THAI-LAOTIAN VOWEL RECOGNITION BASED ON BARK SCALE SPEECH FEATURES

KHAM KHANTHAVIVONE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF ENGINEERING IN ELECTRICAL ENGINEERING

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2007

หัวข้อวิทยานิพนธ์	วิธีการรู้จำเสียงสระภาษาไทยและลาวโดยการใช้ลักษณะของเสียงบนสเกลบาร์ค
นักศึกษา	นายคำ ขันทะวีวอน
รหัสนักศึกษา	46060214
ปริญญา	วิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
พ.ศ.	2550
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ. ดร.ไกรสิน ส่งวัฒนา

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอวิธีใหม่สำหรับการรู้จำเสียง โดยการนำใช้ลักษณะของเสียงบนสเกลบาร์กร่วมกับลักษณะของพลังงานเสียงและหน่วยเสียงวรรณยุกต์ ซึ่งสเกลบาร์กเป็นสเกลที่ใช้ในการวัดสเปกตรัมในการได้ยินของมนุษย์ (Psychoacoustics) ซึ่งจะมีความใกล้เคียงกับระบบการได้ยินของมนุษย์มาก โดยค่าความเข้มพลังงานตามสเปกตรัมบนสเกลบาร์กนั้นจะแสดงถึงระดับการตอบ สนองต่อเสียงที่ความถี่นั้นๆของระบบการได้ยิน ลอการิทึมของค่าความเข้ม (log(CBI)) นำมาแปลงโคซายน์แทรนซฺฟอร์ม และค่าสัมประสิทธิ์ที่ได้ถูกใช้แทนคุณลักษณะเพื่อการรู้จำร่วมกับการคำนวณลักษณะการเปลี่ยนแปลง (differentials) ของค่าสัมประสิทธิ์ตามเวลา ทำให้วิเคราะห์ลักษณะของเสียงพูดได้ชัดเจนมากขึ้น นอกจากนั้นการถดถอยพลังงาน (regression energy) เสียงพูดยังได้แสดงถึงช่วงเวลาสั้น-ยาวตามความหมายของเสียงและลักษณะพลังงานนี้ได้ถูกแปลงเป็นสัมประสิทธิ์เพื่อการแยกแยะเสียงสั้น-ยาวของสระ นอกจากนี้เสียงสระของภาษาไทยและลาวจะควบด้วยเสียงวรรณยุกต์เพื่อบอกความหมายเสียงพูดอย่างครบถ้วน คุณลักษณะที่ใช้ทั้งหมดนี้แทนด้วย 16 พารามิเตอร์ของค่าสัมประสิทธิ์ DCT(log(CBI)) และ 16 พารามิเตอร์ของค่าการเปลี่ยนแปลงของค่าสัมประสิทธิ์ตามเวลา (delta DCT(log(CBI)) ร่วมกับหนึ่งพารามิเตอร์จากเสียงวรรณยุกต์และหนึ่งพารามิเตอร์จากการถดถอยพลังงานของเสียงพูด นอกจากนี้ การวิเคราะห์ลักษณะของเสียงนี้ยังไม่มีขบวนการจำกัดสัญญาณรบกวน ทำให้ประสิทธิภาพในการรู้จำไม่ดีเท่าที่ควรในสภาพแวดล้อมที่มีเสียงรบกวน ดังนั้นในวิทยานิพนธ์นี้จึงได้นำวิธีการสำหรับกรองสัญญาณรบกวนมาเพิ่มประสิทธิภาพการรู้จำทำใช้ให้ดีขึ้น โดยวิธีการใหม่คือ Running Spectrum Filterซึ่งใช้ตัวกรองความถี่ชนิดแถบความถี่ผ่าน (band-pass filter) บนความถี่มอล์ดุเลชัน(modulation frequency) ในวิทยานิพนธ์ได้เสนอการออกแบบตัวกรองที่มีลำดับการคำนวณต่ำทำให้การคำนวณใช้เวลาน้อยและเร็วขึ้น

Thesis Title	Thai-Laotian vowel recognition based on Bark scale speech features
Student	Mr. Kham Khanthavivone
Student ID.	46060214
Degree	Doctor of Engineering
Program	Electrical Engineering
Year	2006
Thesis Advisor	Assoc. Prof. Dr. Kraisin Songwatana

ABSTRACT

We propose a new speech recognition system using speech spectrum envelop on Bark scale, the regression on voice energy and a quantized pitch for speech features. Bark scale is a psychoacoustics measurement on human hearing property and all critical bands are defined on Bark scale. We have estimated the frequency selectivity of the hearing system by calculating critical band intensities (CBI). The regression on the speech energy is used to help classifying short and long vowels. Each tested phonemes including consonants and vowels. All vowels are automatically extracted and recognized in our experiments. The speech features are represented by 16 parameters vectors of logarithm CBI into discrete cosines transform (DCT(log(CBI))), 16 parameters of their differentials, a parameter of regression on the speech energy (RE) and a parameter of quantized pitch (QP). Feature, we have focused on the improvement of the accuracy on vowel recognition using advanced robust speech detection technique, i.e., running spectrum filter(RSF), in the presence of noise using frequency response masking running spectrum filter (FRMRSF). This filter has been designed from model filters and masking filters. It has low numbers of FIR filter coefficients while realizing a narrow transition bandwidth. In this thesis, the modified FRM design based on band-pass filter is introduced and it is applied to the speech recognition system. The new design, RSF, can be used for robust speech recognition with low calculation cost.

Acknowledgments

Firstly, I would like to express my deepest gratitude to my advisor, **Assoc. Prof. Dr. Kraisin Songwatana** and co-advisor, **Prof. Dr Yoshikazu Miyanaga**. He has inspired, encouraged, guided, and supported me every means throughout the duration of my research. I also would like to express my appreciation to AUN/SEED-Net for their financial support throughout this research.

Secondly, I would like to thank patient guidance, friendship and knowledge throughout the duration of my study. I also thank all Telecommunication Engineering's (KMITL) staffs, who made this enjoyable experience. In addition, I would like to thank all my fellow students in the speech recognition research laboratory for their kindness and friendships.

Thirdly, I would like to thank patient guidance, friendship and knowledge throughout the duration of my study. I also thank all Graduate school of information science and technology, Hokkaido University, who made this enjoyable experience. In addition, I would like to thank all my fellow students in the speech recognition research laboratory for their kindness and friendships.

Finally, I would like to thank all Lao undergraduate student and staffs at department of Electronic Engineering, National University of Laos and my Lao's friends at Laos and KMITL, for their kindness and help to collect the data.

Contents

Contents (continuous)

Contents (continuous)

List of Tables

List of Tables (continuous)

List of Figures

List of Figures (continuous)

Chapter 1

Introduction

1.1 Background

The speech recognition has motivated many researchers to develop machines that can accept the human speech and respond properly. Spoken language processing research intends to develop and implement algorithms for a machine to be able to generate, recognize, and understand a spoken language. In order to implement such a machine, speech analysis, speech synthesis, speech recognition, natural language processing, and human interface technology are incorporated in spoken language processing system. The spoken language systems have been developed for a wide variety of applications, ranging from a small set of vocabulary to a large set of vocabulary. It realizes communications between humans and machines, i.e., voice dialing in mobile phones, aviation information retrieval, weather information retrieval, automated reservation, etc.. there application have been applied to practical device such car navigation systems, video games, pet robots, etc.. In the field of speech processing, such demands have induced various research activities: speech recognition, speaker recognition, speech coding, speech synthesis and acoustic processing. The choice of speech features is vary important. They directly affect the accuracy of a speech recognizer. Moreover, the technology is quite essential for future advantaged applications represented by automatic speech translation systems and artificial intelligence. It is expected that speech recognition systems will be embedded into electrical devices. In the context of speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. A basic question then is how to deal with the redundancy and variability carried by the speech signal.

The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front-end. It performs some kind of spectrum temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer.

Basically, spectrum-based features have been widely used in speech recognition tasks. The

speech spectrum envelops from minimum mean square estimation (MMSE) technique has provided good way to generate speech patterns for classification [1], [2], [3]. In addition, tonal information which is derived for fundamental frequency, .i.e., pitch information, is employed into several speech recognition systems [1],[4]. In conventional speech recognition systems, speech spectrum envelops and cepstrum coefficients are calculated from auto-regressive model (AR model) using MMSE, MFCC and PLP. The log energy measures are also extracted as features directly and appended to MFCC or PLP [1], [5], [6],[7].

As a unique spectrum distortion measure, Bark scale has been studied [1], [7], [8]. This scale is based on human physiological and psychological property. Bark scale is recognized as a suitable scale for recognizing many auditory phenomena, such as perception of loudness and timbre. These processes provide good performance in many languages. An auditory-based warping of the frequency axis derived from Bark scale has been widely used in speech recognition, i.e., PLP [6] method. The method employs the critical band spectral, the equal-loudness curve and the intensity loudness power.

1.2 Objective

Our proposed technique focuses on the improvement of the accuracy on vowel recognition using new speech features from the critical band intensity on the Bark scale, a polynomial regression on the voice energy function and a quantized pitch concept. These concepts provide good performance in tone languages .i.e., Thai and Laotian. For new features, we can estimate the frequency selectivity of the hearing system by approximating its critical band intensities (CBI). The MMSE spectrums are mapped onto the Bark scale and the respective CBI are used as feature vectors has been studied [9], [10]. In addition, a regression method on the voice energy has been used to classify short and long vowel has been studied [11], [12]. The quantized pitch is important feature that allows gender-free modeling of the recognition has been studied [13], [14]. These features are estimated for isolated word speech recognition experiment using HMM.

However, speech recognition systems are used in various practical environments where the robustness of recognition systems becomes more essential [15]-[20], [31]-[39]. In real environments, input speeches may include various noises derived from sound sources such as cars, speech babble, etc.. Therefore, the robustness of speech recognition system is considerably needed for practical use. We proposed an advanced noise robust speech adaptive with running spectrum filter. The modified frequency response masking design based on band-pass filter is

applied to the design of running spectrum filter. These filter have been designed from Finite impulse response (FIR) filters allowing low number of filter taps while realizing narrow transition bandwidth. In this thesis, Using the new design, we have reduced the calculation cost. Experiments indicates speech recognition in the presence of noise robustness can be improved.

1.3 Thesis overview

The remainder of this thesis is organized into seven chapters. Chapter 2 describes the rules of spoken languages and grammatical structure words and phrases are selected and ordered. Consonants, vowels, and tones for Thai and Laotian languages consists consonant, vowel and tone. Chapter 3, principal algorithm of implemented speech recognition system is described. We provide the linear prediction coding, Bark scalar, Mel scale, vectors quantization and hidden Markov modeling (HMM). Chapter 4, feature extraction techniques are explained. Our proposed speech focuses its on the improvement of the accuracy on vowel recognition using new speech features: critical band intensity on the Bark scale and the robustness of recognition systems, a regression method on the voice energy and a quantized pitch concept. These concepts provide good performance in tone languages such as Thai and Laotian. Chapter 5, the noise robust techniques used for noise reduction are the Running Spectrum Filter(RSF). In additional, modification frequency response masking design based on band-pass filter is introduced and it is applied to the design of running spectrum filter for reduced the calculation cost. Chapter 6, several experiments are carried out. The conventional recognition system consists of ordinary feature extraction based MFCC, PLP or DCT(log(CBI)). Then, FRMRSF/DRA and RSF/DRA are applied together.This chapter is summarized the performance results of recognition, Chapter 6 is conclusions.

Chapter 2

Thai and Laotian Language Structure

Thai and Laotian language belongs to the Tai language family which also includes Thai, Shan, and languages spoken by smaller related ethnic groups in Laos, Thailand, Burma, southern China, and northern Vietnam [21]. The languages in the Tai family all share a common grammar and tone structure, called *"Tonal Language"*. Thai and Lao language has many regional varieties. The main difference between these varieties is tonal, different varieties will have some changes in tone and vocabulary from region to region. However the Bangkok and Vientiane variety are considered as the unofficial national language. This can be seen in the capital where people from all over the country live. Since the syllable is principally considered a fundamental unit for acoustic phonetic analysis, it is important to have a good understanding about Thai and Lao syllables. Each syllable sound consists of consonant, vowel and tone. They are described in the structure as follow.

2.1 The language structure for Lao

2.1.1 Lao Consonants

There 27 original consonants realized Lao alphabetical order. These consonants are divided into three classes, 6 for high consonants, 8 for middle consonants, and 13 for low consonants. Laotian language is formed by a serial construction of these syllables. Representing 27 original sounds as shown in Table 2.1. Note that, a special Lao consonant (*) is not defined in any class. Since the consonant class is one of the critical factors in determining a syllable's tone, the consonant class has to be known in order to correctly pronounce a Thai and Lao syllable or a word. Normally, all Lao consonants can all be used at the beginning of a syllable, namely *"Initial Consonant"* and at the end of a syllable, called *"Final Consonant"*. Understanding this system helps in understanding Lao tones and enable learners to write Lao correctly.

2.1.2 Lao Vowels

The Lao language has a complex vowel system. It is consisted a total of 27 vowels for Lao. Lao vowels have 12 sounds vowels are monophthong, 12 vowels are diphthong, and addition only 3 sounds but 4 forms of script are special vowel of Lao, as shown in Table 2.2. Lao vowels

Table 2.1 Three consonant classes

High consonants		Middle consonants		Low consonants	
Letter	**sound**	**Letter**	**sound**	**Letter**	**sound**
ຂ	/k/	ກ	/g/	ຄ	/k/
ສ	/s/	ຈ	/j/	ງ	/ng/
ຖ	/th/	ດ	/d/	ຊ	/s/
ຜ	/ph/	ຕ	/t/	ຍ	/ny/
ຝ	/f/	ບ	/b/	ທ	/th/
ຫ	/h/	ປ	/p/	ນ	/n/
		ຢ	/y/	ພ	/ph/
		ອ	/z/	ຟ	/f/
				ມ	/m/
				ລ	/l/
				ວ	/w/
				ຮ	/h/
				ຣ	/r/*

Note: *. (r) is not used as the main consonant in Lao syllable. It is used only when words from foreigner languages are pronounced in Lao

divided into two groups as 12 short so (short duration vowel) and 12 long sounds (long duration vowel). The short vowels have shorter duration and more immediate stop of vocal cord input, and while the long vowels are repetitions of the vowels over a longer input, with gradual release of air from the lung. Remember vowel length changes meanings. It is vary important pronounce correctly, pay attention to the speaker's lips. Note that, the special vowels are not defined to both short and long vowels. The syllables comprise of a special vowel are not allowed to pronounce associated with any final consonants. However, they are categorized as long vowels for tone rule purpose.

2.1.3 Lao Tones

As described above, Thai and Laotian language is a tonal language of Tai language family. A tonal language or tone language is one in which changes in pitch of syllable or word, lead to changes in syllable or word meaning. Example of tone languages are Thai, Chinese, Burmese,

Table 2.2 Vowel categories, monophthong, diphthong and special vowel

Monophthongs				Diphthongs				Special vowels	
Short Vowel		Long Vowel		Short Vowel		Long Vowel			
xะ	/a:/	xา	/aa/	เxีอ	/ωa:/	เxืออ	/ωa/	ไx	/ai/
xิ	/i:/	xี	/ii/	xัวะ	/ua:/	xัว	/ua/	ใx	/ai/
xึ	/ω:/	xื	/ωω/	เxียะ	/ia:/	เxีย	/ia/	เxา	/ao/
xุ	/u:/	xู	/uu/					xำ	/am/
เxะ	/e:/	เx	/ee/						
แxะ	/E:/	แx	/EE/						
โxะ	/o:/	โx	/oo/						
เxาะ	/O:/	xั	/OO/						
เxิ	/ɛ:/	เxิ	/ɛɛ/						

Table 2.3 Lao tone mark

Category	Tone Mark	Name
Dynamic	x◌่	mai-zeek
	x◌้	mai-thoo
Static	x◌๊	mai-tii
	x◌๋	mai-jattavaa

Vietnamese, Lao, and some European and African languages. Most languages use tone to convey grammatical structure or emphasis, but this does not make them tonal languages in this sense. In these cases, tones can change how the audience is intended to interpret a word. But in tonal languages, the tone is an integral part of award itself.

In Thai and Laotian language, tone is an integral component of a syllable, where tone information is an essential lexical meaning of Thai and Lao utterance. For tones of Lao words are determined by the tone chart as show in Table 2.3. All languages in the Tai family follow the tone system explained here, with tones integrated into other aspects of pronunciation: initial consonants, final consonant sounds, and vowel length. Lao writing system has 4 tone marks,

Table 2.4 Tone chart of Lao spoken in Vientiane

Syllable / Initial consonant	Syllable*			Syllable**	
	Inherent Tone	x (low tone mark)	x (falling tone mark)	Long vowel	Short Vowel
high cons., /ฝ/ส/ฟ/ฑ/ฃ/ฅ/	Low Rising(4)	Mid(0)	Low Falling(1)	High Falling(1)	Low Rising(4)
middle cons., /อ/บ/ด/ย/ป/ต/ก/	Low Rising(4) (or Low falling)				
Low cons., /ฟ/ร/ล/ฯ/ฬ/ห/ฮ/ /ง/ญ/ม/ว/ย/	High Rising(3)				Mid(0)

Notes: * A syllable consists of long vowel or ending with sonorant finals.

** A syllable consists of short vowel or ending with stop finals.

- The number 0,1,2,3 and 4 are made up to represent for five Lao tone types in thesis only.

categorized as dynamic tones and static tones as show in Table 2.4. However, there are more than 4 tone sounds in Lao pronunciation. Ancient Lao spoken language system has 5 tone sounds. In recent years, advance research has found that, there are perhaps more than 5 tones in Lao spoken language, depending on the region pronunciation. For example, there are five tones of Luangphapang pronunciation (Northern of Laos), six tones of Pakse pronunciation (Southern of Laos). Vientiane population is emigrated from many region of Laos. However, Vientiane pronunciation with five tones is commonly used as the official spoken language of Laos. Thus, we are used Vientiane tone is only one that has to be studied in this thesis.

From history, five Vientiane tone chart has been presented by Brown in 1965; Reinhorn, 1970-1971; Strecker in 1980 (unpublished); Chittavoravong in 1980 (unpublished); Enfield in 2000 and Senglathsamy Chanthamenavong in 2004. Since, Hoshino, Marcus in 1973; and Levy 1980 have presented six tones of Vientiane pronunciation. However, the tone chart of Crisfield-Hartmann and Enfield, as illustrated on the Table 2.4 are used in this thesis.

A tone is a feature of pitch or fundamental frequency movement within a syllable. Figure 2.1, shows the average of pitch contours, extracted from male and female voiced (Vientiane speaker) of a syllable, which has different tones.

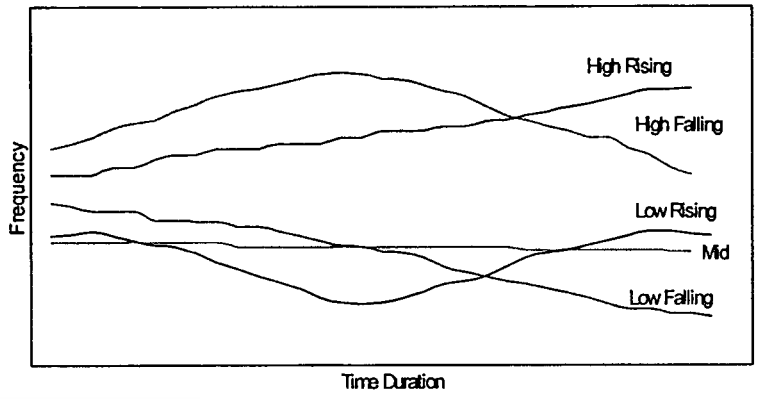


Figure 2.1 Average pitch contours over syllables of Vientiane speaker (Laotian), which are represented five tones

Table 2.5 Phonetic symbol of Thai consonant

High consonants		**Middle consonants**		**Low consonants**	
Letter	**sound**	**Letter**	**sound**	**Letter**	**sound**
ข,ฃ	/kh/	ก	/g/	ค,ฅ,ฆ	/k/
ฉ	/ch/	จ	/j/	ง	/ng/
ฐ	/t/	ด	/d/	ช,ฌ	/ch/
ถ	/t/	ต	/t/	ซ	/s/
ผ	/p/	บ	/b/	ญ,ย	/y/
ฝ	/f/	ป	/p/	ฑ,ฒ,ท,ธ	/t/
ศ	/s/	อ	/o/	ณ,น	/n/
ษ	/s/	ฎ	/d/	พ,ภ	/p/
ส	/s/	ฏ	/t/	ฟ	/f/
ห	/h/			ม	/m/
				ร	/r/
				ล,ฬ	/l/
				ว	/w/
				ฮ	/h/

2.2 The language structure for Thai

2.2.1 Thai Consonants

There are 44 consonantal letters in Thai. These letter represent 21 phonemes, grouped by traditional Thai grammarians into 3 classes; the high class, the middle class and the low class. These classes are very important in determining the tone of a syllable. These consonants are divided into three classes, 11 for high consonants, 9 for middle consonants, and 24 for low consonants. Representing 44 phonetic symbol of Thai consonant as shown in Table 2.5.

2.2.2 Thai vowels

Thai has 18 monophthongs, nine short and nine long. Thai vowels divided into two groups as 12 short (short duration vowel) and 12 long sounds (long duration vowel). The short vowels have shorter duration and more immediate stop of vocal cord input, and while the long vowels are repetitions of the vowels over a longer input, with gradual release of air from the lung. A pair of short and long monophthongs is quantitatively different but speech spectrum quite similar. The Vowel categories as shown in Table 2.6.

Table 2.6 Phonetic symbol of Thai vowel

Monophthong				**Diphthong**			
Short Vowel		Long Vowel		Short Vowel		Long Vowel	
อะ	/a:/	อา	/aa/	เอือะ	/ɯa:/	เอือ	/ɯa/
อิ	/i:/	อี	/ii/	เอียะ	/ia/	เอีย	/ia:/
อึ	/ɯ:/	อื	/ɯɯ /	อัวะ	/ua:/	อัว	/au/
อุ	/u:/	อู	/uu/				
เอะ	/e:/	เอ	/ee/				
แอะ	/E:/	แอ	/EE/				
โอะ	/o:/	โอ	/oo/				
เอาะ	/O:/	ออ	/OO/				
เออะ	/ɛ:/	เออ	/ɛɛ /				

2.2.3 Thai Tones

Thai language is a tonal language of Tai language family. A tonal language or tone language is one in which changes in pitch of syllable or word, lead to changes in syllable or word meaning.

Similarly Lao tones as described in subsection 2.1.3. For tones of Thai and Lao words are determined by the tone chart as show in Table 2.3. There are five different lexical tones in Thai: the mid, the low , the falling , the high, and the rising [11],[12].

All five different tones are found only on sonorant ending syllables, i.e., open syllables with long vowels and syllables ending with nasals. Obstruent ending syllables, i.e., open syllables ending with short vowels and syllables ending with stops, are restricted to specific tones; we found only the low, the fall, and the high but the fall in this type of syllable is scarce. For syllables with long vowels ending with stops, we found only the low, the fall, and the high but the high in this type of syllable is also scarce.

Tones integrated into other aspects of pronunciation: initial consonants, final consonant sounds, and vowel length. Thai writing system has 4 tone marks, the tone chart of the Transliteration's list from the Royal Institution and PhD. Wit Thiengburanathurm, Thai-English Dictionary, and Issues in Thai Text-to-Speech Synthesis by the NECTEC in 2000, as illustrated on the Table 2.7 are used in this thesis.

Table 2.7 Tone chart of Thai spoken

Vowel	Consonant			
	Monophthong		Diphthong	
	Short vowels	Long vowels	Short vowels	Long vowels
Live syllable	Mid, Low, Falling, High	Mid, Low, Falling, High, Rising	-	Mid, Low, Falling, High, Rising
Dead syllable	Mid, Low, Falling, High	Mid, Low, Falling, High	-	Low, Falling, High
Open syllable	Mid, Low, Falling, High	Mid, Low, Falling, High, Rising	Low, Falling, High	Mid, Low, Falling, High, Rising

Chapter 3

Speech Recognition System Fundamentals

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. Speech recognition systems can be characterized by many parameters. A more general form of speech recognition process includes speech pre-processing, feature extraction, spectral analysis of speech, dynamic time warping and recognition by discrete Hidden Markov Modeling. In addition, applications cover vowels and tone recognition, speaker dependent and independent recognition, modeling, and descriptions of some well-known algorithms.

When the speech signal is generated and propagated to the listener, the speech perception process begins. First the listener processes the acoustic signal along the basilar membrane in the inner ear, which provides a running spectrum analysis of the incoming signal. A neural transducer process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process.

3.1 Speech production system

The speech generation process begins when the talker formulates a message that he wants to transmit to the listener via speech. The human vocal tract begins at the opening of the vocal cords, or glottis, and ends at the lips. It consists of the pharynx and the mouth. In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area of the vocal tract, determined by the position of the tongue, lips, jaw and velum, varies from zero to about 20 cm^2. And the nasal tract begins at the velum and ends at the nostrils. When the velum (a trapdoor-like mechanism at the back of the mouth cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

A schematic diagram of the human vocal mechanism is shown in Figure 3.1[22]. Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the windpipe, the tensed vocal cords within the larynx vibrate in the mode of a relaxation oscillator by the air flow. The air flow is chopped into quasi-periodic pulses which are then modulated in frequency in passing through the pharynx, the mouth cavity and possibly the nasal cavity. Depending on the positions of the various articulator i.e., jaw, tongue, velum, lips, mouth,

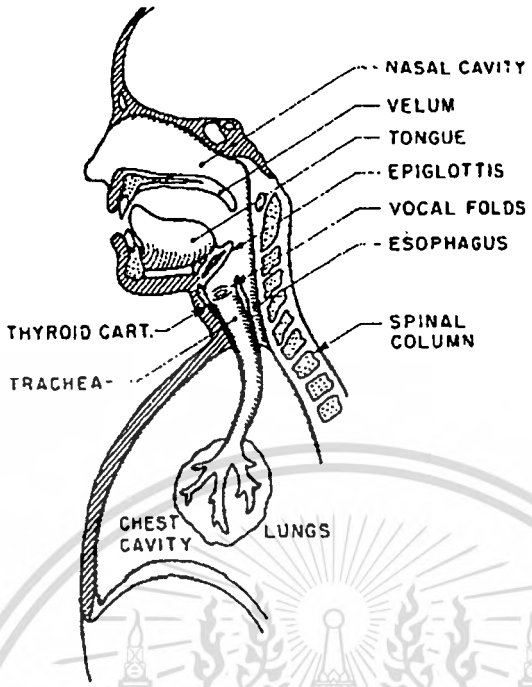


Figure 3.1 Schematic view of the human vocal mechanism [22]

different sounds are produced.

When the vocal cords are tensed, the air flow causes them to vibrate, producing the voiced speech sounds. When the vocal cords relax, in order to produce a sound, the air flow must either pass through the constriction in the vocal tract and thereby become turbulent, producing the unvoiced speech sounds, or it can build up pressure behind a point of total closure within the vocal tract, and when the closure is opened, the pressure is suddenly and abruptly released, causing a brief transient sound. Speech is produced as a sequence of sounds. Hence the state of the vocal cords, as well as the positions, shapes and sizes of the various articulators, changes over time to reflect the sound being produced.

3.2 Fundamental frequency and pitch

Fundamental frequency (f_0) estimation, also referred to as pitch detection, has been found in many research topic for many years, and is still being investigated today [11]-[12]. Most research into this area goes under the name of pitch detection, although what is being done is actually f_0 estimation. Because the psychological relationship between f_0 and pitch is well known, it is not an important distinction to make.

Larynx is an energy provider that serve inputs to the vocal tract. The volume of air determines the amplitude of the sound. The vocal cords at the base of larynx, and glottis triangular-shaped space between the vocal cords are the critical parts from speech production point of view. They separate the trachea from the base of vocal tract. The types of sounds are determined by the action of vocal cords, and we call it excitation. Normally excitations are characterized as phonation, whispering, friction, compression, vibration, or a combination of these. Speech produced by phonated excitation is called *voiced*, produced by the cooperation between phonation and frication is called mixed voiced, and produced by other types of excitation is called *unvoiced* [23].

Voiced. Voiced speech is generated by modulating the air stream from the lungs, and the generation is performed by periodically open and close vocal folds. The frequency of vocal cords vibration is called the *fundamental frequency* (f_0), and it depends on the physical characters of vocal cords. Vowels and nasal consonants belong to voiced speech.

Unvoiced. Unvoiced speech is generated by a constriction of the vocal tract narrow enough to cause turbulent airflow, which results in noise or breathy voiced. It includes fricatives, sibilants, stops, plosives and affricates. Unvoiced speech is often regarded and modeled as white noise.

3.3 Linear predictive coding

The technique of linear prediction is based upon the assumption that sample values of speech may be approximated by a linear combination of the preceding p samples. Mathematically,

$$s'(n) = a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) \cdots a_p s(n-p) \tag{3.1}$$

$$= \sum_{k=1}^{p} a_k s(n-k) \tag{3.2}$$

where $s'(n)$ is the predicted sample at time n and $a_1, a_2 ... a_p$ are the predictor coefficients. Generally it will not be possible to exactly predict the signal, leading to an error $e(n)$ for each sample:

$$e(n) = s(n) - s'(n). \tag{3.3}$$

The coefficients are determined by solving a set of linear simultaneous equations so as to minimize the mean squared error, E, between the predicted signal and the actual signal.

$$E = \sum_n e^2(n) = \sum_n [s(n) - s'(n)]^2 = \sum_n \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2 \tag{3.4}$$

where n is the number of samples over which the error is to be minimized. We need to find a_k such that

$$\frac{\delta E}{\delta a_j} = -2 \sum_n s(n-j) \cdot \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right] = 0 \tag{3.5}$$

$$j = 1,2,\ldots, p$$

which gives

$$\sum_{k=1}^{p} a_k \sum_n s(n-j) \cdot s(n-k) = \sum_n s(n) \cdot s(n-j) \tag{3.6}$$

$$j = 1,2,\ldots, p$$

A set of p linear equations for the set of p unknowns a_k. The choice of p is a compromise between modeling accuracy and computation time. p is generally therefore between 10-20 orders, and solving this system of equations is not trivial. Two efficient methods exist for finding the solution. The auto-correlation method and the covariance method. Again these are both covered in most signal processing texts and will not therefore be covered here.

Once the predictor coefficients are known they may be used to estimate the vocal tract response.

The error signal may be calculated if the predictor coefficients are known

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{3.7}$$

and it follows that the original signal may be reconstructed if the error signal and predictor coefficients are known:

$$s = e(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{3.8}$$

Taking z-transforms

$$S(z) = E(z) + \left[\sum_{k=1}^{p} a_k z^{-k}\right] S(z) \tag{3.9}$$

$$S(z) = E(z) / \left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) \tag{3.10}$$

$$= E(z) H(z) \tag{3.11}$$

where $E(z)$ and $S(z)$ are the z-transforms of $e(n)$ and $s(n)$. $H(z)$ is the transfer function of an all pole filter and equation 3.11 shows that the speech signal may be viewed as the output of this filter when the error signal, $E(z)$ is input. From a physical point of view, $E(z)$ describes the vocal tract excitation and $H(z)$ the response of the vocal tract. An approximation of the vocal tract response may be obtained by substituting $z = e^{j\omega T}$ in $H(z)$:

$$H(\omega) = 1 / \left(1 - \sum_{k=1}^{p} a_k e^{-j\omega kT}\right) \tag{3.12}$$

and evaluating $|H(\omega)|$ at various values of ω as shown in Figure 3.2.

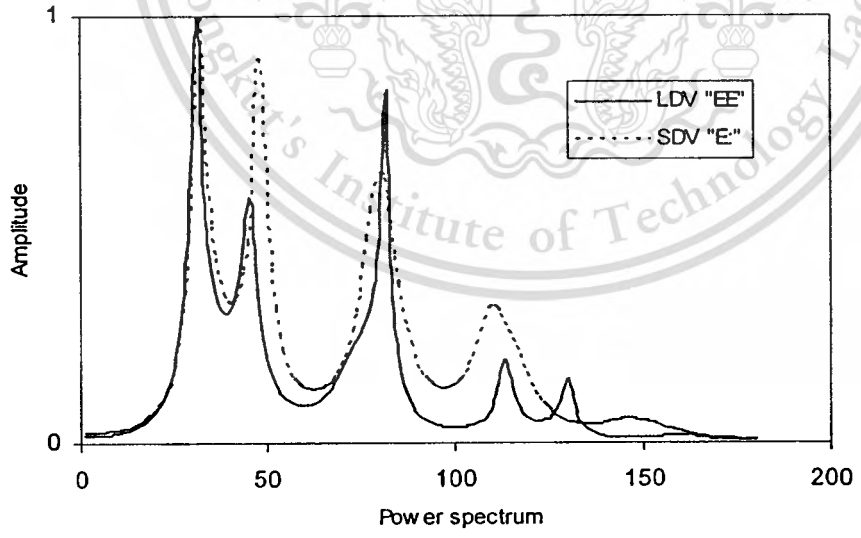


Figure 3.2 Approximation of the vocal tract frequency response obtained using LPC analysis.

3.4 Bark scale

Critical Band Rate scale or Bark scale is a reasonable estimation of the width of the critical band. Below 200 Hz, it seems that the method using the delectability of phase effects when switching from FM to AM is the most reliable one. Although the lowest critical bandwidth in the audible range from 0 Hz to 20 Hz to that 80 Hz, it is attractive to add the inaudible range from 0 Hz to 20 Hz to that critical band, and to assume that the lowest critical bands ranges from 0 Hz to 100 Hz as shown in Figure 3.3 Although there is a small tendency for the critical band to increase somewhat for levels above about 70 dB, the curve given in Figure 3.3 represents a good approximation for critical bandwidth as a function of frequency. The critical bandwidth remains near 100 Hz up to a frequency about 500 Hz. Above that, the critical bandwidth increases a little slower than in proportion to frequency and for frequencies above about 3 kHz a little fast. It is useful to assume constant bandwidth of 100 Hz up to a center frequency of 500 Hz, and a relative bandwidth of 20% for center frequency above 500 Hz. More exact value are given in Table 3.1, which gives the lower and upper limit of the critical bands if they are accumulated in such a way that the upper cut-off frequency of the lower critical band is identical to the lower cut-off frequency of the next higher critical band.

The critical-band concept is important for describing hearing sensations. It is used in so many models and hypotheses that a unit was defined leading to the so-called *critical-band rate* scale. This scale is based on the fact that our hearing system analyses a broad spectrum into parts that correspond to critical band.

Adding one critical band to the next in such a way that the upper limit of the lower critical band corresponds to the lower limit of the next higher critical band, leads to the scale of critical-band rate. If the critical bands are added up this way, then a certain frequency corresponds to each crossing point (see Table 3.1). The procedure is illustrated in Figure 3.3. The first critical band spans the range from 0 to 100 Hz, the second from 100 to 200 Hz, the third from 200 to 300 Hz and so on up to 500 Hz. The frequency range of each critical band increases. Plotting the ordinal number of each critical band lined up as a function of frequency produces a series of dots plotted in Figure 3.4. It can be seen that the audible frequency range to 16 kHz can be subdivided into 24 critical bands. The series of dots does not mean that critical bands exist only between two neighboring dots: rather, they should be though of as able to be shifted continuously along a scale produced by a curve through the dots. The scale produced in this way is called critical-band rate. It grows from 0-24 and has the unit "Bark" (in memory of Barkhausen, a scientist who introduced

the "phone", a value describing loudness level for which the critical band plays an important role). The relation between critical-band rate, z, and frequency, f, is important for understanding many characteristics of the human ear.

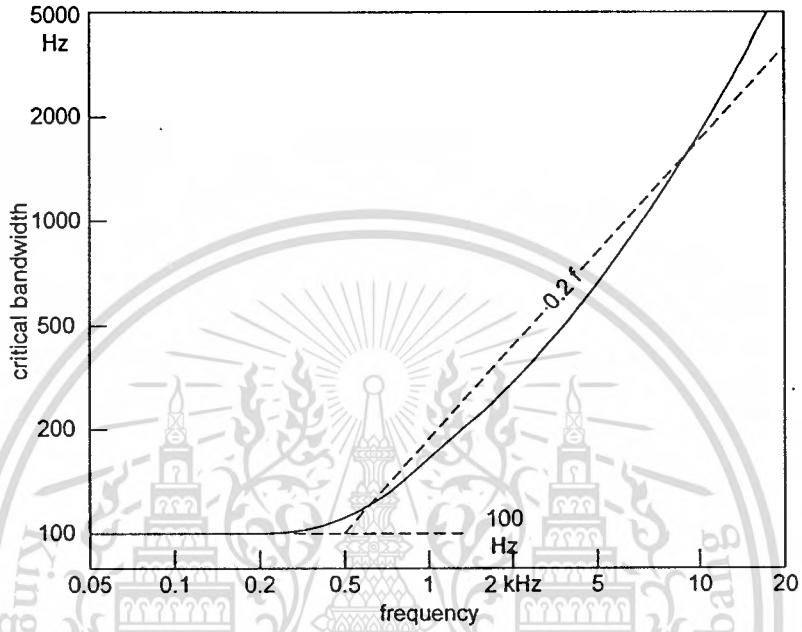


Figure 3.3 Critical bandwidth as a function of frequency. Approximations for low and high frequency ranges are indicated by broken lines

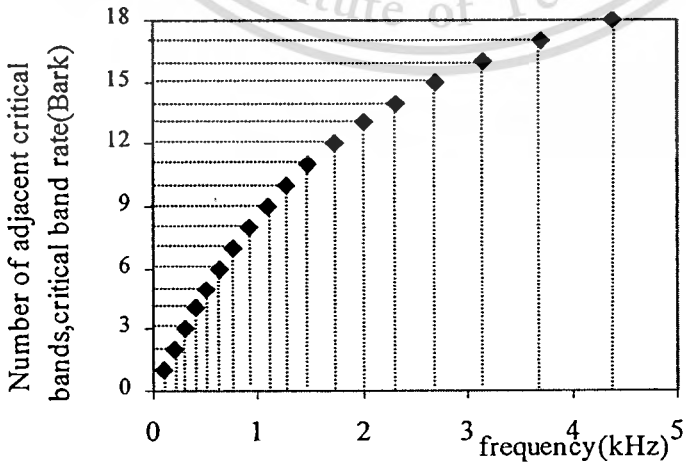


Figure 3.4 Plot of frequencies and Bark equivalents

Table 3.1 Critical band B, Lower (f_l) and Upper (f_u) frequency limit of critical band, Center frequency f_c and Band-Width Δf center at f_c

Z Bark	f_{Lower} / f_{Upper} Hz	f_c Hz	f_c Bark	Bandwidth Hz	Z Bark	f_{Lower} / f_{Upper} Hz	f_c Hz	f_c Bark	Bandwidth Hz
	0					1720			
0		50	0.5	100	12		1850	12.5	280
	100					2000			
1		150	1.5	100	13		2150	13.5	320
	200					2320			
2		250	2.5	100	14		2500	14.5	380
	300					2700			
3		350	3.5	100	15		2900	15.5	450
	400					3150			
4		450	4.5	110	16		3400	16.5	550
	510					3700			
5		570	5.5	120	17		4000	17.5	700
	630					4400			
6		700	6.5	140	18		4800	18.5	900
	770					5300			
7		840	7.5	150	19		5800	19.5	1100
	920					6400			
8		1000	8.5	160	20		7000	20.5	1300
	1080					7700			
9		1170	9.5	190	21		8500	21.5	1800
	1270					9500			
10		1370	10.5	210	22		10500	22.5	2500
	1480					12000			
11		1600	11.5	240	23		13500	23.5	3500
	1720					15500			

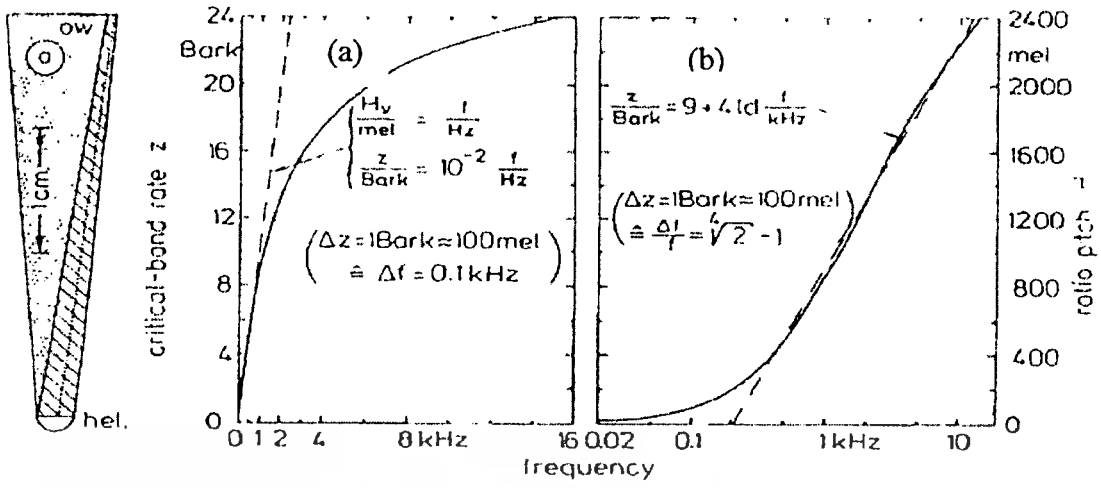


Figure 3.5 (a-b). (a) Frequency on linear scale and (b) Frequency on logarithmic scale

The critical-band rate is closely related to several other scales that describe characteristics of the hearing system. For example, both the just-noticeable increment in frequency and the threshold for frequency modulation are closely related to critical bandwidth. Furthermore, critical bandwidth seems to bear a relation to the function relating frequency to the position of maximal stimulation on the basilar membrane. For comparing critical bandwidth, just-noticeable variations in frequency and the position of maximal stimulation on the basilar membrane, it is convenient to advance in constant step sizes (0.2 mm) along the basilar membrane, and to plot the increment in frequency, Δf , as a function of frequency corresponding to each point. At low frequencies, the step of 0.2 mm leads to a frequency increment of about 15 to 20 Hz. At high frequencies near the oval window, however, a step size of 0.2 mm produces a frequency increment, Δf , of about 500 Hz.

The relationship between frequency on one hand, and length of the basilar membrane or critical-band rate or ratio pitch of tones on the other is important. This relationship is outlined in Figure 3.5 using different frequency scales, one divided linearly and the other logarithmically. Sometimes approximations may be useful, especially if only the low frequency or the high frequency ranges are considered. These approximations, shown as broken straight lines in the figures, are also given numerically. On the left of Figure 3.5, the uncoiled inner ear, including the basilar membrane from helicotrema to oval window, is shown. The dotted line drawn along the center of the basilar membrane may be assumed to be row of the inner hairless. Part (b) shows frequency on linear abscissa scale, and critical-band rate as the ordinate, also on linear scale.

Ratio pitch is given as the ordinate on the right. Because critical bandwidth at low frequencies is 100 Hz, and because frequency and ratio pitch are linearly correlated at low frequencies with the factor of proportionality being unity, the approximation given in part (b) for low frequencies becomes evident: 1 Bark is equal to 100 mel. The approximation of direct proportionality shown as the broken line in part (b) indicates the range within which the ratio pitch in mel is equal to the frequency in Hz. This is the range that governs harmony in music. The critical-band rate in Bark is proportional too, but 100 times smaller than frequency (Hz) in this range. An increment of the critical-band rate of 1 Bark corresponds to a change in ratio pitch of 100 mel.

A logarithmic frequency scale is used in part (c) as the abscissa. The straight broken line indicates that a logarithmic relation between critical-band rate and frequency is a very useful approximation for frequencies above 500 Hz. This approximation leads to the relation between increments in Bark (or increments of 100 mel) to a relative frequency change of about 20%.

In many cases an analytic expression is useful to describe the dependence of critical-band rate (and of critical bandwidth) on frequency over the whole auditory frequency range. The following two expressions have proven useful:

$$z/Bark = 13\arctan(0.76f/kHz) + 3.5\arctan(f/7.5kHz)^2 \tag{3.13}$$

and

$$\Delta f_G/Hz = 25 + 75[1 + 1.4(f/kHz)^2]^{0.69} \tag{3.14}$$

The frequency selectivity of our hearing system can be approximated by subdividing the intensity of the sound into parts that fall into critical bands. Such an approximation leads to the notion of critical-band intensities. If instead of an infinitely steep slope of the hypothetical critical-band filters, the actual slope produced in our hearing system is considered, then such a procedure leads to an intermediate value called excitation. Mostly, these values are not used as linear values but as logarithmic values similar to sound pressure level. The critical-band level and the excitation level are the corresponding values that play an important role in many models as intermediate values. The critical-band intensity, I_G, can be calculated by the following equation that takes into account the frequency dependence of critical bandwidth:

$$I_G(f) = \int_{f-0.5f_G(f)}^{f+0.5f_G(f)} \frac{dI}{df} df \tag{3.15}$$

We have already seen that the critical-band rate is useful in describing the characteristics of our hearing system. Because critical-band rate, z, is a definite function of frequency, (3.15) can also be expressed in critical-band rates:

$$I_G(z) = \int_{z-0.5f_G(z)}^{z+0.5z_G(z)} \frac{dI}{dz} dz . \tag{3.16}$$

In logarithmic expressions, and using $I_0 = 10^{-12}$ W/m^2 as reference value, the critical-band level, L_G, is defined as

$$L_G = 10 \cdot \log \frac{I_G}{I_0} \text{ dB} \tag{3.17}$$

Critical-band intensity can be seen as that part of the overall weighted sound intensity that falls within a frequency window that has the width of a critical band. The transformation of frequency in critical-band rate transfers the frequency-dependent window width into a window width of 1 Bark, independent of critical-band rate. This window of 1-Bark width can be continuously shifted along the critical-band scale. Consequently, a critical-band wide narrow-band noise produces a critical-band intensity which is a function of critical-band rate, and which shows the form of a triangle with a base width of 2 Bark. A sinusoidal tone, however, produces a function with a rectangular shape and the width of 1 Bark.

3.5 Filter Banks

The information needed in each frame is a description of the frequency distribution, i.e. how the power of the signal is distributed over different frequencies. A filter bank, in its simplest form, is a set of band-pass filters with different frequencies covering the interesting part of the spectrum. The output of the filters during a frame can be used as features. The center frequencies of the filters can be chosen in several ways. Usually they are set according to some perceptually motivated scale. The perceived pitch of a sound is not equal to the actual frequency. A popular approximation of the real mapping is the mel scale, the mel frequency (f_{mel}):

$$f_{mel} = 2595 \log_{10}(1 + f/700) \tag{3.18}$$

where f is the actual frequency. An increase in frequency is easier to recognize in the lower register than in the higher.

The human auditory system can not distinguish between frequencies that are close to each other. The higher the frequency the bigger are these critical bands (the intervals within which the frequencies can not be separated from the center frequency). Using for instance the mel frequency the size of the critical band is approximated by:

$$BW_{Critical} = 25 + 75\left[1 + 1.4\left(f_{mel}/1000\right)^2\right]^{0.69} \tag{3.19}$$

In a critical band filter bank the band-pass filters are linearly spaced on a perceptually motivated scale (for instance the mel scale). The bandwidth of the filters are set to the critical bandwidth for the center frequency. That is, they are logarithmically spaced. Figure 3.6 shows the critical band filter bank and mel-filter bank, including the pre-emphasis usually connected to them.

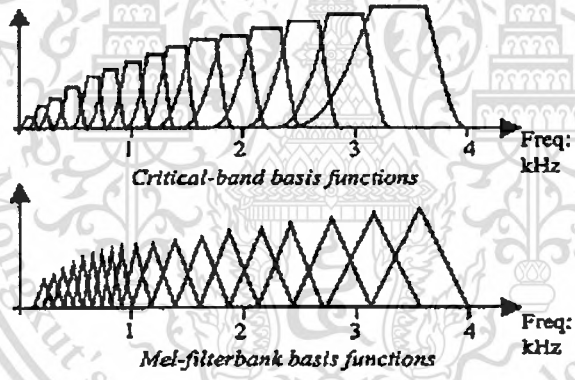


Figure 3.6 Critical-band and mel-filter bank basis functions

Another way to a accomplish a corresponding feature vector is to use the Fourier transform to sample the signal at the desired frequencies (like for instance the centers of the filters in the mel-filter bank).

3.6 Cepstrum coefficients

MFCC is a representation defined by cepstrum of a short-time signal by using discrete furrier transform (DFT) [24]. The DFT of input signal is given by

$$X_a(k) = \sum_{n=0} (N-1)s(n)e^{-j2\pi nk/N}\ ,\ 0 \le k \le N. \tag{3.20}$$

We define a filter bank with M filters ($m = 1,2,...,M$), where filter m is a triangular filter, given energy band by:

$$Y_m(k) = \begin{cases} \dfrac{2(k - f_{mel}(m-1))}{(f_{mel}(m+1) - f_{mel}(m-1))(f_{mel}(m) - f_{mel}(m-1))} \\ \dfrac{2(f_{mel}(m-1) - k)}{(f_{mel}(m+1) - f_{mel}(m-1))(f_{mel}(m) - f_{mel}(m-1))} \\ 0, \qquad \text{Otherwise} \end{cases} \tag{3.21}$$

These filters are computed by the average spectrum of multiple frequency bands. The structure of frequency bands is shown as Figure 3.6. $Y_m(k)$ is defined from the lowest and highest frequency of the mel filter bank.

The contribution from the vocal tract tends to be slowly varying (low frequency) while that from the excitation source is of higher frequency. Hence the contributions are separable by means of a linear filtering operation on the log magnitude spectrum. We compute the log energy at the output of each filter as

$$S(m) = \log\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right),\ 0 < m \le M. \tag{3.22}$$

Taking the inverse transform of the log magnitude spectrum gives the cepstral coefficients of the speech signal. The component due to the periodic excitation source may be removed from the signal by simply discarding the higher order coefficients. The mel frequency cepstrum is the discrete cosine transform (DCT) of the M filter outputs:

$$c(n) = \sum_{k=0}^{N-1} S(m)\cos\left(\frac{\pi n\left(m - \frac{1}{2}\right)}{M}\right),\ 0 < n \le M. \tag{3.23}$$

where M varies for different implementations from 20 to 40. In this study, 12 coefficients are used to feature for speech recognition system.

3.7 Energy Measures

Amplitude of a speech wave is a peak of a speech waveform. In other words, the amplitude is a maximum displacement of a vibration of a mass, which is displaced from its rest position and moving back and forth between two positions that mark the extreme limits of its motion (Denes and Pinson, 1963). In speech recognition, an absolute acoustic energy contour could be directly computed from a short-time analysis of speech waveform using the following relation as

$$E(n)=\sum_{m=1}^{M}\left[s(m)\right]^{2} \tag{3.24}$$

where, $E(n)$ is an absolute energy value of frame n, $s(m)$ are speech sample of frame n, and M is a frame length.

3.8 Addition of dynamic coefficients

In the recognition methods to be described, no use is made of the fact that consecutive frames of speech are likely to be highly correlated, since the articulators may only move a limited distance in the 10 ms gap between frames [25]. Dynamic features, that is, values which attempt to explain the way in which the speech signal is varying between successive frames, such as those presented in [26] is therefore appended to the static coefficients. The following was used to calculate the first order dynamic coefficients, known as velocity, or delta coefficients:

$$D_t=\frac{\sum_{\theta=1}^{N}\theta\left(c_{t+1}-c_{t-1}\right)}{2\sum_{\theta=1}^{N}\theta^{2}} \tag{3.25}$$

where D_t is the delta coefficient at time t and c_t is the static coefficient at time t and N is the width of the window. Since this formula relies on the current samples preceding and subsequent samples, at the beginning and end of the speech the first and last parameters are copied to fill the required regression window. Second order, known as acceleration, or delta-delta coefficients are obtained by applying the same formula to the delta coefficients. In this study the window size used was two.

3.9 Statistical speech pattern recognition

Any method of modeling speech must account for the fact that the information in the signal is carried by the temporal ordering of the sounds. The model must also be able to describe the variation within sounds, while identifying the differences between them. A stochastic process is able to perform both these requirements. Such a method, and one which has become extremely popular in the modeling the speech data is that of hidden Markov modeling (HMM).

Here a brief description of the general principles behind the method is given, followed by a discussion of how HMMs may be used in the classification of unknown speech signals. More detailed descriptions are given in [14, 45, 54, 67, 69].

3.9.1 Application of HMMs to speech recognition

If we make the assumption that the speech articulators, while generating a given sound are moving between a series of target positions, and at each position they generate a characteristic output for a varying length of time, the correlation between this and HMM is clear. Each 'target position' becomes a state in the model, and the 'sound generated' (actually the vector output by our front end at that time frame) is represented by the output symbol. With our present example the outputs must be discreted into a finite number of symbols by, for example, vector quantisation of the speech vectors. The model can however be extended to allow for a continuous set of output symbols defined by a probability density function, which is more appropriate for modeling speech sounds. There are two problems associated with the application of HMMs to speech recognition:

The training problem: Given a set of utterances, labeled at some level of speech unit, generate a set of models (i.e. estimate the values of $\boldsymbol{A}$, $\boldsymbol{B}$ and π) each of which represents one of the units of speech.

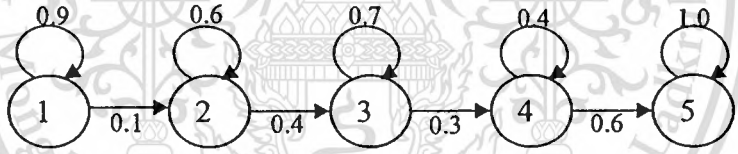
The recognition problem: Given a sequence of speech frames whose classification is unknown, and a set of well trained models, identify the most likely model for each input vector.

The training problem is the more difficult of the two. However an algorithm exists (the Baum-Welch and the Viterbi algorithm) which guarantees to produce a locally optimal model for a given set of training data and recognition. This procedure is covered in detail elsewhere [1],[27] and will not be discussed here. We will assume that a well-trained set of models exist for the speech we wish to recognize. The recognition problem may be solved by means of maximum

likelihood classification. That is we find the model, or series of models which has the highest probability of having produced the given unknown observation sequence.

3.9.2 Hidden Markov Models

Figure 3.7 shows an example hidden Markov model. The model consists of a number of *states*, shown as the circles in Figure 3.7. At time t the model is in one of these states and outputs an *observation*. At $t+1$ time the model moves to another state, or stays in the same state and emits another observation. The transition between states is probabilistic and is based on the transition probabilities between states which are given in the state transition matrix, $\boldsymbol{A}$, where a_{ij} is the probability of being in state i at time t and moving to state j at time $t+1$. Notice that in this case $\boldsymbol{A}$ as upper triangular. While in a general HMM transitions may occur from any state to any other state, for speech recognition applications transitions only occur from left to right i.e., the process cannot go backwards in time, effectively modeling the temporal ordering of speech sounds. Since at each time step there must always be a transition from a state to a state each row of $\boldsymbol{A}$ must sum to a probability of 1.



$$A=[a_{ij}]=\begin{bmatrix} 0.8 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.4 & 0.0 & 00 \\ 0.0 & 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}, \quad B=[b_{jk}]=\begin{bmatrix} 0.1 & 0.8 & 0.1 & 0.0 \\ 0.0 & 0.2 & 0.8 & 0.0 \\ 0.7 & 0.0 & 0.1 & 0.2 \\ 0.0 & 0.2 & 0.3 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

$$\pi=\begin{bmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Figure 3.7. A 5 state left right, discrete HMM with 4 output symbols.

The output symbol at each time step is selected from a finite dictionary. This process is again probabilistic and is governed by the output probability matrix $\boldsymbol{B}$, where b_{jk} is the probability of being in state j and outputting symbol k .Again since there must always be an output symbol at time t, the rows of $\boldsymbol{B}$ sum to 1.

· Finally, the entry probability vector, π , is used to describe the probability of starting in each of the i states of the model. π_i being the probability of starting in state i

The model is fully described by the parameter set $\lambda = [\pi, A, B]$.

3.9.3 The three basic problems of HMM

There are three basic problems of interest that must be solved for the model to be useful. These problems are the following:

3.9.3.1 The evaluation problem

Given the observation sequence $O = o_1, o_2, ..., O_T$, and the model $\lambda = [\pi, A, B]$, how to compute $P(O|\lambda)$, the probability that the observation sequence is produced by the model. This problem can be also viewed as given several competing models and a sequence of observations, how to choose the model which best matches the observations for the purpose of classification or recognition.

3.9.3.2 The decoding problem

Given the observation sequence $O = o_1, o_2, ..., O_T$, and the model $\lambda = [\pi, A, B]$, what the most likely state sequence $Q = q_1, q_2, ..., q_T$ according to some optimality criteria is. This problem is the one to uncover the hidden part of the model to find the correct state sequence. Apart from the degenerate model, there is no correct state sequence to be found. Hence for practical situations, an optimality criterion is employed to solve this problem as best as possible. Unfortunately, there are several reasonable optimality criteria that can be imposed, and therefore, the choice of criterion is a strong function for the uncovered state sequence. Typical uses might be to learn about the structure of the model, to find the optimal state sequences for specific task, or to get average statistics of individual states.

3.9.3.3 The estimation problem

Given the observation sequence $O = o_1, o_2, ..., O_T$, how to adjust the model parameters $\lambda = [\pi, A, B]$, to maximize $P(O|\lambda)$. The problem concerns how to optimize the model parameters so as to best describe how a give observation sequence. The observation sequence used to adjust the model parameters is called a training sequence. The estimation problem is the crucial one for most applications of HMM, since the model parameters can be optimally adapted to observed data for real phenomena.

3.9.4 Solutions to the three basic problems of HMM

3.9.4.1 Solution to the evaluation problem

To calculate the probability of an observation sequence $O = o_1, o_2, ..., O_T$, given the model $P(O|\lambda)$. The most straightforward way is to enumerate every possible state sequence of length T (the number of observations). For every fixed state sequence $Q = \{q_1, q_2, ..., q_T\}$, where q_1 is the initial state. The probability of the observation sequence O for this state sequence is

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda) \tag{3.26}$$

From the output-independent assumption, the observations are assumed statistically independent. This probability can be written as

$$P(O|Q,\lambda) = b_{q1}(o_1)b_{q2}(o_2)...b_{qT}(o_T). \tag{3.27}$$

By applying Markov assumption, the probability of the state sequence Q is

$$P(Q|\lambda) = P(q_1|\lambda)\prod_{t=1}^{T} P(q_t|q_{t-1},\lambda) \tag{3.28}$$

$$= \pi_{q_1} a_{q_1q_2} a_{q_2q_3} ... a_{q_{T-1}q_T} \tag{3.29}$$

$$= a_{q_0q_1} a_{q_1q_2} ... a_{q_{T-1}q_T} \tag{3.30}$$

where, $a_{q_0q_1}$ denotes π_{q_1} for simplicity.

The joint probability of O and Q, which O and Q occur simultaneously, is simply the product of the above two terms

$$P(O,Q|\lambda) = P(O,Q|\lambda)P(Q,\lambda) \tag{3.31}$$

The probability $P(O|\lambda)$ is obtained by summing this joint probability over all possible state sequences q giving

$$P(O|\lambda) = \sum_{allQ} P(O,Q|\lambda)P(Q,\lambda) \tag{3.32}$$

$$= \sum_{allQ} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{3.33}$$

The interpretation of the computation in the above equation is the following. A transition starts from an initial state q_1 with probability $a_{q_0q_1}$, and generates the symbol o_1 in this state with probability $b_{q_1}(o_1)$. Then, a transition is made from the initial state q_1 to state q_2 with transition probability $a_{q_1q_2}$, and generates the symbol o_2 with output probability $b_{q_2}(o_2)$ attached to the corresponding state q_2. This process continues in this manner until the last transition from state q_{T-1} to state q_T with transition probability $a_{q_{T-1}q_T}$, and output probability generating $b_{q_T}(o_T)$ symbol o_T is reached.

The computation of $P(O|\lambda)$, according to its direct definition (Eq.3.33) involves on the order of $O(N^T)$ calculations. At every time $t=1, 2, ..., T$, there are N possible states with can be reached. Therefore there are N^T possible state sequences. This calculation is computationally unfeasible, even for small values of N and T.

Clearly, a more efficient procedure is required to solve the Estimation Problem. Fortunately, such a procedure exists and is called the forward-backward procedure.

3.9.4.1.1 The forward procedure

Consider the forward variable $\alpha_t(t)$ defined as

$$\alpha_t = P(o_1 o_2 ... o_t, q_t = S_i | \lambda) \tag{3.34}$$

This is the probability of the partial observation sequence O and state S_i at time t, given the model λ. This probability can be inductively calculated as follows:

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N \tag{3.35}$$

2. Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{3.36}$$

3. Termination

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{3.37}$$

In the first step, the forward probabilities are initiated as the joint probability of S_i and initial observation o_1. The induction step, which is the most important forward calculation, is illustrated in Figure 3.8. This figure shows how S_j can be reached at time $t+1$ from the N possible states, S_i, $1 \le i \le N$, at time t. Since $\alpha_t(t)$ is the probability of joint event that $o_1, o_2, ..., o_t$ are observed, and the state at time t is S_i, the product $\alpha_t(i)a_{ij}$ is then the probability of joint event that $o_1, o_2, ..., o_t$ are observed, and the state at time t is S_i, the product $\alpha_t(i)a_{ij}$ is then the probability of joint event that $o_1, o_2, ..., o_t$ are observed, and S_j is reached at time $t+1$ via S_i at time t. Summing this product over all the N possible states S_i, $1 \le i \le N$ at time t results in the probability of S_j at time $t+1$ through all the previous partial observations. By multiplying the summed quantity by the probability $b_j(o_{t+1})$, $\alpha_{t+1}(j)$, the probability of the new observation sequence $o_1, o_2, ..., o_t, o_{t+1}$ is obtained in S_j. The computation of the induction step is performed for all S_j, $1 \le j \le N$, for a given t. This computation is then iterated for $t = 1,2,...,T-1$. Finally, the termination step gives the desired calculation of $P(O|\lambda)$ as the sum of the terminal forward variables $\alpha_T(i)$.

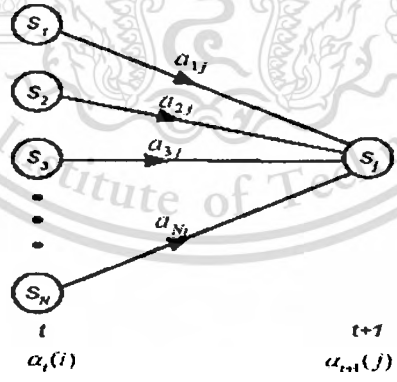


Figure 3.8 Computation of the forward variable

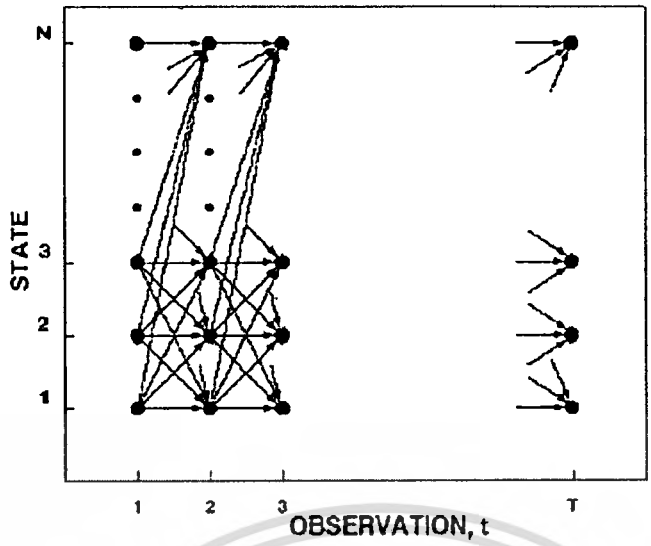


Figure 3.9 Implementation of the computation of $\alpha_t(i)$ in terms of a lattice

The computation in the calculation of $\alpha_t(i)$ requires only on the order of $O(N^2)$ rather than $O(N^T)$ as required by direct calculation. The forward probability calculation is based on the lattice (trellis) structure depicted in Figure 3.9. Since there are only N states (nodes) at each time slot in the lattice, all possible state sequences will remerge in these N nodes, no matter how long the observation sequence. At time t=1, the first time slot in the lattice, the value of $\alpha_1(i)$, $1 \le i \le N$, is calculated. At time t=1, 2,...,T, the only values of $\alpha_t(j)$, $1 \le i \le N$, are needed to compute. Each calculation involves only N previous values of $\alpha_{t-1}(i)$ because each of N grid point is reached from the same N grid points at the previous time slot.

3.9.4.1.2 The backward procedure

In the similar way, a backward variable $\beta_t(i)$ can be defined as

$$\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_T \mid q_t = S_i, \lambda) \tag{3.38}$$

which is the probability of the partial observation sequence from t+1 to the end, given state S_i at time t and the model λ. This backward variable can be also solved inductively in the manner similar to the forward variable as follows:

1. Initialization

$$\beta_T(i) = 1, \quad 1 \le i \le N \tag{3.39}$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \qquad t = T-1, T-2, ..., 1, \ 1 \le j \le N \tag{3.40}$$

The initialization arbitrarily defines $\beta_t(i)$ to be 1 for all i. In order to be in S_i at time t, and to account for the rest observation sequence, a transition from S_i to every one of the possible states at time $t+1$ must be made (the a_{ij} term), which accounts for the observation symbol o_{t+1} in S_j (the $S_j(o_{t+1})$ term), and then accounts for the remaining partial observation sequence from S_j (the $\beta_{t+1}(j)$ term).

The computational complexity of $\beta_t(i)$ is similar to that of $\alpha_t(i)$, which also produces a lattice with observation length and state number. The induction step is illustrated in Figure 3.10. As mentioned above, both the forward and backward procedures can be applied to compute $P(O|\lambda)$ for the evaluation problem. They can also be used together to formulate a solution to the problem of model parameter estimation as discussed in the next section.

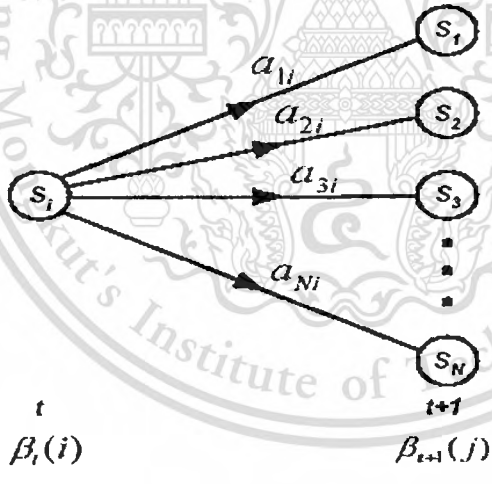


Figure 3.10 Computation of the backward variable

3.9.4.2 Solution to the decoding problem

The hidden part of HMM, which is the state sequence, cannot be uncovered, but can be interpreted in some meaningful ways. A typical use of the recovered state sequence is to learn about the structure of the model, and to get average statistics within individual states. There are several possible ways to find the optimal state sequence associated with the given observation sequence. One possible optimality criterion is to choose the states q_t, which are in the best path

with the highest probability. A formal technique for finding this single best state sequence is called the Viterbi algorithm, which is very similar to the Dynamic Time Warping (DTW) algorithm.

Firstly, the variable $\gamma_t(i)$, the probability of being in state S_i at time t, given the model λ and the observation sequence, is defined as

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) \tag{3.41}$$

This variable can be simply expressed in terms of the forward-backward variables as

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{3.42}$$

$\alpha_t(i)$ accounts for the partial observation sequence $o_1, o_2, ..., o_t$ and the S_i at time t, while $\beta_t(i)$ accounts for the remainder of the observation sequence $o_{t+1}, o_{t+2}, ..., o_T$ and the S_i at time t. The normalization factor $P(O \mid \lambda)$, makes $\gamma_t(i)$ a probability measure so that

$$\sum_{i=1}^{N} \gamma_t(i) = 1. \tag{3.43}$$

Using $\gamma_t(i)$, the individually most likely state q_t at time t can be solved as

$$q_t = \arg\max_{1 \le i \le N} [\gamma_t(i)], \quad 1 \le t \le T \tag{3.44}$$

Although the above equation maximizes the expected number of correct states by choosing the most likely state for each t, there could be some problems with the resulting state sequence. For example, when the HMM has state transitions, which have zero probability, the optimal state sequence may not even be a valid state sequence. This problem occurs because the solution in equation (3.44) simply determines the most likely state at every instant, without regard to the probability of occurrence of sequences of states.

One solution to the above problem is to modify the optimal criterion. The most widely used criterion is to find the single best state sequence to maximize $P(Q, O \mid \lambda)$, which is

equivalent to maximizing $P(Q,O|\lambda)$. A formal technique for finding this single best state sequence is called the Viterbi algorithm.

3.9.4.2.1 The Viterbi algorithm

To find the single best state sequence, $Q=\{q_1,q_2,...,q_T\}$, for the given observation sequence $O=\{o_1,o_2,...,o_t\}$ the quantity $\delta_t(i)$ is needed to define

$$\delta_t(i)=\max_{q_1,q_2,...q_{t-1}} P[q_1q_2....q_{t-1}=S_i,o_1o_2...o_t|\lambda] \tag{3.45}$$

where $\delta_t(i)$ is the best score along a single path at time t, which accounts for the first t observations and end in S_i. By induction, the equation (3.45) becomes to

$$\delta_{t+1}(j)=\left[\max_i \delta_t(i)a_{ij}\right]b_j(o_{t+1}) \tag{3.46}$$

To actually retrieve the state sequence, we need to keep track of the argument that maximized equation (3.46), for each t and j. We do this via the array $\psi_t(j)$. The complete procedure for finding the best state sequence can now be stated as follows:

1. Initialization

$$\delta_1(i)=\pi_i b_i(o_1), \quad 1\le i\le N \tag{3.47}$$

$$\psi_1(i)=0, \quad 1\le i\le N \tag{3.48}$$

2. Induction

$$\delta_t(j)=\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right]b_j(o_t), \quad \begin{matrix}2\le t\le T\\ 1\le j\le N\end{matrix} \tag{3.49}$$

$$\psi_t(t)=\arg\max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}\right], \quad \begin{matrix}2\le t\le T\\ 1\le j\le N\end{matrix} \tag{3.50}$$

3. Termination

$$P^*=\max_{1\le i\le N}\left[\delta_T(i)\right] \tag{3.51}$$

$$q_T^*=\arg\max_{1\le i\le N}\left[\delta_T(i)\right] \tag{3.52}$$

4. Path (state sequence backtracking)

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right),\ t = T-1, T-2, \ldots, 1 \tag{3.53}$$

The Viterbi algorithm (except for the backtracking step) is similar in implementation to the forward calculation. The major difference is the maximization over the previous states in equation (3.53), which is used instead of the summing procedure of the forward variable calculation. Moreover, a lattice or trellis structure efficiently implements the computation of the Viterbi procedure.

3.9.4.3 Solution to the estimation problem

The most difficult problem in HMM is to determine a method to adjust the model parameters $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$ maximize the probability of the observation sequence given the model. There is no known way to analytically solve for the model, which maximizes the probability of the observation sequence. Actually, given any finite observation sequence, there is no optimal method of estimating the model parameters. However, by choosing $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \pi)$. $P(O \mid \lambda)$ is locally maximized, an iterative algorithm or gradient technique for optimization is used. In this section, one iterative algorithm known as Baum-Welch algorithm is described.

3.9.4.3.1 Baum-Welch re-estimation algorithm

The mathematical foundations of the Baum-Welch algorithm for the maximum likelihood estimation. An iterative method for monotonically increasing value of an arbitrary homogeneous polynomial $P(X)$ with non-negative coefficients of degree d in variables x_{ij}, $i = 1,2,\ldots p$, $j = 1,2,\ldots q$, $j = 1,2,\ldots,q$, defined over a stochastic domain, $x_{ij} \geq 0$, $\sum_{j=1}^{q_t} x_{ij} = 1$, through a series of transformations performed on $\{x_{ij}\}$, was firstly purposed. The transformation is defined as

$$T\left(x_{ij}\right) = \frac{x_{ij}\dfrac{\partial P(X)}{\partial x_{ij}}}{\sum_{j=1}^{q_t} x_{ij}\dfrac{\partial P(X)}{\partial x_{ij}}} \tag{3.54}$$

and is often referred to a growth transformation of $P(X)$. A special case of the re-estimation procedure for probabilistic functions of Markov chains with discrete observations was described.

Later, the method was generalized to functions of Markov chains with continuously distributed observations. Recently, an analysis, which extends the algorithm to accommodate a large class of distributions and mixture distributions, was presented. For the discrete output distribution, transition and observation parameters are both re-estimated according to equation (3.54) in the following. However, the re-estimation formulas for the parameters of a continuous density HMM will be described later.

The purpose of the solution to the estimation problem is to obtain the model from observations. If the model parameters are known, the forward-backward algorithm can be used to evaluate probabilities produced by given model parameters for given observations.

In order to describe the procedure for re-estimation of HMM parameters, $\xi_t(i,j)$ the probability of being in S_i at time t and S_j at time $t+1$, given the model and observation sequence, is introduced.

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) \tag{3.55}$$

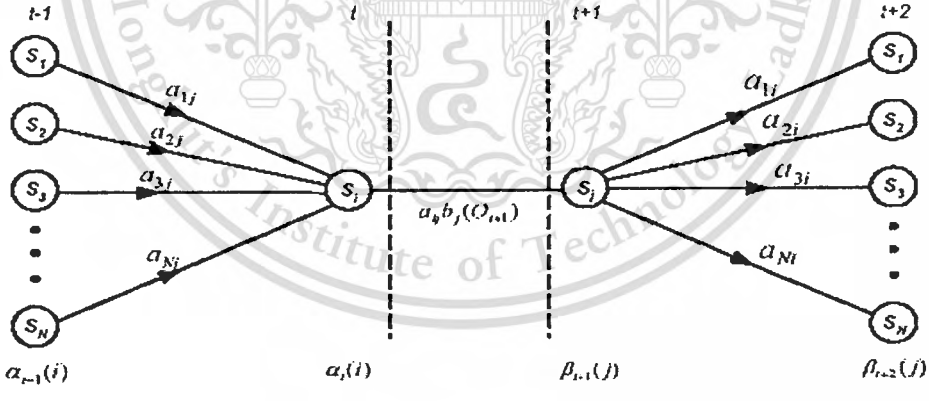


Figure 3.11 Computation of the joint event that the system is in S_i at time t and S_j at time $t+1$

The sequence of events leading to the conditions required by equation (3.55) is illustrated in Figure 3.11. From the definition of the forward and backward variables, $\xi_t(i,j)$ can be written in the form

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O \mid \lambda)} \tag{3.56}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{t=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \tag{3.57}$$

where the numerator term is just $P(q_t = S_i, q_{t+1} = S_i \mid O, \lambda)$ and the division by $P(O \mid \lambda)$ gives the desired probability measure.

Since $\gamma_t(i)$, the probability of being in state S_i at time t, given the observation sequence and the model, is previously defined $\xi_t(i,j)$, can be related to $\gamma_t(i)$ by summing over j, giving

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{3.58}$$

If $\gamma_t(i)$ is summed over the time index t, a quantity, which can be interpreted as the expected number of times that state S_i is visited, or equivalently the expected number of transitions made from S_i, is obtained. Similarly, summation of $\xi_t(i,j)$ over t from $t=1$ to $t=T-1$ can be interpreted as the expected number of transitions from S_i to S_j. That is

$$\sum_{t=1}^{T-1}\gamma_t(i) = \text{expected number of transitions from } S_i \tag{3.59}$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) = \text{expected number of transitions from } S_i \text{ to } S_j. \tag{3.60}$$

Using the above formulas and the concept of counting event occurrences, a method for re-estimation of the HMM parameters is given. A set of re-estimation formulas for $\boldsymbol{A}$, $\boldsymbol{B}$, and π are

$$\bar{\pi}_i = \text{ expected frequency in state } S_i \text{ at time } \gamma_1(i) \tag{3.61}$$

$$\bar{a}_{ij} = \frac{\text{expected number of transition from } S_i \text{ to } S_j}{\text{expected number of transitions from } S_i} \tag{3.62}$$

$$= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{3.63}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in } S_j \text{ and observing symbol } v_k}{\text{expected number of times in } S_j} \tag{3.64}$$

$$= \frac{\sum_{t=1}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}_{s.t.\, o_t = v_k}. \tag{3.65}$$

From equation (3.61) to (3.65), it can be proven that either:

1. The initial model λ defines a critical point of likelihood function, where new estimates equal old ones,
2. Model $\bar{\lambda}$ is more likely than model λ in the sense that $P(O|\bar{\lambda}) \geq P(O|\lambda)$

Thus, if $\bar{\lambda}$ is iteratively used to replace λ and repeats until the above re-estimation calculation, $P(O|\lambda)$ can be improved until some limiting point is reached. The final result of this re-estimation procedure is call a maximum likelihood estimation of the HMM. It should be pointed out that the forward-backward algorithm leads to local minimum only, and that in the most problems of interest, the optimization surface is very complex and has many local minimum.

3.9.5 Continuous density HMM

If the observation does not come from a finite set, but from a continuous space, the discrete output distribution discussed in the previous sections can be extended to the continuous output probability density function (Maneenoi E., et al., 2003). This implies that the vector quantization technique, which maps observation vectors from the continuous space to the discrete space, is no longer necessary. Consequently, the inherent error can be eliminated. The Baum-Welch re-estimation algorithm discussed in subsection 3.8.4), can be extended to estimate continuous probability density function with the auxiliary Q function. The generalized method to continuous output density functions can be applicable to the Gaussian, Poisson, and Gamma distributions but not to the Cauchy distribution. Furthermore, the estimation algorithm is expanded to cope with finite mixtures of strictly log concave and elliptically symmetric density functions. This section will discuss general re-estimation formulas for the continuous HMM, which is applicable to a wide variety of elliptically symmetric density functions.

Using continuous probability density functions, the first candidate for a type of output distributions is the multivariate Gaussian, since

1. Gaussian mixture density functions can be used to approximate any continuous probability density functions in the sense of minimizing the error between two density functions.

2. By the central limit theorem, the distribution of the sum of a large number of independent random variables tends towards a Gaussian distribution.

3. The Gaussian distribution has the greatest entropy of any distribution with a given variance.

The most commonly used distribution is the continuous Gaussian density function defined as

$$N(O,\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{1}{2}(O-\mu)^r\sum^{-1}(O-\mu)} \tag{3.66}$$

where n is the dimensionality of the observation vector O, μ and Σ are the mean vector and the covariance matrix respectively. The advantage of normal distributions is that the parameters of Gaussian can be easily and reliably estimated from a large number of data. In order to obtain more accurate approximations, Gaussian mixtures are used. With enough components, such mixtures can approximate any density function with an arbitrary precision. The probability density of the multiple Gaussian mixtures is defined as

$$b_j(0_t)=\sum_{m=1}^{M}c_{jm}N(o_t;\mu_{jm},\Sigma_{jm}) \tag{3.67}$$

where M is the number of mixture components and m is the mixture weight for the mixture component in state j. The mixture weights satisfy the stochastic constraint

$$\sum_{m=1}^{M}c_{jm}=1, \qquad 1\le j\le N \tag{3.68}$$

$$c_{jm}\ge 0, \qquad 1\le m\le M \tag{3.69}$$

For the continuous probability density functions, the likelihood of an input observation is expressed as

$$P(O|\lambda) = \sum_{allQ} P(O,Q|\lambda) \tag{3.70}$$

$$= \sum_{allQ} P(Q,\lambda)P(O|Q,\lambda). \tag{3.71}$$

An information-theoretic Q-function, which is considered a function of $\bar{\lambda}$ in the maximization procedure, is applied to derive the re-estimation formulas as

$$Q(\lambda,\bar{\lambda}) = \frac{1}{P(O|\lambda)} \sum_{allQ} P(O,Q|\lambda)\log P(O,Q|\bar{\lambda}). \tag{3.72}$$

By using an auxiliary Q-function, re-estimated HMM parameters for the multi-modal Gaussian distributions are

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)}{\sum_{t=1}^{T}\sum_{m=1}^{M} \gamma_t(j,m)} \tag{3.73}$$

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m) o_t}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{3.74}$$

$$\bar{\Sigma}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)(o_t - \mu_{jm})(o_t - \mu_{jm})'}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{3.75}$$

where prime denotes vector transpose and $\gamma_t(j,m)$ is the probability of being in state j at time t with the m^{th} mixture component for o_t

$$\gamma_t(j,m) = \left[\frac{\alpha_t(i)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(i)\beta_t(j)}\right]\left[\frac{c_{jm}N(O,\mu,\Sigma)}{\sum_{m=1}^{M} c_{jm}N(O,\mu,\Sigma)}\right]. \tag{3.76}$$

The term $\gamma_t(j,k)$ generalizes to $\gamma_t(j)$ of equation (3.41) in the case of a simple mixture, or a discrete density. The re-estimation formula for a_{ij} is identical to the one used for discrete

observation densities. The interpretation of equation (3.73-3.76) is fairly straight forward. The re-estimation formula for c_{jm} is the ratio between the expected number of times the system is in the state j using the m^{th} mixture component, and the expected number of times the system is in state j. Similarly, the re-estimation formula for the mean vector μ_{jm} weights each numerator term of equation(3.74) by the observation, giving the expected value of the portion of the observation vector accounted for by m^{th} mixture component. A similar interpretation can be given for the re-estimation term for the covariance matrix Σ_{jm}.

Chapter 4

Feature Extraction Techniques

This chapter introduces the feature extraction and robust technique for automatic speech recognition (ASR). Feature extraction is an important part of ASR system that transforms speech into vectors of features that are suitable for further processing. The purpose of this extraction is to parameterize the incoming speech signal. There are two main processes. Firstly, we represent the speech spectrum on a Bark scale, tone analysis using pitch quantization techniques and regression on the voice energy. Secondly, we improve the running spectrum filtering (RSF) techniques, frequency response masking (FRM) techniques, and dynamic range adjustment (DRA) to the extracted features from the speech signal.

4.1 Extraction techniques based on Bark scale

The speech data are first segmented into frame of 300 samples where its time length is 27.21 ms with 11.025 kHz sampling rate. Figure 4.1 shown the proposed system for speech spectrum on the Bark scale. Each frame of speech is represented by the parameters vector from (DCT(log(CBI)). This features are the applied with Hidden Markov Modeling technique for training and recognition.

The Bark scale is a psychoacoustics measurement on human hearing property and speech features extraction processes consists of four steps : (1) speech preprocessing, (2) auto-regressive model (AR model), (3) critical band intensity (CBI), (4) discrete cosines transform on the logarithm of CBI (DCT(log(CBI)).

4.1.1 Speech Pre-processing

Signal processing is vitally important for optimal speech recognition. The purpose of signal processing is to derive a set of parameters to represent speech signals in form, which is suitable for consequential processing. Various techniques of signal processing and feature extraction are commonly used for speech recognition. However, only some of those techniques, which are correspond to the framework of this thesis. The thesis proposed a process for speech recognition as follows.

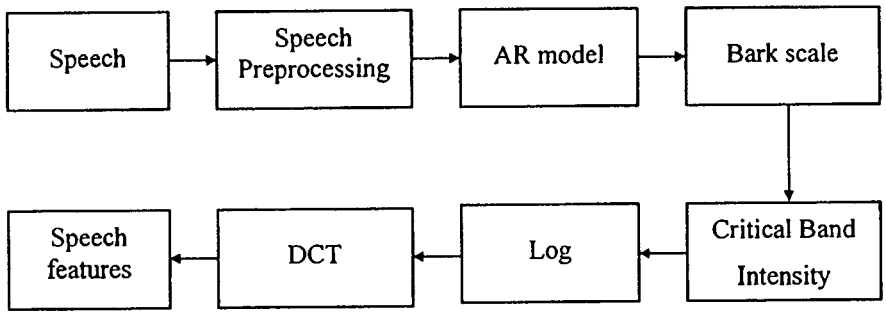


Figure 4.1 Block diagram of a proposed speech feature on a Bark scale.

Pre-emphasis: The pre-emphasis is applied to smooth spectrum of input speech signal $s(n)$. The input signals are passed though first order FIR filter transfer function defined by (4.1) and (4.2), where α is the coefficient of filter, $\tilde{s}(n)$ is the product of Signal pre-emphasis at sequence n and $s(n)$ is the input speech signal at sequence n.

$$H(z)=1-\alpha z^{-1} \tag{4.1}$$

$$\tilde{s}(n)=s(n)-\alpha s(n-1) \tag{4.2}$$

The coefficient of filter α, has the value in the range from 0.95 to 0.99 [1]. It is recommended at 0.97[23] Comparison of the pre-emphasized speech waveform and original speech waveform are indicated in Figure 4.2.

Frame Blocking: speech waveform is blocked into frame of N samples, with shifting every M samples for each frame. This process continues until all the speech data is accounted for within once or more frames. l is the frame index and L is the total number of frame:

$$x_l(n)=\tilde{s}(Ml+n), \quad \begin{matrix} n=1,2,...,N \\ l=0,1,2,...,L-1 \end{matrix} \tag{4.3}$$

The short-term analysis principle is an accepted approach to speech processing. The speech signal changes continuously due to the movements of vocal system, and it is intrinsically non-stationary. Nonetheless, in short segments, typically 20 to 40ms, and overlap of 50% to 75%, speech could be regarded as pseudo-stationary signal [23].

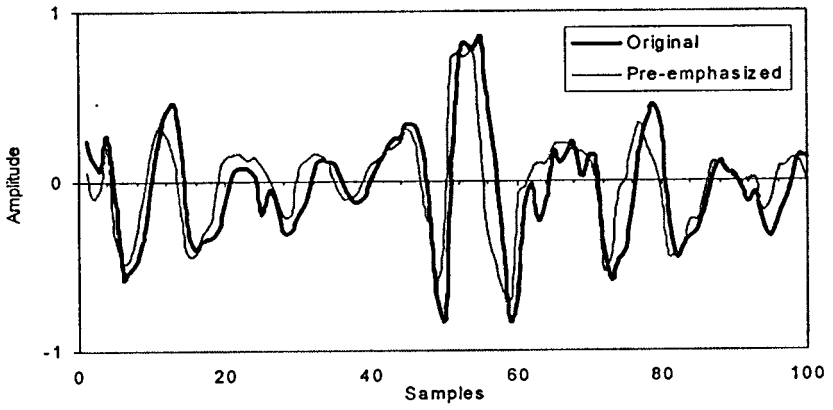


Figure 4.2 Example of Pre-emphasized speech waveform

In this thesis, we have defined the frame to have a duration of 27.21 ms, with an overlap of 1/3 of the frame duration. The sampling rate of the speech data is 11.025 kHz. The corresponding value of N and M are respectively, 300 and 100 samples for optimal performance of speech recognition in our experiments.

Windowing: Hamming window (4.5) is applied to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. The concept is identical to the one discussed with regard to the frequency domain interpretation of short-time spectral analysis. Which depends on windowing of speech waveform. The results depend on the properties of the specific window function. With a window of finite time duration, the window can move progressively along the speech signal to select short sections for analysis.

Consider $w(n)$ as a window function, when $o \le n \le N-1$, where N is window size. The extracted signal with window function can be defined by

$$\tilde{x}_l(n) = x_l(n)w(n), \quad o \le n \le N-1 \tag{4.4}$$

Since, *Hamming Window* is famously used as the window function of speech analysis. The hamming window is given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad o \le n \le N-1 \tag{4.5}$$

The Figure 4.3 shows the output signal of windowing process with the frame blocking: $N = 300$ samples

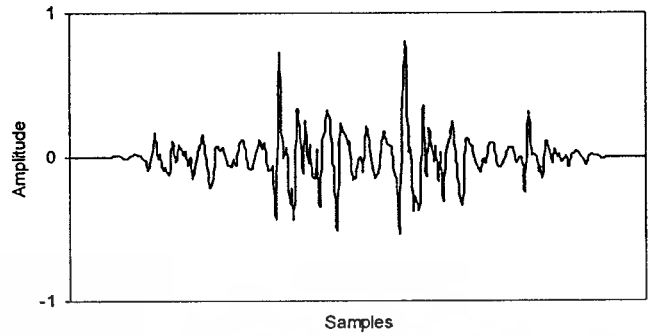


Figure 4.3 Example signal of windowing processed

4.1.2 Auto-Regressive Model (AR model)

In the linear acoustics model of speech production , speech spectrum envelops are calculated from auto-regressive model (AR model) using minimum mean square estimation (MMSE) as described in section 3.3. The speech signal is produced by filtering an excitation signal with a time-varying linear filter (the vocal tract) $H(z)$ as shown in Figure 4.4. An Auto-Regressive (AR) Model can be described by the transfer function

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{4.6}$$

The magnitude $\left|H\left(e^{j\omega}\right)\right|$ models the spectrum of the analyzed signal, when $z^{-1} = e^{-j\omega}$ and $\omega = 2\pi f$. The value f denotes frequency [Hz]. This model is commonly used in linear predictive coding (LPC) as described in section 3.3. In Equation (4.6) p denotes the model order. The coefficients a_k ($k = 1,..., p$) are calculated from a frame of N samples of the input signal.

The magnitude represents the spectrum envelop of the vocal tract. we can estimate the frequency selectivity of the hearing system by approximating its critical band intensities (CBI). The critical band intensity is mapping as a spectrum envelop on Bark scale to speech feature [9],[10]. As an example for Laotian, the spectrum envelops for short duration vowel (SDV) "E:" and long duration vowel (LDV) "EE" are shown in Figure 4.5.

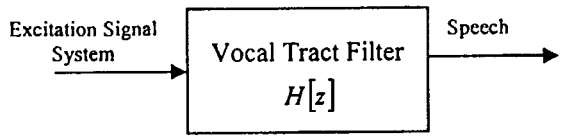


Figure 4.4 Linear prediction model of speech

4.1.3 Bark scale and Critical Band Intensity (CBI)

The Bark scale is a psychoacoustics spectrum measure whose property corresponds to human hearing. In other words, it is based on the fact that our hearing system analyzes speech with critical bands intensity (CBI). The concept of critical band has been developed [6]-[10]. CBIs are extracted to speech feature in this thesis as show in Figure 4.1.

Some experiments have shown that critical bands are narrower at the region of low frequencies than at the region of high frequencies. The critical bands are analogous to the band of a spectrum analyzer with variable center frequencies and bandwidth as described in section 3.4.

The range of human auditory frequency spreads from 20 to 20,000 Hz. It covers approximately 25 critical bands on Bark scale. In this thesis, the underlying sampling rate is set to be 11,025 kHz with a bandwidth of 5.5 kHz. Accordingly, there are 18 critical bands as listed in Table 3.1. The mapping between frequency scale and Bark scale is given in Figure 3.4, the CBI of the *m-th* Bark (α_m) can be calculated by

$$\alpha_m = \int_{f_{l,m}}^{f_{u,m}} \frac{d\alpha}{d\beta} d\beta \tag{4.7}$$

where $f_{l,m}$ and $f_{u,m}$ are the lower and upper band frequencies of the m-th critical band, respectively. α is defined as the intensity of voice in the critical band and it is represents by the spectrum energy of MMSE spectrum envelop. In other words, α_m represents the integrated power of MMSE spectrum envelops in the m-th critical band.

As an example for Laotian vowel, the spectrum envelops and CBI on Bark scale for short duration vowel "E:" and long duration vowel "EE" are shown in Figure 4.5(a) and (b), respectively. The spectrum of MMSE is shown in Figure 4.5 (a). The corresponding 18 CBI are shown in Figure 4.5 (b). According to Figure 4.5(a) and (b), we can see that, there are no large difference between them.

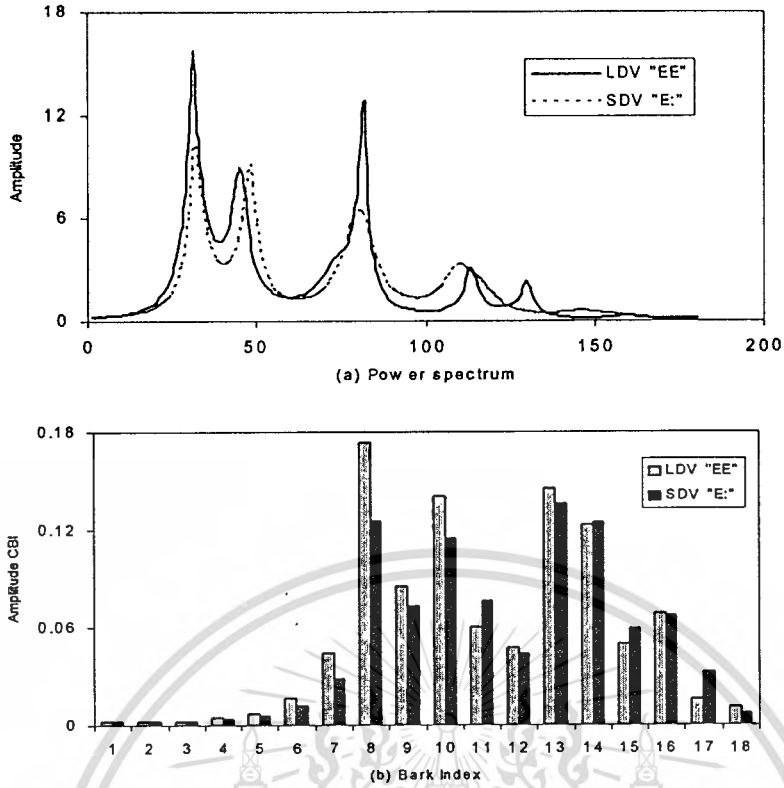


Figure 4.5 Examples of Short-Duration Vowel /E/ and Long-Duration Vowel /EE/. (a) Speech spectrum envelope, (b) Critical Band Intensities

4.1.4 Logarithm CBI on discrete cosines transform (DCT(log (CBI))) and delta

In the linear acoustics model of speech production the composite speech spectrum, as measured by z -transfer function. We refer to CBIs computed critical band. The CBI is very important characteristics of speech feature. The dynamic range of CBI indicates the difference between maximum and minimum of values in critical band. However, The dynamic range is compressed using an amplitude compression scheme by logarithm. The log CBI δ_m as

$$\delta_m = \log_{10}(\alpha_m) \tag{4.8}$$

where m is parameter feature vector is a set of Bark scale (m =1 to 18 on 1-5.5kHz).

A logarithm function, it is generally believed to be sensitive to certain types of noise and signal distortions. The amplitude of $\log_{10}$ CBI is reduced in maximum value comparing amplitude of CBI and increase in minimum value as show Figure 4.6.

In final operation of extraction processes, DCT is applied to δ_m to feature vector.

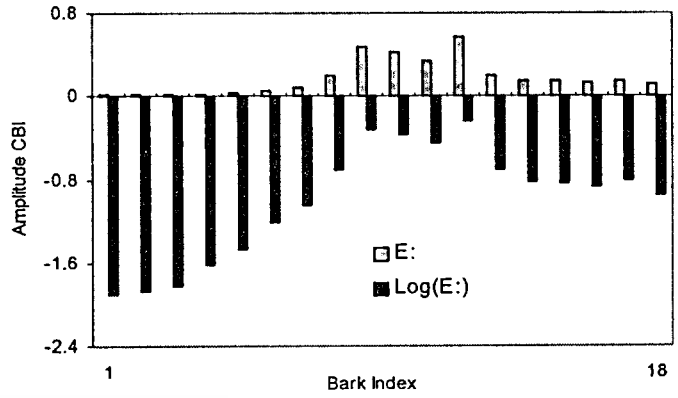


Figure 4.6 Critical Band mapping between linear frequency scale and Bark scale

$$\mu_m = \gamma_m \sum_{k=0}^{M-1} \delta_k \cos\left(\frac{\pi(2k+1)m}{2M}\right), \quad m = 0,1,\ldots,M-1 \tag{4.9}$$

$$\gamma_m = \begin{cases} \sqrt{\frac{1}{N}}, & m = 0 \\ \sqrt{\frac{2}{N}}, & m \neq 0 \end{cases}$$

where M is the number of CBI. In our case, $M = 18$.

The DCT basis were modified with speech features. The spectrum at low order of coefficient was emphasized over that at high order of coefficient. The spectrum was applied to simulate human hearing. DCT vector is chosen as basis function for its high energy compression ratio and energy preservation feature.

The performance of a speech recognition system can be greatly enhanced by adding time derivatives to the basic static parameters. The delta coefficients are computed using the following regression formula as described in section 3.7. The value of θ is set using the configuration parameter. This thesis used θ =2. The component of feature vector the calculate eighteen DCT(log(CBI)) and eighteen differential DCT(log(CBI)).

4.2 Tone analysis

Thai and Laotian has been classified to be monosyllabic which uses tone feature of the syntax as part of the semantic as described in section 2.3. There are 5 intonation levels as shown in Figure 2.1. The different tone occurs from the different rate of vibration of the vocal fold in

human's vocal tract. The rate of this vibration is represented by the fundamental frequency, F_0 of the audio signal. If F_0 is constant, the utterance will be monotonic. If F_0 changes, the utterance will have a changing tone, as in musical voice.

This thesis proposes a method for implementing a tonal recognition model for Thai and Laotian speech recognition. the pitch or F_0 is calculated for each frame of audio signal using auto-correlation with center clipping method. After which, the sequence of F_0 is preprocessed using median filter to filter out unnecessary discontinuities, followed by quantization of data sequence in term of raising, decreasing or remain the same state in the sequence. Each frame of speech is represented by the parameters vector from quantized pitch as shown in Figure 4.7. The sequences are used as additional component in the feature vectors. This features are applied to HMM for training and recognition.

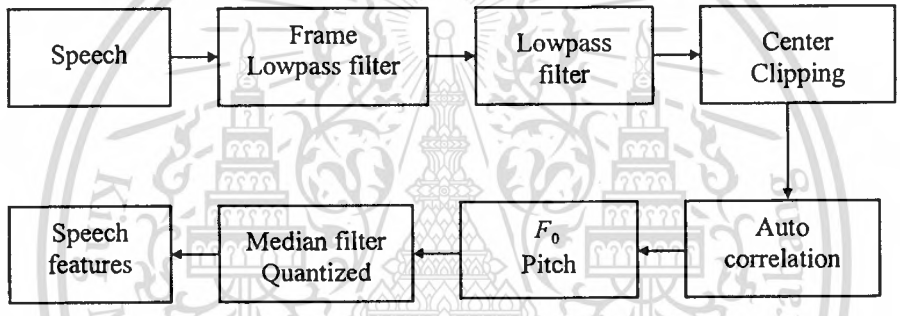


Figure 4.7 Block diagram of quantized pitch analysis.

4.2.1 Fundamental frequency and Pitch

In Figure 4.7, the sequence of audio signal, $s(n)$ is first segmented into frame (300 samples) with each frame overlapped the next by 200 samples. Each frame as shown in Figure 4.8(a) is into low-pass filter with bandwidth of 900 Hz to reduce the damping oscillation effect of the vocal tract response as show in Figure 4.8(b). The low-passed signal is then center-clipped as shown in Figure 4.8(c), to locate only the signal's peak. The clipping threshold value, T_c is a fixed percentage (60%) of the smaller maximum absolute values [12]. The signal $s_c(n)$ is then clipped using the following conditions.

$$s_c(k)=\begin{cases} 1 & s(k) > T_c \\ 0 & |s(k)| \le T_c \\ -1 & s(k) < -T_c \end{cases} \tag{4.10}$$

The auto-correlation of the clipped signal is calculated to find the pitch for each frame. It is

$$r(k) = \sum_{i=0}^{N-1-k} s_c(i)s_c(i+k) \tag{4.11}$$

where $k = 1,2,...,N$. The value N is the number sample in each frame.

The auto-correlation function of the periodic signal has the same periodicity to the observed periodic signal. Let us assumed its period is P. The local maximum of the auto-correlation function occurs at k = 0, P.... The highest auto-correlation peak $r(0)$ is located at k = 0 and the next highest peak $r(P)$ is located at k = P. The fundamental frequency is thus calculated as $F_0 = F_s / P$, where F_s is the sampling frequency of the signal. The fundamental frequency of each tone for Laotian are shown Figure 4.9 (a).

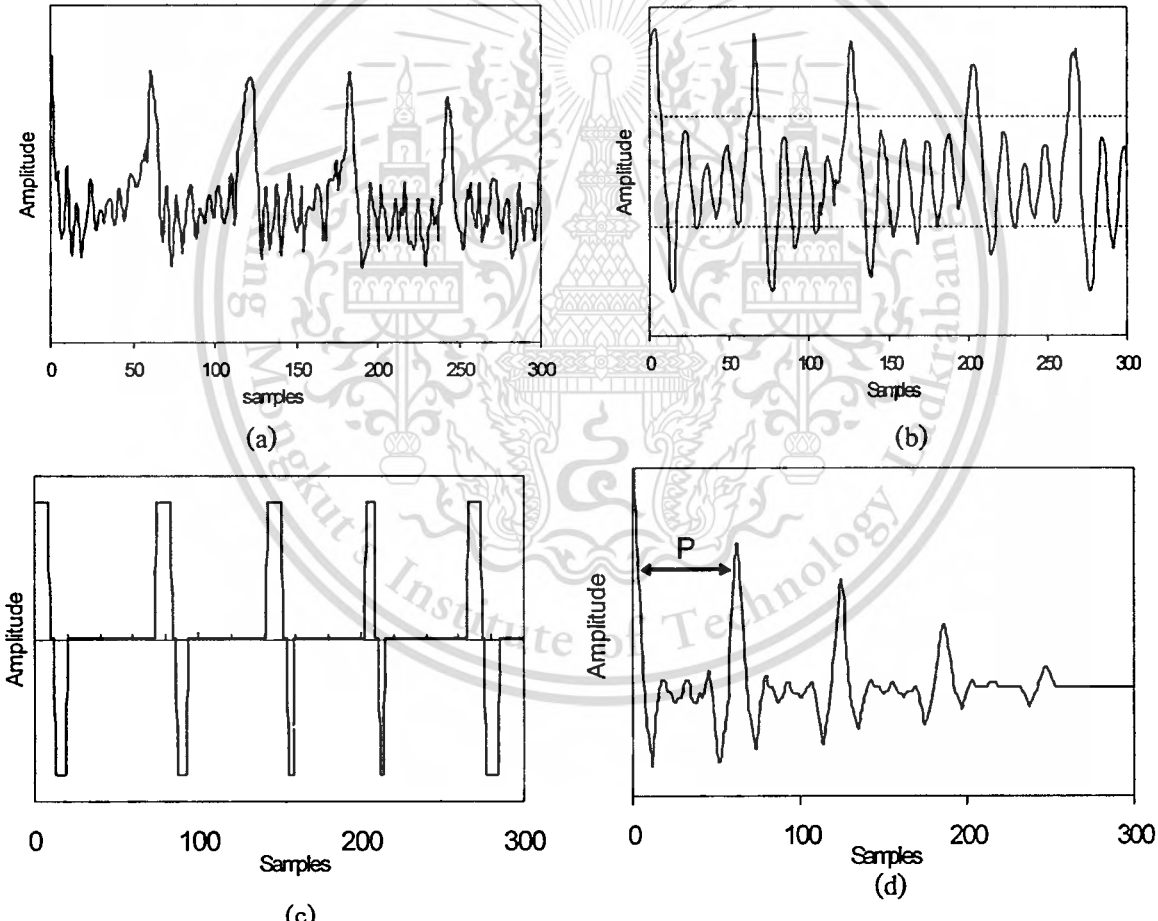


Figure 4.8 The sequence of the audio signal analysis, (a) original signal, (b) set clipping level, (c) center clipping and (d) locate of pitch.

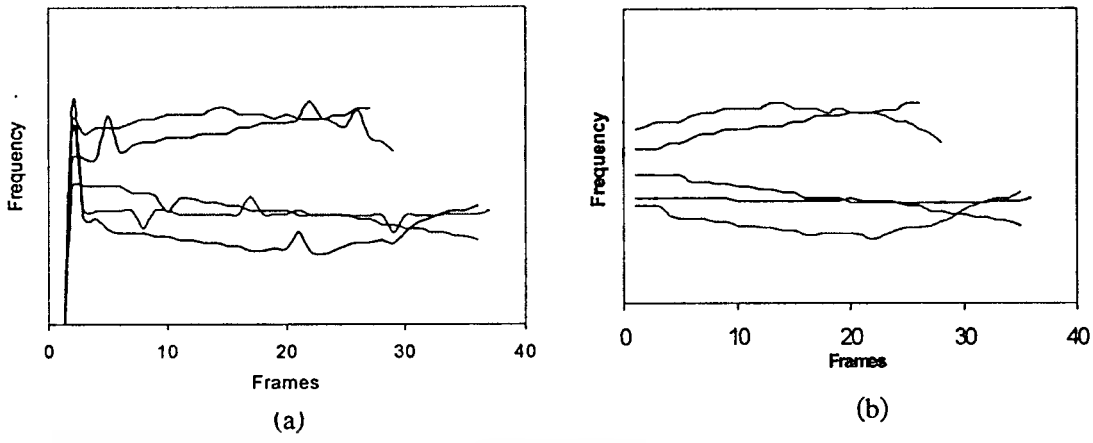


Figure 4.9 Fundamental frequencies, (a) original F_0, (b) median filter F_0

In Figure 4.9(a), the contours is not smooth. The median filtering helps enhancing the continuity in the sequence of the fundamental frequencies. The median filter is done by the following.

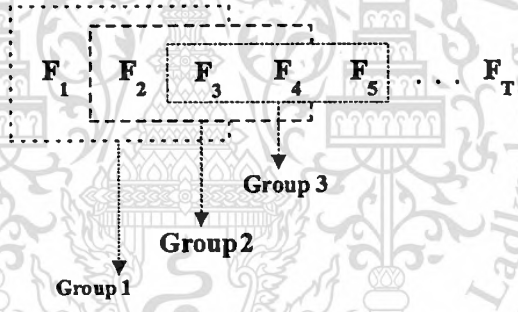


Figure 4.10 Separate three F_0 to group

where F_n is fundamental frequency (F_0) of n-th frame $1 \le n \le T$; T is the number of frames. In the Figure 4.10, suppose the three value of F_0, in each group are a is a, b and c, where value a is assumed minimum of F_0, b is assumed value between a and c, and c is assumed maximum of F_0. Then filtered F_0 is selected to be when $a \le b \le c$ as shown in Figure 4.9(b). The effect of median filtering is shown in Figure 4.9(b). The discontinuities at the beginning of the sequence disappear after median filtering.

The sequence of F_0 as in Figure 4.9(b) is then quantized into three levels. The three levels are defined as +1(positive change), 0(no change) and -1(negative change). The sample of quantized pitch sequences for five tones is shown in Figure 4.11. The quantization process is implemented to

extract relative changes of pitch contour. As an important performance, it can realize gender free model of the recognition engine.

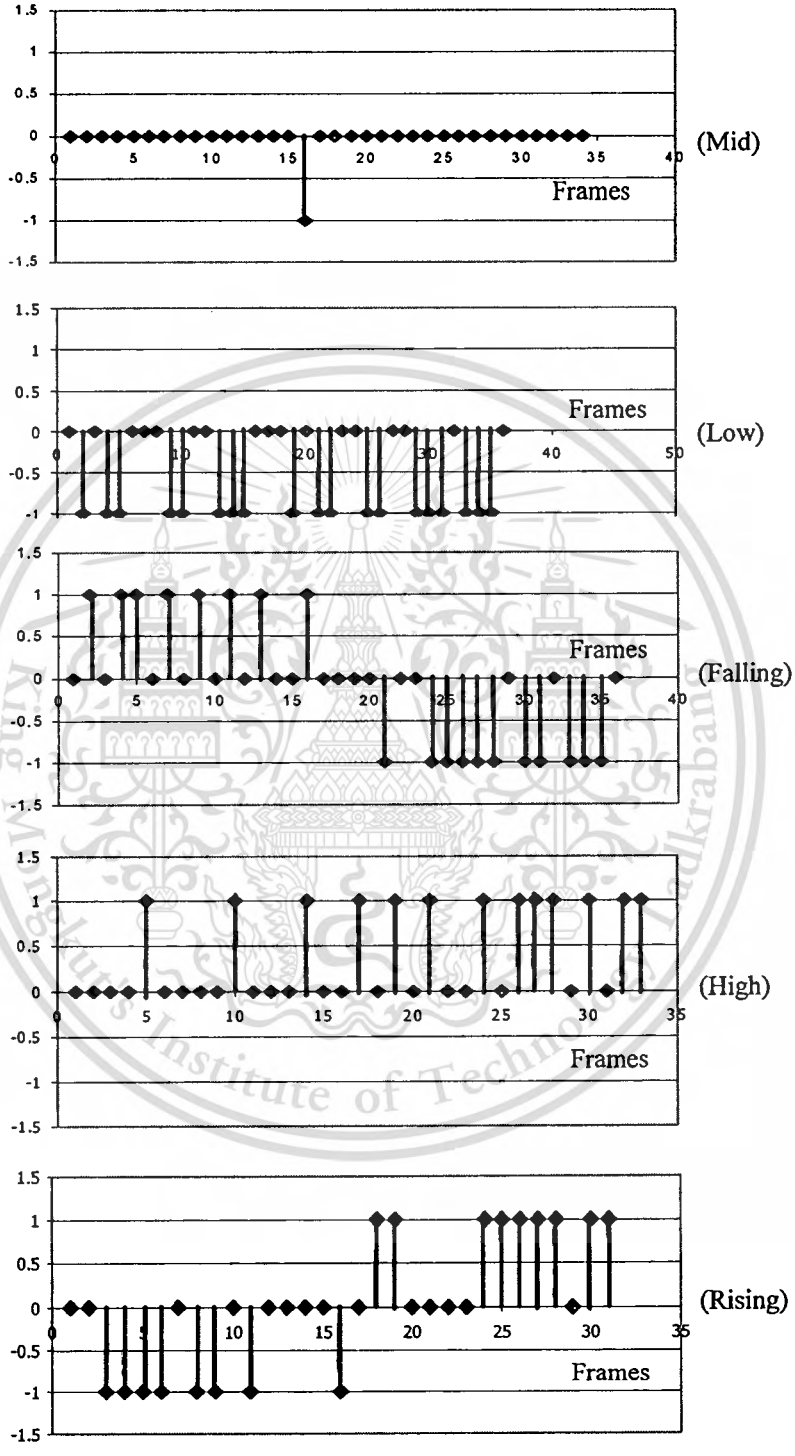


Figure 4.11 Quantized pitch sequences for five tones

4.3 Regression on the voice energy

In some conventional speech recognition system, the time variations in energy are calculated as feature parameter. In addition, its regression combined with cepstrum coefficients has also been used in the classification of speaker-independent isolated word [29].

In this thesis, a regression method on the speech energy is applied and the estimated coefficient is used as a feature to classify either smooth or erupted and of each word and, hence, the short or long vowels in Thai/Laotian spoken language can be distinguished.

The energy of each speech frame is calculated from observed speech data. Assume E_t is the energy in frame, $t = 1,2,...T$. T is the number of frames. The energy is given as

$$E_t = \frac{1}{N}\sum_{i=1}^{N} s_t^2(i) \tag{4.12}$$

where N is the number of samples in the frame.

In the polynomial regression, the parameter of a polynomial is estimated to fit a trajectory of E_t. The set of E_t in Figure 4.12 is one of the examples. The points are approximated by polynomial function.

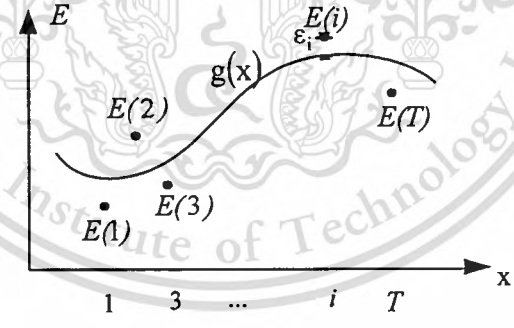


Figure 4.12 Polynomial regression

Let us assume the estimated polynomial is

$$g(j) = a_0 + a_1 j + a_2 j^2 + ... + a_r j^r \tag{4.14}$$

where $a_0, a_1, ..., a_r$ are the polynomial coefficients. r is the order of polynomial.

The polynomial is estimated by minimizing the following criterion:

$$\varepsilon = \sum_{j=1}^{T} [E(j) - g(j)]^2 \tag{4.14}$$

Using (4.12) and (4.14), $a_i(i = 0,1,...,r)$ are estimated optimally when ε is minimized. The estimated parameters satisfy

$$\begin{bmatrix} T & \sum_{j=1}^{T} j & \cdots & \sum_{j=1}^{T} j^r \\ \sum_{j=1}^{T} j & \sum_{j=1}^{T} j^2 & \cdots & \sum_{j=1}^{T} j^{r+1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=1}^{T} j^r & \sum_{j=1}^{T} j^{r+1} & \cdots & \sum_{j=1}^{T} j^{2r} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_r \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{T} E(j) \\ \sum_{j=1}^{T} jE(j) \\ \vdots \\ \sum_{j=1}^{T} j^r E(j) \end{bmatrix} \tag{4.15}$$

The polynomial regression on the energy of speech represents the slope of the time function of each word. The order r of the polynomial regression is determined as $r = 2$ in our preliminary experiment. In addition, the corresponding feature parameter is derived from the 2nd order regression function and it is used as one component of a feature vector in every frame. By using this method, the proposed system can eliminate some noise and personal factors embedded into the time trajectory of speech power.

Examples for the voice energy of short and long duration vowels are shown in Figure 4.13. The pause of SDV is more immediate than that of LDV. The example of second order polynomial regression parameters to the voice energy for SDV and LDV is shown in Figure 4.13. Each curve represents different trend. The curve fitted to SDV is steeper than that of LDV. The regression parameter represents the changes of the time function of the voice energy in a syllable.

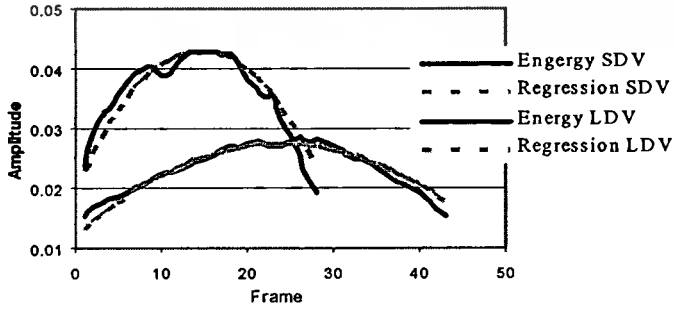


Figure 4.13: Energy and its regression parameters

Chapter 5

Robust speech recognition

As robust speech recognition under additive noisy circumstances, some conventional method have been already proposed [15]-[20]. A linear filtering method reduces additive white noise by using a winner filter. As a widely used method, a spectral subtraction method (SS method) analyzed a noise spectrum and then an estimated noise spectrum is subtracted from an observed noisy speech spectrum. As other sophisticated methods, a hidden markov model (HMM) is used in robust speech recognition where HMM has been used as a modeling method for stochastic speech features. A speech/noise HMM technique is designed with the combination of a speech HMM and a noise HMM on the spectrum domain. A HMM de-convolution analyzes speech data by using a speech HMM model and a noise HMM model separately. Using these methods, noise robust speech recognition has been realized.

On the other, robust methods for convolution noise, i.e., a relative spectral processing (RASTA) and a cepstrum mean subtraction method (CMS method), have been developed [30]-[33]. The convolution components for observed speech are considered as some distortions of a recorder and propagation/radiation. A convolution noise can be represented as an additive component in the log spectrum domain. RASTA employs a low order infinite impulse response (IIR) filter for the reduction of convolution noises. In CMS, an estimated noise spectrum is eliminated on the log spectrum domain.

With recent development of speech recognition systems, various related products have been presented, e.g., a car navigation. However, when the developed speech recognition systems are used under real environments, the actual recognition performance is considerably deteriorated by background noises and unexpected microphone characteristics. Accordingly, several noisy robust systems have been proposed [15-20].

When we consider the property of noises, there are two categories of noises, i.e., an additive noise such as several recorded acoustic sounds and a system noise which indicates a transfer function on a microphone recording system. In this thesis, the modulation spectra which influence speech recognition performance greatly are explored and the important characteristics for speech recognition are emphasized on the modulation spectrum domain.

In addition to the above conventional methods, we have proposed other noise robust

techniques [31],[32]. These techniques are used for both of additive and convolution noises. They are based on FIR filtering in running spectrum domain which is a kind of speech power spectrogram. It was called as a running spectrum filtering (RSF) method. Using this method, we can realize noise robust speech recognition. In particular, under 0 dB- 20 dB SNR circumstances its performance is recognition as higher than other conventional methods. Over 20 dB SNR, its performance is a little higher or same as others.

The RSF method requires usually high calculation cost since a high order FIR filter is repeatedly used many times in several stages. When we apply this method to speech recognition, there is an issue on the calculation cost. Normally a specially designed hardware, e.g., full custom LSI has been required and we have designed a low power LSI system [33] in order to execute this method within real time. When such LSI systems can not employed, the cost reduction should be considered. As a valuable approach for the cost.reduction of a filter, frequency response masking (FRM) techniques have been developed [34]-[36]. Using this technique, we can realize sharp transition of a filter with low calculation cost. However, when the band-pass (BP) filter whose transition areas are sharp and whose properties of right and left edges are different should be required, e.g., BP filter of RSF, conventional design techniques are not suitable for such desired design.

RSF techniques is first explored. The merit of nonlinear RSF is discussed at the theoretical point of view. In addition, in order to reduce the total calculation cost of RSF, a new FRM technique is proposed. Using this modified FRM, RSF can be designed with low calculation cost.

5.1 Modulation spectra and noise circumstances

The whole observed speech data are grouped into frames. Each frame including speech signals is determined as 15msec- 25msec length normally. In many cases, a frame is overlapped the next frame with 50% length. Once a frame is given, the short time Fourier spectrum can be calculated in a frame. Since there are many frame in observed data, we can get many spectra or the running spectrum domain stands for time series of their calculated spectra on a frame axis. On the running spectrum domain, there are time waveforms at all frequencies. For the specific time series at the fixed frequency, a modulation spectrum can be calculated from the time series by using Fourier transform. For a specific time waveform at a fixed frequency, a modulation spectrum can be calculated that is the Fourier transform as follows Figure 5.1.

When the Fourier spectra of an additive noise, a convolution noise and a speech signal are defined as $A_i(n)$, $H_i(n)$ and $S_i(n)$, respectively, an observed signal $X_i(n)$ can be represented as

$$X_i(n) = H_i(n)[S_i(n) + A_i(n)] \tag{5.1}$$

where i and n represent the $i-th$ frequency component of FFT spectrum and the $n-th$ frame, respectively. The $H_i(n)$ means the distortion in a speech recorder and any propagation/ radiation. The power spectrum of (5.1) is given by

$$\begin{aligned} |X_i(n)|^2 &= |S_i(n)H_i(n) + N_i(n)|^2 \\ &= |S_i(n)H_i(n)|^2 + |N_i(n)|^2 + 2|H_i(n)||S_i(n)||N_i(n)|\cos(\theta_i(n)) \end{aligned} \tag{5.2}$$

where $N_i(n)$ is $H_i(n)\,A_i(n)$. The value of $\theta_i(n)$ is phase difference between $S_i(n)$ and $A_i(n)$. In this paper, $|X_i(n)|^2$ is called the $n-th$ running power spectrum where $n = 1,2,\ldots,N$ and N is the total number of speech frames.

If consider the Fourier spectrum of $|X_i(n)|^2$ on $n = 1,2,\ldots,N$ where frequency component i is fixed, the following modulation spectrum can be calculated:

$$X_{i,s}^m = \sum_{n=1}^{N} |X_i(n)|^2 e^{\frac{-j2\pi sn}{N}} \tag{5.3}$$

where s is the $s-th$ modulation frequency component and $s = 1,2,\ldots,N$

Because of the additive noise, i.e., equation(5.2) and (5.2), the energies of all modulation frequency band increase. It stems from the terms of $2|H_i(n)|^2|S_i(n)||A_i(n)|\cos(\theta_i(n))$ and $|N_i(n)|^2$. Especially, a large power is added on low frequencies, i.e., 0Hz-1Hz, in the modulation frequency domain when the component of $|N_i(n)|^2$ is approximately constant. In other words, the time trajectory of $|N_i(n)|^2$ dose change rapidly and thus $|N_i(n)|^2$ is located on about 0Hz-1Hz in the modulation frequency domain.

Equation (5.2) is converted in the running log-power spectrum domain:

$$\log|X_i(n)| = \log|H_i(n)| + \log|S_i(n) + A_i(n)| \tag{5.4}$$

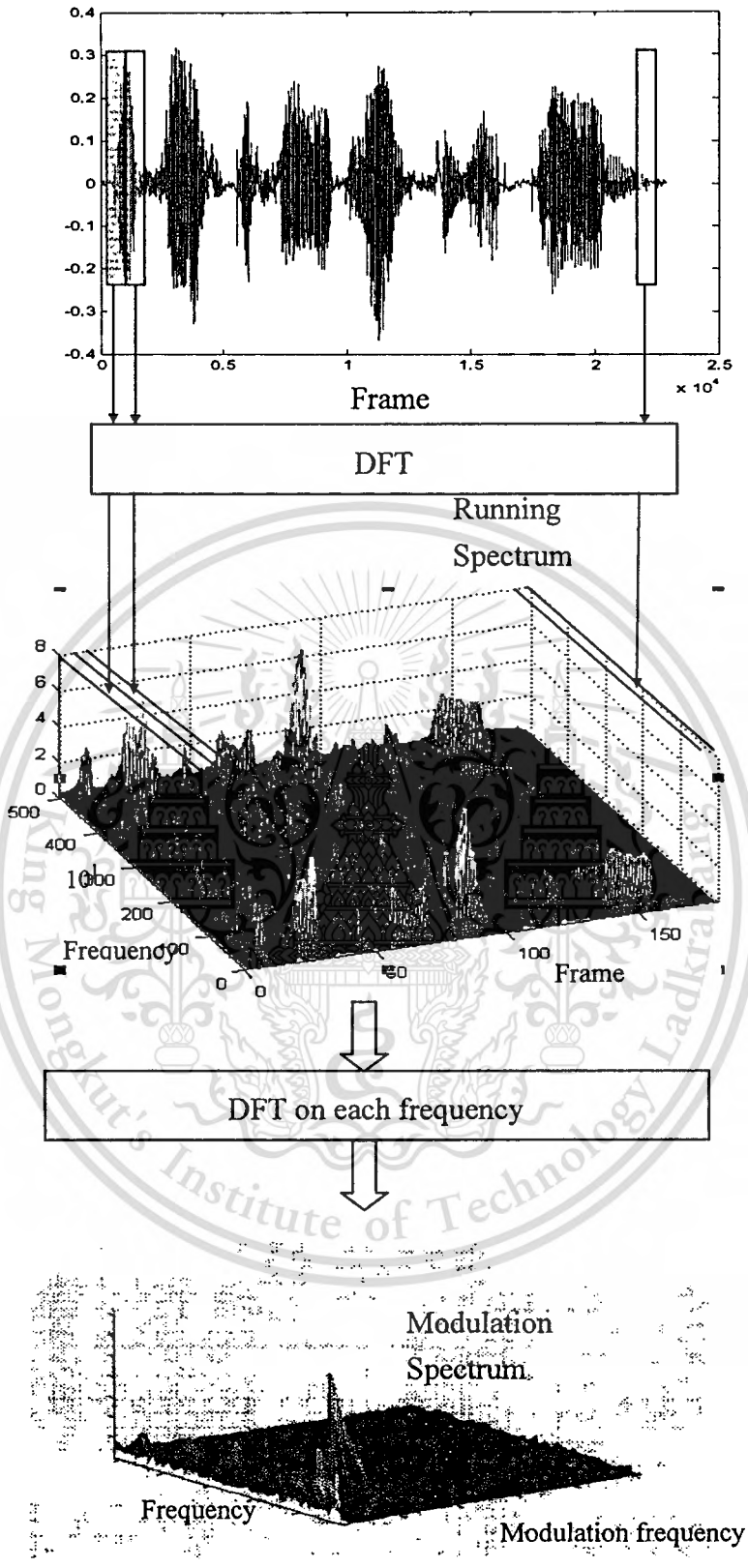


Figure 5.1: The process for obtaining modulation spectrum [17]

The tendency in modulation log-power spectra is the same as in modulation power spectrum. The energy in about 0Hz-1Hz increases. It stems from the convolution noise, i.e., log $|H_i(n)|$. If time trajectory of log $|H_i(n)|$ can be assumed to be constant when we compare it with time trajectory of $\log|S_i(n)+A_i(n)|$, its energy is located at 0Hz-1Hz in the modulation frequency domain.

The modulation spectrum of speech usually concentrates its energy in the band less than 10 Hz in the modulation frequency domain [16]-[20]. It turns out that the quite important part for speech recognition can be discriminated from others on the modulation power/log-power spectra. Almost all of important speech features used for speech recognition exist into 1Hz-10Hz and the useful properties on the variation of every phoneme are located in 2Hz- 4Hz. According to some speech recognition experiments, even if only the band under about 7 Hz is considered, the recognition performance is still good enough [31]-[33].

5.2 Running spectrum filter (RSF)

RSF focuses on modulation spectrum which shows the characteristics of time trajectory on each frame. time speech characteristics in frequency domain are obtained by applying windowing and Fourier Transform to speech waveform in time domain. Therefore, the time trajectory in specific frequency is obtained by tracing its values in each time. The time trajectory of value in frequency domain is the running spectrum, and what is obtained from its frequency analysis is the modulation spectrum. Most of the important information required in speech recognition resides in the range between 1 and 16 Hz in the modulation frequency [16]. The usage of band-pass filter with these frequency bands can suppress noise components. Taking account of filter group delays and computation cost, relative spectra (RASTA) [37] employing low-order IIR is implemented in conventional systems. However, RASTA occurs the degradation of recognition performance by phase deformation in some cases. Our proposed RSF employing a high order FIR separates speech and noise components sharply with a narrow pass-band and has better performance than RASTA. Figure 5.2 shows the modulation frequency properties in RASTA and RSF where the filter order of RASTE and RSF is 20 and 240, respectively. It has been reported that the components under 1Hz cause the degrading of recognition performance in noisy environments. RSF can reduce these components considerably.

The important properties of speech in the modulation spectrum domain are described as follows:

5.2-1. The part on 2Hz- 7Hz band is important on speech since it relates to the frame variation of phonemes.

5.2-2. The influence of additive noise covers almost whole frequency band on the modulation power spectra. Especially, there are a lot of additive noise components on 0Hz- 1Hz band in the modulation power spectra.

5.2-3. The convolution noise seems to be time invariant and it concentrates within 0 Hz – 1 Hz in modulation log-power spectra.

Under the consideration of the above properties, we can get high robust speech recognition system when the effective filtering technique can be applied to the running spectrum and log-spectrum. At first, we apply a law-pass filter to time trajectory of running power spectrum to suppress additive noise, the property of (5.2-2) is assumed for this processing. Secondly, we apply a band-pass filter to time trajectory of running log-power spectrum to suppress convolution noise, the property of (5.2-3) is assumed to this processing.

It has been reported [38] that speech components in modulation frequency domain are dominant around 4Hz and out of the range from 1Hz up to 12Hz can be regarded as noise and unnecessary components. RSF realizes effective feature extraction and can be applied in practical speech recognition system. The process of RSF is as follows. Noisy speech signal $X(n)$ as (5.1) converted to frequency domain by FFT as

$$X(n) = H(n)S(n) + H(n)A(n) \tag{5.5}$$

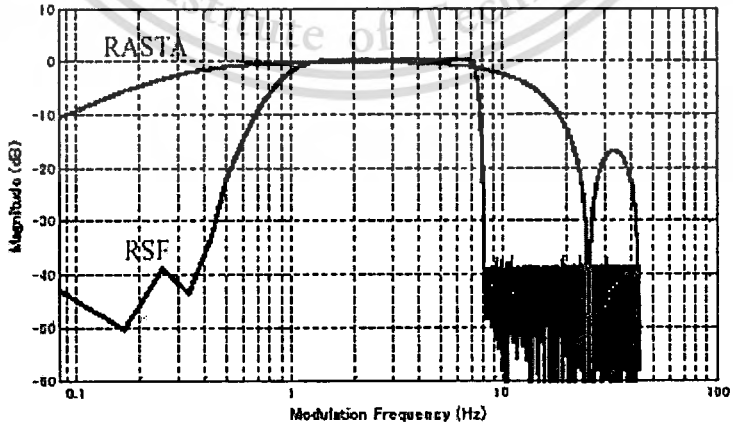
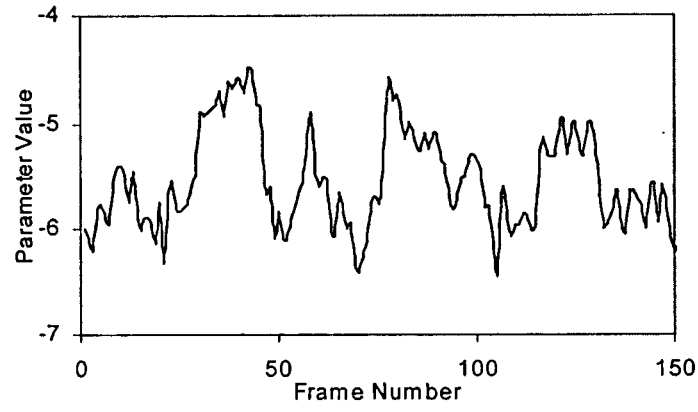
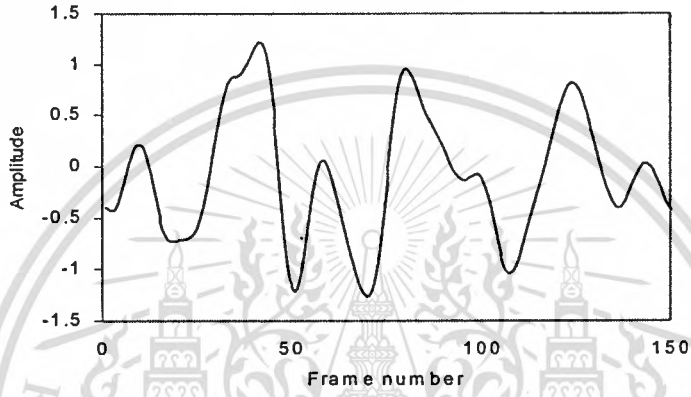


Figure 5.2 Modulation frequency properties in RASTA and RSF.



(a)



(b)

Figure 5.3 Indicates speech feature of the 1-th order of DCT(log(CBI)), (a) value of speech feature in 0 dB SNR and (b) feature after RSF.

In (5.5), $H(n)A(n)$ is additive noise component and the time trajectory of its spectrum is slower than that of speech component. Therefore, it can be removed with low-pass filtering on time spectrum domain. Then the logarithmic power spectrum without the additive noise component is written as

$$\begin{aligned} \log|X(n)| &= \log|H(n)S(n)| \\ &= \log|H(n)| + \log|S(n)|, \end{aligned} \tag{5.6}$$

and this system noise component $H(n)$ can be removed by applying band-pass filtering to the time trajectory of logarithmic power spectrum.

The speech signal with SNR of 0dB is shown in Figure 5.3(a). After the process of RSF, a cleaner speech features is shown in Figure 5.3(b). RSF helps to remove unnecessary parts (speaker characteristic and background noise) of the speech required for recognition.

5.3 Frequency response masking (FRM) technique [15-20]

FRM finite impulse response (FIR) filters can be designed from model filters and masking filters. Since the model filters are interpolated and masked by the masking filters, FRM filters have very narrow transition bandwidth shown in Figure 5.4. Even if we consider a model filter with broad transition shown in Figure 5.4(a), it is possible to get a model filter with narrow transition in (b) and (c) and the same number of coefficients by using an interpolation technique. For example, assume that the filter of (a) was designed with 20 order FIR filter. If the interpolation was 3, i.e., M=3, we can design a FIR filter with 1/3 narrower transition band than its original band and, in addition, the number of FIR filter coefficients was still 20. It means that we can design sharp transition band FIR filter with low calculation cost.

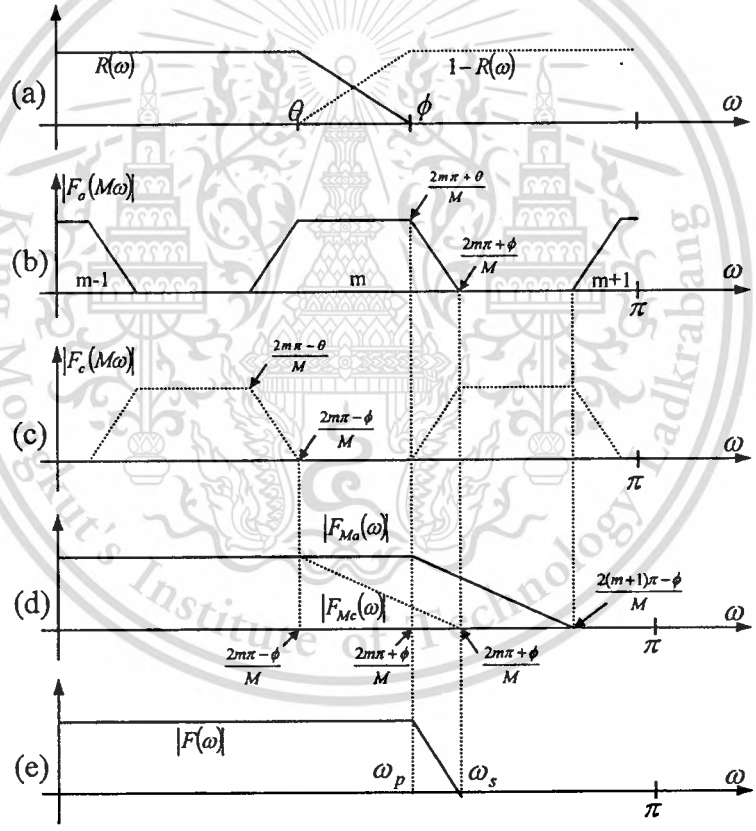


Figure 5.4 Power spectra of FRM-FIR filters. (a) Model filters, (b),(c) Interpolated model filters: $F_a\left(e^{jM\omega}\right)$ and $F_c\left(e^{jM\omega}\right)$, (d) Masking filters and (e) Desired filter

If a FIR filter is required with quite narrow transition band and low calculation cost, the FRM technique can provide its filter sufficiently.

Note that we have several isolated spectra in 0- π by using interpolation shown in Figure

5.4 (b) and (c). When a certain filter, e.g., Figure 5.4 (d) where $F_{Ma}(\omega)$ masks (b) and $F_{Mc}(\omega)$ mask (c). The designed filter is finally given in Figure 5.4 (e).

$$F(z) = F_a(z^M)F_{Ma}(z) + F_c(z^M)F_{Mc}(z) \quad (5.7)$$

$$F_a(e^{j\omega}) = e^{-j(N-1)\omega/2}R(\omega) \quad (5.8)$$

$$F_c(e^{j\omega}) = e^{-j(N-1)\omega/2}(1 - R(\omega)) \quad (5.9)$$

$$F_c(z) = z^{-(N-1)/2} - F_a(z) \quad (5.10)$$

where M is positive integer and indicates interpolation number. In other words, the sampling frequency in Figure 5.4 (b) increases M times higher than (a). The function $R(\omega)$ shows only amplitude and thus $F_a(e^{j\omega})$ consists of the amplitude and a liner phase component, i.e., $e^{-j(N-1)\omega/2}$, where N stands for the filter order of $F_a(e^{j\omega})$.

The filter, $F_{Ma}(z)$ and $F_{Mc}(z)$, are masking filter. The $F_a(z^M)$ and $F_c(z^M)$ are model filter

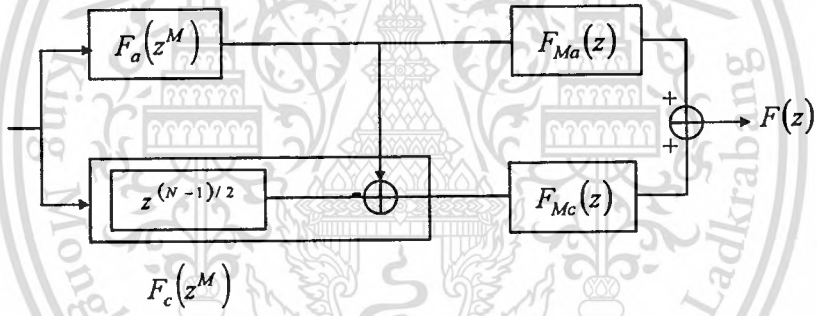


Figure 5.5 Structure of FRM-FIR filters

and $F_c(z^M)$ work as the complementary filter for $F_a(z^M)$. The transition of a finally designed filter is mainly determined by the transition of $F_a(z^M)$.

$F_c(z)$ can be expressed using $F_a(z)$, i.e., (5.9), $F(z)$ is described as the following equation and it is expressed as Figure 5.5.

$$F(z) = F_a(z^M)F_{Ma}(z) + \left(z^{-(N-1)/2} - F_a(z)\right)F_{Mc}(z) \quad (5.11)$$

When we design a low pass filter, e.g., Figure 5.4(e), the pass-band and stop-band transition of the desired filter are first determined. After that, m, M, θ and ϕ in Figure 5.4 are calculated.

The above design has been applied in FRM filter design [34], [35]. However, if an arbitrary

band-pass (BP) filter should be designed, a low-pass model filter in Figure 5.6 gives limited specification to BP design. We introduce a modified FRM technique using a band-pass model filter and realize lower calculation cost on the proposed band-pass FIR than the band-pass FIR which is designed by the conventional FRM technique.

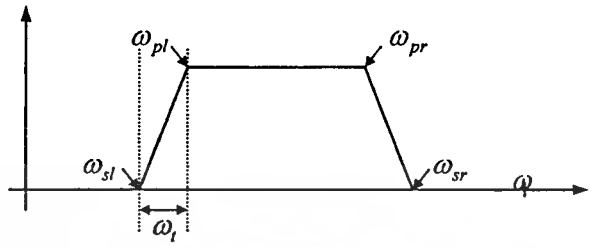


Figure 5.6 The desired BP filter

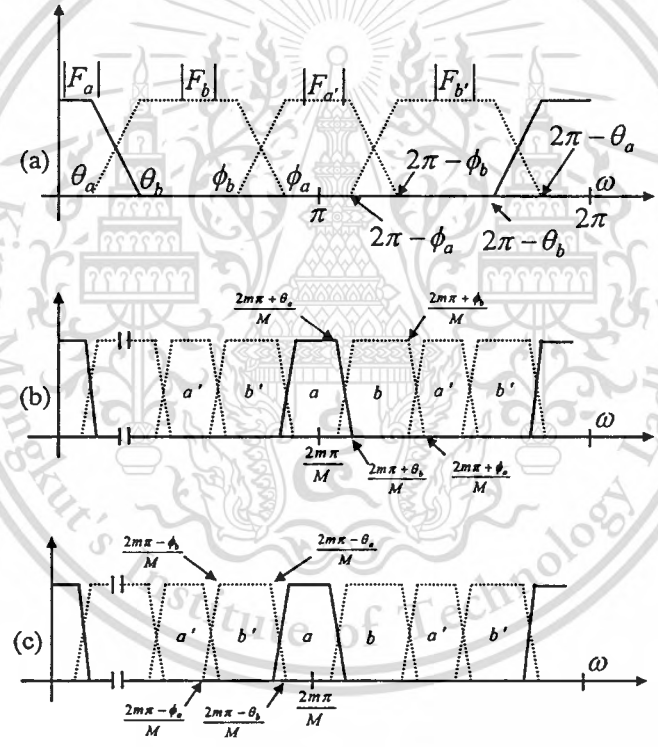


Figure 5.7 Power spectrum of the proposed model filters; (a) The model filters, (b) The interpolated filters in $[2m\pi/M,(2m\pi+\pi)/M]$ and (c) the interpolated filters in $[(2m\pi-\pi)/M,2m\pi/M]$.

5.4 RSF using modified FRM

To avoid design limitation on band-pass (BP) filter using conventional FRM explained in Subsection 5.1 and 5.3, we use a BP filter as a model filter. It has 4 edge parameters (two BP

edges and two stop-band(SP) edges). The desired BP filter can be designed by the model filter.

Let us assume that the required BP filter is given in Figure 5.6. Note that the specification of the right edge and the left edge can be different from each other. For this BP filter, we introduce some model filter in Figure 5.7. In Figure 5.7 $F_b(z)$ is defined as the BP filter whose right and left transition properties are M times wide than the filter of Figure 5.6. It is used as a model filter. In addition, $F_a(z)$ is also given as a complementary filter, i.e., $1-F_b(z)$, as a model filter. Accordingly, (θ_a,ϕ_a) and (θ_b,ϕ_b) be SP and BP edges of the model filter $F_b(z)$, respectively. They are defined as Figure 5.7(a). The characteristics of the right edge in $F_b(z)$ can be difference from that of the left edge. In Figure 5.7(a), in order to make its difference clear, the other notation, i.e., $F_{a'}(z)$ and $F_{b'}(z)$, are introduced. From these model filters, the up-sampling Z-transfer function in Figure 5.4 (b) and (c) are easily calculated. In these figures, the M-th interpolation is applied.

If the desired PB FIR is given as Figure 5.6, the following model filter should be designed:

$$\theta_a = 4\omega_{sl} \tag{5.12}$$

$$\phi_a = 4\omega_{sr} \tag{5.13}$$

$$\theta_b = 4\omega_{pl} \tag{5.14}$$

$$\phi_b = 4\omega_{pr} \tag{5.15}$$

Where M-th interpolation is applied, we can directly obtain the desired BP FIR. However, in order to eliminate other redundancy with whole frequency band, a masking filter is also designed. The masking filter can be simply designed as a low-pass FIR filter with wide transition edge and with low filter order. In this case, our modified FRM filter is designed in Figure 5.8 where $F_{Mb}(Z)$ is defined as a masking filter.

According to our required BP filter, the specification of the BP filter is given as follows:

5.4-1 On modulation spectrum domain, the lower edge of BP filter is given at about 1 Hz.

5.4-2 The upper edge of BP filter is set around 7 Hz.

5.4-3 The sharp edges are required. The lower edge normally requires narrower transition than the higher edge or both of them are required to be narrow transition.

Compared with the whole band, the required BP filter only passes narrow band in low frequency. From the above specification, we can design a model filter as Figure 5.9. The designed BP filter needs the 85 order of FIR filter since the left transition edge needs sharp. Using M=3,

we can get interpolated model filter in Figure 5.10. As mentioned above, its pass band is located in lower frequency. The masking filter is easily designed as sample low pass FIR filter with 19 order as shown in Figure 5.11. Finally, the FRMRSF has the power spectrum of Figure 5.12. Its structure is given as Figure 5.8. Note that we can get the quite sharp edge and obtain the attenuation of -40dB.

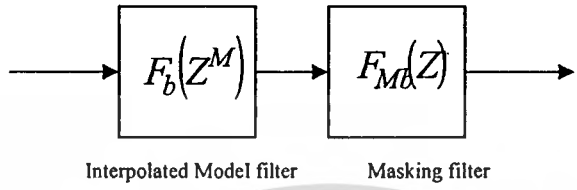


Figure 5.8 The process of modified FRM band-pass filter

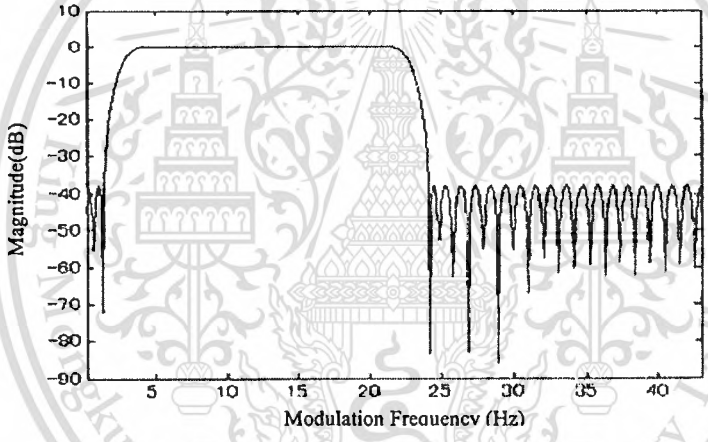


Figure 5.9 The model filter $F_a(z)$

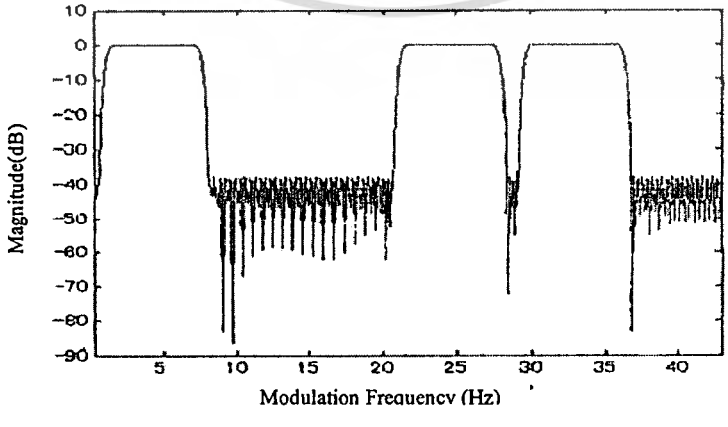


Figure 5.10 The model filter $F_a(z^M)$

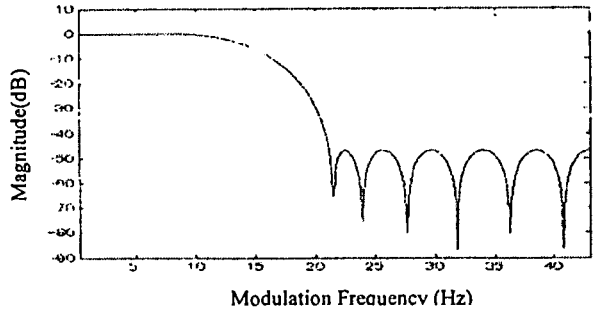


Figure 5.11 The model filter $F_{Ma}(z)$

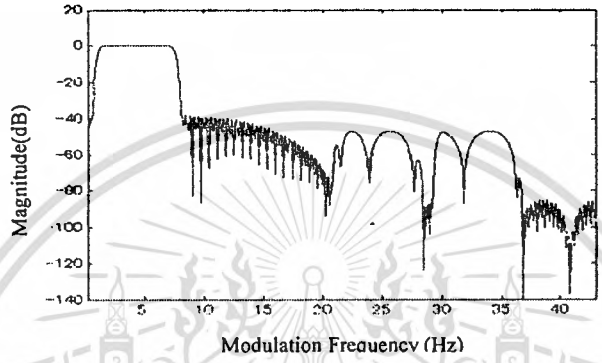


Figure 5.12 The FRM designed filter $F(z)$ for FIR RSF

5.5 Dynamic Range Adjustment (DRA)

Usually, when a white noise is added into speech, the dynamic range of feature parameters estimated from noisy speech is different from clean speech and tends to decrease. In addition, when RSF is applied two times in running power/log-power spectra, the dynamic range is reduced in almost all cases.

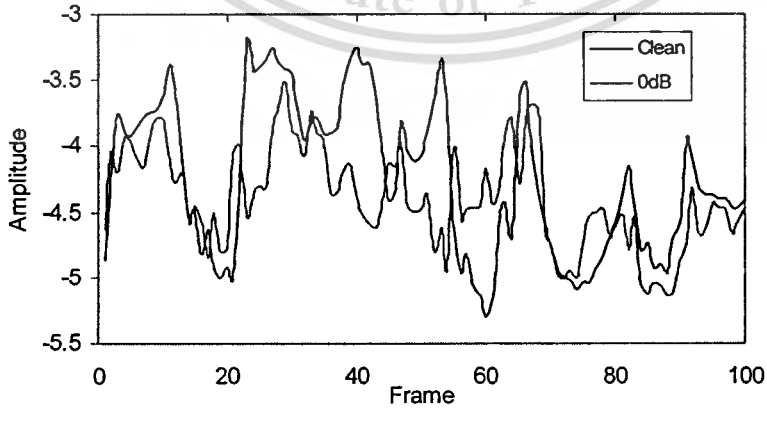


Figure 5.13 A comparison of trajectories of the 1st order DCT(log(CBI)) of clean speech and noise speech 0 dB SNR.

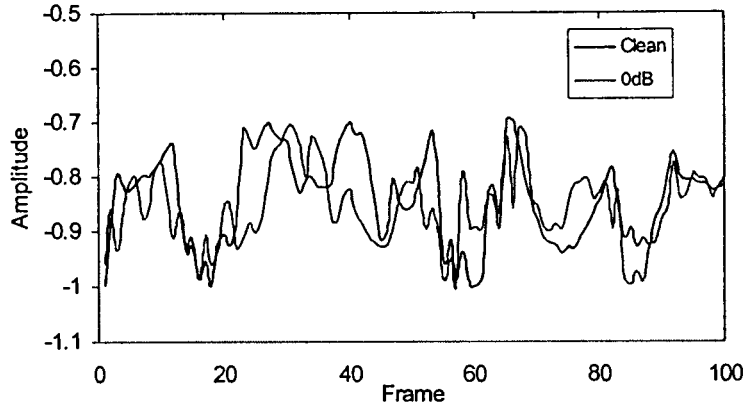


Figure 5.14 A comparison of trajectories of the 1st order DCT(log(CBI)) after DRA of clean speech and noise speech 0 dB SNR.

One of the other causes of noise corruption is derived from the differences in the dynamic ranges of coefficient for DCT(log(CBI)) or cepstrum of MFCC and PLP. The dynamic range of cepstrum indicates the difference between maximum and minimum of cepstral values in each order. Both the peaks maximum and minimum show the important characteristics of speech. However, as shown in Figure 5.13, the coefficient amplitude of peak are reduced comparing to the amplitude of noise free speech and characteristics are degraded. Considering that speech recognition is a kind of pattern matching, these differences can be compensated by normalizing both amplitudes of clean speech and noisy speech. Then, using DRA, the difference of cepstrum dynamic range is adjusted as shown in Figure 5.14 and the cepstrum from noisy speech is adjusted to the clean speech.

In the DRA, each coefficient of a speech feature vector is adjusted in proportion to its maximum by normalizing the amplitude of speech features, the following dynamic range adjustment (DRA) has been proposed [17], [19]-[20]:

$$\overline{p}_i(t) = p_i(t) / \max_{j=1,\ldots,m} |p_j(t)|, \quad i = 1,\ldots,m \tag{5.16}$$

where $p_i(t)$ denotes an element of the feature vector, m denotes the dimension, t denotes the frame number and $\overline{p}_i(t)$ is defined as normalized parameter of feature.

Chapter 6

Experiments for Speech Recognition

6.1 Introduction

A wide variety of techniques are used to perform speech recognition. Generally, there are two parts of speech recognition: speech analysis and speech understanding. Typically speech analysis starts with the digital sampling of speech as is acoustic signal processing. Most techniques include spectral analysis; e.g. LPC analysis (Linear Predictive Coding), MFCC (Mel Frequency Cepstral Coefficients), PLP (Perceptual Linear Predictive) and many more. As a widely used method. Speech spectrum have also been collected to produce a reasonable estimation of the spectrum from Bark scale and filter bank, as described in Section 3.4 and 3.5. In speech understanding, recognition of phonemes, groups of phonemes and words are required. It can be achieved by many processes: DTW (Dynamic Time Warping); HMM (hidden Markov modeling); NNs (Neural Networks); expert systems and combinations of techniques. HMM-based systems are currently the most commonly used and most successful approach. In this thesis, we propose a speech recognition system based on new features of speech spectrum on a Bark scale and HMM technique.

As one of the quite important factors for a speech recognition, the extraction of robust speech features are also considered. It has been known that speech features are corrupted by noise. Practical speech recognition system systems are used in various real environments with noises and noises are known to corrupt speech features and causes serving recognition error.

There are various widely used noise robust methods: noise robust LPC analysis, HMM decomposition and composition and the extraction of the dynamic cepstrum, etc.. However, speech spectrum without adaptation to noise may cause deterioration of speech features such as musical noise. This indicates that the estimation of an accurate noise status by spectral subtraction becomes difficult in some circumstances. In this research, we also explore the robustness of speech features and propose new speech recognition techniques.

6.2 Speech data collection

In this thesis, a vowel speech recognition system for Thai and Laotian language are implemented. All the sample data are recorded from Thai and Lao speakers. We have taken isolated utterances of 24 and 27 vowels respective in the context of initial consonants-vowel (CV),

while the tone information is superimposed on the vowel portion. The initial consonants (Thai 21, Laotian 26 consonants) have been described in Section 2.1-2.3. All the selected vowels are twice recorded per each speaker in order to study the system generalization for speaker-dependent and speaker-independent. The recording configurations are detailed as below:

6.2-1 The Speakers are both male and female, from age around 25-30 year old.

6.2-2 All recorded speeches are in Bangkok pronunciation for Thai and Vientiane pronunciation for Laotian.

6.2-3 Recording carried out in quiet office environment.

6.2-4 Sample speech data have been recorded with mono-channel, at 11.025 kHz sampling rate and 16 bits quantizing resolution.

To create acceptable vowel models, a number of training samples must be large enough. A total of 7,056 utterances for Thai vowel were recorded from 14 speakers (7 males and 7 females) and 9,828 utterances for Laotian vowel were recorded from 14 speakers (7 males and 7 females). The sample data of 4 males and 4 females are observed for training set. The sample data of other 3 males and 3 females are used for speaker-independent testing set. The recognition methods are used MFCC, PLP and Proposed DCT(log(CBI)) process.

6.3 Mel-frequency cepstral coefficients (MFCC) method

MFCC is the extraction spectral characteristics of the speech waveform into a sequence of acoustic vectors suitable for acoustic model processing. Figure 6.1 shows the stages of this transformation.

In the frame blocking process, input speech signal is blocked into frame of duration. Each frames overlap the next by 1/2 of the frame duration. The pre-emphasis process spectrally flattens the frame using a first order filter. After that, the pre-emphasized signals are Hamming windowed to minimize the effect of discontinuities at the edges of the frame duration. MFCC are feature representations defined by cepstrum of a short-time power spectrum. DFT is then applied follows by the me filter bank. These filters are computed by the average spectrum of multiple frequency bands. The structure of triangular filter is defined from the lowest to highest frequency of the Mel scale. The range of the values generated by the Mel filter bank is further emphasized by replacing each value with its natural logarithm. Taking the inverse discrete cosine transform of the log magnitude spectrum gives the cepstral coefficients of the speech signal. The component as a result of the periodic excitation source may be removed from the signal by simply discarding the higher

order coefficients. This representation is called the Mel-cepstrum. We compress the 40-element vector into a 12-elements cepstral vector by discrete cosine transform. The traditional solution is to augment cepstral vector with its first and second differentials. The Mel cepstral vector has 12-elements. After adding the differentials the acoustic vector has 24-elements with first differential and 36-elements with second differential.

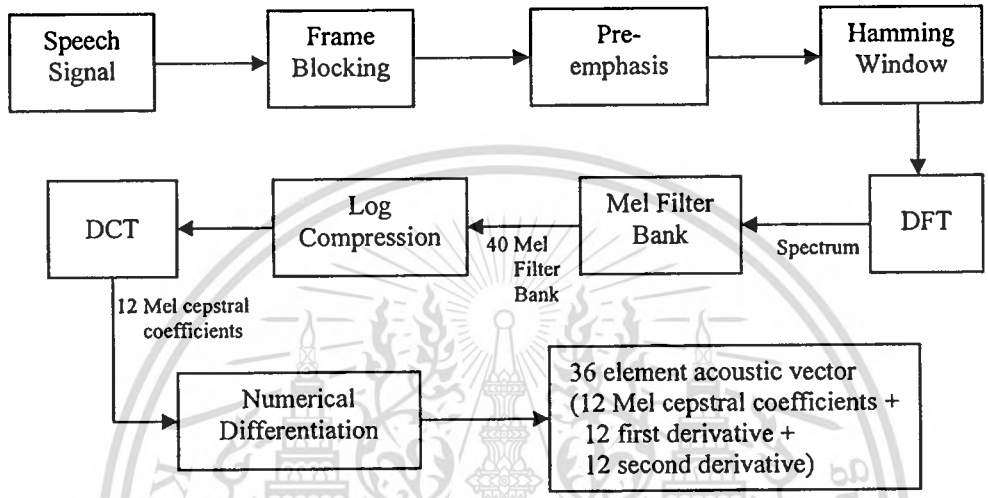


Figure 6.1 Signal processing for MFCC techniques

To evaluate the speech recognition, isolated word speech recognition using HMM has been curried out. The conventional recognition system consists of ordinary feature extraction based on MFCC and HMM. The recognition part is implemented using the HTK Toolkit [39]. The acoustic models has ten states, one with mixture per state HMMs state. The database is as described in Section 6.2. The speech feature vectors have 36-dimension parameters consisting of 12 cepstral coefficients, 12 first differential of cepstral coefficients and 12 second differential cepstral coefficients. Recognition results is shown in Table 6.1.

Table 6.1 Results of vowels recognition with MFCC techniques

Word recognition	Accuracy (%) from vary vectors dimensional parameters		
	MFCC(12)	MFCC+Delta(24)	MFCC+Delta+Delta(36)
Thai	84.82	90.75	92.05
Laotian	85.90	91.18	92.65

Table 6.1 shows the result of the experiments. Every value shows the correct recognition rate (%). "MFCC(12)" indicates 12 cepstral coefficients are used. The results shows an average recognition accuracy above 84%. In "MFCC + Delta (24)" 12 cepstral coefficients and 12 first derivative cepstral coefficients we used. The average recognition accuracy is above 90%, 6% improvement over MFCC(12). The "MFCC+Delta+Delta(36)" used 12 cepstral coefficients and 12 first derivative cepstral coefficients and 12 second derivative cepstral coefficients. The average recognition accuracy is above 92%, an improvement of 8% over MFCC(12).

6.4 Perceptual linear predictive (PLP) method

Perceptual linear predictive (PLP) analysis for speech [6] is a formulated method for deriving a more auditory-like spectrum based on linear predictive (LP) analysis of speech as described in Section 3.3. Conventional LP analysis approximates the high energy areas of the spectrum and smoothes out the finer harmonic structure. This estimation is done equally well at all frequencies which is inconsistent with human hearing. The auditory like spectrum in PLP is achieved by making some engineering approximations of the psychophysical attributes of the human hearing process. This technique uses three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: (1) the critical-band spectral resolution, (2) the equal-loudness curve and (3) the intensity-loudness power law. Figure 6.2 illustrates the process for deriving the auditory spectrum from which all pole modeling and cepstal analysis is performed.

PLP analysis obtains short term power spectrum by analyzing Fourier transforming frame of signal similar to MFCC. PLP analysis employs an auditory based warping of the frequency axis derived from the frequency sensitivity of human hearing. PLP is based on the Bark scale as described in Section 3.4. The amplitudes of the critical band filters are applied to quantize the frequency spectrum. The signal is Fourier transformed (spectral analysis) and mapped to a physiologically motivated frequency scale (critical-band analysis). Then the unequal sensitivity of human hearing across frequency is compensated for by pre-emphasis (equal-loudness pre-emphasis). Last the non-linear relation between intensity and perceived loudness is modeled by taking the cubic root of the intensity (intensity-loudness power law), this operation is approximation to the power law of hearing (Stevens,1975) . In the final operation of PLP analysis, critical band spectrum is approximated by the spectrum of all-pole model(auto-regressive modeling) using the auto-correlation method. The auto-regressive coefficients could be further transformed into cepstral coefficients of all pole model.

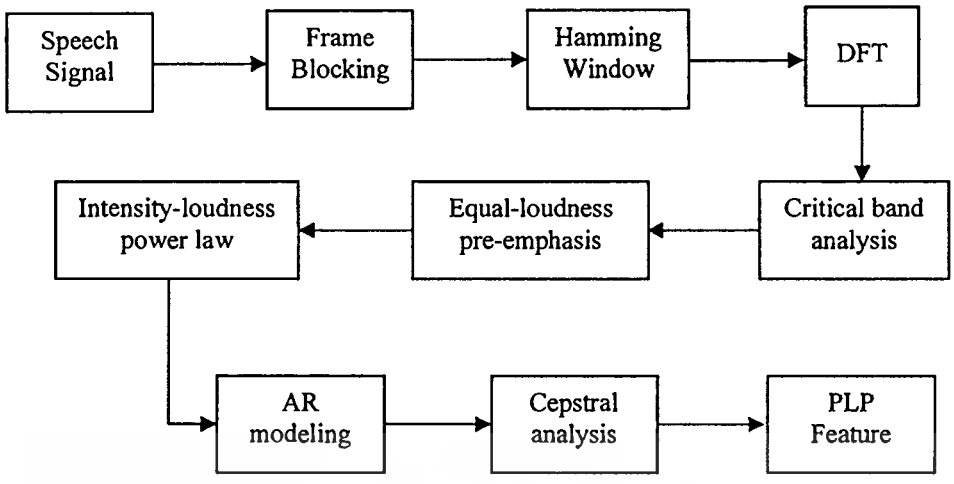


Figure 6.2 Signal processing for PLP techniques

To evaluate the speech recognition, isolated word speech recognition using HMM has been carried out. The conventional recognition system consists of ordinary feature extraction based on PLP and MFCC. The recognition part is implemented using the HTK Toolkit [39]. The acoustic models has ten state, one mixture per HMMs state. The whole database is as described in Section 6.2. The feature vectors have 36-dimension parameters; 12 cepstral coefficients, 12 first differential cepstral coefficients and 12 second differential cepstral coefficients. Recognition results is shown in Table 6.2.

Table 6.2 Results of vowels recognition with PLP techniques

Percentage accuracy of word recognition	Accuracy (%) from vary vectors dimensional parameters		
	PLP(12)	PLP+ Delta(24)	PLP+Delta+Delta(36)
Thai	85.12	90.45	92.23
Laotian	85.90	90.62	92.67

Table 6.2 shows the result of the experiments for PLP techniques. "PLP(12)" refers to sample test in which 12 PLP cepstral coefficients. The average recognition accuracy is above 85%. "PLP+ Delta(24)" 12 cepstral coefficients and 12 first derivative cepstral coefficients of PLP. The average recognition accuracy is above 90%, an 5% improvement over PLP(12) cepstrals. "PLP+Delta+Delta(36)" uses 12 cepstral coefficients and 12 first derivative cepstral coefficients and 12 second derivative cepstral coefficients. The average recognition accuracy is above 92% a 7% improvement over PLP(12).

6.5 Proposed DCT(log(CBI)) process

We propose a speech recognition system based on new features of speech spectrum on a Bark scale. Speech features extraction processes consists of four steps : (1) auto-regressive model (AR model), (2) critical band intensity (CBI), (3) discrete cosines transform on the logarithm of CBI (DCT(log(CBI)). The processing stages has been described as Section 4.1. Extraction features techniques based on Bark scale follows Figure 4.1.

The speech waveform is blocked into first segmented into frame of 300 samples where its time length is 27.21 ms with 11.025 kHz sampling rate and Each frames overlap the next by 1/3 of the frame duration. The pre-emphasis is applied to smooth spectrum of input speech signal. Hamming window is applied to each individual frame to minimize the signal discontinuities at the beginning and end of each frame.

An Auto-Regressive (AR) Model can be described by the transfer function defined in (3.12). This model is commonly used in linear predictive coding (LPC) through the application minimum mean square estimation (MMSE) to provide auto-correlation coefficients. The transfer function is computed from to LPC speech spectrum. The spectrum envelop represents an approximation of a linear speech production model. The critical band intensities, we then mapped from the LPC spectrum using the Bark scale. In this thesis, the underlying sampling rate is set to be 11,025 kHz with a bandwidth of 5.5 kHz. Accordingly, there are 18 critical bands. The dynamic range of CBI is indicated by the difference in maximum and minimum values. For linear acoustic model, the dynamic range is expanded using logarithm CBI. We can increase the dynamic rang by enhancing the lower values of CBI. In the final operation of extraction processes, DCT is applied to the log(CBI). The outcome DCT coefficients are the final extracted features to be used in the HMM for speech recognition.

We also experimentally compared Thai and Laotian vowel recognition test results between CBIs, log(CBI) and DCT(log(CBI)). The whole database is as described in Section 6.2. Speech feature vectors have 18-dimensionnal parameters. The acoustic models are ten state, one mixture per state HMMs state. The recognition accuracy are shown in Table 6.3.

In the Table 6.3, it is found that the DCT(log(CBI)) coefficient provides accuracy that are significantly higher (3.5%) than that of CBIs alone and (2.6%) than that of log(CBI).

Different experiments were, carried out to find the optimal number of DCT(Log(CBI)) coefficients. The accuracy steadily increase when the number of DCT(log(CBI)) parameters vary from 10 to 18 coefficients. The results are as shown in Figure 6.3.

Table 6.3 Percentage accuracy for vowels recognition with CBI

Features	Dimension	Recognition following features	
		Thai (%)	Laotian(%)
CBI	18	82.34	82.48
Log(CBI)	18	83.17	83.32
DCT(log(CBI))	18	85.81	86.49

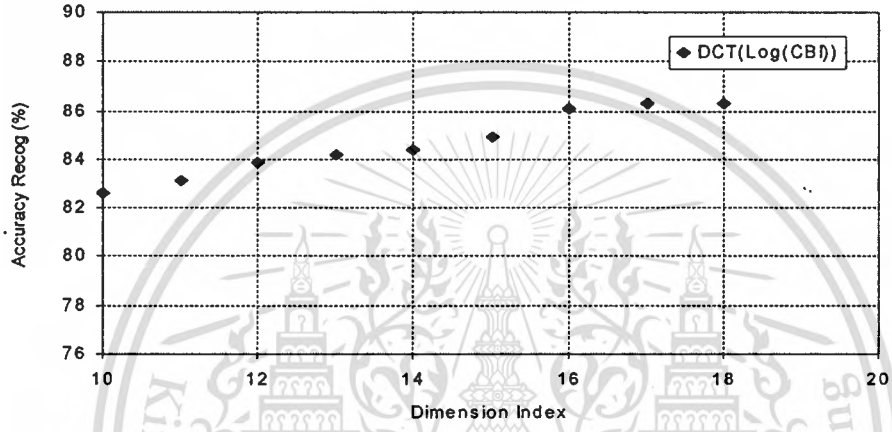


Figure 6.3 Accuracy with dimension DCT(log(CBI)

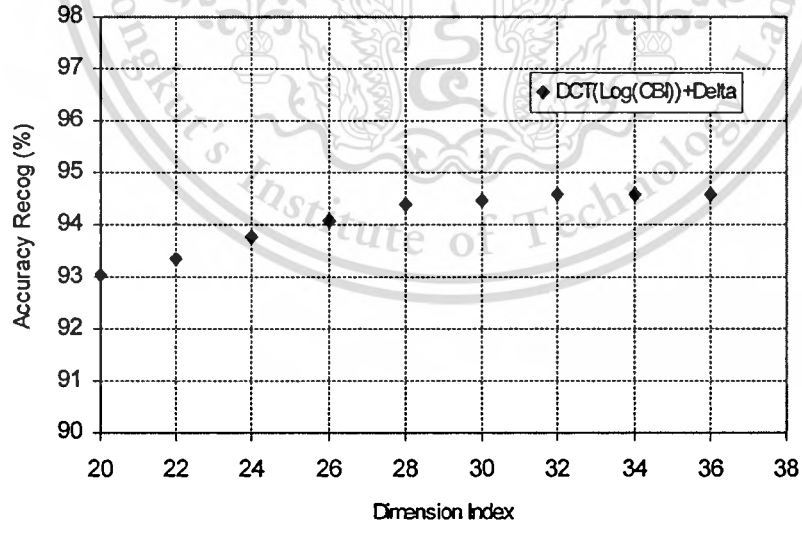


Figure 6.4 Accuracy with dimension DCT(log(CBI) with differential

The average accuracy is indicated in Figure 6.3. It increase up to maximum at 86% and remain constant from 16 coefficients onward. The optimum number of coefficients of

DCT(Log(CBI)) is 16 coefficients. The traditional solution is to augment the feature vectors with its first and second differentials.

Further we append DCT(Log(CBI)) with their differentials and second differentials and test for the optimal number of coefficients. With first differentials, number parameters of acoustic vector are varied from 20 to 36 coefficients. With second differentials, the number parameters are varied from 30 to 54 coefficients. These features are used with first differential, the average accuracy results are as shown in Figure 6.4. It shows improve with performance of 94% accuracy and beyond. The optimum number of coefficients of DCT(Log(CBI)) is more 32. The DCT(log(CBI)) coefficient and first differential (32-parameters) give significantly higher accuracy (8%) than for 16 DCT(log(CBI)) coefficients only.

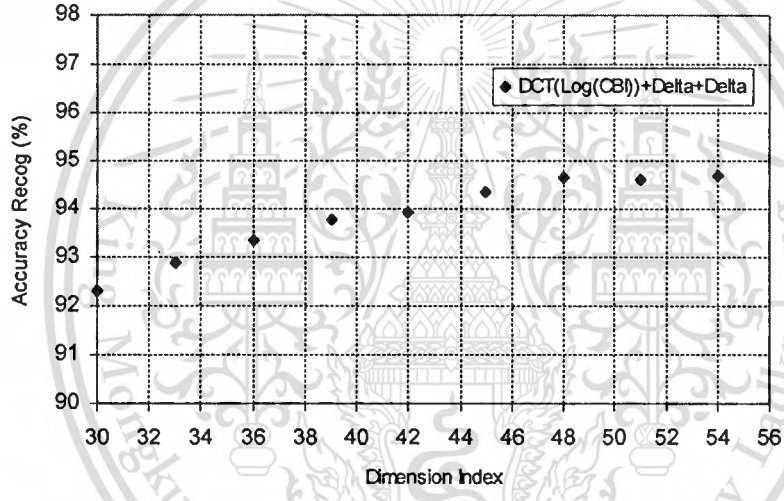


Figure 6.5 Accuracy with dimension DCT(log(CBI) with two differentials

Figure 6.5 shows the average accuracy using DCT(log(CBI)) with first and second differentials with the number of parameters varied from 30 to 54. The average accuracy increases up to maximum above 94% and remains constant at 46 parameters. The optimum number of coefficients of DCT(log(CBI)) is there fore 46.

There is no significant improvement for second differential over the accuracy of first differentials alone. Therefore we use DCT(log(CBI)) together with the first differential. The performance results are summarized in Table 6.4.

In Table 6.4 shows the result of the experiments using DCT(log(CBI)). Every value shows the correct recognition rate (%). "16 DCT(log(CBI))" refer to test in with only 16 DCT(log(CBI))

coefficients are used. The average recognition is accuracy above 85%. "16 DCT(log(CBI)) + 16 Delta" refers to test in with 16 coefficients and their 16 first derivative of DCT(log(CBI)) are you. The average recognition accuracy is above 94% an 8% improvement over that of 16 DCT(log(CBI)) coefficients only. Finally, "16 DCT(log(CBI))+16 Delta+16Delta" refer to test in which 16 coefficients, 16 first differential coefficients, 16 second differential coefficients of DCT(log(CBI)) are used. The average recognition accuracy is only slightly higher then DCT(log(CBI)) coefficients with their first differential.

Table 6.4 Results of vowel recognition with (DCT(Log(CBI))) techniques

Vowel recognition	Accuracy (%) from vary vectors dimensional parameters		
	16 DCT(log(CBI))	16 DCT(log(CBI)) + 16 Delta	16 DCT(log(CBI)) +16 Delta + 16 Delta
Thai	85.81	94.12	94.35
Laotian	86.49	94.44	94.71

Since MFCC, PLP and DCT(log(CBI)) have no explicit used tone information, the accuracy of their recognition systems can be improved for the application of Thai and Laotian speech recognition. The tone characteristics are based on fundamental frequency and thus a part of personality, gender and age considerably affects estimated tones. Therefore, the modeling methods of tones can improve its recognition accuracy. In Thai and Laos spoken languages, the combination of both 5 difference tones and 2 different time duration are used in human speech recognition. While there are proposed methods for tone recognition for other languages. They are met suitable for Thai and Lao. In particular, the difference between short time and long time duration vowels should be accurately evaluated for the vowels of Thai and Lao.

Figure 6.6 shown the total system proposed in this thesis. The feature extraction is shown in the first part for this system. Each frame of speech is represented by the parameters vector from DCT(log(CBI)) in 6.A, the regression of voice energy in 6.B and the quantized pitch of speech in 6.C. These features are applied to HMM for training and recognition.

As, we have carried out the experiments comparing DCT(log(CBI)), MFCC and PLP, we have found that the accuracy of the recognition systems are not enough for the application to Thais and Lao vowels recognition. Therefore, we are combined these speech features with regression on the voice energy in 6.B and a quantized pitch in 6.C.

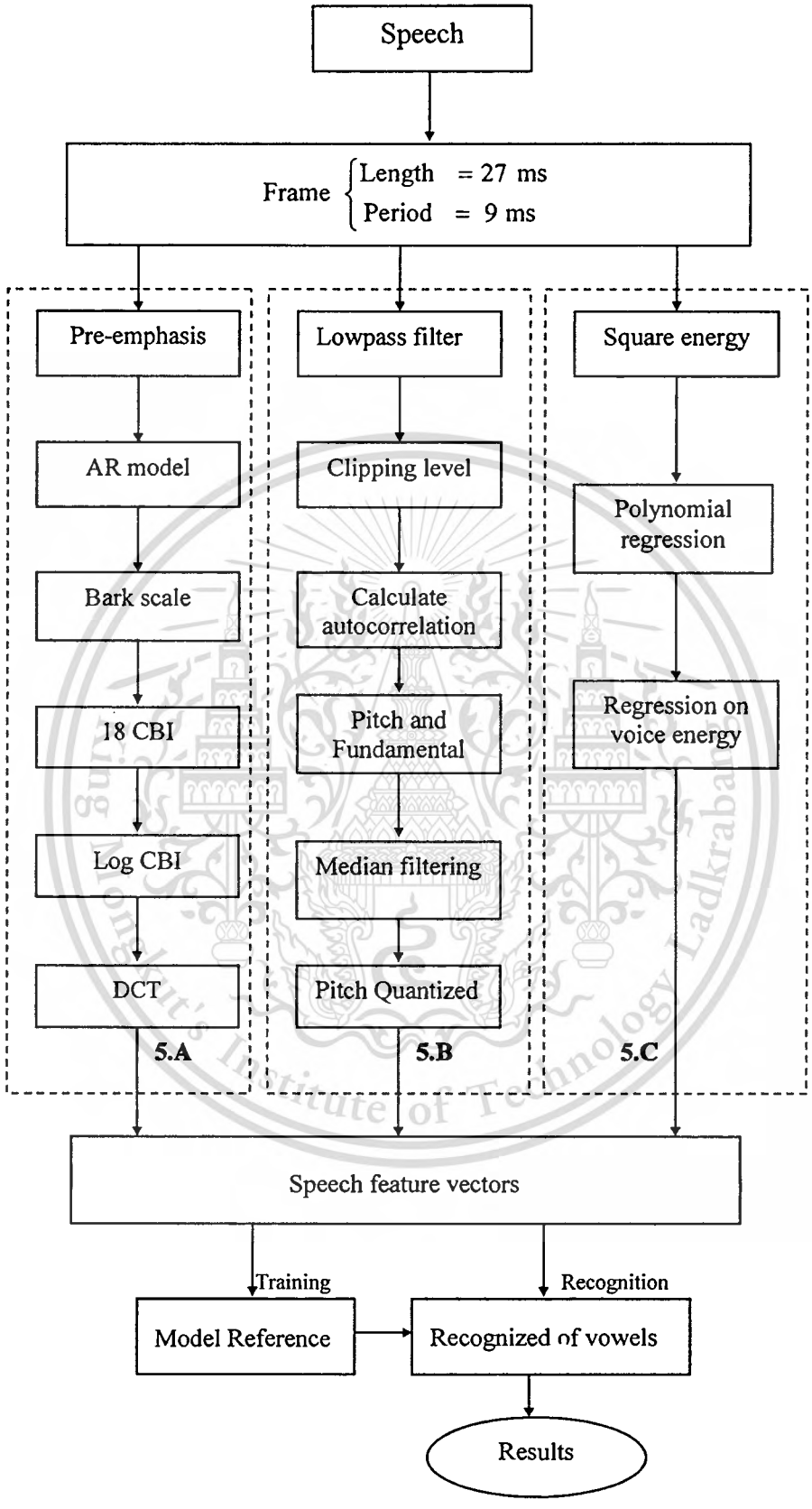


Figure 6.6 Block diagram of a proposed speech recognition system

6.B In quantized pitch (QP) Feature, the estimated pitch is quantized with three levels. A lot of detail features of pitch are eliminated and thus the estimated feature becomes quite simple. However, only the dynamism of pitch on time axis can be expressed. It becomes independent of speakers and gender.

6.C In the Regression of Voice Energy Feature (RE), time variation of speech energy is approximated with a simple polynomial. This is a useful feature in Thai/Laotian language where its time variation identifies different vowel and hence meaning. Using this feature into HMM, short and long duration vowel can be distinguished accurately.

In Figure 6.6, the utterance is represented by feature vectors of DCT(log(CBI)), Energy (E), Regression Energy (RE) and a quantized pitch (QP) sequence. We are carried out in 6 different experiments with feature vectors as listed. These feature parameters are tested with Thai and Laotian vowel. The results of the average accuracy recognition are shown in Table 6.5.

6.5-1. 18- parameters of CBI

6.5-2. 18- parameters of DCT(log(CBI))

6.5-3. DCT(log(CBI)) plus energy (19- parameters of DCT(log(CBI)) +E)

6.5-4. DCT(log(CBI)) plus regression voice energy (19- parameters of DCT(log(CBI)) +RE).

6.5-5. DCT(log(CBI)) plus energy and quantized pitch (QP). (20-parameters of DCT(log(CBI)) +E+QP)

6.5-6. DCT(log(CBI)) plus regression voice energy and quantized pitch (20-parameters of DCT(log(CBI)) + RE+QP)

Table 6.5 indicate the percentage of vowels which are correctly recognized Thai and Laotian. The result shows that an accuracy is above 82% using 18-dimension CBI alone. DCT(log(CBI)) improved the accuracy is above 85%. DCT(log(CBI)) with energy gets better than 86%. DCT(log(CBI)) with regression energy gets better than 88%. DCT(log(CBI)) with regression energy and quantized pitch gets better than 89%. DCT(log(CBI)) with differential and regression energy and quantized pitch gets better than 96%. Sixteen number of parameters for DCT(log(CBI)) and 16 numbers of their delta are enough to increase the accuracy beyond 96%.

When, feature is combined DCT(log(CBI)), differential of DCT(log(CBI)) with E/RE and/or QP. We are carried out in 4 different experiments with feature vectors as listed. These feature parameters are tested with Thai and Laotian vowel.

6.5-7. DCT(log(CBI)) and first differential of DCT(log(CBI)) plus energy (33-parameters of 16 DCT(log(CBI)) + 16 Delta + E).

6.5-8. DCT(log(CBI)) and first differential of DCT(log(CBI)) plus regression voice energy (33-parameters of 16 DCT(log(CBI)) + 16 Delta + RE).

6.5-9. DCT(log(CBI)) and first differential of DCT(log(CBI)) plus energy and quantized pitch (34-parameters of 16 DCT(log(CBI)) + 16 Delta + E+QP).

6.5-10. DCT(log(CBI)) and first differential of DCT(log(CBI)) plus regression voice energy and quantized pitch (34-parameters of 16 DCT(log(CBI)) + 16 Delta + RE+QP).

These feature parameters are tested with Thai and Laotian vowel. The results of the average accuracy recognition are shown in Table 6.6.

Table 6.5 Results of vowels recognition with combination features

Test	Feature parameter	Parameters	Recognition following features	
			Thai (%)	Laotian(%)
6.5-1	CBI	18	82.34	82.48
6.5-2	DCT(log(CBI))	18	85.81	86.49
6.5-3	DCT(log(CBI)) +E	19	86.33	86.75
6.5-4	DCT(log(CBI)) +RE	19	88.47	89.15
6.5-5	DCT(log(CBI))+E+QP	20	87.46	87.54
6.5-6	DCT(log(CBI)) +RE+QP	20	89.83	90.97

Table 6.6 Feature is combined DCT(log(CBI)), delta of DCT(log(CBI)) with E/RE and QP

Test	Feature parameter	Parameters	Recognition following features	
			Thai (%)	Laotian(%)
6.5-7	DCT(log(CBI))+ delta+E	33	94.12	94.45
6.5-8	DCT(log(CBI))+ delta+RE	33	95.47	95.52
6.5-9	DCT(log(CBI))+delta+E+QP	34	94.47	94.67
6.5-10	DCT(log(CBI))+ delta+RE+QP	34	96.12	96.34

In Table 6.6, we show the experiment results using 16 parameters of DCT(log(CBI)) and their. The results were done with Energy feature, Regression Energy and/or Quantized Pitch features. The results shows an 2% increase in accuracy using RE then E; and 1% increase with QP.

Since, MFCC and PLP are the will known speech features, as described in Section 6.3-6.4. We have carried out experiments by combining MFCC and PLP together with voice energy (E), regression on the voice energy (RE) and a quantized pitch (QP). We have carried out 8 different experiments with three feature vectors combination as listed. The results of the average recognition accuracy are shown in Table 6.7.

6.5-11 The 37 parameters MFCC consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients and 1 parameter of E.

6.5-12 The 37 parameters PLP consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients and 1 parameter of E.

6.5-13 The 37 parameters MFCC consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients and 1 parameter of RE.

6.5-14 The 37 parameters PLP consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients and 1 parameter of RE.

6.5-15 The 38 parameters MFCC consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients, 1 parameter of E and 1 parameter of QP.

6.5-16 The 38 parameters PLP consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients, 1 parameter of E and 1 parameter of QP.

6.5-17 The 38 parameters MFCC consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients, 1 parameter of RE and 1 parameter of QP.

6.5-18 The 38 parameters PLP consists of 12 cepstrum coefficients, 12 delta, 12 delta-delta coefficients, 1 parameter of RE and 1 parameter of QP.

Table 6.7 shows the result of the experiments using MFCC and PLP. The value are the correct recognition rate (%). MFCC and PLP feature refer in the test have 36 cepstrum coefficients. "36 MFCC+E" and "36 PLP+E" refers to test in with 36 cepstrum coefficients with additional voice energy. The average recognition accuracy is above 92%, regression on the voice energy improves by 1% over that of MFCC, PLP with voice energy. The result obtained using the combination of the MFCC and PLP with regression on the voice energy and quantized pitch improves above 4% over that of 36 MFCC, 36 PLP cepstrum coefficients (look at Table 6.1 and 6.2).

Table 6.7 Feature is combined MFCC, PLP with E/RE and QP

Test	Feature parameter	Parameters	Recognition following features	
			Thai (%)	Laotian(%)
6.5-11	36 MFCC +E	37	92.42	92.68
6.5-12	36 PLP + E	37	92.57	92.89
6.5-13	36 MFCC + RE	37	94.25	94.36
6.5-14	36 PLP + RE	37	94.56	94.15
6.5-15	36 MFCC +E + QP	38	94.07	94.44
6.5-16	36 PLP + E + QP	38	94.83	94.97
6.5-17	36 MFCC +RE + QP	38	95.35	95.48
6.5-18	36 PLP + RE + QP	38	95.86	95.75

6.6 Robust speech recognition the processes of noise

6.6.1 Nonlinear Running Spectrum Filter(NRSF) [33]

When we consider high SNR, the noise components of equation (5.2) are small and can be neglected. In this case, it is obvious that the first RSF should not be applied to an observed speech have been described in Chapter 5. In case of high SNR, speech features existing in high modulation frequency band are useful for speech recognition. When the first RSF is employed under high SNR, they are disturbed. In order to avoid the performance drop of RSF because of high SNR, we have introduced NRSF as follows:

The nonlinear method is described in Figure 6.7 [33]. This diagram only shows the robust speech analysis part by which noise-reduced speech characteristics, i.e., robust mel-frequency cepctrum component (robust MFCC), are calculated. The MFCC has been used as speech features in speech recognition and thus our features are also based and extended on MFCC. The basic idea of this processing is simple. According to estimated SNR, the system decides whether the first RSF should be used or not.

Just after the calculation of Mel-spectra by using Mel-filter banks, noise powers at every Mel-frequency are estimated in Mel-spectrum domain from several speech-less flames, i.e., the first several frames of the observed whole frames. The estimated noise power at the specific Mel-frequency is different from others. During its observed speech period, the noise power at the specific

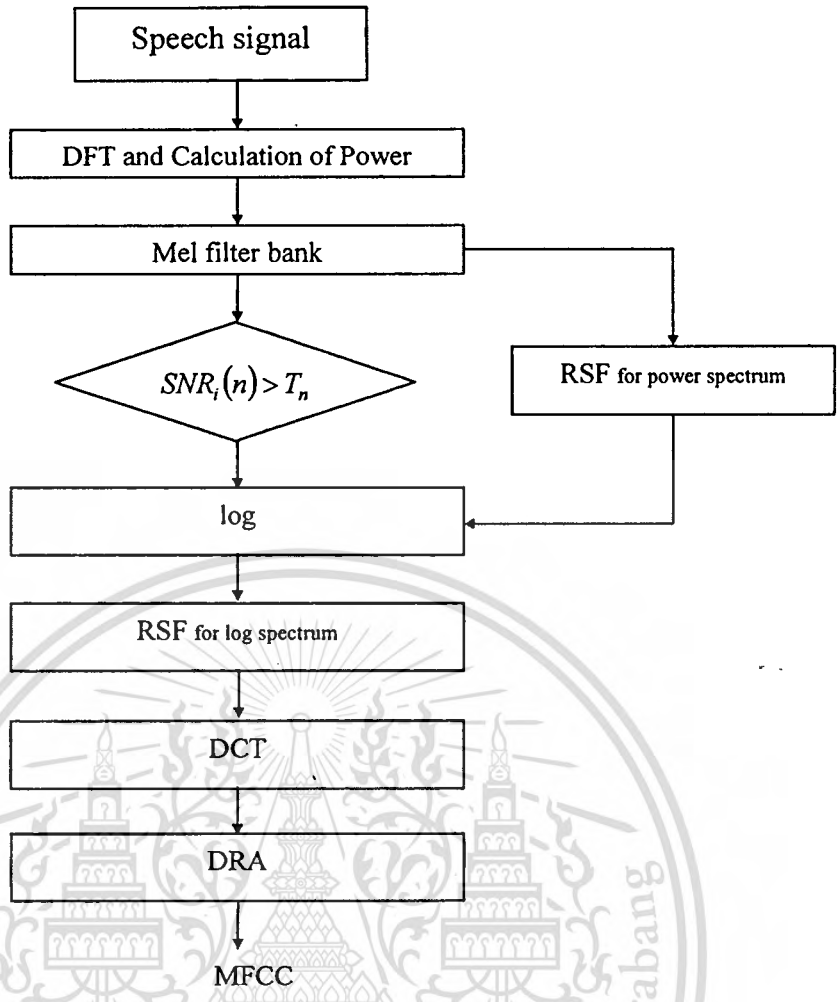


Figure 6.7 Robust speech recognition by RSF, RFMRSF with DRA

Mel-frequency is assumed to be steady on time axis. Using this processing, we can get estimated noise powers at all Mel-frequencies.

In speech frame, we calculate the power spectrum, i.e., $X_i(n)$, which includes speech components and noises at the $n-th$ frame. The estimated RSF used in the nonlinear RSF is given

$$SNR_i(n) = \begin{cases} 20 \cdot \log_{10} \dfrac{X_i(n) - \hat{N}_i}{\hat{N}_i} & \text{if } X_i(n) > 1.1 \cdot \hat{N}_i \\ -20 & \textit{otherwise} \end{cases}$$

where $\hat{N}_i$ is defined an estimated noise power at $i-th$ Mel-spectrum component. It is unnecessary to check its SNR less than -20dB. The threshold of $1.1 \cdot \hat{N}_i$ is used for the calculation of only SNR over -20 dB.

Note that $SNR_i(n)$ is usually changed at every frame. If $SNR_i(n)$ is less than a threshold, the first RSF filter is applied. Otherwise, it is not applied. In other words, the use of the first RSF

filter is decided at every frame and at every Mel-spectrum frequency. The threshold is determined as 5 dB in our experiments. In many experiments, it turns out that a threshold from 0 -10 dB show good results and any threshold among this limit can be applied with good accuracy. Note that $X_i(n)$ is not an average speech power. Accordingly $SNR_i(n)$ dose not indicates actual SNR.

By using this method, we can apply the same processing in both training stage and recognition stage. In addition, the system can realize high performance of speech recognition despite various noise conditions where SNR is considered over 0 dB.

The experiment of isolated Japanese and Thai word recognition was tried. The speech data was given from Japanese pre-feature names in Japanese common voice data (names of places) delivered by the Japan Electric Industry Development Association. and name of public institution in Thailand with 16 bit AD and 11.025 kHz sampling frequency. The number of words Japanese and Thai was limited into 100 and 72 word, respectively. The HMM method was used for speech recognition. in the stage of training, speech data ware given from 40 male speakers for Japanese. And 16 speakers for Thai. In the stage of recognition, there were 10 unspecific speakers for Japanese and 8 speakers in this experiment. We explore the noise robust property of the total system and thus several noise circumstances were considered, i.e., from 0 dB SNR to 20 dB SNR. The 15 additive noises was selected among NOISEX-92. The analysis conditions and the selected characteristics were described in Table 6.8. The HMM with 32 states was used in our system.

Table 6.8 Analysis conditions for NRSF

Window length	23.2ms (256 point)
Frame shift	11.6ms (128 point)
Windowing	Hamming window
Pre-emphasis	$1-0.97z^1$
Feature vectors	MFCC 12, 12 Δ, 12 $\Delta\Delta$, 1 energy, $\Delta\Delta$ 1 energy

Table 6.9 and 6.10 show the result of the experiments. Every value shows the correct recognition rate (%). "BASELINE" means a sample speech recognition method in which there is no noise robust algorithm. "Conventional RSF" means that only the second filtering of RSF is used in training stage and both of two filter are applies in speech recognition stage. "NONLINEAR RSF" is given from the previous sub-section. "Nonlinear RSF (Ideal)" shows NRSF with known

noise variances as a prior information. In other words. We assumed that noise variances were given as a prior and thus ideal SNR could be calculated. However, it is impossible to use it into actual circumstances. Compared with "NONLINEAR RSF", the NRSF present similar performance at speech recognition ability in any circumstances.

Table 6.9 Comparison between recognition performances with conventional method and with proposed method for Japanese

Type of noises / **SNR[dB]**	**BASELINE**			**CONVENTIONAL RSF**			**NONLINEAR RSF**			**NONLINEAR RSF (deal)**		
	0dB	**10dB**	**20dB**	**0dB**	**10dB**	**20d**	**0dB**	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**
White noise	0.8	33.2	92.4	47.9	85.6	97.5	50.0	87.1	97.7	50.2	87.1	97.7
Pink noise	0.7	53.5	96.9	53.3	91.5	97.6	52.3	92.6	98.1	52.7	92.8	98.4
HF Channel noise	3.3	26.6	82.1	49.3	85.1	95.5	48.6	86.1	96.9	48.6	86.2	96.3
Speech babble	7.8	58.2	91.2	20.2	77.0	95.7	18.9	78.9	95.5	22.2	82.5	96.6
Factory floor noise 1	2.0	62.9	97.4	33.4	88.5	98.4	33.3	89.9	98.9	35.8	91.2	98.7
Factory floor noise 2	15.0	80.0	98.1	75.2	96.3	98.1	75.5	97.6	98.9	76.3	97.2	98.9
Jet cockpit noise 1	0.9	49.4	95.5	45.5	88.5	97.9	44.7	88.7	97.6	46.0	88.8	97.9
Jet cockpit noise 2	0.2	41.4	95.5	51.7	90.0	97.7	51.5	92.3	98.4	52.6	92.3	98.3
Destroyer engine room	3.8	55.6	94.7	61.0	90.7	97.0	60.1	93.6	98.0	61.7	93.1	98.0
Destroyer operation room	4.1	76.1	98.5	58.9	94.7	98.5	54.2	94.9	98.5	57.8	94.9	98.9
F-16 cockpit noise	1.7	69.4	97.3	56.2	92.1	97.8	57.2	93.7	98.5	57.3	93.6	98.4
Military vehicle noise	52.9	87.5	98.4	88.4	94.4	97.5	90.6	96.0	98.5	90.5	96.0	99.0
Tank noise	29.7	91.2	98.5	80.7	97.7	98.7	82.6	98.0	99.2	83.7	98.2	99.4
Machine gun noise	80.1	92.6	97.5	84.6	92.5	96.8	84.7	92.8	97.5	87.5	94.1	97.8
Car interior noise	93.9	99.0	99.2	97.0	98.5	98.9	97.8	99.0	99.7	98.4	99.1	99.6
Average	**19.8**	**65.1**	**95.5**	**60.2**	**90.9**	**97.6**	**60.1**	**92.1**	**98.1**	**61.4**	**92.5**	**98.3**
Clean	99.5			98.7			99.3			99.3		

Table 6.11 shows some comparisons, i.e., cepstrum mean subtraction method (CMS method) [16]-[20] as described in Chapter 5 and, CMS/DRA and proposed method. The conditions of this experiment are the same as that Table 6.9 and 6.10. Every value shows the recognition

correctness. When we consider the average recognition accuracy at each SNR, our proposed method shows higher scores than others. In addition, when we consider the circumstance of clean condition where this condition is hardly realized under ordinary circumstances, our method shows similar performance to others.

When we compare the results of nonlinear RSF in Table 6.9 with the results of nonlinear RSF using modified FRM in Table 6.10, the rates are the same. It turns out that FRM-NRSF (frequency response masking techniques and nonlinear running spectrum filtering) has the same performance as NRSF.

Table 6.10 Comparison between recognition performances with conventional method and with proposed method for Thai word

Type of noises	**BASELINE**			**CONVENTIONAL RSF**			**NONLINEAR RSF**			**NONLINEAR RSF (deal)**		
SNR[dB]	**0dB**	**10dB**	**20dB**	**0dB**	**10dB**	**20d**	**0dB**	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**
White noise	1.85	43.9	96.7	40.1	87.3	96.3	38.5	84.4	94.3	38.8	86.7	96.7
Pink noise	2.32	54.5	96.2	43.6	58.8	87.7	44.1	59.5	89.7	45.7	62.3	91.5
HF Channel noise	2.77	31.7	87.4	28.4	76.7	97.5	27.1	78.8	92.5	29.8	79.4	93.1
Speech babble	7.87	72.2	98.4	58.6	88.3	92.8	58.2	89.6	94.8	58.6	89.8	94.7
Factory floor noise 1	3.24	79.1	99.5	39.8	82.5	98.6	41.2	87.2	96.6	39.8	79.6	94.3
Factory floor noise 2	2.31	30.0	88.6	37.8	84.2	96.0	38.5	81.5	96.3	37.8	83.7	97.8
Jet cockpit noise 1	1.85	25.0	90.1	41.2	61.1	90.2	41.6	63.4	97.2	40.4	59.8	94.7
Jet cockpit noise 2	2.77	38.1	91.4	43.7	70.5	92.4	43.5	72.5	94.4	41.5	71.6	94.7
Destroyer engine room	3.24	66.7	93.0	47.2	84.7	96.6	46.6	83.0	94.6	46.9	86.3	96.2
Destroyer operation room	16.6	79.8	97.7	65.3	88.1	98.2	65.6	88.2	98.6	67.7	90.4	98.3
F-16 cockpit noise	2.31	37.9	86.6	30.7	80.4	95.0	29.4	82.5	96.8	29.1	78.5	93.2
Military vehicle noise	42.3	88.4	97.0	85.6	91.4	98.7	85.1	90.4	96.2	84.8	90.5	96.7
Tank noise	36.1	83.9	98.7	46.9	88.3	98.2	45.8	87.7	96.9	45.9	89.3	98.6
Machine gun noise	62.0	94.5	100	81.5	92.6	98.6	80.1	91.4	97.4	81.0	93.6	98.5
Car interior noise	68.0	95.7	99.4	88.2	98.8	99.4	88.4	97.5	99.1	87.8	99.4	99.4
Average	**17.0**	**61.4**	**94.7**	**51.9**	**82.1**	**95.7**	**51.5**	**82.5**	**95.7**	**51.7**	**92.7**	**95.8**
Clean	97.6			97.1			96.8			96.9		

Table 6.11 Comparison between recognition performances with CMS, CMS/DRA and with proposed method for Japanese

Type of noises / **SNR[dB]**	**CMS**			**CMS/DRA**			**NONLINEAR RSF using MODIFIED FRM**		
	0dB	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**	**0dB**	**10dB**	**20dB**
White noise	0.6	55.4	95.5	28.2	82.4	93.1	49.3	87.6	97.5
Pink noise	0.5	58.8	97.1	26.4	85.5	98.2	52.4	92.6	98.3
HF Channel noise	2.0	51.3	95.0	26.8	81.4	96.6	48.8	86.3	96.7
Speech babble	4.1	67.1	97.1	12.9	79.3	98.0	19.2	79.0	95.5
Factory floor noise 1	1.2	59.7	97.1	19.2	73.9	98.0	34.1	89.5	98.5
Factory floor noise 2	7.9	83.6	97.6	52.2	93.7	98.6	75.5	97.5	98.9
Jet cockpit noise 1	0.6	54.1	96.6	15.2	82.7	97.7	44.4	88.7	97.5
Jet cockpit noise 2	0.3	59.0	97.1	26.5	83.6	98.1	52.0	92.4	98.0
Destroyer engine room	2.0	59.0	96.2	34.6	88.2	97.5	59.8	93.4	98.0
Destroyer operation room	3.1	80.2	99.0	11.8	83.8	98.9	54.8	94.7	98.7
F-16 cockpit noise	1.9	61.0	97.3	32.2	89.0	97.7	57.7	93.8	98.5
Military vehicle noise	73.9	97.2	98.9	87.4	97.1	99.0	90.5	96.1	98.4
Tank noise	18.1	91.1	98.9	63.1	96.4	98.8	82.4	98.2	99.3
Machine gun noise	85.4	94.0	98.2	82.5	92.1	98.2	84.9	92.8	97.4
Car interior noise	94.0	98.8	99.3	98.4	99.1	99.4	97.6	98.9	99.5
Average	**19.7**	**71.4**	**97.4**	**40.3**	**87.3**	**98.1**	**60.2**	**92.1**	**98.1**
Clean	99.5			99.5			99.3		

6.6.2 Robust speech recognition

The experiment of isolated vowel recognition carried out using the speech data from Section 6.2. We explore the noise robust property of the total system. Several noise circumstances were considered, i.e., from 0 dB SNR to 20 dB SNR. The 12 additive noises was selected from NOISEX-92. The noise robust techniques used for noise reduction are the Running Spectrum Filter(RSF) and running spectrum filter using frequency response masking (FRMRSF).

The analysis conditions and the selected characteristics were described with block in Figure 6.1 of MFCC, Figure 6.2 for PLP and Figure 6.6 (5.A) of DCT(log(CBI)). Ten-states HMM is used in our system to evaluate the noise robustness of the proposed techniques.

Figure 6.8 Applied robust speech recognition

The total system is depicted in Figure 6.8. Recent studies [16], [20] have demonstrated that filters which remove slow variations in the feature vectors used in speech recognition can yield significantly improved recognition rates. The filters used in these techniques can be implemented in various forms, which have band-pass frequency responses. The noise robust techniques are based on our proposing speech feature extraction using FRMRSF, RSF and DRA [20],[32]. FRMRSF or RSF focus on the modulation spectrum obtained from the time trajectory of spectrum and extract speech components by applying band-pass filtering. We employ FIR filtering as described in Chapter 5, the stability and the accuracy. RSF applies filtering twice and after log-process to and eliminate both addition noise and mintiplicative noise. DRA as described in Section 4.5, normalizes the maximum amplitudes of feature parameters and corrects the differences of dynamic ranges between that of trained data and observed speech data.

The conventional recognition system consists of ordinary feature extraction based MFCC, PLP or DCT(log(CBI)). Then, FRMRSF/DRA and RSF/DRA are applied together. Several experiments of recognition are evaluated in these conditions. MFCC and PLP have speech feature vectors 36-parameters which consist 12 cepstral coefficients, 12 first differential cepstral coefficients and 12 second differential cepstral coefficients whereas DCT(log(CBI)) has 32-parameters which consist 16 coefficients and 16 first differential coefficients respectively.

Table 6.12 shows the results of recognition rates versus white noise power. At a first glance same tendency of recognition rates are obtained in white noise. RSF, FRMRSF and DRA improves recognition performances. Comparing recognition performances of RSF, FRMRSF and DRA, the average accuracy of RSF and FRMRSF are better than DRA whore in white noise. The best performance is seen when both RSF and FRMRSF combined with DRA are applied. With no noise robust technique, the average accuracy for SNR of 0 dB 10 dB and 20 dB are above 2%, 34% and, 90% respectively. When, DRA is applied in speech features, the average accuracies improve by

Table 6.12 Recognition rates versus power using various noise robust features with DCT(log(CBI))

Speech Feature	Recognition rates [%]		
	0dB	10dB	20dB
Conventional	2.65	34.76	90.07
DRA	3.15	36.03	90.58
RSF	20.18	82.58	91.35
FRMRSF	19.24	80.49	89.87
RSF/DRA	21.43	87.49	90.68
FRMRSF/DRA	20.73	85.87	91.76

Table 6.13 Comparison between recognition MFCC performances with proposed RSF/DRA and FRMRSF/DRA

Type of noises	MFCC			MFCC with RSF/DRA			MFCC with FRMRSF/DRA		
SNR[dB]	0dB	10dB	20dB	0dB	10dB	20dB	0dB	10dB	20dB
White noise	2.65	34.76	90.75	21.43	87.49	89.68	19.73	85.87	91.76
Pink noise	5.64	65.68	93.54	56.89	83.53	92.96	56.96	82.53	92.96
HF channel noise	9.32	16.86	88.49	18.12	80.54	93.87	22.65	78.88	88.75
Speech babble	7.87	74.32	94.44	30.64	79.62	89.62	28.69	79.62	92.98
Factory floor noise	16.32	86.21	95.53	48.85	88.89	91.18	48.85	89.2	93.08
Jet cockpit noise	12.34	30.09	97.68	11.53	79.97	96.53	10.84	84.44	90.79
Destroyer engine room	12.77	24.65	87.67	43.45	83.26	94.59	43.45	83.05	96.65
F-16 cockpit noise	9.86	48.14	94.44	46.87	85.04	91.83	45.74	82.59	90.79
Military vehicle noise	37.69	86.75	92.07	78.53	83.32	95.37	68.53	83.64	94.81
Tank noise	65.43	79.81	90.74	82.89	92.21	94.98	84.41	91.84	96.64
Machine gun noise	17.84	51.94	97.68	26.63	86.32	92.64	33.06	83.51	90.79
Car interior noise	58.43	68.42	94.67	85.74	94.32	96.96	85.84	92.64	97.62
Average	21.34	55.63	93.14	45.96	85.37	93.35	45.72	85.31	93.13
Clean (No noise)	93.50			93.48			93.36		

Table 6.14 Comparison between recognition PLP and DCT(log(CBI)) performances with proposed RSF/DRA and FRMRSF/DRA

Type of noises / **SNR[dB]**	PLP			PLP with RSF/DRA			PLP with FRMRSF/DRA		
	0dB	10dB	20dB	0dB	10dB	20dB	0dB	10dB	20dB
White noise	1.84	28.5	88.64	16.81	85.18	90.04	20.86	90.87	90.84
Pink noise	3.64	39.54	95.27	26.52	79.36	93.18	41.84	83.53	90.98
HF channel noise	12.86	58.43	87.58	58.59	88.38	90.66	52.05	80.88	90.64
Speech babble	5.79	36.07	90.64	20.46	79.97	89.64	20.58	79.62	89.75
Factory floor noise	7.64	38.86	93.95	31.73	83.58	94.97	31.85	81.2	92.96
Jet cockpit noise	18.06	63.94	94.32	46.85	89.57	96.75	40.84	89.44	97.43
Destroyer engine room	10.65	40.64	91.75	39.45	83.04	95.85	39.45	83.05	93.97
F-16 cockpit noise	18.76	67.48	96.02	50.46	85.84	96.95	47.02	91.43	96.95
Military vehicle noise	34.02	76.75	96.65	65.12	83.92	96.29	62.53	83.64	97.05
Tank noise	64.65	89.02	95.16	73.05	92.16	96.85	71.54	91.78	96.85
Machine gun noise	26.73	61.22	97.13	46.89	86.94	93.64	43.22	82.51	95.78
Car interior noise	68.38	88.61	96.39	80.43	94.05	95.73	82.94	92.03	96.38
Average	22.75	57.42	93.62	46.36	85.98	94.12	46.22	85.83	94.04
Clean (No noise)	94.45			94.42			94.40		

Type of noises / **SNR[dB]**	DCT(log(CBI))			DCT(log(CBI)) with RSF/DRA			DCT(log(CBI)) with FRMRSF/DRA		
	0dB	10dB	20dB	0dB	10dB	20dB	0dB	10dB	20dB
White noise	4.27	35.58	89.84	23.09	81.85	90.67	23.86	80.84	90.14
Pink noise	5.84	38.62	96.32	25.24	83.53	94.85	25.84	83.23	95.45
HF channel noise	11.04	46.53	92.02	53.45	87.96	92.85	52.86	87.18	91.84
Speech babble	5.69	39.14	89.14	24.86	78.53	88.52	19.47	79.06	88.26
Factory floor noise	13.48	58.43	94.73	36.89	85.75	93.96	33.29	85.34	90.84
Jet cockpit noise	15.28	61.98	97.94	44.24	89.91	97.52	44.84	89.54	97.26
Destroyer engine room	5.97	36.73	90.38	24.86	85.63	93.96	24.42	82.92	93.87
F-16 cockpit noise	22.6	66.96	95.92	45.67	84.63	98.52	47.84	88.87	97.83
Military vehicle noise	37.93	69.32	95.46	64.29	87.52	96.97	65.89	87.92	96.83
Tank noise	55.39	86.93	96.42	73.55	89.75	95.12	71.57	90.74	97.62
Machine gun noise	34.05	68.42	95.67	70.46	87.01	94.73	69.89	85.28	95.03
Car interior noise	66.92	89.13	96.08	76.38	92.75	98.92	82.57	92.84	98.47
Average	23.20	58.14	94.16	46.91	86.23	94.71	46.86	86.14	94.73
Clean (No noise)	94.53			94.50			94.48		

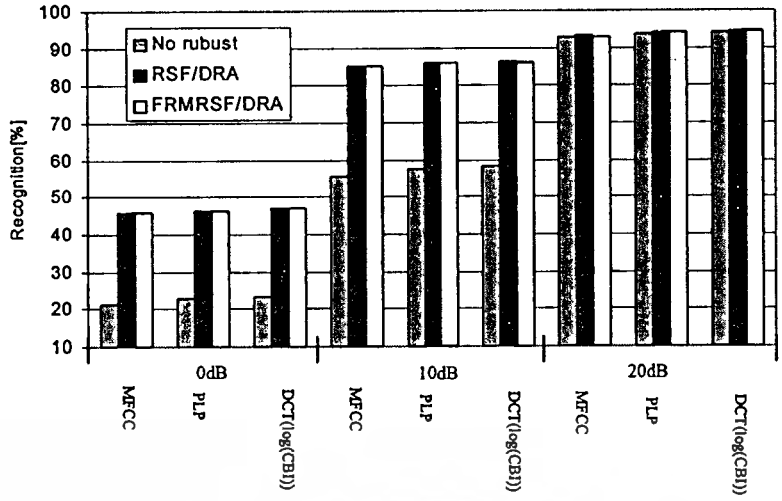


Figure 6.9 Recognition by RSF, RFMRSF with DRA

0.5%, 2% and 0.5 %, respectively. The best performance is seen when both RSF, FRMRSF and DRA are applied. The average accuracy for the latter are above 19%, 87% and 90 %, respectively. An improvement of 21%, 51% and 0.7%, respectively. The advantage of FRMRSF/ DRA and RSF/DRA are there fore confirmed.

The second last row on the Table 6.13 and 6.14 show the accuracy averaged from 12 types of noise from 0 dB SNR to 20 dB SNR using feature extraction based MFCC, PLP and DCT(log(CBI)) respectively. FRMRSF/DRA and RSF/DRA are applied. With no noise robust technique, the average accuracy for SNR of 0 dB, 10dB and 20 dB SNR are above 21%, 55% and, 93% respectively. When, FRMRSF/DRA and RSF/DRA are applied in speech features, performances improve by of 24 %, 29% and 0.21 %, respectively.

Figure 6.9 are summarized the performance results of recognition from Table 6.13 and 6.14. The average accuracy for SNR of 0 dB, 10dB and 20 dB SNR using feature extraction based MFCC, PLP and DCT(log(CBI)) respectively. FRMRSF/DRA and RSF/DRA are applied. With no noise robust technique, the average accuracy for SNR of 0 dB, 10dB and 20 dB SNR are above 21%, 55% and, 93% respectively. When, FRMRSF/DRA and RSF/DRA are applied in speech features, performances are above 45 %, 85% and 93 %, respectively.

6.7 Compares MFCC, PLP and DCT(log(CBI)) performances

We also compare the performance of our proposed method to other well-known front-end analysis, i.e., MFCC and PLP. MFCC and PLP as described in Section 6.3 and 6.4 Both front-end

analyses have been successfully used in many speech recognition systems. The proposed DCT(log(CBI)) as speech features is described as Section 6.4. Three different experiments are carried out. The performance results extracted from Table 6.1, 6.2, 6.6 and 6.7 are summarized in Table 6.15, it is found that the 16 of DCT(log(CBI)) coefficients with16 of differential coefficients provide accuracies that are 2% higher than that of 36 of MFCC coefficients and 36 of PLP coefficients. When, these features are then used with voice energy (E) and regression on the voice energy (RE), as shown in the Table 6.16, it is found that the 16 of DCT(log(CBI)) coefficients with16 differential coefficients and 1 of voice energy provide accuracies that are 0.5% higher than that of 36 of MFCC coefficients and 36 of PLP coefficients. When voice energy (E) is replaced by regress in energy(RE), the average recognition of DCT(log(CBI)) is accuracy above 95%, 3%

Table 6.15 Comparison MFCC, PLP and DCT(log(CBI)) performances

Feature parameter	Dimension	Recognition following features	
		Thai (%)	Laotian(%)
12 MFCC +12 delta + 12 delta	36	92.05	92.65
12 PLP +12 delta + 12 delta	36	92.23	92.67
16 DCT(log(CBI))+16 delta	32	94.34	94.71

Table 6.16 Comparison MFCC, PLP and DCT(log(CBI)) with proposed plus E/RE

Feature parameter	Parameters	Recognition following features	
		Thai (%)	Laotian(%)
12 MFCC +12 delta + 12 delta +E	37	92.42	92.68
12 PLP +12 delta + 12 delta + E	37	92.57	92.89
16 DCT(log(CBI))+16 delta + E	33	93.12	93.45
12 MFCC +12 delta + 12delta +RE	37	94.25	94.36
12 PLP +12 delta + 12 delta + RE	37	94.56	94.44
16 DCT(log(CBI))+16 delta + RE	33	95.47	95.52

Table 6.17 Comparison MFCC, PLP and DCT(log(CBI)) plus E/RE performances and time analysis of process

Feature parameter	Parameters	Times (Second)	Recognition Following features	
			Thai (%)	Laotian(%)
12MFCC +12delta +12 delta +E+QP	38	46	94.07	94.41
12 PLP +12 delta + 12 delta + E+QP	38	48	94.83	94.89
16 DCT(log(CBI))+16 delta + E+QP	34	212	95.56	95.84
12 MFCC+12delta+12delta+RE+QP	38	50	95.35	95.48
12 PLP+12 delta+12 delta+ RE+ QP	38	52	95.86	95.75
16 DCT(log(CBI))+16delta+ RE+QP	34	216	96.16	96.34

improvement 95%, 3% improvement over that of 36 of MFCC coefficients and 36 of PLP coefficients and 2% than that DCT(log(CBI)) plus E. Table 6.17 shown better accuracy using MFCC, PLP or DCT(log(CBI)) with regression on the voice energy and quantized pitch. 16 of DCT(log(CBI)) coefficients with16 of differential coefficients,1 of regression on the voice energy and 1 of quantized pitch obtain average accuracy above 96%, 2% higher than that of 36 of MFCC, and 1% higher than that of PLP combined RE and QP. As for processing speed is shown as third column in Table 6.17, we measured processing time of MFCC, PLP and DCT(log(CBI)).

6.8 Summary

In this report, the new techniques for Thai and Laotian vowel recognition. we propose a speech recognition system based on new features of speech spectrum on a Bark scale, the regression on the voice energy and a quantized pitch for tone features. The Bark scale is a psychoacoustics measurement on human hearing property and speech features extraction processes consists of three steps : (1) auto-regressive model (AR model), (2) critical band intensity (CBI), (3) logarithm CBI into discrete cosines transform (DCT). The regression on the voice energy is used to identify smooth or erupted changes in energy, while the changes in pitch are quantized to extract the tone feature. The detailed information feature is extracted by performing DCT sixteen parameters vectors , sixteen parameters of differential , a parameter of regression on the voice energy (RE) and a parameter of quantized pitch (QP). The average accuracy are carried out in

vowel recognition experiments shows the best performance. This result indicates that DCT(log(CBI)) combined ER and QP extracts speech characteristic more effectively than MFCC and PLP.

This thesis explores the extraction of speech features aiming noise robustness for speech recognition. The noise robust techniques for detection noise are the running spectrum filter using frequency response masking (FRMRSF) and running spectrum filter(RSF). The noise robust are used into the total system and thus several noise circumstances were considered, i.e., from 0 dB SNR to 20 dB SNR. From several experiments, it is shown that the performance of FRMRSF and RSF robust speech recognition system is superior to other robust speech recognition.

Chapter 7

Conclusions

In this thesis, we propose a speech recognition system based on new features on Bark scale together with noise robust speech feature extraction. Vowel speech recognition system for Thai and Laotian language are implemented. All the samples are recorded from Thai and Laotian speakers. We have taken isolated utterances of 24 and 27 vowels respectively in the context of initial consonant vowels. The new speech feature and robust speech techniques are summarized as follows:

7-1. We propose a vowel recognition system based on new features of speech spectrum on a Bark scale, the regression on the voice energy and a quantized pitch for tone. The Bark scale is a psychoacoustics measurement on human hearing property and speech features extraction processes consists of three steps: (1) auto-regressive model (AR model), (2) critical band intensity (CBI), (3) logarithm CBI into discrete cosines transform (DCT). The regression on the voice energy is used to identify smooth or erupted changes in energy, while the changes in pitch are quantized to extract the tone feature. The detailed information feature is extracted by performing DCT on the logarithm of CBI. Additionally, tones are constructed from the quantized pitch sequences with 3 possible outcomes [-1, 0, 1]. The pitch quantization allows the tone recognition engine to be gender-independent. A regression method on the voice energy is applied and the estimated coefficient is used as a feature to classify either smooth or erupted change for each word and hence, the short or long vowels in Thai/Laotian spoken language can be distinguished. Finally, the speech feature is represented by the 16 parameters of DCT(log(CBI)) coefficients, 16 of differential coefficients, 1 parameter of RE and 1 parameter of QP. It was found that, the proposed parameters are performance the conventional MFCC and PLP confer parts.

7-2. We have proposed a robust noise filtration system. The approach has been adapted and utilized in the features of MFCC, PLP and DCT(log(CBI)) to reduce both additive and multiplicative noises and hence, improves the recognition accuracy.

7-3. We have proposed and advanced noise robust speech recognition technique. The modified frequency response masking design based on band-pass filter is introduced

and it is applied to the design of running spectrum filter. The filters have been designed from Finite Impulse Response(FIR) filter allowing low number of filter taps while realizing narrow transition bandwidth. Using the new design, we have reduced the calculation cost by 60%.

We believe the result of this thesis is another step forward practical speech recognition system. Future work can extent the techniques for continuous speech recognition for Thai and Lao languages.

Reference

[1] Rabiner L., and Juang B.H., "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[2] Rabiner L. R., and Schafer R. W., "Digital Processing of speech signal", Prentice-Hall, p443, 1978.

[3] Lin-Shan L., "Voice dictation of Mandarin Chinese", Signal Processing Magazine", IEEE vol. 14, pp. 63 - 101, July 1997.

[4] Rabiner L.R., "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", Proceedings of the IEEE, vol. 77, no. 2 , pp. 257-287, Feb. 1989.

[5] Hermansky H., "Perceptual linear predictive (PLP) analysis for speech", J. Acoust. Soc. Amer., Vol. 87, No. 4, pp. 1738 – 1752, 1990.

[6] Zwicker E., and Fastl H., "Psychoacoustics: Facts and Models" Second Edition, Springer, pp 158-170, 1999.

[7] Smith, J. O., and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEE Trans. Speech & Audio Proc., vol. 7, no. 6, pp. 697-708, Nov. 1999.

[8] Songwatana K., Dejhan K., Miyanaga Y., and Khanthavivone K., "A vowels recognition model for Laotian language using transfer function on bark scale and hidden Markov modeling", NSIN 2005, pp. 36-39, May 2005.

[9] Songwatana K., and Kongkavitool W., "Unmixed Vowels Recognition in Thai spoken language using Vocal Tract Transfer functions on Bark Scale." WPMC'00 ,3rd , vol. 1, pp. 224- 227, 2000.

[10] Songwatana K., and Khanthavivone K., "Tone Recognition Model For Laotian Language Using Hidden Markov Model Technique quantized Pitch and Hidden Markov Model", Ladkrabang Engineering Journal, vol.19. No. 4, pp. 1-6, 2002.

[11] Chanthamenavong S., Maneenoi E., and Jitapunkul S., "Robust Method of Continuous Speech Recognition for A Tonal Language", ECTI International Conference (ECTICON2005), Trans., 2005.

[12] Songwatana K., Arungsrisangchai I., and Charumit C., " An Implementation of Speaker independent tone Recognition Model For Thai Language Using Hidden Markov Model Technique on quantized Pitch Sequence", vol.1. ROVPIA'99, pp. 211-217,1999.

[13] Songwatana K., Dejhan K., Miyanaga Y., and Khanthavivone K., "Short and Long Vowel Classification for Laotian Spoken Language using 3rd Order Polynomial Regression on the Voice Energy Function", ECTI-CON2005, vol.1, pp. 111-114, 2005.

[14] Songwatana K., and Sittiprasert K., " Short and Long Vowels Classification in Thai Spoken Language Using 2nd Order Polynomial Curve Fitting on the Voice Energy function", Ladkrabang Engineering Journal, Vol. 21, No. 2, pp 1-6, June 2004.

[15] Boll S., "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[16] Hermansky H., and Morgan N., "RASTA processing of speech," IEEE Trans. Speech Audio Process., vol. 2, pp. 578–589, 1994.

[17] Wada N., Hayasaka N., Yoshizawa S., and Miyanaga Y., "Robust Speech Recognition with Feature Extraction Using Combined Method of RSF and DRA", ISCIT2004, pp. 1001-1004, 2004.

[18] Hayasaka N., Miyanaga Y., and Wada N., "Running spectrum filtering in speech recogmition," SCIS Signal Processing and communication with Soft Computing, Oct 2002.

[19] Wada N., Yoshizawa S., Hayasaka N., and MiyanagaY., "Robust Speech Feature Extraction using RSF/DRA and Burst Noise Skipping" ECTI Trans. on Electrical Eng., vol. 3, No.2, pp 100-107, Aug. 2005.

[20] Hayasaka N., and Miyanaga Y,. "Spectrum Filtering with FRM for Robust Speech Recognition", IEEE Processing of International Symposium on Circuits and Systems, no. B3P-W.1, pp.3285 – 3288, May 2006.

[21] Noulnavong O., Sisuvhan P., Paphaphan B., Sengsulin B., Sihalaat S., and Sisawan K., "Advance Lao Grammar", Lao Education Ministry, UNICEF, 2003.

[22] Flanagan J.L., Speech Analysis, Synthesis, and Perception, 2nd ed., Springer-Verlag, New York,1972.

[23] Ling F., "Speaker Recognition", Technical University of Denmark Informatics and Mathematical Modelling, Kgs. Lyngby, 2004, IMM-THESIS: ISSN 1601-233X.

[24] Davis S.B., and Mermelstein P., "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences, " IEEE Trans. Acoust., Speech and Signal Processing, vol.28, No.4, pp. 357-366, 1980.

[25] Furui S., "Speaker independent isolated word recognition using dynamic features of the speech spectrum". IEEE Transactions on Acoustics, Speech and Signal Processing, 34:52–59, Feb 1986.

[26] Harte N., Vaseghi S. V., and Milner B., "Dynamic features for segmental speech recognition," In Proc. ICSLP'96, volume 2, pages 933–936, Philadelphia, PA, Oct. 1996.

[27] Cox S. J., " Hidden markov models for automatic speech recognition". British Telecom Technical Journal, 6(2):105–115, April 1988.

[28] Nossair Z., and Zahorian, S., "Dynamic Spectral Shape Features as Acoustic Correlates for Initial Stop Consonants," *J. Acoust. Soc. Amer.*, vol. 89, pp. 2978-2991, 1991.

[29] Furui S., "Speaker-independent isolates word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust. Speech Signal Process, Vol. ASSP-34, No. 1, pp. 52-59,1986.

[30] Atal B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoustic, Soc. Am., vol.55, no.6, pp.1304-1312, June 1974.

[31] Wada N., Miyanaga Y., Norinobu Y., and Yoshizawa S., " A consideration about an extraction of features for isolated word speech recognition in noisy environments" IEEE ISPACS2002, no. DSP2002-33, pp.19-22, 2002.

[32] ZHU Qi., Ohtsuki N., Miyanaga Y., and Yoshida N., "Noise-Robust Speech Analysis Using Running Spectrum Filtering", IEICE Transactions on Fundamentals of Electronics, Communication and Computer Science, vol. E-88-A, no.2, pp. 541-548, February 2005.

[33] Hayasaka N., Yoshizawa S., and Miyanaga Y., "Robust Isolated Word Recognition with Spectral Domain Filtering Based on the Estimated SNR", IEICE Transactions on Information and System (Japanese), vol. J89-D, no.10,pp.2296 – 2304, October 2006.

[34] Lim Y.C., "Frequency-Response Masking Approach for the Synthesis of Sharp Linear Phase Digital Filters," IEEE Trans. On Circuits and System, vol. CAS-33, no. 4, Apr. 1986.

[35] Yang R., Yong-Ching L., and Sydney R., "Design of Sharp Linear-phase FIR Band-Stop Filters using the Frequency response-masking Technique," Journal of Circuits, Systems and Signal Processing, 17, 1, 1-27, 1998.

[36] Sakata M., Miyanaga Y., and Yoshida N., "A new Design Method of Band-pass Filters Based on a Frequency-Response-Masking Technique," Proc. International Symposium on Intelligent Signal Processing and communication 2003, De c.2003.

[37] Kanedera N., Arai T., and Funada T., "Robust Automatic Speech Recognition Emphasizing Important Modulation Spectrum," I E ICE Trans, Vol. J84-D-11, No.7, pp.1261-1269, July 2001

[38] Yoshizawa S., Wada N., Hayasaka N., and Miyanaga Y., "Hardware Implementation of a Noise Robust Speech Recognition System Using RSF/DRA Techinique," Technical report of IEICE, CAS2003-42, pp.127-132, June 2003

[39] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., and Woodland P., The HTK Book (for HTK Version 3.0), Cambridge University Technical Services, Cambridge, UK, 2000.

List of publications

1. Songwatana K., Miyanaga Y., and Khanthavivone K., "A vowels Recognition Method for Laotian language using Transfer Function on Bark scale and classification by KNN technique" The 1st KMITL International Conference. Vol.1, pp. 274-250, 2004.
2. Songwatana K., Dejhan K., Miyanaga Y., and Khanthavivon K., "A vowels recognition model for Laotian language using transfer function on bark scale and hidden Markov modeling", Nonlinear Signal and Image Processing(NSIP 2005), pp. 36-39, May 2005.
3. Songwatana K., Dejhan K., Miyanaga Y., and Khanthavivone K., "Short and Long Vowel Classification for Laotian Spoken Language using 3rd Order Polynomial Regression on the Voice Energy Function", ECTI-CON2005, vol.1, pp. 111-114, 2005.
4. Suktangman N., Khanthavivone K., and Songwatana K., "Optimizing Vowel Recognition in Thai Spoken Language using Reduced LPC Spectrum and Reduced Feature Set of Critical Band Intensities", International Symposium on Communication and Information Technologies (ISCIT 2006), p. W3A-5, 2006.
5. Khanthavivone K., Hayasaka N., Miyanaga Y. and Songwatana K., "A Low Cost Running Spectrum Filter for Speech Recognition Using Modified Frequency Response Masking Technique", Journal of Signal Processing, vol. 11, No.3, July 20,2006.

About the Author

Kham Khanthavivone received the B.E. degree in Electronics Engineering and M.B. degree in Telecommunications Engineering from KMITL, Thailand in 2001 and 2003, respectively. Since 2004, he has been active in the research of new speech feature and noise-robust speech recognition.