text in either training or testing. In this thesis, attention is focused to the text-independent speaker identification problem.

1.3. Contributions of this Thesis

The contributions of the thesis are threefold:

1) Explore the full interpretations of the EMD and the Hilbert transform for complicated data. Associated properties of the Marginal spectrum and various definitions of the

Degree of Stationarity also are explored and some practical problems are pointed out in the thesis.

2) In the scope of speech processing applications, as a new technique, the EMD method and the Hilbert transform are used in the pitch and formant analysis. Positive and negative remarks are discussed with the experiment result.

3) For extensive testing, a text-independent speaker identification system is developed to test the performance of the new methods in two aspects: one is marginal spectrum instead of Fourier spectrum being used in feature extraction and the other one is new pitch detection method by using the Degree of Stationarity.

1.4. Thesis Layout

The structure of this thesis is as follows: Chapter 2 contains a general discussion and analysis of the EMD and the Hilbert transform including implementation details. Chapter 3 introduces two successful applications of the EMD and the Hilbert transform. Chapter 4 points out the partition problem and compares the marginal spectrum with Fourier spectrum. Chapter 5 describes some of background information of speech processing and discussion the utility of the EMD and the Hilbert transform in pitch and formant analysis. Details of the comparison are presented along with the discussion of their advantages and limitations. Additionally, the details that use the EMD and the Hilbert transform in a text-independent speaker identification system are introduced in chapter 6. Finally, in chapter 7, some closing remarks about the utilities of the EMD and the Hilbert transform techniques as well the future works are proposed.

Chapter 2

THE EMPIRICAL MODE DECOMPOSITION AND THE HILBERT TRANSFORM

Before introducing the EMD and the Hilbert transform, let's briefly review the definitions of nonlinear and non-stationary here. According to the traditional definition, a time series, $x(t)$, is stationary (or periodic) in the wide sense, if, for all t,

$$E(|x(t)^2|) < \infty, E(x(t)) = m, \text{ and } C(x(t_1), x(t_2)) = C(x(t_1 + \tau), x(t_2 + \tau)) = C(t_1 - t_2),$$

in which $E()$ is the expected value defined as the ensemble average of the quantity, and $C()$ is the covariance function [1]. Therefore, a time series $c(t)$ is non-stationary if, for some m, the joint probability distribution of $c_i, c_{i+1}, \ldots, c_{i+m1}$ is dependent on the time index i [45]. Nonlinear time series is a natural extension of linear time series. A nonlinear time series is one that is not linear, and the equation is not linear if it has nonzero coefficients on the higher-order terms [46]. An ideal analysis technique for nonlinear, non-stationary signals should be local (to tackle non-stationarity) and adaptive (to tackle nonlinearity). Empirical Mode Decomposition algorithm was proposed specially for the study of nonlinear and non-stationary signals.

2.1. Instantaneous Frequency

The starting point of the Empirical Mode Decomposition is to clarify the definition of the instantaneous frequency. Instantaneous frequency is interpreted in the time-frequency literature as the average frequency at each time in the signal [2]. Common definition of instantaneous frequency $\omega(t)$ is the derivative of the phase $\theta(t)$ of the analytic signal $Z(t) = a(t)e^{i\theta(t)}$:

$$\omega(t) = d\theta(t)/dt$$

The analytic signal $Z(t)$ is computed via the Hilbert transform of the original signal, which will be explained in the section 2.3.

With such definition of instantaneous frequency, at any given instant, obviously, there is only one frequency value; therefore, instantaneous frequency is valid only for mono-component signals. Unfortunately, there is no clear definition of the "mono-component" signal to judge whether a function is or is not "mono-component". For lack of a precise definition, "narrow band" was adopted as a limitation on the data for the instantaneous frequency to make sense [30]. In the study of the probability properties of the signals and waves, the processes are assumed to be stationary and Gaussian. Then, the bandwidth can be defined in terms of spectral moments as follows. In [31][32][33][34][35], a parameter, v, was defined to offer a standard bandwidth measure:

$$v = \sqrt{\frac{m_4 m_0 - m_2^2}{m_2 m_0}} = \pi\sqrt{N_1^2 - N_0^2}$$

where $N_0 = \frac{1}{\pi}(\frac{m_2}{m_0})^{1/2}$ is the expected number of zero crossings per unit time,

$N_1 = \frac{1}{\pi}(\frac{m_4}{m_2})^{1/2}$ is the expected number of extrema per unit time and m_i is the ith moment of the spectrum.

Through the definition of v, we can easily find that for a narrow band signal $v=0$, the expected numbers of extrema N_1 and zero crossings N_0 have to be equal.

Another restrictive condition to obtain meaningful instantaneous frequency was proposed by a simple example in [1]. Let us consider a simple sine function, $x(t)=\sin(t)$. Its Hilbert transform is $-\cos(t)$. So $z(t)=\sin(t)-i\cos(t)$ is a unique analytic signal of x and can be rewritten as $z(t)=e^{i(t-\pi/2)}$. From this notation we directly see that the phase is the linear function $\theta(t)=t-\pi/2$. So the instantaneous frequency is a constant, which was also to be expected. If we move the mean off by an amount α, then, $x(t)=\alpha+\sin(t)$. Then the analytic signal of x is given by $z(t)=\alpha+\sin(t)-i\cos(t)=a(t)e^{i\theta(t)}$ with amplitude $a(t)=\sqrt{\alpha^2+2\alpha\sin(t)+1}$ and phase $\theta(t)=\arctan\frac{-\cos(t)}{\alpha+\sin(t)}$. The corresponding instantaneous frequencies $\omega(t) = d\theta(t) / dt$ are shown in figure 2.1. In the figure, we can see that the instantaneous frequencies of $0.5+\sin(t)$ is non-constant; and the instantaneous frequencies of $1.5+\sin(t)$ even is sometimes negative, which is meaningless.

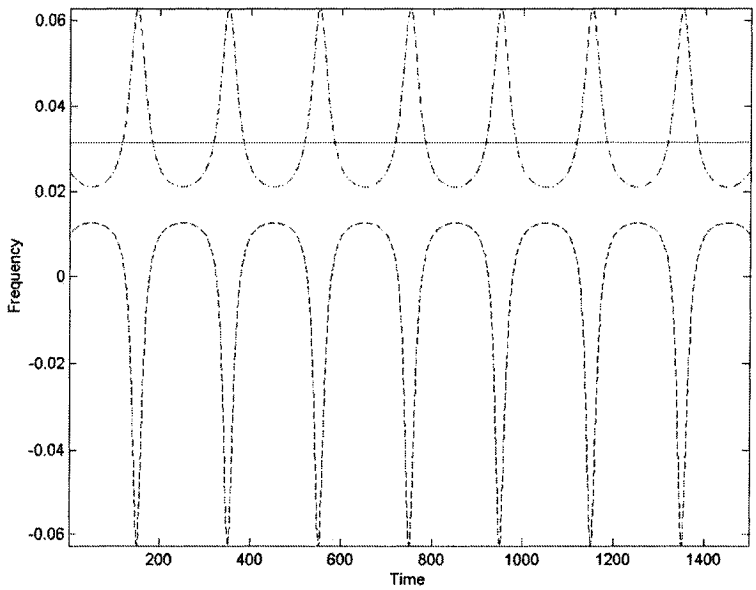


Figure 2.1: Instantaneous frequencies of $\alpha+\sin(t)-i\cos(t)$, with $\alpha=0$ (solid), $\alpha=0.5$(dash-dot), and $\alpha=1.5$ (dashed).

This example illustrates physically that, for a simple signal such as a sine function, the instantaneous frequency can be defined only if we restrict the function to be symmetric locally with respect to the zero mean level. For general data, any riding waves would be equivalent to the case of $\alpha>1$ locally; any asymmetric waveform will be equivalent to the case of $\alpha<1$, but non-zero, locally [1].

Based upon the above explanations, Huang et al [1] proposed two precise conditions of certain kind of functions, based on its local properties, to define meaningful instantaneous frequencies everywhere:

1) In the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one;

2) At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The kinds of functions that can satisfy the upper two conditions are called as Intrinsic Mode Functions (IMFs).

2.2. The Empirical Mode Decomposition Algorithm

Hilbert et al [1] introduced the Empirical Mode Decomposition (EMD) method to decompose the nonlinear non-stationary signals into Intrinsic Mode Function (IMF) components.

Given an arbitrary signal $x(t)$, the effective algorithm of EMD can be summarized as follows:

1) Initialize: $r_0(t) = x(t)$ (the residual), $i = 1$ (index number of IMF);
2) Extract the i-th IMF:
   (a) Initialize: $h_0(t) = r_{i-1}(t)$, $j = 1$ (index number of the iteration),
   (b) Extract the local extrema of $h_{j-1}(t)$,
   (c) Interpolate the local maxima and the local minima by cubic splines to form upper envelope $emin(t)$ and lower envelope $emax(t)$ of $h_{j-1}(t)$,
   (d) Calculate the mean of the upper and lower envelopes $m_{j-1}(t) = (emin(t) + emax(t))/2$,
   (e) Update $h_j(t) = h_{j-1}(t) - m_{j-1}(t)$,
   (f) If $h_j(t)$ is a IMF then set $c_i(t) = h_j(t)$ else go to (b) with $j = j + 1$,
3) Update residual $r_i(t) = r_{i-1}(t) - c_i(t)$;
4) If $r_i(t)$ still has at least two extrema then go to 2) with $i = i + 1$ else the decomposition is finished and $r_i(t)$ is the residue.

In the EMD algorithm, the method for decomposing any general signal into a set of IMFs is also called sifting. The sifting process serves two purposes: to eliminate riding waves and to make the wave profiles more symmetric. But too many sifting cycles, taking the mean and subtracting could reduce all components to a constant amplitude signal with frequency modulation only. Then the IMF components would lose all their physical significance [4]. To guarantee that the IMF components retain enough physical sense of both amplitude and frequency modulations, the number of times that the sifting process repeats has to be limited.

Therefore, Standard Deviation ($SD_j = \sum_{t=0}^{T} \frac{|h_{j-1}(t) - h_j(t)|^2}{(h_{j-1}(t))^2}$) computed from the two consecutive loops results replaces the two conditions of IMFs as stopping criterion in the step 2-(f). The flowchart of the practical EMD process is given as Figure 2.2.

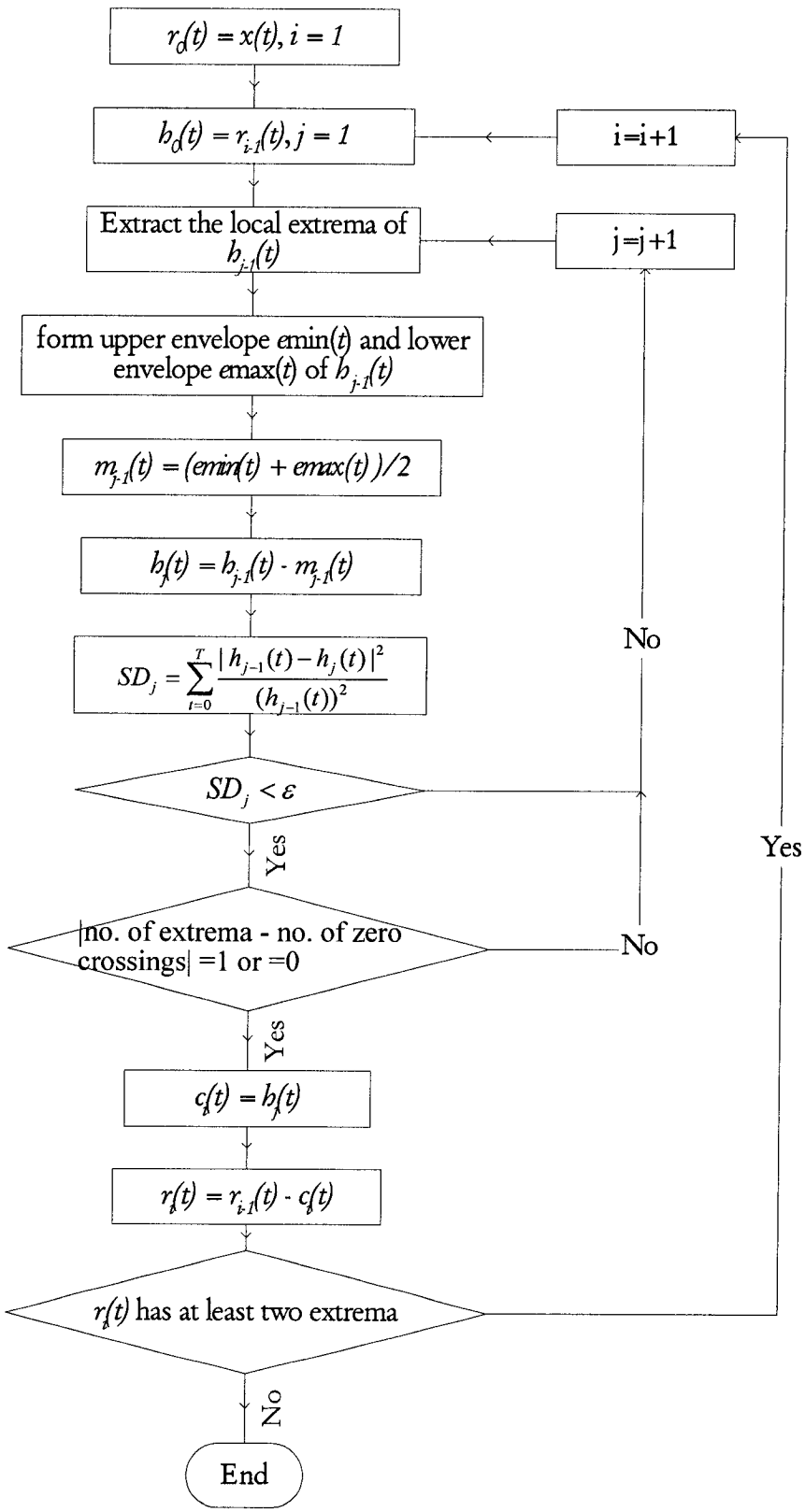


Figure 2.2: Flowchart of the EMD algorithm.

A standard deviation value of 0.2-0.3 for the sifting procedure is a very rigorous limitation for the difference between siftings. So usually the threshold value ε for SD can set between 0.2 and 0.3 [1].

Thus, the original signal *x(t)* can eventually be expressed as follows:

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

where n is the number of IMFs, $c_i(t)$ is the IMF and $r_n(t)$ is the residue.

To illustrate the whole EMD and the Hilbert transform, consider the following example. Let y_1 and y_2 be given by $y_1 = 5\sin(60\pi t)$ and $y_2 = 7\sin(40\pi t)$. Concatenate y_1 and y_2 to yield y_3 i.e. $y_3 = [y_1 \; y_2]$. Then corrupt y_3 with some zero-mean random noise, we get a nonlinear and non-stationary signal y that is noisy, narrow band oscillation around central frequencies, modulated both in amplitudes and frequencies. The signal y can be implemented in MATLAB by the following codes:

```
y1 = 5*sin(2*pi*30*t(1:halfIndex));
y2 = 7*sin(2*pi*20*t(halfIndex+1:end));
y3=[y1 y2];
randVal=2;
y = y3 + randVal*randn(size(t));
```

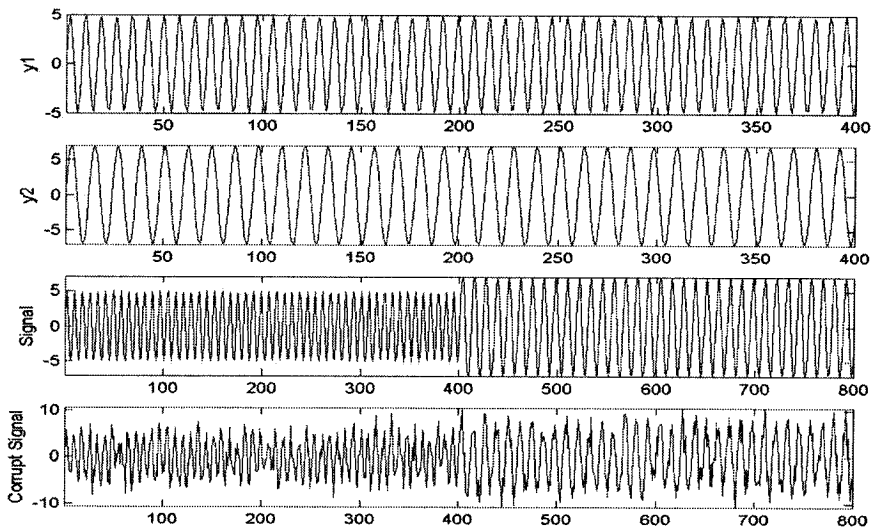


Figure 2.3: Signals y_1, y_2, y_3 and corrupt signal y.

Figure 2.3 shows the original signals y_1, y_2, y_3 and y. After performing the EMD algorithm, all the IMFs, residue and the signal y are shown in figure 2.4.

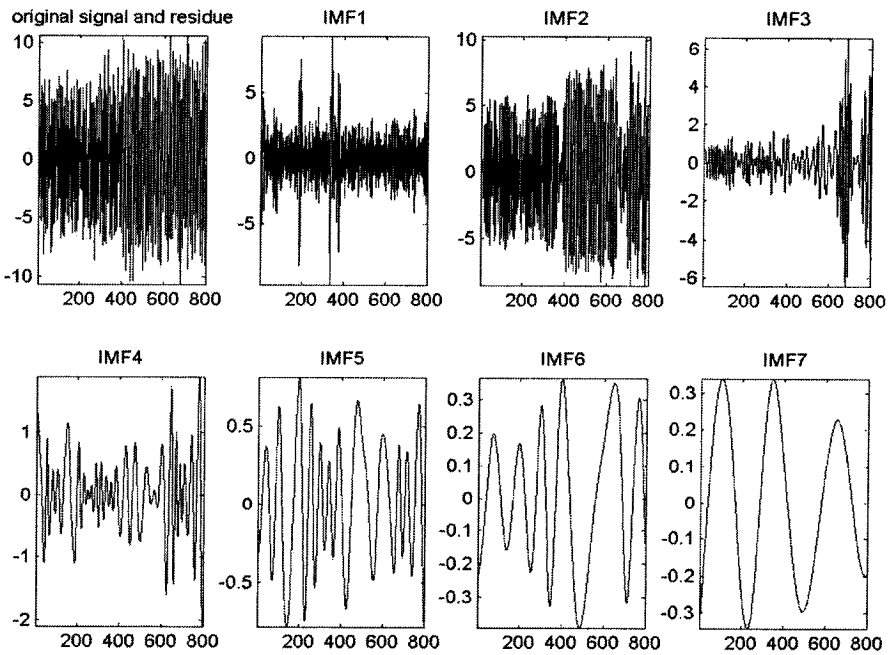


Figure 2.4: Original signal, residue and all the IMFs

2.3. The Hilbert Transforms

For a given real signal $c(t)$ we look for a decomposition into simpler signals (modes) $c(t) = \sum_{j=1}^{M} a_j(t)\cos\phi_j(t)$, where $a_j(t)$ is the amplitude and $\phi_j(t)$ is the phase of the j-th component. Each of the components has to have physical and mathematical meaning. Let $c(t)$ be "mono component" signal, i.e. we can find representation of the form $c(t)=a(t)\cos\phi(t)$ that is both physically ($\phi'(t) >= 0$) and mathematically meaningful. There are infinitely many ways to construct such representations but it is often advantageous to write the signal in complex form $Z(t)=c(t)+ic_i(t)=a(t)\,e^{i\theta(t)}$ and to take the actual signal to be the real part of the complex signal. The imaginary part $c_i(t)$ of $Z(t)$ has to be chosen to achieve a sensible physical and mathematical description . If we can fix the imaginary parts we can then unambiguously define the amplitude and the phase by $a(t) = [c^2(t) + c_i^2(t)]^{1/2}$, $\theta(t) = \arctan[\, c_i(t) \,/\, c(t)\,]$. There are many ways of defining $c_i(t)$; the Hilbert transform of $c(t)$ is one of definitions.

In the past, applications of the Hilbert transform have been limited to narrow band data; otherwise, the results are only approximately correct [4]. After the IMFs are extracted by using EMD method, because each IMF component admits well-behaved Hilbert transforms, to analyze arbitrary nonlinear and non-stationary signals in both time and frequency domain, the Hilbert transform was used as a tool to obtain the instantaneous frequency of each IMF component. This is the most direct method for determination of instantaneous frequency, and is easy to implement [29].

Given a time series data $c(t)$, the corresponding **analytic signal** that still retains the same amplitude and frequency content as the original real data is defined to be:

$$Z(t) = q(t) + i\,H(q(t)) = a(t)\,e^{i\theta(t)},$$

in which

$$H[c(t)] = \frac{1}{\pi} P\left[\int_{-\infty}^{\infty} \frac{c(u)}{t-u} du\right],$$

$$a(t) = [\,c^2(t) + H^2(q(t))\,]^{1/2},$$

$$\theta(t) = \arctan[\,H(c(t)) / c(t)\,],$$

$$\omega(t) = d\theta(t) / dt,$$

Here, the imaginary part $H[\,q(t)\,]$ in the analytic signal is the Hilbert transform of $q(t)$, the notion P indicates the Cauchy principal value of the integral. $a(t)$, $\theta(t)$ and $\omega(t)$ are the instantaneous amplitude, phase and frequency of the original data $q(t)$.

2.4. Various Definitions

After decomposing an arbitrary signal into a number of IMF components, instantaneous frequency values can be assigned to each IMF by using the Hilbert transform. For emphasis here, we must point out that for a complicated signal there is more than one instantaneous parameter at an instant.

The Hilbert Spectrum

With these instantaneous parameters defined, the frequency-time distribution of the amplitude in a three-dimensional space is designated as the Hilbert amplitude spectrum, $H(\omega(t), t)$, or simply Hilbert spectrum.

The Marginal Spectrum

By adopting the Hilbert spectrum, the marginal spectrum, $h(\omega)$, can be defined as:

$$h(\omega) = \int_0^T H(\omega(t), t)\,dt\,.$$

The marginal spectrum offers a measure of total amplitude (or energy) contribution from each frequency value. It represents the cumulated amplitude over the entire data span in a probabilistic sense. As pointed out by Huang et al. [1], the frequency in either $H(\omega(t), t)$ or $h(\omega)$ has a totally different meaning from the Fourier spectral analysis. Moreover, it should be pointed out that the marginal spectrum should not be used for any non-stationary data, for the marginal spectra are the projections rather than the substance of the real frequency-energy-time distribution [4].

After performing the Hilbert transform on the IMFs of the signal y, figure 2.5 shows the first five IMFs and related marginal spectrum of each IMF. From accumulating all the marginal spectrum of each IMF, the marginal spectrum of the original signal y is shown in the figure 2.6, comparing with the Fourier spectrum of the same signal.

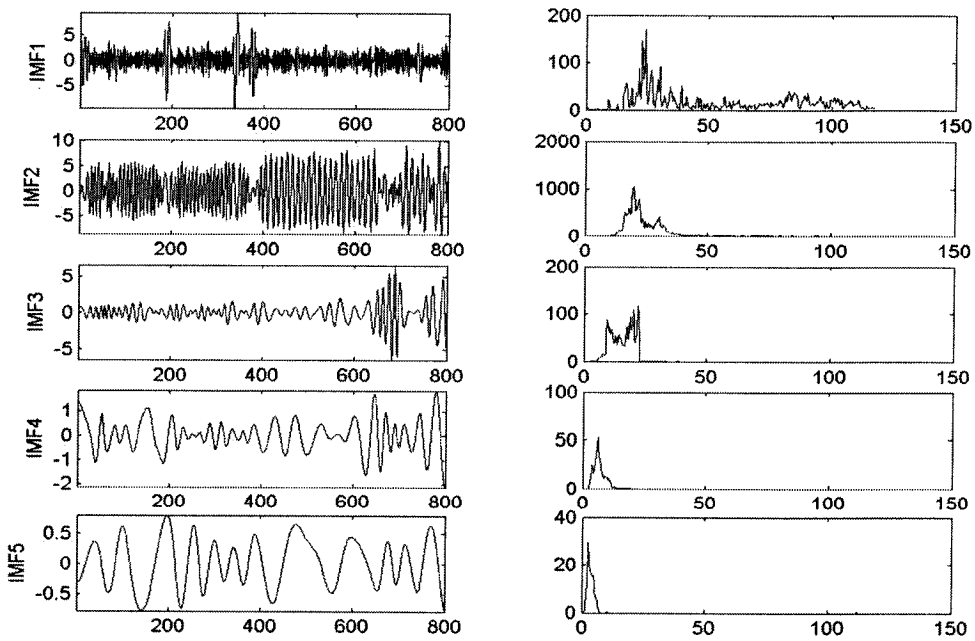


Figure 2.5: The first five IMFs (left column) and related Marginal Spectrum (right column).

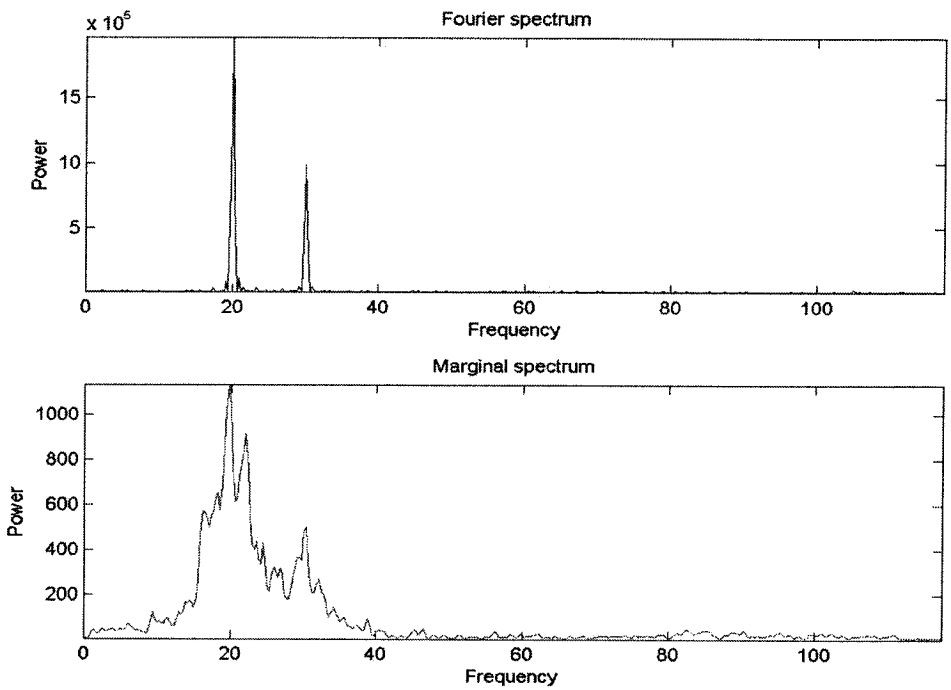


Figure 2.6: The Marginal Spectrum vs. Fourier Spectrum.

The Instantaneous Energy Density Level

Similar to the marginal spectrum, we can also define the Instantaneous Energy density level (IE) as $IE(t) = \int_{\omega} H^2(\omega,t)d\omega$.

Obviously, this IE also depends on time; it can be used to check the energy fluctuation.

The Degree of Stationarity

For arbitrary non-stationary signal, it is possible that certain frequency components can be non-stationary while other components remain stationary. Having established the Hilbert spectrum, Huang et al [1] introduced a new definition, degree of stationarity, to quantify the stationarity of nonlinear non-stationary signal.

Degree of stationarity, $DS(\omega)$, is defined as

$$DS(\omega) = \frac{1}{T}\int_0^T (1 - \frac{H(\omega(t),t)}{h(\omega)/T})^2 dt = \frac{T\int_0^T H^2(\omega,t)dt}{h^2(\omega)} - 1$$

where $H(\omega(t), t)$ is Hilbert amplitude spectrum and $h(\omega)$ is marginal spectrum.

Huang et al [1] referred to the intermittency, used in the turbulence analysis [36] to define the degree of stationary. Therefore the definition of degree of stationary is very similar to the intermittency. As a function of frequency, degree of stationarity is an index in frequency domain that gives a quantitative measure of how far the process deviates from stationarity. The closer to zero the $DS(\omega)$ value, the more stationary is the process. Because of this character of Degree of Stationarity, We proposed to use this new definition in pitch detection.

The Degree of Statistic Stationarity

Sometimes signal can be piecewise stationary. For example, date can be locally stationary while in a long time sense non-stationary; likewise, for a singular outburst in an otherwise stationary signal, the process can be regarded as almost stationary in a long time sense, but locally non-stationary near the outburst. To reflect the fact that signal can be piecewise stationary, Huang et al [1] also introduced another definition to quantify the stationarity of signal on a certain time scale: Degree of Statistic Stationarity, $DSS(\omega, \Delta T)$, which is defined as

$$DSS(\omega, \Delta T) = \frac{1}{T}\int_0^T (1 - \frac{\overline{H(\omega(t),t)}}{h(\omega)/T})^2 dt$$

where the over line indicates averaging over a definite but shorter time span, ΔT, than the overall time duration of the data, T. Even with the difficulty to choose the time scale ΔT, the definition for $DSS(\omega, \Delta T)$ could be more useful in characterizing random variables from natural phenomena.

2.5. Summary

As a summary here, now we can use the following flowchart to illustrate the whole EMD method and the Hilbert transform:

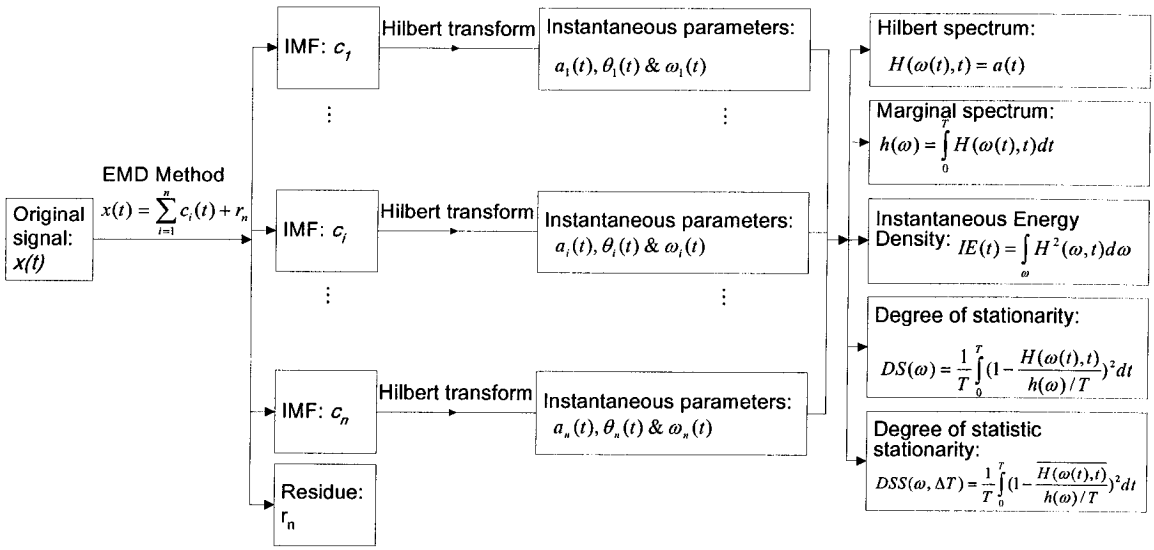


Figure 2.7: Flowchart of the EMD and the Hilbert transform.

As showed in the figure 2.7, after performing the EMD and the Hilbert transform, the input arbitrary signal $x(t)$ can be represented as a three-dimensional distribution plot: Hilbert Spectrum. With the Hilbert Spectrum defined, Marginal Spectrum, Instantaneous Energy Density, Degree of Stationarity and Degree of Statistic Stationarity are defined for descriptions of different natures of the original signal. We expect that these meaningful descriptions can be used in various signal-processing applications.

Chapter 3

APPLICATIONS OF THE EMPIRICAL MODE DECOMPOSITION AND THE HILBERT TRANSFORM

As a pretty new technique specifically designed for analysis of nonlinear non-stationary data, the EMD algorithm and the Hilbert transform can be used in plenty of areas that include earthquake engineering [6] [7], damage detection in structures, fluid dynamics [4] [8], economic data analysis and biomedical engineering [5] etc. In this chapter, two successful applications are introduced: one is Freak Wave Analysis by using the Hilbert spectrum [8]; the other one is Artifact Reduction in Electrogastrogram by using the marginal spectra [5].

3.1. Freak Wave Analysis

Water wave are a non-stationary and non-linear physical phenomenon. Freak waves, as a special water wave, are defined as transient waves existing in one particular location in one certain instant in time [8]. Due to the superposition of a finite number of dispersive wave components, a freak wave occurs and is characterized by an enormous wave height and high velocities underneath the crest of the wave.

In Schkurmann et al's paper [8], freak waves were generated in laboratory. A narrow banded freak wave with its corresponding Wavelet Morlet and Hilbert spectra from [8] are shown in figure 3.1.

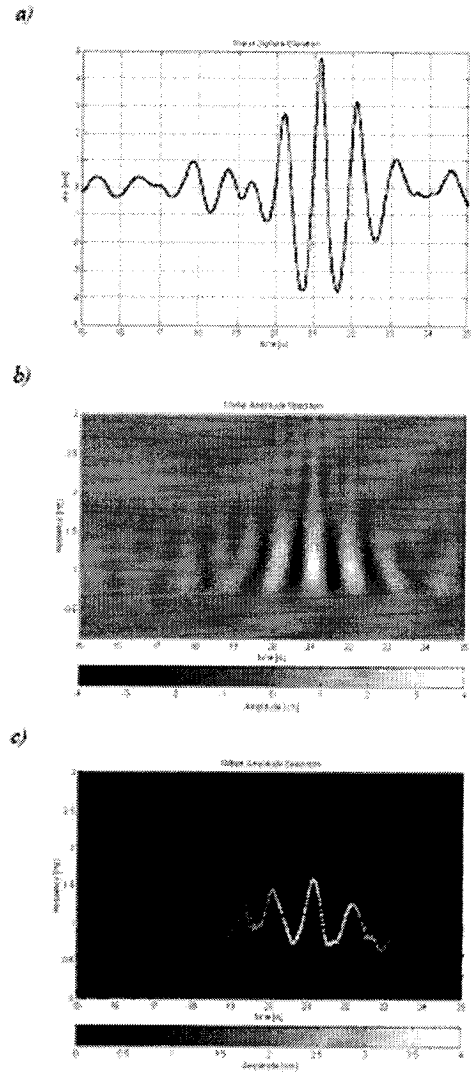


Figure 3.1: Narrow banded freak wave. a) Water surface elevation, b) Morlet amplitude spectrum, c) Hilbert amplitude spectrum. (Reprinted from Schlurmann et al. [8])

In figure 3.1 c, comparing the waves before and after the freak wave, there is no difference concerning the amplitude but a significant difference in the frequency domain. The frequencies of the wave following the freak wave are lower than those of the wave before. This effect has never been mentioned before, as the classical Fourier analysis is not able to resolve such non-stationary phenomena. The Morlet spectrum (in figure 3.1 b) also foreshadows this effect, but due to the difficulty to localize the energy to one frequency, it is not so obvious like in the

Hilbert spectrum. Hilbert spectrum gives a much sharper resolution in frequency and a more precise location in time. As a summary, Schkurmann et al [8] concluded that comparing with Fourier analysis and Wavelet analysis, the Hilbert spectrum based on the EMD algorithm is the only way to focus energy sufficiently in the frequency domain and interpret freak wave effects physically correct.

3.2. Artifact Reduction in Electrogastrogram

Electrogastrogram (EGG) is a cutaneous measurement of electrical activity of the stomach. Severe contamination of gastric signal in the EGG by respiratory, motion, cardiac signals and possible myoelectrical activity from other organs remains a serious problem for EGG interpretation and analysis. So without appropriate artifact/noise reduction it is almost impossible to extract a clean gastric signal from EGG [5].

Using conventional frequency filtering based on Fourier analysis cannot eliminate these artifact contaminations without affecting the gastric signal. Because at first, the conventional filtering is hardly capable to separate signals from broadband signals, e.g. motion artifact; secondly, the conventional filtering may also distort waveforms of the gastric signal by filtering out harmonics of the fundamental frequency of the gastric signal.

Liang et al [5] presented a successful application in artifact reduction in cutaneous EGG by using the EMD algorithm and the Hilbert transform. At first, the given EGG data is decomposed into a finite and often small numbers of IMFs. Then the next step is to perform the Hilbert transform to each of the decomposed IMF component to obtain their

instantaneous frequencies. Based on the prior knowledge about the frequency range of the gastric signal, it is easy to extract the clean gastric signal from the IMF components of the EGG data. Liang et al's experiments [5] on real EGG data showed that the EMD method does yield more efficient artifact reduction in the EGG and keep the gastric signal less affected.

Figure 3.2 [5] illustrates the schematic diagram of this application.

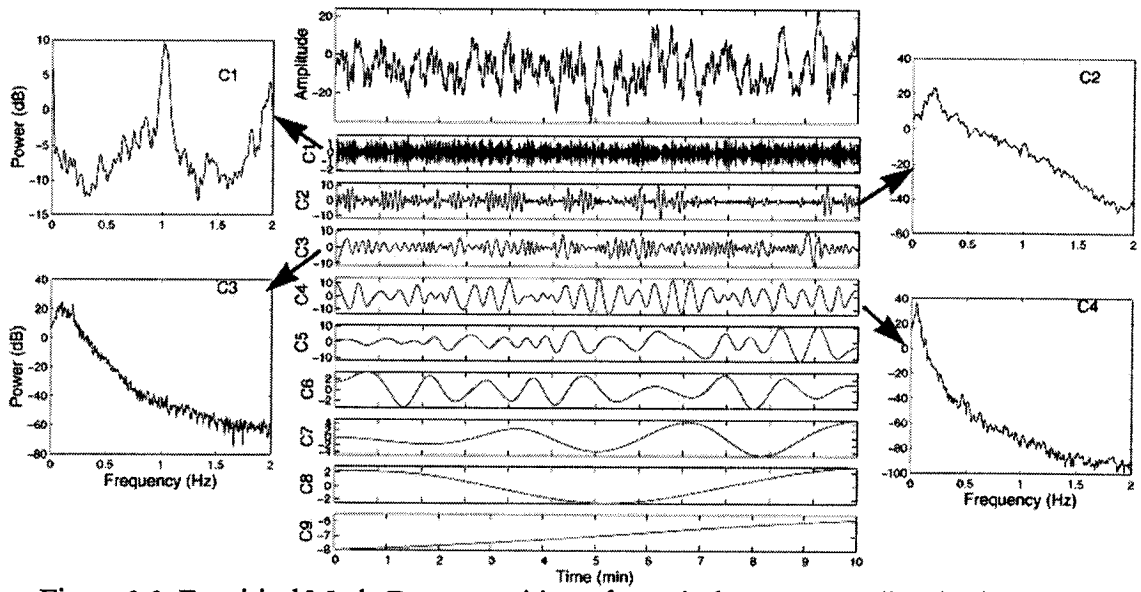


Figure 3.2: Empirical Mode Decomposition of a typical EGG recording (top) into nine components (C1 to C9), each with clear physically meaning. Left & right panels: the marginal spectra of the first four components. The peak frequencies of components C1 to C4 are successively 1.02 Hz, 0.2 Hz, 0.1 Hz and 0.05 Hz, which correspond to the heartbeat, respiratory artifact, harmonic signal and gastric slow wave, respectively. (Reprinted from Liang et al. [5])

This application shows that the IMF components carry physical significance. However, as we can find in the section 4.2, due to the IMF components overlap in frequency domain, individual IMF does not guarantee a well-defined physical meaning.

Chapter 4

PARTITION PROBLEM AND MARGINAL SPECTRUM UTILITY

After implementing the EMD algorithm and the Hilbert transform, two problems arise when we attempt to use the new techniques in speech processing. At first, if we partition the signal into smaller sections, will the sections still retain the properties of the original data? The second question is whether marginal spectrum better then Fourier spectrum in these applications? In this chapter we try to find the answers to these questions.

4.1. Partition Problem

Performing the EMD algorithm on a long serial data is extreme time-consuming. For example, it could cost 3 hours to compute a 30-second speech signal that the simple frequency is 8000 kHz in MATLAB environment. Therefore, partitioning original signal into smaller sections and performing the EMD in each section is an intuitive choice. However, after partition, for the same certain instant in both original signal and truncated signal, the number of IMFs could be different consequently. Figure 4.1 illustrates this phenomenon. In this figure, after the EMD and the Hilbert transform, for the same particular instant, there are n relevant instantaneous parameters in truncated signal, but m relevant instantaneous parameters in original signal (m could be different from n).

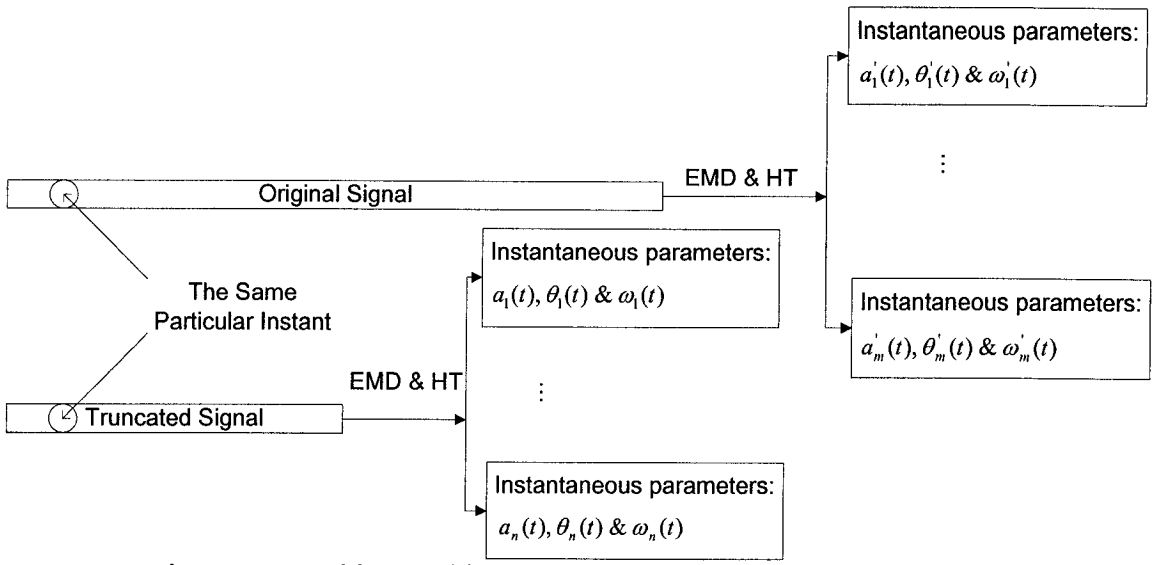


Figure 4.1: Partition Problem in the EMD and the Hilbert Transform

To examine this problem, we use the following experiment to compare the marginal spectra derived from the original signal and truncated signal. The testing signal is yielded in MATLAB defined by the function $y = 15\sin(2\pi 100t)+8\sin(2\pi 60t)+12\sin(2\pi 10t)$ and corrupted with some zero-mean random noise as follow:

```
y = 15*sin(2*pi*100*t)+8*sin(2*pi*60*t)+12*sin(2*pi*10*t);
randVal=8;
y = y + randVal*randn(size(t));
```

Partition this nonlinear quasi-periodic signal y into two same-length sections (i.e. Original signal= Section 1 + Section 2). The first section and original signal y are shown in the figure 4.2-(a). After EMD, original signal y generates eight IMFs, but the first section generates seven IMFs. Both of them are shown in the figure 4.2-(b) and figure 4.2-(c). Correspondingly, the first five IMFs and their marginal spectra also are shown in the figure 4.2-(e) and figure 4.2-(f). The marginal spectra of the two signals are compared in the figure 4.2-(d). Moreover, table 4.1 compares the computational time on the signal y and their sections. From this table, we can find the computational time really improves 33.67% after partition. Next, let us examine the

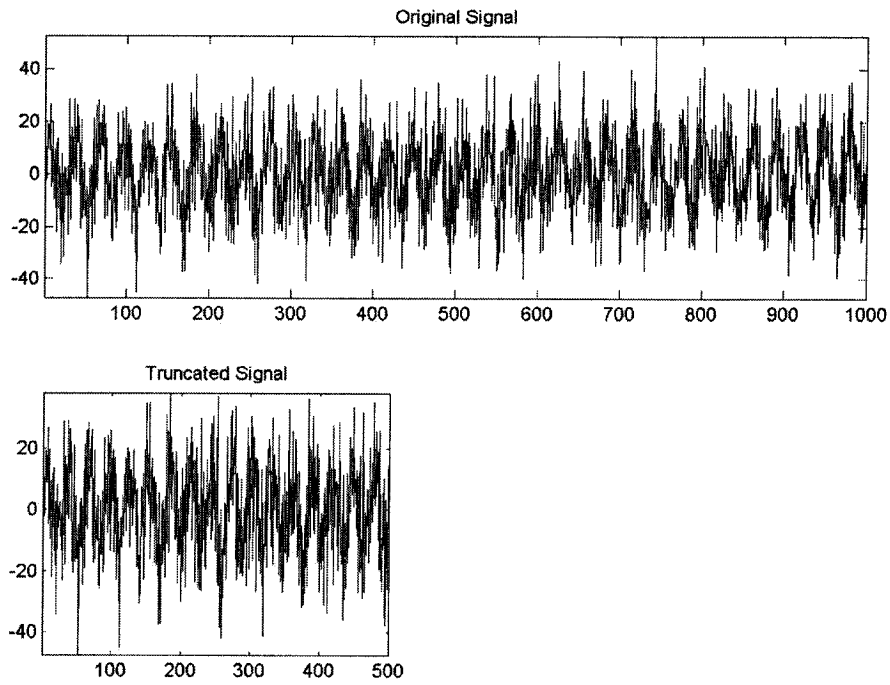
validation of the partition in the EMD and the Hilbert transform. Obviously, we find that the number of IMFs of the first section is not equal to the number of IMF of original signal. However, from the Marginal Spectrum in Figure 4.2-(d), we observe that the two spectra are similar even they are derived from different IMFs. Another observation shows that the individual marginal spectra of the first five IMFs of the two signals also have very similar peak frequencies (the frequencies that take the maximum power value). Furthermore, Hilbert spectrum was calculated for comparison. The results showed in figure 4.4 indicate that even the numbers of IMFs and related instantaneous parameters are different; however, the spectrum characters almost have no change between the truncated signal and the original signal. Therefore, we conclude that partitioning a signal into smaller sections and performing the EMD and the Hilbert transform on the each section still preserve the main characters of the original signal. At the same time the efficiency of the algorithm increases. In the later applications, we use this approach to facilitate the computation.

	Original signal	Original signal	
		Section 1	Section 2
Number of IMFs	8	7	8
Computational Time (second)	4.67	1.16	1.93
		3.09	

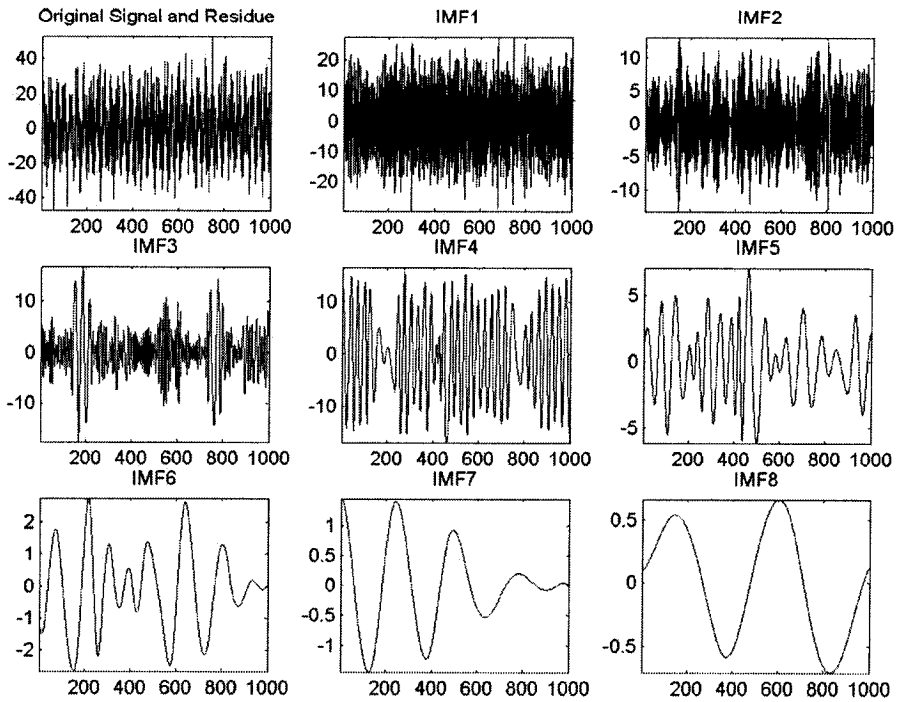
Table 4.1: Computation of two sections of the signal

Additionally, comparing the individual marginal spectra of IMFs in figure 4.2-(e) and figure 4.2-(f) with the accumulated marginal spectra in figure 4.2-(d), we can find that the individual marginal spectrum of each IMF has better frequency resolution than the accumulated marginal spectrum of the original signal. In the accumulated marginal spectrum some important frequency components are submerged by the other harmonic components. However, in the

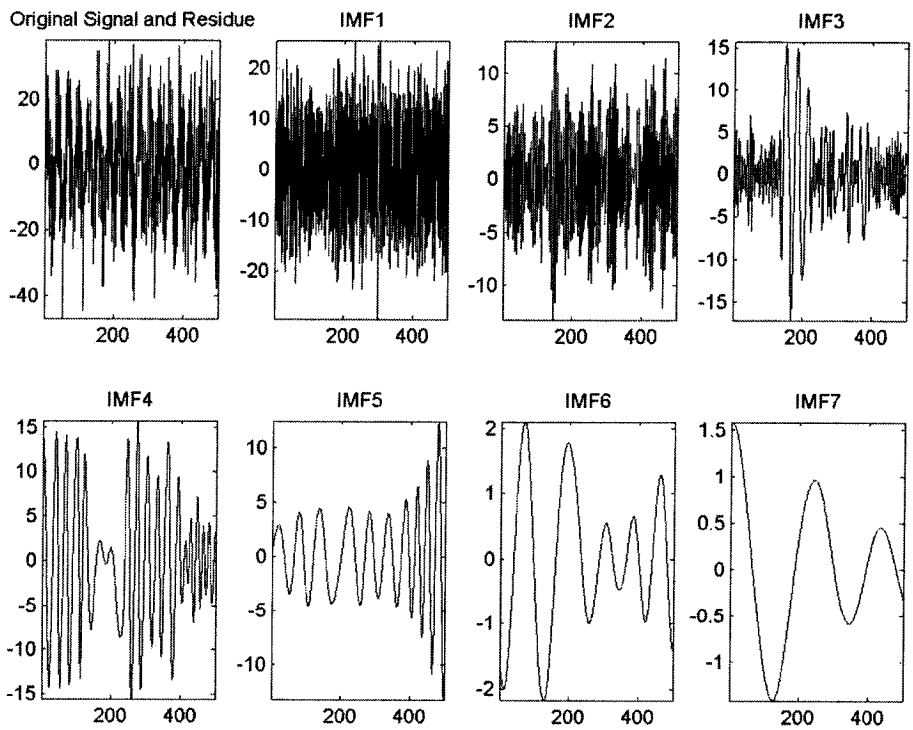
individual marginal spectrum of each IMF, because the EMD has decomposed the signal into IMF components in frequency domain, the peak frequencies in the marginal spectrum of the first several IMFs are easily located and carry well-defined physical meaning. In this example, the peak frequencies in the marginal spectrum of the first three IMFs are approximately 100, 60 and 10 kHz (in the figure 4.2-(e) and figure 4.2-(f)). Obviously, these peak frequencies correspond with the basic frequencies of the original signal that are very difficult to locate in the accumulated marginal spectrum. Therefore, roughly speaking, the first several IMFs do carry physical significance.



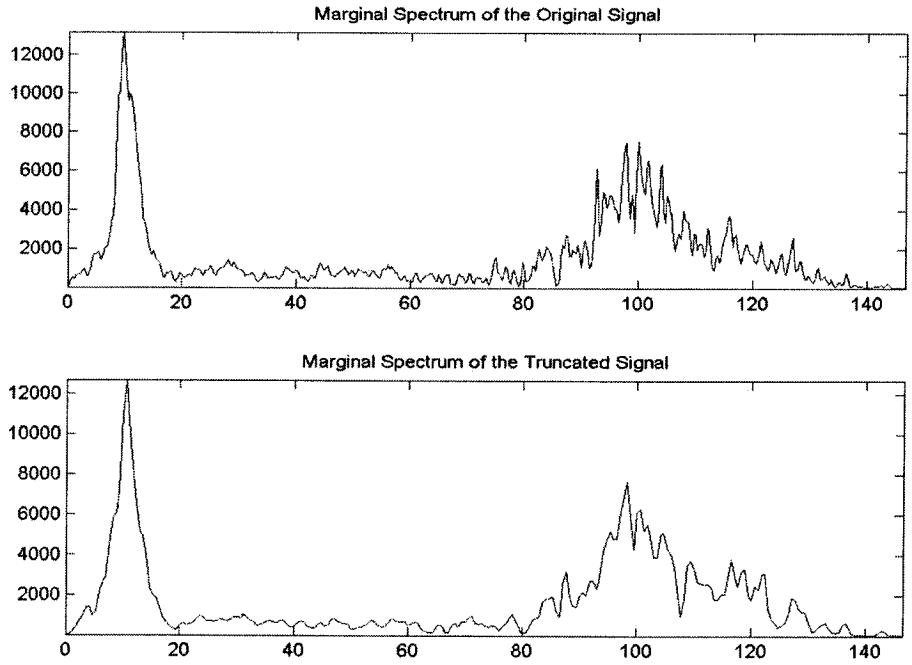
(a)


Original Signal and Residue
IMF1
IMF2
IMF3
IMF4
IMF5
IMF6
IMF7
IMF8

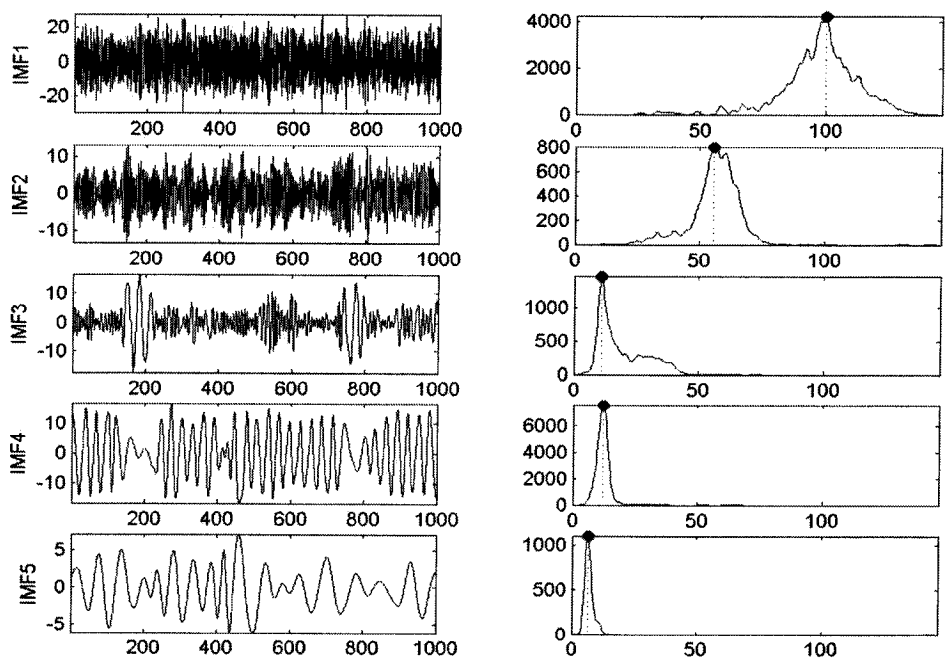
(b)


Original Signal and Residue
IMF1
IMF2
IMF3
IMF4
IMF5
IMF6
IMF7

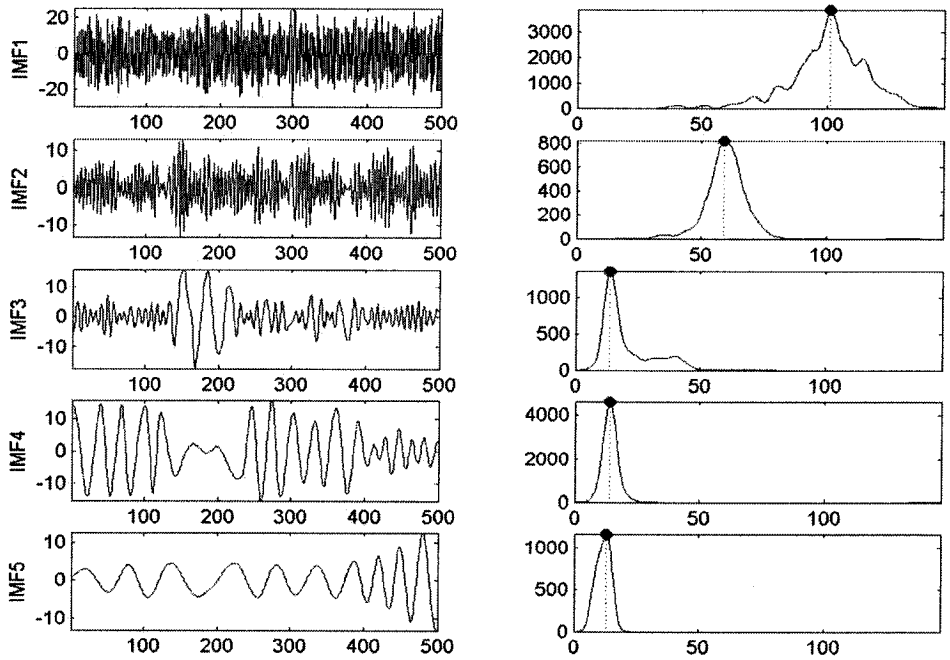
(c)


Marginal Spectrum of the Original Signal
12000
10000
8000
6000
4000
2000
0
20
40
60
80
100
120
140
Marginal Spectrum of the Truncated Signal
12000
10000
8000
6000
4000
2000
0
20
40
60
80
100
120
140

(d)


IMF1
IMF2
IMF3
IMF4
IMF5
200
400
600
800
1000
50
100

(e)



(f)

Figure 4.2: Partition EMD (for all marginal spectrum in this figure, the X-axis is frequency and Y-axis is power)

4.2. Marginal Spectrum vs. Fourier Spectrum

Some applications [8] [1] have shown that marginal spectrum has better performance then Fourier spectrum in frequency domain. However, if we try more examples to compare marginal spectrum and Fourier spectrum, we get different result. For example, using another function: $y = 3\sin(2\pi 50t) + 5\sin(2\pi 120t)$, the result is not exciting. Figure 4.3 shows the Fourier spectrum and marginal spectrum of the signal yielded by the above function.

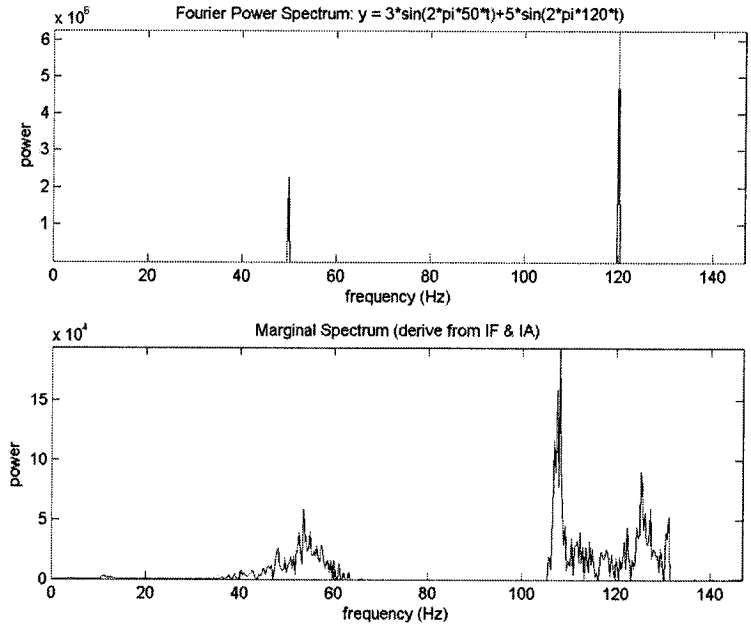


Figure 4.3: Fourier spectrum vs. marginal spectrum.

This example shows that comparing with Fourier spectrum, marginal spectrum not always has more precise presentation of the signal in frequency domain. The experiment of replacing Fourier spectrum by marginal spectrum in text-independent speaker identification also indicated that marginal spectrum has worse performance compare with Fourier spectrum. Here is the comparison table:

	Conventional approach	New approach (EMD and Hilbert Transform)
Features	MFCC	MFCC
Speaker Mode	GMM	GMM
Recognition rate	90.3% (28 speakers)	82.7% (28 speakers)

Table 4.2: Comparison of marginal spectrum and Fourier spectrum in text-independent speaker identification.

In the table 4.2: a text-independent speaker recognition system that uses Mel Frequency Cepstral Coefficients (MFCCs) as feature and Gaussian Mixture Model (GMM) as classification model is used to test the new method (more details will be given in chapter 6).

The recognition rate decreases when marginal spectrum is used in MFCCs extraction instead of the Fourier spectrum. In fact, in [4] the authors pointed out that the marginal spectrum should not be used for any non-stationary data, for the marginal spectra are the projections rather than the substance of the real frequency-energy-time distribution. In [1], another simple explanation is that marginal spectrum represents the cumulated amplitude over the entire data span in a probabilistic sense and the frequency in the marginal spectrum indicates only the likelihood that an oscillation with such a frequency exists. The exact occurrence time of that oscillation is given in the full Hilbert spectrum.

From another point of view, we propose another explanation. As discussed before, instantaneous frequency is computed by the formula:

$\omega(t) = d\theta(t)/dt$
where $\theta(t)$ is the instantaneous phase.

The reverse equation of it is: $\theta(t) = \int \omega(t)dt$. However, because negative frequency has no physical meaning, $d\theta(t)$ in the equation $\omega(t) = d\theta(t)/dt$ must add $2k\pi$ to make sure that no negative frequency exists. Hereby, the result of equation of $\theta(t) = \int \omega(t)dt$ is not the original instantaneous phase $\theta(t)$. In another word, in the real applications, the equation $\omega(t) = d\theta(t)/dt$ is not reversible. Instantaneous frequencies and instantaneous amplitudes cannot uniquely deduce the original signal without instantaneous phases. That is $x(t) \neq \sum_{j=1}^{n} a_j(t) e^{i\int \omega_j(t)dt}$. Consequently, marginal spectrum derived from instantaneous frequencies and instantaneous amplitudes cannot precisely describe the properties of the

original signal due to the uncertainty relation between the instantaneous frequencies and the instantaneous phase.

Additional explanation is about the Hilbert transform. Because the numerical method to implement the Hilbert transformation is based on the Fourier transform, the practical implementation of the Hilbert transform bears some intrinsic disadvantage of the Fourier transform, for example, the end effects and Gibbs phenomena. In MATLAB environment, to approximate the analytic signal, the Hilbert transform function calculates the Fast Fourier Transform (FFT) of the input sequence, replaces those FFT coefficients that correspond to negative frequencies with zeros, and calculates the inverse FFT of the result. In detail, the Hilbert transform function uses a four-step algorithm:

1) Calculates the FFT of the input sequence, storing the result in a vector x.
2) Creates a vector h whose elements $h(i)$ have the values:
   - 1 for $i = 1, (n/2)+1$,
   - 2 for $i = 2, 3, \ldots, (n/2)$,
   - 0 for $i = (n/2)+2, \ldots, n$,
3) Calculates the element-wise product of x and h.
4) Calculates the inverse FFT of the sequence obtained in step 3 and returns the first n elements of the result.

Using this algorithm, after the EMD and the Hilbert transform, the marginal spectrum can only approximate the theoretical values. This is another reason that why marginal spectrum can't have stable performance in different applications.

Chapter 5

APPLICATIONS IN SPEECH PROCESSING

Examples from the numerical results of the classical nonlinear equation system and data representing natural phenomena are given to demonstrate the power of the EMD and the Hilbert transform [1]. A number of successful applications ([4][5][6][7][8]) using this new method have been published recently. Speech signal is a typical nonlinear and non-stationary signal. Intuitionally, we expect the EMD and the Hilbert transform can perform well on speech processing.

5.1. Acoustics Model of Speech Production

An understanding of speech production mechanism will help us to analyze the speech sounds. To find acoustic measurements from a speech signal, people use the acoustics model of speech production to extract and represent the desired information. The observation about how speech is produced will motivate the features commonly used in speech processing. Vocal tract is generally considered as the speech production organs that mainly consists of pharynx, nasal cavity and oral cavity, which is shown in figure 5.1. An adult male vocal tract is approximately 17 cm long. Another important acoustical organ is glottis. Glottis is the opening between the vocal cords.

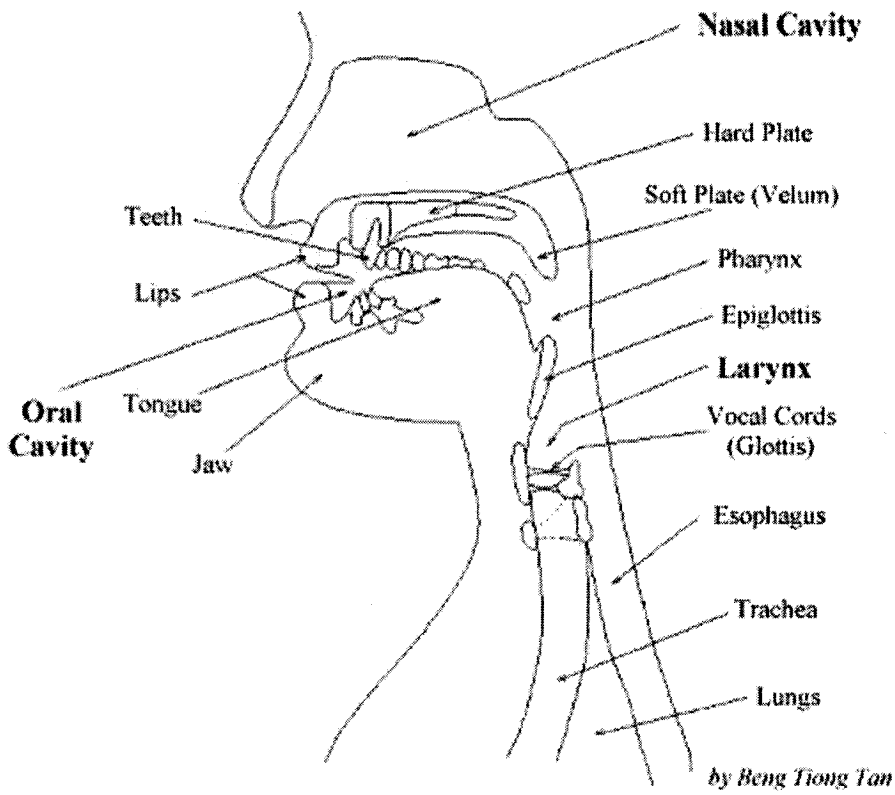


Figure 5.1: Human Vocal System (Reprinted from Flanagan [23])

Utterances are initiated by the excitation that is generated by airflow from lungs through glottis. As the airflow passes through the vocal tract, its frequency content is altered by the resonances of the vocal tract. To produce different sounds, the vocal tract moves into different configurations that change its resonance structure.

Mathematically, this model can be described as follows: $s(t) = g(t) \otimes v(t)$, where *s(t)* is the speech signal, *g(t)* denotes the excitation signal, *v(t)* denotes the vocal tract impulse response, and "⊗" denotes convolution. After taking Fourier Transform of both sides, the frequency domain representation of this process is: $S(f) = G(f) \bullet V(f)$ In this equation, multiplication replaces convolution operation. If we take the logarithm of both sides, we have:

$$Log(S(f)) = Log(G(f) \bullet V(f)) = Log(G(f)) + Log(V(f))$$

Because the excitation *g(t)* varies much more quickly than the vocal shaping v(t), in the log domain, the excitation and the vocal tract shape can be separated using conventional signal processing .

This model is sufficient for most speech processing applications, but we also need to know that this model can only model part of phonation. For example, fricatives and nasals are the exception of this model.

5.2. Pitch

Formant analysis and Pitch analysis are the basic and most important problem in speech processing. Formant analysis help to identify the word being uttered since it is heavily based on the resonances of the vocal tract and shape of the vocal tract creation. Pitch analysis makes it possible to recognize the speaker and to recognize the expressive way of speaker speaking. Firstly, we explore pitch detection and then the EMD and the Hilbert transform were applied in these two basic speech analysis problems.

The frequencies at which the vocal cords vibrate during a voiced sound are called pitch (or fundamental frequency or F_0). Pitch is the smallest unit modeled by impulse response [43]. In atonal languages like English, pitch does not carry information about phoneme identity, although it does carry prosodic information about questions or emphasis [10]. Pitch can take values from 50 to 800 Hz and fluctuates according to the stress the speaker poses to his

phrasing. Pitch is an independent parameter so that it can be used jointly with other spectral features. Three typical speakers' pitches are shown in figure 5.2.

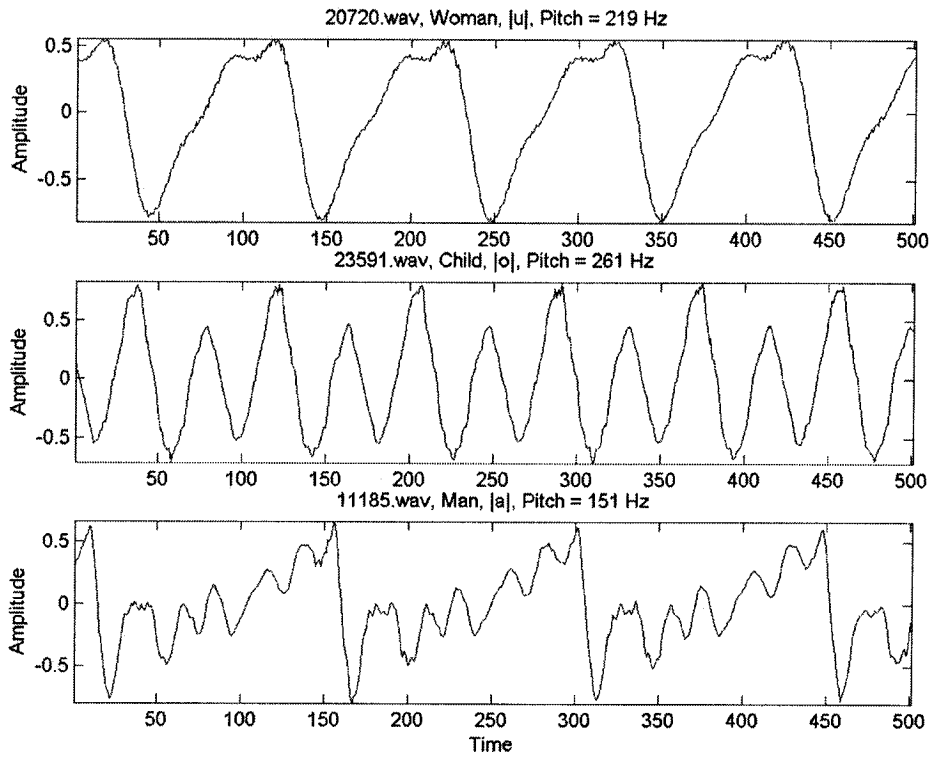


Figure 5.2: Comparison of pitches.

5.2.1. Voiced/Unvoiced Detection

Unlike speech recognition system which make use of both voiced and unvoiced portions of speech, it is sufficient for speaker recognition systems to rely on voiced portions only, since they are robust to additive noise [11].

All voiced speech originates as vibrations of the vocal cords. Its primary characteristic is its periodic nature. Vowels sounds are one example voiced speech; for example, the /aa/ sound in father or the /ow/ sound in boat. Unvoiced speech does not have the periodicity associated with voiced speech. The vocal folds are held open for these sounds. For example, /f/ in fish,

and /s/ in sound. Roughly speaking, many vowels are voiced sounds and many consonants are unvoiced sounds. The conventional voiced/unvoiced speech detection algorithms are similar to the end point detection algorithm that will be discussed in section 6.3.

Pitch is dependent on the size and tension of the speaker's vocal folds at any given instant. Pitch indicate whether the phone is voiced or unvoiced, but do not contain any other phone specific information. Everyone has a "habitual pitch level", which is a sort of "preferred" pitch that will be used naturally on the average. In the paper [24], the author indicated that the pitch varies between 90 and 175 for male sounds, 185 and 320 for female sounds and 350 to 440 for children sounds. Therefore we plan to use pitch as an additional feature and combine it with Mel Frequency Cepstral Coefficients (MFCCs) features in speaker recognition system. A Voiced/Unvoiced Detection sample is showed in the following Figure.

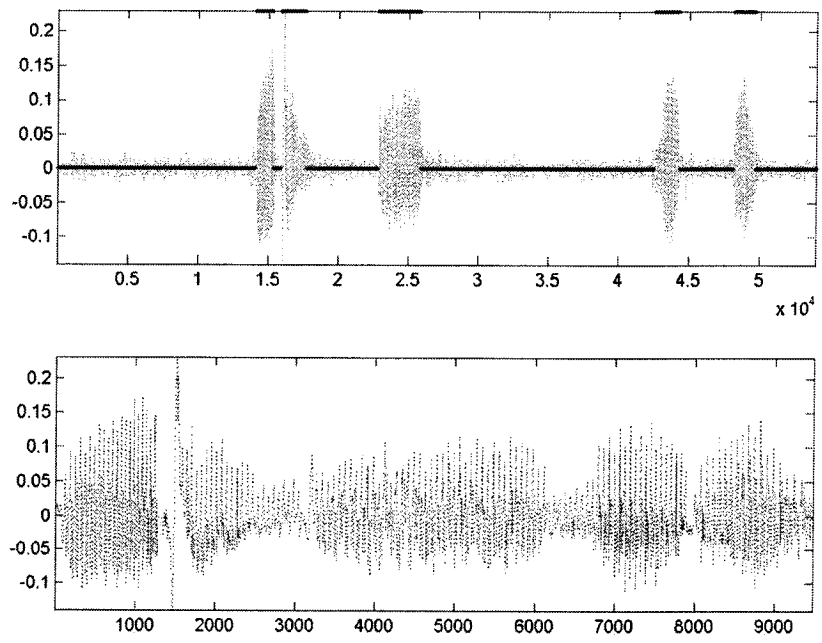


Figure 5.3: Voiced/Unvoiced Detection. Divide the speech into two classes: voiced speech and unvoiced speech. Then combine the voiced speech into a new speech signal.

5.2.2. Autocorrelation method

Autocorrelation method is a conventional approach for pitch detection. The basics idea of this method is how to reliably extract the periodicity of quasi-periodic signals.

The autocorrelation function of a deterministic sequence is given by:

$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$. If the signal is periodic with period P, then $x(m) = x(m+p)$ and so

the autocorrelation function is also periodic with period P:

$$\phi(k+P) = \sum_{m=-\infty}^{\infty} x(m)x(m+k+p) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) = \phi(k)$$

Also, $\phi(k)$ has a maximum at $k = 0$. This fact, together with the fact that the autocorrelation of a periodic signal is also periodic, suggests that the autocorrelation function has peaks at each integer multiple of the period P. Therefore, any autocorrelation-based pitch estimator simply chooses the period as the lag (over all possible pitch periods) which maximizes the autocorrelation function. When dealing with speech signals, one typically calculates a short-term autocorrelation function (i.e. on a frame-by-frame basis) according to:

$$R_n(k) = \frac{1}{N} \sum_{m=-\infty}^{\infty} s(m)\omega(n-m)s(m+k)\omega(n-m-k)$$

where $\omega(n)$ is a window of length N. For a symmetric window centered on the origin, the short-term autocorrelation can be written as:

$$R_n(k) = \frac{1}{N} \sum_{m=0}^{N-k-1} s(n+m)\omega(m)s(n+m+k)\omega(m+k)$$

When calculating the short-term autocorrelation of a speech signal, it is important to include at least two full pitch periods to allow for accurate estimation of the period. Because as the lag increases there are fewer terms involved in the summation. To remedy the problem, the unbiased short-term autocorrelation function was introduced:

$$R_n(k) = \frac{1}{N-|k|} \sum_{m=0}^{N-k-1} s(n+m)\omega(m)s(n+m+k)\omega(m+k)$$

The lag that maximizes $R_n(k)$ over all possible pitch periods is chosen as the pitch estimate for the frame centered at time n.

5.2.3. Pitch Used in Speaker Recognition

Due to physiological considerations such as the length and thickness of the vocal folds, and respiratory muscle patterns, the phonation of a particular vowel with "normal effort" may result in differing rates of vocal fold vibration (corresponding to the acoustical correlate of fundamental frequency) for different speakers. For example, a child will have a high fundamental frequency compared to an adult because of the child's smaller vocal folds.

Recent work carried out on gender identification indicates that a speaker's gender can be identified with 98% accuracy using the mean pitch parameter alone [22]. This led us to believe that useful information about a speaker's identity may be contained in the speaker's mean pitch, even the use of pitch features alone could not give enough recognition performance [44]. From another point of view, although fundamental frequency, along with intensity and duration, is a controllable attribute of stress and intonation which may vary widely, each person appears to have a mean fundamental frequency value which, if averaged

over a sufficiently long period of time, is relatively constant over a reasonable time span and is independent of linguistic content [12]. In addition, the standard deviation of the fundamental frequency over a long interval of time may carry important speaker-dependent information. For example, if the speaker were judged to be a monotone speaker, then the standard deviation would be expected to be relatively small. However, if the speaker were thought to be an "expressive" or "forceful" speaker, it would be expected to be relatively large. Reference to these proposed methods; a log pitch was appended to the conventional feature MFCC to make the feature vector in our project, which is showed in Figure 5.4.

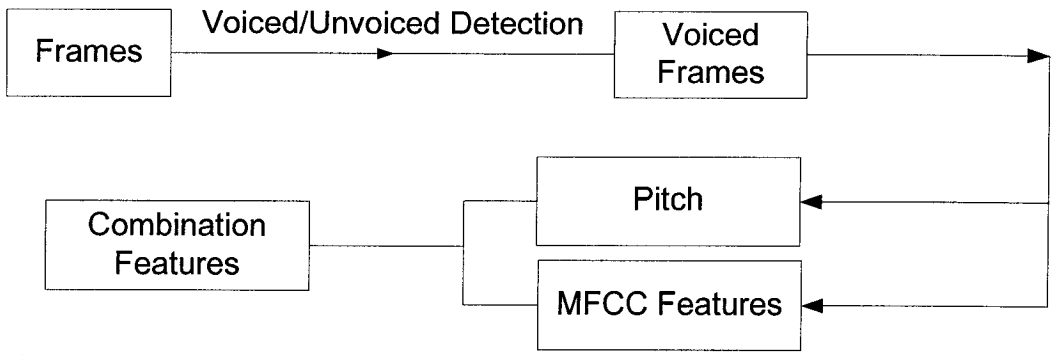


Figure 5.4 Combine pitch and MFCC as the features of speaker Identification system.

The Microsoft .wav files of vowel speeches that come from Dieter Maurer's research [18] are analyzed to test the new method that uses Degree of Stationarity. The main experimental approach of Dieter Maurer's research was to investigate the spectrum and the spectral envelope of isolated sound fragments of vowels. A large sample of the Swiss German vowels /u, o, a, ä, ö, e, ü, i/ was investigated in the research, which include approximately 18,700 recordings that were made of 35 men, 44 women, and 20 children. In Dieter Maurer's research, Fourier and Linear Predictive Coding (LPC) analysis were performed and the spectra and their envelopes were visually inspected for pitches analysis. The Internet presentation of

the research including part of original sounds can be found in the Internet [18]. The pitches in this research are considered as the baseline of comparison. Table 1 shows the comparison of results of difference methods as follow:

No	File		Speaker	Reference F_0	F_0 (Autocorrelation method)	Error rate	F_0 (DS method)	Error rate
1	10881.wav	/a/	woman	176	173.9	1.19%	185	5.11%
2	23591.wav	/o/	child	261	250	4.21%	268.6	2.91%
3	28014.wav	/u/	child	497	470.6	5.31%	496	0.20%
4	16794.wav	/o/	woman	355	173.9	51.01%	366.9	3.35%
5	17690.wav	/i/	woman	713	666.7	6.49%	683.7	4.11%
6	12387.wav	/e/	woman	207	205.1	0.92%	218.1	5.36%
7	12815.wav	/i/	woman	368	347.8	5.49%	367.6	0.11%
8	11185.wav	/a/	man	151	148.148	1.89%	159.059	5.34%
9	21656.wav	/a/	child	500	470.588	5.88%	510.71	2.14%
10	23737.wav	/o/	child	390	190.476	51.16%	396.785	1.74%
11	15163.wav	/u/	woman	400	380.952	4.76%	409.998	2.50%
12	15477.wav	/a/	man	146	145.455	0.37%	153.513	5.15%
13	20720.wav	/u/	woman	219	210.526	3.87%	222.018	1.38%
17	10128.wav	/i/	woman	302	285.714	5.39%	287.519	4.80%
18	10135.wav	/i/	woman	405	380.952	5.94%	405.733	0.18%
19	25133.wav	/o/	child	262	250	4.58%	268.562	2.50%
20	25135.wav	/o/	child	299	285.714	4.44%	299.945	0.32%
						9.58%		2.78%

Table 5.1: Autocorrelation method compare with DS method.

Refer to the table the conventional method: Autocorrelation method yields 9.58% error rate, but for the new method the error rate just has 2.78%. The error rate has reduced. The Degree of Stationarity (DS) method is proved to be a more accurate method than the autocorrelation method.

Figure 5.5 and Figure 5.6 depict the DS of the two vowel speeches:

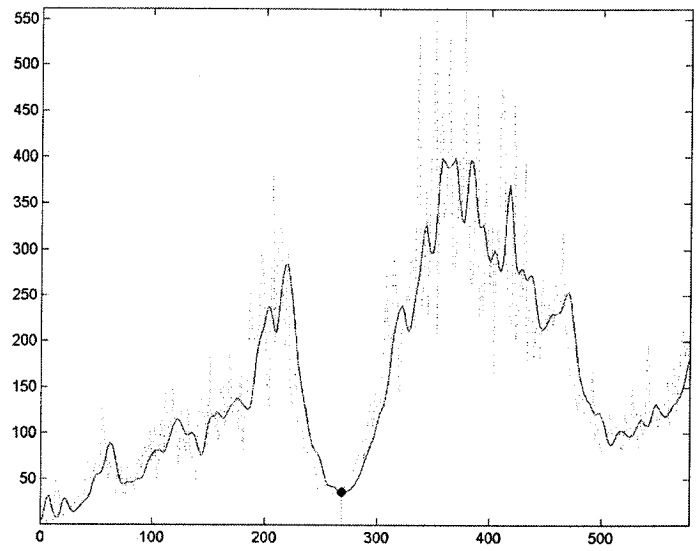


Figure 5.5: DS of the wav file: 25133.wav, F_0: 268.562 Hz (In this figure the X-axis is frequency and the Y-axis is the value of DS)

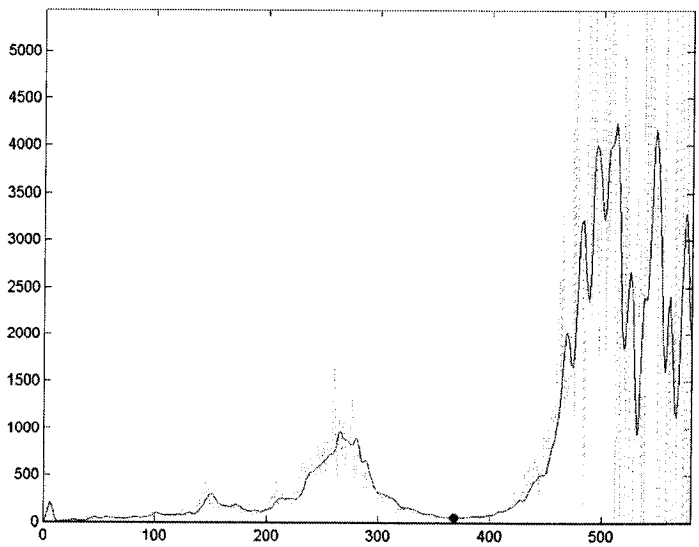


Figure 5.6: DS of the wav file: 28014.wav, F_0: 496 Hz (In this figure the X-axis is frequency and the Y-axis is the value of DS)

5.3. Formants

In Linear acoustics model of speech production, speech sounds are the product of an airflow passed through the glottis, producing resonances in the vocal tract. To produce different

sounds, the vocal tract moves into different configurations that change its resonance structure. The resonance frequencies of the vocal tract are called formants. Voiced sounds, especially vowels, generally have three formants, which are called the first, second and third formants, beginning with the lowest frequency component. They are usually written as F_1, F_2 and F_3. Formant frequencies usually appear as peaks in the spectrum and take values from 250 to 5000 Hz.

Some successful applications, especially the application: Artifact Reduction in EGG that we mentioned in chapter 3, encourage us to utilize the EMD and the Hilbert transform in formant detection. We try to locate peak frequencies in Marginal Spectrum of each IMF to indicate the formants of the speech. Table 5.2 and table 5.3 show that the peak frequencies (bold font in table 5.3) can generally indicate the F_1 (bold font in table 5.2) but can't indicate the other F_2 and F_3.

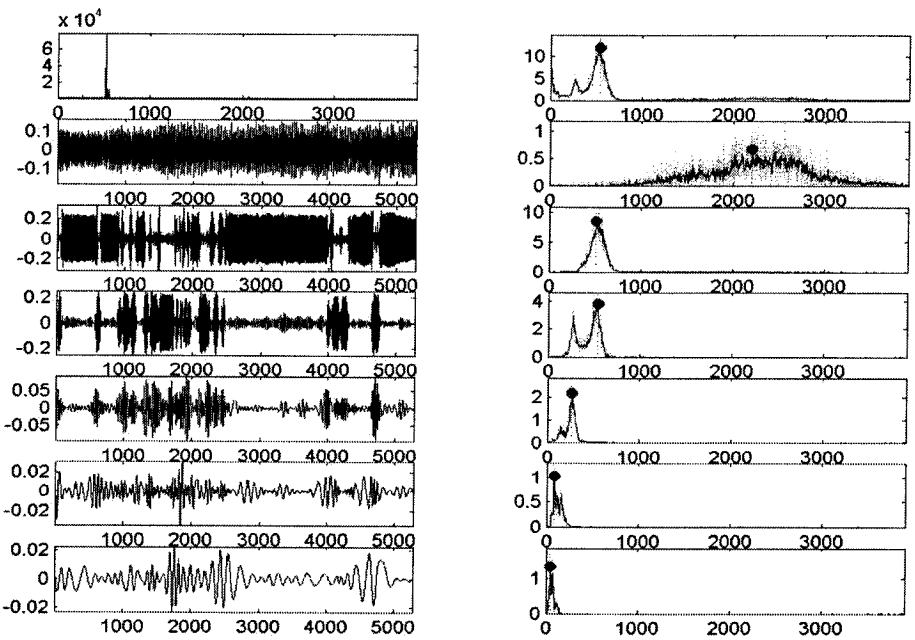


Figure 5.7: Peak frequencies of each IMFs

File		Gender	F_1	F_2	F_3
10881.wav	/a/	Woman	**559.00**	2014.00	2971.00
23591.wav	/o/	Child	**517.00**	2012.00	2941.00
28014.wav	/u/	Child	**499.00**	1982.00	2979.00
17690.wav	/i/	Woman	**714.00**	2853.00	3572.00
12387.wav	/e/	Woman	**398.00**	2649.00	3334.00
12815.wav	/i/	Woman	**369.00**	2587.00	3290.00

Table 5.2: Reference formants

Table 5.2 is the reference F_1, F_2 and F_3 formants from Dieter Maurer's research [18]. Table 5.3 indicates the peak frequencies of the accumulated marginal spectrum and the peak frequencies of the marginal spectrum of the first five IMFs for the same vowel files in the table 5.2. Compare the two tables, we cannot find the necessary relationship between the peak frequencies of each marginal spectrum of the IMF and formants.

File		IMF1	IMF2	IMF3	IMF4	IMF5	Peak Freq
10881.wav	/a/	2856.92	2220.31	**536.26**	297.92	185.03	185.03
23591.wav	/o/	2205.27	**516.03**	540.32	267.12	83.48	541.83
28014.wav	/u/	2766.44	517.37	**495.99**	239.44	132.55	495.99
17690.wav	/i/	3440.22	1530.38	**705.61**	332.42	144.26	3440.22
12387.wav	/e/	2856.99	1443.46	**414.86**	209.57	220.26	211.71
12815.wav	/i/	2837.71	358.76	**367.58**	202.90	89.69	367.58

Table 5.3: Comparison of peak frequency of each IMF and formants

Chapter 6

SPEAKER RECOGNITION

Speaker recognition is classified into two specific types: speaker identification and speaker verification. Speaker identification is to determine which speech in a known group of speeches best matches the speaker, whereas the speaker verification is to determine if the speaker is who he or she claims to be. In speaker recognition speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). Text-independent means that the identification procedure should work for any text in either training or testing. In our project, we focused our attention to the text-independent speaker identification problem. For extensive testing, we try to use pitch combined with Mel Frequency Cepstral Coefficients (MFCCs) as feature in a complete text-independent speaker identification system. As we mentioned in chapter 1, the purpose of the speaker identification system is just for testing and comparison of the new method, not for improvement of the performance of the speaker identification system itself. We hope to use the recognition rate of the speaker recognition system to indicate the performance of the new algorithm.

6.1. Cepstral Coefficients

Cepstral coefficients are derived from an Inverse Discrete Fourier Transform of logarithm of short-term power spectrum of a speech segment *s(t)* as:

$$C(q) = F^{-1}(\log | F(s(t)) |)$$

where $F(\ )$ denotes the discrete-time Fourier transform, $F^{-1}(\ )$ denotes its inverse, $s(t) = g(t) \otimes v(t)$, $s(t)$ is the speech signal, $g(t)$ denotes the excitation signal, $v(t)$ denotes the vocal tract impulse response, and "⊗" denotes convolution.

The cepstral features contains the information due to the slowly varying vocal shaping $v(t)$ in its first coefficients and the information due to the faster varying excitation impulses $g(t)$ in its later coefficients.

Since the physical characteristics of the vocal tract vary from person to person, parameters that are dependent on the vocal tract shape can be considered as features in speaker recognition. Cepstral features are the kind of parameters that can be used to get the shape of the vocal tract.

For example, consider the signal $f(t) = \cos(50t) + \cos(500t)$. Obviously, the $\cos(500t)$ term varies much more quickly than the $\cos(50t)$ term. In order to get rid of the $\cos(500t)$, we could perform a Fourier transform to get rid of the higher frequency terms, and then transform back.

Acoustics model equation $(Log(S(f)) = Log(G(f)) + Log(V(f)))$ is similar in form to $f(t) = \cos(50t) + \cos(500t)$ in the upper example. There is a sum of two terms where one term $(\log(G(f)))$ varies much more rapidly than the other $(\log(V(f)))$. A similar operation can be performed to filter out the faster varying component. Then taking the IFFT gives the "cepstrum" (the reason this is called "cepstrum" is because the first four letters of the word "spectrum" were transposed) $c(n)$.

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(\omega)| e^{j\omega n} d\omega$$

The cepstral features contain the information due to the slowly varying vocal shaping in its first coefficients and the information due to the faster varying glottal impulses in its later coefficients. Therefore, generally only first few (perhaps 12) low cepstral coefficients of the cepstrum are retained.

6.2. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are a variant form of cepstral coefficients.

$$C(q) = F^{-1}(Log(T_{Mel}(|F(s(t))|^2)))$$

where T_{Md} Denotes Mel Frequency Warping

Acoustic Researches have shown that humans pay relatively more attention to the lower frequencies than the higher ones.

The Mel frequency scale can be understood approximately as linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. We can use the following approximate formula to compute the Mels for a given frequency f in Hz: $Mel\ frequency = 2595 \log_{10}(1 + \frac{f}{700})$. Figure 6.1 shows the relation between the Mel frequencies and regular frequencies.

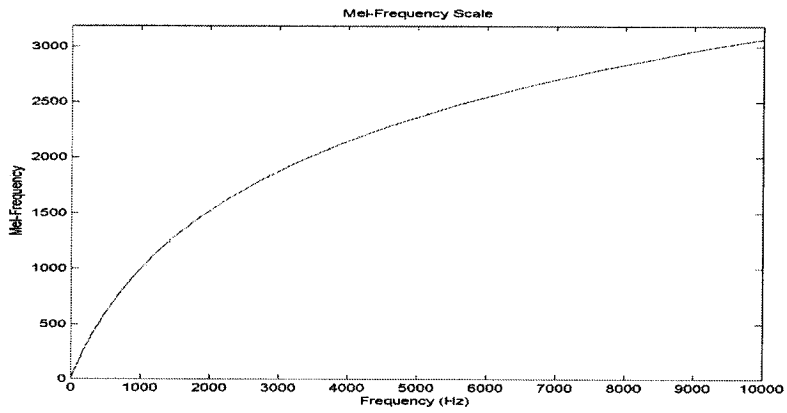


Figure 6.1: Mel Frequency Scale

There are a number of variants of cepstral coefficients. The most common choice is Mel Frequency Cepstral Coefficients that is based on a filter bank model.

6.3. Preprocessing

Before calculating the MFCC features, some preprocessing algorithms usually impose on the speech signal, so that the feature extraction module performed in the next can be more accurate. These preprocessing algorithms include: End Point Detection, Pre-emphasis and Frame Blocking, which are illustrated in Figure 6.2.

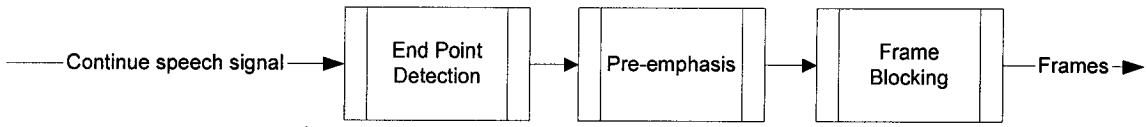


Figure 6.2: Preprocessing of speaker recognition

End Point Detection

Reducing the frontal and appended non-voiced parts is the first step for an efficient speech-processing algorithm. The main motivation behind endpoint detection is that the processing of

these non-voiced parts is bound to undermine the performance of our system. The features extracted from these segments cannot characterize the speaker's identity and can mislead the recognizer. Moreover, we reduce the total processing time of the system by removing these unwanted parts.

There are two parameters that are normally used to identify a spoken utterance from the background noise present in a recorded utterance *x(n)*. The first one is the logarithm of the frame's energy and the other one is the number of zero-crossings.

- Logarithm of the frame's energy

The log energy of each frame is obvious by the equation: $LE = Log(\sum_{n=1}^{N} x(n)^2)$. The logarithm function makes a non-linear compression to the amplitude of the signal and the weak portions of the signal have the opportunity to reveal their details sufficiently [12].

- Number of zero-crossings

The following equation gives the mathematical definition of the number of zero-crossings: $CX = \sum_{n=1}^{N} |\operatorname{sgn}(x(n) - \operatorname{sgn}(x(n-1))|$. Actually, we use zero-crossings rate to correct the endpoints defined by the energy criterion and correctly depict the unvoiced phonemes. The zero-crossings rate can be an efficient tool for that operation, as we know that unvoiced phonemes feature greater high frequency terms.

The endpoint detection algorithms employ the above two parameters to define the endpoints of an utterance. Normally we can use successive tests to choose the thresholds of them. It depends much on the noise level in the input speech waveforms.

Pre-emphasis

Because hearing is more sensitive above the 1 kHz region of the spectrum, the pre-emphasis filter amplifies this area of the spectrum, assisting the spectral analysis algorithm in modeling the most perceptually important aspects of the speech spectrum [14]. From another point of view, pre-emphasis is similar to the auricle (pinna) does in the human auditory system. A Finite Impulse Response (FIR) high pass filter described by the following transfer function is applied to increase the relative energy of the high-frequencies spectrum.

$$H_{pre}(z) = 1 + a_{pre} z^{-1}$$

Normally, a_{pre} takes values from –1.0 to –0.9.

Frame Segmentation and Overlapping

The input signal is segmented into frames of constant length that overlap each other. In these frames, the parameters of the vocal tract model stay unchanged.

The overlapping of speech frames is used in order to increase the redundancy of the input signal, so as to provide more speech data to the feature extraction algorithm. Moreover, we can capture the changes in the vocal tract more accurately. A common choice for overlapping is 50%.

6.4. Calculating MFCC

After preprocessing, a block diagram of the structure of an MFCC processing is given in Figure 6.3. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As has been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFCs are shown to be less susceptible to variations.

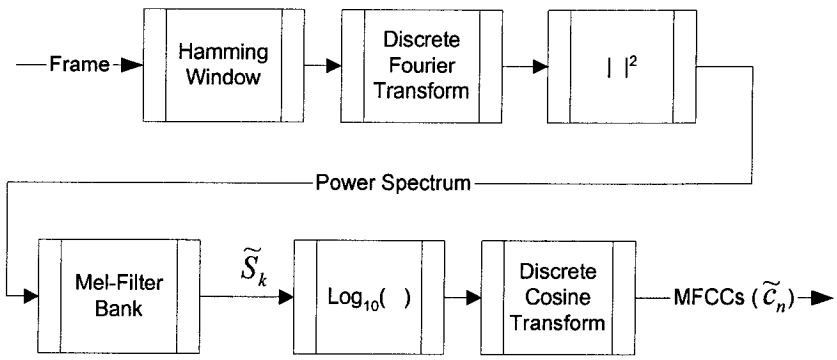


Figure 6.3: Steps of MFCCs extraction

Windowing

The first step for frames in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where N is the number of samples in each frame, then the result of windowing is the signal

$y_l(n) = x_l(n)w(n), 0 \le n \le N-1$. Typically the Hamming window is used, which has the form: $w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), 0 \le n \le N-1$.

Fast Fourier Transform

The next processing step is the Fast Fourier Transform (FFT), which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) that is defined on the set of N samples $\{x_n\}$, as follow: $X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, n = 0, 1, ..., N-1$

In general x_n's are complex numbers. The resulting sequence $\{x_n\}$ is interpreted as follow: the zero frequency corresponds to $n = 0$, positive frequencies $0 < f < F/2$ correspond to values $1 \le n \le N/2-1$, while negative frequencies $-F/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, F denotes the sampling frequency. The result after this step is often referred to as Fourier spectrum.

Mel Frequency Wrapping

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel Frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale (see Figure 6.4). That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The modified spectrum of $S(w)$ thus consists of the output power of these filters when $S(w)$ is the input. The number of Mel spectrum coefficients, K, is typically chosen as 20. This filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle-shape windows in the Figure 6.4 on the spectrum. A useful way of thinking about this Mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

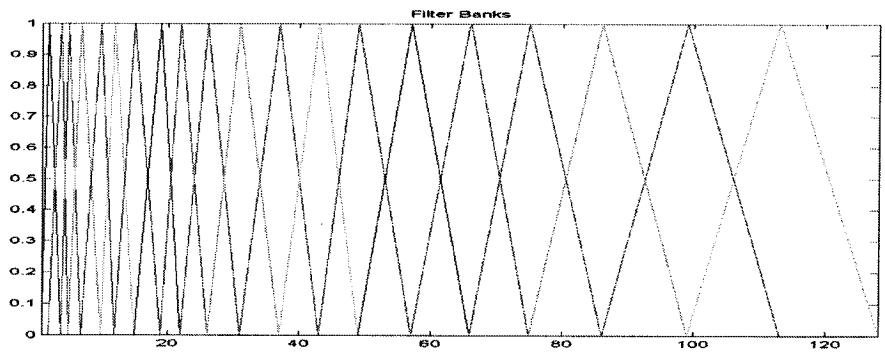


Figure 6.4: An example of Mel-spaced filter banks ($K=20$)

Cepstrum

In this final step, we convert the log Mel spectrum back to time. The result is called the Mel Frequency Cepstral Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore

if we denote those Mel power spectrum coefficients are $\tilde{S}_k$, $k = 1,2,...,K$ we can calculate the MFCCs, $\tilde{c}_n$, as $\tilde{c}_n = \sum_{k=1}^{K} (\log_{10} \tilde{S}_k) \cos(n(k-1/2)\pi / K)$, $n = 1, 2, ..., k$

Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal that carried little speaker specific information.

The main advantage of DCT is that it is an orthogonal transformation, which decorrelates the spectral coefficients very efficiently, that is, converts statistically dependent spectral coefficients into independent cepstral coefficients.

By applying the procedure described above, for each speech frame with overlap, a set of Mel Frequency Cepstral Coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a Mel frequency scale.

6.5. Gaussian Mixture Models

The Gaussian model is a basic parametric model that has merit by itself and can be the basis of the other more sophisticated models.

The use of the Gaussian mixture density for speaker identification is then motivated by two interpretations. First, the individual component Gaussian in a speaker-dependent Gaussian Mixture Models (GMM) are interpreted to represent some general speaker-dependent vocal tract configurations that are useful for modeling speaker identity. Second, a Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker. Finally, the

maximum-likelihood parameter estimation and speaker identification procedures are described [16].

A Gaussian mixture density is a weighted sum of M component densities given by the equation: $p(x \mid \lambda) = \sum_{i=1}^{M} p_i b_i(x)$

where x is a D-dimensional random vector, $b_i(x)$ are the component densities and p_i are the mixture weights. Each component density is a D-variate Gaussian function of the form:

$$b_i(x) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \exp(-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1}(x-\mu_i))$$

with mean vector μ_i and covariance matrix $\sum_i$ The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The mean vectors, covariance matrices and mixture weights from all component densities parameterize the complete Gaussian mixture density. The notation collectively represents these parameters: $\lambda = \{p_i, \mu_i, \Sigma_i\}\ i = 1, ..., M$. For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ.

The GMM can have several different forms depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component (nodal covariance), one covariance matrix for all Gaussian components in a speaker model (grand covariance), or a single covariance matrix shared by all speaker models (global covariance). The covariance matrix can also be full or diagonal. In this thesis, nodal, diagonal covariance matrices are primarily used for speaker models, except as noted for some experiments. This choice is based

on initial experimental results indicating better identification performance using nodal, diagonal variances compared to nodal and grand full covariance matrices. In our project, we utilized GMM functions from Data Clustering and Pattern Recognition Toolboxes [37] for implementations in MATLAB,

6.6 Speech Database: TIMIT

We applied the method to a number of speech utterances from TIMIT database [27]. TIMIT contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform files for each utterance. The performance was studied on speakers from the TIMIT database. The files were down sampled to 8kHz from the original 16kHz for accommodating real world situations and reducing the processing time. Considering the fact that only voiced sections were being used for training and testing, a 28-speaker subset was considered. All the speech files from a speaker were concatenated for this purpose. The first 5 sentences of individual speakers were used to model the GMMs using the MFCC and logarithm of pitch. The last 5 sentences were used for test.

6.7. Experiment Results

Firstly, we attempt to use Marginal spectrum instead of Fourier spectrum in the feature extraction modules. The result (Table 4.2) shows that Marginal spectrum has poorer performance than the Fourier spectrum. As mentioned in chapter 3, one of the reasons is the

instantaneous parameter; especially instantaneous frequencies and instantaneous amplitudes have an uncertain relation with the marginal spectrum. Marginal spectrum only represents the cumulated amplitude over the entire data span in a probabilistic sense and can't ensure to have satisfied performance in every practical application.

When use degree of stationarity (DS) in pitch detection, the result is exciting. Therefore, as an independent feature, we can use pitch derived from DS combined with conventional MFCC features in text-independent speaker identification. In the experiment, 28 speakers were chosen; each speaker has 10 sentences for training and testing. Without pitch feature, the average recognition rate is 83.57% for 14 MFCCs, then use additional pitch feature, the recognition rate increased to 88.57%.

Chapter 7

CONCLUSION AND FUTURE WORKD

7.1. Conclusions

In this thesis we discussed a new method for time-frequency analysis: the Empirical Mode Decomposition and the Hilbert transform. Two typical successful applications were introduced in chapter 3. At the same time, two practical problems and related explanations were pointed out in chapter 4. Next, in chapter 5 applications in pitch and formant detection by using the new method were proposed. Corresponding experiment results were discussed. For extensive testing of the new methods, a text-independent speaker identification system also was tested in chapter 6. The results include negative and positive areas.

Comparing with Fourier transform and Wavelet transform, the EMD and Hilbert transform only involves the signal itself; there are no analyzing functions used in the Fourier transform and the wavelet transform. However, comparing with Fourier Transform and Wavelet Transform, the EMD and Hilbert transform are highly non-linear methods that are very time-consuming.

Additional, we get the following conclusions:

- The individual marginal spectrum of each IMF has better frequency resolution than the accumulated marginal spectrum of the original signal.

- In most cases, the first several IMFs contain certain physical significance. However, these meaningful content cannot ensure to demonstrate the whole physical significance each time.

- Degree of Stationarity can be used successfully in pitch detection and have promising experimental results.

7.2. Future Works

Keele pitch extraction reference database [25] is a small database that was used widely for pitch tracking. For example, in [26] [27] [28] Keele pitch database was deemed as "ground truth" for the evaluation of the pitch-tracking algorithm. Keele pitch database can be downloaded from: ftp://ftp.cs.keele.ac.uk/pub/pitch. We plan to use Keele pitch database to test our pitch detection method for further analysis and estimation.

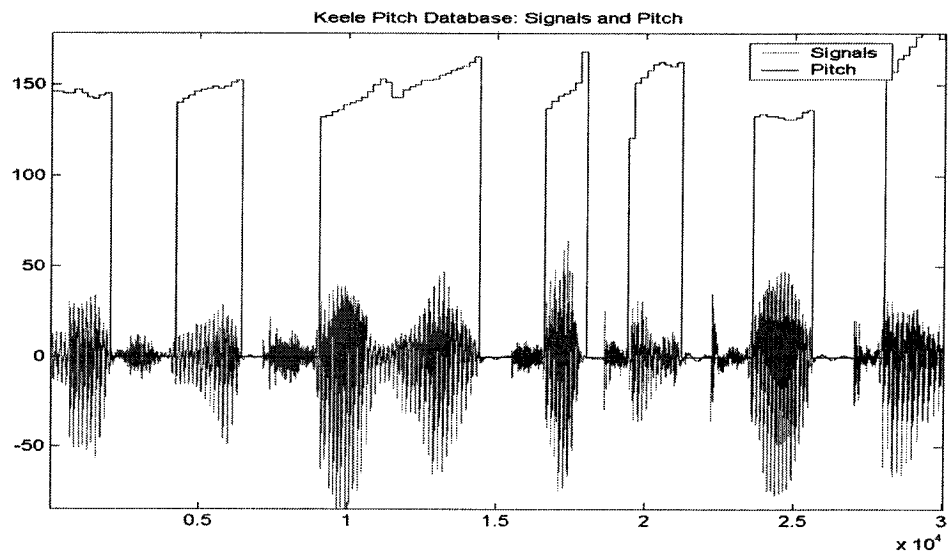


Figure 7.1: Keele pitch database: signals and pitch

Because this thesis focus on the application aspects of the EMD and the Hilbert transform, future work will continue on finding sophisticated mathematics to describe the EMD procedure and to come to a more efficient algorithm.

As we mentioned in section 2.4, the definition of degree of stationary is very similar to the intermittency used in the wavelet analysis proposed in [36]. But Huang et al [1] didn't give a detailed description of the intermittency. In fact, Wavelet Intermittency $\tilde{I}(x.r)$ defined in turbulence analysis [36] measures local deviations from the mean spectrum of *f(x)* at every position *x* and scale *r*:

$$\tilde{I}(x.r) = \frac{|\tilde{f}(x,r)|^2}{\int_{\Re^2} |\tilde{f}(x,r)|^2 \, d^2x}$$

where $\tilde{f}(x.r)$ is the two-dimensional wavelet transform of *f(x)*.

Comparing the two definitions, $DS(\omega) = \frac{T\int_0^T H^2(\omega,t)dt}{h^2(\omega)} - 1$ and $\tilde{I}(x.r) = \frac{|\tilde{f}(x,r)|^2}{\int_{\Re^2} |\tilde{f}(x,r)|^2 \, d^2x}$, the similarities of them are obvious. Consequently, further study about the intermittency in turbulence analysis would be an assistance to improve the definitions of the Degree of Stationarity (DS) and the Degree of Statistic Stationarity (DSS), which can hopefully be a more accurate analysis method to explore arbitrary nonlinear non-stationary signal. In addition, more comparison work needs to be done in speech processing area, for example, comparing EMD with methods using the wavelet transform.

The EMD combined with Hilbert fransform is a highly non-linear method that is very time-consuming. Therefore, improving the efficiency of EMD is what is most important step. Some job has been done, for example in [38], a quicker method called Local Mean Mode Decomposition (LMMD) was introduced to calculate the mean value of the upper and lower envelopes by a two-tap adaptive time-varying filter. It would be interesting to study this in future work. This might gain a further understanding of the EMD method and may assist to prove it.

BIBLIOGRAPHY

[1] N. E. Huang; S. Zheng; L. R. Steven; W. C. Manli; S. H. Hsing; Z. Quanan; Y. Naichyuan; T. Chichao; L. H. Henry. *The empirical mode decomposition and theHilbert spectrum for nonlinear and non-stationary time series analysis*. Proc. R. Soc. Lond. A 454, 903-995. 1998.

[2] L. J. Patrick; T. Berkant. *Comments on the Interpretation of Instantaneous Frequency*. IEEE Signal Processing Letters, Vol. 4, NO. 5. May 1997.

[3] B. Boashash. *Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals*. Proc. IEEE 80, 520-538. 1992.

[4] N. E. Huang; S. Zheng; L. R. Steven. *A New View of Nonlinear Water Waves: The Hilbert Spectrum*. Annu. Rev. Fluid Mech. 31: 417-57. 1999.

[5] L. Hualou; L. Zhiyue; W. M. Richard. *Artifact Reduction in Electrogastrogram Based on the Empirical Mode Decomposition Method*. Medical & Biological Engineering & Computing, 38(1), 35-41, http://www.sahs.uth.tmc.edu/hliang/research.htm. 2000.

[6] C. Chauhuei; L. Chengping; T. Taliang. *Surface-Wave Dispersion Measurements Using Hilbert-Huang Transform*. TAO, Vol. 13, No. 2, 171-184. June 2002.

[7] P.J. Oonincx. *Empirical Mode Decomposition: A New Tool for S-Wave Detection*. PNA-R0203, ISSN 1386-3711. 2002.

[8] T. Schlurmann; S. Schimmels; T. Dose. *Spectral Analysis of Freak waves Using Wavelet Spectra (Morlet) and Hilbert Spectra (EMD)*. 4th Int. Conf. on Hydro-science & Engineering, Seoul, South Korea, 26-29. Sep. 2000.

[9] J.Flanagan. *Speech Analysis and Perception*. 2nd ed. New York and Berlin: Springer-Verlag p. 10, Fig. 1972.

[10] S. Roweis. *Speech Processing Background*. http://www.cs.toronto.edu/~roweis/notes.html. Nov. 1998.

[11] S. Krishnakumar; K.R.P. Kumar; N. Balakrishnan. *Pitch maxima for robust speaker recognition*. Acoustics, Speech and Signal Processing. (ICASSP '03). IEEE International Conference on, Pages: II - 201-4 vol.2. April 2003.

[12] J. Markel; B. Oshika; A. J. Gray. *Long-term feature averaging for speaker recognition*. Acoustics, Speech, and Signal Processing, IEEE Transactions on, Volume: 25, Issue: 4, Pages: 330 – 337. Aug 1977.

[13] F. Sadaoki. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, INC. New York. 1989.

[14] J. W. Picone. *Signal modeling techniques in speech recognition*. Proceedings of the IEEE, Volume: 81, Issue: 9, Pages: 1215 – 1247. Sept. 1993.

[15] N. Mitianoudis. *A graphical framework of the Evaluation of Speaker Verification systems*. MSc Thesis, University of London, http://egnatia.ee.auth.gr/~mitia/page4.htm. Sept 2000.

[16] D. A. Reynolds; R. C. Rose. *Robust text-independent speaker identification using Gaussian mixture speaker models*. Speech and Audio Processing, IEEE Transactions on, Volume: 3, Issue: 1, Pages: 72 – 83. Jan. 1995.

[17] S. M. Griebel. *Multi-channel wavelet techniques for reverberant speech analysis and enhancement*. Technical Report 5, HIMMEL, Harvard University, Cambridge, MA, http://citeseer.ist.psu.edu/griebel99multichannel.html. Feb. 1999.

[18] D. Maurer. *Acoustics of the vowel*. University Hospital Zurich, Department of Neurology, Neuropsychology Unit, http://www.neurol.unizh.ch/psychologie/associates/maurer.

[19] H. Rosler. *A Study on the Empirical Mode Decomposition.* Master's Thesis, Faculty of Natural Sciences, University of Amsterdam. Dec 2002.
[20] P. Kuchi. *Gait Recognition Using Empirical Mode Decomposition Based Feature Extraction.* Master's Thesis, Arizona State University. Dec 2003.
[21] *TIMIT.* Acoustic-Phonetic Continuous Speech Corpus, National Inst. Standards Technol. Speech Disc 1-1.1, NTIS Order PB91-505 065. Oct. 1990.
[22] M. J. Carey; E. S. Parris; T. H. Lloyd; S. Bennett. *Robust prosodic features for speaker identification.* Spoken Language, ICSLP 96. Proceedings, Fourth International Conference on, Pages: 1800 - 1803 vol.3. Oct. 1996.
[23] F. Minyue. *Speech Production and Acoustic-Phonetics.* Tutorials on Speech Technology, Department of Electrical and Computer Engineering, The University of Newcastle, http://murray.newcastle.edu.au/users/staff/speech/home_pages/tutorial_acoustic.html. Jul 1997.
[24] A. Cherif. *Pitch and formants extraction algorithm for speech processing.* Electronics, Circuits and Systems, ICECS. The 7th IEEE International Conference on, Volume: 1, 17-20 Pages: 595 - 598 vol.1. Dec. 2000.
[25] F. Plante; G. Meyer; W. A. Ainsworth. *A pitch extraction reference database. EUROSPEECH'95*, Madrid, pp. 837-840. 1995.
[26] C. Wang; S. Seneff. *Robust pitch tracking for prosodic modeling in telephone speech.* Acoustics, Speech, and Signal Processing. ICASSP'00. Pages: 1343 - 1346 vol.3. June 2000.
[27] K. Kasi; S.A. Zahorian. *Yet another algorithm for pitch tracking.* Acoustics, Speech, and Signal Processing. (ICASSP '02). Pages: I-361 - I-364 vol.1. May 2002.
[28] E. Mousset; W. A. Ainsworth; J.A.R. Fonollosa. *A comparison of several recent methods of fundamental frequency and voicing decision estimation.* Spoken Language. ICSLP 96. Pages: 1273 - 1276 vol.2. Oct. 1996.
[29] M.I.Todorovska. *Estimation of Instantaneous Frequency of Signals using the Continuous Wavelet Transform.* Report CE 01-07, University of Southern California, Department of Civil Engineering, Los Angeles, California. December 2001.
[30] M. Schwartz; W. R. Bennett; S. Stein. *Communications systems and techniques.* New York: McGraw-Hill. 1966.
[31] S. O. Rice. *Mathematical analysis of random noise.* Bell Sys. Tech. Jl 23, 282-310. 1944.
[32] S. O. Rice. *Mathematical analysis of random noise II. Power spectrum and correlation functions.* Bell Sys. Tech. Jl 23, 310-332. 1944.
[33] S. O. Rice. *Mathematical analysis of random noise III. Statistical properties of random noise currents.* Bell Sys. Tech. Jl 24, 46-108. 1945.
[34] S. O. Rice. *Mathematical analysis of random noise IV. Noise through nonlinear devices.* Bell Sys. Tech. Jl 2 4, 109-156. 1945.
[35] H. M. S. Longuet. *The statistical analysis of random moving surface.* Phil. Trans. R. Soc. Lond. A249, 321-387. 1957.
[36] M. Farge. *Wavelet Transforms and Their Applications to Turbulence.* Annual Reviews of fluid mechanics, Vol. 24: 395-457. 1992.
[37] J. R. Jyhshing. *Data Clustering and Pattern Recognition Toolboxes for MIR Lab.* CS Dept., Tsing Hua University, Taiwan, http://neural.cs.nthu.edu.tw/jang/matlab/toolbox/. 2004.
[38] G. Qiang; M. Xiaojiang; Z. Haiyong; Z. Yankun. *Processing time-varying signals by a new method.* Radar, CIE International Conference on, Proceedings, 15-18 Pages: 1011 – 1014. Oct. 2001.

[39] R. Balocchi; D. Menicucci; M. Varanini. *Empirical mode decomposition to approach the problem of detecting sources from a reduced number of mixtures*. Engineering in Medicine and Biology Society. Proceedings of the 25th Annual International Conference of the IEEE, Volume: 3, 17-21 Pages: 2443 - 2446 Vol.3. Sept. 2003.

[40] A. Bouzid; N. Ellouze. *Empirical mode decomposition of voiced speech signal*. Control, Communications and Signal Processing. First International Symposium on, Pages: 603 – 606. March 2004.

[41] A.O. Andrade; P.J. Kyberd; S.D. Taffler. *A novel spectral representation of electromyographic signals*. Engineering in Medicine and Biology Society. Proceedings of the 25th Annual International Conference of the IEEE, Volume: 3, 17-21 Pages: 2598 - 2601 Vol.3. Sept. 2003.

[42] C. Chengjie; L. Weixian; S. F. Jeffrey; L. Yilong. *Doppler frequency extraction of foliage penetration radar based on the hilbert-huang transform technology*. Radar Conference. Proceedings of the IEEE, Pages: 170 – 174. 2004.

[43] Y.J. Kim; J.H. Chung. *Pitch synchronous cepstrum for robust speaker recognition over telephone channels*. Electronics Letters, Volume: 40, Issue: 3, Pages: 207 – 208. Feb. 2004.

[44] C. Miyajima; Y. Hattori; K. Tokuda; T. Masuko; T. Kobayashi; T. Kitamura. *Speaker identification using Gaussian mixture models based on multi-space probability distribution*. Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '01). IEEE International Conference on, Volume: 1, 7-11 May 2001 Pages: 433 - 436 vol.1. 2001.

[45] E. W. Weisstein. *Nonstationary Time Series*. MathWorld - Wolfram Research, Inc. http://mathworld.wolfram.com/NonstationaryTimeSeries.html. 2004.

[46] G. P. J. Dwyer. *Nonlinear Time Series and Financial Applications*. Federal Reserve Bank of Atlanta, http://www.dwyerecon.com/pdf/lectnlin.pdf. January 2003.