VOICE RECOGNIZER FOR PERSONAL IDENTIFICATION

VARAPORN PHOMVI-IN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2000

ISBN 974-663-997-8

Thesis

entitled

VOICE RECOGNIZER FOR PERSONAL IDENTIFICATION

Varaporn Phomvi-in

.....

Miss Varaporn Phomvi-in
Candidate

.....

Assoc.Prof. Supachai Tangwonsan, Ph.D.
Major-Advisor

.....

Asst.Prof. Damras Wongsawang, Ph.D.
Co-advisor

.....

Prof. Liangchai Limlomwongse,
Ph.D.
Dean
Faculty of Graduate Studies

.....

Assoc.Prof. Supachai Tangwonsan, Ph.D.
Chairman
Master of Science Programme in
Computer Science
Faculty of Science

Thesis

entitled

VOICE RECOGNIZER FOR PERSONAL IDENTIFICATION

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Computer Science)

on
April 28, 2000

Varaporn Phomvi-in
.....
Miss Varaporn Phomvi-in
Candidate

.....
Assoc.Prof. Supachai Tangwonsan, Ph.D.
Chairman

.....
Asst.Prof. Damras Wongsawang, Ph.D.
Member

.....
Assoc.Prof. Chinda Achariyakul, Ph.D.
Member

S Phongsuphap
.....
Lect. Sukanya Phongsuphap, Ph.D.
Member

.....
Prof. Liangchai Limlomwongse, Ph.D.
Dean
Faculty of Graduate Studies
Mahidol University

.....
Prof. Amaret Bhumiratana, Ph.D.
Dean
Faculty of Science
Mahidol University

ACKNOWLEDGEMENT

I would like to gratefully thank to Dr. Supachai Tangwongsan, for his guidance, invaluable advice, supervision, encouragement and constructive criticism, which enabled me to carry out the thesis successfully. I am also grateful to my reviewing committee members, Dr. Damras Wongsawang, Dr. Chinda Achariyakul for their constructive guidance.

Great appreciation also goes to Unocal Thailand, Ltd., for providing me with an education aids. Moreover, I want to express my gratitude to Mr. Kreuze, Jim, Unocal Thailand's English Instructor, who helped with the thesis proof-reading and grammar editing.

I wish to express my appreciation to all of my teachers for their teaching and advice during my graduate studies at Mahidol University. Also, all my graduate friends are thanked for their friendly and warm working during study.

Most of all, I am deeply thankful to my mother and my young sister whose unconditional love.

Varaporn Phomvi-in

3837393 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc. (COMPUTER SCIENCE)
KEY WORD : SPEAKER RECOGNITION / SPEAKER IDENTIFICATION
VARAPORN PHOMVI-IN : VOICE RECOGNIZER FOR PERSONAL IDENTIFICATION. THESIS ADVISOR : SUPACHAI TANGWONGSAN Ph.D., DAMRAS WONGSAWANG Ph.D. 65 p. ISBN 974-663-997-8

This research presents a model of a voice recognizer for personal identification. It is the process of automatically determining who the speaker is by matching his/her speech pattern to reference speakers in a database, which is obtained from a group of known speakers. The outcome of speaker identification is the one whose speech pattern is the best match to the known speech samples.

The voice recognizer system in this study consists of three major parts, namely: feature extraction, clustering, and identification. In the feature extraction, we initially detect voiced segments of speech samples by using ZCR (Short-Time Average Zero-Crossing Rate), then given as input to perform DFT (Discrete Fourier Transform) and compute the PSD (Power Spectrum Density). Next, the PSD vectors are used as training samples in a pattern-based training model and also used to calculate the distribution of the acoustic features by employing ML (Maximum Likelihood Procedure) to represent the statistical-based training model. In addition, gender classification was implemented in order to improve the performance during testing. This is based on the pitch frequency value that can be computed by using Auto-correlation function.

From the clustering method, we pre-detected the possible set of speakers who are acoustically similar to the test speaker by combining the scoring procedure and the Mahalanobis distance measure to optimize the distribution of each phonetic data. Finally, in the identification process, we employed the square-error pattern matching to determine the cohort set which is the last and smallest set of possible speakers. Then the decision algorithm is performed to obtain the most likely speaker from the cohort set.

The system is implemented in a stand-alone personal computer, in which Thai language is chosen as the word utterance for the evaluation. For noise-free speech recording environment, the voice recognizer is able to achieve the identification within accuracy up to 95%, while the identification error down to 5%. Concerning the computational time, the system is able to determine the speaker within 30 seconds, which is quite short in the sense of real environment. Moreover, the system requires minimal disk space because of using a small number of speech feature representatives, which significantly reduces the system pre-processing time.

3837393 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

วราภรณ์ พรหมวิอินทร์ : ระบบตรวจรู้เสียงพูดเพื่อระบุตัวบุคคล (VOICE RECOGNIZER FOR PERSONAL IDENTIFICATION). คณะกรรมการควบคุมวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., ดำรัส วงศ์สว่าง , Ph.D. 65 หน้า. ISBN 974-663-997-8

งานวิจัยนี้เป็นการนำเสนอการออกแบบต้นแบบของระบบตรวจรู้เสียงพูดเพื่อระบุตัวบุคคล คือกระบวนการที่ทำการบ่งชี้อย่างอัตโนมัติว่าใครคือผู้พูดโดยใช้การเทียบเคียงรูปแบบเสียงของผู้พูดทดสอบเข้ากับรูปแบบเสียงของผู้พูดอ้างอิงในฐานข้อมูลซึ่งเป็นการจัดเก็บรูปแบบเสียงจากกลุ่มผู้พูดที่ระบบรู้จัก ผลลัพธ์ที่ได้จากการตรวจรู้เสียงผู้พูดคือคนที่มีรูปแบบเสียงเข้ากันได้ดีที่สุดกับรูปแบบเสียงของผู้พูดทดสอบ

การศึกษาระบบตรวจรู้เสียงพูดนี้ประกอบด้วย 3 ส่วนหลักคือ การสกัดตัวแปรจำเพาะ (feature extraction) การแบ่งส่วน (clustering) และการบ่งชี้ (identification) ในการสกัดตัวแปรจำเพาะ ขั้นแรกจะตรวจจับเอาเฉพาะสัญญาณที่เป็นเสียงโดยใช้ ZCR (Short-time average zero-crossing rate) แล้วส่งสัญญาณที่ได้เข้า DFT (Discrete Fourier Transform) และทำการแปลงเป็น PSD (Power spectrum density) จากนั้นค่า PSD จะถูกนำไปเป็นตัวอย่างในแง่รูปแบบเสียงและถูกนำไปคำนวณหาค่าการกระจายของคุณลักษณะเสียงโดยใช้ ML (Maximum likelihood procedure) เพื่อเป็นตัวอย่างในแง่สถิติ นอกเหนือจากนี้เรายังได้พัฒนาการระบุเพศของผู้พูดโดยศึกษาการหาค่า pitch frequency value ที่สามารถคำนวณได้จากสมการ Auto-Correlation สำหรับการแบ่งส่วนใช้กลยุทธในการในการค้นหากลุ่มตัวอย่างของผู้พูดที่น่าจะมีรูปแบบเสียงคล้ายกับผู้พูดทดสอบ ขั้นตอนสุดท้ายเราจะใช้การเทียบเคียงรูปแบบเสียงของผู้พูดแต่ละคนในกลุ่มตัวอย่าง (cohort set) เข้ากับรูปแบบเสียงของผู้พูดทดสอบ และกระบวนการตัดสินใจจะทำการเลือกคนที่น่าจะใช่มากที่สุดออกมา

งานวิจัยนี้ได้พัฒนาบนเครื่องคอมพิวเตอร์ส่วนบุคคล โดยเลือกภาษาไทยเป็นคำที่ใช้ทดสอบ ในสภาพแวดล้อมที่เงียบปราศจากเสียงรบกวน ระบบตรวจรู้เสียงพูดให้ผลถูกต้องสูงถึง 95% ขณะที่ผลผิดพลาดจะอยู่ในราว 5% ในแง่ของเวลาในการคำนวณ ระบบสามารถระบุผู้พูดได้ภายใน 30 วินาที ซึ่งก็ถือว่าไม่มากเมื่อเทียบกับสถานะการณ์จริงของคน นอกจากนี้ระบบยังต้องการพื้นที่จัดเก็บฐานข้อมูลน้อย ซึ่งเป็นเหตุผลในแง่ประสิทธิภาพการคำนวณดีขึ้นอีกด้วย

CONTENTS

CONTENTS (Cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Speaker identification is the process of automatically determining who is a speaker by matching his/her speech pattern to speech records of known speakers. It involves the task of extraction of characteristic information from the speaker speech waveforms. A speech pattern of unknown speaker is analyzed and compared with samples of reference speakers, and attempted to determine the most likely speaker which is the best match to known samples. There are many applications required the usage of speaker identification, such as telephone banking and shopping, database access control, voice mail, remote access to computer, and physical access control to privileged locations in secure areas. It is classified in two types: text-independent and text-dependent speaker identification. The former requires the user to utter a set of pre-defined words, while the latter allows the user to utilize any vocabularies which are rather difficult for processing speaker identification.

The main objective of this research work is to develop a model of text-dependent speaker identification system, using pre-defined words in Thai language. We start to record the speech feature that requires minimal computation and the input speech is recorded under the pre-defined conditions. To obtain the significant acoustic features, the input speech is digitized and then segmented voiced/unvoiced signal by the Short-Time Average Zero-Crossing Rate (ZCR). Next, it is modulated by the Hamming window of approximately 25 msec in duration time for signal smoothing. The discrete Fourier Transform (DFT) is then performed and computed the square component-wise to obtain the Power Spectrum Density (PSD) of speech interval. The PSD vectors are used as training samples in pattern-based training model and also used to calculate the distribution of the acoustic feature by employing ML (Maximum Likelihood Procedure) to represent the statistical-based training model. In addition, the speech interval is also used as input to compute the pitch frequency value using Auto-correlation function. The results will be sent to the clustering process to seek for the set of possible speakers against the statistical-based speaker model. Next, the speech pattern of the possible speaker is used to match against the known speaker data using Square-Error Pattern Matching, and end up with the cohort set. Finally, the square-error scores are used for identification decision process to select the most likely speaker.

Our system is implemented on a PC system, running under Windows 95 platform. We use C and Visual basic as the software tools. The experimental result is quite satisfactory and the performance is impressive in term of computational time and storage usage. Essentially we apply the speaker clustering approach, which

include scoring and distance calculation. We search for the optimal set of possible speakers and calculate the most likely one, which allows us to achieve very good performance.

The thesis is organized in seven chapters as follows:

Chapter 1 : Introduction

Chapter 2 : Problem statement, the problem definition is investigated and literature survey is reviewed. The information of speech signal processing is presented. In the final section, the thesis objective and problem scope are defined.

Chapter 3 : Approach, this chapter explains the method to implement speaker identification, which consists of two parts including acoustic approach and the recognition approach.

Chapter 4 : System design and implementation, this chapter explains the formulas and algorithms that utilizing the concept mentioned in chapter 3. The implement of the system is presented by both the examples of calculation and the program flow chart.

Chapter 5 : Experimental results, this chapter presents the experimental results of the Speaker Identification System. We first describe how to setup the system using in this experiment, the result of gender determination, and the results of identification are presented in the final.

Chapter 6 : Discussion and conclusion, the discussion and conclusion on the experiment results presented in the chapter 5.

Chapter 7 : Suggestion for the future work, this chapter proposes the future work in connection with our research and discuss the improvement area which is needed.

CHAPTER II

PROBLEM STATEMENT

2.1 Problem identification

At present, the human voice is considered to be the key for the accessing control technology because it is more users friendly, secure, and it does not involve loss of keys or copying of passwords. One such method is speaker identification. Ideally, it would provide several advantages in the following areas:

1. Information Access Control: Many applications can use the speaker identification for better services, including telephone banking and shopping, database access and information services, voice mail, and remote access to computer.

2. Physical Access Control: Speaker identification can be used for physical access control to privileged locations from secure areas; for example, computer rooms, hotel rooms, and residential homes.

3. Speech Recognition: Additionally speech recognition sometimes requires information of who is speaking because the intention of an utterance becomes different depending on the speaker [2]. Therefore, speaker identification can be used in conjunction with speech recognition to improve its recognition performance.

2.2 Literature survey

Although this thesis addresses the problem of designing a system to perform speaker identification, most of our survey is a study of speech production and perception, a search for the speech features of our interest, and an understanding of signal representation and processing. The details are given in this section.

2.2.1 Survey of speech production and perception [2,3,6,7]

2.2.1.1 Speech production

When a person produces speech sounds, many kinds of information are embedded in the speech signal, including the identity of the speaker, the language spoken, the presence and type of speech pathologies, and the emotional state of the speaker. Speech sounds produced upon a physical process consist of two parts:

1. Sound source including the capacity of the individual lung, size of the vocal tract and the vocal folds.

2. Filtering including tongue, the lips, teeth, nasal cavity, etc.

During speech production, the lungs act as the air supply by blowing the air through the vocal tract. This flow of air causes the vocal folds to vibrate. The vocal folds are two membranes situated in the larynx. These membranes allow the area of the glottis to be varied. The way in which the vocal folds are vibrated, the shape of the vocal tract or the site of the constriction can all be varied in order to produce the range of speech sounds with which we are familiar. Figure 2.1 shows the vocal tract.

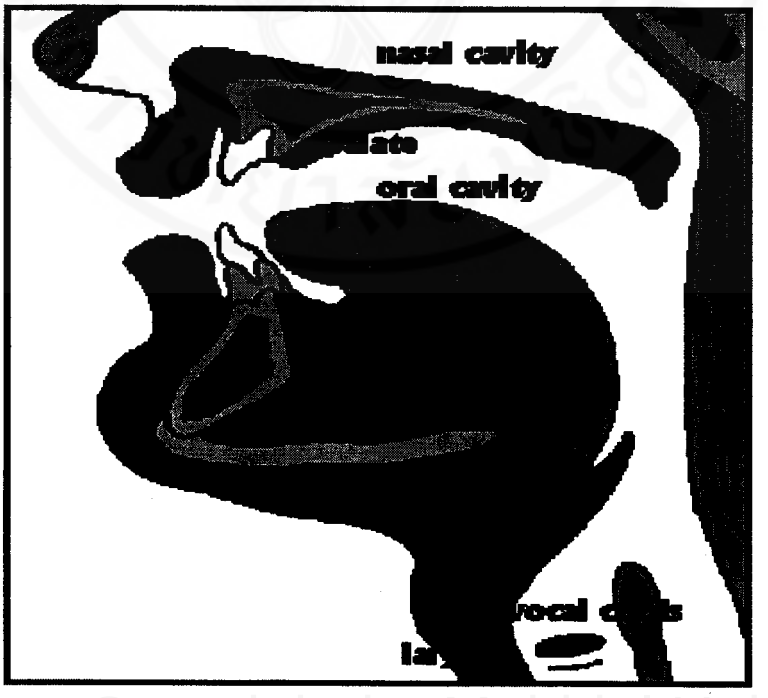


Figure 2.1 The vocal tract

2.2.1.2 Speech perception

The human ear consists of three main sections, the outer, the middle and the inner ears. The outer ear consists of the ear lobe and the external auditory canal which is to channel sounds into the middle ear. The middle ear adjusts the pressure levels between the outer and inner ear. Muscles attached in the middle ear will suppress the vibration of the sound air if it is too violent and so protect the inner ear. This protection only works for sounds below 2kHz and it does not work for impulsive sounds. The Eustachian tube connects the middle ear to the vocal tract and removes any static pressure difference between the middle ear and the outer ear. If a significant pressure difference is detected, then the Eustachian tube opens and the differences are removed. The inner ear or cochlea contains the sensitive apparatus. Inside the cochlea, there is a hair-lined membrane called the Basilar membrane. This membrane is used to convert the sound energy into neural signals for the brain. Different frequencies excite different portions of this membrane allowing a frequency analysis of the signal to be carried out. So, the ear is essentially a spectrum analyzer that responds to the energy of the signal. These three sections can be clearly seen in Figure 2.2.

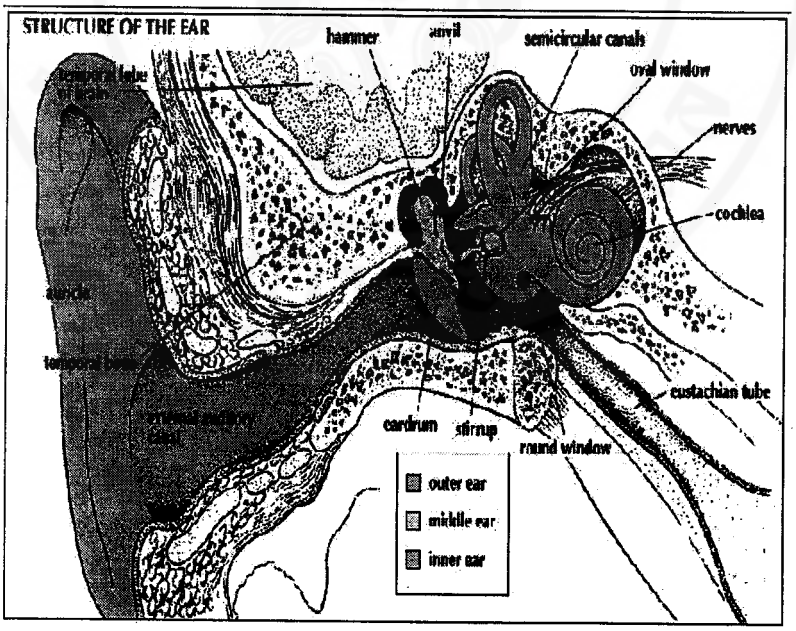


Figure 2.2 Inside the ear

Therefore, the ear is the receiver which has limitations of sensitivity, called the threshold of audibility. If sounds are too weak, they will not be detected. Different speech sounds are distinguished on the basis of their short time spectra and how these spectra evolve with time.

2.2.1.3 Spectrum of speech

Speech can be broken into two classes, called voiced and unvoiced. During the voiced sound, like the vowel, the vocal tract acts as a resonant cavity. This resonance produces large peaks in the resulting speech spectrum, known as formants. The formants contain almost all information contained in the signal, including the quality and naturalness of speech. Figure 2.3 shows frequency-amplitude sections for the vowels. The formants are marked and numbered. A resonant frequency, amplitude, and bandwidth can characterize each formant. Generally, the lower three formants characterize the phonetic quality of the sound itself, while the extra resonance adds to the accuracy and naturalness of speech. It should be noted that the formant structure appears to peak down above 4 kHz as noise introduced by the turbulent flow of air through the vocal tract begins to dominate.

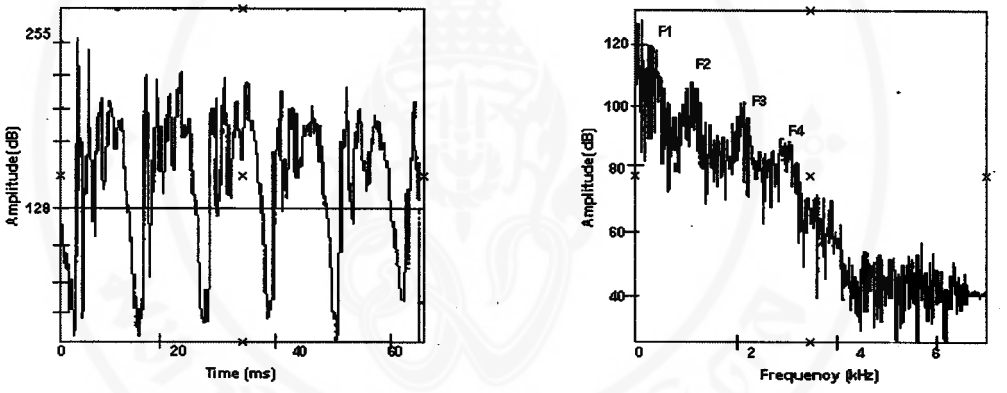


Figure 2.3 The spectrum of voiced speech

Unvoiced sounds, like ซอบ, เสีย, are generated by constricting the vocal tract close to the lips. Unvoiced speech tends to have a nearly flat or a high-pass spectrum. The formants that are so obvious in voiced speech are gone and so is the fine pitch structure. The energy in the signal is also much lower than that in voiced speech. An example of unvoiced speech is given in Figure 2.4

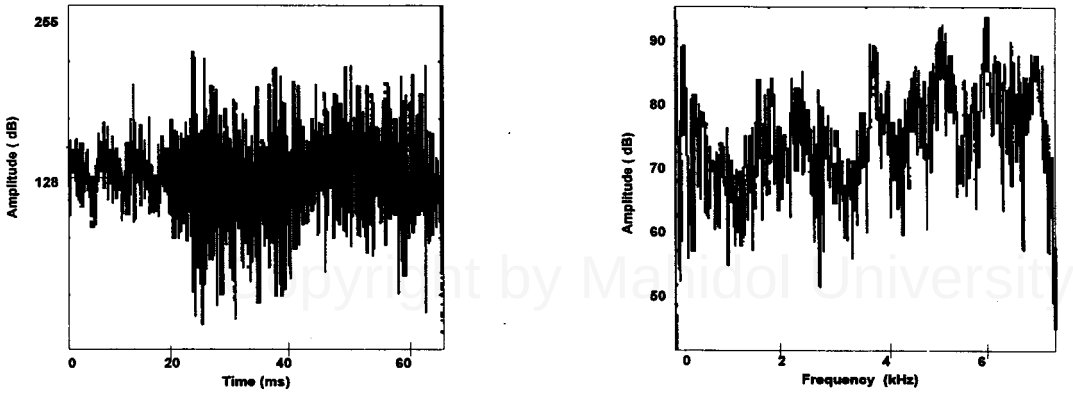


Figure 2.4 The spectrum of unvoiced speech

2.2.2 Speech features

Speaker identity from speech signal is the main focus of our observation. Speaker identity is correlated with physiological and behavioral characteristics of the speaker. For example, variations in the size of vocal tract cavities produce differences in the 'characteristic resonance of the spectrum of the speech signal. Size of the vocal folds is associated with changes in frequency of voiced speech, and the size of the nasal cavities produce spectral differences in nasalized speech sounds. All of these characteristics are often difficult to measure.

Voiced sound measurements have been mentioned as good candidates for characterizing speakers. As mentioned in 2.2.1.3, voiced sound is the resonance that produces large peaks in the resulting speech spectrum. These peaks are called formants and are numbered from the bottom up, as F1, F2, F3, etc. The numbers of peaks are from 2 positions up to 6 positions, depending on the speaker. The positions of the formants are different for different sounds, and they can often be predicted for each phoneme. Table 2.1 shows the average range of the formant frequency.

Table 2.1 The range of Formant frequency

Formant	Range of Frequency	Energy	Frequency
F1 (lowest Formant)	200 – 1000 Hz	▲ High	▼ Low
F2	500 – 2500 Hz		
F3	1500 – 3500 Hz	Low	High

To estimate the position of formant frequency is very difficult. The operation of the Fourier Transform can result in an estimate of the power spectrum of the speech segment. The advantage of a power spectrum is that it is a real function of frequency, and the power of signal components is sometimes a closer indication of the perceptual importance (peak of formants) than the amplitude in the spectrum.

2.2.3 Signal representation [5,6,7]

The human voice signal as we have seen above is very complex and has a lot of information encoded into it. With modern electronics technology, the speech can be represented in speech waveform, which is shown in Figure 2.5. The signal was commonly sampled at 10 kHz with 8 bit precision which is mentioned effective for extracting the speech features. The analysis of speech signal can be implemented in two parts, a time domain and frequency domain. With our observation, we considered 2 speech features to the represent the speech signal that can be used in our speaker identification research, one is prosodic feature analysis on time domain and the another is the power spectral feature analysis on frequency domain.

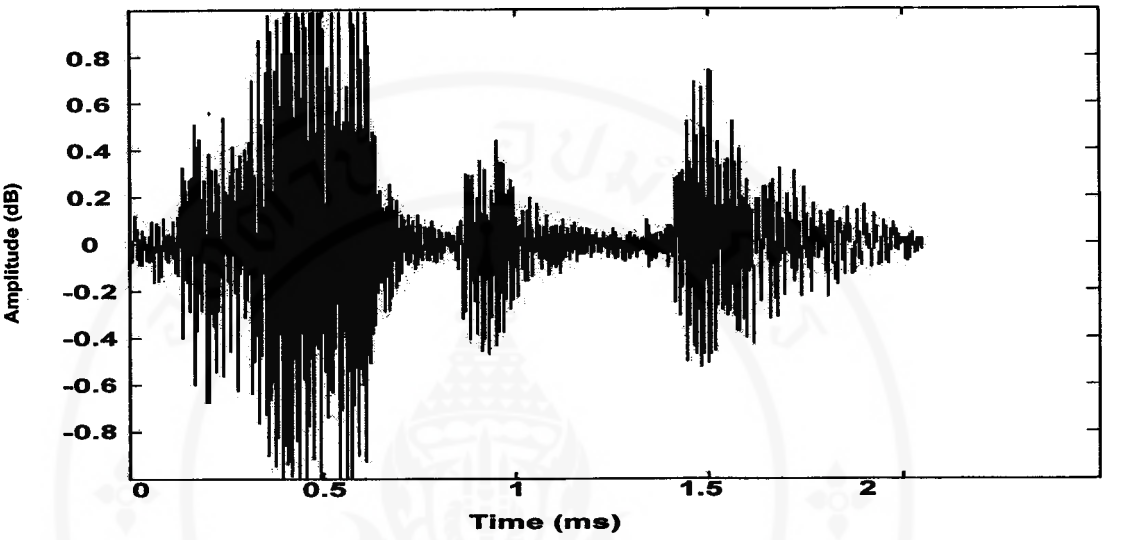


Figure 2.5 The speech waveform

2.2.3.1 Prosodic feature

Prosodic features are considered as combine of stress, intonation, duration and juncture. These features have a very important role to play in communicating meaning. One of feature fall in "intonation" is the fundamental frequency or "pitch" of the signal that can be presumably used to differentiate the speaker gender. To estimate the pitch of the signal, the analysis performs on time domain signal using short-time auto-correlation function for a selected frame of sampled data, which can measure the pitch of the voiced speech quite accurately. It shows the correlation between one sample and a few of its predecessors. The correlation is large if at some delay the two signals have similar waveform. This involves finding the distance between the most significant large value of the correlated signal, which corresponds to the pitch period.

2.2.3.2 Power spectral feature

As noted earlier, power spectral density represents the frequency spectrum of speech signal. The frequency spectrum of the voiced source is mixed with the vocal tract information, which contain almost all information to differentiate the speaker.

To obtain the power spectral feature, speech samples are initially modulated by Hamming Window and the discrete Fourier Transform (DFT) is applied to produce

the spectral coefficients of approximately 25 msec in duration. Then each frame was squared component-wise to obtain the power spectral coefficients.

2.2.4 Research publications [5,7,8]

There have been several studies for speaker identification for many years. Basic operation of a speaker recognition system is shown in Figure 2.6. The interested speech features are extracted from an input utterance of an unknown speaker. The measurement is carried out by comparison with every speaker model and used the matching score of comparison to make an identification decision.

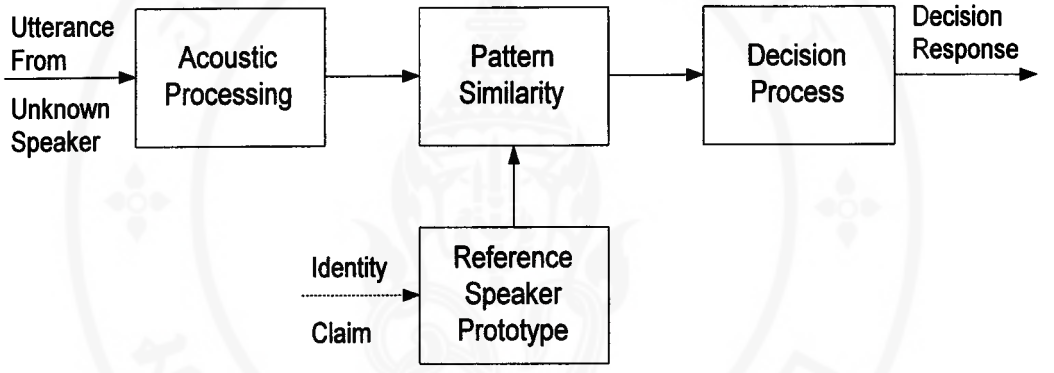


Figure 2.6 Basic block diagram of speaker recognition process

The above mentioned speaker identification was start investigated by Wolf in 1972, he carried out an experiment to rate measurements of various types of speech events, segmented manually, for his speaker discriminative capacity. In his test trial, target phonetic events are detected in test utterances and the measurements compared with prototype measurements for reference speakers. This approach is potentially powerful and in opposite its taken with long-term statistics which represent generalized measurements carried out over a large number of phonetic events. However, there is a lack of practical implementations based on this approach because of the difficulty of the accurate and consistent detecting target phonetic events.

In the 1980s, many researchers began using new approach called Vector Quantization (VQ) codebook to compress the training feature vectors to small set of representative points. The VQ codebook is simply a distillation of all the training feature vectors available for a speaker. It is intended to be an optimal representation of the training vectors but the codebook vectors do not have an obvious association with explicit speaker characteristics.

In early 1990s, Poritz proposed using a statistical model, called Hidden Markov Modeling (HMM) for test-independent speaker recognition. The HMM

representation can be considered in speech events, e.g., word, sub-word phone link units, or acoustic segment. This approach has become very successful in speech recognition for many years later. Many investigators tried to apply this approach to a speaker identification experiment by expanding Poritz's idea. Here again there are no explicit representations of speaker characteristics.

Recently, new approach, such as neural networks has been applied to speech recognition experiment. The computing architecture of neural network consists of massive and parallel inter connection of simple neural processors and also multiple layers of interconnected cells. Neural network has a lot of successful in speech and hand writing recognition. For speaker recognition, each speaker represented by a unique neural network, and the identification is based on the weight of each speaker's model. It is observed that the computation intensive and the development time and resources are expensive.

2.3 Research objectives

As mentioned problems in section 2.1 and the literature survey we have been made on the speaker recognition in section 2.2. Our objectives and goals of speaker identification research can define as follows:

1. We hope to develop a speaker identification system which is requires minimal computation. Thus, the investigation will be focus on the acoustic features that have been proven to be robustness and no extraction complexity, anyhow, because of we do not intend to investigate on robustness issues specifically.

2. Study the method to extract the studied optimal acoustic features from the speech signal and the method to transform its to be the appropriate speaker model format.

3. To keep our computational efficient and not depend on the number of reference speakers.

4. We study the technique to pre-detected a small set of possible speaker using statistical approach instead of comparison with every speaker model.

5. Due to the pattern of acoustic feature can be well represented the explicit speaker characteristics. We study to combines two practicable approaches to speaker identification: statistical and pattern matching.

6. After pattern matching process, it is possible to appear that more than one people who have the score well against to the test speaker. We study to define the algorithm to determine the only one most likely speaker.

2.4 Problem scope

To achieve our goal of speaker identification research, we search for the optimal acoustic features that can represent the characteristics and naturalness of the speaker by studying the physical of the human speech production apparatus, which includes pitch frequency and power spectral density. We investigated the method of using statistical to represent the speaker model as well as detect a small set of possible speaker. Finally, we employed the decisions making algorithm to identification process, which allows us to modify the existing pattern scoring calculation to conduct a search for the most likely speaker. As the results, our research scope would cover only the topics mentioned above. Details of our speaker identification system and its design are given in chapter 3.

CHAPTER III

APPROACH

As mentioned in section 2.3, most of the problem found in the previous research is the system has no explicit representations of speaker characteristics. In this chapter, we will discuss the approaches to solve the problem.

When we analyze the production and perception of speech in human beings as shown in Figure 3.1, the speaker brain formulates the message and converts it into natural language code. Then, it executes the neuromascular commands to make the vocal tract vibrate and produce the acoustic waves. The listener processes the acoustic signal along the basilar membrane in the inner ear to produce a continuous spectral analysis of the signal. The neural transducer converts the output of the basilar membrane into characteristic patterns, and then transforms it into activity signal in the brain to recall who is the speaker.

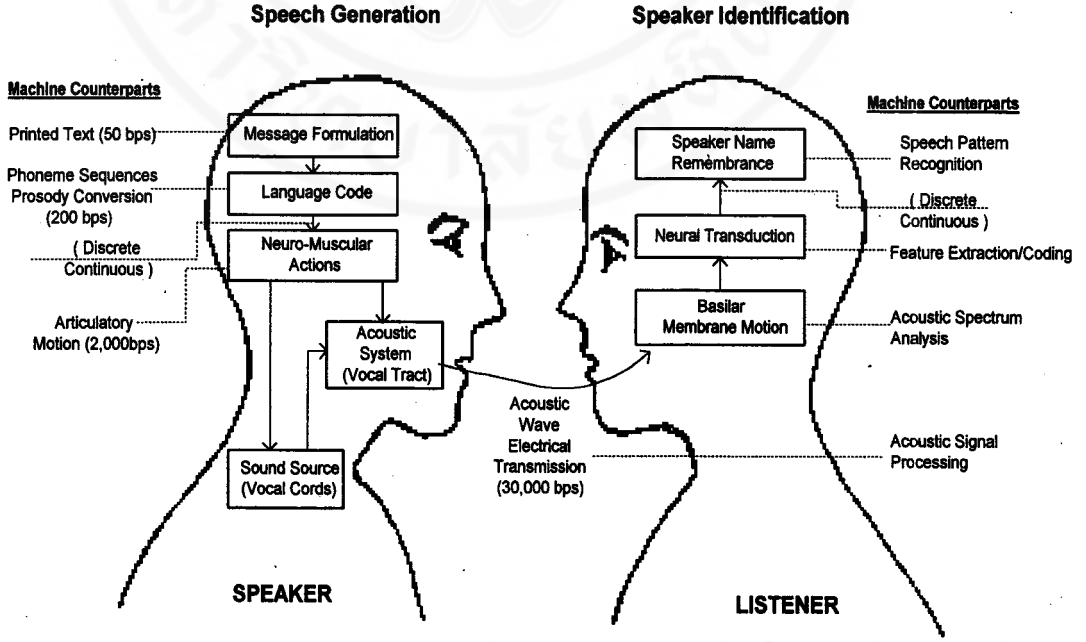


Figure 3.1 Schematic diagram of speech production and perception process [011]

However, the way to enable the machine to recognize speech as the human is a combination of different approaches. The speaker identification method of the listener as shown in Figure 3.1 leads our study to the following two different major recognition approaches:

1. Acoustic-phonetic approach
2. Pattern-recognition approach

3.1 Acoustic-phonetic approach [003,004]

In this approach, a speech signal is decoded to a digital signal before performing the analysis. Most small computer are able to digitize speech and store the sample sequence directly to disk. Commonly used appropriate formats for speech files are sampling at 10 kHz with 8 bit analog-to-digital converter. The major problem associated with this approach is that it requires extensive knowledge of the acoustic properties in term of obtaining the observed acoustic features as described below:

3.1.1 Recording conditions

The training speech data should be recorded with a highly sensitive microphone. The speakers should speak carefully and be closely monitored by the experimenter. The original speech is recorded as an audio wave file format directly onto the disk. Moreover, the speech should be recorded in a sound proof room with less environmental noise and the microphone positioned three inches from the speaker's mouth.

3.1.1 Acoustic spectrum analysis

The recorded speech was carried out in order to relate acoustic phonetic properties, which can be classified into 6 broad classes consisting of vowels, nasals, weak fricatives, strong fricatives, stops, and silence. Table 3.1 shows the phone distribution of 6 board classes.

3.1.1.1 Speech Perception

The human ear consists of three main sections, the outer, the middle and the inner ears. The outer ear consists of the ear lobe and the external auditory canal, which is to channel sounds into the middle ear. The middle ear adjusts the pressure levels between the outer and inner ear. Muscles attached in the middle ear will suppress the

vibration of the sound air if it is too violent and so protect the inner ear. This protection only works for sounds below 2kHz and it does not work for impulsive.

Table 3.1 Phone distributions of broad classes

Class	Phones (in English)	Phones (in Thai)
Vowels	aa,ay,ao,eh,ae,oy,ey,er,uh, ux,iy,ih,ow,r,el	อา อี อื อู เอ โอ
Stops	B,d,g,p,t,k	บ ด พ ท ก
Nasals	M,em,n,en,nx,ng,eng,dx,q	ม หม น ณ หน
Strong Fricatives	S,sh,z,zh,ch,jh	ส ศ ษ ซ ช ฉ
Weak Fricatives	F,th,dh,v,hh,hv	ฝ ฟ ห ฮ
Silence	Pcl,tcl,kcl,bcl,dcl,gcl,pau	ล พ หล

For the purposes of this study, we tried to measure the explicit acoustic characteristics of speech. As mentioned in section 2.2, vowel or voice source measurements were mentioned as good candidates for characterizing speakers' speech. Vowel sound properties correspond to the vocal tract resonance peaks which can be represented in terms of a small number of formants. The five formants highest in frequency can determine the characteristics of speech in terms of good quality, intelligibility, and naturalness of the speech result.

Figure 3.2 shows an example of the amplitude of different frequency components as indicated by the density of marking. Figure 3.3 shows the formants pattern of the vowels at the point marked with arrows in Figure 3.2.

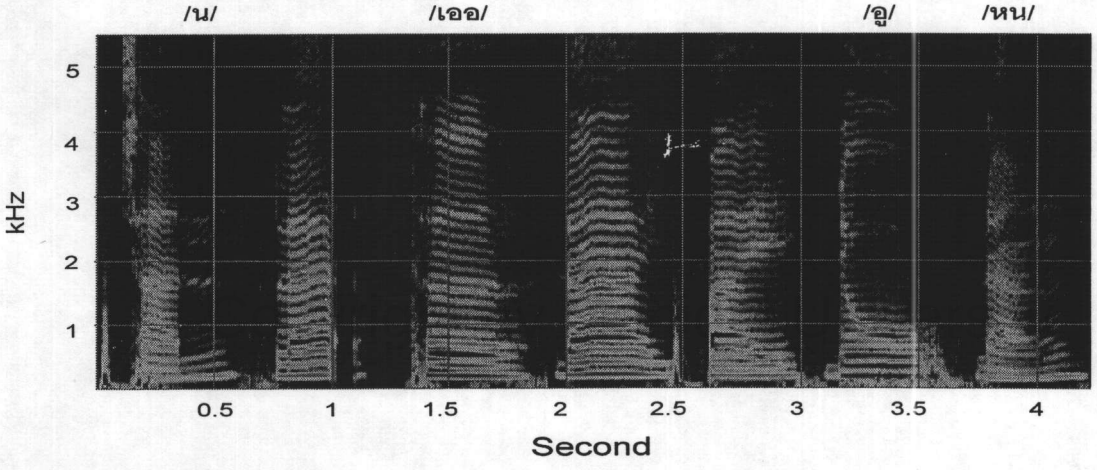


Figure 3.2 A frequency / time spectrogram for the utterance "ฉันและเธอได้ไปดูหนัง"

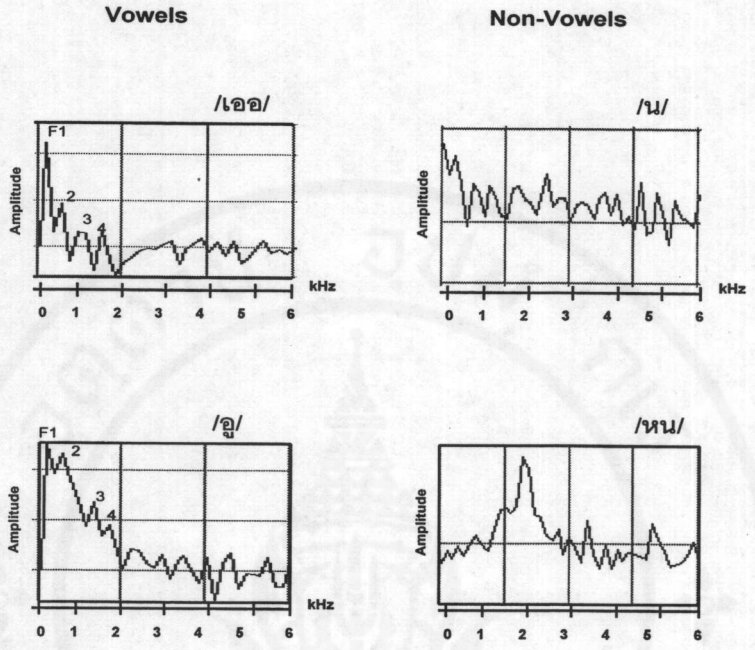


Figure 3.3 Formant frequency of vowels and non-vowels

3.1.2 Feature extraction/coding analysis [006]

Past observations have shown that the Power Spectral Density (PSD) is useful for speaker identification. The PSD is the real function of frequency, and the power of signal components is sometime a closer indication of the perceptual importance than the amplitude.

The PSD pattern corresponds to the estimation of the resonant frequency and their amplitude. To extract PSD from the speech waveform, there are many steps as follows:

- *Voiced/unvoiced detection:* Speech samples can detect voiced segments by using Short-Time Average Zero-Crossing Rate (ZCR). The ZCR can help determine whether speech is voiced. The difference between the voiced and unvoiced segments of speech is the rate at which the signal changes sign. High ZCR corresponds to unvoiced speech, while a low ZCR corresponds to the voiced speech.

- *Windowing:* Since we assume the signal is short-time stationary, before performing a Fourier transform on the speech segment, we multiple the signals by window function to taper smoothly down to zero at each end of

the range. Hamming window acts as a low pass filter. We then modulate the speech segment by Hamming Window of approximately 25 msec in duration.

- *Power spectral density:* Next the PSD of speech segment can be estimated with the discrete Fourier transform (DFT), and squared component-wise to obtain the power spectral density.

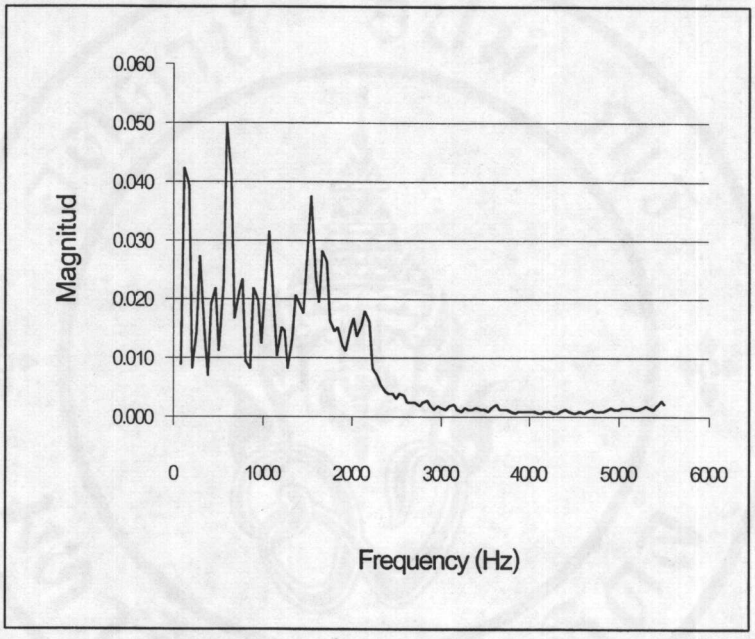


Figure 3.4 The power spectrum density

The details of the signal processing described above are summarized in the block diagram below:

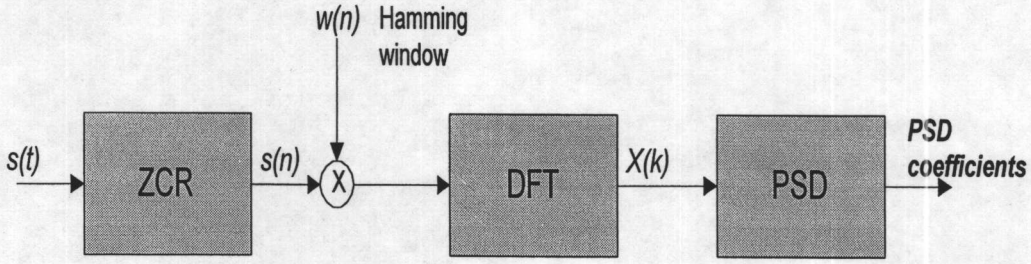


Figure 3.5 Block diagram for computing PSD

3.2 Pattern-recognition approach

This approach involved two parts: *training* and *recognition*. Training establishes a model or reference of speech patterns, while the recognition tries to find the most likely speaker by comparing the unknown pattern to all reference patterns.

Applying pattern recognition methods to speaker identification system involves several steps, including the method of detecting possible speaker, similarity comparison, and a decision. In this section, we will present classical statistical technologies and the decision approach applied in our pattern recognition area.

3.2.1 Maximum likelihood estimation

In the first step of this approach, the probability density functions (pdfs) is chosen to represent the distribution of the acoustic feature sub-spaces. During both training and testing sessions, the means and variances of the first sub-spaces of each speaker acoustic feature are calculated using the maximum likelihood (ML) estimation procedure.

3.2.2 Clustering

To keep the computation efficient before performing pattern matching, we approached the pre-detected possible set of speakers who are acoustically similar to the unknown speaker. It is reasonable to assume that the probability distribution of the similar acoustic feature sub-spaces will be raised into the same class. Therefore, the combination of clustering and distance measure is chosen to pre-detect a set of possible speakers. The algorithm has two steps:

- *Step1-the scoring:* The likelihood probability scores are used to cluster all speaker data models. The scores are then sorted for each speaker, and the speaker in the range of pre-defined threshold will be accepted.

$$p_a \in SP_j \quad \text{if} \quad S\text{-}N \geq p \geq S\text{+}N \tag{3.1}$$

Where SP_j is the likelihood probability j of the speaker model, S is the likelihood probability score of the unknown speaker, and N is the predefined threshold we varied in the experiment.

- *Step2-the distance measuring:* The searching for possible speakers is repeated by calculating the distance between two speakers using Mahalanobis distance metric. This metric is applied to all possible speakers generated from the previous step. The calculated distances are then sorted, and the speaker is accepted if his/her distance is in top N scores.

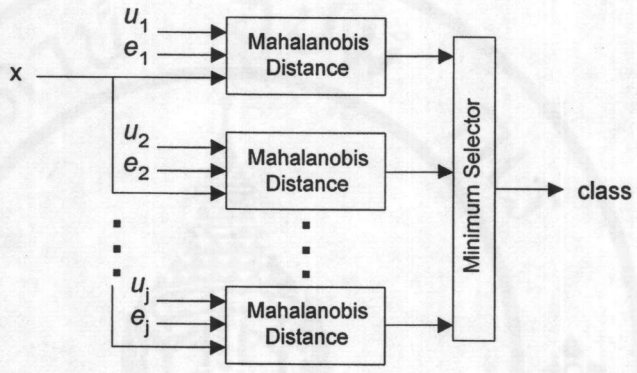


Figure 3.6 Mahalanobis computing procedure

3.2.3 Pattern similarity comparison and identification

The last step of this approach, the square-error pattern matching, will be performed to detect the last small set of speakers, called claimed speakers, who have the speech pattern similar to the test speaker. All found N nearest neighbors (claimed speakers) are added to the list called the cohort set. Speakers outside the cohort set are considered outliners that have low probabilities of scoring well and will be eliminated before getting into the square-error pattern matching process. Also, the number of nearest neighbors in the cohort set can be reduced if their number of claimed is less than a pre-defined threshold.

CHAPTER IV

SYSTM DESIGN AND IMPLEMENTATION

In this chapter, we will first give an overview of the methodology used for implementing the speaker identification model in Section 4.1. Section 4.2 described the system model, which consists of the feature extraction, speaker search, and identification process. Finally, the components and environment for implementation of the system model are presented and shown on the flow chart given in Section 4.3.

4.1 System overview

We will divide the components of our model into four parts as shown in Figure 4.1 the block diagram of the speaker identification system model. The input speech is sent digitally under pre-defined conditions to the feature extraction subsystem to obtain acoustic features. These features will be sent to the speaker search subsystem to search for the possible set of speakers against the statistical-based speaker model. Next, the speech pattern of the possible speaker is used to perform the matching against the unknown speaker data to find the cohort set. Finally, the square-error scores are used to classify speakers, and to make an identification decision. The details of each part are given in the next section.

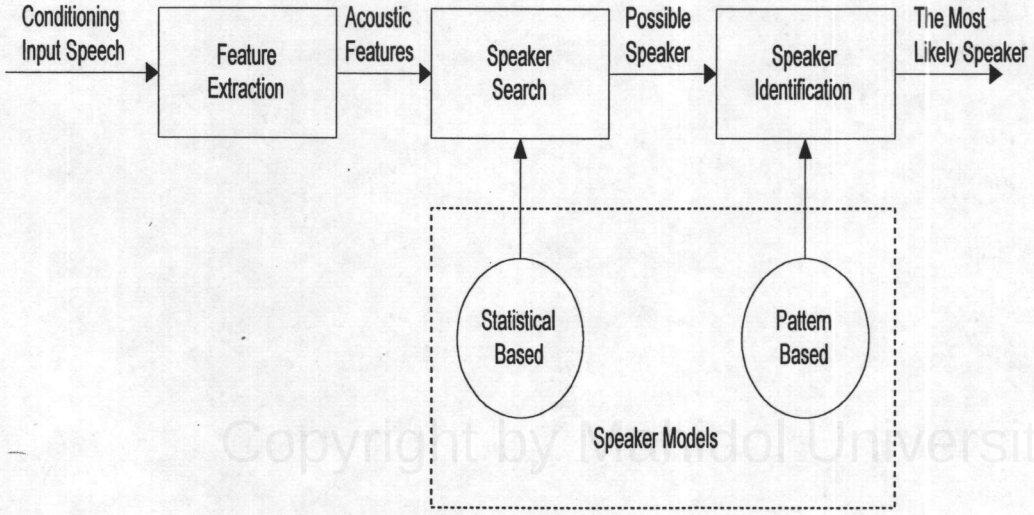


Figure 4.1 Block diagram of the implementation model

4.2 System design

In the overview, we briefly described the system model of speaker identification. In this section, we will breakdown each part and discuss it in detail. We will describe the design of the speech feature extraction, the speaker search, the identification process, and the details of how to develop the speaker models.

4.2.1 Features extraction [3,8]

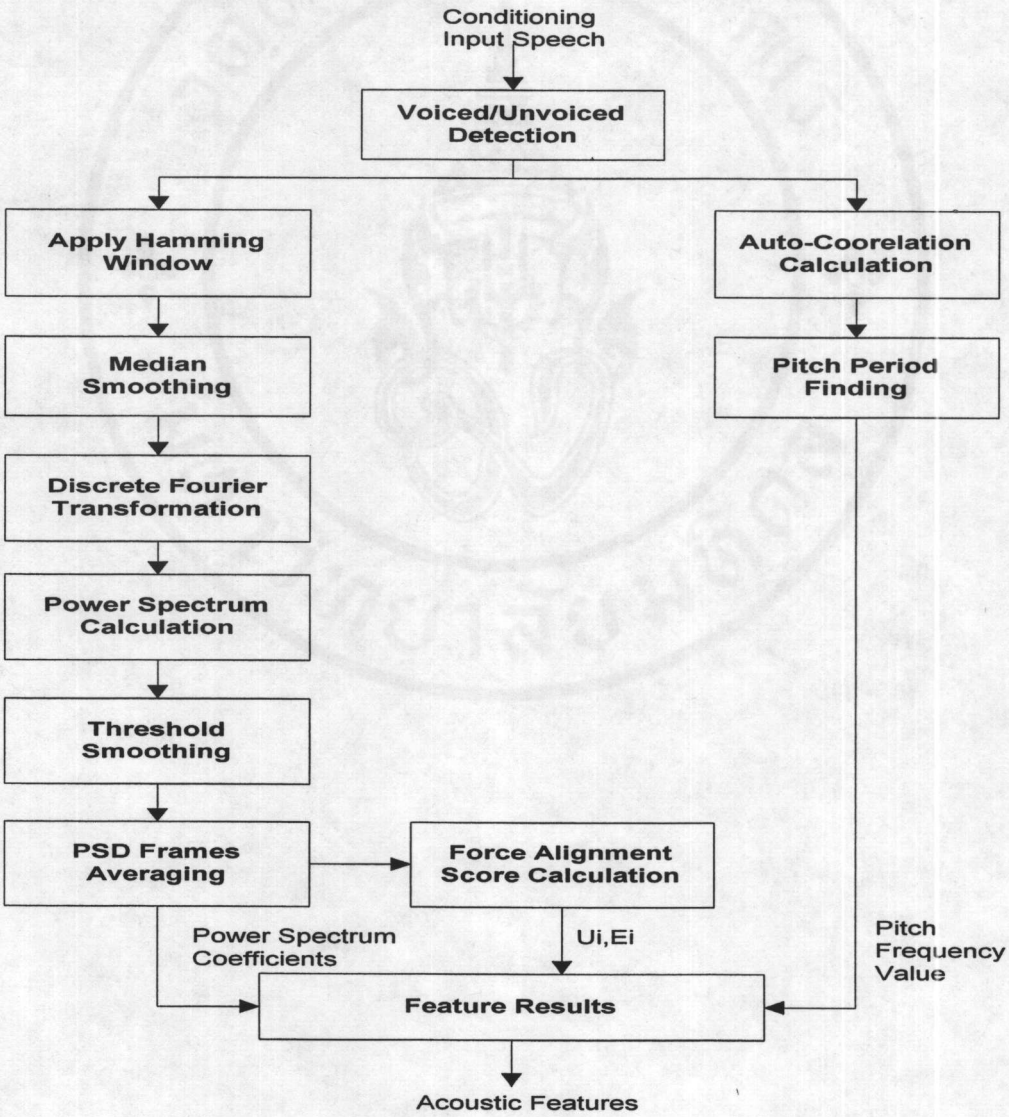


Figure 4.2 Speech features extraction block diagram

As shown in the diagram of the feature extraction process given in Figure 4.2, the input speech is recorded under the pre-defined conditions. The process produced the output acoustic features which consist of power spectrum coefficients and pitch frequency value. Each subsystem of feature extraction process is described as follows:

4.2.1.1 Conditioning input speech

The input speech is raw wave data recorded using Sound Blaster card. The operational details used for conversion of the data to digital format are as follows:

Number of bits per sample: 8
Mode of operation : Mono mode
Sampling rate : 11.025 kHz

4.2.1.2 Voiced/unvoiced detection

Input : Input speech
Output : Voiced speech data
Description : Use Short-Time Average Zero-Crossing Rate (ZCR) formula to find the voiced segment, which the rate of signal changes sign is about 0.5 crossing/ms.
Details : The ZCR or $Z(n)$ defined as:

$$Z(n) = \frac{1}{N} \sum_{m=1}^{N} |sign(x(m)) - sign(x(m-1))| \quad (4.1)$$

Where $$sing(x(m)) = \begin{cases} 1, & \text{if } x(m) \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

N = Number of samples
n = The n^{th} ZCR frame

4.2.1.3 Hamming window

Input : Voiced speech data
Output : Segmented voiced speech data
Description : The speech samples are modulated by a Hamming Window of approximately 25 msec in duration.
Details : The most common in speech analysis is the Hamming Window:

$$W(n) = (0.54 - 0.46)\cos\left(\frac{2\Pi n}{N-1}\right) \qquad 0 \leq n \leq N-1 \tag{4.2}$$

Where N = DFT length

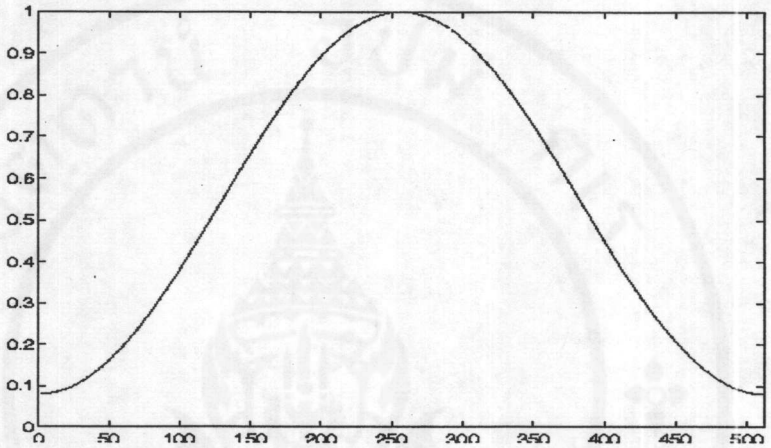


Figure 4.3 The Hamming window

4.2.1.4 Median smoothing

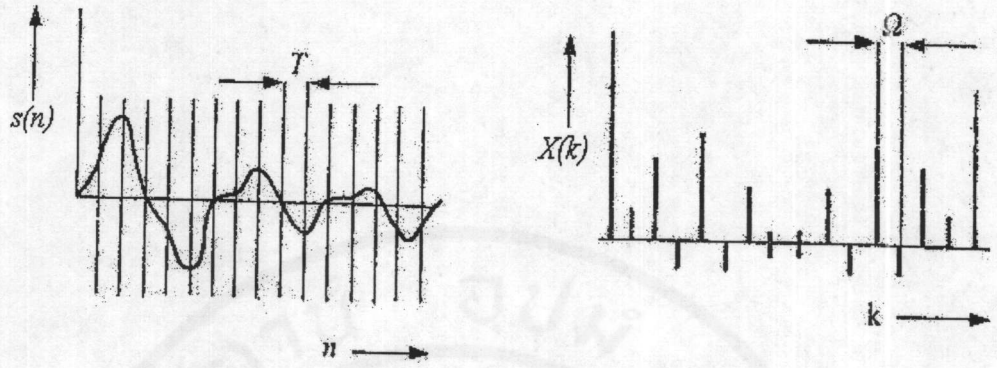
Input : Segmented voiced data
Output : Smoothed voiced data
Description : Convert time series signal to be the data point scale. Then subtract out the mean value from the signal.

4.2.1.5 Discrete Fourier transformation

Input : Voiced speech data
Output : 256 points DFT
Description : Perform discrete Fourier Transform (DFT) of the segmented interval of speech data.
Details : The DFT defined as:

$$X(k) = \sum_{n=1}^{N-1} s(n)\ e^{-j2\Pi n k / N} \tag{4.3}$$

Where a set of N sample values is indexed by n in the frequency domain, which is the transformation of a periodic signal sampled in the time domain with index k.



Sampled waveform Discrete Fourier transform

Figure 4.4 Discrete Fourier transform of a periodic signal

4.2.1.6 Power spectrum calculation

Input : 256 point DFT
Output : Power spectrum frames
Description : Each DFT frame was squared component-wise to obtain the power spectrum density.
Details : The PSD is defined as:

$$PSD(k) = Sqr(X(k)) \quad (4.4)$$

4.2.1.7 Threshold smoothing

Input : Power spectrum frames
Output : Normalized power spectrum frames
Description : Smooth the curve of each power spectrum coefficient by normalizing it to fit in the maximum magnitude of the threshold 0.05.

4.2.1.8 Power spectrum frames averaging

Input : Normalized power spectrum frames
Output : A frame of power spectrum coefficients
Description : Average the N frames of power spectrum whose values are in the frequency range 0-4000Hz.

4.2.1.9 Forced alignment score calculation

Input : First Sub-band of PSD
Output : Forced Alignment Scores (S)
Description : Use the first sub-band of power spectrum of unknown speaker to estimate the scores using Maximum Likelihood (ML) procedure.

4.2.1.10 Auto-Correlation calculation

Input : Voiced speech data
Output : Auto-correlation speech segment
Description : The Auto-correlation is the correlation of the waveform with itself. It is the repetition rate of pulses produced when the vocal cords vibrated, which can determines the pitch period. To determine the fundamental pitch period, uses a voiced speech segment of approximately 40 msec. Then compute the correlation between one sample and a few of its predecessors. The high correlation shows the peaks in the auto-correlation function.
Details : The Auto-correlation function C(k) is define as:

$$C(k) = \sum_{m=1}^{N-1} s(m)\ s(m+k) \qquad (4.5)$$

Where samples is sequenced from s(0) up to s(m-1+k) and k is number sample of delay.

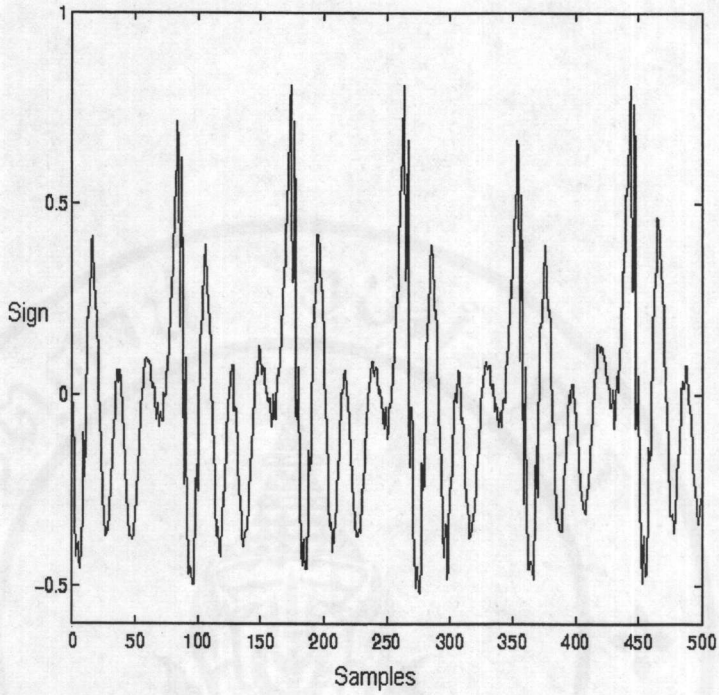


Figure 4.5 The Auto-correlation function

4.2.1.11 Pitch period finding

Input	: Auto-correlation curve
Output	: Pitch Frequency
Description	: Calculate to find the distance between the most significant maximum of the signal to compute the pitch frequency, while the process also calculate to reject numerous smaller peaks in each cycle.
Details	: For estimation of pitch period, there are four steps as follows:

1. *Cube* - Each data point is cubed to accentuate large values and attenuate small values. This odd power is used to preserve the sign of the data.

2. *Threshold clipping* - We considered only the data points rising above the threshold region. Then we smoothed the clipped data by finding only the maximum data point for each period.

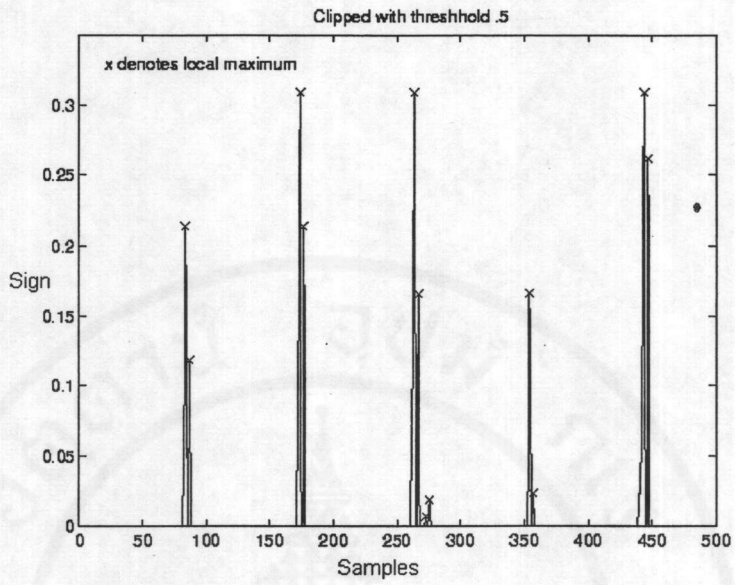


Figure 4.6 Threshold clipping

3. *Local maximum finding* - We compute the number of samples between each local maximum and those distance which are greater than one standard deviation from the me an distance will be thrown out.

4. *Calculate pitch frequency* - The distance between two significant peaks is obtained in msec and converted to frequency as:

   Pitch frequency(Hz) = (1/pitch period) * 1000

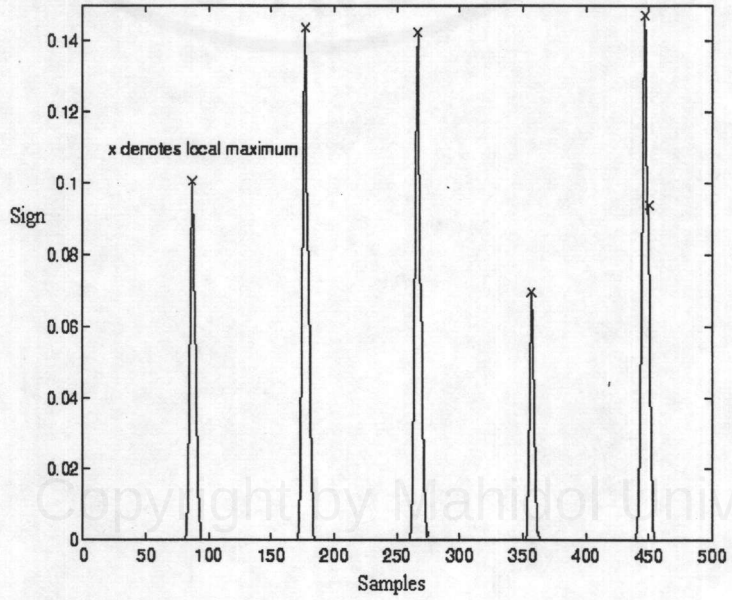


Figure 4.7 Peak smoothed

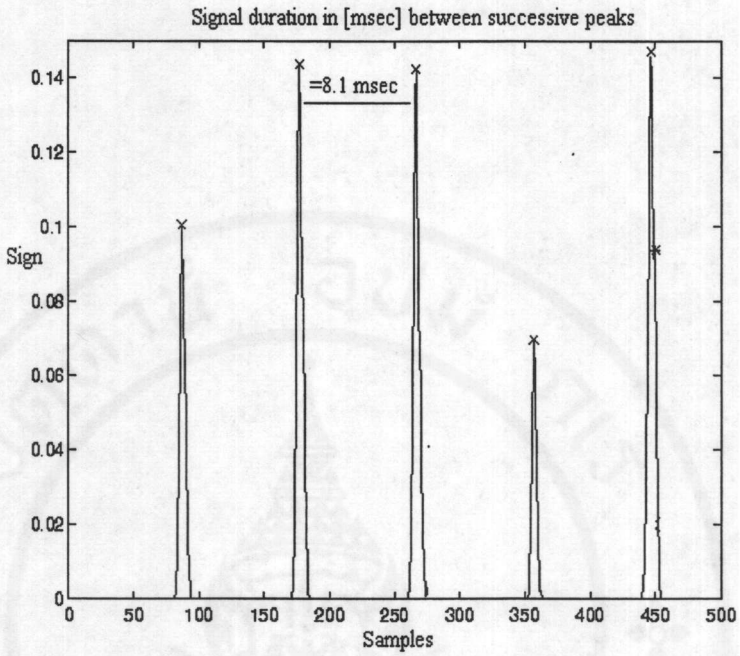


Figure 4.8 Pitch period

4.2.2 Speaker Search [10]

As shown in Figure 4.8, the pitch frequency is used to select male or female sub data of the statistical speaker model. The forced alignment scores (Ui, Ei) are then used to pre-detect the possible speakers, the speaker is accepted if his/her score is in the S±N results. After that, Mahalanobis distance is calculated once again against the pre-detected possible speaker scores to post-detect the set of speakers. Each subsystem of speaker search process is described as follows:

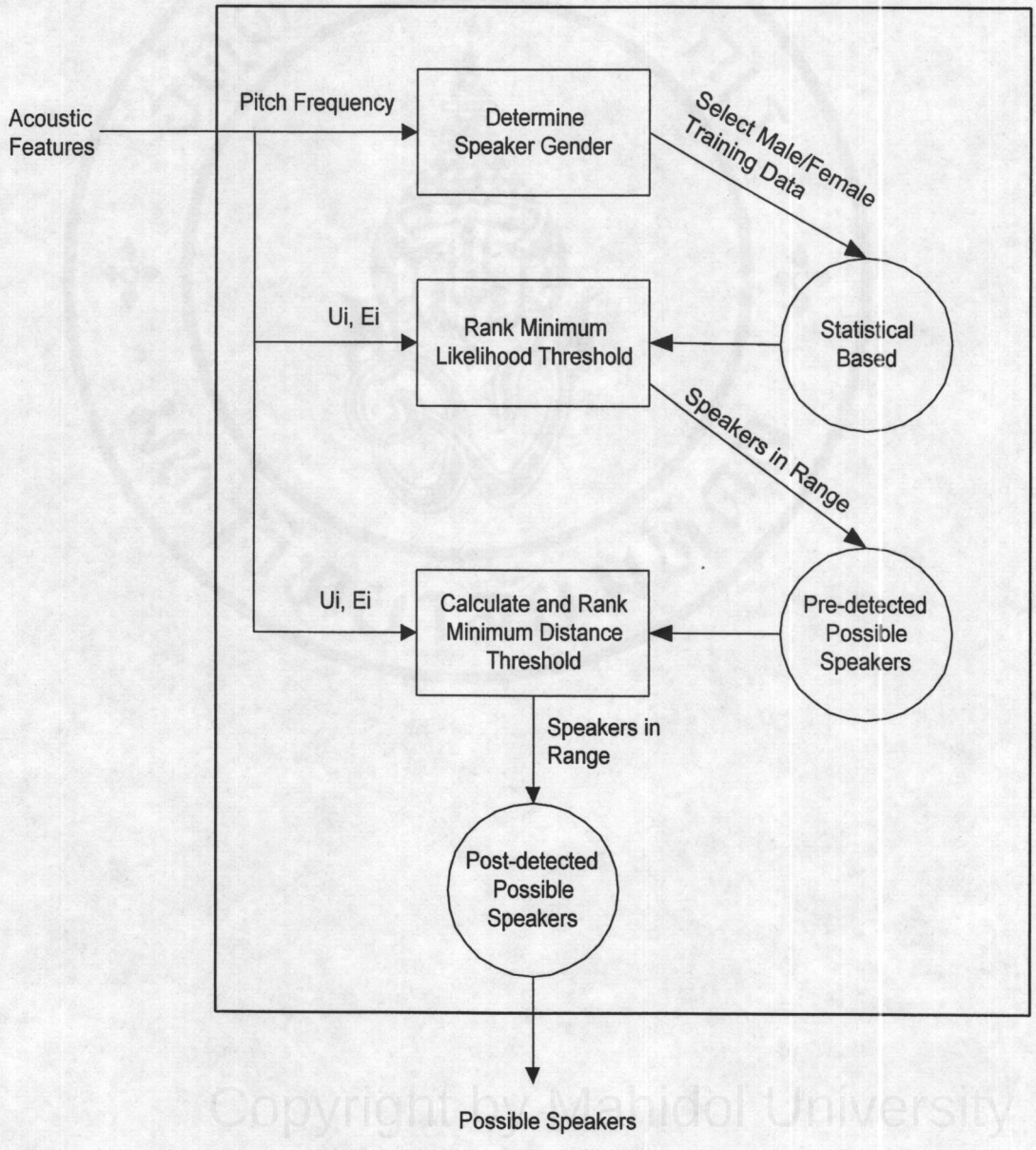


Figure 4.9 Speaker search block diagram

4.2.2.1 Determine speaker gender

Input : Pitch Frequency
Output : Gender Flag
Description : Determine the speaker gender using pitch frequency. Then a gender flag is produced to select type of speaker model (male or female or both).

4.2.2.2 Rank minimum likelihood threshold

Input : Forced Alignment Scores (S)
Output : Pre-detected Possible Speakers
Description : The score of the speaker model is sorted, and the speaker whose score is in S±N will be accepted.
Details : The estimation of Maximum Likelihood is defined as:

$$u_j = \frac{1}{n_j} \sum_{k=1}^{k_j} x_{j,k} \tag{4.6}$$

$$e_j = \frac{1}{n_j} \sum_{k=1}^{n_j} (x_{j,k} - u_j)^2. \tag{4.7}$$

Where n is the total number of the feature vectors of the sub-space
j is the j^{th} sub-space
$x_{j,k}$ is the k^{th} feature vectors for sub-space j
u_j is the ML estimate of the mean for sub-space j
e_j is the ML estimate of the variance for sub-space j

4.2.2.3 Calculate and rank minimum distance threshold

Input : Forced Alignment Scores (S)
Output : Post-detected Possible Speakers
Description : The Mahalanobis distance is calculated against the pre-detected possible speaker scores. The speaker is accepted if the score is in the top N of the minimum distance value.
Details : Mahalanobis distance estimation (D^2)

$$r^2 = (x - u_j)' \overset{-1}{e_j} (x - u_j) \quad (4.8)$$

Where u is the mean vector of feature spaces j, x is the mean vector of unknown speaker feature space, and e is the variance j for two speakers.

4.2.3 Speaker identification

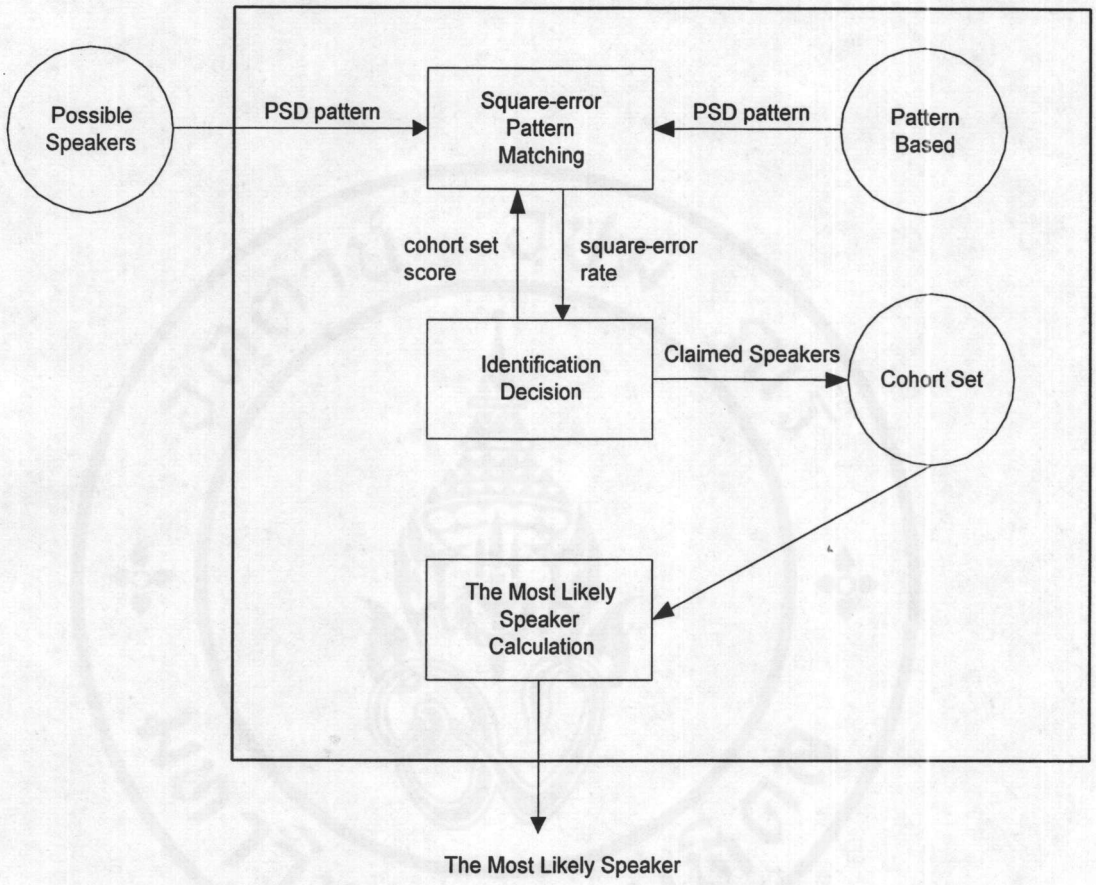


Figure 4.10 Speaker identification block diagram

As shown in Figure 4.9, the square-error pattern matching is performed to compare each possible speaker pattern to the unknown speaker pattern. The accepted speaker is the one who scores best against the unknown speaker data called the claimed speaker. At the same time, the identification decision process is also computed to eliminate the speaker with a low probability score. In this way, only high probability speakers can get into the pattern matching process to keep our computation inexpensive and to reduce number of claimed speakers which we added to the list called the cohort set.

The last step of identification is the most likely speaker selection which very important. We implemented a technique called the speaker identification decision. For each test utterance, 10 utterances were tested one after from another. After the square-error pattern matching, we detected the last small set of speakers, called claimed speakers, who are the speech pattern similar to the test speaker. All found N nearest neighbors (claimed speakers) are added to the list called cohort set. As the results, we

need to obtain the most likely who have acoustic best match to the test speaker. However, computation became more expensive as a number of nearest neighbors are added to the cohort set. In designing a decision algorithm, through the first three testing utterances we will keep all of those claimed speakers in to the cohort list. Then the fourth test utterance and so on, speakers outside the cohort set are considered outliners that have low probabilities of scoring well and will be eliminated out before getting in to the square-error pattern matching process. And we usually reduce number of nearest neighbors in cohort set if their number of claimed less than a threshold i.e.

$$y \in C_i \quad \text{if count}(R_y) \geq U_n * (1/x) \tag{4.9}$$

Where C_i is the cohort at testing *i* and *x* is the parameter we varied in the experiment. We also propose the use of a test statistic $S_n(y;k)$ for searching the most likely speaker as:

$$S_n(y;k) = \text{Max} \frac{1}{U_n} \sum R_{yk} \tag{4.10}$$

Where R_{yk} is an observation correct rate at test utterance *k* for nearest neighbor *y*, and U_n is the total number of tested utterances. The function $S_n(y;k)$ will identify a speaker that attains the highest correct rate score among all competing speakers.

Each subsystem of speaker identification process is described as follows:

4.2.3.1 Square-error pattern matching

Input : PSD Patterns
Output : Square-error Rate
Description : Perform Square-error pattern matching to compare each possible speaker pattern to unknown speaker pattern.
Details : Square-error calculation is defined as:

$$SE = \sum_{i=1} (x_i - \nu_i)^2 \tag{4.11}$$

Where ν_i is a column vector of sample data
and σ_i is the i-th diagonal element of Σ

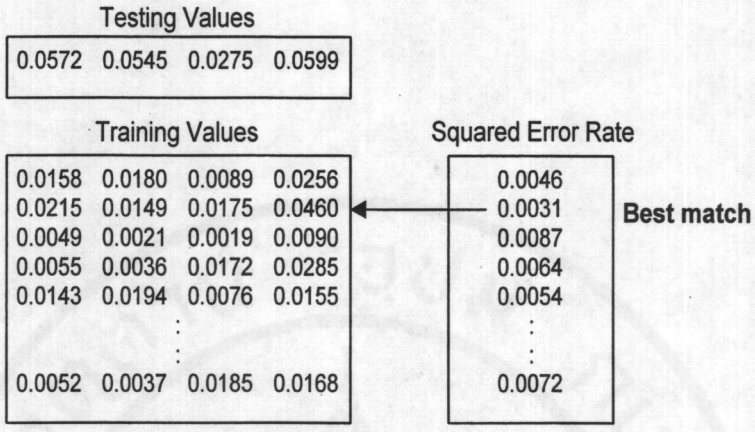


Figure 4.11 Square-error calculation example

4.2.3.2 Identification decision

Input : Square-error rate
Output : Cohort Set Score and Claimed Speakers
Description : Calculate the score of each speaker in cohort set to eliminate low probability speaker and to reduce the overhead of square-error pattern matching process.
Details : Identification decision estimation is defined as:

$$y \in C_i \quad \text{if } \text{count}(R_y) \geq U_n / (1/x) \tag{4.12}$$

Where C_i is the cohort at testing *i* and *x* is the parameter we varied in the experiment.

4.2.3.3 The most likely speaker calculation

Input : Cohort Set Scores
Output : The Most Likely Speaker
Description : Calculate the statistical value of speaker in cohort set for searching the most likely speaker.
Details : The propose statistic calculation as:

$$S_n(y;k) = \text{Max} \sum \frac{R_{yk}}{U_n} \tag{4.13}$$

Where R_{yk} is an observation correct rate at test utterance *k* for claimed speaker *y*
and U_n is the total number of tested utterances

4.2.4 Statistical Based Training Samples

Speaker models consist of a statistical-based model and a pattern-based model. We choose to represent the probability density values for statistical models because they do not require many parameters to train and accurately estimate the possible speakers. Power spectrum coefficients were implemented for the pattern-based model because this feature fits well with the nature of speech characteristics.

The probability density that models the acoustic features for each speaker are developed using maximum likelihood (ML) estimation procedure as:

$$\mu_j = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{j,k} \tag{4.14}$$

$$e_i = \frac{1}{n_j} \sum_{k=1}^{n_j} (x_{j,k} - \mu_j)^2 \tag{4.15}$$

Where n is the total number of the feature vectors of the sub-space
j is the j^{th} sub-space
$x_{j,k}$ is the k^{th} feature vectors for sub-space j
u_j is the ML estimate of the mean for sub-space j
e_j is the ML estimate of the variance for sub-space j

4.3 Speaker identification implementation

4.3.1 System components

The system was designed to operate in a Microsoft operating system environment on a stand-alone personal computer. Sound Blaster card used to record the input speech, as well as the use of C and Visual Basic as the programming languages for implementing the system model.

4.3.2 Speaker model initialization

The database used in this system consists of 1,000 training sample utterances spoken by 47 speakers, 22 females and 25 males. Each speaker uttered a set of 10 words, 4 time repeated. The Thai words were "ฉัน และ เธอ ได้ ไป ดู หนัง ที่ ดี มาก". Figure 4.12 shows the example speech waveform of all words spoken in Thai. Two utterance sets of each speaker were used for creating Speaker Models and the others two sets were for testing.

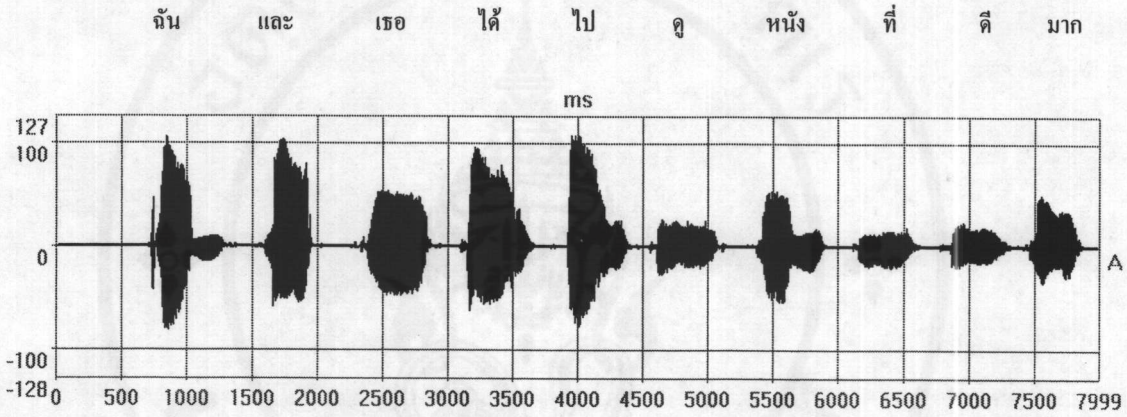


Figure 4.12 Speech waveform of male

During training, speaker models are consisted of pattern based models and statistical models. To obtain the pattern based models, all utterances were separated word by word. Speech samples were initially modulated by Hamming window of approximately 25 msec in duration and the discrete Fourier transform (DFT) of the modulated speech interval was computed. Then, each of the coefficients was squared component-wise to produce the power spectral density (PSD or energy) for each frame. Thereafter, the normalization of the PSD was computed and the results of 128 point spectral coefficients were obtained. Finally, the statistical models of the first sub-brand (84 – 1000 Hz) of pattern based models were developed for each speaker using the maximum likelihood (ML) estimation procedure.

An example of the statistical model developed using the ML procedure is shown in Table 4.1 and Figure 4.13 illustrates a graph of the speaker's pattern data.

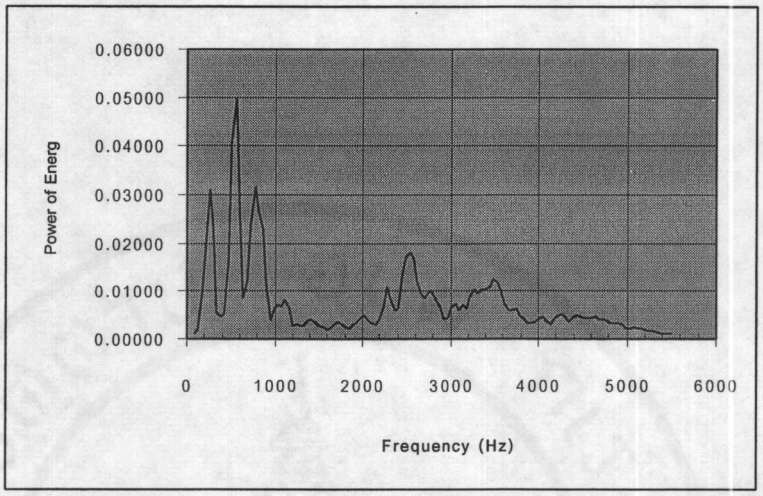


Figure 4.13a Speech pattern of the word "เธอ" of a female

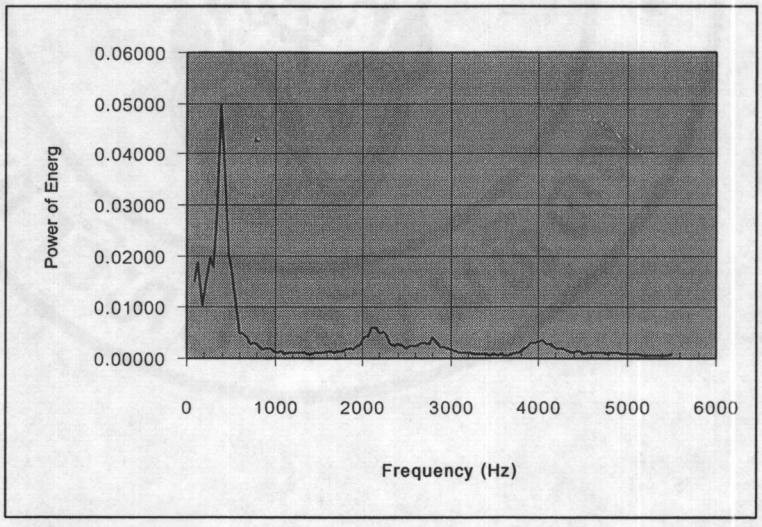


Figure 4.13b Speech pattern of the word "เธอ" of a male

Table 4.1 Example values of statistical model

Sex	Word	Mean	Variances
Male#1	เธอ	0.023119	0.000157
Male#1	ดี	0.014266	0.000238
Male#2	เธอ	0.016402	0.000178
Male#2	ดี	0.012766	0.000190
Female#1	เธอ	0.018258	0.000209
Female#1	ดี	0.013051	0.000189
Female#2	เธอ	0.013022	0.000132
Female#2	ดี	0.013700	0.000183

4.3.3 System iplementation

This section will present the flow chart of each module from 4.2 as follows:

4.3.2.1 Voiced/unvoiced detection

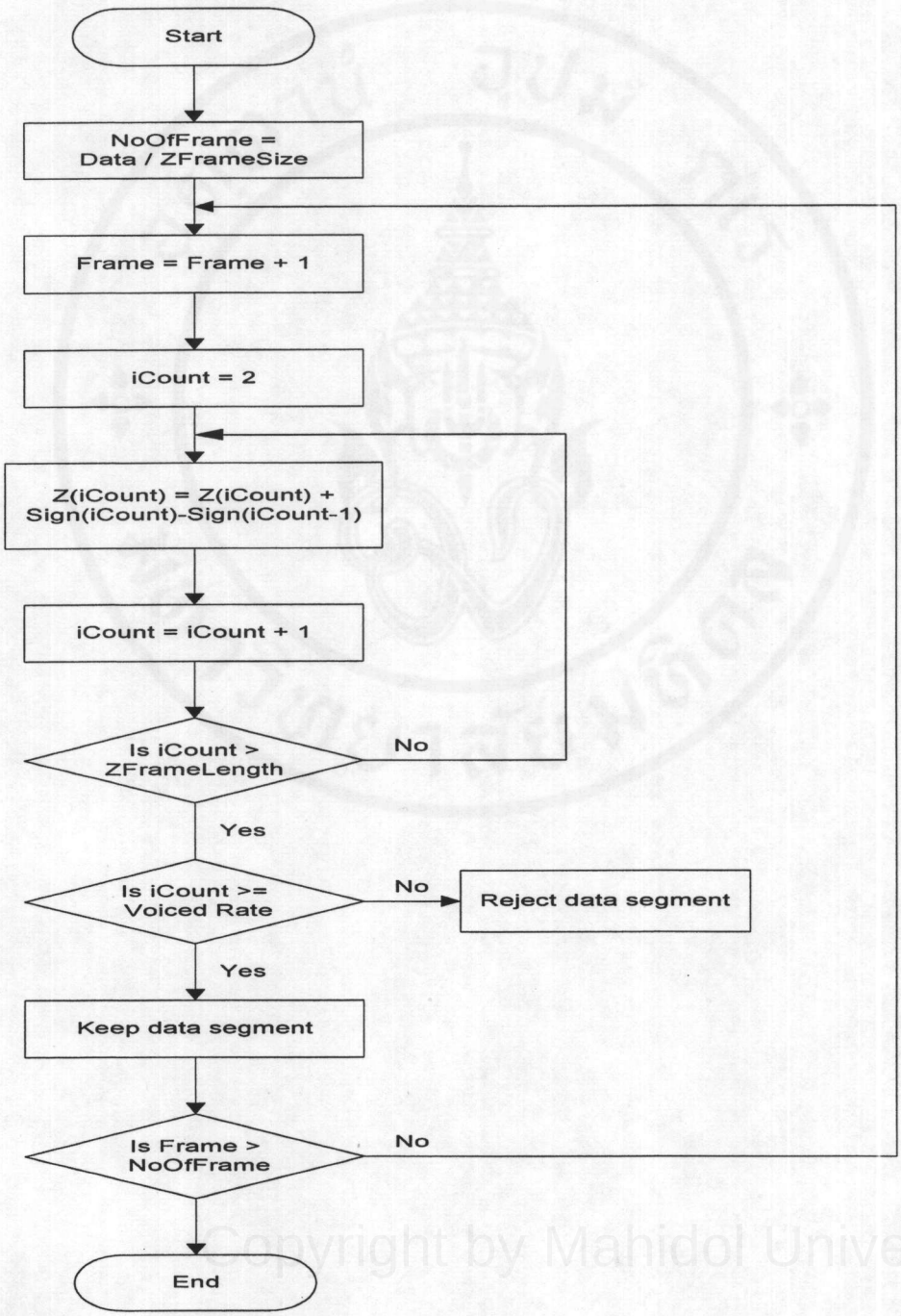


Figure 4.14 Flow chart of voiced/unvoiced detection

4.3.2.2 Hamming window and median smoothing

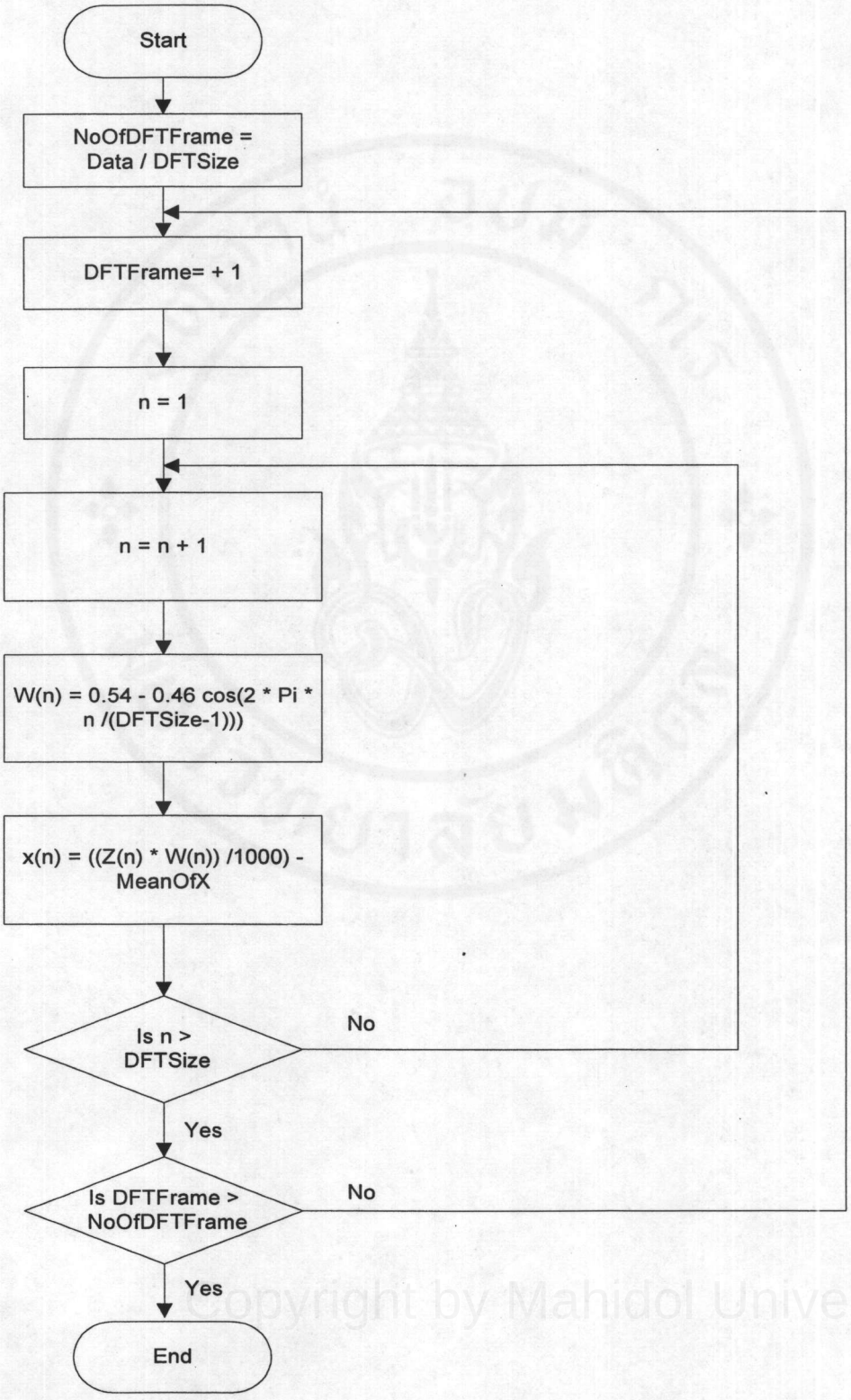
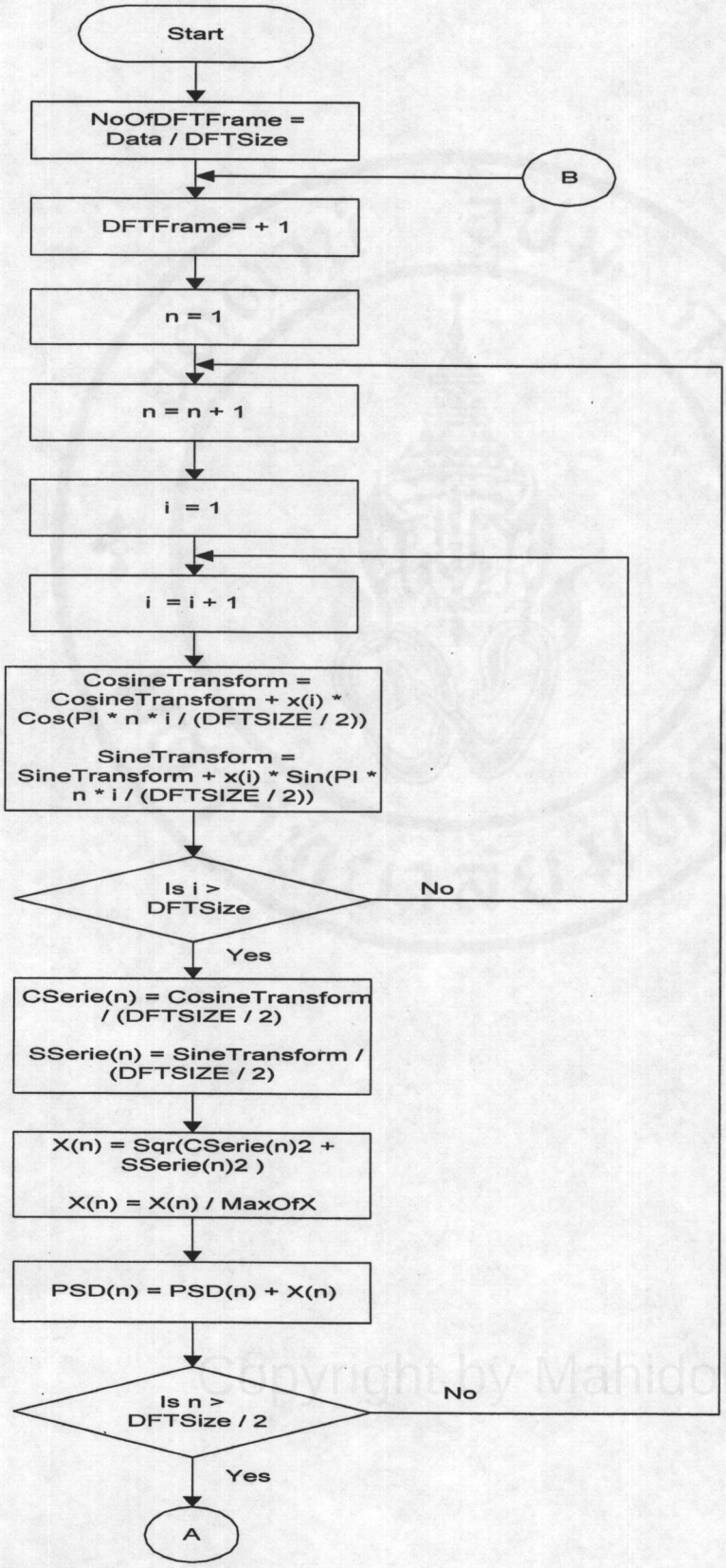


Figure 4.15 Flow chart of Hamming window and median smoothing

4.3.2.3 Power Spectrum calculation



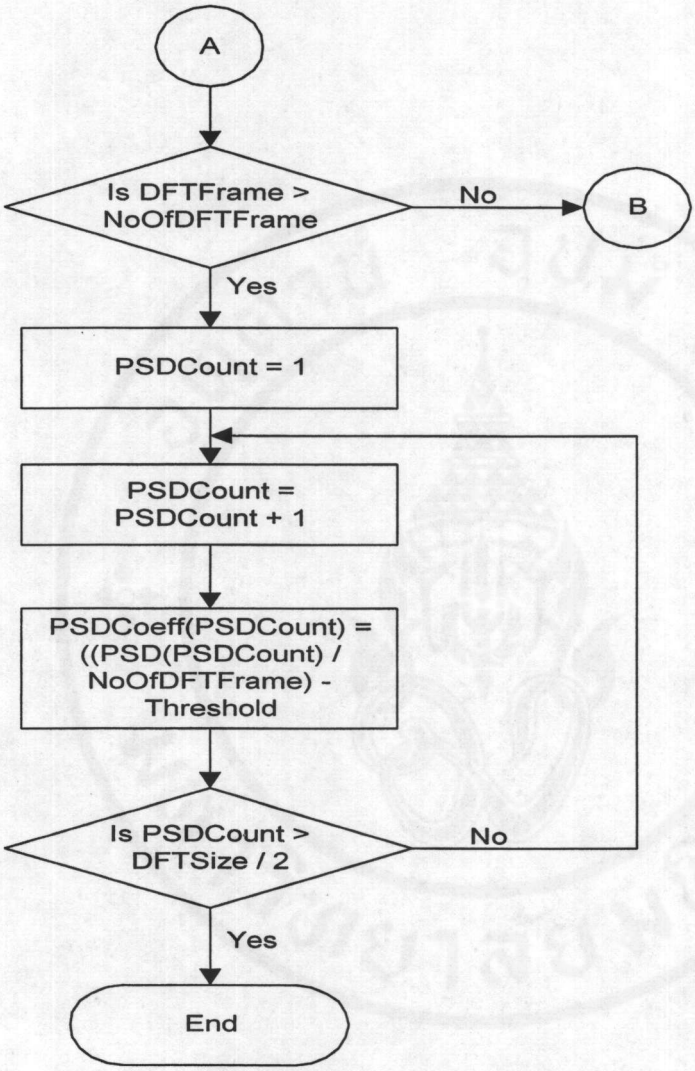


Figure 4.16 Flow chart of Power Spectrum calculation

4.3.2.4 Maximum Likelihood calculation

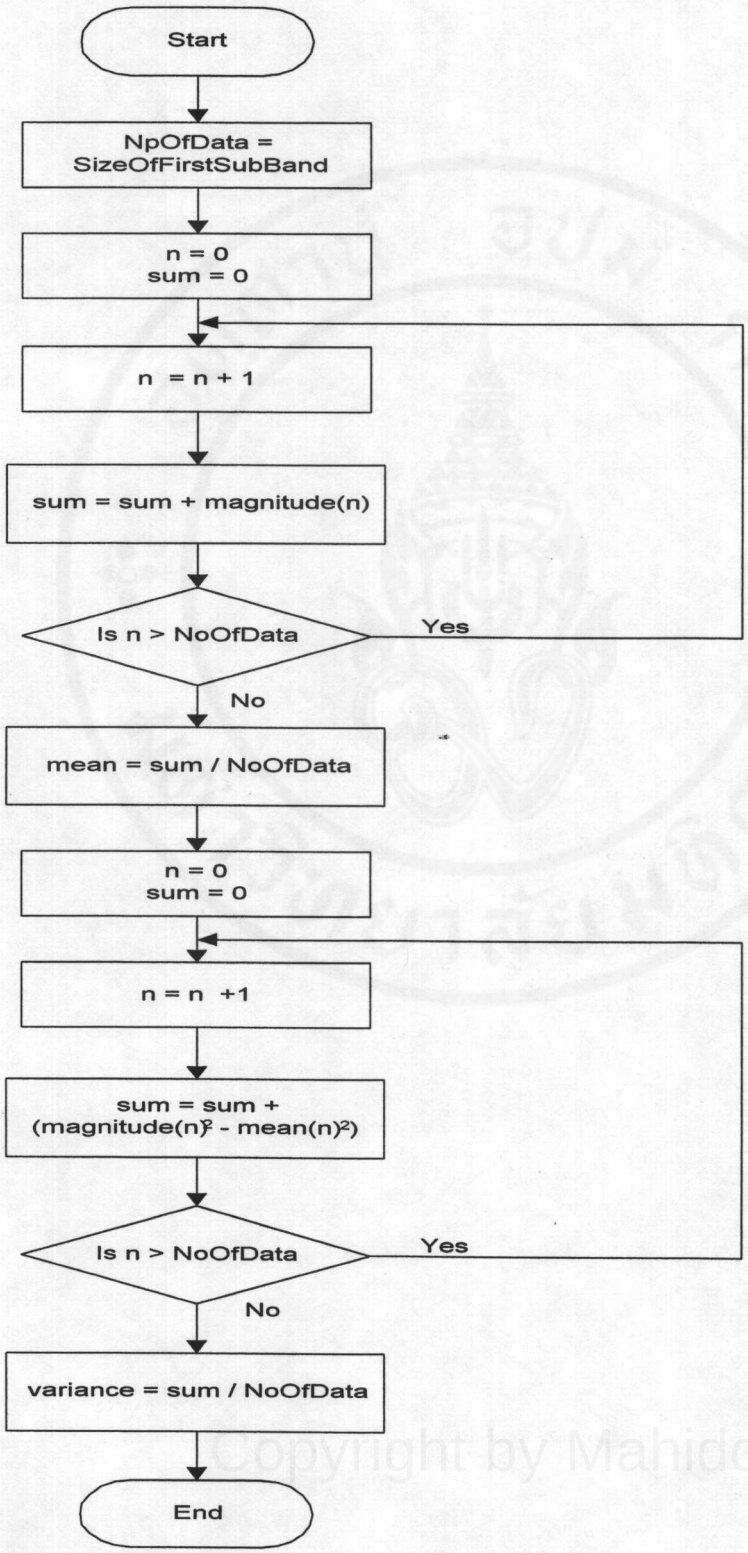
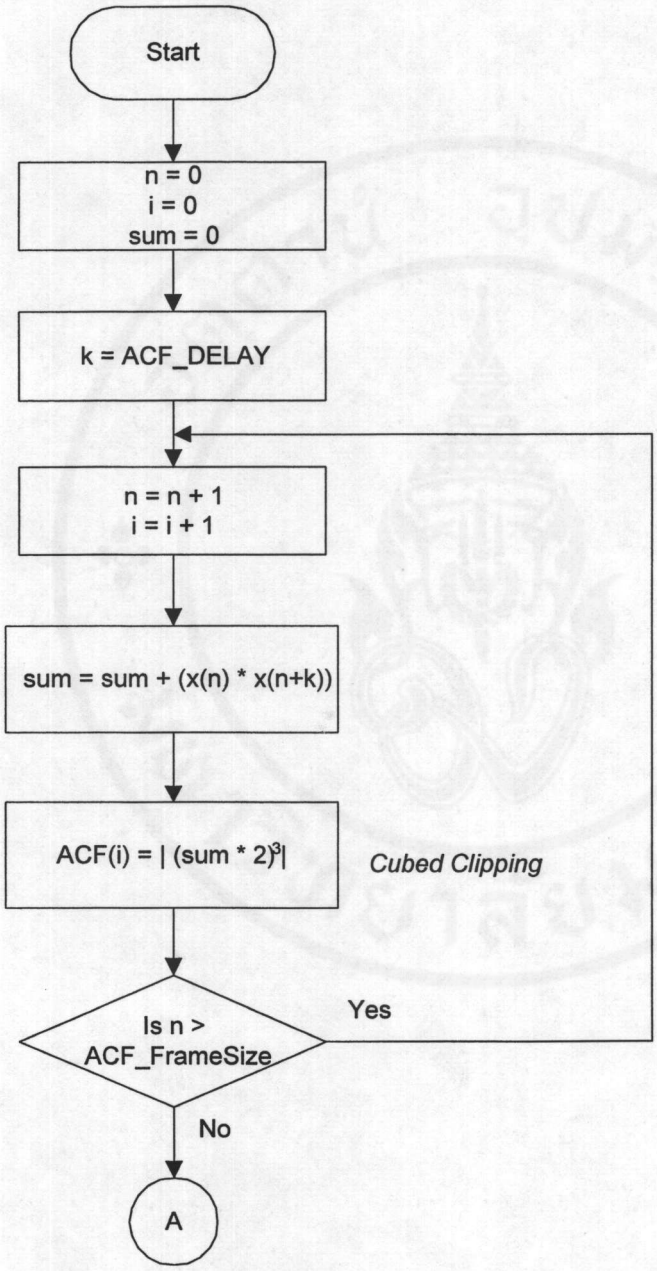


Figure 4.17 Flow chart of Maximum Likelihood calculation

4.3.2.5 Auto-Correlation calculation



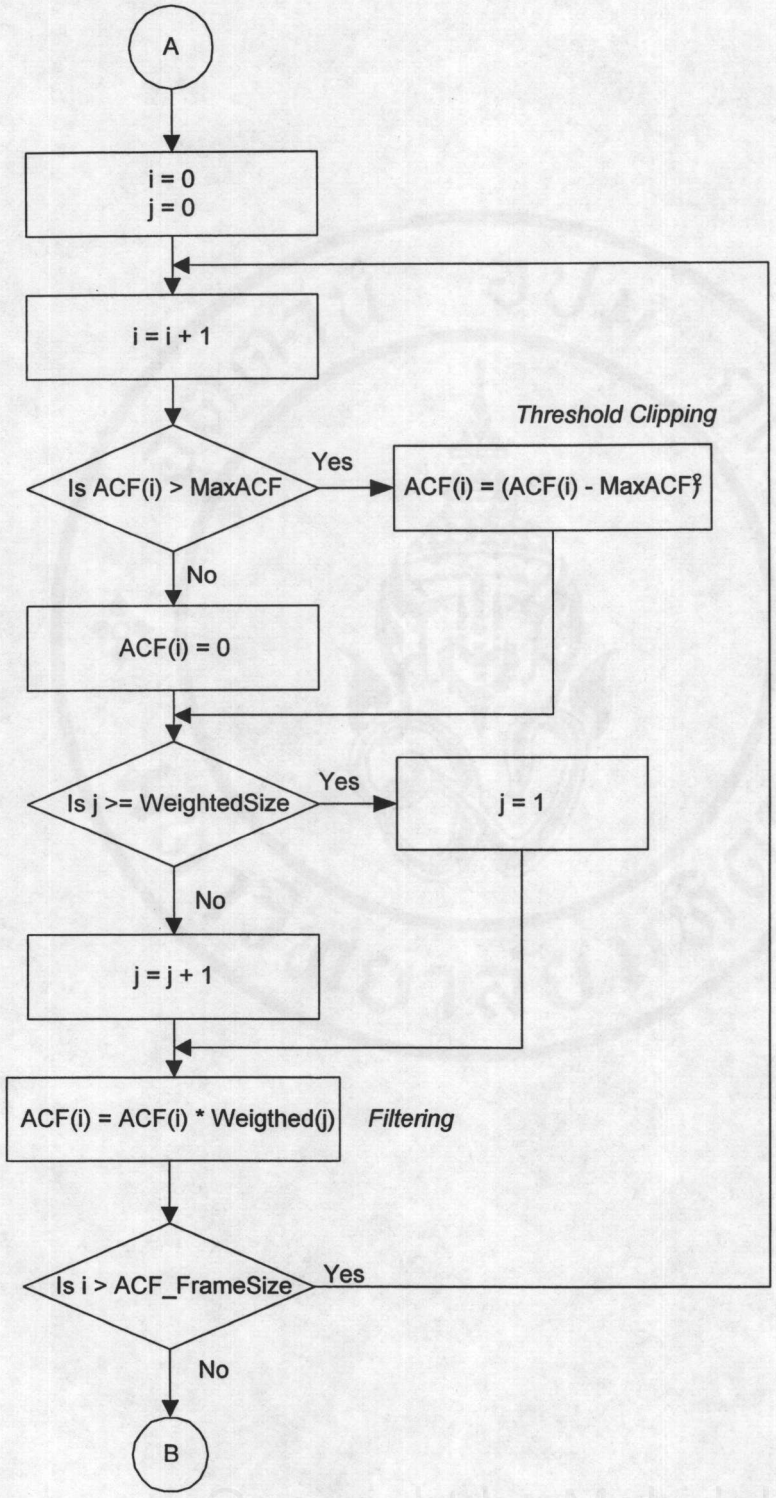
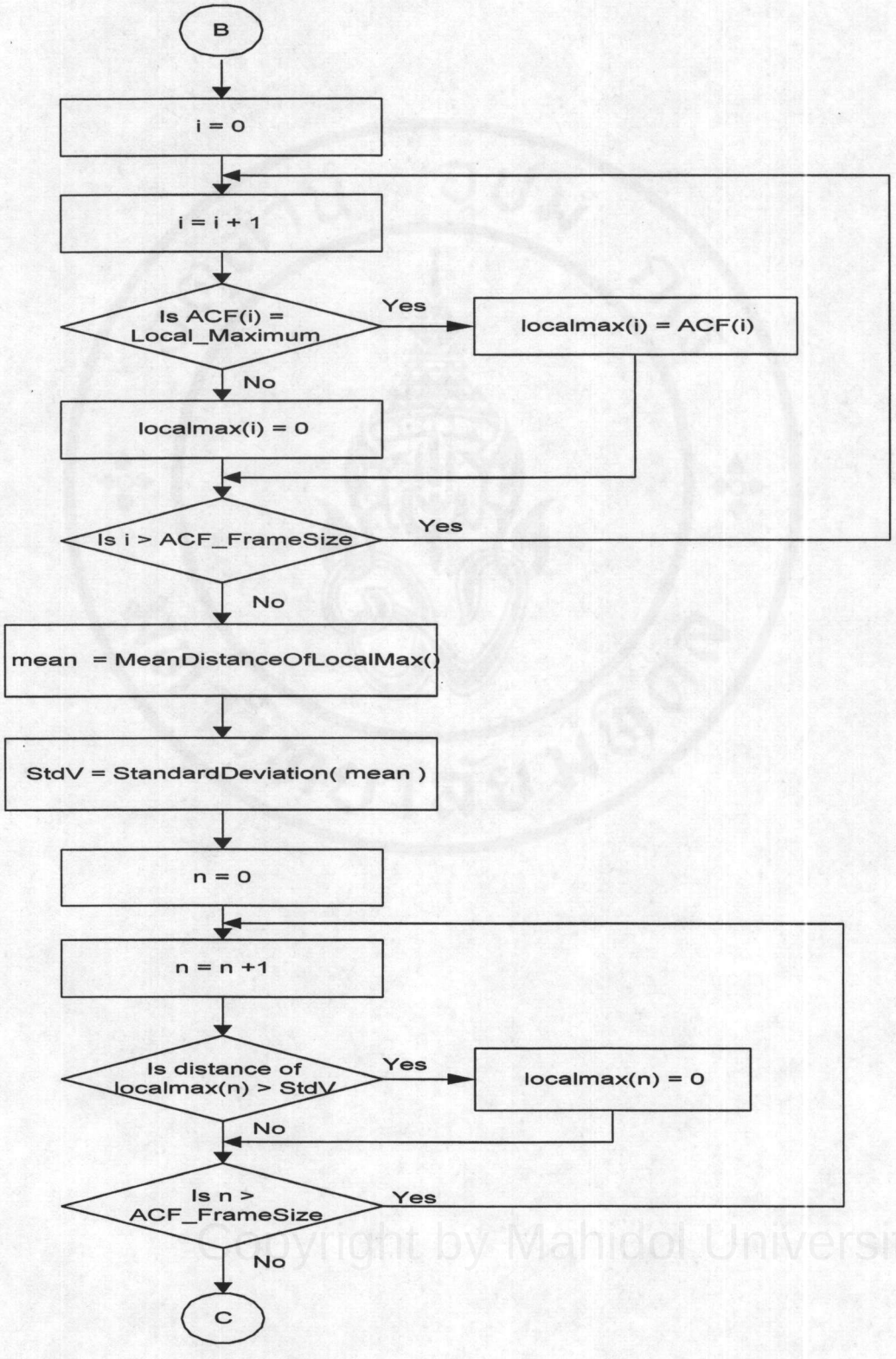


Figure 4.18 Flow chart of Auto-Correlation calculation

4.3.2.6 Pitch period finding



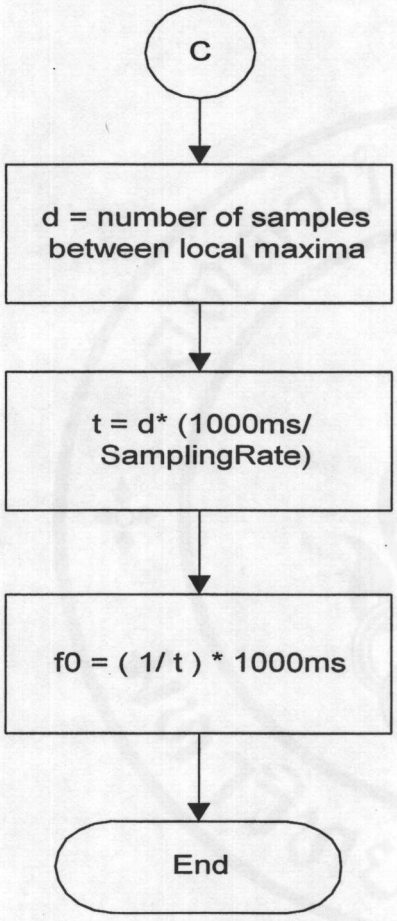
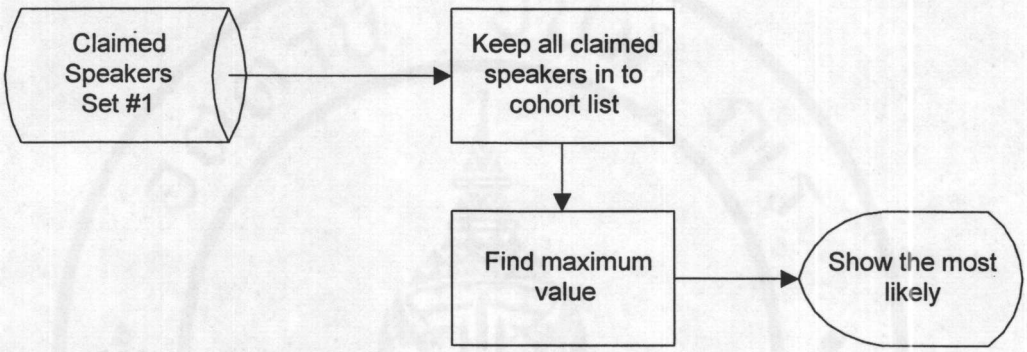


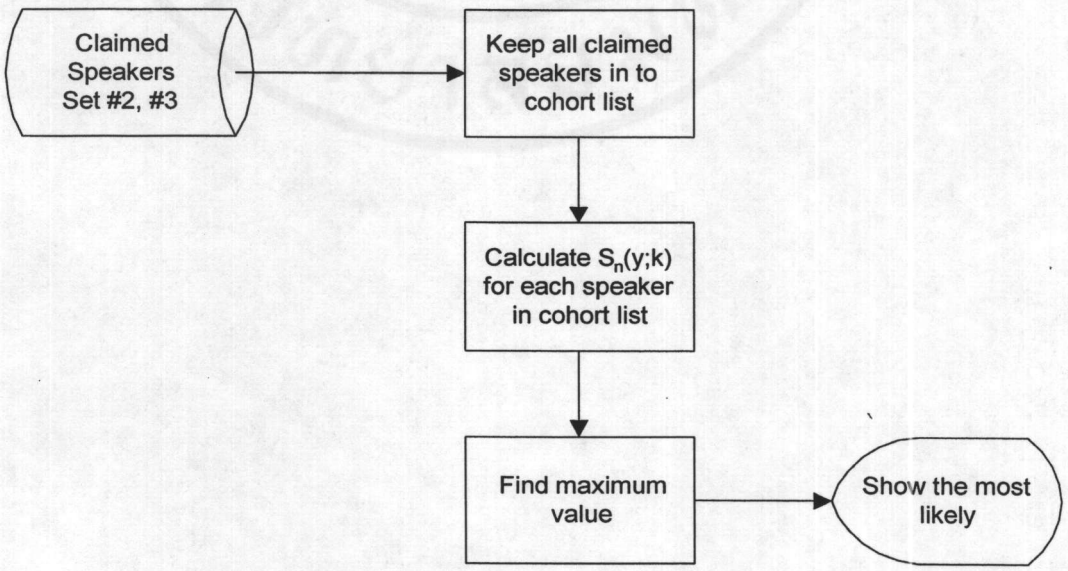
Figure 4.19 Flow chart of pitch period finding

4.3.2.7 Identification decision

Test utterance 1



Test utterance 2 and 3



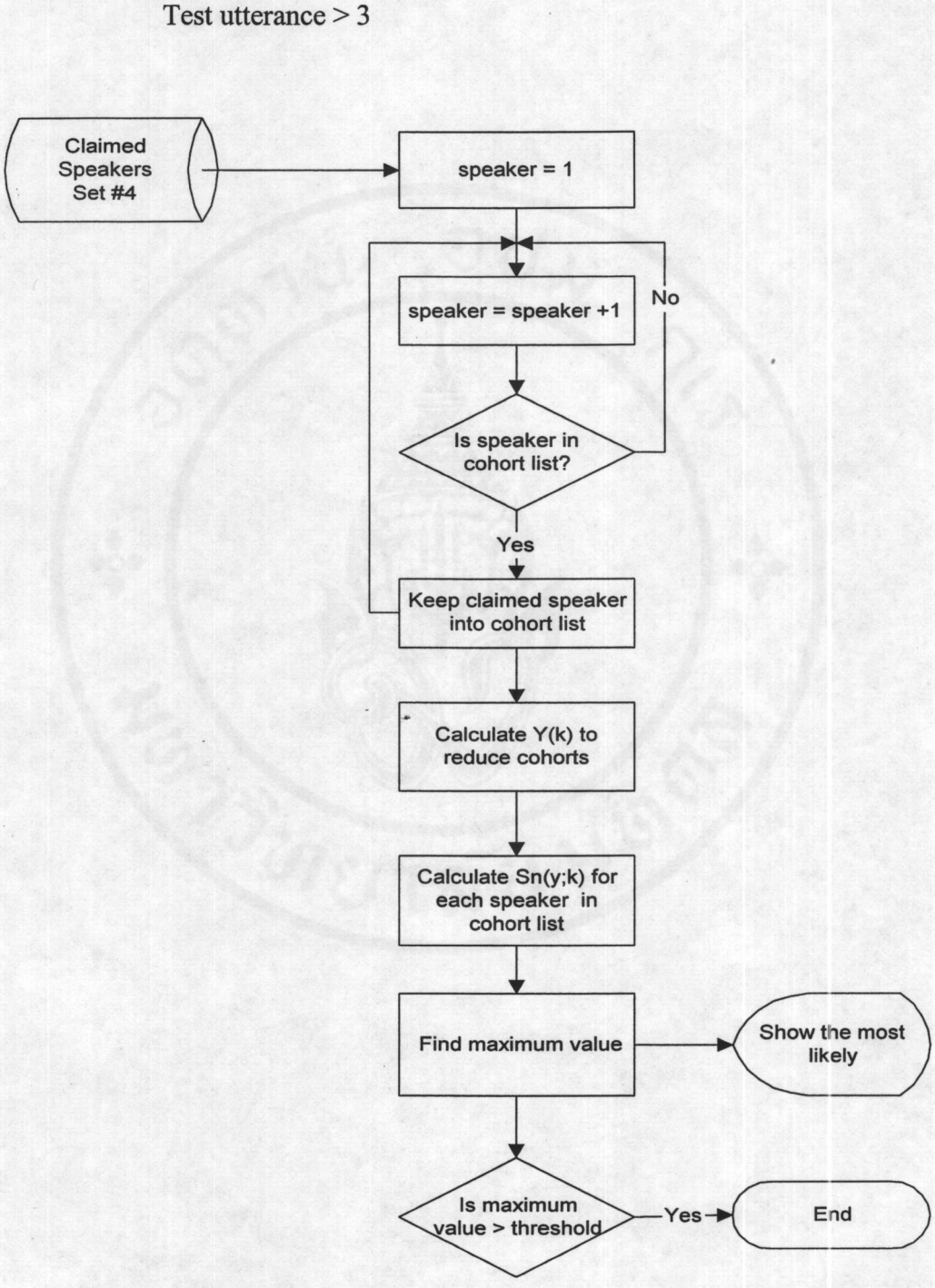


Figure 4.20 Flow chart of identification decision

CHAPTER V

EXPERIMENTAL RESULTS

This chapter presents the experimental results of the Speaker Identification System. We first describe how to setup the system using in this experiment in section 5.1, then the Gender Determination is presented in section 5.2.

Next, in sections 5.3 and 5.4, the experimental results of identification measures are presented. These sections attempt to estimate the identification accuracy of the system. It is also necessary to observe how each component or threshold conditions affect the performance of the overall system. For this reason, we present the results of the overall speaker identification process using sets of known speakers and unknown speakers to ensure that the system accuracy and the performance are carried out simultaneously. The experiment was conducted on a set of 500 and 1,000 training samples respectively.

5.1 System setup

The system designed to operate in a Windows environment, because the choice of a compatible audio card and microphone. Implementation of the system on a stand-alone personal computer would be relatively quick and easy to perform. The speech signal has been recorded as the digital audio data file format, the extraneous information such as characteristics and recording parameters were eliminated. Then the raw signal was used to pre-process of the observed feature and send to the identification process.

There are two tasks that the speaker identification system must perform including training and testing:

- During training, the speaker spoke a set of pre-defined words. His or her pronunciation of the words is then processed and the statistical based and pattern based training samples were constructed.

- During the use as testing, the test speaker is presented a set of pre-defined words. The test speaker speech is then preprocessed to obtain the parameter for clustering process. The pattern of the test speaker then compared with the training samples of the selected possible speakers for identification.

5.2 Gender determination

The Auto-correlation method was used to estimate the pitch frequency of the speech signal to determine the gender of the speaker. One frame of time domain signal at analysis rate of 500 samples or around 40 msec was calculated for the correlation value. Then, the distances between two peaks were marked and calculated the F0 frequency. The average range of pitch periods and F0 frequencies of males, females, and children are shown in Table 5.1. The experiment was conducted using 2 males and 2 females, and the results are shown in Table 5.2, which illustrates the numeric determination values in terms of frequency.

Table 5.1 Average of F0 frequency

	Average Pitch Period Value	Voicing Frequency
Males	7 ms	≅ 80 – 200 Hz
Females	3.5 ms	≅ 150 – 400 Hz
Children	Shorter than Female	> 400 Hz

Table 5.2 Gender determination results

Speaker #	Sex	ฉัน (Hz)	และ (Hz)	เธอ (Hz)	ได้ (Hz)	ไป (Hz)	ดู (Hz)	หนัง (Hz)	ที่ (Hz)	ดี (Hz)	มาก (Hz)
1	Female	202	251	266	322	242	203	194	337	244	330
2	Female	200	222	254	248	202	178	176	300	308	246
3	Male	111	126	111	130	244	139	101	111	121	117
4	Male	113	141	218	142	122	222	123	176	264	228

Our observation is that women tend to produce a higher average pitch frequency than men since their vocal folds are shorter. The results indicated that the algorithm correctly identifies females much more often than males. We found that, for males, their fault determination when attempting to identify the gender of a voice saying the word lies in the nasal sound class such as "ดี" or "ไป".

It is reasonable to conclude that the pitch curve is obviously visible at points where the word makes the vocal chords strongly vibrate with a wide open- nasal cavity. Therefore, the nasal sounds of males might produce a higher value overlap than the range of female pitch frequencies.

5.3 Identification of known speakers

In this section, we presented the identification results through two series of evaluations. The first experiment we have evaluated our identification algorithm on a database of 500 training samples using 2 male and 2 female known speakers for testing. (The "known" test speaker was always one of the known reference speaker) In the second experiment, a database of 1,000 training samples has been used, then we used the same set of known speakers from the first experiment for testing.

Each utterance of test speaker was passed through the same training algorithm to obtains the speech pattern and probability values of the test speaker. Its probability values then use as the forced alignment scores. The probability scores of training data are then sorted and Mahalanobis distances are then calculated against the test speaker scores, and the speaker is accepted if the score of his/her model is in the S±N results[1]. After obtaining the smaller set of possible speakers, the Square-Error Pattern Matching algorithm was performed to find the most-likely speaker. Table 5.3 shows the condition of speech analysis.

Table 5.3 Condition of speech analysis

Sampling rate	11025 Hz
By 40 speakers	20 males and 20 females
10 word utterances	ฉัน และ เธอ ได้ ไป ดู หนัง ที่ ดี มาก
Window type	Hamming
Auto-Correlation frame length	40 ms
DFT frame length	25 ms
Feature Analysis	Power spectral density
Feature parameter	128 points PSD, Probability values of the first sub band
Forced alignment threshold	±20%
Mahalanobis distance threshold	±5%
Square-error rate threshold	>= 92%

5.3.1 Result using 500 training samples

The first experiment, the algorithm was tested on 500 training samples. Test data consisted of 10 words per speaker and it was measured sequentially. In the Table 5.4, pitch frequency, number of the possibly speakers, and the averaged correct rate of the most likely speaker are listed. During testing, all speech pattern of the possible speaker are compared to speech pattern of the test speaker. Then computed average square-error rates are ranked, and shown the most likely speaker, the one who obtained highest values of average correct rate.

The results presented in Table 5.4, the table shown the identification for the input test of the ten words for the four speakers, 2 males and 2 females. The test and the training of speech utterances of all the four speakers were recorded in the same environment, the algorithm produced 100% correct results. We observed that the clustering and the identification performance are variant depending on the performance of gender determination process, since the fault determination of gender lead to the wrong usage of the group of the reference speakers. From 10 utterances of each test speaker, we found more than 30% of fault gender determination, especially fault determination for the male saying the high tone and the female saying the low tone. Moreover, we found the first sub-band of Power Spectral Density performed well in clustering, the member of possible set almost always contains the correct speaker, which is trends to the correct of identification result.

Table 5.4 Results of using 500 training samples

Unknown speaker			Uttered word	Pitch frequency	Determined gender	Force alignment scores	Clustering (possible speaker)	Speakers in Cohort set	The most likely name	Average correct rate of the most likely (%)
No	Sex	Name								
1	M	Akachai	ฉัน	142.86	Male	0.028, 0.00015	5	0	-	-
			และ	146.02	Male	0.018, 0.00015	9	2	Pornpong	97.27
			เธอ	203.98	Male/Female	0.018, 0.00013	19	3	Pornpong	48.64
			ได้	145.13	Male	0.013, 0.00017	11	3	Akachai	65.02
			ไป	141.56	Male	0.015, 0.00015	11	4	Akachai	73.01
			ดู	356.91	Male/Female	0.011, 0.00013	10	4	Akachai	73.01
			หนัง	111.46	Male	0.017, 0.00012	11	2	Akachai	73.01
			ที่	333.59	Male/Female	0.010, 0.00019	9	2	Akachai	73.01
			ดี	50 <	Male/Female	0.011, 0.00014	11	2	Akachai	77.50
			มาก	145.18	Male	0.019, 0.00015	11	1	Akachai	77.50
2	M	Pracha	และ	109	Male	0.027, 0.00011	4	0	Pracha	-
			เธอ	100	Male	0.014, 0.00012	10	0	Pracha	-
			ได้	110.06	Male	0.016, 0.00014	11	1	Pracha	99.41
			ไป	116.15	Male	0.027, 0.00010	4	2	Pracha	49.71
			ดู	111.11	Male	0.028, 0.00013	6	2	Pracha	65.67

			หนัง	50<	Male/Female	0.013, 0.00013	21	5	Pracha	73.73
			ที่	155.76	Male	0.024, 0.00013	10	4	Pracha	97.51
			ดี	201.49	Male/Female	0.011, 0.00016	20	1	Pracha	98.05
			มาก	111.11	Male	0.015, 0.00022	11	1	Pracha	97.8
3	F	Sureeporn	ฉัน	198.01	Male/Female	0.020, 0.00015	21	0	Sureeporn	99.19
			และ	201.43	Female	0.022, 0.00022	18	1	Sureeporn	98.05
			เธอ	275.85	Female	0.012, 0.00017	10	3	Sureeporn	98.01
			ได้	252.74	Female	0.016, 0.00025	8	4	Sureeporn	97.95
			ไป	204.34	Male/Female	0.016, 0.00025	19	5	Sureeporn	98.34
			ดู	439.27	Female	0.008, 0.00013	9	2	Sureeporn	98.34
			หนัง	332.83	Female	0.016, 0.00016	10	2	Sureeporn	98.34
			ที่	262.75	Female	0.010, 0.00014	9	2	Sureeporn	98.24
4	F	Sririnut	ฉัน	242.64	Female	0.017, 0.00017	10	4	Sririnut	97.58
			และ	258.27	Female	0.012, 0.00016	10	4	Sririnut	98.55
			เธอ	212.16	Male/Female	0.011, 0.00013	17	4	Sririnut	65.7
			ได้	204.64	Male/Female	0.014, 0.0002	19	4	Sririnut	49.27
			ไป	251.5	Female	0.008, 0.00017	4	3	Sririnut	49.27
			ดู	172.02	Male	0.018, 0.00013	11	3	Sririnut	49.27
			ดี	228.7	Male/Female	0.009, 0.00022	8	3	Sririnut	59.23
			มาก	226.64	Male/Female	0.014, 0.00017	20	3	Sririnut	56.33

5.3.2 Result using 1000 training samples

In the second experiment we increase the size of database from 500 to 1000 training samples. Using the same conditions of speech analysis as the first experiment and tests the same set of test speaker. The added number of training samples for this experiment does not include any training samples of the four test speakers.

As the first experiment, during testing we computed the pitch frequency, number of the possibly speakers, the averaged correct rate of the most likely speaker and then listed. The Table 5.5 shows almost the same results as using 500 training samples, the identification obtained the correct result even used the larger size of training samples; however, the ranks of testing results for the most likely speaker get changed a bit.

Table 5.5 Results of Using 1000 Training Samples

Unknown speaker			Uttered word	Pitch frequency	Determined gender	Force alignment scores	Clustering (possible speaker)	Speakers in Cohort set	The most likely name	Average correct rate of the most likely (%)
No	Sex	Name								
1	M	Akachai	ฉัน	142.86	Male	0.028, 0.00015	14	0	-	-
			และ	146.02	Male	0.018, 0.00015	22	5	Pornpong	97.27
			เธอ	203.98	Male/Female	0.018, 0.00013	43	7	Pornpong	48.64
			ได้	145.13	Male	0.013, 0.00017	22	7	Akachai	65.02
			ไป	141.56	Male	0.015, 0.00015	23	8	Akachai	73.01
			ดู	356.91	Male/Female	0.011, 0.00013	21	8	Akachai	73.01
			หนัง	111.46	Male	0.017, 0.00012	23	3	Akachai	73.01
			ที่	333.59	Male/Female	0.010, 0.00019	20	3	Akachai	73.01
			ดี	50 <	Male/Female	0.011, 0.00014	43	3	Akachai	77.50
			มาก	145.18	Male	0.019, 0.00015	24	2	Akachai	77.50
2	M	Pracha	และ	109	Male	0.027, 0.00011	10	1	Naruk	96.08
			เธอ	100	Male	0.014, 0.00012	20	4	Naruk	48.04
			ได้	110.06	Male	0.016, 0.00014	23	5	Pracha	33.14
			ไป	116.15	Male	0.027, 0.00010	10	6	Pracha	24.85
			ดู	111.11	Male	0.028, 0.00013	15	3	Pracha	49.25
			หนัง	50<	Male/Female	0.013, 0.00013	42	1	Pracha	98.31

			ที่	155.76	Male	0.024, 0.00013	22	1	Pracha	97.91
			ดี	201.49	Male/Female	0.011, 0.00016	42	1	Pracha	98.05
3	F	Sureeporn	ฉัน	198.01	Male/Female	0.020, 0.00015	44	0	-	-
			และ	201.43	Female	0.022, 0.00022	40	2	Sureeporn	97.19
			เธอ	275.85	Female	0.012, 0.00017	22	6	Sureeporn	98.05
			ได้	252.74	Female	0.016, 0.00025	19	8	Sureeporn	98.01
			ไป	204.34	Male/Female	0.016, 0.00025	40	9	Sureeporn	97.95
			ดู	439.27	Female	0.008, 0.00013	14	5	Sureeporn	98.34
			หนัง	332.83	Female	0.016, 0.00016	22	3	Sureeporn	98.34
			ที่	262.75	Female	0.010, 0.00014	20	3	Sureeporn	98.24
4	F	Sririnut	ฉัน	242.64	Female	0.017, 0.00017	22	7	Sririnut	97.58
			และ	258.27	Female	0.012, 0.00016	22	8	Sririnut	98.55
			เธอ	212.16	Male/Female	0.011, 0.00013	42	8	Sririnut	65.70
			ได้	204.64	Male/Female	0.014, 0.0002	46	9	Sririnut	49.27
			ไป	251.5	Female	0.008, 0.00017	18	6	Sririnut	58.05
			ดู	172.02	Male	0.018, 0.00013	24	6	Sririnut	58.05
			ดี	228.7	Male/Female	0.009, 0.00022	20	5	Sririnut	49.36
			มาก	226.64	Male/Female	0.014, 0.00017	45	5	Sririnut	56.33

5.4 Test of unknown speakers using 1000 training samples

To ensure that the identification would get similar results from the test speakers, we conducted an identification experiment using unknown speaker. This set of unknown speaker does not have any training sample contain in our reference database. We used the database of 1000 training samples and used the same condition of speech analysis as the previous experiment.

The results of test of unknown speaker are given in Table 5.6. In our search, the algorithm stills indicated the most likely speaker whose speech pattern is most similar to the unknown speaker without knowing that the unknown is not the member of the speaker in the database. We realized that it is difficult to distinguish among the speech pattern of each speaker. In fact, the average distance between the power spectrum pattern from computation is very small. As this result, the unknown speech pattern will possibly get close matched to one of the speaker in reference database.

Table 5.6 Results test of unknown speaker using 1000 training samples

Unknown speaker			Uttered word	Pitch frequency	Determined gender	Clustering (possible speaker)	Speakers in Cohort set	The most likely name	Average correct rate of The most likely (%)
No	Sex	Name							
1	F	Sririnut	ฉัน	207.18	Male/Female	40	1	Sureeporn	96.13
			และ	248.87	Female	20	1	Sureeporn	96.13
			เธอ	244.22	Female	21	7	Sririnut	49.17
			ได้	258.44	Female	20	7	Sunisa	65.34
			ไป	247.27	Female	21	8	Sunisa	49.01
			ดู	304.61	Female	21	3	Sridawan	72.13
			หนัง	310.46	Female	21	3	Sridawan	72.13
			ดี	50<	Male/Female	23	3	Sridawan	77.54
			มาก	50<	Male/Female	13	3	Sridawan	77.54
2	M	Naruk	ฉัน	105.26	Male	14	1	Pracha	97.23
			และ	113.82	Male	23	4	Pracha	48.62
			เธอ	115.15	Male	19	6	Pracha	32.41
			ได้	133.33	Male	22	7	Pracha	48.44
			ไป	112.37	Male	15	2	Pracha	64.59
			ดู	100.38	Male	18	2	Pracha	64.59
			หนัง	128.35	Male	19	2	Pracha	72.90
			ที่	50<	Male/Female	28	1	Pracha	97.20
			ดี	128	Male	23	1	Pracha	97.20

CHAPTER VI

DISCUSSION & CONCLUSION

This chapter is the discussion and conclusion on the experiment results presented in the previous chapter.

6.1 Discussion

The speaker identification system was developed and implemented in stand-alone personal computer environment. Three algorithms – feature extraction, speaker search, and identification were developed to be used for indicating the prospect speaker. Feature extraction was extracted the power spectrum density (PSD) from the speech input. We used PSD of each speaker to develop speaker references, which are statistical-based and pattern-based. Clustering algorithm finally used the observed features to calculate for the forced alignment scores and search for the set of possible speaker before being applied to speaker identification and decision making processes.

6.1.1 Complexity

It has been found that during study of the speech feature extraction approach, the sampling rate used while recording (11025 Hz) also can be a source of error. The appropriate rate is ambiguous from many past experiments, it is not exactly known whether this is too small or too big. Also, the environmental background noise may significantly have reduced the system identification quality. Moreover, we also discovered that the variations of Thai phonetic alphabet may be needed more learning cycle, which gives more efficient word utterance in the experiment.

6.1.2 Speed

In the experiment results, the response time is rather short and it should be considered as acceptable. We found that the response time of one utterance can show the identification result in 30 seconds. However, it would depend on the hardware performance and software operation. For hardware we used in the experiment, it is based on Pentium 133 MHz processor, with 30 MB RAM and 1GB Hard Disk. It is running under Windows 95.

6.1.3 Storage

The system requires almost 2MB of storage for storing both data files and source program, which was implemented by using Microsoft Visual Basic 6.0. The database files are stored in ASCII format since it is suitable for sequential search, faster, and requires minimal disk space.

6.1.4 Quality

In the prototype of speaker identification system, we provided the function to add new reference speaker into the speaker model. However, as the number of reference speakers increase, it would effect the identification quality and performance.

6.2 Conclusion

The result obtained in this project can be considered as good in term of performance and accuracy. Ten words, uttered by 40 speakers were used for developing speaker references, under the condition of low noise environment with uttering carefully of the speaker. More than 95% identification was achieved. Although the fault identification is not error free, it is lower than 5%, we have performed the speaker search in quite a minimal computational time. Moreover, the system uses small number of speech features, which significantly reduce system pre-processing time.

From our observation, the key factors that cause an identification error should be:

- The system currently obtains a population of N speakers, the one who is not the member in the population would be identified as a nearest match to one of those within the population. The system in unable to conclude a speaker as unidentified, but it will be present the name of the most likely one with low rate of possibility.

- Speech recording environment is also a very important factor, it should be quiet and the speakers should speak patiently and carefully. Moreover, the normalization process to remove the background noise from the speech should be included, since this would improve the identification performance.

- Another important factor is that the pitch contours of each speaker uttered had sometime shifted beyond the range of appropriate frequency, causing fault identification of their gender. This indicates the fact that the pitch frequency estimation is required in proper word utterance for proper determination of gender.

CHAPTER VII

SUGGESTION FOR THE FUTURE WORK

This section, we proposes the future work in connection with our research as the following:

1. The speech record environment should be isolated with background noise. The speakers should speak consistently and the distance between the speaker and the microphone should be fixed.

2. The hardware components for recording the speech should be acquired to use more high quality. For example, use the high sensitive head-microphone.

3. The pre-defined words for identifying may needs more length of learning cycle for different speakers.

4. The normalization process should be added to the system, which aimed to solve the problem of speaking at different level of volumes and to remove background noise prior to filtering.

5. The experiments should be undertaken to discover the optimum sampling rate and number of bits for recording speech data.

6. The extension of feature extraction study should be further investigated, this is for searching the optimal features to be introduced in the system in order to improve the identification performance.

REFERENCES

1. Ramalingam H. Extraction of tone of speech: An application to the Thai language.
Master Thesis, Asian Institute of Technology 1995.

2. Ian H. Witten. Principles of computer speech. 1982.

3. Chris Rowden. Speech processing. McGraw-Hill Book Company 1992.

4. Sarma, Sridevi V. A segment-based speaker verification system using SUMMIT. Master Thesis Department of Electrical Engineering and Computer Science, M.I.T. 1997.

5. A Syrdal, R. Bennett, S. Greenspan. Applied speech technology. CRC Press,Inc. 1995

6. Aaron E. Rosenberg and Frank K. Soong. Recent research in automatic speaker recognition. AT&T Bell Laboratories, Murray Hill, New Jersey.

7. Tony Robinson. Speech analysis.
http://svr-www.eng.cam.ac.uk/%7Eajr/SA95/SpeechAnalysis.html

8. Pitch determination.
http://www-dsp.rice.edu/~akira/digitalbb/pitch.html

9. Mahalanobis classifiers.
http://www-engr.sjsu.edu/~knapp/HCIRODPR/PR_Mahal/PR_Mahal.htm

10. Yasuo Ariki, Speaker rcognition rbust for tme dfference bsed on subspace method. Department of Electronics and Informatics, Faculty of Science and Rechnology, Ryukoku University.

BIOGRAPHY

NAME Miss Varaporn Phomvi-in
DATE OF BIRTH 11 March 1968
PLACE OF BIRTH Bangkok, Thailand

INSTITUTIONS ATTENDED

Rajamangala Institute of Technology,
1989-1990:
Bachelor of Electrical Engineering
(Minor – Electronics & Computer)
Mahidol University, 1995-1999:
Master of Science
(Computer Science)

POSITION & OFFICE

1992-1992	Software Engineer, Computer Department National Semiconductor (Thailand), Ltd.
1993-1995	Computer Program/Analyst, Computer Department, Computer Integrated Manufacturing Corp.
1996-Present	Database Systems Professional, Information Technology Department, Unocal Thailand, Ltd.