A PROTOTYPE OF THAI SPEECH RECOGNITION SYSTEM FOR BASIC VOICE COMMANDING

KRIENGKRAI NANTANITIKRON

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2008

Thesis
Entitled

A PROTOTYPE OF THAI SPEECH RECOGNITION SYSTEM FOR BASIC VOICE COMMANDIND

……..……
Mr. Kriengkrai Nantanitikron
Candidate

……..……
Asst. Prof. Pisit Phokharatkul,
D.Eng. (Electrical Engineering (Computer))
Major-Advisor

……..……
Lect. Songpol Ongwatanakul,
Ph.D. (Electrical and Computer Engineering)
Co-Advisor

……..
Lect. Kanat Poolsawasd,
M.Sc. (Informatics)
Co-Advisor

……
Prof. Banchong Mahaisavariya, M.D.
Dean
Faculty of Graduate Studies

……..…
Lect. Pronchai Chanyagorn,
Ph.D. (Computer Engineering)
Chair
Master of Science Programme in Technology of Information System
Faculty of Engineering

Thesis
Entitled

A PROTOTYPE OF THAI SPEECH RECOGNITION SYSTEM FOR BASIC VOICE COMMANDIND

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
February 25 , 2008

……..……
Mr. Kriengkrai Nantanitikron
Candidate

……..……
Assoc. Prof. Supachai Phaiboon
D.Eng. (Electrical Engineering)
Chair

……..
Asst. Prof. Pisit Phokharatkul,
D.Eng. (Electrical Engineering (Computer))
Member

……..……
Lect. Songpol Ongwatanakul,
Ph.D.(Electrical and Computer Engineering)
Member

……..…
Lect. Kanat Poolsawasd,
M.Sc. (Informatics)
Member

……
Assoc. Prof.Vittaya Tipsuwanporn,
M.Eng. (Electrical Engineering)
Member

……..……
Prof. Banchong Mahaisavariya, M.D.
Dean
Faculty of Graduate Studies
Mahidol University

……
Asst. Prof. Rawin Raviwongse, Ph.D.
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENT

I would like to express my gratitude and sincere appreciate to those who help me to complete this thesis.

The first person, I would like to express my profound gratitude to my thesis advisor, Asst. Prof. Dr. Pisit Phokharatkul, for his invaluable advice, revision of this thesis, numerous suggestions, time and continuous support on my thesis which were the key to the successful accomplishment of this study. He was always around to listen and give advice. He meticulously reviewed all related documents and supported the choices I made toward the completion of this thesis.

My deep appreciation goes to my very kind co-advisor, Dr. Songpol Ongwattanakul, who showed me some helpful tips, techniques, and skills that help me to overcome technical difficulties during the research. He also help me to revise this thesis.

Special gratitude to all committees including Lect. Kanat Poolsawasd, and my thesis chair committee, Asso. Prof. Dr.Supachai Phaiboon, who gave very useful comments, suggestions, reviews and revisions of my thesis. Moreover, I would like to express my appreciation the external committee, Asso. Prof. Vittaya Tipsuwanporn, who gave me some insightful comments, reviews and revisions my work.

My sincere thanks to the staffs at the Technology of Information System Management program for their services during the past few years at the Faculty of Engineering, Mahidol University. Their kind supports have been of great inspiration during the down time. I am grateful for my IT’46 friends and glad to be one of them.

Finally, I wish to extend my appreciation and thank to my beloved parent for everlasting support and care throughout my whole life.

Kriengkrai Nantanitikron

A PROTOTYPE OF THAI SPEECH RECOGNITION SYSTEM FOR BASIC VOICE COMMANDING

KRIENGKRAI NANTANITIKRON 4637194 EGTI/M

M.Sc.(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORS: PISIT PHOKHARATKUL, D.Eng.,
SONGPOL ONGWATTANAKUL, Ph.D., KANAT POOLSAWASD, M.Sc.

ABSTRACT

Most recent Thai speech recognition systems that have been continually researched often use feature extraction from Fourier transform based methods and some learning algorithm such as Artificial Neural Networks (ANN), Hidden Markov Models (HMM), and Linear Prediction Codes (LPC).

This study presents a Thai speech recognition system based on Fourier transform and a set of filter banks as a feature extraction which is called double filter banks. The goal of this study was to develop a prototype of a real-time single-word Thai speech recognition system that can recognize some common Thai words. The study encompass speech capturing, analog-to-digital conversion, automatic speech marking, identification of starting and ending points, pre-emphasis, speech feature extraction, speech feature matching, and displaying the output to user.

The author create an application to analyze a comprehensive set of speech data and created a multi-dimensional speech feature that includes speech achieving operations such as read, write, and load into the main memory in JAVA language.

The system was evaluated for its accuracy and stability in performing various conditions. The accuracy was validated by an experiment with 9,000 speeches from several volunteers. The average accuracy rate is 94.6% in an offline test. Finally, the result shows that the evaluation was beyond satisfaction for every aspect.

KEY WORDS: THAI SPEECH RECOGNITION /SPEECH RECOGNITION/ PATTERN RECOGNITION / FILTER BANK /AI

71 pp.

ระบบต้นแบบการรู้จำเสียงภาษาไทยสำหรับการสั่งงานด้วยเสียงขั้นพื้นฐาน
(A PROTOTYPE OF THAI SPEECH RECOGNITION SYSTEM FOR BASIC VOICE COMMANDING)

เกรียงไกร นันทนิธิกร 4637194 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการควบคุมวิทยานิพนธ์ : พิศิษฏ์ โภคารัตน์กุล, D.Eng., ทรงพล องค์วัฒนกุล, Ph.D., ฆนัท พูลสวัสดิ์, M.Sc.

บทคัดย่อ

งานวิจัยที่ค้นคว้าในเรื่องราวของการรู้จำเสียงหรือการรู้จำคำพูดภาษาไทยนั้น ได้มีการทำวิจัยต่อมาอย่างสืบเนื่อง โดยส่วนมากมักใช้ผลการแปลงฟูเรียร์เป็นพื้นฐานในคุณลักษณะของเสียง และทำการรู้จำหรือจัดกลุ่มต่ออีกครั้งโดยวิธีการนิวรอลเน็ตเวิร์คซึ่งเป็นวิธีที่นิยมกันมาก หรือโดยวิธีรูปแบบของฮิดเดนมาคอฟ ที่นำเอาคุณลักษณะมาออกแบบเป็นกฎในการรู้จำ หรือวิธีการพยากรณ์แบบเชิงเส้น

ในส่วนของงานวิจัยนี้จะนำเสนอระบบต้นแบบการรู้จำเสียงภาษาไทยที่ดัดแปลงการจัดกลุ่มของเสียงโดยหาคุณลักษณะด้วยผลการแปลงฟูเรียร์ และผ่านกระบวนการฟิลเตอร์แบงก์ (Filter bank) จำนวน 2 ครั้ง ซึ่งเป็นวิธีที่ยังไม่มีผู้ใดทดลองทำมาก่อน ซึ่งจุดมุ่งหมายในการศึกษาระบบนี้ เพื่อพัฒนาระบบต้นแบบรู้จำเพื่อตรวจสอบคำพูดภาษาไทยว่าเป็นคำใด โดยระบบจะประกอบด้วยส่วนรับคำพูด ส่วนแปลงเป็นดิจิตอล ส่วนการค้นหาหัว-ท้ายคำ ส่วนการเร่งความถี่ ส่วนการเฟ้นหาคุณลักษณะของเสียง ส่วนการเปรียบเทียบคุณลักษณะของเสียง และส่วนแสดงผลลัพธ์ว่าคำพูดที่ตรวจสอบนั้นเป็นคำพูดใด

ผู้วิจัยได้พัฒนาระบบนี้ ไม่ว่าจะเป็นส่วนต่างๆที่ใช้คำนวณข้อมูลที่ซับซ้อน การจัดเก็บข้อมูล เช่น คุณลักษณะของเสียงลงคอมพิวเตอร์ หรือกระทั่งการจัดการหน่วยความจำของคอมพิวเตอร์ ด้วยภาษาจาวา

ระบบนี้ได้มีการประเมินโดยยึดความถูกต้องและความเสถียรในการทำงาน เพื่อทดสอบความถูกต้องเหมาะสมของส่วนต่างๆ จากการสอนระบบด้วยคำพูด 11,000 ตัวอย่าง และทดสอบด้วยคำพูด 20 คำ จำนวน 9,000 ตัวอย่างที่ไม่เคยสอน พบว่ามีความถูกต้องอยู่ที่ระดับ 94.6% ซึ่งแสดงให้เห็นว่าในทุกส่วนของระบบได้ผลการทำงานอยู่ในระดับที่ดี เป็นไปตามวัตถุประสงค์ที่กำหนดไว้

71 หน้า

CONTENTS

CONTENTS (CONT.)

LIST OF TABLE

Page

Table

LIST OF FIGURES

LIST OF FIGURES (CONT.)

LIST OF FIGURES (CONT.)

CHAPTER 1
INTRODUCTION

1.1 Background and Problem Statement

Nowadays, computer technology has become an important thing in human life. The computer technology can enable people with effective, efficient, yet easier working and living life styles. However, the way to command a computer remains largely through the use of keyboard and mouse which are considered unnatural to human. In recent years, there has been a growing popularity in making a more natural user interface between human and computer. As a result, many researches around the world pursuit speech recognition as a mean to communicate with a computer in a natural way.

Speech or speaking is one of the human general communication methods. Communication between human and machine through speech is highly desirable. It enhances the human-computer interaction with less effort. Therefore, speech recognition is a natural interface to convey commands to a computer instead of typing on the keyboard or clicking on the mouse.

Speech recognition system allows the users to provide input to an application via their voice. Similar to clicking on the mouse button, typing on the keyboard, or pressing a key on the touchtone pad, the speech recognition allows computer user to provide input by using speech commanding. Therefore, in doing so, this kind of computers may require an audio capture interface and a microphone, instead of keyboard and mouse.

Nowadays, the improvement of speech recognition technology is growing faster and faster, and many new techniques are more and more discovered. In the field

of speech recognition, the researches move very fast, but still could not reach a satisfaction level due to the variety of languages in the world. Usually, a new speech recognition system is tuned to its native language. Unique language such as Thai language contains many phoneme combinations which post additional challenges in Thai speech recognition. More researches on Thai speech recognition are needed to make computers to recognize Thai speech. Therefore, Thai people can communicate with the computer by speaking Thai.

1.2 Objective of Study

The object of this study is

1.2.1. Study theory of pronouncing words and phonemes.

1.2.2 Study theory of speech recognition.

1.2.3 Develop a prototype of Thai speech recognition and basic voice commanding system.

1.3 Scope of Study

The scope of this study will

1.3.1 Use isolated utterance in 20 samples speech type.

1.3.2 Use DFT basis.

1.3.3 Use Filter Bank basis.

1.3.4 Classify and verify data to information from analog speech to digital representation.

1.3.5 Classify the sample speech to the correct identify group of speech for the system basic needed.

1.3.6 Row signal data cutting, selecting and justify by system.

1.3.7 Operate on microcomputer with Microsoft Windows XP.

1.3.8 Developed by JAVA Second Edition Version 1.4.1_02

1.3.9 Speech recognition not depended on subject and except feeling related.

1.4 Expected Result

The outcome of this study will be a prototype of a Thai speech recognition system that can

1.4.1 Recognize common Thai words through the use filter bank feature.

1.4.2 Detect voice command.

1.4.3 Give insight to the future development of such system.

CHAPTER 2
LITERATURE REVIEW

2.1 Linguistics and Phonetics

2.1.1 Linguistics

Linguistics is the scientific study of language, which can be theoretical or applied. Someone who engages in this study is called a linguist.

Theoretical (or general) linguistics encompasses a number of sub-fields, such as the study of language structure (grammar) and meaning (semantics). The study of grammar encompasses morphology (formation and alteration) of words and syntax (the rules that determine the way words combine into phrases and sentences). Also a part of this field are phonology, the study of sound systems and abstract sound units, and phonetics, which is concerned with the actual properties of speech sounds (phones), non-speech sounds, and how they are produced and perceived [1].

2.1.2 Phonetics

Phonetics (phonê) that mean "sound" or "voice" is the study of the physical sounds of human speech. It is concerned with the physical properties of speech sounds (phones), and the processes of their physiological production, auditory reception, and neurophysiologic perception [2].

The task of any phonetic theory is to determine the form of a phonetic component by establishing the internal and external constraints on that component. The phonetic component itself converts linguistic knowledge of the structure of the

speech act into time-varying commands suitable for control of the articulatory mechanism. Performing involves knowledge, and this knowledge must be expressed in a form accessible to the speaker operating in time. Knowing how to use knowledge of performance constraints involves manipulation of the conversion from segmental notional time embodied in simple sequencing to timing of muscular control. A solution to the handling of this time conversion is discussed.

2.2 Fundamentals Theory of Human Speech Production

The speech system in which to review into the way of anatomy shown by midway through the upper torso as look on from the right side. The gross components of the system are the lungs, trachea (windpipe), larynx (organ of speech production), pharyngeal cavity (throat), oral or buccal cavity (mouth), and nasal cavity (nose). The larynx is commonly referred to as the "voice box." It continues the basic line of conduction that leads the laryngophraynx into the trachea [3]. Level with the fourth through the sixth vertebrates, the larynx is located along the anterior midline of the neck.

The larynx is responsible for two basic functions, to protect the trachea and the lungs from food and fluids as well as permitting the passage of air into the respiratory system. The larynx is also vital in the production of sound, and has a triangular-box-like shape. In technical discussions, the pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract.

Accordingly, the vocal tract begins at the output of the larynx (vocal cords, or glottis) and terminates at the input to the lips. The nasal tract begins at the velum and ends at the nostrils. When the velum (a trapdoor-like mechanism the back of the oral cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the trachea, the tensed vocal cords within the larynx are

caused to vibrate by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the throat, the oral cavity, and possibly nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, mouth), different sounds are produced.

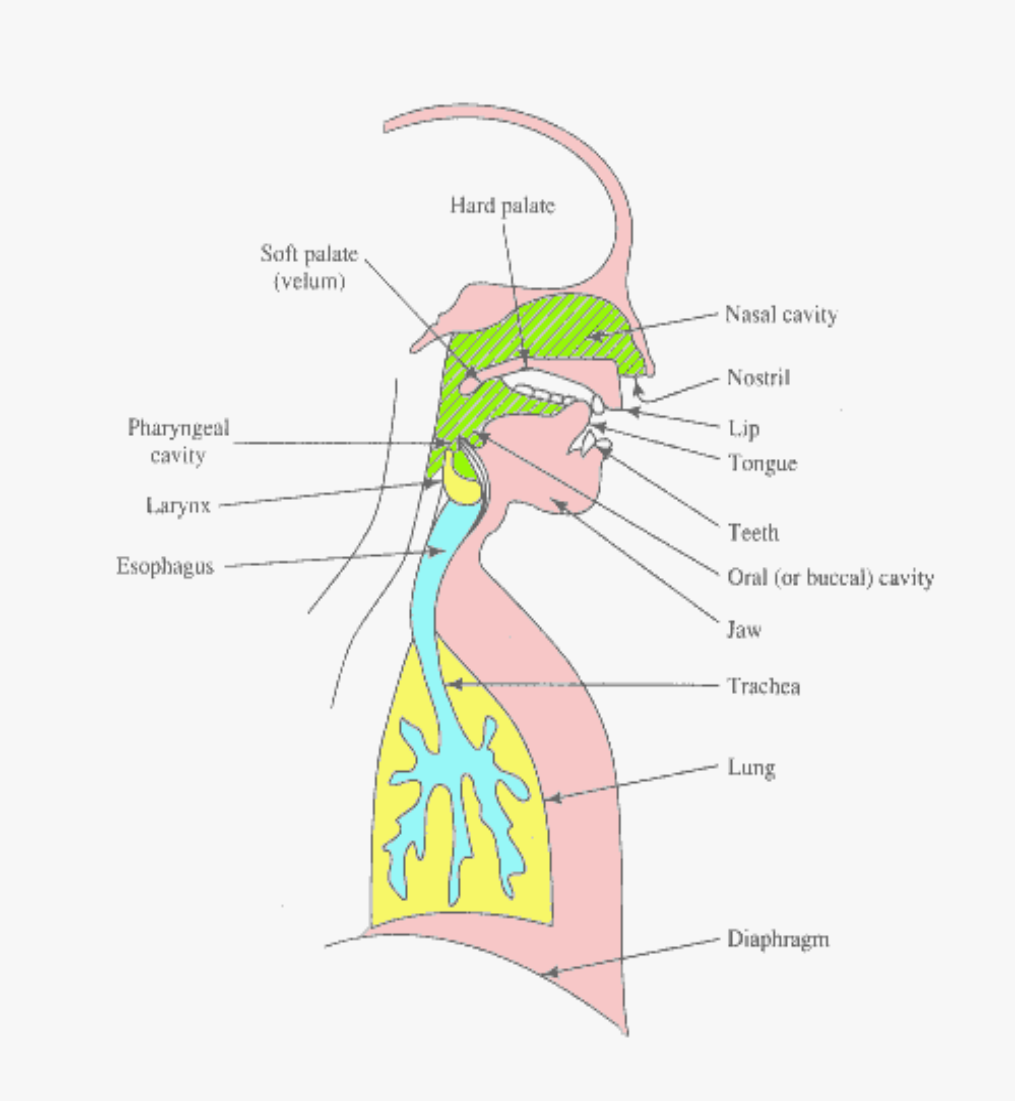

Figure 2.1 Schematic view of human speech production mechanism.

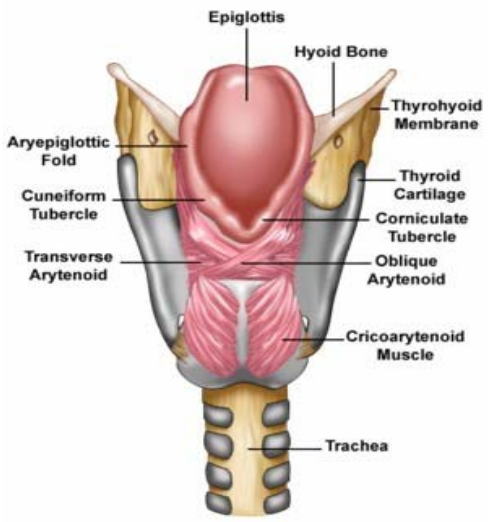

Figure 2.2 Larynx mechanisms.

A simplified representation of the complete physiological mechanism for creating speech is shown in Figure 2.1.The lungs and the associated muscles act as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs (shown schematically as a piston pushing up within a cylinder) and through the trachea. When the vocal cords are tensed, the air flow causes them to vibrate, producing so-called voiced speech sounds.

When the vocal cords are relaxed, in order to produce a sound, the air flow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sounds, or it can build up pressure behind a point of the total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly release, causing a brief transient sound [4].

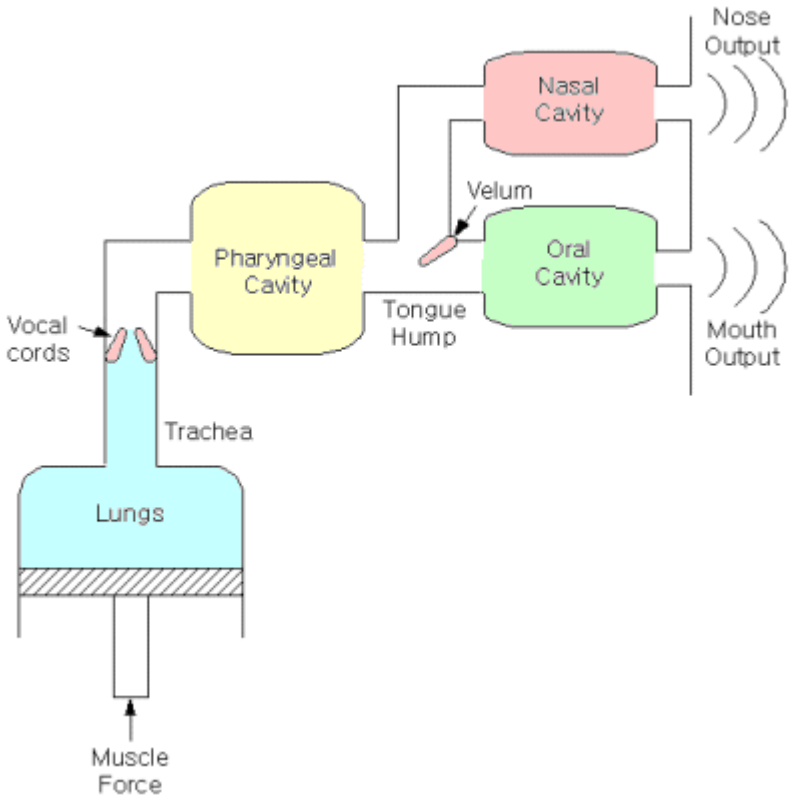

Figure 2.3 Block diagram of human speech production.

2.3 Prosody

Finding correct intonation, stress, and duration from written text is probably the most challenging problem for years to come. These features together are called prosodic or suprasegmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The intonation means how the pitch pattern or fundamental frequency changes during speech. The prosody of continuous speech depends on many separate aspects, such as the meaning of the sentence and the speaker characteristics and emotions. The prosodic dependencies are shown in Figure 2.4 However, with some specific control characters this information may be given to a speech synthesizer.

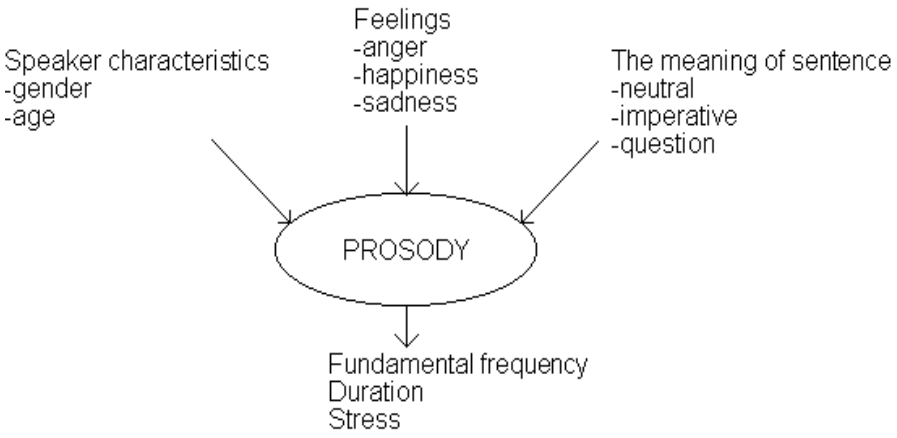

Figure 2.4 Prosodic dependencies

The mechanisms of speech production models, as shown in Fig. 2.5, in a speech production model, a phoneme symbol (equivalent to a pronunciation key) for a spoken word is specified, a motor goal (task) for the articulator movement is generated, and a motor trajectory for the vocal organ is calculated for the motor task assigned. The vocal tract shape is determined from the position of the vocal organs, and speech is produced by controlling the speech production model using the vocal tract area [5].

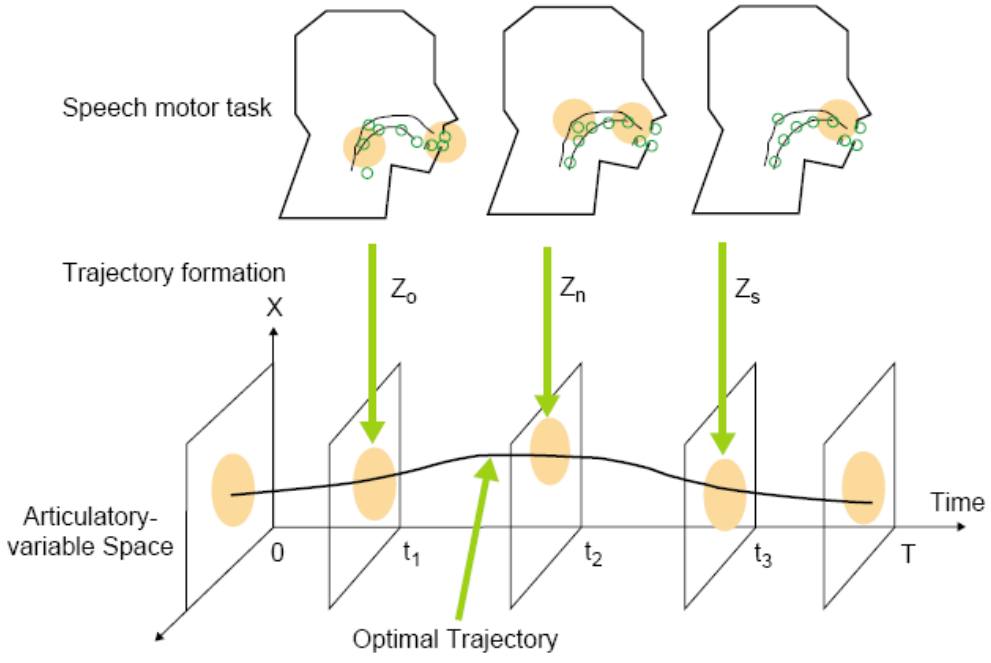

Figure 2.5 Trajectory formation models of articulator movements.

The mechanism of speech is composed by following in processes: language processing, in which the content of an utterance is converted into phonemic symbols in the brain's language center for generation of motor commands to the vocal organs in the brain's motor center and articulator movement for the production of speech by the vocal organs based on these motor commands and the emission of air sent from the lungs in the form of speech [5].

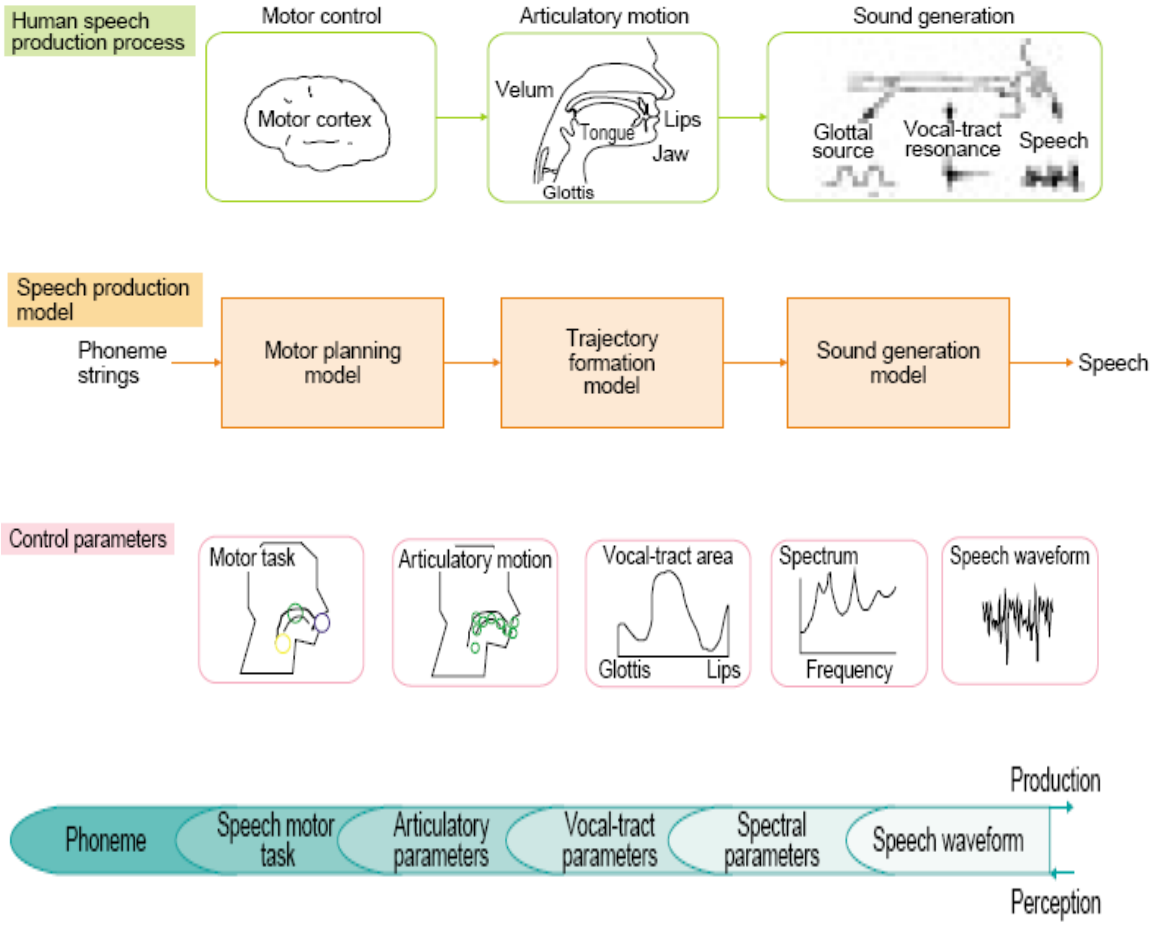

Figure 2.6 Step of Sound generation

2.4 Digital signal processing (DSP)

Digital signal processing is the study of signals in a digital representation and the processing methods of these signals. DSP and analog signal processing are subfields of signal processing. DSP includes subfields like: audio signal processing, control engineering, digital image processing and speech processing. RADAR Signal processing and communications signal processing are two other important subfields of DSP.

In general cases, the signal of interest is initially in the form of an analog electrical voltage, produced for example by a microphone or some other type of transducer. In some situations, such as the output from the data readout system of a CD (compact disc) player, the data is already in digital form. An analog signal must be converted into digital form before DSP techniques can be applied. An analog electrical voltage signal, for example, can be digitized using an electronic circuit or software called an analog-to-digital converter or A/D Converter or ADC. This generates a digital output as a stream of binary numbers whose values represent the electrical voltage input to the device at each sampling instant.

Signals commonly need to be processed in a variety of ways. For example, the output signal from a transducer or microphone may well be contaminated with unwanted electrical voltage, "noise", or other unused information. Processing the signal using a filter circuit can be remove, reduce the unwanted part of the signal or pickup the important information. The filtering of signals to improve signal quality or to extract important information is done by DSP techniques rather than by analog electronics.

Since the goal of DSP is usually to measure or filter continuous real world analog signals, the first step is usually to convert the signal from an analog to a digital form, by using an analog to digital converter. Often, the required output signal is another analog output signal, which requires a digital to analog converter.

The algorithms required for DSP are sometimes performed using specialized computers, which make use of specialized microprocessors called digital signal processors.

2.5 Signal Sampling

A digital signal is often a numerical representation of a continuous signal. This discrete representation of a continuous signal will generally introduce some error in to the data. The accuracy of the representation is mostly dependent on two things; sampling frequency and the number of bits used for the representation. The continuous signal is usually sampled at regular intervals and the value of the continuous signal in that interval is represented by a discrete value. The sampling frequency or sampling rate is then the rate at which new samples are taken from the continuous signal. The number of bits used for one value of the discrete signal tells us how accurately the signal magnitude is represented. Similarly, the sampling frequency controls the temporal or spatial accuracy of the discrete signal.

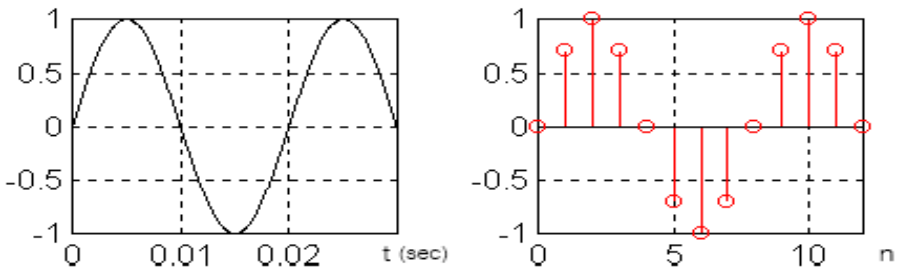

Figure 2.7 Example to show discrete point after digitization the signal [6]

Sampling rate determines the sound frequency range (corresponding to pitch) which can be represented in the digital waveform. The range of frequencies represented in a waveform is often called its bandwidth. Waveforms sampled at a high sampling rate can represent a broad range of frequencies and hence have broad bandwidth. In fact, the maximum bandwidth of a sampled waveform is determined exactly by its sampling rate; the maximum frequency represent able in a sampled waveform is termed its Nyquist frequency, and is equal to one half the sampling rate. Thus, for example, a waveform sampled at 16,000 Hz can represent all frequencies up to its Nyquist frequency of 8,000 Hz.

A problem called aliasing occurs when a signal to be sampled contains energy at frequencies above the sampling Nyquist frequency [7]. The figure 2.8 illustrates how aliasing would occur when the sampling rate is much too low for the frequency of an input signal. The solid curve represents the analog signal at a comparatively high frequency. Circles show where samples were taken at a relatively low sampling rate. The dotted line illustrates the apparent frequency of the sampled waveform, completing about two cycles in the period that the original signal completed 20 cycles [8].

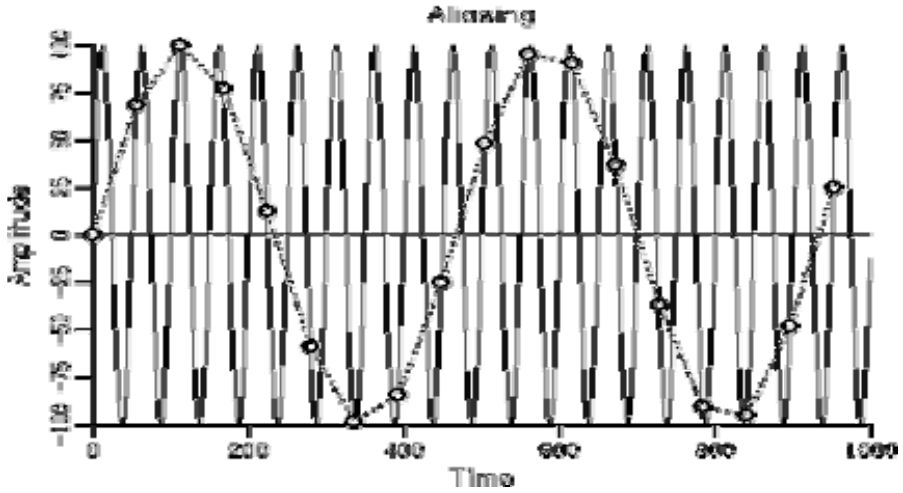

Figure 2.8 Example to show point of sampling

Obviously, aliasing has the effect of producing sounds of lower frequency from sounds that are higher in frequency than the Nyquist frequency. Once aliasing has occurred, it is absolutely impossible to distinguish a component generated by aliasing from one that was actually present in the input signal. This effect is one of the most common sources of distortion in digitized waveforms. Fortunately, most modern computer hardware for digitizing sound has built in filters which are tuned to remove sound energy at frequencies beyond the Nyquist frequency for whatever sampling rate is being used.

2.6 Time-domain

Time-domain is a term used to describe the analysis of mathematical functions, or real-life signals, with respect to time. In the time-domain, the signal or function's value is known at various discrete time points; or for all real numbers, for the case of continuous time.

Sound in time domain

Sound is perceived when fluctuations in air pressure cause structures inside our ears to vibrate. These air pressure fluctuations can be quite small or large and can occur slowly or rapidly. We refer to the rate at which pressure fluctuates cyclically from higher to lower to higher and so forth as its frequency.

Typically, we express frequency in cycles per second or equivalently Hertz. The following figure is a graph of two "cycles" of fluctuation. As shown in figure 2.9 the Amplitude of air pressure variations relative to mean air pressure (in no particular units) as a function of Time (expressed in milliseconds or thousandths of a second). Thus, 0 on the Pressure scale corresponds to the mean air pressure. In this figure the pressure starts at the average air pressure, increases to a value of 100 at a time corresponding to about 1.25 milliseconds, decreases to -100 at 3.75 milliseconds and returns to zero at 5.0 milliseconds before starting the second cycle.

The length of each cycle in time is called the period of the waveform because the shape of the waveform repeats periodically at this interval. Since the period of this waveform is 5.0 milliseconds, there would be 200 periods or cycles in one second. The frequency of this sound is thus 200 cycles per second or 200 Hertz (which we will abbreviate as Hz hereafter). More generally, the frequency of a periodic waveform is the inverse of its period; F = 1/P or in this example, 200 = 1.0 / 0.005. In addition to the frequency of a sound, we can describe its amplitude.

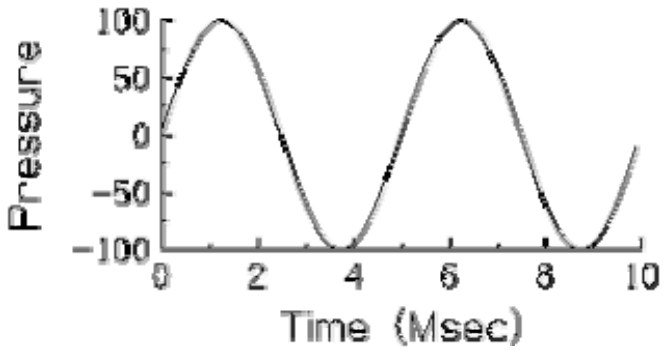

Figure 2.9 Example the amplitude of sound signal at 200 Hertz

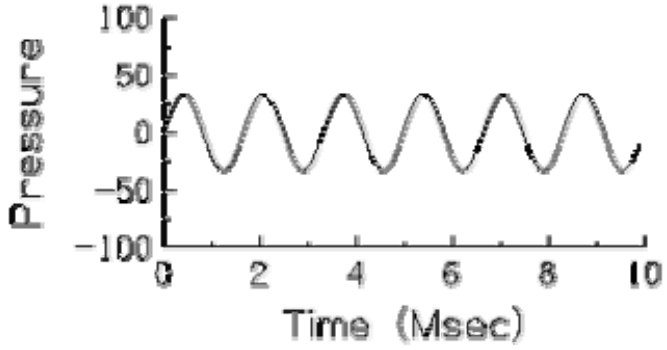

Figure 2.10 Another difference frequency of sound signal from signal in figure 2.9

In general, small variations in pressure produce weak (or quiet) sounds while large variations produce strong (or loud) sounds. The figure 2.10 shows another sound which is lower in amplitude that the previous example because the pressure varies less extremely over time.

In another word the figure 2.10 shows a sound which also differs in frequency from the sound illustrated in the previous figure. Note that frequency and amplitude vary independently. Although the amplitude in figure 2.9 is lower than figure 2.10, the pressure fluctuations are more rapid than in the figure 2.9; six cycles occur within ten milliseconds so this tone has a frequency of 600 Hz. Consequently, this sound, figure 2.9, is higher in frequency but lower in amplitude than the sound depicted in the figure 2.10.

One other property called phase is important in describing the physical properties of sound. To illustrate what is meant by phase, the figure 2.11 shows two 200 Hz sinusoids, one drawn with a solid line and the other drawn with a dotted line. The two sinusoids are identical except that they are differently aligned with respect to the time axis. These two sinusoids are said to differ in phase while having the same amplitude and frequency. This is a good moment to point out that the notion of `beginning' and `ending' needs some qualification here. The figures drawn on this page have waveforms which obviously begin and end within the limits of the graph. However, they represent snippets of functions which do not have beginning and ending points. Thus, the phase differences shown in the present figure do not reflect the notion that one function started at a different time than the other. Rather, the phase differences represent the way the two functions are aligned with respect to each other at all times, including those which lie outside the bounds of the present graph [9].

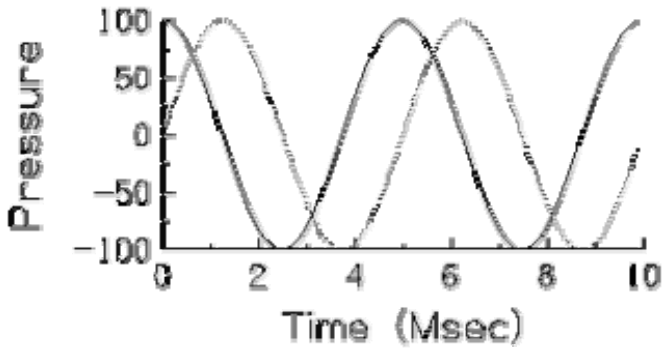

Figure 2.11 Example shown the difference phase of sound signal

2.7 Frequency –domain

Frequency domain is a term used to describe the analysis of mathematical functions or signals with respect to frequency.

A time domain graph shows how a signal changes over time, whereas a frequency domain graph shows how much of the signal lies within each given frequency band over a range of frequencies. A frequency domain representation can also include information on the phase shift that must be applied to each sinusoid in order to be able to recombine the frequency components to recover the original time signal.

The frequency domain relates to the Fourier transform or Fourier series by decomposing a function into an infinite or finite number of frequencies. This is based on the concept of Fourier series that any waveform can be expressed as a sum of sinusoids (sometimes infinitely many).

The representation of sound in the time domain is important to understand, but in some ways it is also hard to identification the feature. For instance, the frequency of a sound is one of its most important physical properties, but determining frequency from a waveform requires making measurements of time intervals and then doing arithmetic. Indeed, for many complex waveforms, where multiple sinusoids of various frequencies are simultaneously present, it is often unclear where the intervals to be measured begin and end. The frequency domain provides an alternative description of sound in which the time axis is replaced by a frequency axis. In the frequency domain, sounds are represented in a frequency by amplitude and/or phase diagram.

The figure 2.12 is a frequency domain representation of the 200 Hz sine wave that saw in the figure 2.9. In the frequency domain, this sound is represented by a line at a point on the frequency axis corresponding to 200 Hz and with a length corresponding to its amplitude. Figures like this are called line spectra.

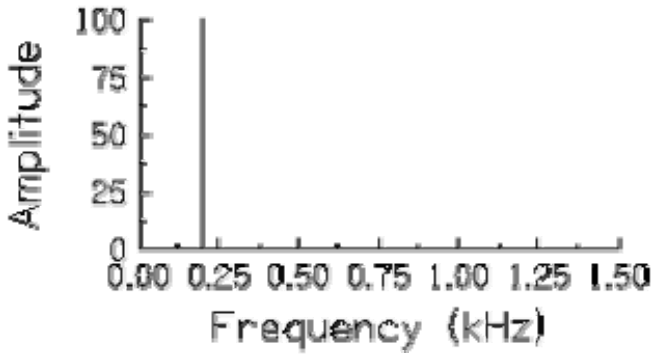

Figure 2.12 Frequency domain representation of the 200 Hz sine wave

There are several things to note in figure 2.12. First, the Y axis is labeled amplitude rather than pressure because the axis now provides a measure of the strength of the pressure changes: neither absolute pressure, nor the direction of relative pressure change is represented. In fact, pressure need not be the physical measure on which amplitude is based here. With sound, we often measure the voltage fluctuations produced by a microphone rather than pressure per second. Consequently, amplitude is a better, more general, term. Second, note that the amplitude axis has no values less that zero. In this spectral representation, called a magnitude spectrum amplitudes cannot be less than zero it is not possible to have negative amounts of sound energy. A third feature to note is the labeling of the frequency axis which is in units of kilohertz or thousands of cycles per second.

One of the most convenient features of frequency domain representations of sound is that sounds of many different frequencies can be plotted simultaneously on the same figure. This figure, for instance, shows all of the components we used above to start an approximation to a square wave. In figure 2.13, each line is one of the harmonics of the 200 Hz fundamental frequency of the square wave. The height of each harmonic line indicates the amplitude of the sinusoid at that frequency. The figure 2.13 does not show us anything about the phase relationships among the harmonics which were obvious in the time-domain figures earlier. Notice that the amplitude reduces very quickly with each successive harmonic in this spectrum. In fact, the apparent differences in amplitude are actually much larger than the

differences we would hear when listening to each of these tones. In this next figure, amplitude is expressed in dB rather than in linear units. The amplitude relations among the harmonics expressed in dB are much closer to the loudness relations we hear among the harmonics [10].

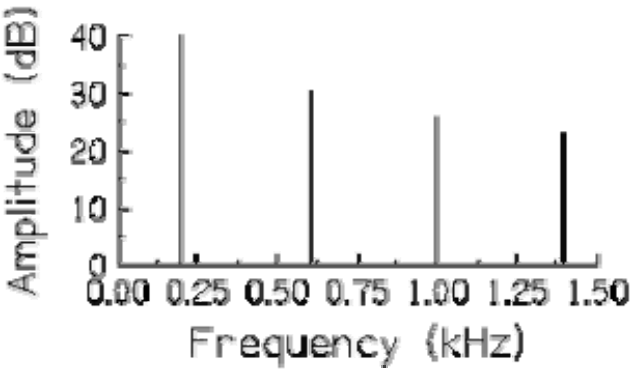

Figure 2.13 Harmonics of the 200 Hz fundamental frequency of the square wave

2.8 Start point/End point of Speech Detection

During speech recognition of words, a precise and strong detection of start/end points of the words must be ensured.

Windows energy based analysis

One method to detect the start point or end point of speech signal is "*Windows energy based analysis*". This model can be function by capture the signal in small time called "*window*" or "*frame*". The window is sliding from start to end of signal and calculate energy for each window. The function to calculate energy follows as:

1. Calculated summary of all signal value in window called "noise" from first 4-5 window and using maximize of "noise" value that called "Energy of noise" (or from experiment by get silent sound from microphone and calculate average noise value).

2. Calculated "Energy" by summary of all signal value in each window to compare with noise value.

3. If “Energy of sound” more than “Energy of noise”, exciting state, then mark position to “start point” and next exciting state mark position to “stop point”.

4. Cutting target sound, speech signal, and save in new file.

5. Repeat until end of signal.

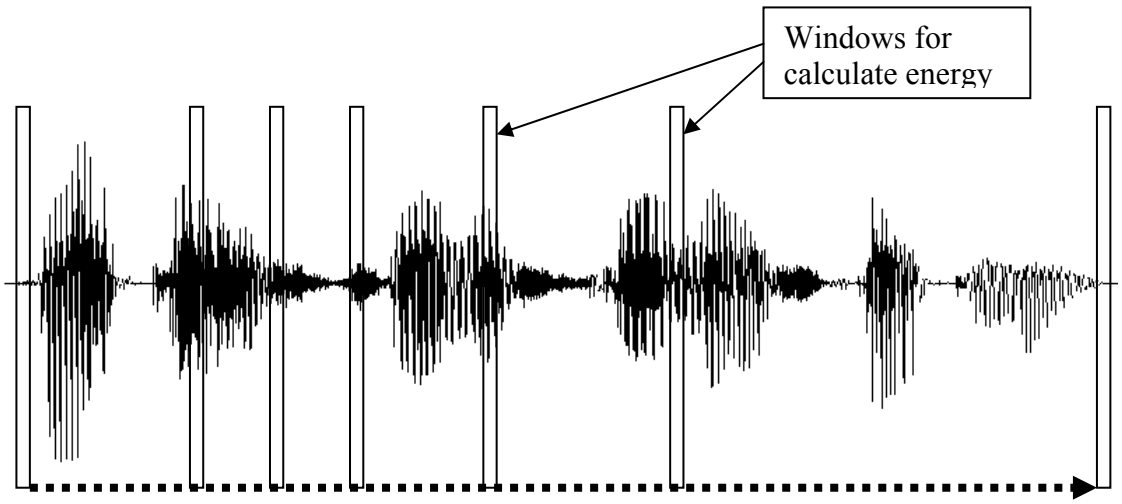

Figure 2.14 Sliding windows for find start point / end point

2.9 Pre-emphasis

In time-domain speech high frequency components of input digital speech samples are emphasized by a pre-emphasis filter. Therefore, speech analysis, whether DFT-based, FFT-based or LPC-based, must be carried out on short segments across which the speech signal is assumed to be stationary. Typically, the feature extraction is performed on 20 to 30 ms windows with 10 to 15 ms shift between two consecutive windows. Pre-emphasis is also traditionally used to compensate for the -6dB/octave spectral slope, used to boost signal spectrum, of the speech signal. The pre-emphasis filter is applied on the input signal before windowing. In the time domain, the pre-emphasized signal is related to the selected digital signal by the relation:

$$X'(n) = X(n) - 0.95\, X(n-1) \quad \text{Where } n = 1,\ldots., N-1$$

2.10 Windowing

For speech processing, method want to assume the signal is short-time stationary and perform a Fourier transform on these small blocks (windows or frames). Solution: multiple the signal by a window function that is zero outside some defined range.

The rectangular window is defined as:

$$w_n = \begin{cases} 1 & 0 \le n < N \\ 0 & \text{otherwise} \end{cases}$$

But consider the discontinuities this can generate, as illustrated in figure 2.9

Figure 2.15 A waveform truncated with a rectangular window

One way to avoid discontinuities at the ends is to taper the signal to zero or near zero and hence reduce the mismatch.

The most common in speech analysis is the Hamming window [11]:

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) & n = 0 \ldots N\text{-}1 \\ 0 \text{ Otherwise} \end{cases}$$

This is simply a raised cosine and is plotted out in figure 2.16. [12]

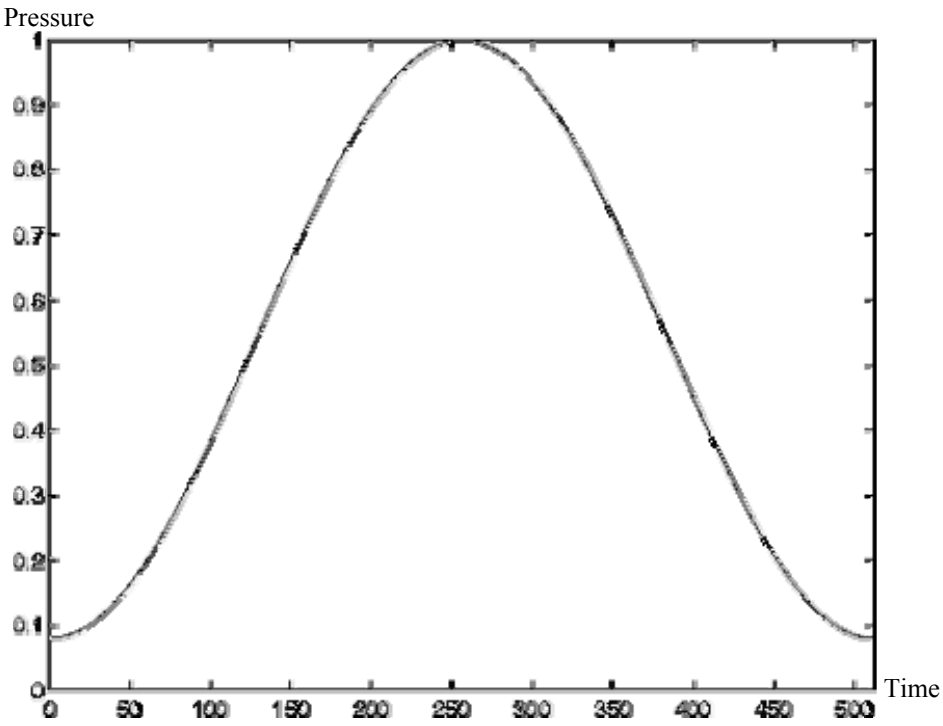

Figure 2.16 The hamming window

2.11 Discrete Fourier transform (DFT)

The discrete Fourier transform (DFT), occasionally called the finite Fourier transform [14], is a transform for Fourier analysis of finite-domain discrete-time signals. It is widely employed in signal processing and related fields to analyze the frequencies contained in a sampled signal, to solve partial differential equations, and to perform other operations such as convolutions.

The sequence of *N complex numbers* x_0, ..., x_{N-1} is transformed into the sequence of *N* complex numbers X_0, ..., X_{N-1} by the DFT according to the formula:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

Where e is the base of the natural logarithm, i is the imaginary unit ($i^2 = -1$), and π is pi. The transform is sometimes denoted by the symbol $\mathcal{F}$, as in or $\mathcal{F}(\mathbf{x})$ or $\mathcal{F}\mathbf{x}$.

In the signal processing literature, it is common to write the DFT and its inverse in the more pure form below, obtained by setting in the previous definition:

$$X(k) \triangleq \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N}, \qquad k = 0, 1, 2, \ldots, N-1$$
$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}, \qquad n = 0, 1, 2, \ldots, N-1$$

Where *x(n)* denotes the input signal at time (sample) *n*, and *X(k)* denotes the *k*-th spectral sample. This form is the simplest mathematically, while the previous form is easier to interpret physically [15].

Note that the normalization factor multiplying the DFT and IDFT (here 1 and 1/N) and the signs of the exponents are merely conventions, and differ in some treatments. The only requirements of these conventions are that the DFT and IDFT have opposite-sign exponents and that the product of their normalization factors be 1/N. A normalization $1/\sqrt{N}$ of for both the DFT and IDFT makes the transforms unitary, which has some theoretical advantages, but it is often more practical in numerical computation to perform the scaling all at once as above.

There are two remaining symbols in the DFT we have not yet defined:

$$j \triangleq \sqrt{-1}$$
$$e \triangleq \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.71828182845905\ldots$$

The $j = \sqrt{-1}$ first,, is the basis for complex numbers.1.1 As a result, complex numbers will be the first topic we cover in this book (but only to the extent needed to understand the DFT).

The second, $e = 2.718\ldots$, is a (transcendental) real number defined by the above limit.

Note that not only do we have complex numbers to contend with, but we have them appearing in exponents, as in

$$s_k(n) \triangleq e^{j2\pi nk/N}.$$

We will systematically develop what we mean by imaginary exponents in order that such mathematical expressions are well defined.

with e, j, and imaginary exponents understood, we can go on to prove Euler's Identity:

$$e^{j\theta} = \cos(\theta) + j\sin(\theta)$$

Euler's Identity is the key to understanding the meaning of expressions like

$$s_k(t_n) \triangleq e^{j\omega_k t_n} = \cos(\omega_k t_n) + j\sin(\omega_k t_n).$$

2.12 Filter Bank

A filter bank is an array of band-pass filters that separates the input signal into several components, each one carrying a single frequency sub-band of the original signal. It also is desirable to design the filter bank in such a way that sub-bands can be recombined to recover original signal. The first process is called analysis, while the second is called synthesis. The output of analysis is referred as sub-band signal with as many sub-bands as there are filters in filter bank [17].

The filter bank serves to isolate different frequency components in a signal. This is useful because for most applications some frequencies are more important than others. For example these important frequencies can be coded with a fine resolution. Small differences at these frequencies are significant and a coding scheme that preserves these differences must be used. On the other hand, less important frequencies do not have to be exact. A coarser coding scheme can be used, even though some of the finer details will be lost in the coding [18].

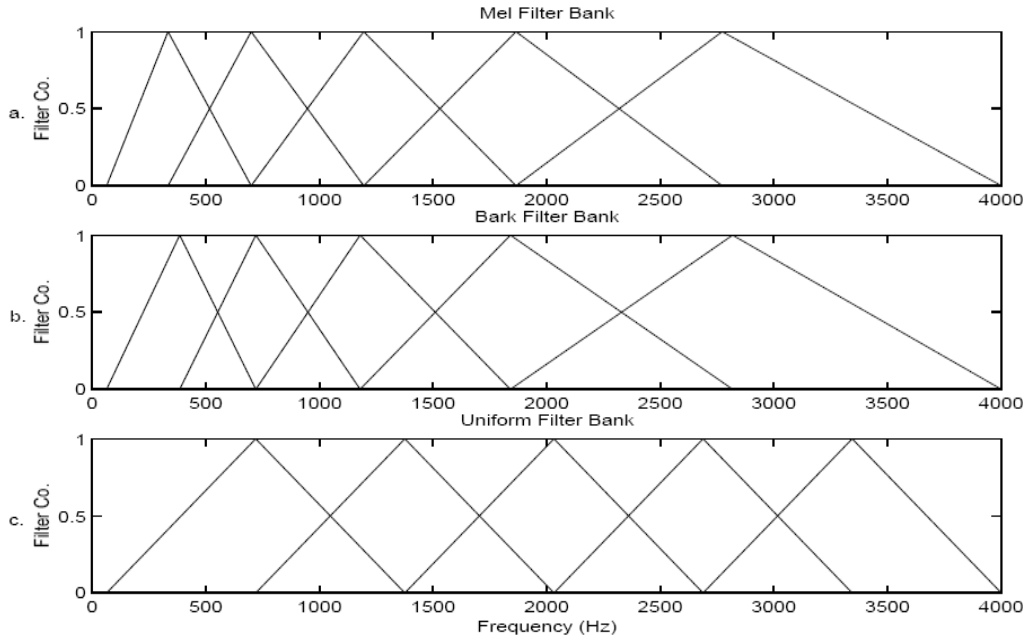

Figure 2.17 Example Filter Banks

2.13 Mel Spectrum and Mel Frequency Filter Bank (Mel– Filter Bank)

The Filters from input power spectrum through a bank of number of Mel-filters. The output is an array of filtered values, typically called Mel-spectrum, each corresponding to the result of filtering the input spectrum through an individual filter. Therefore, the length of the output array is equal to the number of filters created.

The triangular Mel-filters in the filter bank are placed in the frequency axis so that each filter's center frequency follows the Mel scale in figure 2.18, in such a way that the filter bank mimics the critical band, which represents different perceptual effect at different frequency bands. Additionally, the edges are placed so that they coincide with the center frequencies in adjacent filters. Pictorially, the filter bank looks like figure 2.19 [20].

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

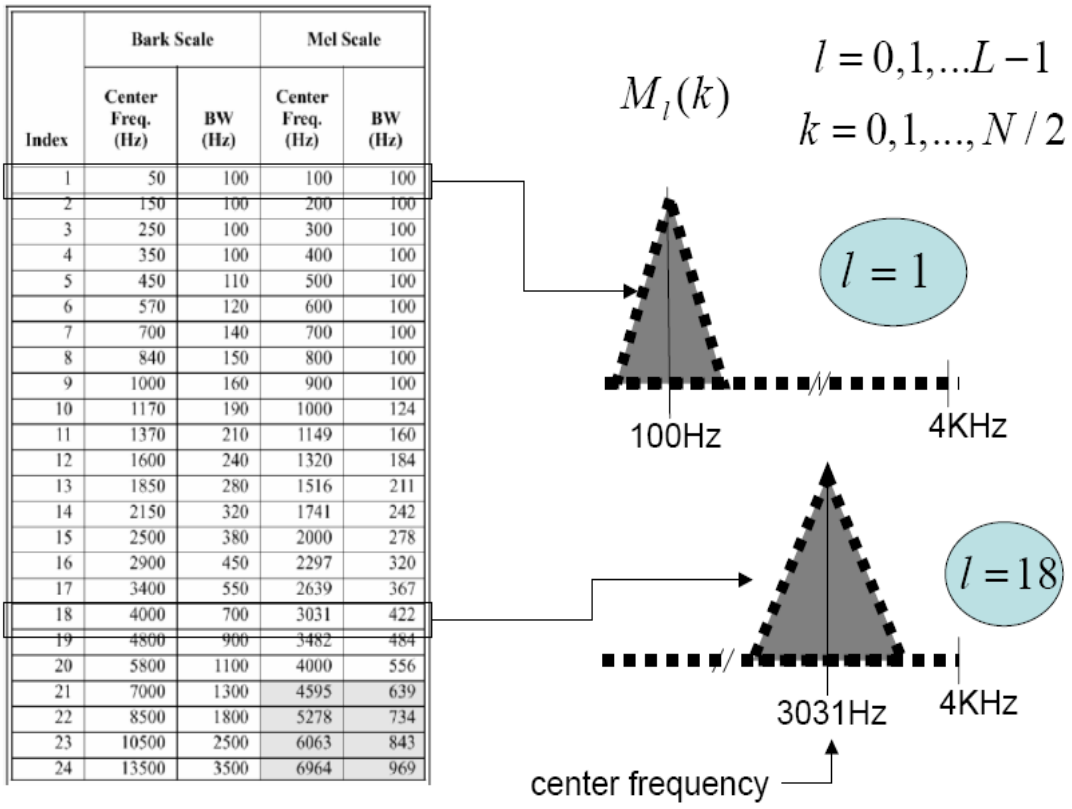

Figure 2.18 Mel Filter bank frequency table

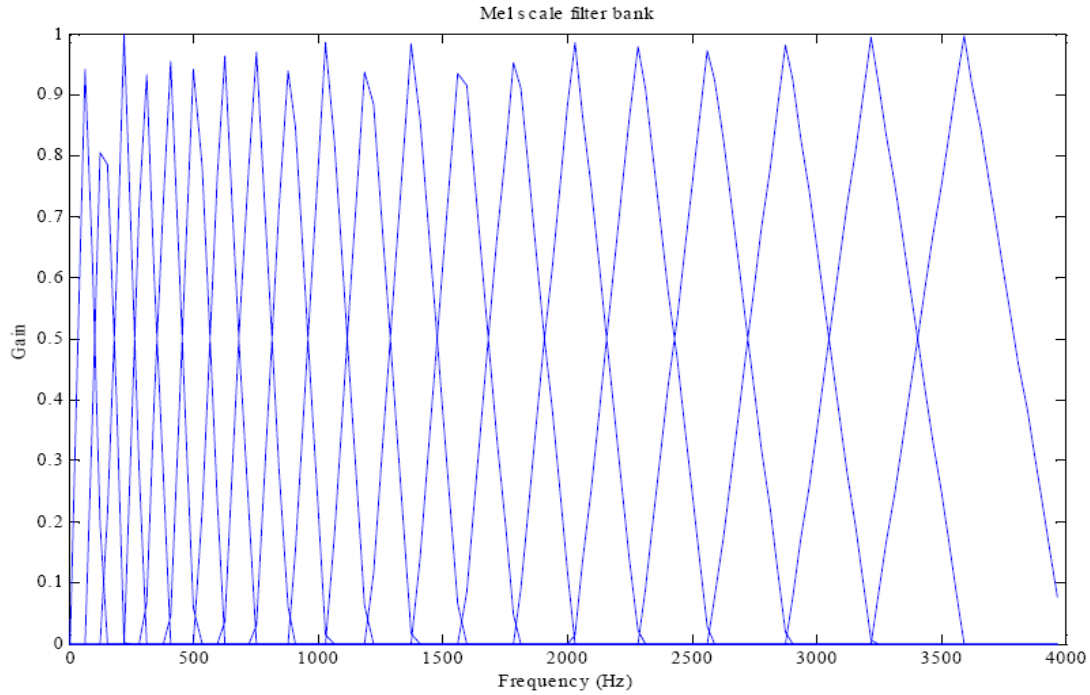

Figure 2.19 Mel Filter bank

Addition, calculate the power spectrum with each of the triangular Mel weighting filters by Mel frequency scale:

$$\text{Mel}(f) = 2595 \text{ Log}_{10} (1+\left(\frac{f}{700}\right))$$

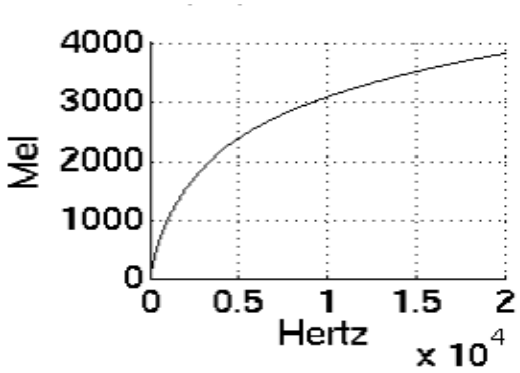

Figure 2.20 Show the wrapping frequency to Mel frequency scale

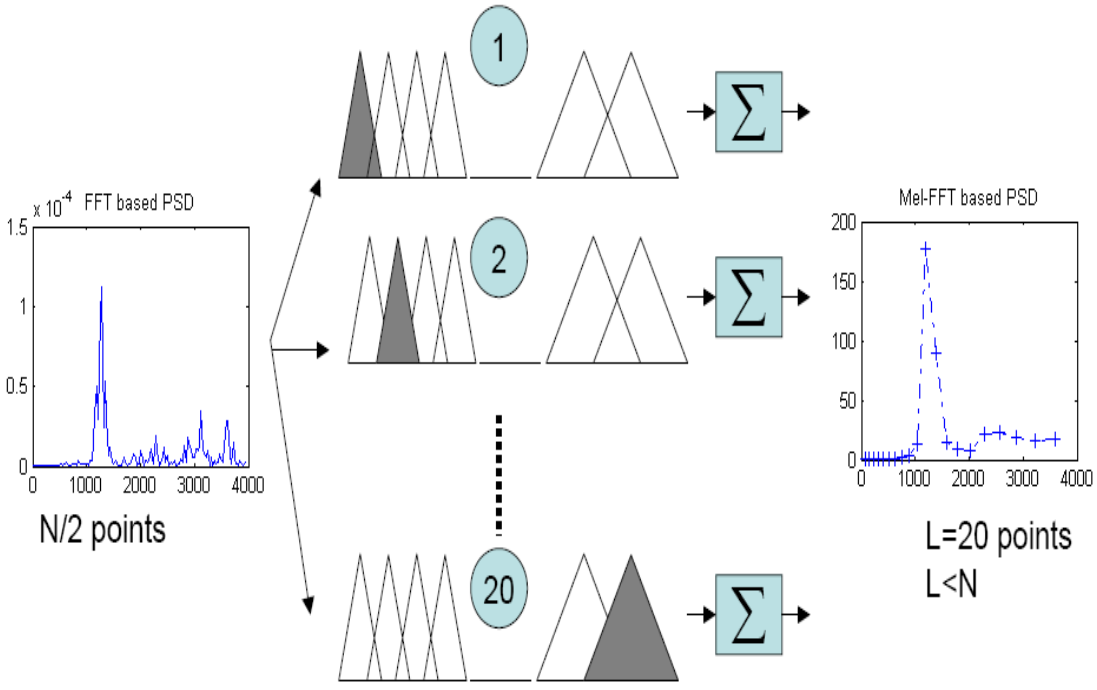

Figure 2.21 Result after use Mel frequency scale call "*Mel Spectrum*"

Finally, the output from Mel-Filter Bank is array of summation value from each Mel weighting filters.

2.14 Related Research

2.14.1 A Thai Speech Recognition

Seri Pansang [21] proposed to apply an artificial neural network with back propagation method. Main objective of this research was focused to recognition in 20 of Thai isolated words with male and age between 20-45 years old only.

The first step of this method was recorded to male isolated word into the cassette tape and using A/D converter with 8Kh, 8bit, 1 channel (mono sound) into computer. Then each isolated word must be transfer in frequency domain and pass to pre-processing method to find the "Voice Characteristic Correlative Relation". Then the "Voice Characteristic Correlative Relation" sent to train with back propagation neural network (BPNN) that set final condition by selected one between "average square error" not over 0.05 or completed with 5000 loop of neural network train loop and the result of this research was shown in table 2.1.

Table 2.1 Accuracy of Back propagation neural network recognition

	The result of accuracy recognition rate																			
Isolated	0	1	2	3	4	5	6	7	8	9	10	open	cloae	rotate	lift	lay	gap	left	right	light	%
words	ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า	สิบ	เปิด	ปิด	หมุน	ยก	วาง	ช่อง	ซ้าย	ขวา	ไฟ	accuracy
ศูนย์ (0)	9													2	1						75
หนึ่ง (1)		9				1								1		1					75
สอง (2)			5						1	1							1		4		41.67
สาม (3)	1			10			1														83.33
สี่ (4)		1			5								4	1						1	41.67
ห้า (5)				4		5				1						2					41.67
หก (6)							9					1					1			1	75
เจ็ด (7)							1	9	1											1	75
แปด (8)						2	2		7							1					58.33
เก้า (9)									1	11											91.67
สิบ (10)	1				1				1		6		3								50
เปิด (open)		1					2					8				1					66.67
ปิด (close)		1						1					10								83.33
หมุน (rotate)		1												10		1					83.33
ยก (lift)					1										9		1		1		75
วาง (lay)							1		1							9		1			75
ช่อง (gap)																	11			1	91.67
ซ้าย (left)							2											10			83.33
ขวา (right)																1			11		91.67
ไฟ (light)					1	1	1													9	75
																				AVG	*71.67*
																				MAX	*91.67*
																				MIN	*41.67*

CHAPTER 3
METHODOLOGY AND MATERIALS

3.1 Research Methodology and Procedure

This study will apply the prototype technique base on System Development Life Cycle (SDLC) to develop a prototype of the Thai speech recognition: case study basic voice Detecting.

The methodology of this concept is follows as:

3.1.1 Preliminary investigation and feasibility study.

3.1.2 Analyze data and system requirement.

3.1.3 Design of systems.

3.1.4 Development of systems.

3.1.5 Integrated and testing the systems.

3.1.6 Evaluation the efficient of systems

3.1.7 Conclude the result

3.1.1 Preliminary investigation and feasibility study.

At present, Thai speech recognition is popular among researchers in Thailand. Some research organizations such as NECTEC maintain the updated publications, voice data, and prototype founded in this investigate. Most data is provided for basic study.

Addition sources are from the Master's thesis from other universities or electronic publications such as IEEE that focus on speech recognition. This investigation shows that there are several potential methods that can be used in the implementation for Thai speech.

From the above, there are two states in this investigation. For the reviewing state, potential methods from various publications are reviewed. The other state is to select a suitable method for the implementation of Thai speech recognition and Thai voice detection.

The final state is to study the theory of

- Fundamentals of speech recognition
- Fundamentals of digital signal processing
- Fundamentals of feature extraction and Vector representation
- Fundamentals of feature analysis
- Speech recognition based feature analysis
- Similarity measurement
- Evaluation result

3.1.2 Analyze the system requirement.

This step is to determine the requirements of the new system. The prototype of Thai speech recognition system needs to refine its requirement. It will identify the inputs, outputs, digital signal processing method, and combination of graphic user interface to complete the functional requirements and qualitative requirements.

The new system is required the following modules,

3.1.2.1 Input module

- Algorithm to connecting between software and microphone
- Algorithm to convert sound Analog to sound Digital (A/D)
- Algorithm to record sound in digital format to storage device

3.1.2.2 Processing module

- Algorithm to cutting suitable length of sound sample
- Algorithm to extract speech feature
- Algorithm to display speech feature for check the correct result

- Algorithm to record vector format that represent speech feature
- Algorithm to analysis feature
- Algorithm to record values of analysis feature result

3.1.2.3 Output module

- Algorithm to measure the similarity of speech feature
- Algorithm to identify the speech input
- Algorithm to display to final identify of speech detecting

3.1.3 Design of systems.

Designing the new system by follow as:

- Use JAVA programming to create application for interface and result monitoring
- Select appropriate and layout Graphic User Interface (GUI) control.
- Identify and divide type of each input or output data into application zone.
- Use the same format in button, text box and other common.
- The sampling, training, monitoring and vector are use in numerical format of Java model.

3.1.3.1 Input data

- Create form to start-capture stop-capture playback and save the input speech for classify and training
- Create form for entry name of Input speech when capture finish

3.1.3.2 Process data

- Create form for select speech to automatic cutting, automatic indexing, and automatic save
- Create form for select speech to run method windowing, Pre-emphasis, Discrete Fourier Transform (DFT) and Filter Bank
- Create form to display Discrete Fourier Transform (DFT) for developer to check result

3.1.3.2 Output data

- Create form for select Unknown sound to identify
- Create form for display identify result

3.1.4 Development of systems.

All analyzed and designed will used for developing the system following as:

- Create system function follow analysis phase.
- Create, page layout by follow designed phase with GUI of JAVA Application.
- Coding in each parts with JAVA code and text database.

3.1.5 Integrated and testing the systems

For stability of system the tested must be process follow as: Unit test is going performed on isolation module.

- Each system function tested.
- Integrated test will test on multiple modules that related.
- System application connected with text database test.
- System test is done to ensure the system in properly running.
- Sampling featuring and training for speech input management.
- Speech detecting and identifying for output.

3.1.6 Evaluation the efficient of systems

Evaluation of system, performed to identify its productivity. The assessments of system function following dimension: Main Operational

- percent of accurate identify unknown speech
- Other Operational :
- properly calculate value of DSP and pattern recognition formula
- response time when running automatic process a large number of speech

- stability when running automatic process a large number of speech

3.1.7 Conclude the result

Conclude the researched result if correctly in line with the objective, present the idea in order to better the system and issue the research document.

3.2 Research Equipment and tools

Hardware

CPU	: Mobile AMD Turion64 MT30 1600 MHz
RAM	: 512 MB
Hard Disk	: At least 5 GB
Monitor	: WXGA TFT LCD
SoundCard	: On borad SiS 7012 Audio Device
Peripheral	: Microphone, Mouse, Keyboard, CD-ROM, Printer

Software

Operating System	: Microsoft windows XP Service Pack 2
Application Tool	: Microsoft Notepad, Editplus2
Application development	: JAVA Second Edition Version1.4.1_02
Database	: Text file
Drawing	: Microsoft Visio2003
Documentation	: Microsoft Word 2003

3.3 Research Time

Table3.1 Research Schedule

Activity	Time											
	2005				2006				2007			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1.Preliminary Investigation												
2.Determination of System Requirement												
3.Design of System												
4.Development of System												
5.System Testing												
6.Impremention and Evaluation												
7.Report Writing Up												

CHAPTER 4

RESULT

This chapter presents the algorithm developed for a prototype of Thai speech recognition system. This chapter is divided into 4 parts: an overview of the proposed system, the input preprocessing, the feature extraction, feature modeling and the feature matching.

4.1 An overview of the proposed system

The overview can show by below

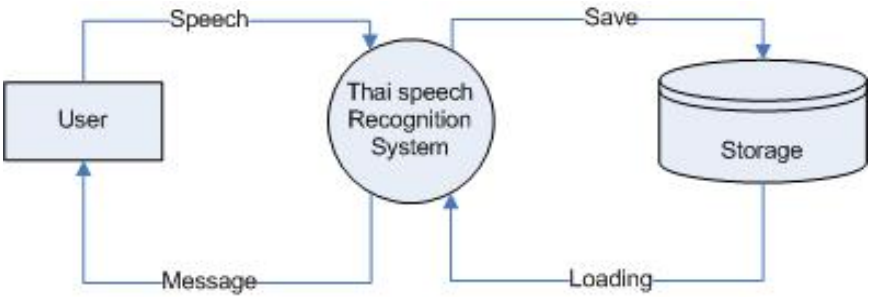

Figure 4.1 Context diagram of Thai speech recognition system

As Shown in figure 4.1 the context diagram user must pronoun the speech or commanding using isolated words via microphone to communicate with a prototype of Thai speech recognition system.

4.1.1 An overall of the speech model preparation system

The Architecture of main process to prepared the vector model for use to compare and identify of unknown speech can shown by diagram in Figure 4.2 and Figure 4.3

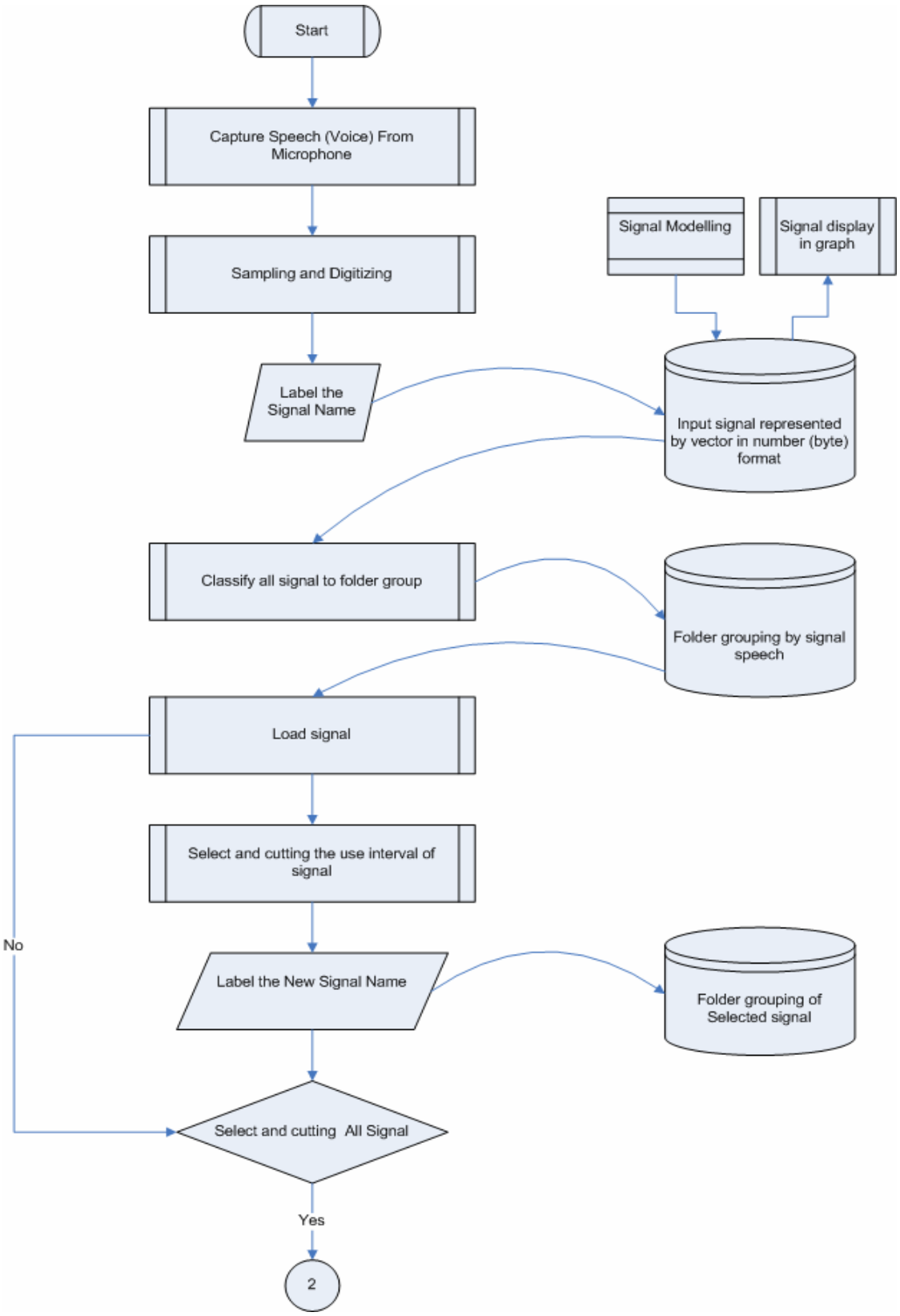

Figure 4.2 Architecture of Thai speech Recognition systems.

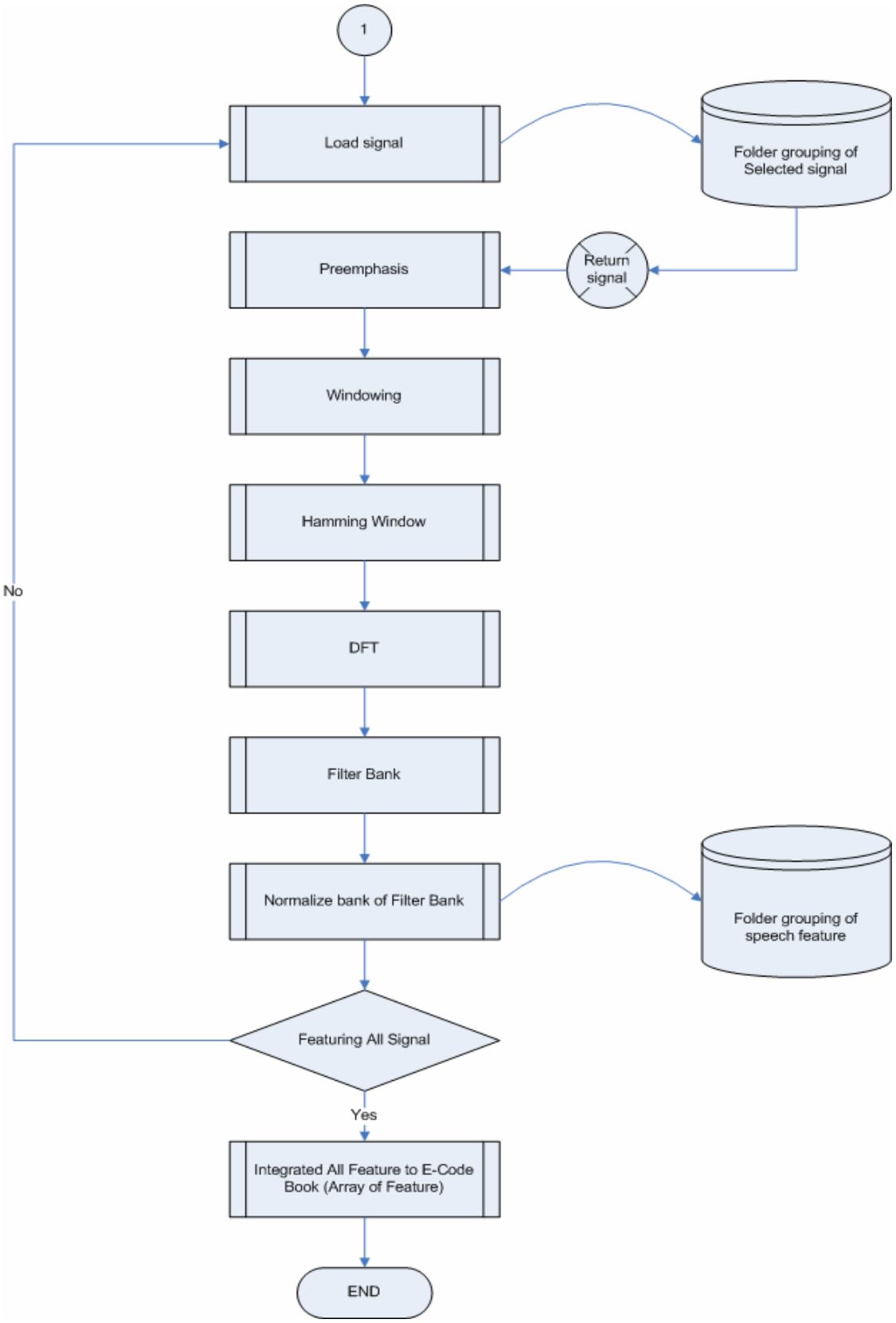

Figure 4.3 Architecture of Thai speech Recognition systems (cont).

4.2 The Input Preprocessing

4.2.1 The Input and Digitization of Speech signal

The input to system must use the I/O stream of programming to connect between microphone and the system. After that the human prototype subject must pronounce the several time with same isolated word, repeated, via microphone to create the system reference of signal speech model or system signal speech code book. The speech signal must be captured in time series on a single file per series per human subject. Simultaneously, this method is also sampling and digitizing the analog signal, the continuous signal, from microphone to digital signal format. The sampling and digitizing using:

Sampling rate: 16000 bit per second (bps)

(Using the large bit rate per second for focus in sound detail provide large scale for the future use if require),

Sampling size signed 8 bit,

Channel one channel (mono).

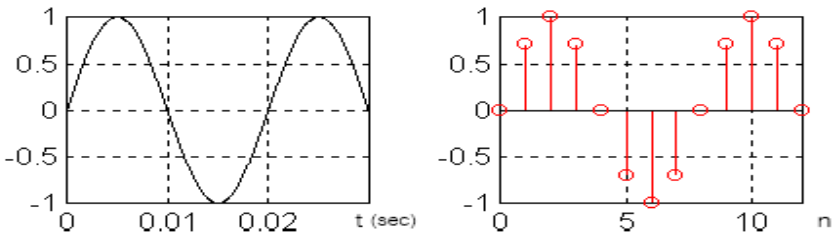

Figure 4.4 Example to show discrete point after digitization the signal [6]

After recorded isolated words and digitization successes each file of speech signal must be label to providing for the next process. The scope of this step can be as process in Figure 4.5

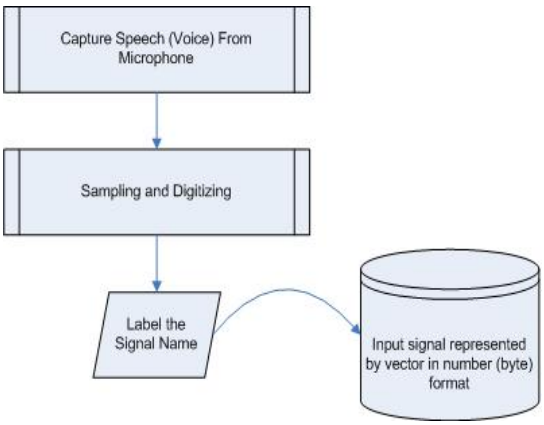

Figure 4.5 Process of input section in Thai speech recognition system

This section of the Thai speech recognition system is use the main module called “Audio Capture and A/D Conversion” to get the speech signal that can use to Capture, Play and Save. Step to use this module is:

1. Click “Record” button.

The system will start capture every signal via microphone. The system will open java stream connect java with microphone and keep signal in stream with all audio signal formal assignment like sampling rate, sampling size, number of channel.

2. Click “Stop” button.

The system will get signal from stream, disconnect stream and convert data for storage into byte format.

3. Click “Play” button (Optional, not a main process).

To playback speech signal for check correct of pronounce from subject by use the data that keep form java I/O stream and convert to audio format.

4. Click “Save” button. To Input name for label speech signal and output will shown in Figure 4.7.

Figure 4.6 Main modules in system to capture input speech that develop follow plan in Figure 4.5

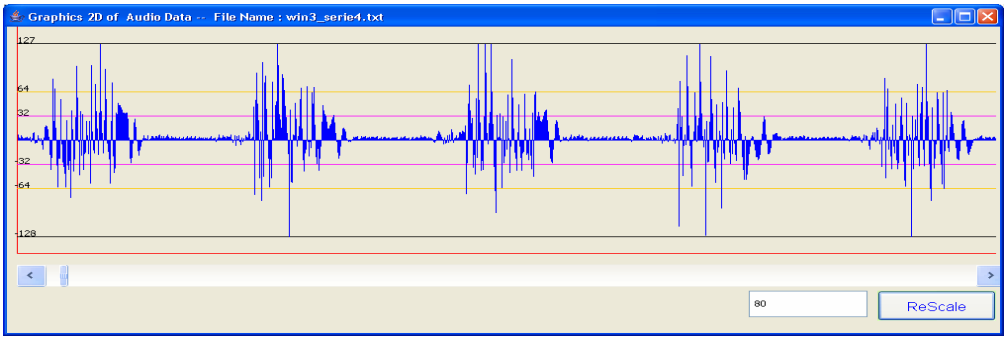

Figure 4.7 Series of the signal that repeated five-time with the same isolated word of one human subject and captured in one file.

4.2.2 The Selected and Preparation of Input signal

4.2.2.1 Digital speech selected using windows energy based

Repeating five-time of isolated word in one file much be selected and extracting into only one of isolated word and recorded again in another file that each file must contain only one of signal of isolated word by using windows energy based analysis.

The windows of energy using window width 400 sample sizes and slide window from start to end using shift between windows width 200 sample sizes, it good sizes performance by experimental . The windows energy based can calculate the energy in interval of energy window to identify the start and stop point of only one isolated word pronouncing.

Now, the loop of isolated words must be extracting into many file by filter from energy of speech interval already. Finally the selected digital isolated word signal must be label to provide for the next process**.**

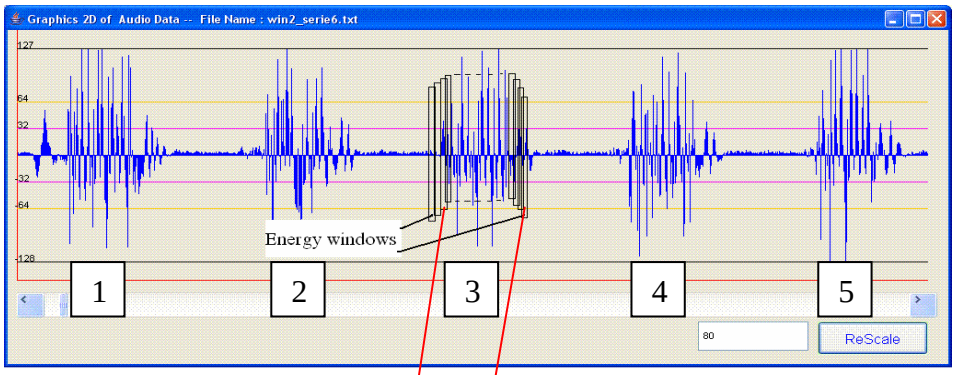

Figure 4.8 Speech signal separated by *windows energy based analysis* method

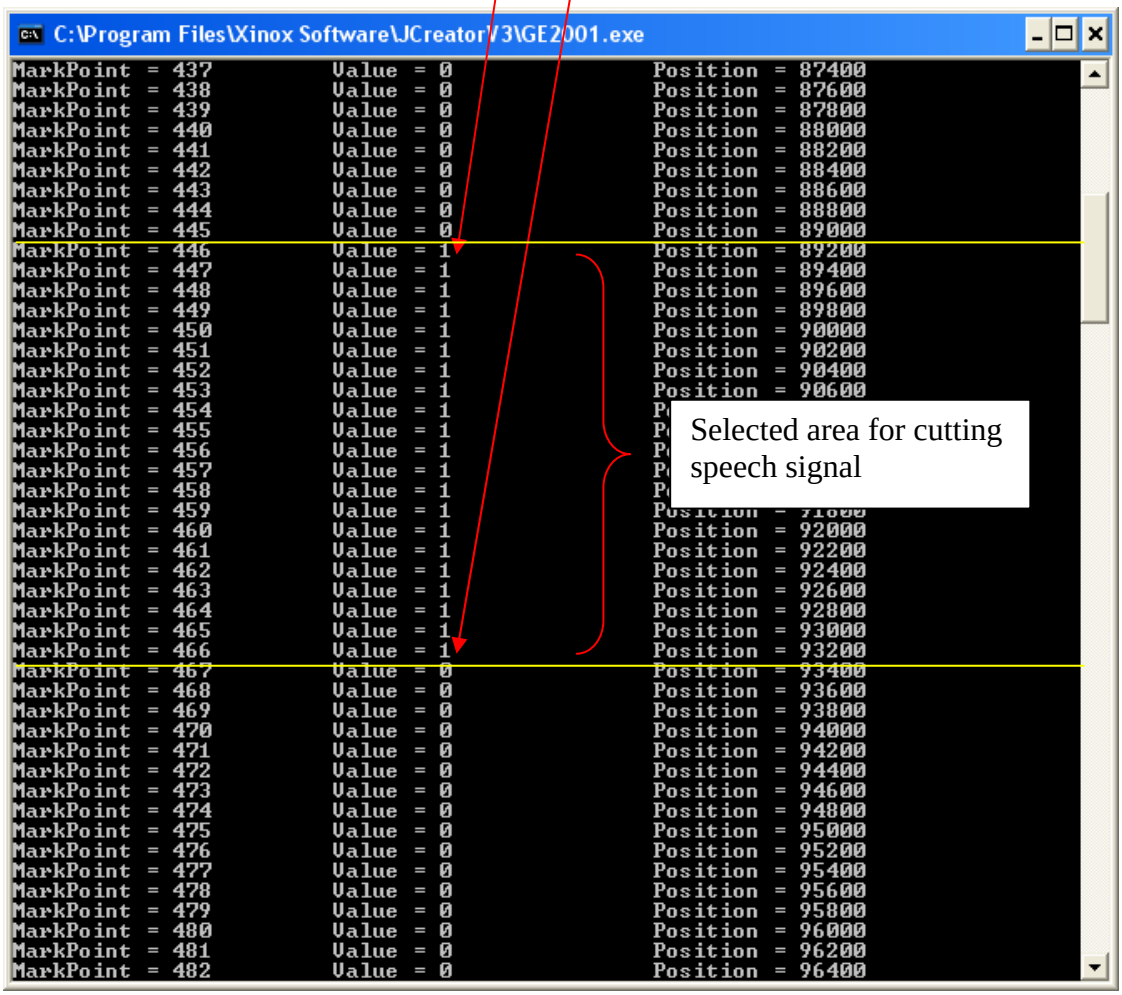

Figure 4.9 Function and mark point of windows energy based analysis

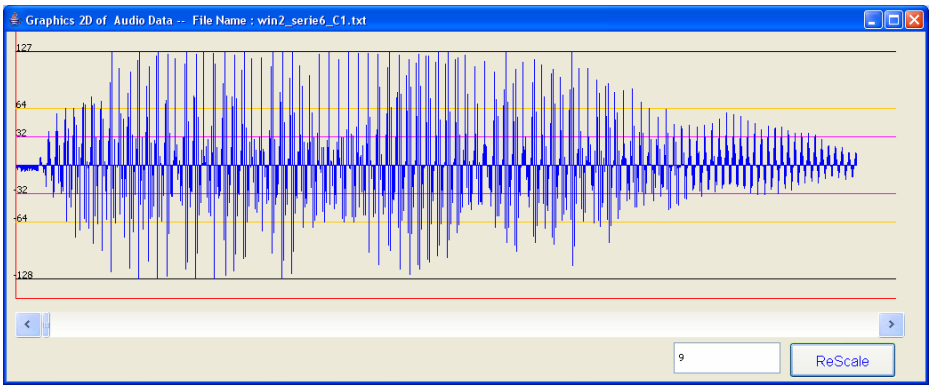

Figure 4.10 Speech signal of isolated word “2” number 1 from 5 that extract from looping speech captured in Figure 4.7

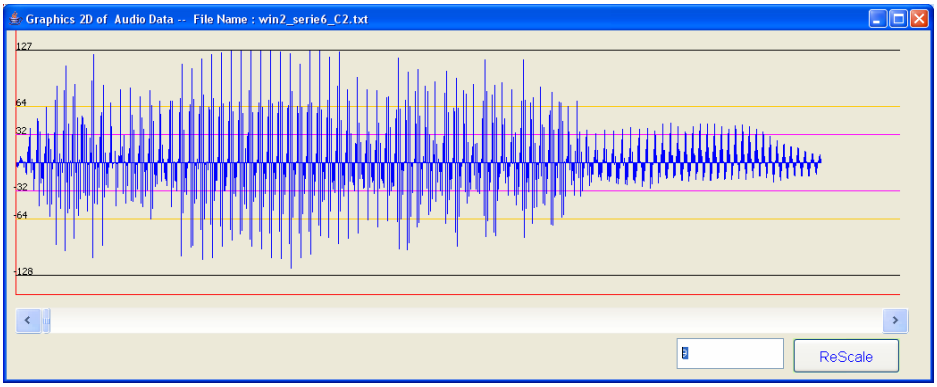

Figure 4.11 Speech signal of isolated word “2” number 2 from 5 that extract from looping speech captured in Figure 4.7

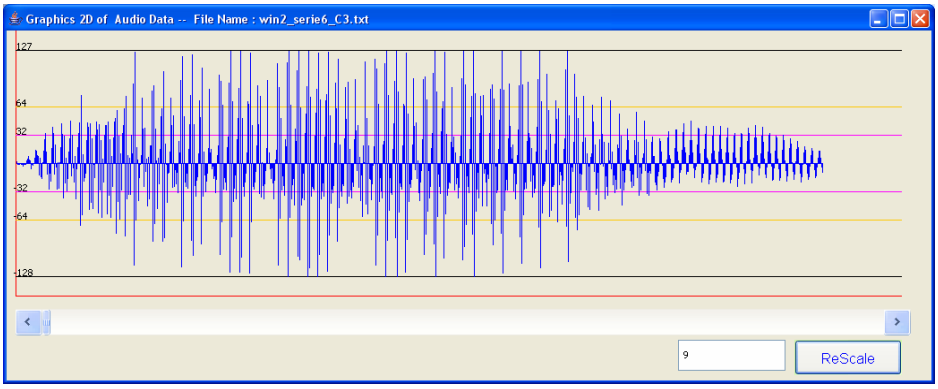

Figure 4.12 Speech signal of isolated word “2” number 3 from 5 that extract from looping speech captured in Figure 4.7

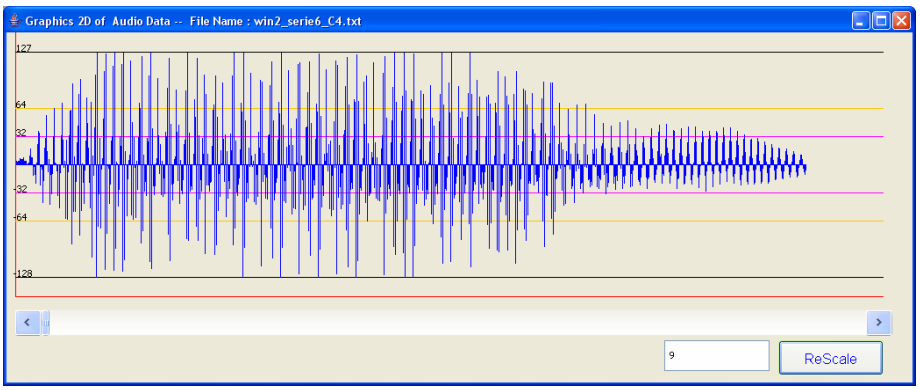

Figure 4.13 Speech signal of isolated word “2” number 4 from 5 that extract from looping speech captured in Figure 4.7

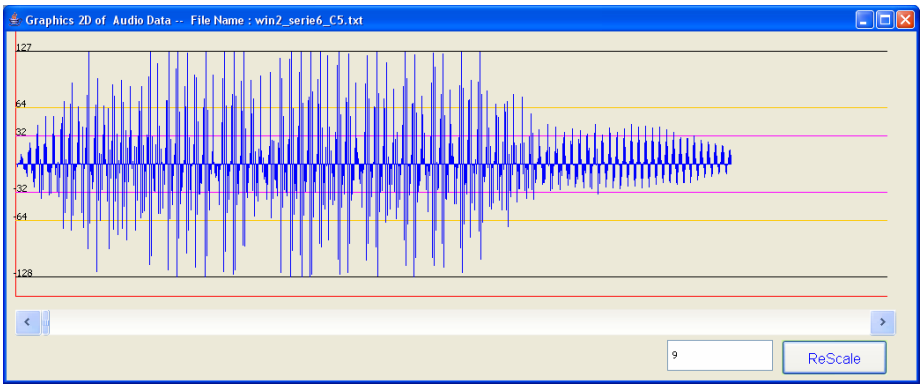

Figure 4.14 Speech signal of isolated word “2” number 5 from 5 that extract from looping speech captured in Figure 4.7

From above (described in section 4.2.2) its show the detail in section of the “selected and preparation of input signal” that shown in Figure 4.16

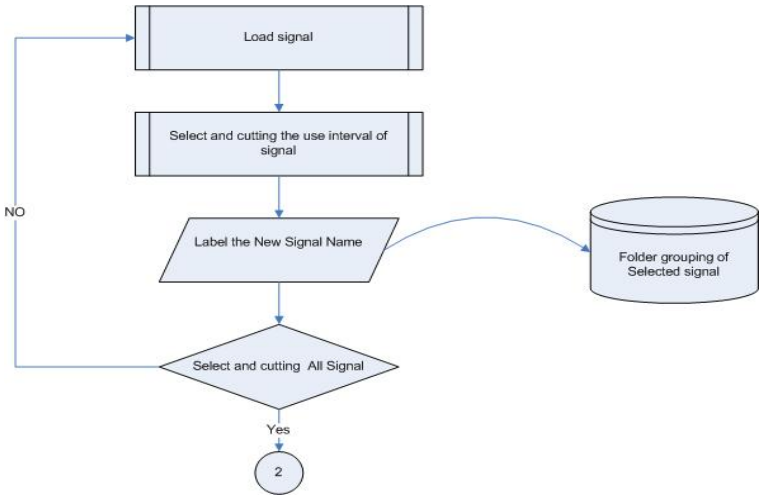

Figure 4.15 Process in section of the selected and Preparation of Input signal

4.3 The Feature Extraction and Feature Modeling

4.3.1 Pre-emphasis and Windowing.

As the shown from above in Figure4.11-Figure4.15, after to completed signal preparation.

The signal of speech would be pre-emphasis to get sharper peaks. A being the pre-emphasis parameter in the time domain, the pre-emphasized signal is related to the selected digital signal by the relation:

$$X'(n) = X(n) - 0.95\,X(n-1) \quad \text{Where } n = 1,\ldots., N-1$$

After that the pre-emphasized signal of speech would be separate into many windows, *The Windowing*, depend on length of signal and fix window width in 800 sample size (50 ms from 16000 bit per second) and must be shift with length 160 sample size (10 ms from 16000 bit per second)

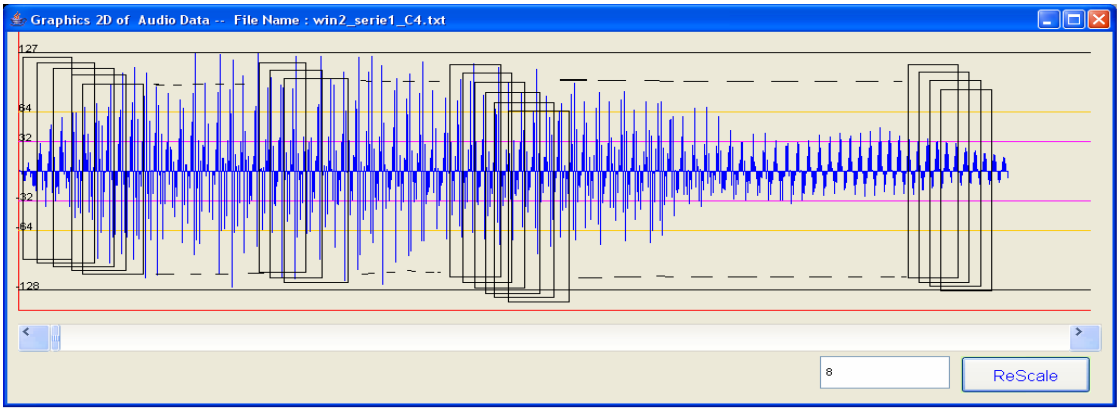

Figure 4.16 Performed pre-emphasize signal and windowing method

Each window or frame from window function, *Windowing*, will process with *Hamming window*. In Speech Recognition the most-used window shape is the *Hamming window*, whose impulse response is a raised cosine impulse:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & n = 0,\ldots., N-1 \\ 0 \quad \text{otherwise} \end{cases}$$

4.3.2 The Feature Perform Using DFT Method

The DFT method must process after completing the *Hamming window* in each frame. The DFT method use to transform each frame from time-domain to frequency-domain. The frequency-domain is the feature to use for speech signal analysis.

The feature perform using DFT method

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

In each frame value of each the frequency, spectral, must have a little change from time to time and use to analysis called "*spectral analysis*". This can show the DFT result for some of frame in below figure.

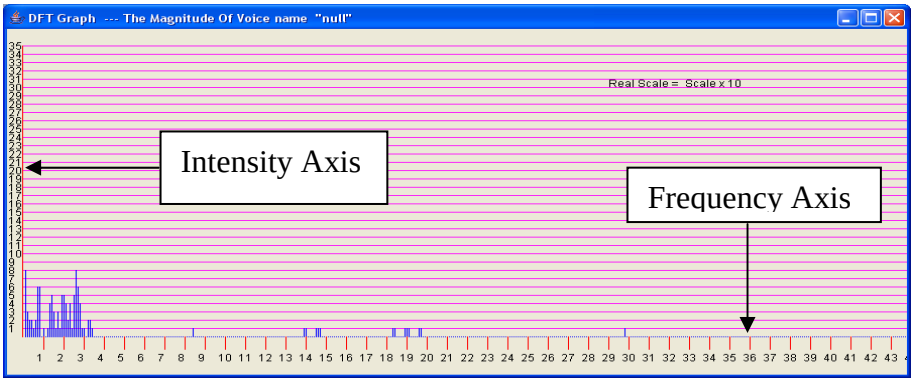

Figure 4.17 The frequency domain of frame number 6 of speech "4"

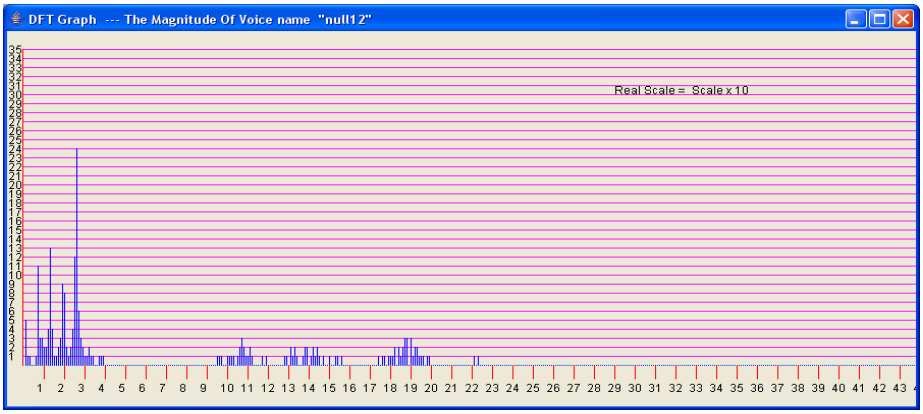

Figure 4.18 Shown a little change from frequency in frame number 12 to compare with frame number 6 (Figure 4.17) of speech "4"

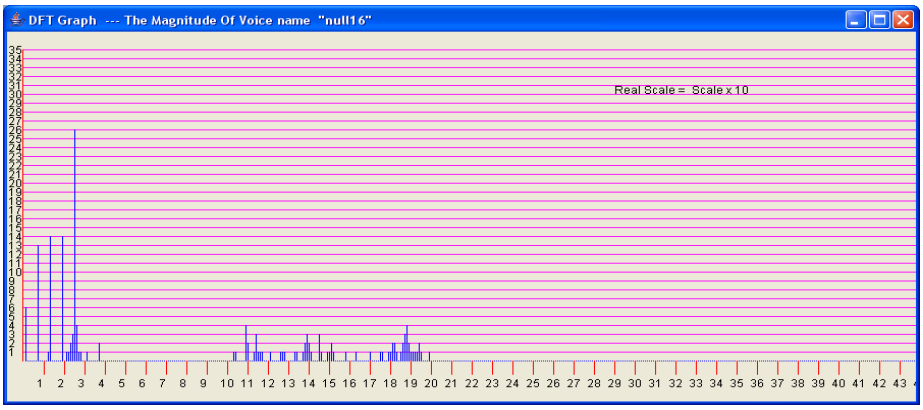

Figure 4.19 Shown a little change from frequency in frame number 16 to compare with frame number 12 (Figure 4.18) of speech “4”

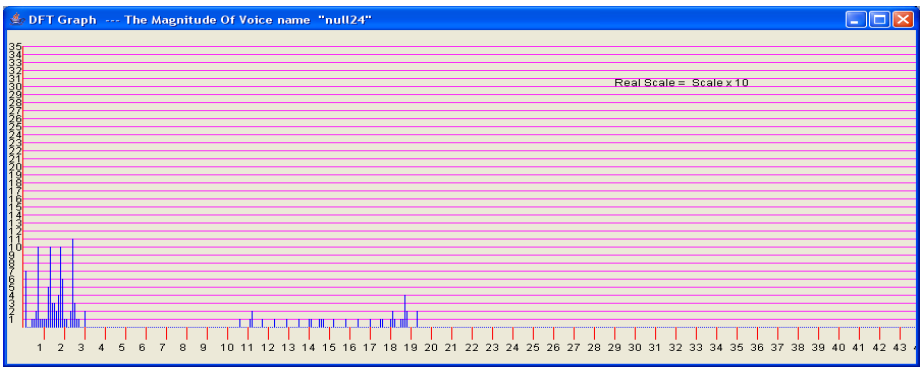

Figure 4.20 Shown a little change from frequency in frame number 24 to compare with frame number 26 (Figure 4.19) of speech “4”

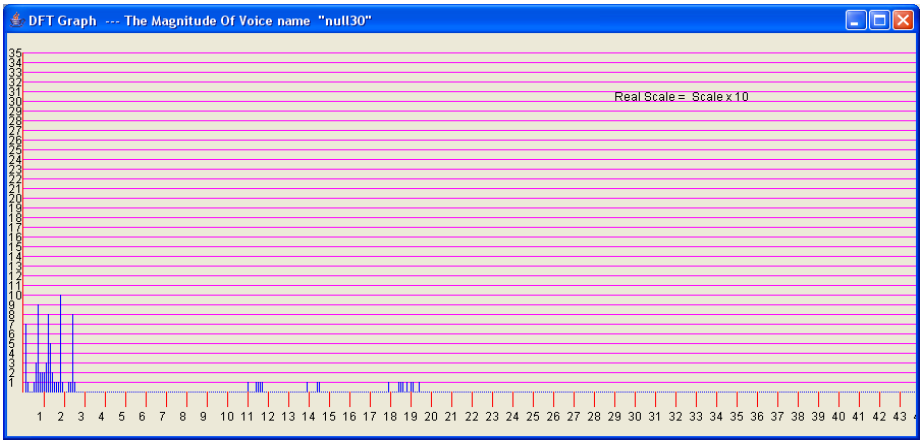

Figure 4.21 Shown a little change from frequency in frame number 30 to compare with frame number 24 (Figure 4.20) of speech “4”

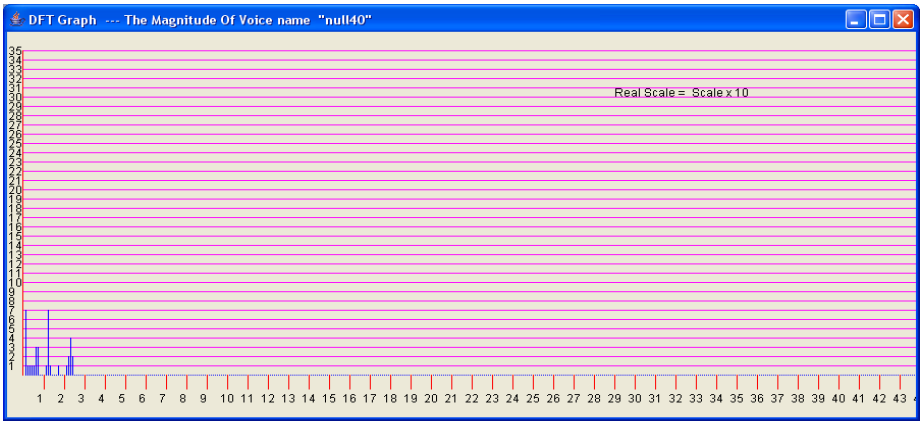

Figure 4.22 Shown a little change from frequency in frame number 40 to compare with frame number 30 (Figure 4.21) of speech “4”

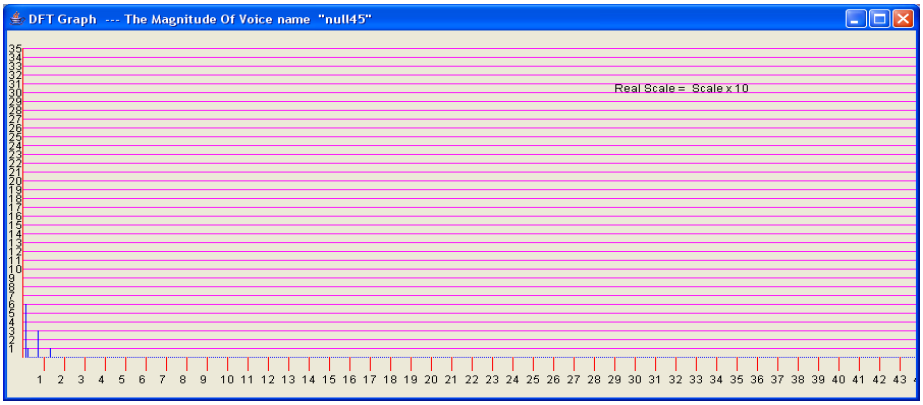

Figure 4.23 Shown a little change from frequency in frame number 45 to compare with frame number 40 (Figure 4.22) of speech “4”

4.3.3 The feature perform using filter bank method

After completed to obtain the frequency domain in each frame of speech signal from DFT method then the system needed to compress or concise the feature for more efficiency to calculate or compare the speech feature, one way to more concisely characterize the signal frequency is by a filter bank. The system can divide the frequency range of interest (range 100-8000 Hz) into N band and measure the overall intensity in each band. This could be done using the incoming frequency be computed from spectral analysis. In a uniform filter bank, each frequency band is

of equal size. For instance if used 8 ranges, the bands might cover the frequency ranges:

100Hz-1000Hz, 1000Hz-2000Hz, 2000Hz-3000Hz, …, 7000Hz-8000Hz

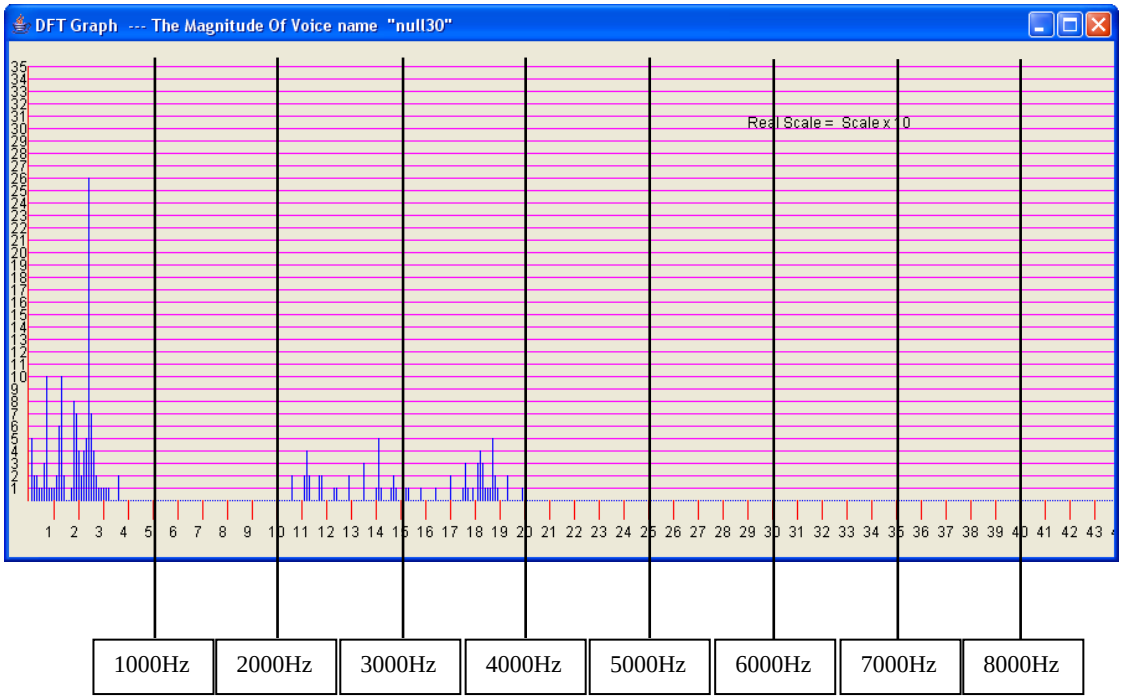

Figure 4.24 Uniform Filter bank on spectra for speech “4”

Consider a uniform filter bank representation of artificially generated spectra similar to that for the speech “4” shown in figure 4.25. With discrete set of sample points, System can measure the intensity in each band by computing the summing value of all values or “power” measure by summing the between the lines of bands.

The system would get a representation of the spectra that consists of a vector of eight numbers in this case, approximate vector is (127, 0, 31, 41, 0, 0, 0, 0)

It’s not good represented vector by uniform filter bank. A better alternative is to organize the ranges using Mel scale this method is to design a non-uniform set of frequency bands that has no simple mathematical characterization but better reflects the response of the ear as determined from experimentation. One very common design is based perceptual studies to define critical bands in linear up to 1000Hz and logarithmic after that. For instance, system might follow to use the ranges start like Mel scale but adjust a little like the table 4.1.

Table 4.1 Mel scale and System using Mel adjusted scale

Index	Mel Scale		System Mel Adjusted Scale	
	Center frequency (Hz)	Band-Width (Hz)	Center frequency (Hz)	Band-Width (Hz)
1	100	100	80	80
2	200	100	160	160
3	300	100	240	160
4	400	100	320	160
5	500	100	400	160
6	600	100	480	160
7	700	100	560	160
8	800	100	640	160
9	900	100	720	160
10	1000	124	800	160
11	1149	160	880	160
12	1320	184	960	160
13	1516	211	1024	160
14	1741	242	1120	160
15	2000	278	1200	160
16	2297	320	1280	160
17	2639	367	1360	160
18	3031	422	1440	160
19	3482	484	1520	160
20	4000	556	1700	400
21	4595	639	1900	400
22	5278	734	2200	400
23	6063	843	2400	400
24	6964	969	2800	800
25	-	-	>3200	

Now, after to use filter bank method the outcome is vector of 25 numbers and amount of vector must equal of frame number. The next step to concise again by use the uniform filter bank that divide in 5 bands. In this step not use summing method in each band but use average the vector and normalization to percent of maximum value in completed average vector of each band.

Figure 4.25 Result normalized vector after use average vector by uniform filter bank

Finally, the final vector is vector of 125 numbers that representation for speech signal

Figure 4.26 Final vectors that representation for speech signal of speech “3”)

4.3.4 Time to use for Feature Extraction

The feature extraction in this research that performed by using DFT and filter bank method must take time for running in each process. The time consuming can shown by table 4.2 and table 4.3.

Table 4.2 Extraction time (DFT with Filter bank) for female speech

(English)	(Thai)	Min	Max	Average	Number of	Total time
Isolated word		(sec)	(sec)	(sec)	isolate words	(Min)
0	ศูนย์	3.7	6.3	4.82	283	22.73
1	หนึ่ง	2.7	3.8	3.28	262	14.32
2	สอง	3.1	6.4	4.41	255	18.74
3	สาม	3.7	6.7	4.54	256	19.37
4	สี่	2.4	6.8	4.58	229	17.48
5	ห้า	3.2	5.2	4.01	251	16.78
6	หก	1.4	3.5	2.36	261	10.27
7	เจ็ด	1.6	3.2	2.23	250	9.29
8	แปด	2.9	5.7	4.15	271	18.74
9	เก้า	3.5	5.8	4.03	249	16.72
10	สิบ	1.3	2.2	1.77	307	9.06
Close	ปิด	1.4	2.6	1.71	323	9.21
Delete	ลบ	2.1	3.4	2.68	261	11.66
Down	ลง	2.4	3.6	2.74	243	11.10
Exit	ออก	2.5	5.2	4.36	271	19.69
Left	ซ้าย	4.2	6.1	5.54	243	22.44
Open	เปิด	3.8	5.9	4.66	256	19.88
Right	ขวา	4.1	7.6	5.57	247	22.93
Save	บันทึก	3.6	6.8	4.98	256	21.25
Space	วรรค	1.8	3.5	2.66	258	11.44
Up	ขึ้น	1.9	5.4	3.13	249	12.99
					Total	336.09

Table 4.3 Extraction time (DFT with Filter bank) for male speech

(English) Isolated word	(Thai)	Min (sec)	Max (sec)	Average (sec)	Number of isolate words	Total time (Min)
0	ศูนย์	2.9	6.0	4.52	208	15.67
1	หนึ่ง	2.2	4.8	3.47	233	13.48
2	สอง	3.3	6.1	4.67	207	16.11
3	สาม	3.2	6.0	4.62	215	16.56
4	สี่	3.0	5.9	4.45	210	15.58
5	ห้า	2.9	5.1	3.89	217	14.07
6	หก	1.9	3.3	2.29	219	8.36
7	เจ็ด	1.8	2.5	1.97	214	7.03
8	แปด	2.6	4.9	3.59	221	13.22
9	เก้า	3.2	5.7	4.20	209	14.63
10	สิบ	1.0	3.0	2.02	290	9.76
Close	ปิด	1.2	2.8	1.77	316	9.32
Delete	ลบ	1.9	3.1	2.43	200	8.10
Down	ลง	2.0	3.5	2.49	204	8.47
Exit	ออก	2.8	5.1	4.41	197	14.48
Left	ซ้าย	3.8	5.7	5.22	213	18.53
Open	เปิด	3.3	5.5	4.56	205	15.58
Right	ขวา	3.6	6.5	5.23	212	18.48
Save	บันทึก	3.5	6.7	4.77	220	17.49
Space	วรรค	1.3	3.1	2.44	203	8.26
Up	ขึ้น	1.4	3.2	2.68	206	9.20
					Total	272.36

4.4 Matching Feature and Speech Identification

The final process and main target of this system is identification the unknown speech signal. This process must use Euclidian distance measure to compare between vector of unknown and vector of source. This step must compare distance of vector between unknown vectors to all source vectors to find the minimum distance. The nearly, minimum distance, vector must be the answer to identification of unknown speech that input to the system.

The overall of the process that used to identification can show in figure below.

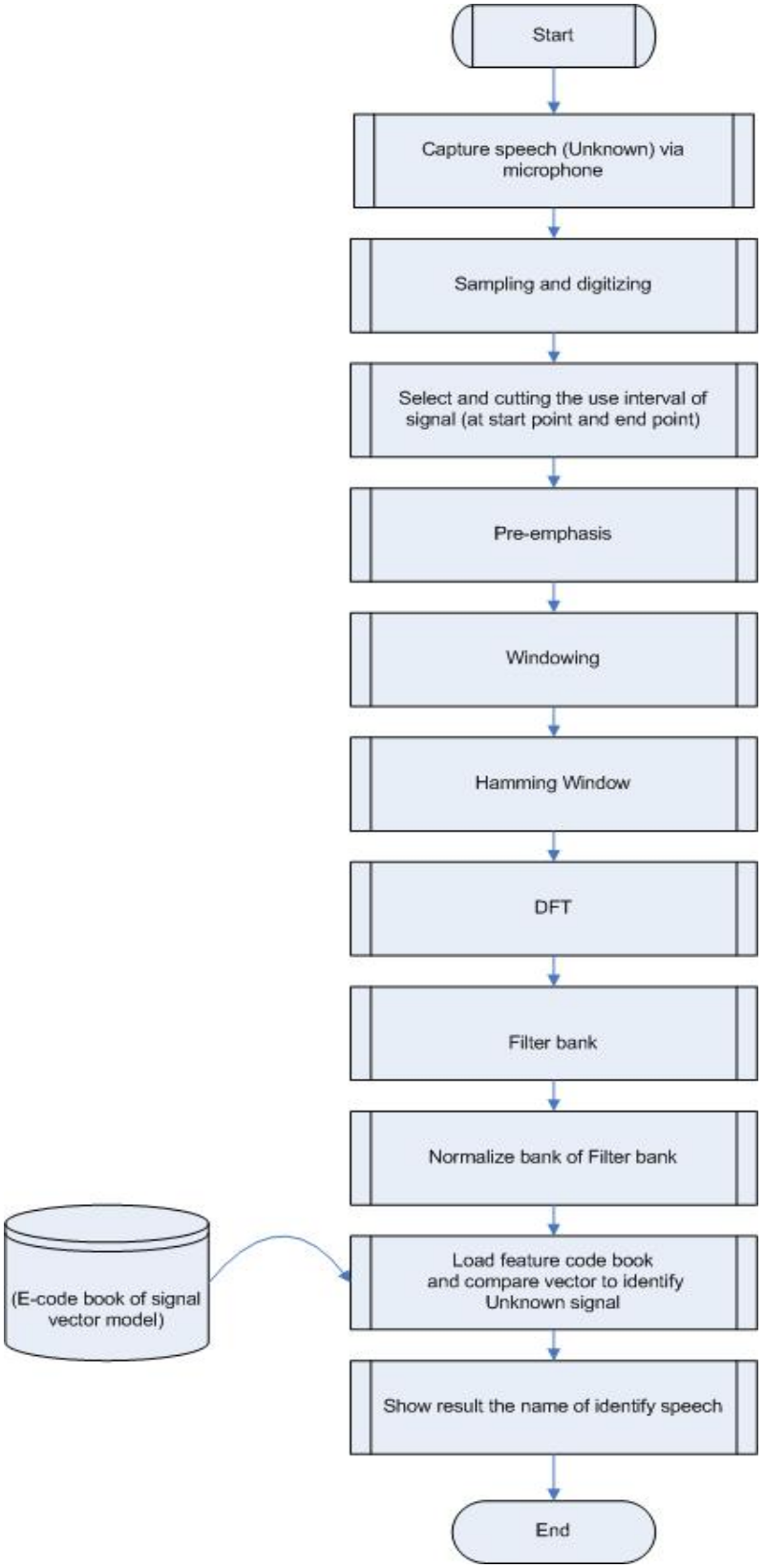

Figure 4.27 Overall the process to identification of unknown speech

4.5 The Result of Speech Identification

The result of this experimental can be show in matrix and graph to represent data for fast and easy to compare accuracy rate of each isolated word. The result will show follow as figure 4.29 and figure 4.30. And the full data detail must show in matrix table 4.4, table 4.5 and table 4.6.

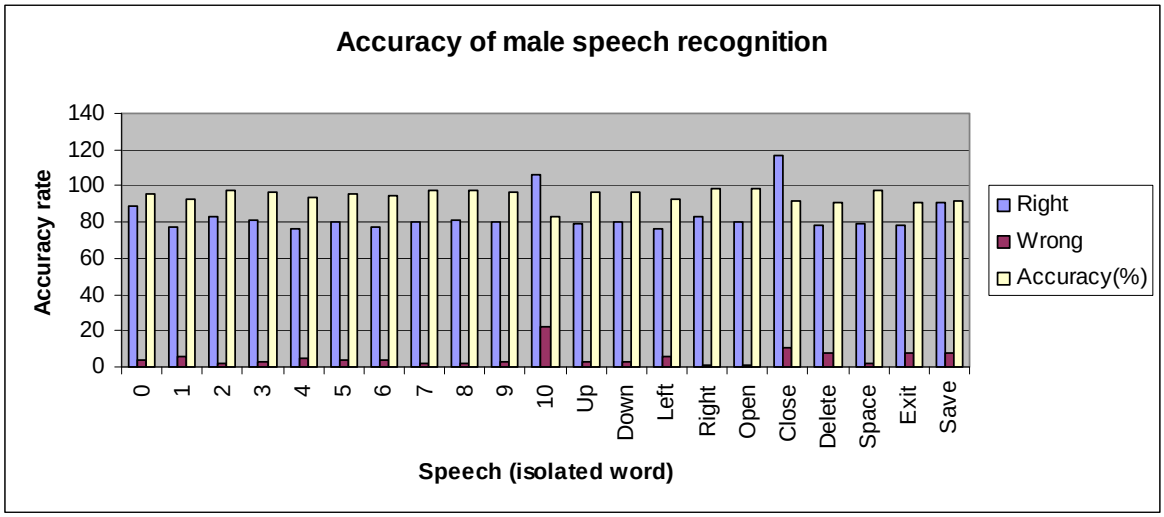

Figure 4.28 Shown the male speech recognition rate

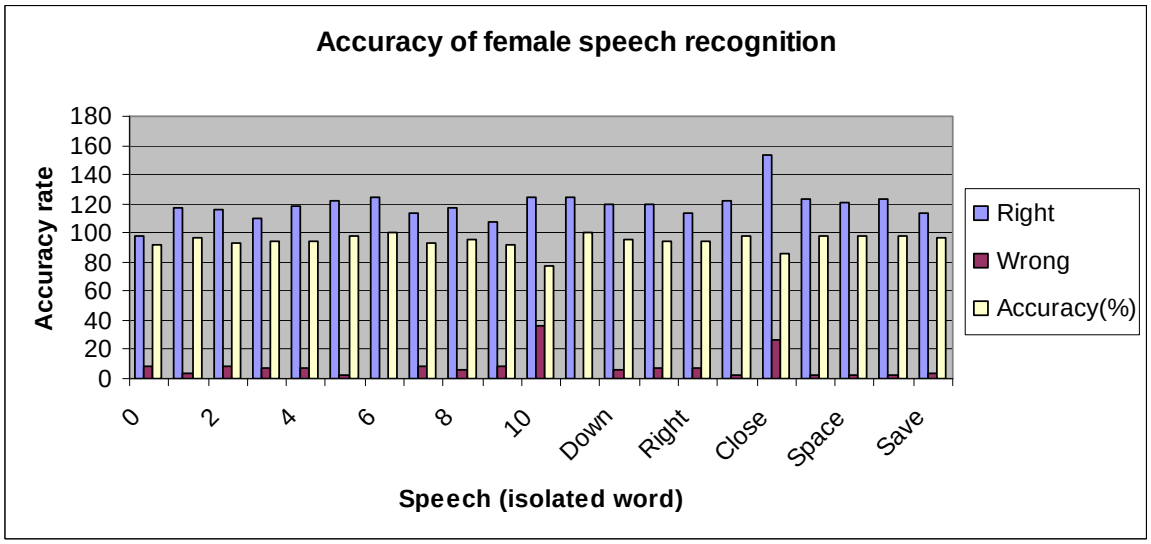

Figure 4.29 Shown the female speech recognition rate

Table 4.4 Male accuracy of propose system

Male Accuracy

All	right	wrong	%	English	Thai	0 ศูนย์	1 หนึ่ง	2 สอง	3 สาม	4 สี่	5 ห้า	6 หก	7 เจ็ด	8 แปด	9 เก้า	10 สิบ	Up ขึ้น	Down ลง	Left ซ้าย	Right ขวา	Open เปิด	Close ปิด	Delete ลบ	Space วรรค	Exit ออก	Save บันทึก
93	89	4	95.7	0	ศูนย์	89	3	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
83	77	6	92.8	1	หนึ่ง	3	77	-	-	-	-	-	1	-	-	-	1	-	-	-	-	1	-	-	-	-
85	83	2	97.6	2	สอง	-	-	83	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-
84	81	3	96.4	3	สาม	-	-	1	81	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-
81	76	5	93.8	4	สี่	-	1	-	-	76	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	2
84	80	4	95.2	5	ห้า	-	-	-	-	-	80	-	-	-	-	-	-	-	2	2	-	-	-	-	-	-
81	77	4	95.1	6	หก	-	-	-	-	-	-	77	-	-	-	-	-	3	-	-	-	-	1	-	-	-
82	80	2	97.6	7	เจ็ด	-	-	-	-	-	-	1	80	-	-	1	-	-	-	-	-	-	-	-	-	-
83	81	2	97.6	8	แปด	-	-	-	-	-	-	-	-	81	-	-	-	-	-	2	-	-	-	-	-	-
83	80	3	96.4	9	เก้า	-	-	-	2	-	-	-	-	1	80	-	-	-	-	-	-	-	-	-	-	-
128	106	22	82.8	10	สิบ	-	-	-	-	3	-	-	2	-	-	106	2	-	-	-	-	15	-	-	-	-
82	79	3	96.3	Up	ขึ้น	1	-	-	-	-	-	-	-	-	-	1	79	-	-	-	-	1	-	-	-	-
83	80	3	96.4	Down	ลง	-	-	-	-	-	-	-	-	-	-	-	-	80	-	-	-	-	3	-	-	-
82	76	6	92.7	Left	ซ้าย	-	-	-	4	-	2	-	-	-	-	-	-	-	76	-	-	-	-	-	-	-
84	83	1	98.8	Right	ขวา	-	-	-	-	-	1	-	-	-	-	-	-	-	-	83	-	-	-	-	-	-
81	80	1	98.8	Open	เปิด	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	80	-	-	-	-	-
128	117	11	91.4	Close	ปิด	-	1	-	-	-	-	-	1	-	-	7	1	-	-	-	-	117	-	-	-	-
86	78	8	90.7	Delete	ลบ	-	-	-	-	-	-	-	-	-	-	-	-	8	-	-	-	-	78	-	-	-
81	79	2	97.5	Space	วรรค	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	79	-	1
86	78	8	90.7	Exit	ออก	-	-	3	-	-	-	1	-	2	-	-	-	-	-	1	-	-	-	1	78	-
99	91	8	91.9	Save	บันทึก	-	-	-	-	1	-	2	-	-	-	2	1	-	-	-	2	-	-	-	-	91
	Average		94.6																							

Table 4.5 Female accuracy of propose system

Female accuracy

All	right	wrong	%	English		0	1	2	3	4	5	6	7	8	9	10	Up	Down	Left	Right	Open	Close	Delete	Space	Exit	Save
					Thai	ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า	สิบ	ขึ้น	ลง	ซ้าย	ขวา	เปิด	ปิด	ลบ	วรรค	ออก	บันทึก
107	98	9	91.6	0	ศูนย์	98	-	-	6	-	-	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-
121	117	4	96.7	1	หนึ่ง	-	117	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
124	116	8	93.5	2	สอง	-	-	116	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	7	-
117	110	7	94.0	3	สาม	-	-	-	110	-	-	-	-	1	-	-	-	-	5	1	-	-	-	-	-	-
125	118	7	94.4	4	สี่	1	-	-	-	118	-	-	-	-	-	-	-	-	1	-	-	5	-	-	-	-
124	122	2	98.4	5	ห้า	-	-	-	-	-	122	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-
125	125	0	100.0	6	หก	-	-	-	-	-	-	125	-	-	-	-	-	-	-	-	-	-	-	-	-	-
122	114	8	93.4	7	เจ็ด	-	-	-	-	-	-	4	114	3	-	-	-	-	-	-	-	1	-	-	-	-
123	117	6	95.1	8	แปด	-	-	2	-	-	-	-	-	117	-	-	-	-	-	-	-	-	-	-	4	-
117	108	9	92.3	9	เก้า	-	-	-	3	-	2	-	-	1	108	-	-	-	1	-	-	-	-	1	1	-
160	124	36	77.5	10	สิบ	-	-	1	-	2	-	-	2	-	-	124	-	-	1	-	-	29	-	-	-	1
124	124	0	100.0	Up	ขึ้น	-	-	-	-	-	-	-	-	-	-	-	124	-	-	-	-	-	-	-	-	-
125	119	6	95.2	Down	ลง	-	-	-	-	-	-	-	-	-	-	-	-	119	-	-	-	-	6	-	-	-
126	119	7	94.4	Left	ซ้าย	-	-	-	3	-	-	-	-	-	-	-	-	-	119	-	-	-	-	-	-	-
121	114	7	94.2	Right	ขวา	-	-	-	5	-	-	-	-	-	-	-	-	-	-	114	-	-	-	2	-	-
124	122	2	98.4	Open	เปิด	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	122	-	-	-	1	-
179	153	26	85.5	Close	ปิด	-	3	-	-	-	-	-	-	-	-	23	-	-	-	-	-	153	-	-	-	-
125	123	2	98.4	Delete	ลบ	-	-	-	-	-	-	1	-	-	-	-	-	1	-	-	-	-	123	-	-	-
124	121	3	97.6	Space	วรรค	-	-	-	-	-	-	-	-	-	1	-	-	-	1	1	-	-	-	121	-	-
126	123	3	97.6	Exit	ออก	-	-	1	-	-	-	-	-	1	-	-	-	-	-	1	-	-	-	-	123	-
118	114	4	96.6	Save	บันทึก	-	-	-	-	-	-	1	-	-	-	1	2	-	-	-	-	-	-	-	-	118
	Average		94.5																							

Table 4.6 Overall accuracy of propose system

Overall Accuracy

All	right	wrong	%	English	Thai	0	1	2	3	4	5	6	7	8	9	10	Up	Down	Left	Right	Open	Close	Delete	Space	Exit	Save
						ศูนย์	หนึ่ง	สอง	สาม	สี่	ห้า	หก	เจ็ด	แปด	เก้า	สิบ	ขึ้น	ลง	ซ้าย	ขวา	เปิด	ปิด	ลบ	วรรค	ออก	บันทึก
200	187	13	93.5	0	ศูนย์	187	3	-	6	1	-	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-
204	194	10	95.1	1	หนึ่ง	3	194	-	-	3	-	-	1	-	-	-	1	-	-	-	-	1	-	-	-	-
209	199	10	95.2	2	สอง	-	-	199	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	9	-
201	191	10	95.0	3	สาม	-	-	1	191	-	-	-	-	1	-	-	-	-	6	2	-	-	-	-	-	-
206	194	12	94.2	4	สี่	1	1	-	-	194	-	-	-	-	-	-	2	-	1	-	-	5	-	-	-	2
208	202	6	97.1	5	ห้า	-	-	-	-	-	202	-	-	-	1	-	-	-	-	2	2	-	-	-	1	-
206	202	4	98.1	6	หก	-	-	-	-	-	-	202	-	-	-	-	-	3	-	-	-	1	-	-	-	-
204	194	10	95.1	7	เจ็ด	-	-	-	-	-	-	5	194	3	-	1	-	-	-	-	-	1	-	-	-	-
206	198	8	96.1	8	แปด	-	-	2	-	-	-	-	-	198	-	-	-	-	-	-	2	-	-	-	4	-
200	188	12	94.0	9	เก้า	-	-	-	5	-	2	-	-	2	188	-	-	-	1	-	-	-	-	1	1	-
288	230	58	79.9	10	สิบ	-	-	1	-	5	-	-	4	-	-	230	2	-	1	-	-	44	-	-	-	1
206	203	3	98.5	Up	ขึ้น	1	-	-	-	-	-	-	-	-	-	1	203	-	-	-	-	1	-	-	-	-
208	199	9	95.7	Down	ลง	-	-	-	-	-	-	-	-	-	-	-	-	199	-	-	-	-	9	-	-	-
208	195	13	93.8	Left	ซ้าย	-	-	-	7	-	-	-	-	-	-	-	-	-	195	-	-	-	-	-	-	-
205	197	8	96.1	Right	ขวา	-	-	-	5	-	1	-	-	-	-	-	-	-	-	197	-	-	-	2	-	-
205	202	3	98.5	Open	เปิด	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	202	-	-	-	1	-
307	270	37	87.9	Close	ปิด	-	4	-	-	-	-	-	1	-	-	30	1	-	-	-	-	270	-	-	-	-
211	201	10	95.3	Delete	ลบ	-	-	-	-	-	-	1	-	-	-	-	-	9	-	-	-	-	201	-	-	-
205	200	5	97.6	Space	วรรค	-	-	-	-	-	-	-	-	-	1	-	-	-	1	1	-	-	1	200	-	1
212	201	11	94.8	Exit	ออก	-	-	4	-	-	-	1	-	3	-	-	-	-	-	1	1	-	-	1	201	-
217	205	12	94.5	Save	บันทึก	-	-	-	-	1	-	3	-	-	-	3	3	-	-	-	-	2	-	-	-	209
	Average		94.6																							

CHAPTER 5
DISCUSSION

According to the objective of this thesis and problem statements is to study, analyze and develop the prototype of Thai speech recognition system case basic voice commanding. This work is the beginning of a line of research focused on how to recognize and identification the Thai speech, Thai isolated words, by the computer.

This implementation integrates several concept and methods in to development of speech signal processing, application design and programming. After the digital signal processing, JAVA programming, speech recognition theory and method were understood, and planning has already well done. As shown the all expected result of implementation that was presented in chapter IV.

In this chapter, the discussion is made in 4 parts follow as:

5.1 The tools for prototype development.

5.2 The result of experiment.

5.3 Problem of this research

5.4 Evaluate the system.

5.1 The tool for prototype development

Java is a programming language originally developed by Sun Microsystems and released in 1995 [24] as a core component of Sun's Java platform. The language derives much of its syntax from C and C++ but has a simpler object model and fewer low-level facilities. Java applications are typically compiled to byte code which can run on any Java virtual machine (JVM) regardless of computer architecture.

"Java" generally refers to a combination of three things: the Java programming language (a high-level, object-oriented programming language); the Java Virtual Machine (a high-performance virtual machine that executes byte code on a specific computing platform, typically abbreviated JVM); and the Java platform, a JVM running compiled Java byte code, usually calling on a set of standard libraries such as those provided by Java Standard Edition (SE) or Enterprise Edition (EE). Though coupled by design, the language does not imply the JVM, and vice versa.

Platform independence

One characteristic, platform independence, means that programs written in the Java language must run similarly on any supported hardware/operating-system platform. One should be able to write a program once, compile it once, and run it anywhere.

This is achieved by most Java compilers by compiling the Java language code halfway (to Java bytecode) – simplified machine instructions specific to the Java platform. The code is then run on a virtual machine (VM), a program written in native code on the host hardware that interprets and executes generic Java bytecode. (In some JVM versions, bytecode can also be compiled to native code, either before or during program execution, resulting in faster execution.) Further, standardized libraries are provided to allow access to features of the host machines (such as graphics, threading and networking) in unified ways. Note that, although there is an explicit compiling stage, at some point, the Java bytecode is interpreted or converted to native machine code by the JIT compiler.

The first implementations of the language used an interpreted virtual machine to achieve portability. These implementations produced programs that ran more slowly than programs compiled to native executables, for instance written in C or C++, so the language suffered a reputation for poor performance. More recent JVM implementations produce programs that run significantly faster than before, using multiple techniques.

One technique, known as just-in-time compilation (JIT), translates the Java byte code into native code at the time that the program is run, which results in a program that executes faster than interpreted code but also incurs compilation overhead during execution. More sophisticated VMs use dynamic recompilation, in which the VM can analyze the behavior of the running program and selectively recompile and optimize critical parts of the program. Dynamic recompilation can achieve optimizations superior to static compilation because the dynamic compiler can base optimizations on knowledge about the runtime environment and the set of loaded classes, and can identify the hot spots (parts of the program, often inner loops, that take up the most execution time). JIT compilation and dynamic recompilation allow Java programs to take advantage of the speed of native code without losing portability [27].

5.2 The result of experiment

From the experiments conducted to verify the system integrity, it found that, if design the system to easy to use for user the developer must take an automatic function to provide and support for related task. The automatic function can meet the need of user but it hard to create and develop. The propose system is try to create more automatic function with high accuracy result but take time too. The focus discussion of experiment is divided into 2 parts as in the following:

Discussion of the Advantage and disadvantage of system using Mel adjusted scale to define a range of each band in filter bank method.

Advantage

Focus in low to middle of frequency not over 3200Hz (shown in table 4.1), because the fundamental frequency is in range 200-400Hz and frequency that more than 3200Hz are less information to use, to track for a step to change of power spectral.

Disadvantage

If the speech have current frequency more then 3200Hz must use Mel or Bark scale for suitable range.

Discussion of the Advantage and disadvantage of 2nd-normalization the Mel adjusted filter bank by uniform filter bank.

Advantage

1. The speech features that obtain from Mel adjusted filter bank ca make more concisely characterize.
2. Can use for normalization the final feature of signal that not ever to normalize the range in time domain.

Disadvantage

Increase the computation time.

5.3 Problem of this research

Since Speech Recognition concept contain the several step of method. Some step may have some problem occur and will show by follow as:

5.3.1 The data collecting problem

The way to collecting the data, speech sound, can affect directly to developer of the propose system. This thesis needed to collect the speech sound to use for convert to speech signal and analysis in next step later.

One problem to collecting the speech is the format of speech sound that use in system because this thesis use the java programming and care for independent platform then must use such as *.wav*, *.au*, *.aiff* that java supported.

The selected of this propose system solution is new define sound format that use in only this system or local sound storing format that support by java such as array of byte, one of java data type, that can save in text file.

The difficulty of collecting speech sound is machine to uses in this case must use the computer that connected with microphone to collect speech sound via software module in propose system only.

5.3.2 Computing time consuming in preparing system

In mathematic many formulas must use the loop to solve the result. In this system is uses many formulas. For one signal the process must use the formulas that consist of sampling and digitizing, start-end point detecting, DFT, Pre-emphasis, windowing, Mel adjusted scale filter bank, normalization by uniform filter bank.

The nested of method must use a computing time for prepare and convert the speech signal to vector that represent the speech signal. This system must to process for 20000 speech samples (20 speech type x 20 subjects x 10 series x 5 words / series) that divide into 2 parts. Part one for prepare the model code book 11000 sample sound and part two 9000 sample speech for testing the accuracy of the system.

Unfortunately, if the data prepare is failed or have error occur it mean that the new waiting time to must used for data preparing (shown in table 4.2 and table 4.3).

5.3.3 Missing identifying of system

The missing deification is main focus that can affect the reliable accuracy rate of system. Major missing is focus to the word “ten” and “close” (in Thai must pronoun in “ปิด” and “สิบ”). The analysts are focused into basic components of isolated word like a tone, start consonant, stop consonant, vowel, and pronoun time period. In this case the analyses are can shown the step follow as:

1. Tone issue: the both are having the same tone.
2. Vowel issue: the both are having the same vowel
3. Pronoun time period issue: the both are short period.

Final, this result is shown that if the isolated word is have the same condition, same component, it is make a harder to identify because the rule base is decrease parameter or condition to use for distinguish the isolated word.

5.4. Evaluate the system

The main point of developed of this system is the percentage of accuracy to identification the unknown speech. This system can be detected and identification correctly in 94.6 percent for isolated speech as shown in figure 5.1. The satisfy accuracy rate of this propose system is define to 85 percent.

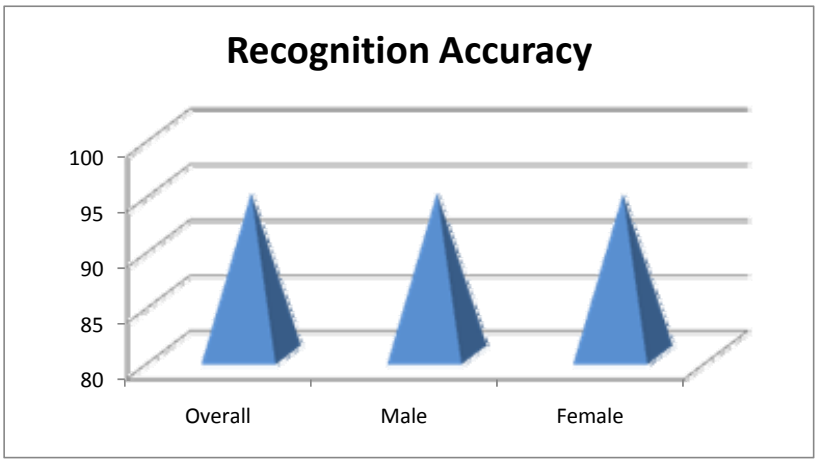

Figure 5.1 Shown the speech recognition accuracy (percent) of system

Finally, the system can be receive the input via microphone and identify the unknown speech in satisfy time and satisfy accuracy in real time uses. I hope that this project will be worthy and can use for case study or use for base of advance Thai speech recognition system.

CHAPTER 6
CONCLUSION

The success of failure of speech recognition implementation is related to several factors such as the system planning and design, project management, programming knowledge, research methodology, logical technique, feature extraction technique, analysis technique and tools, and tools of implementations.

This thesis presents a methodology that is different from other existing method very a little. According to the previous Thai speech recognition system researcher used many feature extraction for speech signal such as DFT, FFT, filter bank, Linear Prediction Code (LPC), Hidden Markov Model (HMM), linear discriminate analysis. The advantage of filter bank that used is follow as:

1.) Perform an average over neighboring frequency

2.) Use the uniform filter bank, Mel scale–filter bank, Bark scale–filter bank or custom adjust scale-filter bank.

3.) Reduce from 256 or more Fourier coefficients to 24 output (widely used) or N output where N is custom define number.

4.) Suitable for isolated words (widely used for speaker recognition).

The filter use for feature extraction and modeling is the key information to represent the speech with familiar information like a vector distance in this thesis is use adaptive parameter in DFT function and adaptive by mix scale as well known like a Mel scale and uniform scale in filter bank analysis.

The storing of digital speech format not same the widely used of sound format as a presented format but it used the data type in array of byte format of Java to make a quickly time to load and save signal for operational in any analysis of this system. And another required is decision-making rule, rule-based, which also has the

potential capability to handle any ad-hoc analysis and consistency with the speech signal format.

Recommendation

There are four recommendations, challenges faced, for further study:

- The prototype of speech recognition if not concern about platform should support for many sound format.
- The DFT must set parameter to 2^n and replace by FFT for faster computation
- The computing time to prepare 11000 sample sound (or more in the future) must be solved to reduce.
- The capability of speech sample number should be increase.
- The continuous Thai speech recognition is highly recommended to develop.

REFERENCES

1. Linguistics [Online]. Available from: http://en.wikipedia.org/wiki/Linguistics (Accessed 2007 September 18).
2. Phonetics [Online]. Available from: http://en.wikipedia.org/wiki/Phonetics (Accessed 2007 September 18)
3. Larynx [Online]. Available from: http://www.medicalook.com/human_anatomy/organs/Larynx - MedicaLook Human Anatomy.html (Accessed 2007 October 12)
4. Speech Production [Online]. Available from: http://ispl.korea.ac.kr/~wikim/research/speech.html (Accessed 2007 September 23)
5. Masaaki Honda. Human Speech Production Mechanisms [Online]. NTT Communication Science Laboratories. Atsugi-shi. Japan. Available from: http://www.analogue.org/network/Speech%20production.pdf (Accessed 2007 September 23)
6. Introduction to Digital Signal Processing [Online]. Available from: http://www.dsptutor.freeuk.com/intro.htm (Accessed 2007 May 11)
7. Sampling [Online]. Available from: http://en.wikipedia.org/wiki/Sampling (Accessed 2007 Feb 12)
8. Sampling Rate[Online]. Available from: http://wagstaff.asel.udel.edu/speech/tutorials/instrument/sam_rat.html (Accessed 2007 Feb 13)
9. Sound in the Time Domain [Online]. Available from: http://www.asel.udel.edu/speech/tutorials/acoustics/time_domain.html (Accessed 2007 May 20)
10. Frequency Domain [Online]. Available from: http://www.asel.udel.edu/speech/tutorials/acoustics/time_domain.html (Accessed 2007 May 18)
11. Claudio Becchetti, Lucio Prina Ricotti. Speech recognition theory and C++ Implementation, John Wiley & Sons, Inc., 1999.
12. Windowing [Online]. Available from: http://svr-www.eng.cam.ac.uk/~ajr/SA95/node26.html (Accessed 2007 Feb 17)

13. Julius O. Smith III. Mathematics of the Discrete Fourier Transform (DFT) with audio applications second edition [Online]. Available from: http://ccrma.stanford.edu/~jos/mdft/mdft.html (Accessed 2007 May 10)
14. Discrete Fourier transform [Online]. Available from: http://en.wikipedia.org/wiki/Discrete_Fourier_transform (Accessed 2007 May 10)
15. The Discrete Fourier Transform [Online]. Available from: http://www.dspguide.com/ch8.htm (Accessed 2007 Feb 25)
16. Ben J. Shannon, Kuldip K. Paliwal. A Comparative Study of Filter Bank Spacing for Speech Recognition, Microelectronic engineering research conference 2003.
17. Umesh, S., Cohen, L. and Nelson, D. Frequency warping and the Mel scale, IEEE Signal Processing Letters, Volume: 9, Issue: 3, pp. 104 -107, 2002.
18. Speech parameterization using the Mel scale Part II [Online]. Available from: http://www.ee.bilkent.edu.tr/~onaran/SP-4.pdf (Accessed 2007 May 25)
19. Molau, S., Pitz, M., Schluter, R. and Ney, H. Computing Mel-frequency cepstralcoefficients on the power spectrum, Acoustics, Speech, and Signal Processing Proceedings, Volume: 1,pp. 73 -76, 2001.
20. Filter bank [Online]. Available from: http://en.wikipedia.org/wiki/Filter_bank (Accessed 2007 Feb 25)
21. เสรี ปานซาง และ ชม กิ้มปาน. การรู้จำเสียงพูดคำไทยเฉพาะบุคคลด้วยนิวรัลเน็ตเวิร์ค. การประชุมวิชาการทางไฟฟ้า ครั้งที่ 18.ชลบุรี มหาวิยาลัยมหานคร.2538
22. Krishnamoorthy, M., George Nagy, Sharad Seth, Mahesh Viswanathan., Syntactic segmentation and labeling of digitized pages from technical journals. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, no7.737-747
23. X. Huang, A. Acero, and H. Hon, Spoken Language Processing. A guide to theory, algorithm, and system development: Prentice Hall, Inc., 2001.
24. What is java [Online]. Available from: http://www.onjava.com/pub/a/onjava/2006/03/08/what-is-java.html (Accessed 2006 June 20)

25. Java [Online]. Available from: http://en.wikipedia.org/wiki/
Java_(programming_language) (Accessed 2006 July 05)

26. The Java Platforms [Online]. Available from: http://www.onjava.com/
pub/a/onjava/2006/03/08/what-is-java.html?page=2#java-platforms
(Accessed 2006 June 25)

27. Platform independence (Java) [Online]. Available from: http://en.wikipedia.org/
wiki/Java_(programming_language)#Platform_independence
(Accessed 2006 July 06)

APPENDIX

APPENDIX A

Storage Format

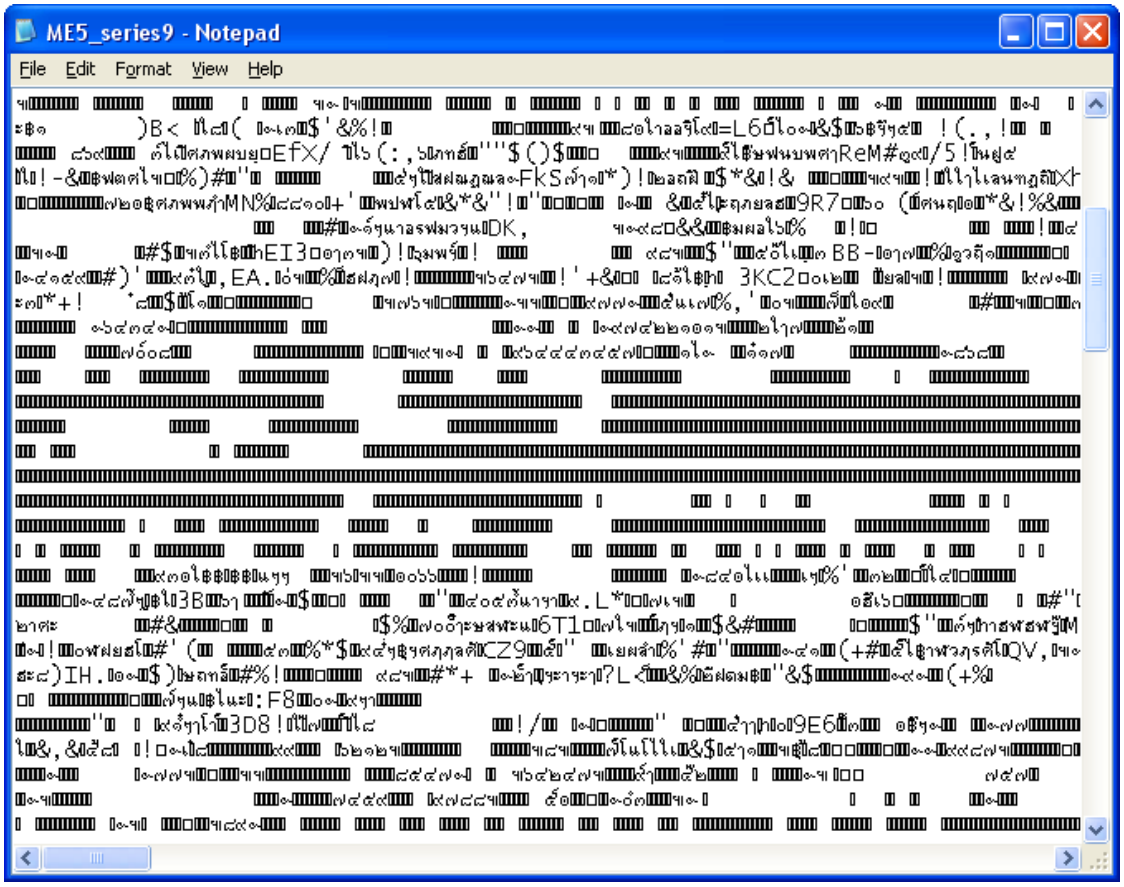

Figure 1 Example of digital data in byte format that represented of analog signal

BIOGRAPHY

NAME	Mr. Kriengkrai Nantanitikron
DATE OF BIRTH	12 September 1979
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENED	Chiang Mai University, 1997-2001: Bachelor of Science (Mathematics) Mahidol University, 2003-2008: Master of Science (Technology of Information System Management)
HOME ADDRESS	429 Armonchai3 Village, Boromarajonani Rd. Salatammasop Sub District, Taveewattana District, Bangkok 10170
RESEARCH GRANT	This thesis is partially supported by Graduate Studies of Mahidol University Alumni Association