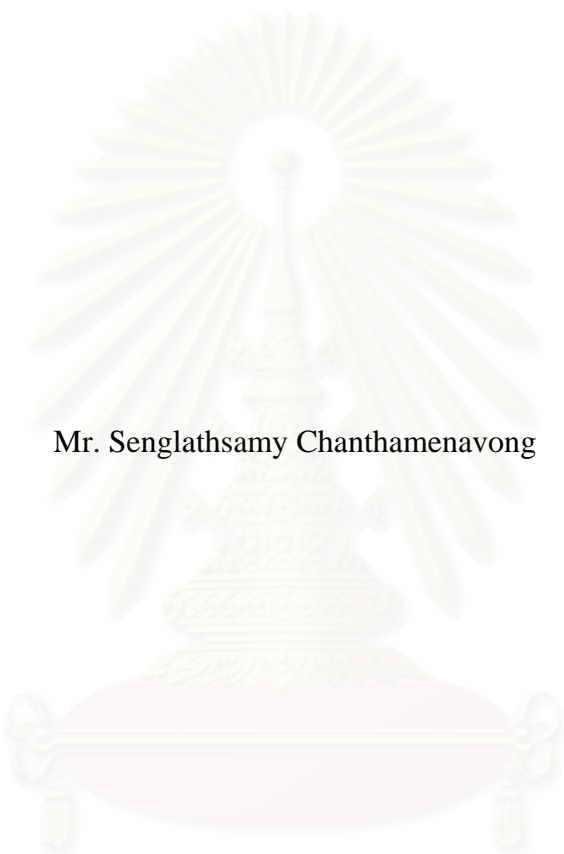


RECOGNITION TONE AND SYLLABLE IN COMBINATION
FOR LAO CONTINUOUS SPEECH



Mr. Senglathsamy Chanthamenavong

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Electrical Engineering

Department of Electrical Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2004

ISBN 974-53-1803-5

การรู้จำเสียงวรรณยุกต์ และ พยางค์ร่วมกันสำหรับเสียงพูดต่อเนื่องภาษาลาว



นาย แสงรัมย์ิ จันทมินาวงศ์

สถาบันวิทยบริการ

จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2547

ISBN 974-53-1803-5

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	RECOGNITION TONE AND SYLLABLE IN COMBINATION FOR LAO CONTINUOUS SPEECH
By	Mr. Senglathsamy Chanthamenavong
Field of study	Electrical Engineering
Thesis Advisor	Associate Professor Somchai Jitapunkul, Dr.Ing.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial
Fulfillment of the Requirements for the Master's Degree

..... Dean of the Faculty of Engineering
(Professor Direk Lavansiri, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Suvit Nakpeerayuth)

..... Thesis Advisor
(Associate Professor Somchai Jitapunkul, Dr.Ing.)

..... Member
(Associate Professor Boonserm Kijisirikul, Dr.Eng.)

..... Member
(Nisachon Tangsangiumvisai, Ph.D.)

4670643721 : MAJOR ELECTRICAL ENGINEERING

KEY WORD: TONAL SYLLABLE RECOGNITION / LAO SPEECH RECOGNITION / CONTINUOUS SPEECH RECOGNITION.

SENGLATHSAMY CHANTHAMENAVONG : RECOGNITION TONE AND SYLLABLE IN COMBINATION FOR LAO CONTINUOUS SPEECH. THESIS ADVISOR : ASSOC. PROF. SOMCHAI JITAPUNKUL, Dr.Ing., 113 pp. ISBN 974-53-1803-5.

This thesis proposed a robust method on tonal syllable recognition for continuous Lao speech recognition, by applying the specific characteristic of Lao language system as the conditional classification. The Lao language was studied in both acoustical and grammatical of Lao spoken. From the acoustical point of view, a syllable consists of initial consonant, vowel, final consonant and a tone. The final consonant is strongly influenced by the vowel duration. In addition, the part of vowel and final consonant is considered as voiced portion. Usually, only the voiced portion can carry tone information. Since, the initial consonant is not affected by the duration of the vowel. From the grammatical point of view, the tone of a syllable will be known, when the initial consonant, vowel and final consonant are know. Furthermore, the meaning of some syllables can change when different tones are applied. Therefore, this research investigated to increase the accuracy and recognition speed of the Lao speech recognition by applying the condition of Lao tonal grammar.

To implement a tonal syllable recognition system for Lao language based on continuous speech, various conventional speech units used in speech recognition systems have been investigated, in order to find out the optimal speech model for Lao speech recognition. In experiments, the existing techniques of tonal syllable recognition have also been experimented to evaluate the effectiveness and compare with the proposed method. As a result, experiment results shown that, the proposed system can achieve higher recognition rate at 66.85% and 81.92%, for speaker-independent and speaker-dependent, respectively. Moreover, the proposed system can recognizes with faster speed than that of baseline system, around 25%.

Department Electrical Engineering

Student's signature.....

Field of study Electrical Engineering

Advisor's signature.....

Academic year 2004

แสงรัศมี จันทมินาวงศ์ : การรู้จำเสียงวรรณยุกต์ และ พยางค์ร่วมกันสำหรับเสียงพูดต่อเนื่อง
ภาษาลาว. (RECOGNITION TONE AND SYLLABLE IN COMBINATION FOR LAO
CONTINUOUS SPEECH) อ. ที่ปรึกษา : รฟ. ดร.สมชาย จิตะพันธ์กุล, 113 หน้า. ISBN 974-
53-1803-5.

วิทยานิพนธ์ ฉบับนี้ นำเสนอระเบียบวิธีการรู้จำพยางค์ เสียงวรรณยุกต์ สำหรับการรู้จำ
เสียงพูดต่อเนื่องภาษาลาว ที่มีความคงทน โดยการประยุกต์ลักษณะสมบัติของระบบภาษาลาว ใน
การจำแนกแบบมีเงื่อนไข มีการศึกษาภาษาพูดลาวทั้งแบบ อะคูสติกและไวยากรณ์ ในมุมมองทางด้าน
อดคูสติก แต่ละพยางค์ ประกอบด้วย พยัญชนะ ต้น สระ พยัญชนะตัวสะกด หรือ พยัญชนะท้าย และ
วรรณยุกต์ พยัญชนะท้ายได้รับอิทธิพล อย่างมากจาก ช่วงการออกเสียงสระ นอกจากนั้น บางส่วนของ
เสียงสระและ พยัญชนะท้าย จัดเป็นส่วนเสียงก้อง (voiced portion) ตามปกติ มีเพียงส่วนนี้เท่านั้น ที่
จะมีข้อสังเกตของ วรรณยุกต์ เพราะว่า พยัญชนะต้น ไม่ได้รับผลกระทบจาก ช่วงการออกเสียงสระ
จากมุมมองทางไวยากรณ์ เมื่อรู้จัก พยัญชนะต้น สระ และ พยัญชนะท้าย จะทำให้รู้วรรณยุกต์ของ
พยางค์ได้อย่างแท้จริง ยิ่งไปกว่านั้น บางพยางค์มีความหมายเปลี่ยนไป ตามเสียงวรรณยุกต์ ดังนั้น
งานวิจัยนี้ ต้องการเพิ่มความเร็ว และอัตราของการรู้จำเสียงพูดภาษาลาว โดยการประยุกต์เงื่อนไข
ของไวยากรณ์ทางวรรณยุกต์ ของภาษาลาว.

เพื่อให้ ระบบรู้จำเสียงพยางค์ที่มีวรรณยุกต์ สำหรับภาษาลาว บนพื้นฐานเสียงพูดต่อเนื่อง
เป็นผล จะทำการศึกษาและทดสอบหน่วยเสียงแบบต่างๆที่มีใช้ในปัจจุบัน เพื่อนำมาสร้างแบบจำลอง
เสียงที่เหมาะสมที่สุด สำหรับระบบการรู้จำเสียงภาษาลาว เทคนิคการรู้จำเสียงพยางค์ที่มีวรรณยุกต์แบบ
ต่างๆ ได้ถูกนำมาทดสอบเพื่อเปรียบเทียบกับวิธีการที่นำเสนอ จากผลการทดสอบแสดงให้เห็นว่า
ระบบที่นำเสนอสามารถเพิ่มอัตราของการรู้จำได้เป็น 66.85% และ 81.92% สำหรับการรู้จำแบบไม่ขึ้น
ต่อผู้พูดและขึ้นต่อผู้พูด ตามลำดับ นอกจากนี้ ระบบที่นำเสนอสามารถรู้จำได้เร็วกว่าระบบเดิม
ประมาณ 25%

ภาควิชา วิศวกรรมไฟฟ้า
สาขาวิชา วิศวกรรมไฟฟ้า
ปีการศึกษา 2547

ลายมือชื่อนิสิต.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....

ACKNOWLEDGEMENTS

Firstly, I would like to express my deepest gratitude to my advisor, Assoc. Prof. Dr. Somchai Jitapunkul. 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 show my thankfulness to all of my thesis committee. The chairman, Mr. Suvit Nakpeerayuth, and the other committees, Assoc. Prof. Dr. Boonserm Kijirikul, and Dr. Nisachon Tangsangiumvisai, have given me some valuable comments regarding my research.

Thirdly, I would like to thank Dr. Ekkarit Maneenoi for his helpful assistance, patient guidance, friendship and knowledge throughout the duration of my study. I also thank all Electrical Engineering's staffs and students, who made this enjoyable experience. In addition, I would like to thank all my fellow students in the DSP research laboratory for their kindness and friendships.

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

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CONTENTS

	Page
ABSTRACT IN ENGLISH	iv
ABSTRACT IN THAI	v
ACKNOWLEDGEMENTS	vi
CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER I: INTRODUCTION	1
1.1 Literature reviews	2
1.2 Objectives	7
1.3 Scope and Goals of Research	7
1.4 Research Procedure	7
1.5 Research Contribution	8
CHAPTER II: FUNDAMENTAL TECHNIQUES FOR SPEECH	
RECOGNITION	9
2.1. Speech Production	9
2.1.1 Excitation Source Production	9
2.1.2 Discrete-Time Filter Modeling	11
2.2 Signal Processing for Speech Recognition	12
2.2.1 Short-Term spectral Analysis	12
2.2.2 Preemphasis	14
2.2.3 Linear Predictive Coding Analysis	14
2.2.4 Mel-frequency of filterbank Analysis	15
2.2.5 Cepstral Analysis	16
2.2.6 Coefficient Weighting	17
2.2.7 Delta Coefficients	18
2.2.8 Fundamental frequency	19
2.2.9 Energy Measures	19
2.3 Hidden Markov Model	20
2.3.1 Definition of the Hidden Markov Model	20

	Page
2.3.2 The Three Basic Problems of HMM	22
2.3.2.1 The Evaluation Problem	22
2.3.2.2 The Decoding Problem	22
2.3.2.3 The Estimation Problem	23
2.3.3 Solutions to the Three Basic Problems of HMM	23
2.3.3.1 Solution to the Evaluation Problem	23
2.3.3.1.1 The Forward Procedure	25
2.3.3.1.2 The Backward Procedure	27
2.3.3.2 Solution to the Decoding Problem	28
2.3.3.3 Solution to the Estimation Problem	31
2.3.4 Continuous Density Hidden Markov Model	36
2.3.4.1 Continuous Parameter Re-estimation	37
2.3.5 Hidden Markov Model for Speech Recognition	39
2.3.5.1 Composite Models for Continuous Speech Recognition	39
2.3.5.2 Multiple Observation Sequence	43
2.4 Large Vocabulary Continuous Speech Recognition	45
2.4.1 Search Algorithm	47
2.4.2 Language Modeling	48
CHAPTER III: THE ANALYSIS OF LAO LANGUAGE	49
3.1 Lao Consonants	50
3.1.1 Initial Consonants	52
3.1.2 Final Consonants	52
3.2 Lao Vowels	55
3.3 Lao Tones	57
3.4 Lao Syllable Structure	59
3.5 Lao Sentence	60
CHAPTER IV: METHODOLOGY	62
4.1 Data Collection	62
4.2 System Procedure	63
4.2.1 Feature Extraction	65
4.2.1.1 Speech Preprocessing	65

	Page
4.2.1.2 LPCC Measurement	67
4.2.1.3 MFCC Measurement	70
4.2.1.4 Energy	72
4.2.1.5 Fundamental Frequency	72
4.2.2 Hidden Markov Model Architectures	73
4.2.3 Speech modeling	75
4.2.4 Codebook	76
4.2.5 Tone Chart Applying	76
CHAPTER V: EXPERIMENTAL RESULTS	78
5.1 Base Syllable recognition	78
5.2 Tone recognition	84
5.3 Tonal Syllable recognition	86
CHAPTER VI: CONCLUSIONS	90
6.1 Summary and Conclusions	90
6.2 Future works	92
REFERENCES	93
APPENDICES	98
Appendix A	99
Appendix B	107
Appendix C	108
VITAE	113

LIST OF TABLES

	Page
Table 1.1 Comparison of continuous speech recognition (CSR) and isolated word recognition (IWR)	3
Table 1.2 Comparison of three techniques for tonal syllable recognition	7
Table 3.1 Twenty-seven original Lao consonant and sounds	50
Table 3.2 Six Consonant clusters and sounds	51
Table 3.3 Three Consonant classes, High, Middle, and Low Consonants	51
Table 3.4 Sonorant and Stop final consonants	52
Table 3.5 Vowel categories, monophthong, diphthong and special vowel	55
Table 3.6 Short vowel and Long vowel	56
Table 3.7 Lao tone mark	58
Table 3.8 Tone chart of Lao spoken in Vientiane	58
Table 5.1 Base syllable recognition based on monophoneme model	79
Table 5.2 Base syllable recognition based on subword model	79
Table 5.3 Base syllable recognition based on initial-final model	79
Table 5.4 Base syllable recognition based on onset-rhyme model	80
Table 5.5 Comparison of using different speech models for both LPCC+ Δ and MFCC+ Δ	81
Table 5.6 Comparison of using different number of cepstral coefficients, in case of applying 4 states HMM of onset-rhyme model	82
Table 5.7 Base syllable recognition based on individual onset and rhyme models	83
Table 5.8 Tone recognition of female speaker only	84
Table 5.9 Tone recognition of male speaker only	85
Table 5.10 Tone recognition of both male and female speakers	85
Table 5.11 Comparison of continuous tone recognition with the various number of HMM states	85
Table 5.12 Recognition of sub-system, including the final result for both speaker-independent and speaker-dependent recognitions	87
Table 5.13 Comparing the result of tonal syllable recognition incase of applying different techniques	88

LIST OF FEATURES

		Page
Figure 1.1	Speech recognition process	1
Figure 1.2	Tonal syllable recognition methods	5
Figure 1.3	Three methods of tonal syllable recognition	6
Figure 2.1	Anatomical structure of human vocal system	10
Figure 2.2	Speech waveform of voiced (a) and unvoiced (b)	11
Figure 2.3	Discrete-time speech production model	12
Figure 2.3	Hamming windows with 64 window length	13
Figure 2.4	The triangular mel-frequency scaled filter banks	15
Figure 2.5	The sequence of operations required for the computation of the forward variable $\alpha_{t+1}(i)$	26
Figure 2.6	Implementation of the computation of $\alpha_t(i)$ in terms of a lattice of observations t and S_i	27
Figure 2.7	The sequence of operations required for the computation of the backward variable $\beta_t(i)$	28
Figure 2.8	The sequence of operations required for the computation of the joint event that the system is in S_i at time t and S_j at time $t+1$	35
Figure 2.9	HMM with non-emitting entry and exit states	42
Figure 2.10	Tee model HMM	43
Figure 2.11	Structure of speech recognition according to information theory	46
Figure 3.1	The waveform and spectrogram of example initial consonant associated with a vowel	53
Figure 3.2	Example of the similarity between final and initial in transitional syllables	54
Figure 3.3	The waveforms and spectrogram of the syllables ending with stop and non-stop final consonants	55
Figure 3.4	The differential of short and long vowel categories	57
Figure 3.5	The general Lao syllable structure	58
Figure 3.6	Average pitch contours over syllables of Vientiane speaker, which are represented five tones	60

	Page
Figure 3.7	the pitch contour of Lao sentence 61
Figure 4.1	Tonal syllable recognition System for Lao language 63
Figure 4.2	The training step for HMM of tones and HMM of base syllables 64
Figure 4.3	Example of Preemphasized speech waveform 66
Figure 4.4	Example signal of windowing processed 67
Figure 4.5	The block daiagram of LPCC processor 68
Figure 4.6	The block daiagram of MFCC feature extraction 70
Figure 4.7	Example of 12 MFCC coefficients extracted 71
Figure 4.8	Energy contour of sample speech waveform 72
Figure 4.9	The block diagram of tone feature extraction 73
Figure 4.10	Conventional left-to-right Hidden Markov Model 74
Figure 4.11	Transition production model 75
Figure 5.1	Base syllable recognition with using different speech models 80
Figure 5.2	Comparison of using different speech models for both LPCC+ Δ and MFCC+ Δ 81
Figure 5.3	Onset-Rhyme recognition with different feature vector sets 82
Figure 5.4	Comparison of 3-5-onset-rhtme model with different number of MFCC coefficients 83
Figure 5.5	Comparison of continuous tone recognition results of using different tone models 86
Figure 5.6	Comparison of tonal syllable recognition results of using different techniques 88
Figure 5.7	Recognition speed of using different techniques in z-score 89

CHAPTER I

INTRODUCTION

Since the histories of human beings, voice is mainly used for human communication until now. Since, speaking is very convenient and natural ways for human communication. Nowadays, the machines have become more and more essential components in human society. However, almost the communication of human to machine is still based on touching method such as switch control, keyboard, mouse and etc; that method will disturb user, when his/her hand is busy. This 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. Applications of human-machine interaction involve in many tasks for example, voice dialing in mobile phones, aviation information retrieval, weather information retrieval, automated reservation, dictation and editing, transcription of broadcast data, etc.

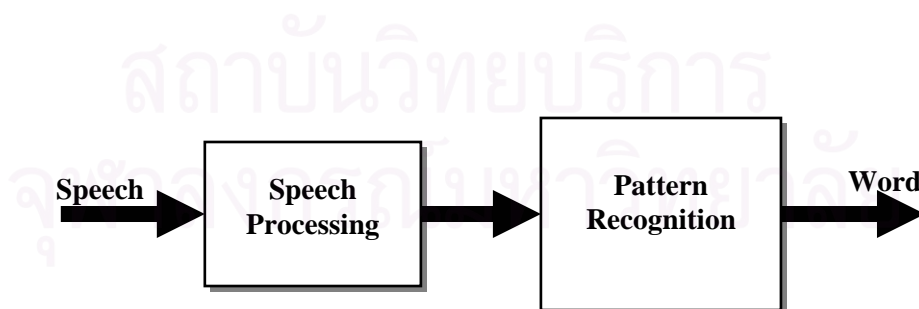


Figure 1.1 Speech recognition process

1.1 Literature reviews

The research of speech recognition has been continuously developed for the last half century. A number of significant advances in the past including signal processing, computational architectures, computer hardware, and programming techniques have contributed to rapid development of speech technology. Development of speech recognition system requires not only knowledge from the computer field, but also from other related fields. Multidisciplinary approaches have been applied in speech recognition research to make the system works effectively, such as: *signal processing, linguistics, acoustic physics, physiology, pattern recognition, computer science, and communication and information technology*, (Rabiner and Juang, 1993). Speech recognition system is a system, which can recognize the variety of utterance and phrase. Speech recognition can be separated in several different classes by describing what type of utterance, they have the ability to recognize and depend upon which mode is used, such as: *Isolated Word, Connected Word, Continuous Speech, Spontaneous Speech, Voice Verification/Identification, and etc.* The isolated word recognizer usually require each utterance to have quiet (Silence) on both sides of the sample. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have silence or none-silence states, where they require the speaker to wait between utterances (usually doing processing during the pauses), this method is suitable for a small vocabulary speech recognition application. Since the continuous speech recognition is a high flexible system, continuous speech recognizers allow users to speak almost naturally. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. The comparison of continuous speech recognition (CSR) and isolated word recognition (IWR) are also indicated in the Table 1.1.

In isolated word recognition system, word boundaries are not affected by any adjacent words, and therefore the utterance can be segmented into words with a short period of silence between words. Then each word is compared with the reference templates to produce isolated word recognition. One of the stochastic processes, *Hidden Markov Model (HMM)*, has been widely employed in the speech recognition system (Lee and Hon, 1989; Young, 1992; Ganpathiraju, et al., 2001; Lee, 1997). This process estimates the parameters of a probabilistic model of the data to produce the

representation of speech, which is robust to the variation in natural speech. Each acoustic model can be concatenated in a series to generate a composite model of a continuous speech. However, other recognizer algorithms of speech recognition such as, *Dynamic Time Warping (DTW)*, *Artificial Neural Network (ANN)*, *Hybrid System (NN-HMM)*, and etc; will not be considered on this thesis.

Table 1.1 Comparison of continuous speech recognition (CSR) and isolated word recognition (IWR)

	IWR	CSR
Complexity	Low	High
Flexibility	Low	High
Vocabulary	Very Small	Very Large
Generalization	No	Yes
Trainability	Poor	Good
Accuracy	Depend on Vocabulary	Good

Initially, speech recognition system utilized a simple pattern matching technique, to recognize word where the reference templates are created based upon the word model. Word recognition systems have reached their limitations on the number of words in the vocabulary to be modeled individually, which training data could not be shared between words (Huang, et al., 2001). Speech recognition system using word-based approach is not productive because it is impossible to implement such a recognizer that covers the whole language.

Presently, most recognition systems use acoustic units corresponding to phonemic units, such as, *syllable*, *monophone*, *Diphone*, *Triphone*, *Initials/Finals*, and *Onset/Rhyme*. Despite of using traditional context-dependent units such as diphone or triphone employed in the English speech recognizer, initials/finals are utilized as a fundamental unit in a Mandarin Chinese dictation machine (Lee, et al., 1997), and onsets/rhymes are utilized as speech unit in a Thai speech recognition (Maneenoi E., et al., 2003). Different syllable structures of those languages will result in using different speech units. The choice of speech unit depends on language structure and

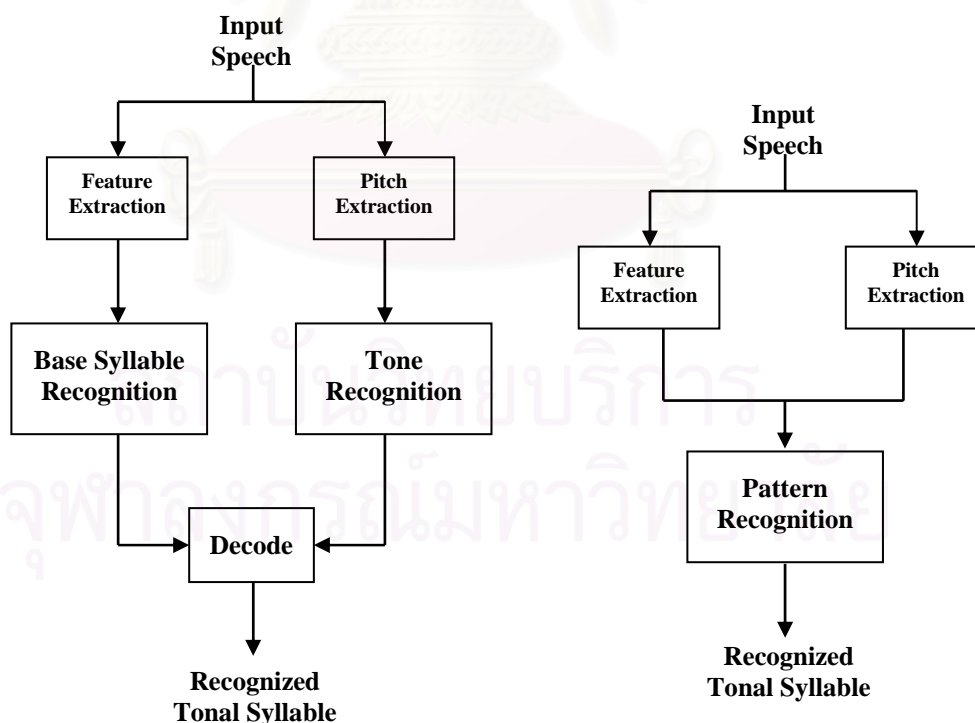
the availability of sufficient training data for constructing effective reference models. Since each language has its own attribute, choosing suitable speech units leads to effective utilization of the training data and a good performance of speech recognizer. Lao syllable structure is similar to some of those, therefore, the studying of those speech units, such as Thai and Mandarin languages can be adapted to Lao speech recognition as well.

In addition, Lao is a tonal language, which tones are lexically significant of word meaning. Therefore, tone information is very essential for speech recognition of Lao languages. Several researches of tone recognition have been developed for tonal language such as, Mandarin, Canton, Thai, Japanese and etc. The tone of a syllable is determined by pitch (*Fundamental frequency*) contour of the entire syllable, where pitch information on the main vowel of a syllable is sufficient to determine the tone of that syllable (Haiping Li et al., 2001). The recognition of Lao tones by using HMM was presented for isolated-syllable speech recognition, where the features have been generalized for speaker-independent by using Three-level quantization technique (Khanthavivone K., et al., 2002). However, there is not any research to present tone recognition for Lao continuous speech. Although, many tone recognition methods have done well in isolated-syllable speech recognition. But it has some difficulties in handling continuous speech. In continuous speech, there are several interacting factors those affect pitch realization of tones. To prevent those effects, such as, tonal assimilation and declination effects, which are compensated by the tone information of neighboring syllables, *Context-dependent-Tone model* (CD-T) and *half-tone model* (H-T), were proposed for continuous Thai tone recognition (Thubthong N., et al., 2001).

During 1980s and early 1990s decades, a popular method for recognition of tonal language was the two-step method (as show in Figure 1.2.(a).). Initially, this method separately starts to recognize the base syllable by its consonants and vowel. And, to recognize tone of the syllable by classifying pitch contour of that syllable. At last, the recognition of tonal syllable is executed in combination of both base syllable and tone recognitions (Wang H.M., et al., 1994). Later, in 1997, the one-step method for continuous speech recognition of tonal language was proposed by Chen C.J., et al., and again, in 2001 by Haiping Li. (as show in Figure 1.2.(b).). Also, comparison of three different methods based on hidden Markov model framework for recognition of tonal syllable of continuous speech, such as, *Joint Detection*, *Sequential Detection*,

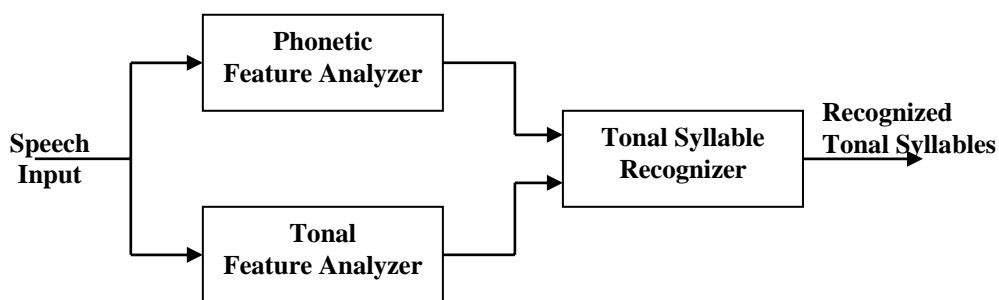
and *Linked Detection* (as shows in Figure 1.3), have been presented (Demeechai T., et al., 2001). The joint detection method is conceptually similar to that of Chen C.J, and Haiping Li, and the sequential detection method is conceptually similar to that of Wang H.M. However, linked detection is a new proposed method by Demeechai T. Although, the computational complexity of sequential detection is higher than that of both linked detection and joint detection, but it has simple architecture, which is suitable to implement. The comparison between three conventional techniques of tonal syllable recognition is shown in Table 1.2. Furthermore, in sequential detection, specific characteristic of a language model can be applied to detect which is possible to recognize tones.

Because of those advantages of sequential detection method as shown in Table 1.2., this thesis has concentrated on improvement of the performance in tonal syllable recognition for Lao language based on sequential detection method, by proposing an algorithm to apply specific characteristic of Lao language, and Lao tones chart as the conditions of recognition process.

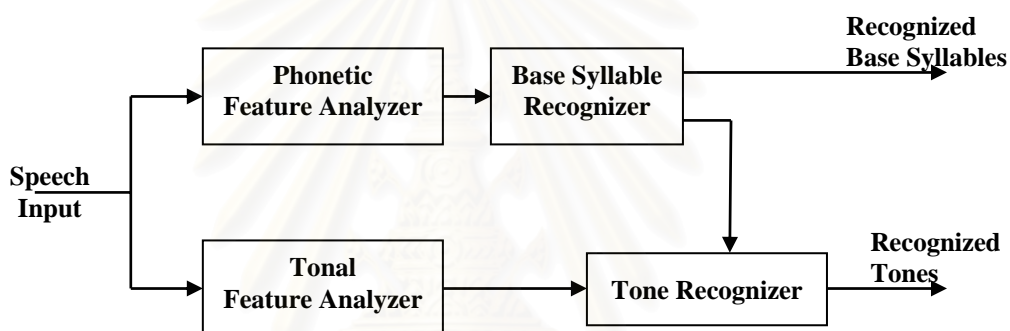


(a). Two-step tonal syllable recognition (b). One-step tonal syllable recognition

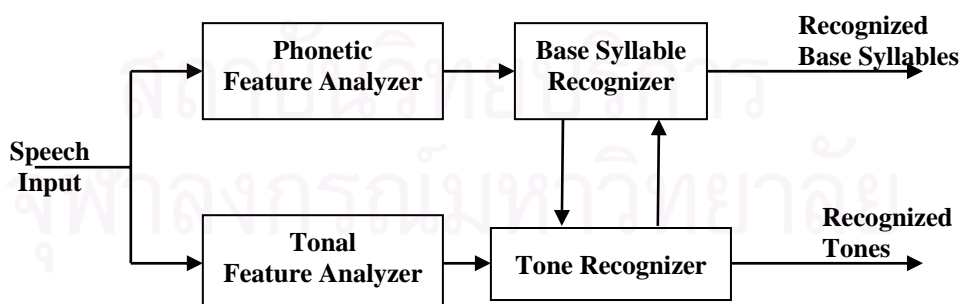
Figure 1.2 Tonal syllable recognition methods



(a). Joint detection of tonal syllable recognition



(b). Sequential detection of tonal syllable recognition



(c). Linked detection of tonal syllable recognition

Figure 1.3 Three methods of tonal syllable recognition

Table 1.2 Comparison of three techniques for tonal syllable recognition

	Joint Detection	Sequential Detection	Linked Detection
Accuracy	high	high	higher
Complexity	low	high	high
Recognition Speed	high	good	low
Memory	large	small	small
Vocabulary	small	large	large
Flexibility	poor	good	good

1.2 Objectives

1. To develop the technique in order to recognize tonal syllable sound for Lao appropriate speech.
2. To study and select a suitable acoustic and language models for Lao continuous speech recognition.
3. To provide basic acknowledge for Lao continuous speech recognition.

1.3 Scope and Goals of Research

1. Recognition of Lao voiced with embedded tone based on continuous speech by using HMM.
2. The sample data should not be less than 50 speakers (30 speakers for training and 20 speakers for recognizing).
3. The proposed method can improve up recognition speed faster than baseline method, and the recognition rate should be over 80.00%.

1.4 Research Procedure

1. Study the characteristic of Lao Linguistics, Lao grammar and sound unit of Lao language.
2. Study the fundamental of speech recognition and speech signal processing.
3. Study the Hidden Markov Model algorithms for continuous speech recognition.

4. Collect the sample data and analysis to select a suitable feature vector for Lao continuous speech recognition.
5. Find out a suitable system's structure for tonal syllable recognition system.
6. Develop and modify system's structure to decreasing system complexity and increase system accuracy.
7. Summarize and prepare the thesis.

1.5 Research Contribution

1. To improve the research on speech recognition for Lao language approach.
2. Give an understanding on the Lao speech recognition mechanism.
3. It's expected that the proposed system can be implemented in real-time application of automatic speech recognition machine.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER II

FUNDAMENTAL TECHNIQUES FOR SPEECH RECOGNITION

This chapter provides a concise introduction to the theory and application of fundamental techniques for speech recognition. The mechanism of speech production is early described in the first part of this chapter. The speech signal processing algorithms for speech recognition will be described to measure feature parameters of speech signal, these have used in this thesis. Then, theory of the hidden Markov model will be elaborated. As last, details of the large vocabulary continuous speech recognition system will be explained.

2.1. Speech Production

Speech production process begins with a thought which shows the initial communication message. Following the rules of spoken languages and grammatical structure, words and phrases are selected and ordered. After the thought constructs into language, brain sends commands by means of motor nerves to the vocal muscles, which move the vocal organs to produce sound (Ling F., et al., 2004).

Speech production can be divided into three principal components: excitation production, vocal tract articulation, and lips and/or nostrils radiation.

2.1.1 Excitation Source Production

Excitation powers the speech production process. It is produced by the airflow from lungs, and then carried by trachea through the vocal cords as indicated by Figure 2.1. During inspiration, air is filled into lungs, and during expiration the energy will be spontaneously released. The trachea conveys the resulting air stream to the larynx.

Larynx refers as an energy provider to serve inputs to the vocal tract, and 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*, (Ling F., et al., 2004).

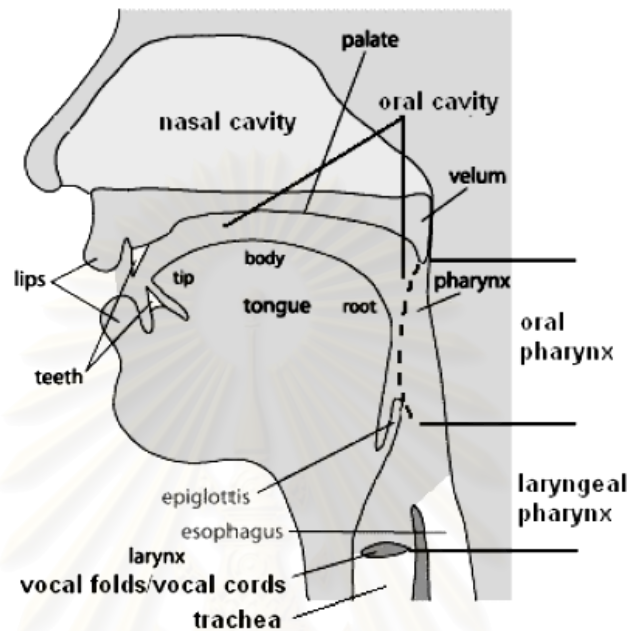


Figure 2.1 Anatomical structure of human vocal system (Ling F., et al., 2004)

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 (show in Figure 2.2.a)). 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 (show in Figure 2.2.b)).

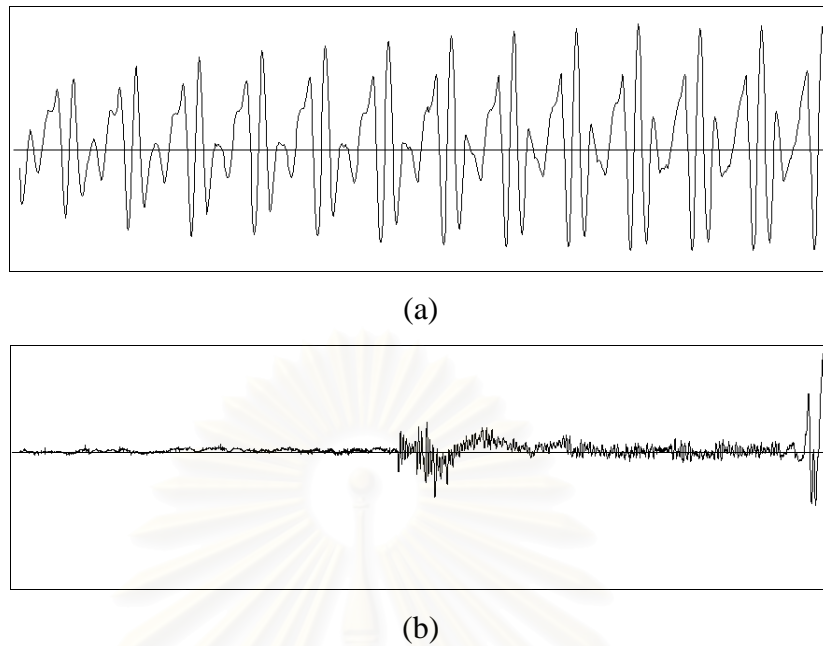


Figure 2.2 Speech waveform of voiced (a) and unvoiced (b)

2.1.2 Discrete-Time Filter Modeling

The speech production is normally divided into three principal components: excitation production, vocal tract articulation and lips and/or nostrils radiation. Which have no coupling between each other, we assume that these three components are linear, separately and planar propagation (Ling F., et al., 2004). Consequently we could construct a simple linear model, discrete-time filter model, for speech production, which consists of excitation production part, vocal tract filter part and radiation part separately, shown in Figure 2.3. The excitation part corresponds to the vibrating of the vocal cords (glottis) causing voiced sounds, or to a constriction of the vocal tract causing a turbulent airflow and thus causing the noise-like unvoiced excitation. Therefore, Sound can be computed as the product of three respective Fourier transfer functions flowing as

$$S(\omega) = U(\omega)H(\omega)R(\omega) \quad (2.1)$$

In time-domain, relation Eq. (2.1) will be presented as

$$s(n) = u(n) * h(n) * r(n) \quad (2.2)$$

where the excitation spectrum $U(\omega)$ and radiation $R(\omega)$ are mostly constant and well known a priori, the vocal tract transfer function $H(\omega)$ is the characteristic part to determine articulation (Ling F., et al., 2004). Therefore it deserves our special attention on how it can be modeled adequately.

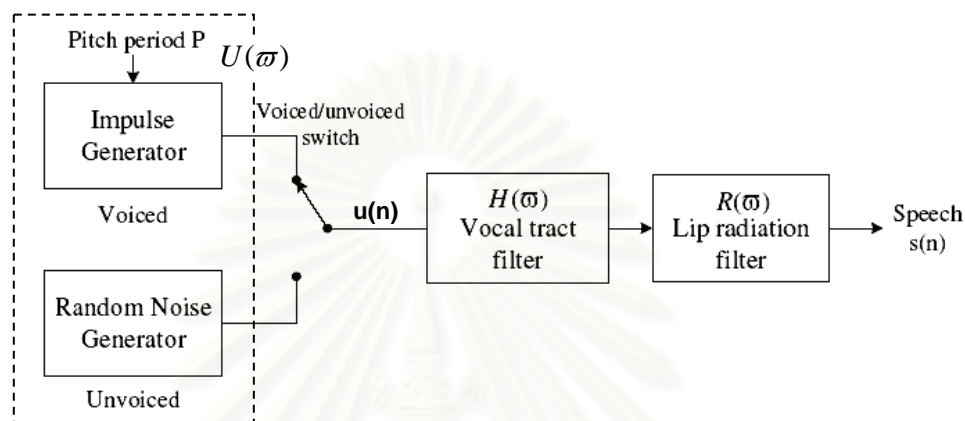


Figure 2.3 Discrete-time speech production model
(adapt from Ling F., et al., 2004)

2.2 Signal Processing for Speech Recognition

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 corresponding to framework of this thesis, will be reported in this section.

2.2.1 Short-Term spectral Analysis

The short-term analysis principle is a valid 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 (Ling F., et al., 2004). Speech analysis is generally carried out in frequency domain with short segments across which the speech signal is assumed to be stationary, and this

kind of analysis is often called *short-term spectral analysis*. In addition, the short-time Fourier analysis is one kind of short-term spectral analysis. It depends on windowing of speech waveform and 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 $0 \leq n \leq N - 1$, where N is window size. The extracted signal with window function can define by

$$\tilde{x}_l(n) = x_l(n)w(n) \quad 0 \leq n \leq N - 1 \quad (2.3)$$

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

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N - 1}\right); \quad 0 \leq n \leq N - 1 \quad (2.4)$$

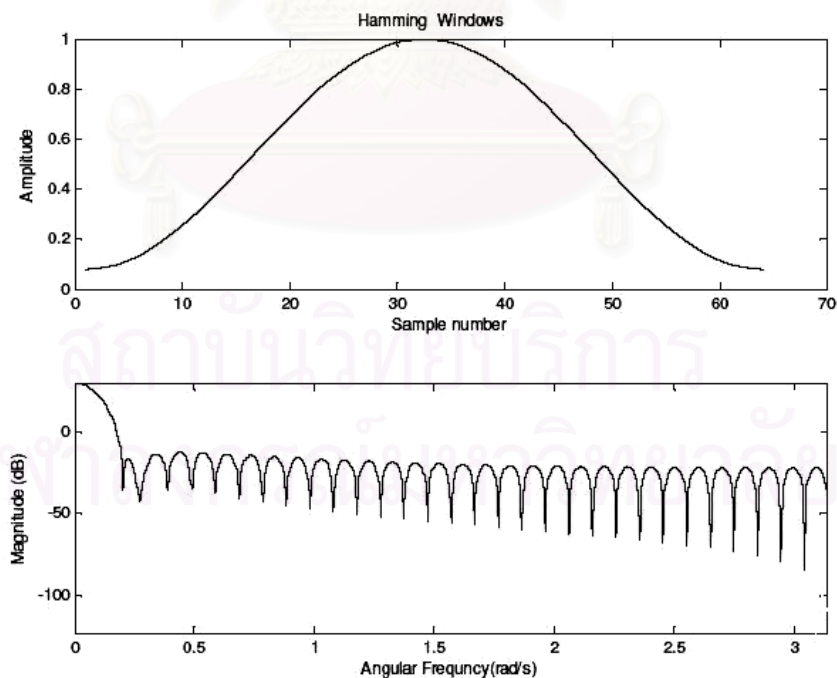


Figure 2.3 Hamming windows with 64 window length

2.2.2 Preemphasis

The preemphasis is early applied to smooth spectrum of input speech signal $s(n)$ by passing first order FIR filter transfer function (E.q. (2.5) and (2.6)), where α is the coefficient of filter, $\tilde{s}(n)$ is the product of Signal preemphasis at sequence n , $s(n)$ is the input speech signal at sequence n , since $s(n-1)$ is the input speech signal at sequence $n-1$.

$$H(z) = 1 - \alpha z^{-1} \quad (2.5)$$

$$\tilde{s}(n) = s(n) - \alpha s(n-1) \quad (2.6)$$

Usually, the coefficient of filter (α) of speech processing is mostly used at 0.95 to 0.99 (Rabiner and Juang, 1993).

2.2.3 Linear Predictive Coding Analysis

Linear Predictive Coding (LPC) can provide a complete description for a speech prediction model at the vocal tract level (Rabiner and Juang, 1993). The basic idea underlying LPC is that each speech sample x_n , can be represented as a linear combination of previous samples, and prediction error can be minimized according to the mean-square value of the prediction error e_n , which is defined by

$$e_n = x_n - \sum_{i=1}^p a_i x_{n-i} \quad (2.7)$$

where p is the order filter of LPC analysis, and a_i are LPC coefficients.

The LPC coefficients, which minimize the mean-square prediction error over a short segment (frame) of the speech waveform, can be obtained by setting the partial derivative of the mean-square prediction error E , with respect to each a_i , equal to zero (Eq. (2.8) and (2.9)).

$$E_n = \sum_m e_n^2(m) \quad (2.8)$$

$$\frac{\partial E_n}{\partial a_i} = 0; \quad i = 1, 2, \dots, p \quad (2.9)$$

By minimizing the prediction error, the LPC technique models the spectrum as a smooth spectrum of an order p all-pole filter (Rabiner and Juang, 1993). The value of p required for adequate modeling of vocal tract depends on the sampling frequency. The LPC coefficients can be obtained by solving the Yule-Walker equation (Maneeno E., et al., 2003). The solution of this equation can be achieved with various algorithms (Rabiner and Juang, 1993). However, the autocorrelation method has been mainly used for this task.

2.2.4 Mel-frequency of filterbank Analysis

Alternatively, the spectral features can be obtained, by passing the speech signal through a bank of bandpass filters. The filterbanks are generally triangular (see figure 2.4.), and they are equally spaced along the mel scale, which is defined by (Maneeno E., et al., 2003)

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.10)$$

where f denotes the real frequency, and $Mel(f)$ denotes the perceived frequency.

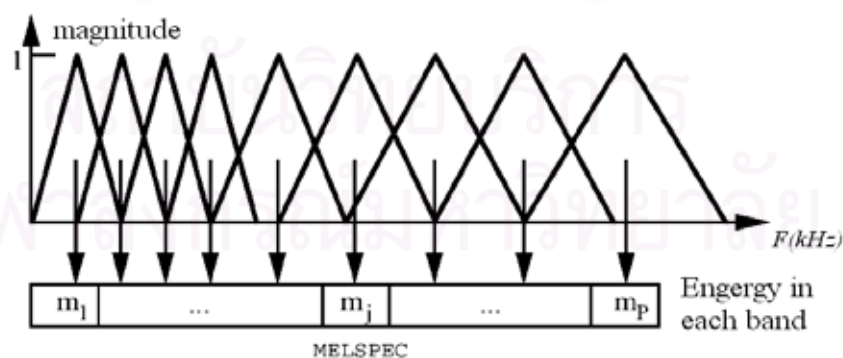


Figure 2.4 The triangular mel-frequency scaled filter banks

The mel scale is a linear frequency spacing below and logarithmic above 1 kHz. This scale is known to be a good scale for approximating the ability of human

auditory system to discriminate frequencies (Ling F., et al., 2004). To implement the filterbank, each segment of speech data is transformed using a Fourier transform and the magnitude is taken. Each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. If the cepstral parameters are computed from the log filterbank amplitude using the Discrete Cosine Transform (DCT) as shown in Eq. (2.11), then, the mel-frequency cepstral coefficients (MFCCs) are obtained.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{i\pi}{N}(j-0.5)\right) \quad (2.11)$$

where N is the number of filterbank channels and m_j is the log filterbank amplitude.

Since, mel-frequency cepstral coefficients (MFCC) are the best known and most commonly used features for not only speech recognition, but speaker recognition applications as well (Ling F., et al., 2004).

2.2.5 Cepstral Analysis

The observed speech sequence is a convolution of the excitation and the vocal tract filter impulse response in the time domain or the product of the excitation and the filter spectral in the frequency domain. In the frequency domain, the product of the excitation and filter spectral is transformed to the summation of these two spectral by logarithmic operation. Then, the transformation from the frequency domain back to the time domain results in the ‘‘cepstrum’’, which has a number of properties suitable to the deconvolution of speech (Maneenoi E., et al., 2003).

There are several variants of cepstral coefficients in use. Two of the most common are linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). The cepstral coefficients, c_n , obtained from LPC analysis, can be computed recursively from the LPC coefficients as

$$c_n = a_n + \sum_{i=1}^{n-1} \frac{i}{n} c_i a_{n-i}; \quad 1 \leq n \leq p \quad (2.12)$$

$$c_n = \sum_{i=1}^{n-1} \frac{i}{n} c_i a_{n-i}; \quad n > p \quad (2.13)$$

where a_n are LPC coefficients, c_n are cepstral coefficients and p is order of LPC coefficients.

2.2.6 Coefficient Weighting

The lower cepstrum coefficients have been found to be strongly affected by speaker-specific characteristics, because the low-order cepstral coefficients are sensitive to overall spectral slopes (Rabiner and Juang, 1993, Maneenoi E., et al., 2003). This speaker-dependent effect on the cepstrum coefficients is undesirable, and needs to be eliminated for speaker-independent speech recognition. Moreover, the high-order cepstral coefficients are sensitive to noise and other forms of noiselike variability. These sensitivities need to be minimized by weighting technique. Weighting the cepstrum coefficients or less emphasis is given on the lower cepstrum coefficients. The process of weighting or windowing the cepstrum coefficients are also known as cepstrum liftering. Several weighting functions or lifting windows have been proposed for speech recognition (Juang, et al., 1987). The raised sine function is one of the liftering windows, $w(i)$, which has been found to work very well in speech recognition. This window is defined as

$$w(i) = 1 + \frac{Q}{2} \sin\left(\frac{i\pi}{Q}\right); \quad i = 1, 2, \dots, N \quad (2.14)$$

where Q is a liftering parameter, which is typically found experimentally. The new weighted coefficients were obtained as

$$\hat{c}(i) = c(i)w(i); \quad i = 1, 2, \dots, N \quad (2.15)$$

2.2.7 Delta Coefficients

The cepstral representation of speech spectrum provides a good representation of the local spectral properties of the signal for the given analysis frame (Maneenoi

E., et al., 2003). These coefficients are considered to be static or instantaneous coefficients, which are computed without taking into account past or future spectrum information. Spectral changes, such as formant transitions, play an important role in speech perception. Therefore, it seems reasonable to incorporate such spectral changes in the features to enhance speech recognition extending the analysis to include information about the temporal cepstral derivative. To introduce the cepstral order into the cepstral representation, the m^{th} cepstral coefficient at time t is denoted by $c_m(t)$. The time derivative of the log magnitude spectrum has a Fourier series representation of the form

$$\frac{\partial}{\partial t} [\log |S(e^{j\omega}, t)|] = \sum_{m=-\infty}^{\infty} \frac{\partial c_m(t)}{\partial t} e^{j\omega} \quad (2.16)$$

It is well known that $c_m(t)$ is a discrete time representation, where t is a frame index, simply using a first or second order difference is inappropriate to approximate the derivative. Hence, a better method to approximate $\frac{\partial c_m(t)}{\partial t}$ is using an orthogonal polynomial fit over a finite length window; that is

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \sum_{k=-K}^K k c_m(t+k) \quad (2.17)$$

where μ is an appropriate normalization constant and $(2K+1)$ is the number of frames over which the computation is performed (Maneenoi E., et al., 2003).

Based on the computation described above, for each frame t , the results of O_t is a vector of N weighted and an appended vector of N time derivative MFCC/ LPCC; that is

$$O_t = \{\hat{c}_1(t), \hat{c}_2(t), \dots, \hat{c}_N(t), \Delta \hat{c}_1(t), \Delta \hat{c}_2(t), \dots, \Delta \hat{c}_N(t)\} \quad (2.18)$$

where O_t is a vector of $\hat{c}_i(t)$ and $\Delta \hat{c}_i(t)$ with N components. $\hat{c}_i(t)$ are the estimated MFCC/LPCC coefficients at frame t , and $\Delta \hat{c}_i(t)$ are the delta coefficients of estimated coefficients at frame t .

2.2.8 Fundamental frequency

The basic property of a vocal cord source is its periodicity expressed by the duration of a complete voice period or by the inverse value of the voice fundamental frequency (F_0), (Fant, 1970). Related to the number of times the vocal folds open and close per second, the frequency of vocal cords vibration directly determines the lowest frequency of the sound, which is produced (Borden and Harris, 1980). The duration of pitch cycle can always vary from one period to the other. This changing of pitch period perceived as pitch pattern or intonation contour of phrase or sentence is particularly effective in expressing differences in attitude and differences in meaning. Fundamental frequency is an important acoustic feature especially in tonal languages. Different fundamental frequency contours indicate different lexical meanings of the syllable. Another important exploiting of fundamental frequency is voiced/unvoiced classification. Since, only voiced sound has quasi-periodic, in other hand, unvoiced sound does not (Maneenoi E., et al., 2003).

Various fundamental frequency extraction techniques are generally grouped into three major categories according to their principal features (Furui, 2001). Firstly, the waveform processing consists of methods for detecting the periodicity peaks in the waveform. Secondly, the correlation processing is composed of methods widely used in digital signal processing of speech. Lastly, spectrum processing comprises the methods for tracking pitch in spectral domain. However, among the correlation techniques, autocorrelation function is become successful in pitch measurement of short-time analysis of speech signal, when it was combined with time domain center clipping (Sondhi, 1968; Rabiner, 1977; and Khanthavesone K., et al., 2002).

2.2.9 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

$$E(n) = \sum_{m=1}^M [s(m)]^2 \quad (2.19)$$

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.

2.3 Hidden Markov Model

The hidden Markov model (HMM) is a powerful statistical approach for the study of time series modeling with many of the classical probability distributions. The HMM approach provides a framework, which includes an automatic supervised training algorithm with mathematically proven convergence, the Baum-Welch algorithm. In addition, an efficient decoding scheme, the Viterbi algorithm, is incorporated in the HMM. Many successful speech recognition systems have employed the HMM approach as a major recognition part. Not only the HMM can be used in speech recognition, but it also can be applied in statistical language modeling, spoken language understanding, machine translation, and so on.

2.3.1 Definition of the Hidden Markov Model

A natural extension to the Markov chain introduces a non-deterministic process that generates output observation symbols in any given state. Thus, the observation is a probabilistic function of the state. This new model is known as a hidden Markov model, which can be viewed as a double embedded stochastic process with an underlying stochastic process or the state sequence not directly observable. The state sequence is hidden, and can only be observed through another set of observable stochastic processes. A hidden Markov model is basically a Markov chain, where the output observation is a random variable generated according to the output probabilistic function associated with each state. A set of output probability distributions of each hidden state can be either discrete probability distributions or continuous probability density functions. To describe the HMM characteristics, the following HMM elements are defined (Maneenoi E., et al., 2003)

- 1) The number of states in the model is N . Generally, the states are interconnected in such a way that any state can be reached from any other

state. The individual states and the state at time t are denoted as $S = \{S_1, S_2, \dots, S_N\}$ and q_t respectively.

- 2) The number of distinct observation symbols per state is M . The observation symbols correspond to the physical output of the system being modeled. The individual symbols is denoted as $V = \{V_1, V_2, \dots, V_M\}$
- 3) The state transition probability distribution is $A = \{a_{ij}\}$, where

$$a_{ij} = P[q_{t+1} = S_j / q_t = S_i], \quad 1 \leq i, j \leq N \quad (2.20)$$

- 4). The observation symbol probability distribution in state is $B = \{b_j(k)\}$, where

$$b_j(k) = P[V_k \text{ at } t / q_t = S_j], \quad 1 \leq j \leq N, 1 \leq k \leq M \quad (2.21)$$

- 5). The initial state distribution is $\pi = \{\pi_i\}$, where

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \quad (2.22)$$

Since a_{ij} , $b_j(k)$, and π_i are all probabilities, they must satisfy the following properties: $a_{ij} \geq 0$, $b_j(k) \geq 0$, $\pi_i \geq 0$ for all i, j , and k ,

$$\sum_{j=1}^N a_{ij} = 1 \quad (2.23)$$

$$\sum_{k=1}^M b_j(k) = 1 \quad (2.24)$$

$$\sum_{i=1}^N \pi_i = 1 \quad (2.25)$$

Given appropriate value of N , M , A , B , and π , the HMM can generate an observation sequence $O = o_1, o_2, \dots, o_T$, where each observation o_t is one of the symbols from V , and T is the number of observations in the sequence. A complete specification of an HMM requires two constant parameters, N and M , representing the total number of states and the size of observation symbols, and three sets of

probability measures, A , B , and π . For convenience, the compact notation is used to represent the complete parameter set of the model

$$\lambda = (A, B, \pi) \quad (2.26)$$

2.3.2 The Three Basic Problems of HMM

Given the definition of HMM, there are three basic problems of interest that must be solved for the model to be useful. These problems are the following:

2.3.2.1 The Evaluation Problem

Given the observation sequence $O = o_1, o_2, \dots, o_T$, and the model $\lambda = (A, B, \pi)$, 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.

2.3.2.2 The Decoding Problem

Given the observation sequence $O = o_1, o_2, \dots, o_T$, and the model $\lambda = (A, B, \pi)$, what the most likely state sequence $Q = q_1, q_2, \dots, q_T$ 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.

2.3.2.3 The Estimation Problem

Given the observation sequence $O = o_1, o_2, \dots, o_T$, how to adjust the model parameters $\lambda = (A, B, \pi)$, to maximize $P(O | \lambda)$. The problem concerns how to optimize the model parameters so as to best describe how a give observation sequence

comes about. 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.

2.3.3 Solutions to the Three Basic Problems of HMM

2.3.3.1 Solution to the Evaluation Problem

To calculate the probability of an observation sequence $O = \{o_1, o_2, \dots, o_T\}$, given the model λ , $P(O | \lambda)$. The most straightforward way is to enumerate every possible state sequence of length T (the number of observations). For every fixed state sequence

$$Q = \{q_1, q_2, \dots, q_T\} \quad (2.27)$$

where q_1 is the initial state. The probability of the observation sequence O for this state sequence is

$$P(O | Q, \lambda) = \prod_{t=1}^T P(o_t | q_t, \lambda) \quad (2.28)$$

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

$$P(O | Q, \lambda) = b_{q_1}(o_1)b_{q_2}(o_2)\dots b_{q_T}(o_T) \quad (2.29)$$

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

$$P(Q | \lambda) = P(q_1 | \lambda) \prod_{t=1}^T P(q_t | q_{t-1}, \lambda) \quad (2.30)$$

$$= \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \quad (2.31)$$

$$= a_{q_0 q_1} a_{q_1 q_2} \dots a_{q_{T-1} q_T} \quad (2.32)$$

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

$$P(O, Q | \lambda) = P(O, Q | \lambda)P(Q, \lambda) \quad (2.33)$$

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

$$P(O | \lambda) = \sum_{\text{all } Q} P(O, Q | \lambda)P(Q, \lambda) \quad (2.34)$$

$$= \sum_{\text{all } Q} \prod_{t=1}^T a_{q_{t-1}q_t} b_{q_t}(o_t) \quad (2.35)$$

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. (2.35)) involves on the order of $O(N^T)$ calculations. At every time $t=1, 2, \dots, T$, 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.

2.3.3.1.1 The Forward Procedure

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

$$\alpha_t(i) = P(o_1 o_2 \dots o_t, q_t = S_i | \lambda) \quad (2.36)$$

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

$$\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \quad (2.37)$$

2. Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), \quad \begin{array}{l} 1 \leq t \leq T-1 \\ 1 \leq j \leq N \end{array} \quad (2.38)$$

3. Termination

$$P(O | \lambda) = \sum_{i=1}^N \alpha_T(i) \quad (2.39)$$

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 2.5. This figure shows how S_j can be reached at time $t+1$ from the N possible states, S_i , $1 \leq i \leq N$, at time t . Since $\alpha_t(i)$ is the probability of joint event that o_1, o_2, \dots, 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, \dots, 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 , $1 \leq i \leq N$ 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 $o_1, o_2, \dots, o_t, o_{t+1}$, is obtained in S_j . The computation of the induction step is performed for all S_j , $1 \leq j \leq N$, for a given t . This computation is then iterated for

$t=1,2,\dots,T-1$. Finally, the termination step gives the desired calculation of $P(O|\lambda)$ as the sum of the terminal forward variables $\alpha_T(i)$.

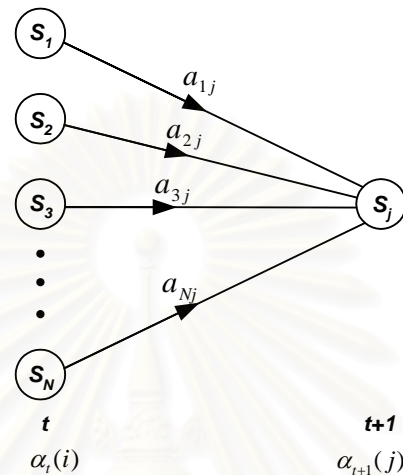


Figure 2.5 The sequence of operations required for the computation of the forward variable $\alpha_{t+1}(i)$

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 2.6. 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 $\alpha_1(i)$, $1 \leq i \leq N$, is calculated. At time $t=1,2,\dots,T$, the only values of $\alpha_t(j)$, $1 \leq i \leq N$, 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.

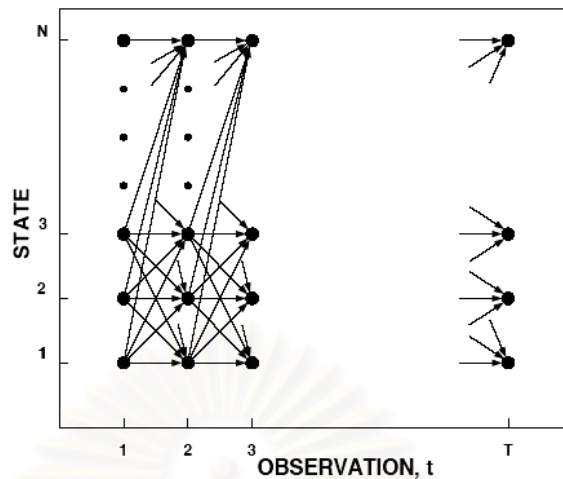


Figure 2.6 Implementation of the computation of $\alpha_t(i)$ in terms of a lattice of observations t and S_i

2.3.3.1.2 The Backward Procedure

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

$$\beta_t(i) = P(o_{t+1}o_{t+2}\dots o_T | q_t = S_i, \lambda) \quad (2.40)$$

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

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \quad (2.41)$$

2) Induction

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad (2.42)$$

$$t = T - 1, T - 2, \dots, 1, \quad 1 \leq j \leq N$$

The initialization arbitrarily defines $\beta_i(i)$ 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 $b_j(o_{t+1})$ term), and then accounts for the remaining partial observation sequence from S_j (the $\beta_{t+1}(j)$ term).

The computational complexity of $\beta_i(i)$ is similar to that of $\alpha_i(i)$, which also produces a lattice with observation length and state number. The induction step is illustrated in Figure 2.7.

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.

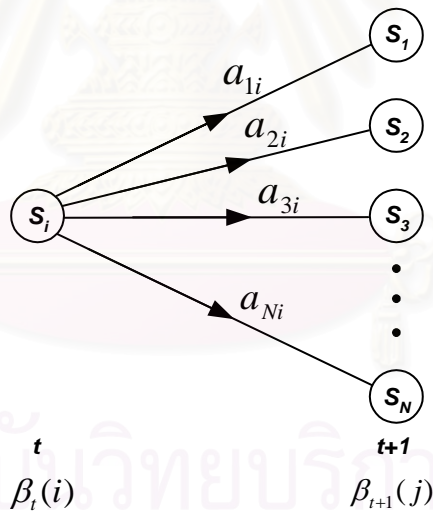


Figure 2.7 The sequence of operations required for the computation of the backward variable $\beta_t(i)$

2.3.3.2 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

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) \quad (2.43)$$

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

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O | \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} \quad (2.44)$$

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

$$\sum_{i=1}^N \gamma_t(i) = 1 \quad (2.45)$$

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

$$q_t = \arg \max_{1 \leq i \leq N} [\gamma_t(i)]; \quad 1 \leq t \leq T \quad (2.46)$$

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 Eq. (2.46) 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 $P(Q|O, \lambda)$, which is equivalent to maximizing $P(Q, O | \lambda)$. A formal technique for finding this single best state sequence is called the Viterbi algorithm.

2.3.3.2.1 The Viterbi Algorithm

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

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P[q_1 q_2 \dots q_{t-1} = S_i, o_1 o_2 \dots o_t | \lambda] \quad (2.47)$$

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 Eq. (2.47) becomes to

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij} \right] b_j(o_{t+1}) \quad (2.48)$$

To actually retrieve the state sequence, we need to keep track of the argument that maximized Eq. (2.48), 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

$$\delta_1(i) = \pi_i b_i(o_1); \quad 1 \leq i \leq N \quad (2.497)$$

$$\psi_1(i) = 0 \quad (2.50)$$

2. Induction

$$\delta_t(j) = \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(o_t); \quad \begin{array}{l} 2 \leq t \leq T \\ 1 \leq j \leq N \end{array} \quad (2.51)$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \quad \begin{array}{l} 2 \leq t \leq T \\ 1 \leq j \leq N \end{array} \quad (2.52)$$

3. Termination

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.53)$$

$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.54)$$

4. Path (state sequence backtracking)

$$q_t^* = \psi_{t+1}(q_{t+1}^*); \quad t = T-1, T-2, \dots, 1 \quad (2.55)$$

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 Eq. (2.51), 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.

2.3.3.3 Solution to the Estimation Problem

The most difficult problem in HMM is to determine a method to adjust the model parameters $\lambda = (A, B, \pi)$ to 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 $\lambda = (A, B, \pi)$ that

$P(O | \lambda)$ 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.

a) Baum-Welch Re-estimation Algorithm

The mathematical foundations of the Baum-Welch algorithm for the maximum likelihood estimation were established by Baum. 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} , $i = 1, 2, \dots, p$, $j = 1, 2, \dots, q$,

defined over a stochastic domain, $D: x_{ij} \geq 0$, $\sum_{j=1}^{q_i} x_{ij} = 1$, through a series of

transformations performed on $\{x_{ij}\}$, was firstly purposed. The transformation is defined as

$$T(x_{ij}) = \frac{x_{ij} \frac{\partial P(X)}{\partial x_{ij}}}{\sum_{j=1}^{qt} x_{ij} \frac{\partial P(X)}{\partial x_{ij}}} \quad (2.56)$$

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 Eq. (2.56) 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 forwardbackward 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.

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) \quad (2.57)$$

The sequence of events leading to the conditions required by Eq. (2.57) is illustrated in Figure 2.8. From the definition of the forward and backward variables, $\xi_t(i, j)$ can be written in the form

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)} \quad (2.58)$$

$$= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \quad (2.59)$$

where the numerator term is just $P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$ and the division by $P(O | \lambda)$ 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

$$\gamma_t(i) = \sum_{j=1}^N \xi_t(i, j) \quad (2.60)$$

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

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } S_i \quad (2.61)$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{expected number of transitions from } S_i \text{ to } S_j \quad (2.62)$$

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 \mathbf{A} , \mathbf{B} , and π are

$$\bar{\pi}_i = \text{expected frequency in state } S_i \text{ at time } \gamma_1(i) \quad (2.63)$$

$$\begin{aligned} \bar{a}_{ij} &= \frac{\text{expected number of transitions from } S_i \text{ to } S_j}{\text{expected number of transitions from } S_i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \quad (2.64)$$

$$\begin{aligned} \bar{b}_j(k) &= \frac{\text{expected number of times in } S_j \text{ and observing symbol } v_k}{\text{expected number of times in } S_j} \\ &= \frac{\sum_{t=1}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \end{aligned} \quad (2.65)$$

From Eq. (2.63) to (2.65), it can be proven that either:

- 1) The initial model λ defines a critical point of likelihood function, where new estimates equal old ones, or
- 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 minima only, and that in the most problems of interest, the optimization surface is very complex and has many local minima.

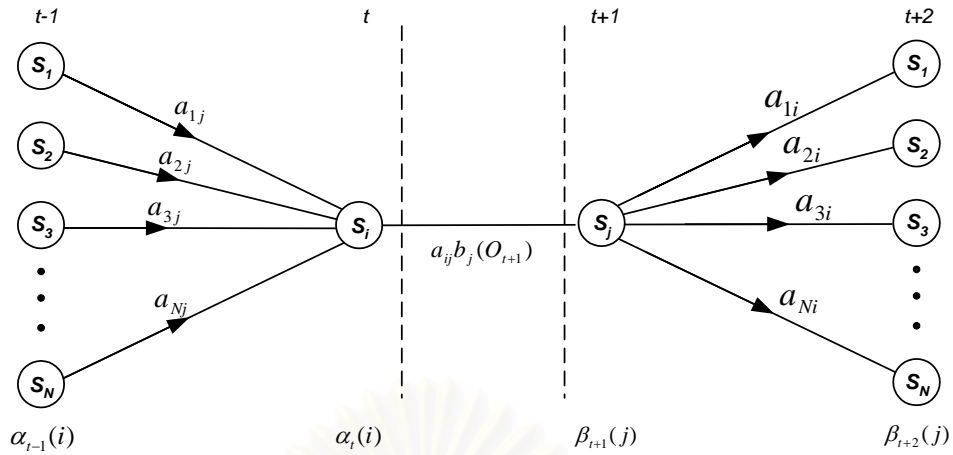


Figure 2.8 The sequence of operations required for the computation of the joint event that the system is in S_i at time t and S_j at time $t+1$

b) Multiple Observation Sequence

Note that a single observation sequence is not enough for re-estimation of the HMM parameters. Hence, in order to have sufficient data to make reliable estimates of all model parameters, multiple observation sequences are used. The re-estimation formulas can be easily extended to such multiple observation sequences. Let a set of L observation sequences denoted as

$$O = [O^{(1)}, O^{(2)}, \dots, O^{(L)}] \quad (2.66)$$

where $O^l = \{o_1^l, o_2^l, \dots, o_{T_k}^l\}$ is the L^{th} observation sequence. Assuming that observation sequences are independent of each other, the parameter estimations of HMM is then based on the maximization of

$$P(O^L | \lambda) = \prod_{l=1}^L P(O^l | \lambda) \quad (2.67)$$

$$= \prod_{l=1}^L P^l \quad (2.68)$$

Since the re-estimation formulas are based on frequencies of occurrence of various events, the re-estimation formulas are modified by adding together the

individual frequencies of occurrence of each sequence. Thus, the re-estimation formula for the transition probability a_{ij} , can be computed:

$$\bar{a}_{ij} = \frac{\sum_{l=1}^L \frac{1}{P_l} \sum_{t=1}^{T_l-1} \alpha_t^l(i) a_{ij} b_j(O_{t+1}^l) \beta_{t+1}^l(j)}{\sum_{l=1}^L \frac{1}{P_l} \sum_{t=1}^{T_l-1} \alpha_t^l(i) \beta_t^l(i)} \quad (2.69)$$

Similarly, the re-estimation formula for the observation symbol probability distribution in state j , $b_j(k)$, can be computed:

$$\bar{b}_j(k) = \frac{\sum_{l=1}^L \frac{1}{P_l} \sum_{t=1}^{T_l-1} \alpha_t^l(i) \beta_{t+1}^l(j)}{\sum_{l=1}^L \frac{1}{P_l} \sum_{t=1}^{T_l-1} \alpha_t^l(i) \beta_t^l(i)} \quad (2.70)$$

s.t. $o_t = v_k$

2.3.4 Continuous Density Hidden Markov Model

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 2.3.3.3.a), 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.

2.3.4.1 Continuous Parameter Re-estimation

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

$$N(O; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(O-\mu)^T \Sigma^{-1} (O-\mu)} \quad (2.71)$$

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

$$b_j(o_t) = \sum_{m=1}^M c_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}) \quad (2.72)$$

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

$$\sum_{m=1}^M c_{jm} = 1 \quad 1 \leq j \leq N \quad (2.73)$$

$$c_{jm} \geq 0 \quad \begin{array}{l} 1 \leq j \leq N \\ 1 \leq m \leq M \end{array} \quad (2.74)$$

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

$$P(O | \lambda) = \sum_{\text{all } Q} P(O, Q | \lambda) \quad (2.75)$$

$$= \sum_{\text{all } Q} P(Q, \lambda) P(O | Q, \lambda) \quad (2.76)$$

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

$$Q(\lambda, \bar{\lambda}) = \frac{1}{P(O | \lambda)} \sum_{\text{all } Q} P(O, Q | \lambda) \log P(O, Q | \bar{\lambda}) \quad (2.77)$$

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

$$\bar{c}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m)}{\sum_{t=1}^T \sum_{m=1}^M \gamma_t(j, m)} \quad (2.78)$$

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) o_t}{\sum_{t=1}^T \gamma_t(j, m)} \quad (2.79)$$

$$\bar{\Sigma}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) \cdot (o_t - \mu_{jm})(o_t - \mu_{jm})'}{\sum_{t=1}^T \gamma_t(j, m)} \quad (2.80)$$

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

$$\gamma_t(j, m) = \left[\frac{\alpha_t(i)\beta_t(j)}{\sum_{j=1}^N \alpha_t(i)\beta_t(j)} \right] \left[\frac{c_{jm}N(O, \mu, \Sigma)}{\sum_{m=1}^M c_{jm}N(O, \mu, \Sigma)} \right] \quad (2.81)$$

The re-estimation formula for a_{ij} is identical to the one used for discrete observation densities.

There are two possible options in the design of the mixtures. Either the Gaussian mixtures are state specific or they are shared (tied) between different states of the HMM. HMM with state specific Gaussian mixtures is called continuous density HMM. HMM that shares Gaussian mixtures among different states is called semi-continuous HMM or tied mixture HMM (Maneeno E., et al., 2003).

2.3.5 Hidden Markov Model for Speech Recognition

2.3.5.1 Composite Models for Continuous Speech Recognition

The parameter estimation and decoding techniques in the previous section are defined to apply to a single HMM mapped onto an isolated word. One of the advantages of the HMM approach is the ease with which it can be adapted to a continuous recognition environment. In order to extend to the continuous model, two modifications are made to the HMM structure. The first modification was already discussed in subsection 2.3.3.1; the addition of the entry and exit states to each model. The entry and exit states are defined as non-emitting states, which take Δt time to traverse, where Δt is negligibly small. Thus, the forward and backward probabilities that correspond to the entry and exit states are those at $t - \Delta t$ and $t + \Delta t$, where t is the time value at the immediately following or preceding state respectively. Therefore, the constraints are

$$a_{11} = 0 \text{ and } a_{Ni} = 0 \quad \forall i \quad (2.82)$$

which simply ensure that the entry and exit states can only be occupied for one transition. The other structural change is the addition of glue models. These models have only one emitting state, plus the entry and exit state, along with a non-zero entry to exit transition probability. These glue models are often called null or tee models

(Young, et al., 1999.). A model with entry and exit states is depicted in Figure 2.9 and a tee model is shown in Figure 2.10. Using tee models and non-emitting entry and exit states, a series of HMMs, with tee model between words, may be linearly combined into a single HMM for training purpose.

The modification required for the training formulas can be generated in a straightforward manner. The notation, a superscript q in parentheses representing the current model, is used as the notation that a training sentence model is represented by Q HMMs placed in sequence. The resulting forward and backward recurrent algorithms can be rewritten directly from the earlier definitions and new model structure. The forward equations are:

Initialization

$$\alpha_1^q(1) = \begin{cases} 1 & q = 1 \\ \alpha_1^{q-1}(1)a_{1N_q}^{q-1} & \text{Otherwise} \end{cases} \quad (2.83)$$

$$\alpha_1^q(j) = \alpha_1^q(1)a_{1j}^q b_1^q(o_t) \quad (2.84)$$

$$\alpha_1^q(N_q) = \sum_{i=2}^{N_q-1} \alpha_1^q(i)a_{iN_q}^q \quad (2.85)$$

Recursion

$$\alpha_t^q(1) = \begin{cases} 0 & q = 1 \\ \alpha_{t-1}^{q-1}(N_{q-1}) + \alpha_{t-1}^{q-1}(1)a_{1N_{q-1}}^{q-1} & \text{Otherwise} \end{cases} \quad (2.86)$$

$$\alpha_t^q(j) = \left[\alpha_t^q(1)a_{1j}^q + \sum_{i=2}^{N_q-1} \alpha_{t-1}^q(i)a_{ij}^q \right] b_j^q(o_t) \quad (2.87)$$

$$\alpha_t^q(N_q) = \sum_{i=2}^{N_q-1} \alpha_t^q(i)a_{iN_q}^q \quad (2.88)$$

The corresponding backward equations are:

Initialization

$$\beta_T^q(N_q) = \begin{cases} 1 & q = 1 \\ \beta_T^{q+1}(N_{q+1})a_{1N_{q+1}}^{q+1} & \text{Otherwise} \end{cases} \quad (2.89)$$

$$\beta_T^q(i) = \beta_T^q(N_q)a_{iN_q}^q \quad (2.90)$$

$$\beta_T^q(1) = \sum_{j=2}^{N_q-1} \beta_T^q(j) a_{1j}^q b_j^q(o_T) \quad (2.91)$$

Recursion

$$\alpha_t^q(N_q) = \begin{cases} 0 & q = 1 \\ \beta_{t+1}^{q+1}(1) + \beta_{t+1}^{q+1}(N_{q+1}) a_{1N_{q+1}}^{q+1} & \text{Otherwise} \end{cases} \quad (2.92)$$

$$\beta_t^q(i) = \beta_t^q(N_q) a_{iN_q}^q + \sum_{j=2}^{N_q-1} \beta_{t+1}^q(j) a_{ij}^q b_j^q(o_{t+1}) \quad (2.93)$$

$$\beta_t^q(1) = \sum_{j=2}^{N_q-1} \beta_t^q(j) a_{1j}^q b_j^q(o_t) \quad (2.94)$$

The Baum-Welch re-estimation equations for transition probabilities will now be split into four categories:

1. Internal transitions between emitting states,
2. Transitions from the entry state into emitting states,
3. Transitions from emitting states into the exit state,
4. Tee transitions from the entry state directly to the exit state, generally zero for non-tee models.

The equations are all similar to the original transition re-estimation formulas, with some primary differences above. The resulting formulas are:

$$a_{ij}^{r(q)} = \frac{\sum_{t=1}^{T-1} \alpha_t^q(i) a_{ij}^{(q)} b_j^{(q)}(o_{t+1}) \beta_{t+1}^{(q)}(j)}{\sum_{t=1}^{T-1} \alpha_t^{(q)}(i) \beta_t^{(q)}(i)} \quad (2.95)$$

$$a_{1j}^{r(q)} = \frac{\sum_{t=1}^T \alpha_t^q(1) a_{1j}^{(q)} b_j^{(q)}(o_t) \beta_t^{(q)}(j)}{\sum_{t=1}^T \alpha_t^{(q)}(1) \beta_t^{(q)}(1) + \alpha_t^{(q)}(1) a_{1N_q}^{(q)} \beta_t^{(q)}(1)} \quad (2.96)$$

$$a_{iN_q}^{r(q)} = \frac{\sum_{t=1}^T \alpha_t^q(i) a_{iN_q}^{(q)} \beta_t^{(q)}(N_q)}{\sum_{t=1}^T \alpha_t^{(q)}(i) \beta_t^{(q)}(i)} \quad (2.97)$$

$$a_{1N_q}^{(q)} = \frac{\sum_{t=1}^T \alpha_t^{(q)}(1) a_{iN_q}^{(q)} \beta_t^{(q+1)}(1)}{\sum_{t=1}^T \alpha_t^{(q)}(1) \beta_t^{(q)}(1)} + \alpha_t^{(q)}(1) a_{iN_q}^{(q)} \beta_t^{(q+1)}(1) \quad (2.98)$$

It can also be seen from examination of the last equation that the last model $q = Q$ in the state sequence cannot have a non-zero tee probability from the entry to exit state. This restriction is generally enforced for the initial model $q = 1$ as well, so that neither the beginning nor end of an utterance sequence can be a tee model.

The underlying Baum-Welch equations for estimating output distributions from Eq. (2.78) to (2.81) do not change once the modifications have been made to the forward and backward probabilities.

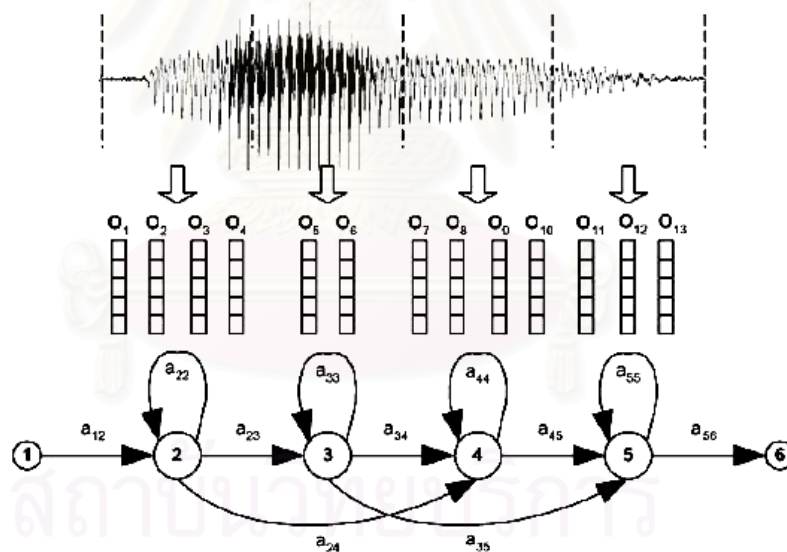


Figure 2.9 HMM with non-emitting entry and exit states

(From E. Maneeni, et al., 2003.)

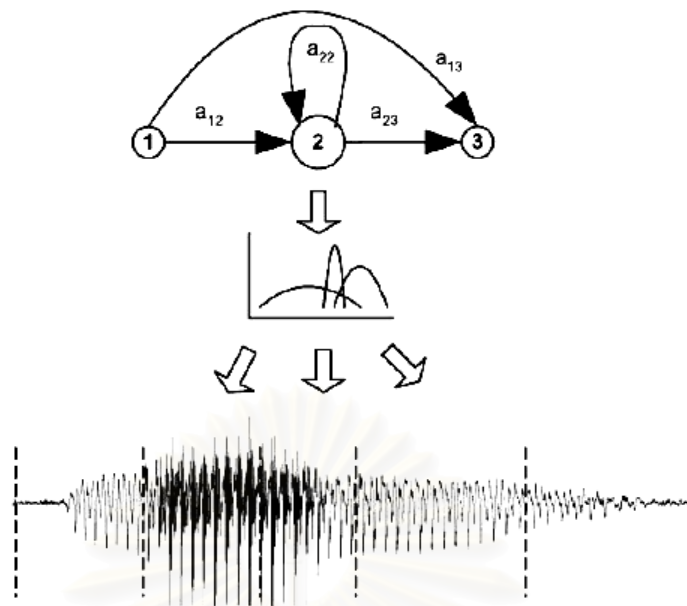


Figure 2.10 Tee model HMM (From Maneenoi E., et al., 2003)

2.3.5.2 Multiple Observation Sequence

In a complex large vocabulary speech recognition system, there may be literally thousands of models representing context-dependent sub-word units or segmental sub-word units. One problem that arises when performing training operation is that the Baum-Welch equations discussed so far are designed to be computed on one training sentence at a time, which is likely to use only a handful of different models just once or twice each, resulting in a very small quantity of training data for each iteration and corresponding poor re-estimation. A simple and accurate approach to solving is to treat the training sentences as a concatenated series of observation sequences assumed to be independent of each other (Maneenoi E., et al., 2003). This concept leads to updating the parameters for each model only one time over the entire training set, where the new parameters are given by continuously summing the numerator and denominator terms of the re-estimation equations throughout training. In the transition probability re-estimations, a $\frac{1}{P_r}$ term, where P_r is the $P(O | \lambda)$ for the r^{th} sentence, is added to the numerator and denominator. The full set of re-estimation equations for the Gaussian mixture distributions with multiple observation sequences, including entry and exit states and tee models, is given below

$$a'_{ij}{}^{(q)} = \frac{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_t^{(q)}(i) a_{ij}^{(q)} b_j^{(q)}(o_{t+1}) \beta_{t+1}^{(q)}(j)}{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_t^{(q)}(i) \beta_t^{(q)}(i)} \quad (2.99)$$

$$a'_{1j}{}^{(q)} = \frac{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(1) a_{1j}^{(q)} b_j^{(q)}(o_t) \beta_t^{(q)}(j)}{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(1) \beta_t^{(q)}(1) + \alpha_t^{(q)}(1) a_{1N_q}^{(q)} \beta_t^{(q)}(1)} \quad (2.100)$$

$$a'_{iN_q}{}^{(q)} = \frac{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(i) a_{iN_q}^{(q)} \beta_t^{(q)}(N_q)}{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(i) \beta_t^{(q)}(i)} \quad (2.101)$$

$$a'_{iN_q}{}^{(q)} = \frac{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(1) a_{1N_q}^{(q)} \beta_t^{(q+1)}(1)}{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_t^{(q)}(1) \beta_t^{(q)}(1)} + \alpha_t^{(q)}(1) a_{1N_q}^{(q)} \beta_t^{(q+1)}(1) \quad (2.102)$$

$$\gamma_t^{(q)}(j, m) = \left[\frac{\alpha_t^{(q)}(j) \beta_t^{(q)}(j)}{P_r} \right] \left[\frac{c_{jm} b_{jm}^{(q)}(o_t)}{b_j^{(q)}(o_t)} \right] \quad (2.103)$$

$$c'_{jm}{}^{(q)} = \frac{\sum_{r=1}^R \sum_{t=1}^T \gamma_t^{(q)}(j, m)}{\sum_{r=1}^R \sum_{t=1}^T \sum_{m=1}^M \gamma_t^{(q)}(j, m)} \quad (2.104)$$

$$\mu'_{jm}{}^{(q)} = \frac{\sum_{r=1}^R \sum_{t=1}^T \gamma_t^{(q)}(j, m) \cdot o_t}{\sum_{r=1}^R \sum_{t=1}^T \gamma_t^{(q)}(j, m)} \quad (2.105)$$

$$\sum'_{jm}{}^{(q)} = \frac{\sum_{r=1}^R \sum_{t=1}^T \gamma_t^{(q)}(j, m) \cdot (o_t - \mu_{jm}) (o_t - \mu_{jm})'}{\sum_{r=1}^R \sum_{t=1}^T \gamma_t^{(q)}(j, m)} \quad (2.106)$$

The implementation of these equations can be made with attention to some cancellations within the terms. In particular, the recursion for $\alpha_t^{(q)}(j)$ contains the term $b_j^{(q)}(o_t)$ within it, which is also in the denominator of the formula for $\gamma_t^{(q)}(j, m)$.

The variable $U_t^{(q)}(j)$ is defined as

$$U_t^{(q)}(j) = \begin{cases} \alpha_t^q(1)a_{1j}^q & \text{if } t=1 \\ \alpha_t^q(t)a_{1j}^q + \sum_{i=2}^{Nq-1} \alpha_t^q(t)a_{1Nq}^q \beta_t^q(1) & \text{Otherwise} \end{cases} \quad (2.107)$$

to represent $\alpha_t^{(q)}(j)$ without $b_j^{(q)}(o_t)$ term. The computation of this latter term is cancelled entirely, giving

$$\gamma_t^{(q)}(j, m) = \frac{1}{P_r} U_t^{(q)}(j) \beta_t^{(q)}(j) c_{jm} b_{jm}^{(q)}(o_t) \quad (2.108)$$

Similar modifications may be made to the distribution re-estimation equations for discrete probability densities so that composite models and multiple observation sequences can be considered, resulting in the equation

$$b'_j(o_t) = \frac{\sum_{r=1}^R \sum_{t=1}^T \gamma_t(j)}{\sum_{r=1}^R \sum_{t=1}^T \gamma_t(j)} \quad (2.109)$$

s.t. S emits o
j t

2.4 Large Vocabulary Continuous Speech Recognition

The performance of a speech recognition system depends on the system's ability to reduce uncertainty about the identity of a spoken word using information from the acoustic signal and past word sequences.

The speech recognition problem can be viewed as a problem in communication theory (Shannon, 1948). A spoken word of known identity w is viewed as passing through an acoustic channel model, which produces a sequence of acoustic observation symbols a (Valtchev, 1995). An acoustic observation a is a sequence feature vector extracted from the acoustic signal generated by the speaker while uttering w . The joint probability of words w and acoustics observation a is

$$P(w, a) = P(a | w)P(w) = P(w | a)P(a) \quad (2.110)$$

The language model component, $P(w)$, provides information about the word sequence in w . The conditional distribution $P(a|w)$ of acoustic given words describes the acoustic channel model, and the conditional distribution $P(w|a)$ defines a probabilistic decoder. For a known sequence of observations, the marginal distribution $P(a)$ is assumed to be constant since it does not depend on the model (Valtchev, 1995). The structure of speech recognition system, according to information transmission theory, is depicted in Figure 2.11.

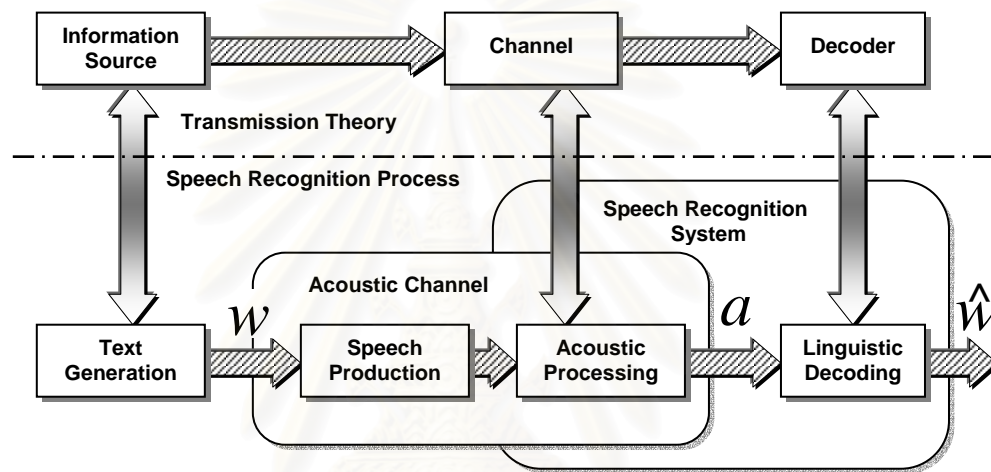


Figure 2.11 Structure of speech recognition according to information theory
(Adapted from Maneenoi E., et al., 2003)

The above definition of the speech recognition problem can be viewed as the following as practical considerations (Maneenoi E., et al., 2003):

Acoustic model structure – The acoustic model is a probabilistic function, which models the phonological and acoustic-phonetic variations in the speech signal. It is extremely difficult for a human expert to devise an accurate and complete acoustic model due to partial knowledge and inability such knowledge in an algorithmic form. For this reason, an acoustic model is defined as a family of parametric distributions with parameter λ . The chosen family of distributions should be based on true assumptions about speech and have a relatively small number of free

parameters. The value of λ identifies a unique acoustic model from the family and is usually estimated from a large sample of speech data.

Parameter estimation – The ultimate goal in parameter estimation is to find a parameter vector λ , which produces a decoder with the lowest possible recognition error rate. To achieve the lowest error rate, some objective function $F(\lambda)$, which relates to the decoder's performance, has to be optimized. The objective function should be such that when $F(\hat{\lambda}) > F(\lambda)$ then $\hat{\lambda}$ will produce a better decoder than λ . Once $F(\hat{\lambda})$ has been chosen, the second problem is to find the parameter set λ , which maximizes it. Complex acoustic models typically employ a large number of parameters, which makes it very unlikely that a globally optimal λ will be found. This means that even with a good function, it is possible to obtain unsatisfactory results if the estimation procedure converges to a bad local maximum.

Probabilistic decoder – A speech decoder is a device, which attempts to find the identity of a word from its acoustic representation. Since the chosen identity \hat{w} is different from the actual identity of the spoken word w then there is a decoding error. The probability of making an error is the most important factor in choosing the decoder. The optimal decoder with regard to minimizing the probability of error is the maximum a posteriori (MAP) decoder, where w is chosen such that

$$\hat{w} = \arg \max_w P(w | a) = \arg \max_w P(a | w) \frac{P(w)}{P(a)} \quad (2.111)$$

2.4.1 Search Algorithm

The two main schemes of decoding most commonly used today are Viterbi decoding using the beam search heuristic and stack decoding (Ravishankar, 1996; Steinbissa, et al., 1995; Robinson, 2002; Maneenoi E., et al., 2003).

Continuous speech recognition is normally performed as a time synchronous Viterbi search in a state space. The search produces the most likely word sequence by matching each frame from the unknown utterance to a network of HMM instances (Valtchev, 1995). The network is compiled corresponding to the grammar of the

language. The search itself is the computationally most expensive part of the recognition system due to the huge number of possible paths. This is a result of the vocabulary size and inherent acoustic ambiguities. In order to reduce the search space, it is customary to limit the scores generated by the acoustic models. Multi-pass recognition systems are another way of making the recognition task more manageable (Maneeno E., et al., 2003).

2.4.2 Language Modeling

The language model is a natural component in the information-theoretic formulation of the speech recognition problem. It is required in a large vocabulary speech recognition system for disambiguating between the large set of alternative confusable words that might be hypothesized during the search (Ravishankar, 1996; Maneeno E., et al., 2003). The language model defines the priori probability of a sequence of a word sequence W . The probability of a sentence, a sequence of words w_1, w_2, \dots, w_n , provided by the language model, is given by

$$\begin{aligned} P(W) &= P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2)P(w_4 | w_1, w_2, w_3) \dots P(w_n | w_1, \dots, w_{n-1}) \\ &= \prod_{i=1}^N P(w_i | w_1, \dots, w_{i-1}) \end{aligned} \quad (2.112)$$

where $P(w_i | w_1, \dots, w_{i-1})$ indicates the probability that the word w_i was spoken given that the word sequence w_1, w_2, \dots, w_i was said. It is practically impossible to obtain reliable estimations given arbitrarily long histories of all the words in a given language since that would require enormous amount of training data (Ravishankar, 1996; Loizou, 1995; Maneeno E., et al., 2003).

CHAPTER III

THE ANALYSIS OF LAO LANGUAGE

Lao (or 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 (Noulnavong O., et al., 2003). The languages in the Tai family all share a common grammar and tone structure, called “*Tonal Language*”. Lao language has many regional varieties in Laos. The main difference between these varieties is tonal, different varieties will have some changes in tone from the Vientiane Lao tone chart (Table 3.8). There are also some differences in vocabulary from region to region. However the Vientiane variety is becoming the unofficial national language. This can be seen in the capital where people from all over the country live. Many people there change their pronunciation or at least recognize that they speak a regional variety. The Vientiane variety is spoken on TV and radio and broadcasted over the whole country. However, Vientiane pronunciation is respected to be official national spoken language of Lao P.D.R. Therefore, this thesis will consider in Vientiane spoken language only (Vientiane is the capital of the Lao P.D.R.).

Since the syllable is principally considered a fundamental unit for acoustic-phonetic analysis, it is important to have a good understanding about Lao syllables. The basic Lao syllable sound consists of consonant sound, vowel sound and tone, where consonant is unvoiced sound and vowel is voiced sound, while tone is music sound and it will be represented by pitch contour over a syllable. Almost Lao spoken words are monosyllabic words, and perform several functions in a sentence. A polysyllabic word is constructed by concatenating each syllable. So, the several combinations of these syllables with tones can produce several words. In addition, a sentence is formed by a serial construction of these syllables.

Notice: Some phonetic symbol of Lao sounds below, have been modified to be convenient in terms of programming and used in this thesis only.

3.1 Lao Consonants

Table 3.1 Twenty-seven original Lao consonant and sounds

Letter	Sound	Example		
		Lao	Pronounce	English
ກ	/g/	ໄກ່	<i>gai0</i>	(chicken)
ຂ	/k/	ໄຂ່	<i>kai0</i>	(egg)
ຄ	/k/	ຄວາຍ	<i>kwai3</i>	(buffalo)
ງ	/ng/	ງົວ	<i>nguu3</i>	(cow)
ຈ	/j/	ຈອກ	<i>jawk1</i>	(glass)
ສ	/s/	ເສືອ	<i>squa4</i>	(tiger)
ຊ	/s/	ຊ້າງ	<i>saang2</i>	(elephant)
ຍ	/ny/	ຍຸງ	<i>nyung3</i>	(mosquito)
ດ	/d/	ເດັກ	<i>dek3</i>	(child)
ຕ	/t/	ຕາ	<i>taa4</i>	(eye)
ຖ	/th/	ຖົງ	<i>thong4</i>	(bag)
ທ	/th/	ທຸງ	<i>thung0</i>	(flag)
ນ	/n/	ນົກ	<i>nok0</i>	(brid)
ບ	/b/	ແບ້	<i>bxx2</i>	(goat)
ປ	/p/	ປາ	<i>paa4</i>	(fish)
ຜ	/ph/	ເຜີ້ງ	<i>phqaeng1</i>	(bee)
ຝ	/f/	ຝົນ	<i>fon4</i>	(rain)
ພ	/ph/	ພູ	<i>phuu3</i>	(mountain)
ຟ	/f/	ໄຟ	<i>fai4</i>	(fire)
ມ	/m/	ມ້າ	<i>maa2</i>	(horse)
ຢ	/y/	ຢາ	<i>yaa4</i>	(medicine)
ລ	/l/	ລິງ	<i>liing3</i>	(monkey)
ວ	/w/	ວີ	<i>wii3</i>	(hand-fan)
ຫ	/h/	ຫານ	<i>haan0</i>	(goose)
ອ	/z/	ໂອ	<i>zoo4</i>	(bowl)
ຮ	/h/	ເຮືອນ	<i>hquan3</i>	(house)
ຮ	/r/*	ຝຣັ່ງ	<i>frang0</i>	(France)

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

There are 27 original consonants and six consonant clusters realized in Lao alphabetical order, representing 21 original sounds (Table 3.1). These consonants are divided into three classes, 12 High Consonants, 8 Middle Consonants, and 12 Low Consonants, as shown in Table 3.2. 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 Lao syllable or a word.

Table 3.2 Six Consonant clusters and sounds

Letter	Sound	Example		
		Lao	Pronounce	English
ຫງ	/ng/	ເຫງ້ນ	<i>ngen4</i>	(civet cat)
ຫຍ (ຫນ)	/ny/	ໃຫຍ່	<i>nyai0</i>	(big)
ໜ (ຫມ)	/n/	ໝູ່	<i>nuu4</i>	(mouse)
ໝ (ຫລ)	/m/	ໝູ່	<i>muu0</i>	(pig)
ຫຼ	/l/	ຫລານ	<i>laan4</i>	(grandchild)
ຫວ	/w/	ຫວີ	<i>wii4</i>	(comb)

Table 3.3 Three Consonant classes, High, Middle, and Low Consonants

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/	ຫ	/th/
ຫ	/h/	ປ	/p/	ນ	/n/
ຫງ	/ng/	ຢ	/y/	ໝ	/Ph/
ຫຍ	/ny/	ອ	/z/	ຟ	/f/
ໜ	/n/			ມ	/m/
ໝ	/m/			ລ	/l/
ຫລ	/l/			ວ	/w/
ຫວ	/w/			ຮ	/h/

Normally, all Lao consonants can all be used at the beginning of a syllable, namely “*Initial Consonant*”. However, some of them can be used at the end of a syllable, called “*Final Consonant*”.

3.1.1 Initial Consonants

Actually, all Lao consonants are the initial consonants, that represented by 21 phonetic sounds, as illustrated in Table 3.1 and 3.2, which is included all, high, mid, and low consonant classes. The initial consonants can also be combination consonants such as, ກວ - /gw/, ຂວ - /kw/, ຄວ - /kw/, ງວ - /ngw/, and etc.

3.1.2 Final Consonants

There are eight basic final consonants sounds, five sonorant (unstop) final consonants and three stop final consonants. The difference of stop and sonorant final consonants is that, sonorant finals are voiced but stop finals are unvoiced. This distinction is important for determining the syllable’s tone. All the sonorant final consonants are low consonant, and all stop consonant are mid consonant as shown in Table 3.4.

Table 3.4 Sonorant and Stop final consonants

Sonorant		Stop	
Letter	sound	Letter	sound
ງ	/-ng/	ກ	/-k/
ນ	/-n/	ດ	/-t/
ມ	/-m/	ປ	/-p/
ຍ	/-y/		
ວ	/-w/		

Notes that, When, ກ, ດ, ປ and ຍ are initial consonants, they are transcribed as /g/, /d/, /b/ and /ny/, respectively. However, when they are final consonants, they are often transcribed as /-k/, /-t/, /-p/ and /-y/.

The waveform and spectrogram of initial consonant associated with a vowel (which a vowel commonly used to pronoun the consonant in Lao language), are example as illustrated in Figure 3.1. Let see in Appendix for more examples.

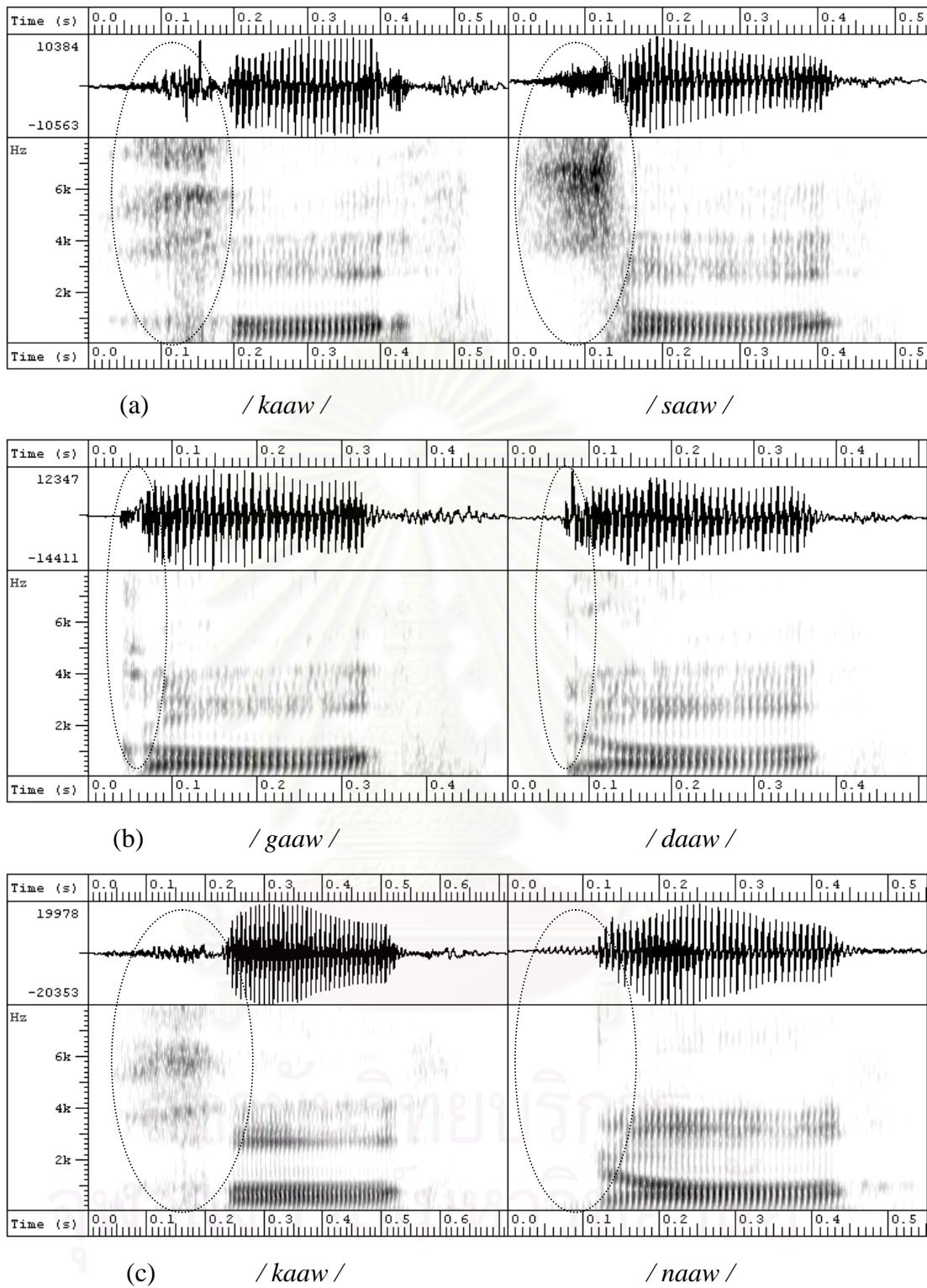


Figure 3.1 The waveform and spectrogram of example initial consonant associated with a vowel, (a) are that of high consonants, (b) are that of middle consonants and (c) are that of low consonants.

In addition, Figure 3.2 shows that, when the syllables are uttered in continuous speech, the feature of final and initial at the transitional syllables may look the same feature, especially for the final consonants, which are the nasal, such as /ng/, /n/ and /m/. The similarity of final (/n/) in a syllable and the initial (/m/) of a next syllable, is shown in Figure 3.2.

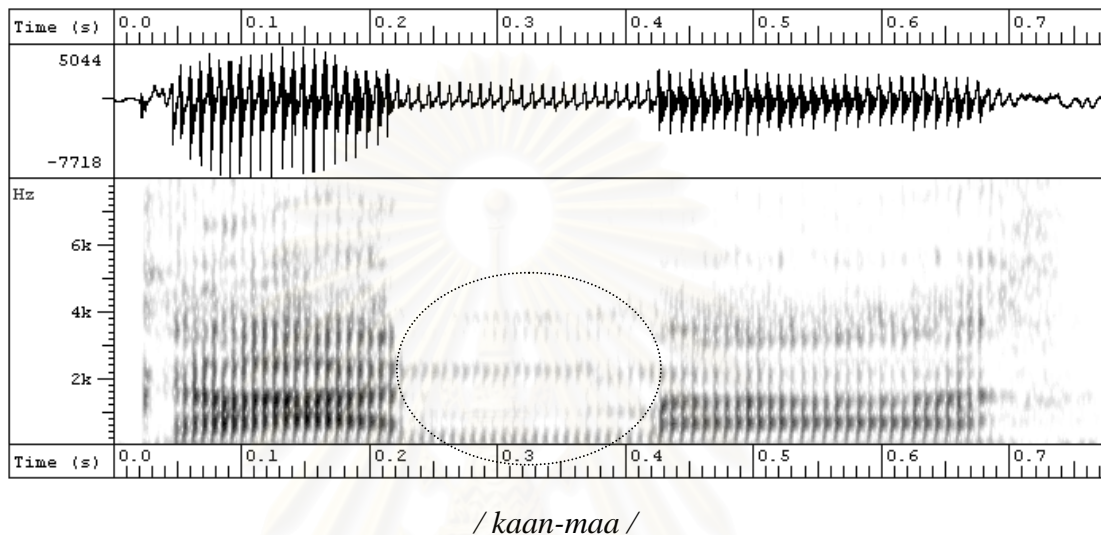
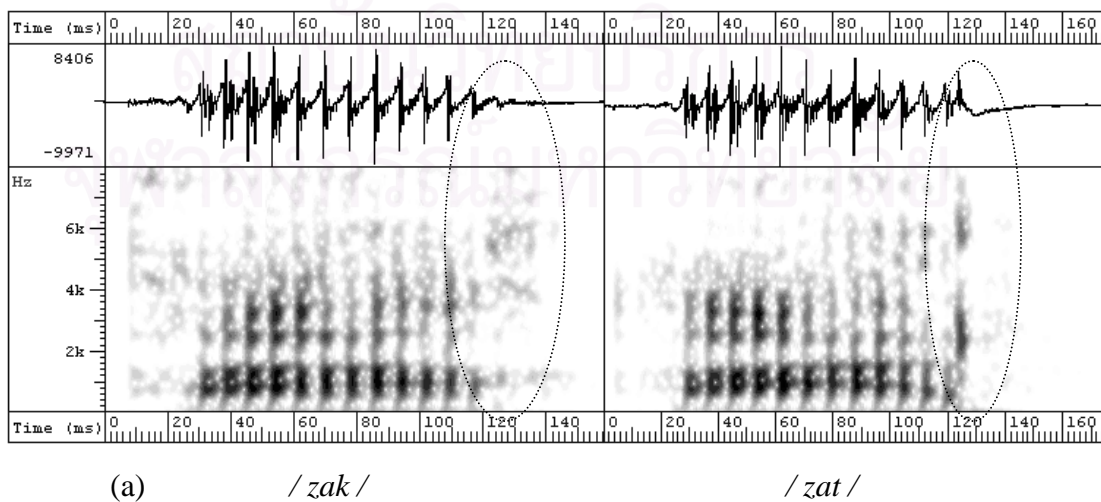


Figure 3.2 Example of the similarity between final and initial in transitional syllables

Since the Figure 3.3, (a) and (b) show the speech waveforms and spectrogram of the sample syllables, which are the syllables ending with stop final and non-stop final consonants, respectively. Let see in Appendix for more examples.



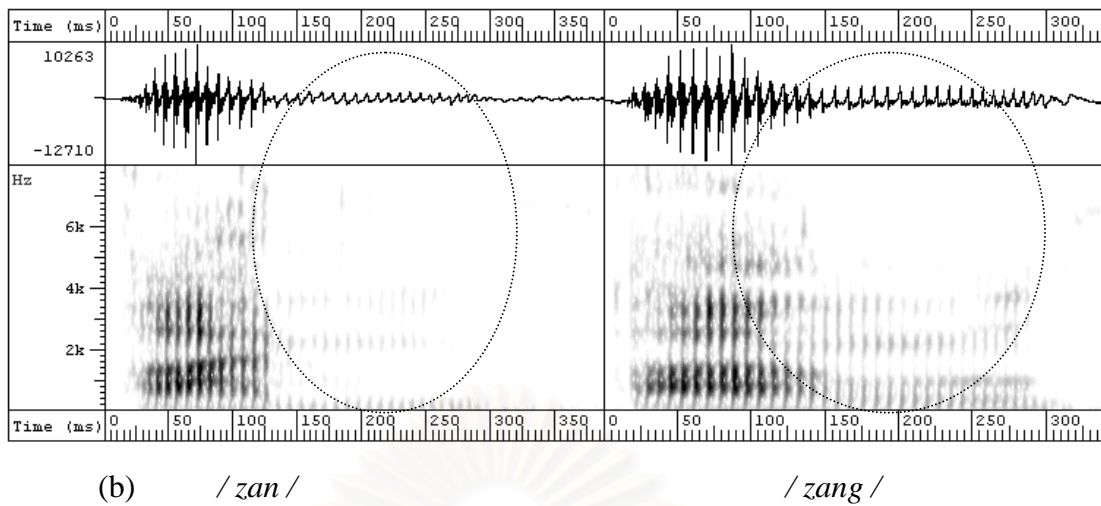


Figure 3.3 The waveforms and spectrogram of the syllables ending with stop and non-stop final consonants

3.2 Lao Vowels

Table 3.5 Vowel categories, monophthong, diphthong and special vowel

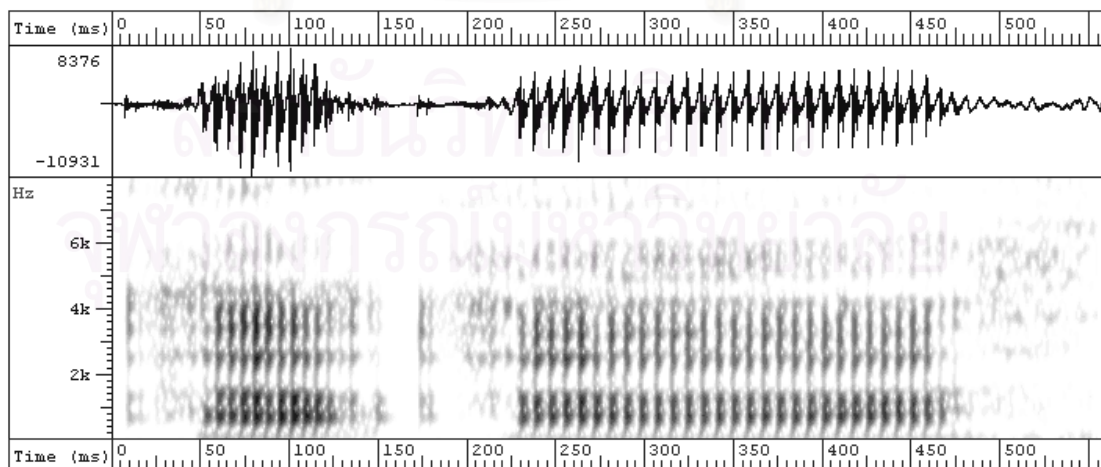
Monophthongs		Diphthongs		Special vowels	
Letter	sound	Letter	sound	Letter	sound
ຂ	/a/	ເຂ	/e/	ໄຂ	/ai/
ຯ	/aa/	ແຂ	/x/	ໃຂ	/ai/
ິ	/i/	ື່	/qa/	ໄຂ່	/ao/
ີ	/ii/	ື່	/qae/	ຂ່	/amh/
ື	/q/	ື່ອ	/qu/		
ື່	/qq/	ື່ອ	/qua/		
ຸ	/u/	ໄຂະ	/o/		
ຸ່	/uu/	ເຂະ	/aw/		
ເຂ	/ee/	ຂ່ະ	/ua/		
ແຂ	/xx/	ຂ່	/uua/		
ໄຂ	/oo/	ຂ່ຍ	/ia/		
ິ່	/aaw/	ເຂຍ	/ia/		

The Lao language has a complex vowel system. It is consisted a total of 28 vowels, representing 27 original sounds, where 12 vowels are monophthong, 12 vowels are diphthong, and other 4 vowels are special vowel, as shown in Table 3.5.

Lao vowels can be divided into two groups as short (short vowel) and long sounds (long vowel), as shown in Table 3.6. Note that, the special vowels are not defined to both short and long vowels. It may be sound either short or long. Since 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.

Table 3.6 Short vowel and Long vowel

Short vowels		Long vowel	
Letter	sound	Letter	sound
⺊	/a/	⺊	/aa/
⺋	/i/	⺋	/ii/
⺌	/q/	⺌	/qq/
⺍	/w/	⺍	/uu/
⺎	/e/	⺎	/ee/
⺏	/x/	⺏	/xx/
⺐	/qa/	⺐	/qae/
⺑	/qu/	⺑	/qua/
⺒	/o/	⺒	/oo/
⺓	/aw/	⺓	/aaw/
⺔	/ua/	⺔	/uua/
⺕	/ia/	⺕	/iia/



(a) / a /

/ aa /

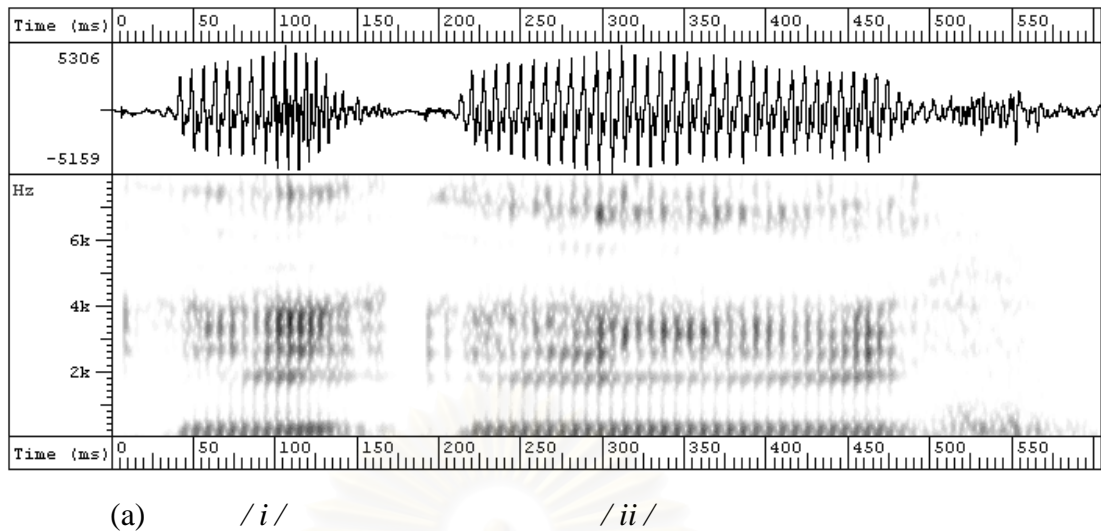


Figure 3.4 The differential of short and long vowel categories

In Figure 3.4 (a) and (b), the differential of short and long vowel categories was represented by time duration in syllable. The spectrogram of long vowel syllable is similar to that of short vowel syllable. Since the duration of long vowel syllable is appeared longer than that of short vowel syllable.

3.3 Lao Tones

As described above, Lao 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, such as, Thai, Chinese, Japanese, Burmese, Vietnamese, Lao, and also some European, African, etc. 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 a word itself.

In Lao language, tone is an integral component of a syllable, where tone information is an essential lexical meaning of Lao utterance. Tones of Lao words are determined by the tone chart (Table 3.7). 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, categorized as dynamic tones and static tones (Table 3.8), there are represented Lao tones in Lao writing system. However,

there are more than 4 tone sounds in Lao pronunciation. Since old Lao language system presented 5 tone sounds to pronounce. In recent years, advance research find out, there are perhaps more than 5 tones in Lao spoken language, it depends on the region pronunciation, for example: there are five tones of Luangphapang pronunciation (Northern of Laos), six tones of Pakse pronunciation (Southern of Laos), and it's very confusing of Vientiane tone categories (Middle of Laos), because, 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, Vientiane tone is only one that has to be studied in this thesis.

Table 3.7 Lao tone mark

Category	Tone Mark	Name
Dynamic	x	<i>mai2-zeek1</i>
	x̃	<i>mai2-thoo3</i>
Static	X̃	<i>mai2-dii1</i>
	x̄	<i>mai2-jat4da0vaa3</i>

Table 3.8 Tone chart of Lao spoken in Vientiane

Syllable Initial consonant class	Live Syllable*			Dead Syllable**	
	Inherent Tone	[˩] (low tone mark)	[˨˨] (falling tone mark)	Long Vowel	Short Vowel
High Class /ຝ/ສ/ຜ/ຖ/ຂ/ຫ/ຫງ/ ຫຍ/ໜ/ໝ/ຫລ/ຫວ/	Low Rising (4)	Mid (0)	Low Falling(1)		Low Rising (4)
Middle Class /ອ/ບ/ດ/ຢ/ປ/ຕ/ກ/ຈ/	Low Rising (4) (or Low Falling)		High Falling (2)		
Low Class /ພ/ຮ/ລ/ຊ/ພ/ຫ/ຣ/ນ/ ງ/ມ/ວ/ຍ/	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.

From history, five Vientiane tone chart has presented by Brown, 1965; Reinhorn, 1970-1971; Strecker, 1980 (unpublished); Chittavoravong 1980 (unpublished); Enfield, 2000 and Crisfield-Hartmann, 2002. Since, Hoshino, Marcus 1973; and Levy 1980 have presented six tones of Vientiane pronunciation. However, this thesis respects to apply as tone chart of Crisfield-Hartmann and Enfield, as illustrated on the Table 3.8.

3.4 Lao Syllable Structure

As above description, Lao syllables are composed of three sound systems, namely consonants, vowels, and tones. The smallest construction of sounds or a syllable in Lao is composed of one monophthong unit or one diphthong, one, two, or three consonants, and a tone (Paphaphan B., et al. 2000). A Lao syllable can be formed as illustrated in Figure 3.5.

$$S = C_i(C_s)V(C_f)T$$

Figure 3.5 The general Lao syllable structure

Where, C_i is an initial consonant, V is a vowel, C_f is a final consonant, T is a tone, and C_s is additional consonants (/w/, /l/ or /r/).

Examples:	ໄປ	/pai4/	(go)
	ກິນເຂົ້າ	/gin4-kao2/	(have meal)
	ຄວາມຄິດ	/kwaam3-kit0/	(idea)
	ກວ້າງ	/gwaang2/	(wide)

The syllable is principally considered a primitive unit for analysis with several reasons. First, the language model originates from this unit. A syllable is composed of sounds, which depends upon the phonological rules of each language. Second, the syllable is an acoustic unit, which is closely connected with human speech perception and articulation. Especially in connected speech, three linguistic factors, stress, tone,

and intonation, are influential in an utterance. The syllable integrates some co-articulation phenomena and represents conversational speech compactly. Therefore, using the syllable as the primitive unit is appropriate and has benefits for prosodic study. Furthermore, a syllable embraces both spectral and temporal dependencies due to its size, which makes the syllable a more stable acoustic unit. The syllable is seemingly good for modeling as an acoustic unit.

When, a tone is a feature of pitch or fundamental frequency movement within a syllable. The Figure 3.6, shows the average of pitch contours, extracted from male and female voiced (Vientiane speaker) of a syllable, which has different tones.

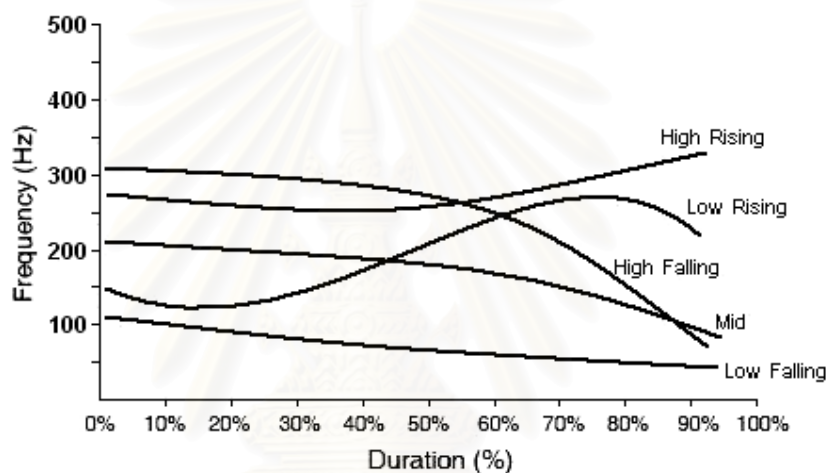
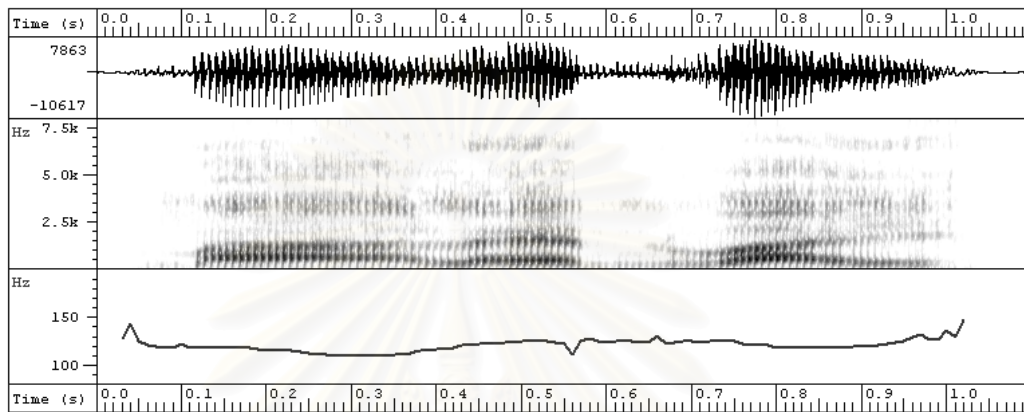


Figure 3.6 Average pitch contours over syllables of Vientiane speaker, which are represented five tones

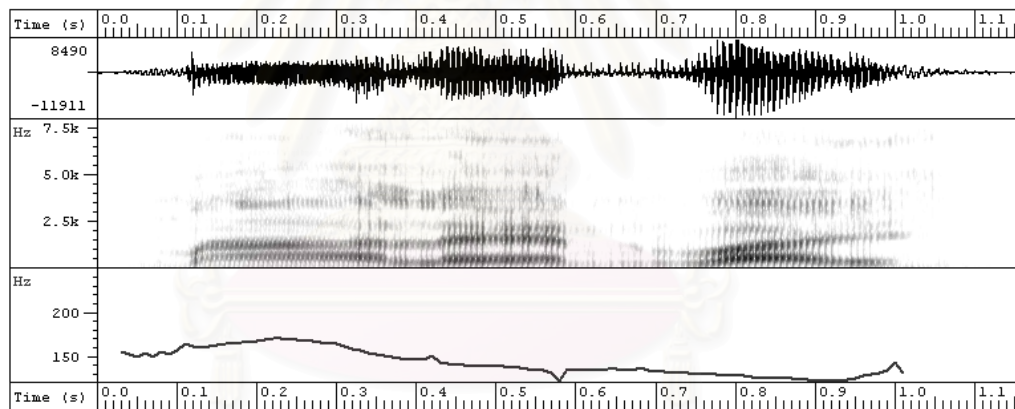
3.5 Lao Sentence

Lao language has more complex system. Unlike Western language, Lao sentence can be performed in several functions with the same meaning. A Lao sentence may compose of a syllable of several syllables, since some time these syllables can be noun, or verb, or etc; furthermore, Lao sentence can be a single sentence or a combination sentence. However, the variety of Lao sentence can be classified into two types, general sentence and special sentence (Paphaphan B., et al. 2000), the different of both is, the general sentence is form as full sentence, but special sentence is form as broken sentence, which is widely used in spoken language. In addition, the meaning of Lao sentence can vary with different tones information, while the tone information is the continuous pitch contour of syllabic components in

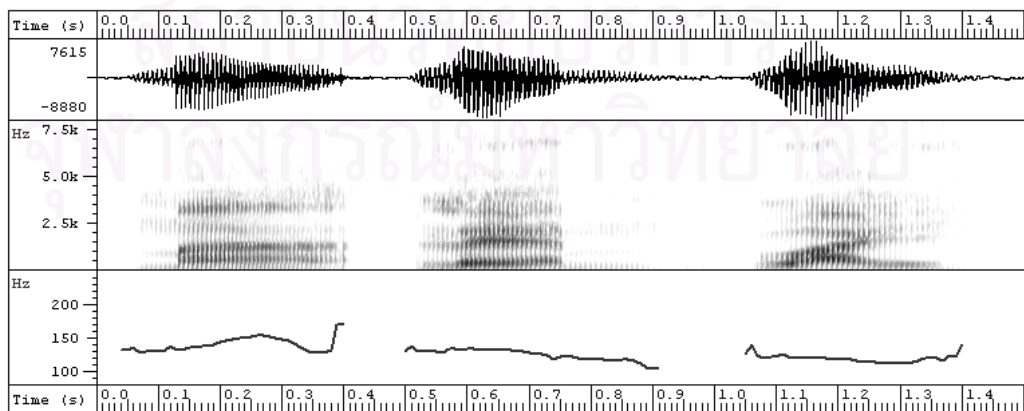
that sentence, as illustrated in Figure 3.7. In Figure 3.7 (a) and (b), actually, they are different syllable meaning, which is the same a phonetic sentence pronouncing with different tones. Furthermore, the pitch contour of each syllabic component in a continuous sentence may difference, when those syllable have pronounced in individual syllable (see Figure 3.7, (b) and (c)).



(a) /maa4-lxxn0-wai3/ (“dog run fast”)



(b) /maa2-lxxn0-wai3/ (“horse run fast”)



(c) /maa2 / - / lxxn0 / - / wai3 /

Figure 3.7 the pitch contour of Lao sentence

CHAPTER IV

METHODOLOGY

The research procedure will be explained in this chapter. As first, the detail of data collection method for training and testing will be described. Then, the proposed tonal syllable recognition system for Lao continuous speech, such as feature extractions and HMM model construction will be explained. The training and testing methodology will also be explained in the last part of this chapter.

4.1 Data Collection

As the purpose of this thesis, a continuous speech recognition system for Lao language will be implemented. All the sample data are recorded from Lao speakers. The reading in Vientiane sound was selected for less significant coarticulatory effects and pronunciation variations of Lao language. All the selected sentences are twice recorded per each speaker in order to study the system generalization for speaker-dependent and speaker-independent. Since, the recording configurations are detailed as below:

1. The Speakers are both male and female, from age around 18-25 year old.
2. All the speakers are familiar in Vientiane pronunciation sound.
3. The sample sentences are described as the stories.
4. Recording the sentences by reading style, in quiet office environment.
5. Sample speech data have been recorded with mono-channel, at 16 kHz sampling rate and 16 bits quantizing resolution.

In addition, to create an initial acoustic and tone models, a number of training samples must be sufficient. In this task, eighty Lao sentences were recorded from 50 speakers (30 males and 20 females), which each sentence is lengthened around 5-14 second. There are 696 syllables in 27 sentences for the total sample data. Since, the sample data of 20 males and 15 females are observed for training set. To evaluate the speech recognition system, the sample data of other 10 males and 5 females are used for speaker-independent testing set. For the speaker-dependent testing set, the sample data is obtained from different recording of training set.

4.2 System Procedure

Tonal syllable recognition system is required and necessary for speech recognition system of tonal language. Several techniques of tonal syllable recognition system has been presented, such as the research of Wang H.M., et al., 1994, Chen C.J, et al., 1997 and Demeechai T., et al., 2001. Also, this thesis has attentively represented a tonal syllable recognition system, associated with the specific characteristic of a tonal language. As the explanation in Figure 4.1, the proposed system will be individually recognized base syllable and tone recognitions. The sample speech signal is initialized using the signal preprocessing algorithms of speech recognition. Then, the set of feature vectors for base syllable recognition can be obtained by using the corresponding feature extraction algorithms. Phonetic feature vectors are putted into base syllables recognizer. In this step, the syllables will be recognized and form as a sentence, based on HMM of base syllable for continuous speech recognition. After base syllable recognition is finished, system will match a context sentence (the result of base syllable recognition) with the list of any available sentences. Where, the list of sentences is contented only the available sentences. Which, those sentences can be changed the meaning with different tones.

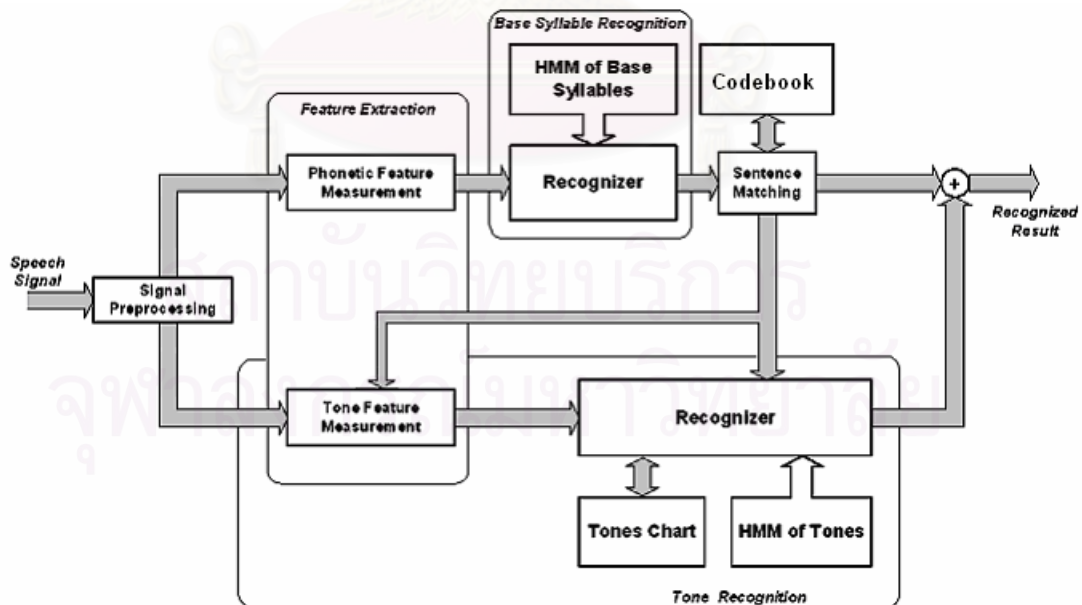


Figure 4.1 Tonal syllable recognition System for Lao language

As the result, if the sentence matching is return found. Immediately, the subsystem of tone recognition is started the process. Tone feature extractor was early applied to obtain the corresponding feature vectors for tone recognition. Since, the tone recognizer is recognized, based on HMM of tone recognition, The tone recognizer will be processed associated with syllable boundaries and time derivative of each syllabic component, which are the result of base syllable recognition then, tone recognizer is classified to the most promising tones, based on available tones chart. Finally, the results of both subsystems are combined together as the tonal syllable recognition.

In addition, all both the HMM of base syllables and tones recognitions were trained as following steps in Figure 4.2. Moreover, more explanation of training step can be found in the HTK manual book (Young S., et al., 2002).

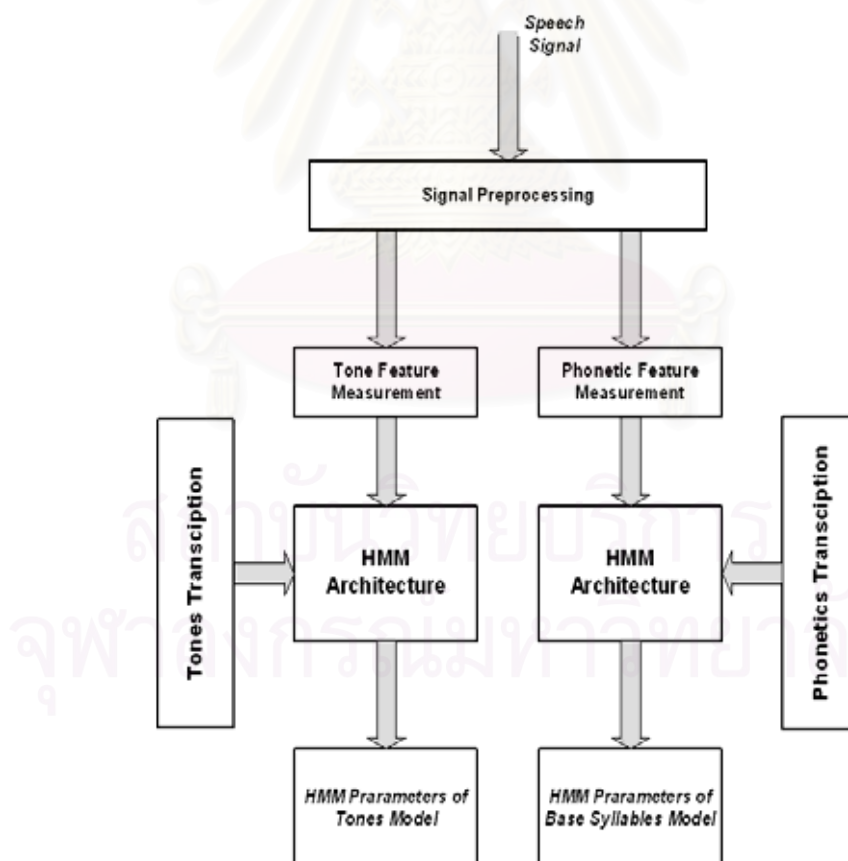


Figure 4.2 The training step for HMM of tones and HMM of base syllables

4.2.1 Feature Extraction

The feature extraction influences greatly the recognition rate so, it is vital for any recognition/classification systems. Feature extraction is to convert an observed speech signal (speech waveform) to some type of parametric representation for further analysis and processing. Since, stationary spectral features have been used in speech recognition systems for many years, such as *Linear Prediction Analysis (LPA)*, *Filterbank analysis*, *Energy measurement*, *Delta coefficients*, *Fundamental frequency*, and etc. The spectral estimate assumes a stationary signal, only a small amount of data is used for each estimate. The amount of data is usually referred to as the window length. In this context, a good feature is obtained with 30 ms frame length, moving every 10 ms. In the experiment, this thesis has also preformed with individual of LPCC and MFCC feature vectors, associated with Energy and delta coefficients, to observe the optimal feature for Lao speech recognition. Since, the feature vector will be normalized during recognition.

4.2.1.1 Speech Preprocessing

Initially, the speech waveform is put through a low-order transfer function, to spectrally flatten and to make it less susceptible to finite precision effects later in the signal processing. Typically, speech preprocessing of speech recognition are executed following:

- 1) **Preemphasis:** As described in section 2.2.2, the speech waveform is smoothing by using first-order FIR filter transfer function (Eq. (2.6).). Since, the preemphasis factor (α) is recommended at 0.97 (Ling F., et al., 2004). Comparison of the Preemphasized speech waveform and original speech waveform are indicated in Figure 4.3.
- 2) **Frame Blocking:** The preemphasized 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. When l^{th} is frame index of speech by $x_l(n)$, and L is the total number of frame, then

$$x_l(n) = \tilde{s}(Ml + n); \quad n = 1, 2, \dots, N; \quad l = 0, 1, 2, \dots, L-1 \quad (4.1)$$

The framework of this thesis will be performed with frame of 30 ms, and overlap of $\frac{1}{3}$ frame duration, when the sampling rate of the speech data is 16 kHz, the corresponding value of N and M are respectively, 480 and 160 samples, which are obtained the optimal performance of Lao speech recognition in the experimentation.

- 3) **Windowing:** This step is applied the hamming window function (Eq. (2.4).) into 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 in subsection 2.2.1. when $w(n)$ is defined as the window function, then the result of windowing is

$$\tilde{x}_l(n) = x_l(n)w(n); \quad n = 1, 2, \dots, N \quad (4.2)$$

The Figure 4.4 is shown the output signal of windowing process, with the frame blocking, $N = 30$ ms.

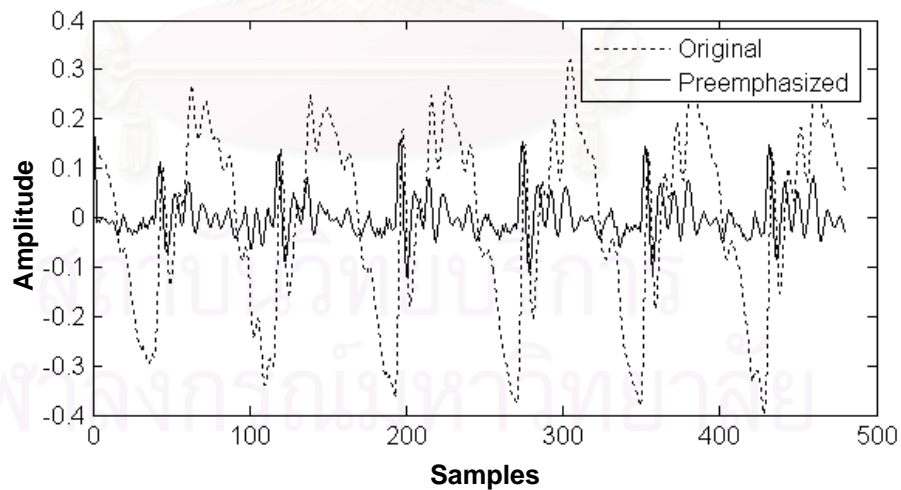


Figure 4.3 Example of Preemphasized speech waveform

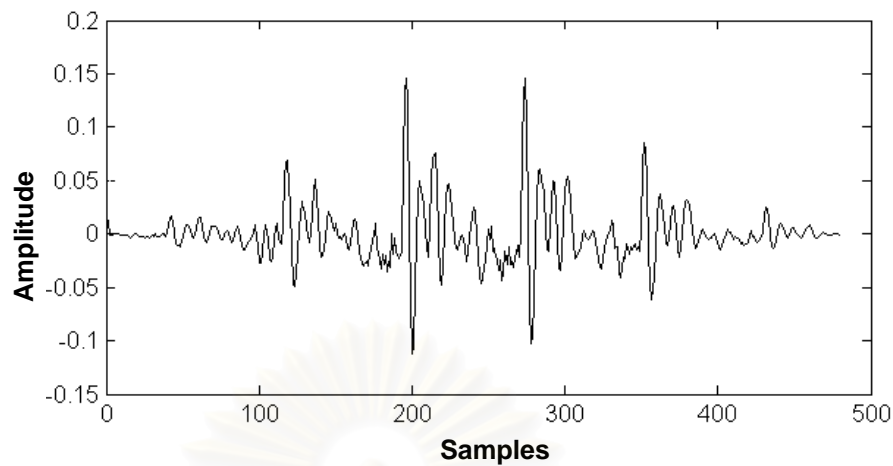


Figure 4.4 Example signal of windowing processed

4.2.1.2 LPCC Measurement

Linear Prediction Coefficients (LPC) can parameterize the speech spectrum quite well. LPC assumes an all-pole speech production model, as shown in Eq. (4.4). In this equation, $X(z)$ is the spectrum of the speech signal and $G(z)$ is the spectrum of the glottal excitation, which is assumed to be white. $1/A(z)$ is the spectrum of the vocal tract, where $A(z)$ is modeled as a polynomial function of z .

$$X(z) = G(z) \frac{1}{A(z)} \quad (4.3)$$

$$= G(z) \frac{1}{1 - a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}} \quad (4.4)$$

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k} \quad (4.5)$$

The LPC coefficients, $\{a_1, a_2, \dots, a_p\}$, are estimated from the current frame of data, given the speech production model, where p is the order of the LPC coefficients.

In LPC extraction, The filter coefficients, a_k , are chosen to minimize the mean square filter prediction error summed e_t (Eq. (2.8), (2.9).), over the window analysis. Since, the autocorrelation method is a common technique to obtain the coefficients of LPC filter. So, it has been also considered for this task.

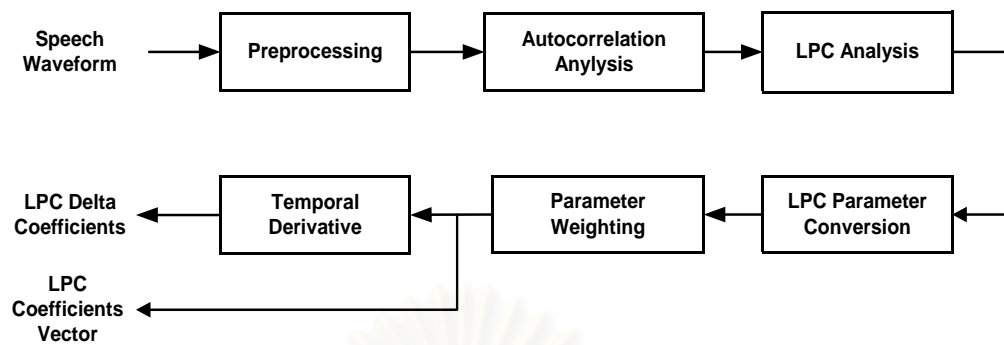


Figure 4.5 The block diagram of LPCC processor

The basic steps in the LPCC processing include the following:

1. **Preprocessing (or speech preprocessing):** This step is a common step which is always used in several techniques of speech processing. The proceeding method of the speech preprocessing has been already described in above section. What is the initialization before apply the LPC feature measurement technique.
2. **Autocorrelation Analysis:** Each frame of windowed signal is next autocorrelated to give

$$R_l(\tau) = \sum_{n=1}^{N-\tau} \tilde{x}_l(n)\tilde{x}_l(n-\tau), \quad \tau = 0,1,2,\dots, p-1 \quad (4.6)$$

where, $\tilde{x}_l(n)$ is the windowed signal. p is the order of LPC analysis. Typically, values of p from 8 to 16 have been used (Rabiner, and Juang, 1993).

3. **LPC Analysis:** This step converts each frame of p autocorrelation into an LPC parameter set (LPC coefficients). As the concept of LPC analysis

$$\sum_k R_l(\tau-k)a_k = R_l(\tau); \quad k = 1,2,\dots, p \quad (4.7)$$

where, a_k ; $1 \leq k \leq p$, are the LPC coefficients, and can be given by using Durbin's method (Rabiner, and Juang, 1993). When, E is the prediction error,

k_i are auxiliary coefficients and α_m are the filter coefficients. Then a filter of order τ can be calculated as following steps:

$$E_l^{(0)} = R_l(0) \quad (4.8)$$

$$k_\tau = \frac{R_l(\tau) - \sum_{j=1}^{\tau-1} \alpha_j^{(\tau-1)} R_l(|\tau - j|)}{E^{(\tau-1)}}; \quad 1 \leq \tau \leq p \quad (4.9)$$

$$\alpha_\tau^{(\tau)} = k_\tau \quad (4.10)$$

$$\alpha_j^{(\tau)} = \alpha_j^{(\tau-1)} - k_\tau \alpha_{\tau-j}^{(\tau-1)} \quad (4.11)$$

$$E^{(\tau)} = (1 - k_\tau^2) E^{(\tau-1)} \quad (4.12)$$

For $\tau = 1, 2, 3, \dots, p$, the final solution of Eq. (4.8)-(4.12) will be given as

$$a_m = \alpha_m^{(p)}; \quad 1 \leq m \leq p \quad (4.13)$$

4. **LPC Parameter Conversion:** it's very importance that LPC parameter set is the LPC cepstral coefficients. The principal advantage of cepstral coefficients is that they are generally decorrelated and this allows diagonal covariances to be used in the HMMs. In the case of linear prediction cepstral coefficients (LPCC), it can be obtained from the Fourier transform representation of the log magnitude spectrum. However, it can be shown that the required cepstral can be more efficiently computed using a simple recursion (Eq. (2.12) and (2.13)).
5. **Parameter Weighting:** This technique will be applied to weight the spectral coefficients by a taper window so as to minimize the sensitivity of the low-order cepstral coefficients to overall spectral slope, and the sensitivity of the high-order cepstral coefficients to noise. The process of weighting or windowing the cepstral coefficients, is also known as cepstral liftering (Eq. (2.14)). Since, the new LPC cepstral coefficients can be given by Eq. (2.15).
6. **Temporal Cepstral Derivative:** This step is applied to obtain LPC delta coefficients ($\Delta \tilde{c}_m(t)$), which it provides a good representation of the local

spectral properties of the signal for the given analysis frame. For the computation detail of cepstral derivative can be found in subsection 2.2.7.

4.2.1.3 MFCC Measurement

The mel-frequency cepstral Coefficients (MFCC) are the best well known and most commonly used features for speech recognition system. The computation of MFCC is based on the short-time analysis and similar to that of Cepstral Coefficients, that have described in subsection 2.2.4. The significant difference lays on the usage of critical bank filters to realize mel-frequency warping. The critical bandwidths with frequency are based on the human ears perception. The block diagram for computing MFCC is given as Figure 4.6.

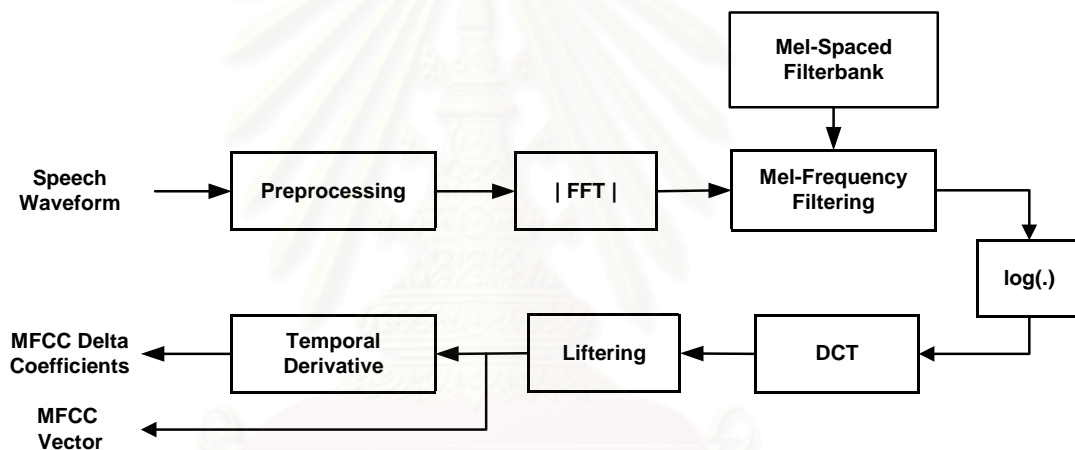


Figure 4.6 The block daiagram of MFCC feature extraction

The computation of MFCC are commonly followed as

1. **Preprocessing:** Sinmilar to the LPCC process, signal preprocessing is applied as the initail of MFCC procedure. The processing of this step can be implemented as following subsection 4.2.1.1.
2. **Fast Fourier Transform analysis (FFT):** To implement the filterbank, each frame of windowed signal is transformed using a Fourier transform to obtain the magnitude cofficients. The FFT magnitude is exemplad as illustrated in the upper-rigth of Figure 4.7.
3. **Mel-Frequency Filtering:** After The FFT was applied, each FFT magnitude spectrum coefficients is multiplied by the corresponding gain of mel-spaced filterbank (Eq. 2.10). and the results accumulated. Thus, each bin holds a

weighted sum representing the spectral magnitude in that filterbank channel (see lower-left of Figure 4.7.). The cepstral parameters are computed from the log filterbank magnitude using the Discrete Cosine Transform (DCT) as shown in Eq. (2.11), then, the mel-frequency coefficients are obtained (see lower-right of Figure 4.7.). The detail of mel-frequency of filterbank analysis has been reported in subsection of chapter 2 (signal processing of speech recognition).

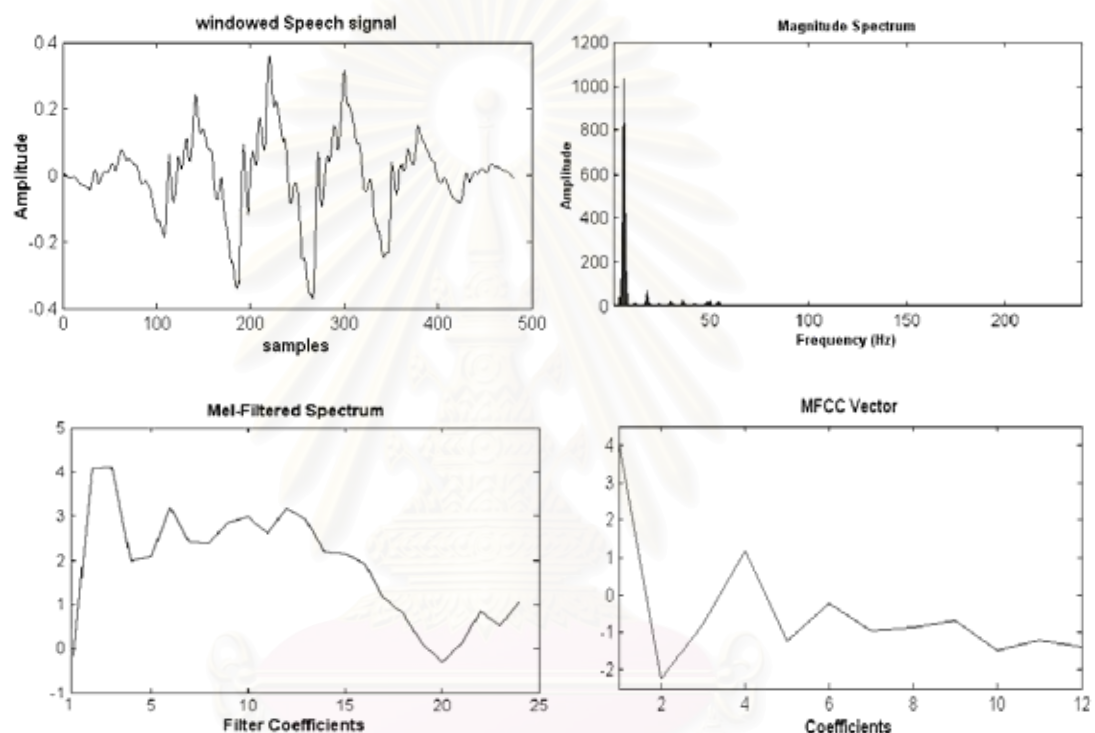


Figure 4.7 Example of 12 MFCC coefficients extracted

4. **Liftering:** As described in section 2.2.6, the liftering or less emphasis is also applied to MFCC coefficients, to obtain robust features for speaker-independent speech recognition. From Eq. (2.14) and (2.15), the new MFCCs is given by following

$$\tilde{c}(i) = \left(1 + \frac{Q}{2} \sin\left(\frac{i\pi}{Q}\right) \right) c(i) \quad (4.14)$$

5. **Temporal Derivative:** it's similar to the PLCC feature extraction, this step is also given the MFCC delta coefficients ($\Delta\tilde{c}_m(t)$), which can be improved the performance of a speech recognition system by adding to the basic static

parameters. For the computation detail of temporal derivative can be found in subsection 2.2.7.

4.2.1.4 Energy

In speech recognition, energy is also an important feature, which is well known to represent voiced and unvoiced sounds portion. An absolute energy can be directly computed from a speech waveform using Eq. (2.19). The absolute energy contour along the speech waveform duration is shown in Figure 4.8.

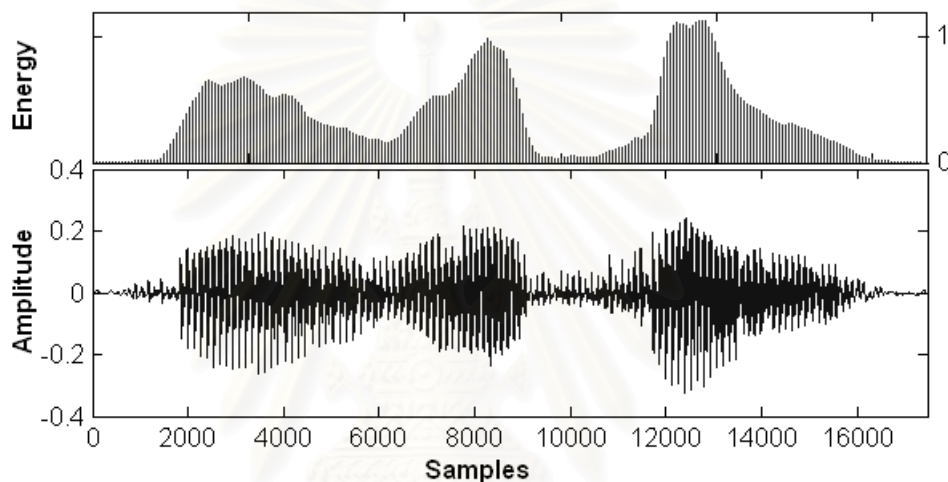


Figure 4.8 Energy contour of sample speech waveform

4.2.1.5 Fundamental Frequency

As introduced in subsection 2.2.8, the vibration frequency of vocal cords is defined as *fundamental frequency* (F_0), an important feature for automatic speech and speaker recognition. Fundamental frequency, or pitch period is robust to noise and channel distortions. Also, the different pitch contour of voice sounds is well known to represent the different tones in Lao language (Kanthvisone K., et al., 2001). However, speech recognition system based on pitch information works well with small number of speakers, but error rate increases significantly with the increasing number of speaker increases. Therefore, a speaker-independent speech recognition system has been done combining pitch information with other features (Ling F., et al., 2004).

There are several existing techniques for fundamental frequency extraction. However, autocorrelation method is applied as the analysis portion in this task, which

short-time autocorrelation function is a simple computation technique, and the autocorrelation method is well done for noise environment (Ling F., et al., 2004). In addition, the accurate fundamental frequency will be improved by using center chipping technique, combining with autocorrelation function (Kanthavisone k., et al., 2001).

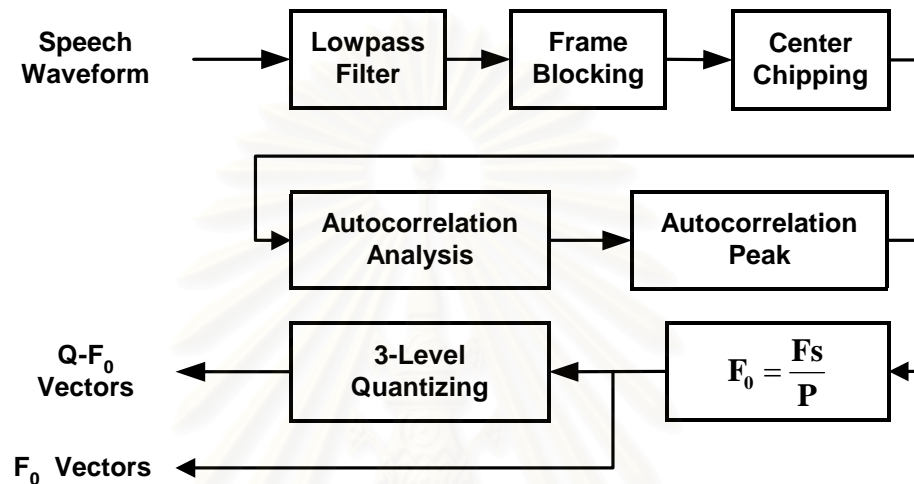


Figure 4.9 The block diagram of tone feature extraction

In this thesis, tone features extraction is performed following as the block diagram in Figure 4.9. Since, the fundamental frequency is lower frequency of human utterance. To minimize the effects from other formants of vocal excitation, lowpass filter is applied with 900 Hz cutoff frequency at initiation. As the result, 3-level quantization technique (Kanthavisone k., et al., 2001) was applied to minimize the different level effect of male and female characteristic.

4.2.2 Hidden Markov Model Architectures

Hidden Markov models (HMMs) are a probabilistic tool, which are popular in speech recognition systems because they are simple enough to implement in a real time system, and also complex enough to capture the basic non-stationary structure of speech. Because their behavior can be described with simple formulas, the full power of mathematics and probability theory can be brought to solve on the speech

recognition problem. The definitive tutorial on the basic HMM formulations can be seen in section 2.3.

In this thesis, five states left-to-right Hidden Markov Model architecture is investigated using for tone recognition. Since, the HMM architecture of base syllable recognition is determined corresponding to the specific characteristic of speech model as has been explained detail in next chapter.

The conventional, five states left-to-right HMM is exemplified as show by the Figure 4.10.

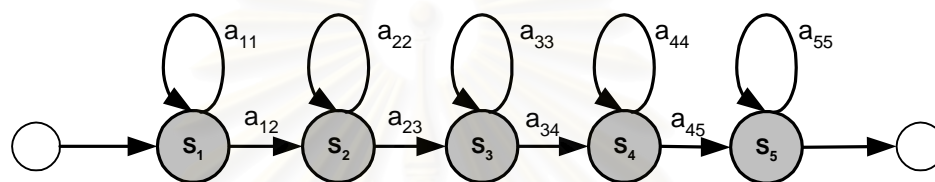


Figure 4.10 Conventional left-to-right Hidden Markov Model

The five states, S_1 , S_2 , S_3 , S_4 , and S_5 , correspond to states that the model can take on. The model takes on the properties of a stationary stochastic process. While, each state is described by its probability distribution function (PDF). This PDF can be modeled by one of several ways.

The two small circles at the beginning and ending of the graph, are represented respectively the entrance and exit states of the model. Since, the model does not produce any output at a time in these states. All of the arrows represent allowed transitions between states, which are the transition probabilities. The summation of transition probabilities in a state is equal to 1. At each time increment, the model can follow only one of the allowed transitions.

Individual HMM models are united in a larger HMM structure, for continuous speech recognition, as illustrated in Figure 4.11, with transitions between the individual HMM models provided by the language model (see section 2.4.). Then, the Viterbi algorithm is used to search the single state sequence path with the highest probability.

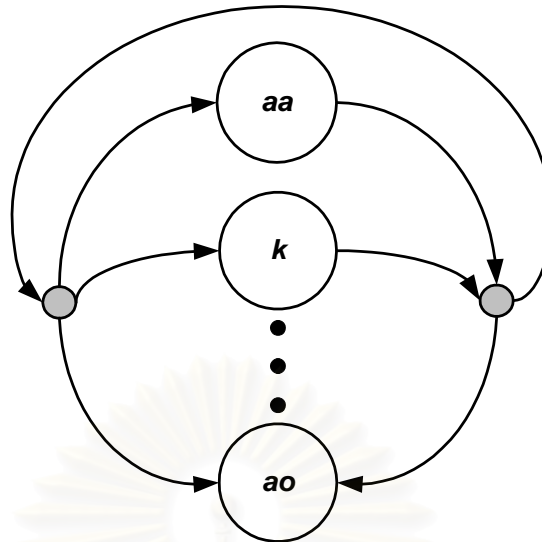


Figure 4.11 Transition production model

Figure 4.11 is presented a large HMM structure associated with transition production model. The large circles are represented individual HMM of each symbol model. The detail of HMM for continuous speech recognition can be found in subsection 2.3.4.

4.2.3 Speech modeling

As the previous chapter, the acoustic-phonetic analysis of Lao language was reported. The characteristics of vowels and consonants were thoroughly explored. Not only does the acoustic-phonetic analysis contribute strong knowledge, but it also provides well knowledge for modeling the appropriate speech unit for Lao language. Various speech models have been presented for speech recognition with very large vocabularies such as, phonetic, syllable, subword, initial-final and onset-rhyme, (Lee K.F., et al., 1990; Lee L.H., et al., 1993; Lee L.S., et al., 1997; Zue, et al., 1989; Rabiner, et al., 1989 and Maneenoi E., et al., 2003), those technical models are widely used in many research of speech recognition area. In this task, various speech models are compared, to select a most suitable speech model for Lao language. As the principle of the proposed system, onset-rhyme models technique (Maneenoi E., et al., 2003) is mainly used as the acoustic model for base syllable recognition. The proposed system is individually recognized base syllables and tones. Therefore, the specific modeling of tone model is required. This research has also studied on various technical tone models, to find out a suitable tone model for continuous Lao speech

recognition. Since, haft-tone model technique (Thubthong N., et al., 2001), is investigated to use for tone recognition part of the proposed system. Haft-Tone model is a technical tone modeling for continuous speech, which is well know to prevent some effects from tone information of neighboring syllables.

4.2.4 Codebook

As the result, base syllable recognition is obtained the estimate syllables form as a sentence. Although, tone recognition has not considered in base syllable recognition, however, a possible tone of each syllabic component may be exactly known by following tone chart, especially, when that syllabic is belong in the group of dead syllable. Furthermore, since the syllabic component of a sentence can be vary with different tones but the meaning of that sentence is not always change. Therefore, all the tonal variable sentences were list as a codebook, called “Codebook”.

4.2.5 Tone Chart Applying

One alternative way, the performance of tone recognition can be improved using conditional of tone chart (see Table 3.8.), which the tone chart is made up corresponding to specific characteristic of each language. Therefore, this research has investigated to improve the performance of tone recognition system by using that tone rule to limit the sequent number of tone model as tone mapping that is defined as bellow

C_i : [H, M, L];

C_f : [naso, stop, non];

V : [short, long];

T : {0, 1, 2, 3, 4};

Tone of a syllable (T_s) will be known depend on the type of initial consonant (C_i), vowel (V) and final consonant (C_f), as presented bellow

H.long.{naso,non} = H.short.naso = {0, 1, 4} ;

H.short.{stop, non} = {4} ;

H.long.stop = {1} ;

M.long.{naso,non} = M.short.naso = {0, 1, 2, 4} ;

$M.short.\{stop, non\} = \{4\}$;
 $M.long.stop = \{1\}$;
 $L.long.\{naso,non\} = L.short.naso = \{0, 2, 3\}$;
 $L.short.\{stop, non\} = \{0\}$;
 $L.long.stop = \{2\}$;

Although, there are 3 classes of Lao consonant sounds (H, M, L), but some of their symbols are represented by the same phonetic symbol such as *f, h, k, s, p* and *t*. Therefore, even though, we have known all the phonetic symbols of syllable but we can not know its tone, exactly. For example: a syllable, $k.a.a.t.=T$, $T=\{1\}$ if *k* is belong in high class (H) and $T=\{2\}$ if *k* is belong in low class (H). So, the condition of this syllable, $k.a.a.t. = \{1, 2\}$.

In addition, half-tone model (H-T) is mainly applied as the tone model for tone recognition in this thesis, which is well known better than other tone models for continuous tone recognition (Thubthong N., et al., 2001). To recognize tone of a syllable, half-tone model technique was separately recognized tone of first half and second half. Where, the first half is consisted of rhyme of the preceding syllable and onset of the considering syllable, while the second half is consisted of rhyme of the considering syllable and onset of the following syllable. In addition, we can exactly know a possible tone of a syllable by following tone chart. Therefore, the requirement sequence of half-tone recognition can be decreased by using conditional of tone chart, i.e. each half of H-T model technique is required 25 sequences. However, it's required only 16 sequences maximum and less than that, with using conditional of tone chart.

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER V

EXPERIMENTAL RESULTS

In this chapter, experimental results will be presented with details. It is divided into 3 sections: base syllable recognition, tone recognition and combining of both base syllable and tone recognition, called *tonal syllable recognition*. A suitable feature set of Lao syllable recognition is obtained from by comparing between LPC cepstral coefficients and MFCC cepstral coefficients. In this chapter, it has also studied the suitable HMM architecture of base syllable and tone recognitions, which is correspond to speech model structure. Since, the comparison between and discussion of baseline system and the proposed system will be preformed at the end of this chapter.

5.1 Base Syllable recognition

The experiments of this section have been performed the training and recognizing with sample data from both male and female speakers. Since, the feature sets of LPC cepstral coefficients and MFCC cepstral coefficients are individually applied to find out which is a suitable feature vector set for Lao speech recognition. In the experiments, the set of LPC cepstral coefficients and energy contour are represented by LPCC, and MFCC was used to represent for the set of MFCC coefficients and energy contour. While the LPCC+ Δ and MFCC+ Δ are respectively the feature sets of LPCC and MFCC, including with their corresponding delta coefficients. As the result, a suitable speech model of base syllable recognition for Lao language is obtained from by comparing and evaluating of base syllable recognition, in cases of applying monophoneme, subword, initial-final and onset-rhyme models. In addition, the variable state of HMM models have also been studied to select most suitable number of HMM states for each speech model.

Table 5.1, 5.2, 5.3 and 5.4 are the results in case of applying monophoneme, subword, initial-final and onset-rhyme models, respectively. As the result, recognition rates are obtained by changing the number of HMM states, from 3 to 10. The results in all tables shown that, the recognition rate of base subword recognition is obtained the optimum with 6 states of HMM architecture. Since, that of base monophoneme, initial-final and onset-rhyme recognitions are obtained the optimum with 3 or/and 4

states of HMM architecture. In addition, the limited number of HMM states is usually depended on the time duration of each model instance, as shown in Table 5.1, 5.2, 5.3 and 5.4. Respectively, monophoneme models are based on phoneme unit, which is smaller unit in any conventional speech units, and the maximum available HMM states for training models, at 5 states. Subword model is based on syllable unit, which its duration is quite long, and the maximum available HMM states of these models is observed over 10 states. Since, the maximum available HMM states of initial-final models is not more than 5 states, consequently initial model of this technical model is also based on phoneme unit. And the last is onset-rhyme models, the similar of initial model, onset model was composed of the initial consonant and its forward transition to the vowel portion, and the maximum available HMM states of onset-rhyme model is limited at 6 states. Furthermore, the results in each table have also indicated that, the MFCC+ Δ feature set is always obtained higher recognition rate than that of other feature sets.

Table 5.1 Base syllable recognition based on monophoneme model

Features	Number of HMM state with Recognition Rates							
	3	4	5	6	7	8	9	10
LPCC	63.62	62.23	62.19	-	-	-	-	-
MFCC	69.40	69.89	67.64	-	-	-	-	-
LPCC+ Δ	74.64	73.78	70.83	-	-	-	-	-
MFCC+ Δ	77.67	77.71	74.15	-	-	-	-	-

Table 5.2 Base syllable recognition based on subword model

Features	Number of HMM state with Recognition Rates							
	3	4	5	6	7	8	9	10
LPCC	52.27	59.40	63.17	64.85	65.86	64.01	62.58	58.89
MFCC	59.90	62.75	69.21	71.48	72.23	71.98	69.71	65.69
LPCC+ Δ	72.15	78.02	82.55	83.22	82.63	81.80	81.12	78.52
MFCC+ Δ	80.54	82.97	85.82	86.41	86.16	84.73	82.89	80.70

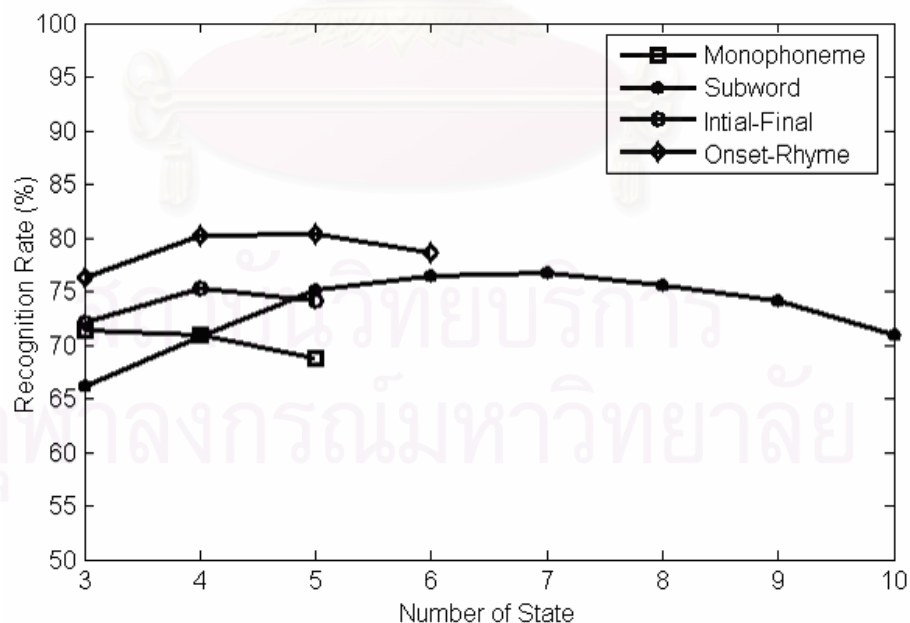
Table 5.3 Base syllable recognition based on initial-final model

Features	Number of HMM state with Recognition Rates							
	3	4	5	6	7	8	9	10
LPCC	60.71	65.40	62.73	-	-	-	-	-
MFCC	66.16	71.01	71.67	-	-	-	-	-
LPCC+ Δ	78.03	80.00	79.14	-	-	-	-	-
MFCC+ Δ	83.23	84.55	83.23	-	-	-	-	-

Table 5.4 Base syllable recognition based on onset-rhyme model

Features	Number of HMM state with Recognition Rates							
	3	4	5	6	7	8	9	10
LPCC	63.89	69.75	68.69	68.23	-	-	-	-
MFCC	71.82	75.96	78.64	76.92	-	-	-	-
LPCC+ Δ	81.82	86.31	85.45	83.18	-	-	-	-
MFCC+ Δ	87.47	88.89	88.79	85.91	-	-	-	-

Figure 5.1 has also presented the comparison of average recognition rates, in case of applying the different speech models. In the figure shown that, the performance of base onset-rhyme recognition are seen to be better than that of other, especially when the number of HMM states are around 4 and 5. However, the limited number of HMM states based on onset-rhyme model is observed maximum at 6 states. Consequently, the time duration of onset models are usually shorter than that of rhyme models. Therefore, the suitable number states of individual onset and rhyme models have been studied as Table 5.7.

**Figure 5.1** Base syllable recognition by using different speech models

In above experiments, the feature sets of LPCC and MFCC plus their corresponding delta coefficients are always given higher performance than that of using LPCC and MFCC only. The results in Table 5.5, which are obtained with using 12 coefficients for both LPC and MFCC feature extraction. As the result, it is that, the feature set of MFCC+ Δ can be given higher performance than that of LPCC+ Δ feature set, especially for the case of using onset-rhyme model (see Figure 5.2). Since, onset-rhyme model will be considered as speech model for next experimental of base syllable recognition.

Table 5.5 Comparison of using different speech models for both LPCC+ Δ and MFCC+ Δ

Features	Acoustic Models with Recognition Rates			
	Monophone	Subword	Initial-Final	Onset-Rhyme
LPCC+ Δ	73.78	78.02	80.00	86.31
MFCC+ Δ	77.71	82.97	84.55	88.89

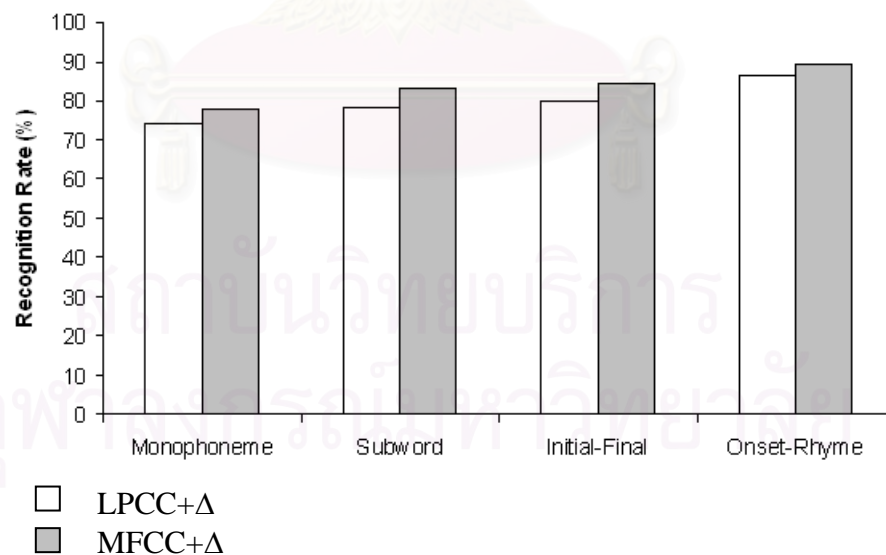


Figure 5.2 Comparison of using different speech models for both LPCC+ Δ and MFCC+ Δ

Table 5.6 Comparison of using different number of cepstral coefficients, in case of applying 4 states HMM of onset-rhyme model

Features	Number of Coefficients with Recognition Rates							
	8	9	10	11	12	13	14	15
LPCC	60.45	63.74	65.96	67.12	69.75	68.69	70.81	70.56
MFCC	75.00	74.09	74.95	75.25	75.96	75.56	76.01	76.16
LPCC+ Δ	82.93	83.79	85.15	86.16	86.31	86.62	85.81	86.62
MFCC+ Δ	88.54	89.24	88.89	88.99	88.89	88.94	88.43	88.59

The number of feature coefficients is also effect to performance of the system. The smaller number of feature coefficients can be decreased computing complexity, and the large enough number of feature coefficients maybe also given higher performance. As Table 5.6, the recognition rates of base syllable recognition based on onset-rhyme model, in case of using LPCC and MFCC feature sets are increased when the number of feature coefficients is usually increased up to 12, 13 and 14. Since, that of using LPCC and MFCC plus their corresponding delta coefficients (LPCC+ Δ and MFCC+ Δ) sets are obtained the optimal recognition rates at around 9 to 12 coefficients. As the result, in this experiment, the feature set of MFCC+ Δ is usually given higher recognition than other feature sets for base syllable recognition, as illustrated in Figure 5.3.

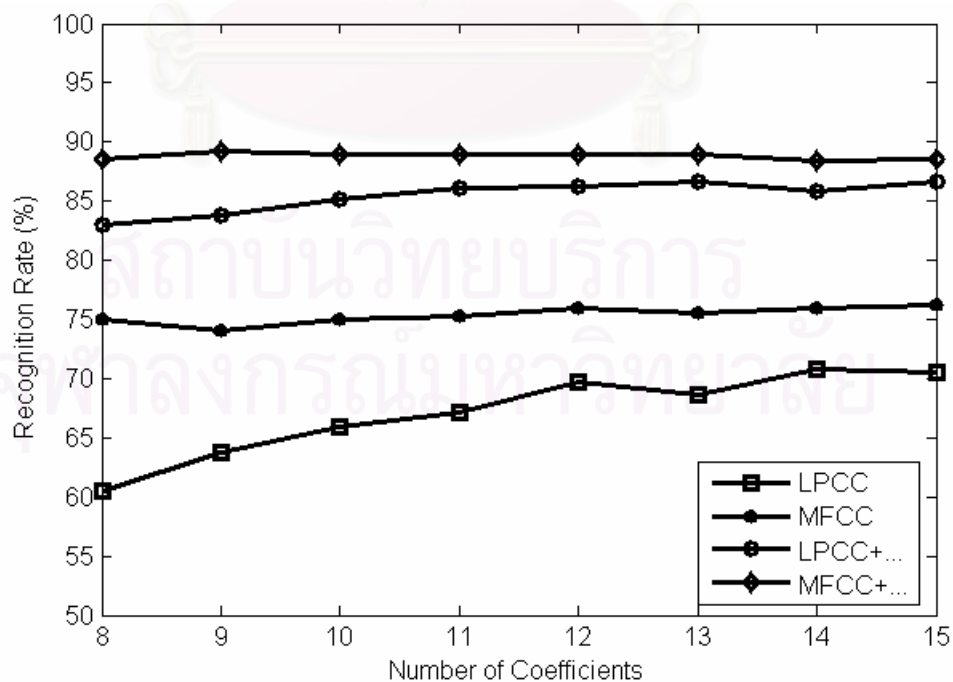
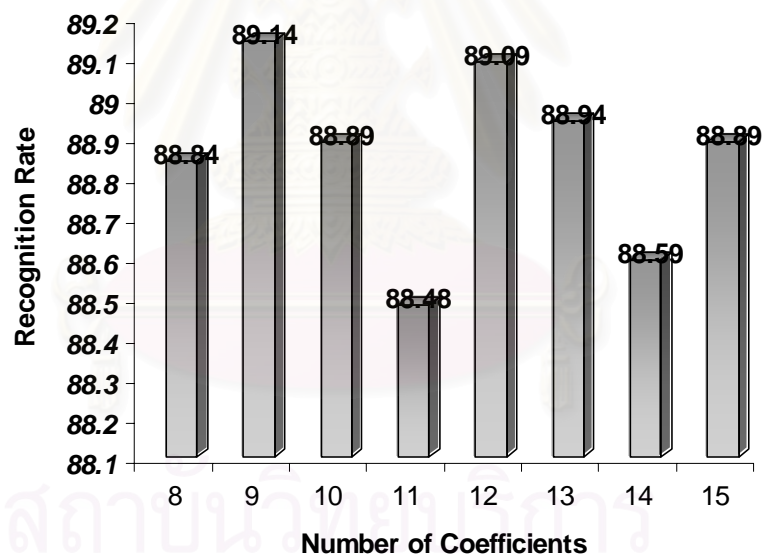


Figure 5.3 Onset-Rhyme recognition with different feature vector sets

Table 5.7 Base syllable recognition based on individual onset and rhyme models

		HMM state of Rhyme model							
		3	4	5	6	7	8	9	10
HMM state of Onset model	3	87.47	88.84	89.09	87.98	88.48	87.37	87.68	86.01
	4	87.68	88.89	89.09	88.38	87.88	86.92	86.46	85.15
	5	87.27	88.74	88.79	87.73	86.72	86.11	84.80	82.68
	6	85.81	86.77	86.77	85.91	85.45	83.99	81.31	79.24
	7	-	-	-	-	-	-	-	-
	8	-	-	-	-	-	-	-	-
	9	-	-	-	-	-	-	-	-
	10	-	-	-	-	-	-	-	-

**Figure 5.4** Comparison of 3-5-onset-rhyme model with different number of MFCC coefficients

The results shown in Table 5.7 are obtained with using 12-MFCC+ Δ feature set. The training models are individually specified number of HMM states for onset and rhyme models. In experiment, the available number of HMM states for rhyme model can be varied up to 10, since that of onset model can be varied up to 6 states

only. However, the optimum recognition rate is obtained with 5 states of HMM of rhyme, and 3 or/and 4 states of HMM of onset. Since, the smaller number of state is expected to be decreasing the computational complexity of a system. Therefore, 3 states of HMM of onset and 5 states of HMM of rhyme (3-5-onset-rhyme model) are introduced for Lao speech recognition base on onset-rhyme model, and was using for the proposed system in this thesis. However, a higher performance of base syllable recognition based on 3-5-onset-rhyme model can be obtained with 9-MFCC+ Δ feature set as shown by Figure 5.3, which is corresponding to the result in Table 5.6.

5.2 Tone recognition

In order, Tone recognition is an important part of this thesis. Since, several techniques have been presented to recognized tones for any tonal language, and a popular feature exaction of those researches is an analysis of pitch (or fundamental frequency). Therefore, in this task has also applied the sets of pitch analysis for tone recognition system. In addition, tone model can also be effect to the performance of tone recognition result. Therefore, the studying of tone model techniques is necessary to select a suitable tone model for tone recognition of Lao. As in Table 5.8, 5.9 and 5.10 are the results of tone recognition in different tone models and tone feature sets. In Table 5.8 is the specific results form female speaker, since the tone recognition of male speaker only are resulted as in Table 5.9, and the results of tone recognition of both male and female speakers are obtained as show in Table 5.10. As the result, tone recognition results in case of specific gender are seen to be better than that of both male and female, especially for in case of applying feature set of included direct fundamental frequency (F_0), consequently, fundamental frequency (or pitch) of different gender are usually performed in different contour levels.

Table 5.8 Tone recognition of female speaker only

Tone Models	Recognition Rates		
	F_0	QF_0	$F_0 + QF_0$
CI-T	52.45	54.43	52.16
CD-T	76.09	83.72	81.55
H-T	86.56	88.02	86.96

Table 5.9 Tone recognition of male speaker only

Tone Models	Recognition Rates		
	F ₀	QF ₀	F ₀ + QF ₀
CI-T	55.09	54.66	55.72
CD-T	78.37	85.10	82.26
H-T	86.36	88.71	87.54

Table 5.10 Tone recognition of both male and female speakers

Tone Models	Recognition Rates		
	F ₀	QF ₀	F ₀ + QF ₀
CI-T	46.44	52.08	50.17
CD-T	66.98	80.00	78.33
H-T	67.67	85.65	82.54

In addition, a suitable HMM architecture of a tone model is as importance to improve the performance result of tone recognition system. As shown in Table 5.11, the results of tone recognition are also varied following the number of HMM states is changed. A suitable number of HMM states for tone recognition based on context independent tone model (CI-T) and half-tone model (H-T), is around 5 states. Since the duration of context dependent tone model (CD-T) is longer than other two models (CI-T and H-T), so 7 states of HMM is more suitable for CD-T model.

Table 5.11 Comparison of continuous tone recognition with the various number of HMM states

Tone Models	Number of HMM state with Recognition Rates							
	3	4	5	6	7	8	9	10
CI-T	45.39	50.98	52.08	51.76	50.18	50.20	47.55	46.31
CD-T	70.41	75.87	80.00	83.20	86.78	86.01	84.77	81.96
H-T	75.64	83.66	85.65	82.23	80.76	80.44	78.02	74.39

The comparison of using different tone models such as CI-T, CD-T and H-T models is indicated in Figure 55. The results show that, CD-T and H-T models can be given higher performance than that of CI-T. Although, the performance results of both CD-T and H-T models can be given higher, however they are observed in different number of HMM states. Since, H-T model is observed in smaller number of HMM states, CD-T model is observed in bigger of that. In order, the sequence of CD-T

model is too larger than that of H-T model. Therefore, it's clearly that, the CD-T model system is obtained high performance with more complex computation than that of H-T model system.

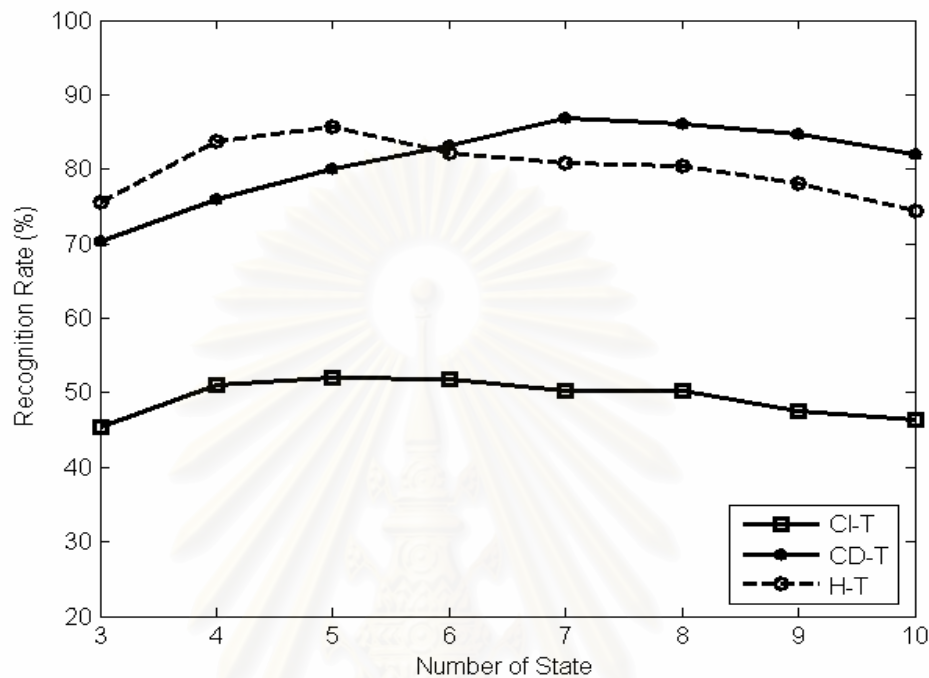


Figure 5.5 Comparison of continuous tone recognition results of using different tone models

5.3 Tonal Syllable recognition

As the proposed of this thesis, to combine both tone and base syllable recognition systems together, this section has been performed the experimental results by applying several techniques of tonal syllable recognition such as, joint detection (J.D.), sequential detection (S.D.) and the proposed method. Also, the experiment has been recognized for both speaker-independent (Speaker-ID) and speaker-dependent (Speaker-DD).

Table 5.12 is shown the results of the proposed system, by individually report for sub-system such as: base syllable and tone recognitions, and shown the final recognition as tonal syllable recognition (see Figure 4.1 for the process step of tonal syllable recognition based on proposed system). As the evaluated result in above section, 3-5-onset-rhyme model technique is used as a proposed model for base

syllable recognition with using 9-MFCC+ Δ set as the feature vectors, which is well known in the previous section that it can give better performance of base syllable recognition for Lao language. In this case, the performance of base syllable and tone recognitions are defined following the description of HTK manual book (S. Young et al., 2002). Since, the performance results of tonal syllable recognition based on the proposed method can be computed by following an equation below,

$$\text{Result}_{\text{final}} = \text{Result}_{\text{base syllable}} - (1 - \text{Result}_{\text{tone}}) \quad (5.1)$$

Where, $\text{Result}_{\text{tone}}$ is considered equal to the result of tone recognition part, when the system allows to recognizing tone. Otherwise, $\text{Result}_{\text{tone}}$ will be considered equal to 1. Since, $\text{Result}_{\text{base syllable}}$ is the result of base syllable recognition part and $\text{Result}_{\text{final}}$ is the final result of tonal syllable recognition system.

Table 5.12 Recognition of sub-system, including the final result for both speaker-independent and speaker-dependent recognitions

	Speaker-ID		Speaker-DD	
	Corr.	Acc.	Corr.	Acc.
Base Syllable	80.14	78.19	89.27	86.97
Tone	86.71	84.84	92.65	90.26
Tonal Syllable (Proposed Method)	66.85	63.03	81.92	77.23

Table 5.13 is shown the performance results of tonal syllable recognition in comparison of applying joint detection, sequential detection, a proposed method and also a proposed method without using Lao tone chart. Joint detection method is a classic method (H. M. Wang, et al., 1994 and T. Demeechai, et al., 2001), by combining of both features for base syllable and tone recognition as a feature packet. Sequential detection method, which is early presented in 1997 by C. J Chen, et al., and again, in 2001 by T. Demeechai, et al, this technique was required to separately recognize base syllable and tone recognitions. Subsystem of tone recognition will be applying when the target syllable is allowed to carry different tones. Similarly, the proposed system will be allowed to recognize tone when the target sentence is allowed to carry different tones on some syllabic components. While, the proposed

method has applied tone chart (Tone chart can be given from each tonal language) to decrease the sequent reference models during tones classification. The experiment has also compared the result of the proposed system [1] without using tone chart.

Table 5.13 Comparing the result of tonal syllable recognition in case of applying different techniques

	Speaker-ID		Speaker-DD	
	Corr.	Acc.	Corr.	Acc.
J.D. Method	63.11	58.93	68.20	64.11
S.D. Method	64.08	59.78	80.65	75.85
Proposed Method ^[1]	63.97	59.16	80.27	76.54
Proposed Method	66.85	63.03	81.92	77.23

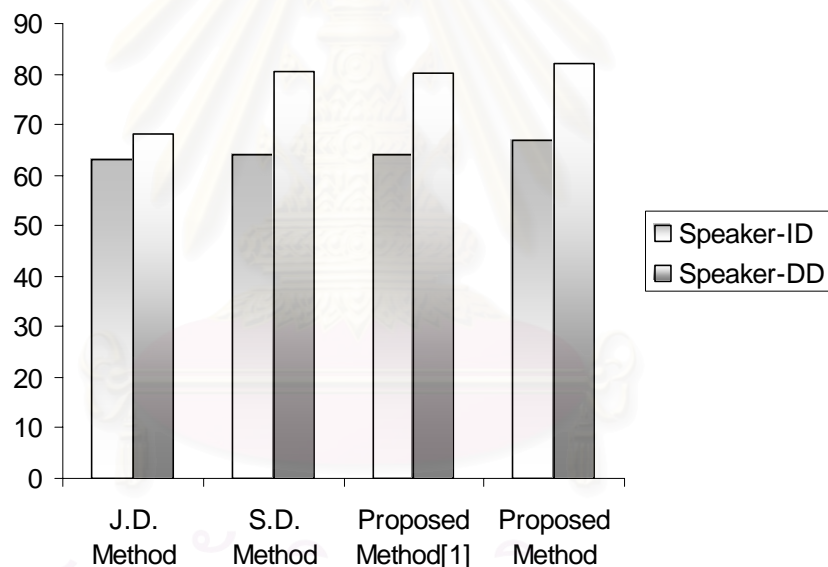


Figure 5.6 Comparison of tonal syllable recognition results using different techniques

In comparison, joint detection method is obtained lower performance result at 63.11% and 68.20% recognition rates, respectively for speaker-independent and speaker-dependent recognitions. As well known, tone information of neighboring syllables can be effected changing the shape of a tone (N. Thubthong, et al., 2001). However, joint detection technique has not applied any algorithm to prevent those effects, because joint detection is directly combined both features of phonetic and

tone as a feature set and recognized tonal syllable as base syllable recognition. While the proposed method is obtained higher performance result than the other, at 66.85% and 81.92% recognition rates, respectively for speaker-independent and speaker-dependent recognitions (see Figure 5.6). Furthermore, the proposed system has also provided faster recognizing than that of sequential detection method in the experiment (see Figure 5.7). Since, the results of sequential detection method and the proposed method are obtained similarity, in term of corrected recognition (see Figure 5.6). However, the recognition speed of proposed method ^[1] is seen to be clearly better than of sequential detection method as well, as illustrated in Figure 5.7.

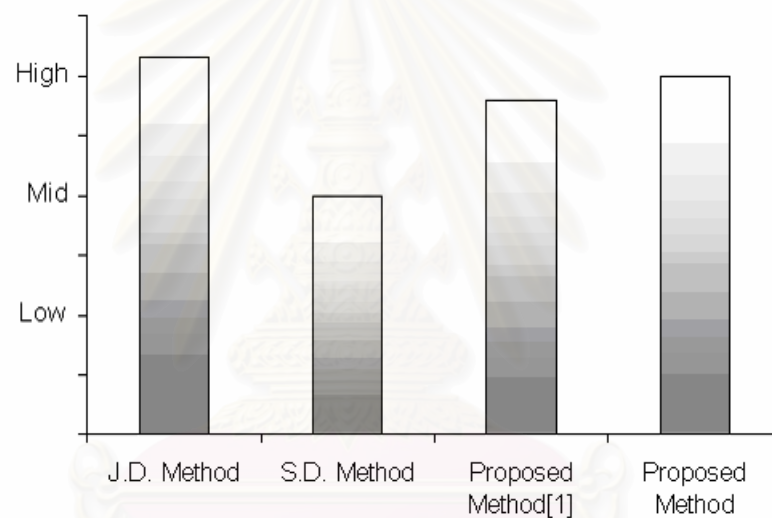


Figure 5.7 Recognition speed of using different techniques in z-score

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER VI

CONCLUSIONS

6.1 Summary and Conclusions

Automatic speech recognition (ASR) system has received considerable attention as a natural interface system for the future communication of human and machine. However, high performance of that system is well known with non-tone language such as: English, French, Italian, etc. Although, there are several speech recognition researches, have been worked on tonal language. The performance of continuous speech recognition is still haven limitation with that of tone recognition. Therefore, this research has proposed a robust method to recognize tonal syllable of continuous speech, implementing based on Lao language. The performance of tonal syllable recognition can also improved with a suitable combination technique of base syllable and tone recognitions. This thesis has proposed a similar method of the two-step tonal syllable recognition technique and sequential detection method. The proposed method is given a convenient method to implementation. Also, it has provided individual subsystem of base syllable and tone recognition that, we can be specifically applied the suitable techniques corresponding to each subsystem as well.

Subsystem of base syllable recognition was performed based on onset-rhyme model technique. Onset-Rhyme model is an optimal acoustic modeling for continuous Lao speech recognition, which is well known from the experimental results of this thesis. Although, onset-rhyme model technique is required large sequence of reference models, but it can be prevent the problem of continuous speech as well. In addition, MFCC coefficients were seem to be a suitable feature vector of base syllable recognition, especially for the set of MFCC and its delta coefficients. Since, the optimal performance of the proposed system is observed at around 9 MFCC coefficients. This number of feature coefficients is also given high performance, corresponding with 3 states of onset model and 5 states of rhyme model. Consequently, the duration of onset is usually less than that of rhyme. Therefore, the suitable HMM architecture for onset and rhyme models should be individually assigned.

Since, the results of base syllable recognition subsystem was obtained without considering tone, it maybe confused in Lao meaning of some word or sentence. Therefore, tone recognition is required to correct that. Tone recognition part of the proposed system is performed with the traditional feature vector of tone recognition system. Fundamental frequency (or pitch) contours of Lao utterance has computed to extract a suitable feature vectors of Lao tone recognition. In experimental result shown that, tone recognition result is obtained high performance with the feature set of 3-level quantization (QF_0). QF_0 feature is well known to prevent the effect of different tone levels of male and female. Although, the performance of using CD-T and H-T are seem very similar, but H-T is given lower sequence of reference tone models, which it can also be reduced the system complexity. Corresponding of H-T model technique, 5-states of HMM model is proposed for tone recognition, which is an optimal number of HMM states of tone model.

The proposed method is mainly proposed of two tasks. Firstly, the conditional division to recognize tone is performed at the sentence level. And secondly, the sequent number of reference tone models will be reduced by using conventional rules of Lao tone. As the result, the proposed method is advantaged on the recognition task. Although, it is requested times for training step, but it is shown high speed in recognition step. Also, the proposed system can obtained higher performance than that of other baseline system.

From the experimental results of this thesis, as expected, the performances of the proposed method are obtained higher than that of joint detection and sequential detection methods. The recognition rates of both cases of speaker-independent and speaker-dependent recognitions are shown at 66.85% and 81.92%, respectively, and 63.03% and 77.23% for accuracies. Since, the average recognition rate of the proposed system can be improved up 8% of joint detection, and 2% of sequential detection. Although, the recognition speed of the proposed method can not be fester than that of joint detection method. Clearly, it's shown faster than that of sequential detection as well. However, the similarity of different tone model and the differential of the same tone in continuous speech can be decreased the recognition rate of the system. Those problems are effect from some difference between Lao tone mark and acoustic tone.

6.2 Future Works

Although, the concept of the proposed method has been developed and observed with high performance, there are some interesting issues worth of investigations.

Some difference between Lao tone mark and acoustic tone may be prevent with other computation of speech signal. Lao tone marks had to present for long time before, the signal or speech processing have been studied, they preferred tone mark by listening from voided sound.

Due to limited resources, the speech data were recorded from only 30 male and 20 female speakers. However, this set of sample data can sufficiently be verified by the proposed tonal syllable recognition in evaluation. To implement a practical Lao continuous speech recognition interface system, more speakers are necessary to train the model variation from various speakers.



REFERENCES

- Borden G.J. and Harris K.S. Speech Science Primer: Physiology, Acoustic, and Perception of Speech, Baltimore, Maryland, U.S.A: Waverly Press, 1980.
- Chen C.J., Gopinath R.A., Monkowski M.D., Picheny M.A. and Shen K. New methods in continuous Mandarin speech recognition, Proceeding of Euro. Conf. Speech Communication Technology 3(1997):1543–1546.
- Chen S.H. and Wang Y.R. Tone recognition of Continuous Mandarin Speech Based on Neural Networks, IEEE Transaction of Speech and Audio Processing 3,2, March 1995.
- Haiping Li., Shen L.Q., and Fu G.K. Recognize Tone Languages using Pitch Information on the Main Vowel of Each Syllable, Proceedings of the 2001 IEEE International Conference on Acoustic, Speech, and Signal Processing 1(2001):61-64.
- Demechai T. and Makelainen K. Recognition of Syllables in A Tone Language, Speech Communication 33(2001):241-254.
- Denes P.B. and Pinson E.N. The Speech Chain. Bell Telephone Laboratories, 1963.
- Fant G. Acoustic Theory of Speech Production. The Netherlands: Mouton & Co., Printers, Hague, 1970.
- Furui S. and Sondhi M. Advances in Speech Signal Processing. New York: Marcel Dekker, 1992.
- Furui S. Digital Speech Processing, Synthesis, and Recognition. New York: Marcel Dekker, 2001.
- Ganapathiraju A. Hamaker J., Picone J., Ordowski M. and Doddington G.R. Syllable-Based large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing 9,4(2001):358-366.
- Haiping Li., Shen L.Q., and Fu G.K. Recognize Tone Languages using Pitch Information on the Main Vowel of Each Syllable, Proceedings of the 2001 IEEE International Conference on Acoustic, Speech, and Signal Processing 1(2001):61-64.

- Hartmann J. Spoken Lao - A Regional Approach, Center for Southeast Asian Studies, Northern Illinois University. SEAsite Laos 2002, Available from: <http://www.seasite.niu.edu/lao>
- Huang X., Acero A. and Hon H.W. Spoken Language Processing. New Jersey, U.S.A: Prentice Hall PTR, 2001.
- James G. Droppo III. Time-Frequency Features For Speech Recognition, Thesis of degree of Doctor of Philosophy, University of Washington, 2000.
- Juang B.H., Rabiner L.R., and Wilpon, J. On the Use of Bandpass Liftering in Speech Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 35,7(1987):947-954.
- Khanthavisone K. and Songwattana K. Tone Recognition Model for Laotian Language Using Pitch Quantization and Hidden Markov Modeling Techniques, Ladkrabang Engineering Journal 19,4(2002):1-6.
- Lawrence R. and Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, (1989):77-2.
- Lee Ch., Hyun D., Choi E., Jinwook Go. and Lee C.Y. Optimizing Feature Extraction for Speech Recognition, IEEE Transactions On Speech And Audio Processing, (January 2003):11-1.
- Lee K.F. Context-dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 38,4(April 1990):559-609.
- Lee L.H., Tseng C.Y., Gu H.Y., Liu F.H., Chang C.H. Lin Y.H., Lee Y., Tu, S.L., Hsieh, S.H. and Chen C.H. Golden Mandarin I – A Real Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary. IEEE Transaction on Speech and Audio Processing 1,2(1993):158-179.
- Lee L.S. Voice Dictation of Mandarin Chinese, IEEE Processing Magazine 14,4(1997),:63-101.
- Lee K.F. and Hon H.W., Speaker-independent Phone Recognition Using Hidden Markov Models, IEEE Transactions on Acoustics, Speech and Signal Processing 37,11(1989):1641-1648.

- Lee L.S., Voice Dictation of Mandarin Chinese, IEEE Processing Magazine 14,4(1997):63-101.
- Lee T., Ching P.C., Chan L.W., Cheng Y.H. and Mak B., Tone Recognition of Isolated Cantonese syllable, IEEE Trans, Speech and Audio Process 3,3(1995):204-209.
- Ling F. Speaker Recognition, Technical University of Denmark Informatics and Mathematical Modelling, Kgs. Lyngby, 2004, IMM-THESIS: ISSN 1601-233X.
- Loizou P.C. Robust Speaker-Independent Recognition of A Confusable Vocabulary, Doctoral dissertation, Arizona State University, 1995.
- Maneenoi E. Thai Vowel Phoneme Recognition using Artificial Neural Networks. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1998.
- Maneenoi E., Luksaneeyanawin S. and Jitapunkul S. An Acoustic Study of Syllable Rhymes: A Basis for Thai Continuous Speech Recognition System, Ph.D.'s thesis, Electrical Engineering, Faculty of Engineering Chulalongkorn University, 2003.
- Maneenoi E., Ahkuputra V., Luksaneeyanawin S. and Jitapunkul, S. A Study on Acoustic Modeling for Speech Recognition of Predominantly Monosyllabic Languages, to be published in Special Issue on Speech Dynamics by Ear, Eye, Mouth, and Machine, IEICE Transaxtion on Information and Systems E87-D, 5, 2004.
- Mingbuapha K. and Poomsan B. B. Lao-English /English-Lao Dictionary. Paiboon Publishing, 2003.
- Ngarmchatetanarom N., Maneenoi E., Assadornwiset W. and Jitapunkul S. Tone Recognition of Thai Continuous Speech using Fujisaki's Model. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 2003.
- Noulnavong O., Sisuvhan P., Paphaphan B., Sengsulini B., Sihalaat S. and Sisawan K. Advance Lao Grammar, Lao Education Ministry, UNICEF, 2003.

- Owens F.J. Signal Processing of Speech, The Macmillan Press LTD. Hampshire RG21 2Xs and London Houndmills, Basingstoke, , 1993.
- Paphaphan B., Noulnavong O., Sisuvhan P., Sengsulini B., Sihalaat S. and Sisawan K. Lao Grammar, Lao Education Ministry, UNICEF, 2000.
- Picone J. Continuous Speech Recognition Using Hidden Markov Models, IEEE ASSP Magazine, 26, 1990.
- Rabiner L.R. and Allen J.B. A Unified Approach to Short-Time Fourier Analysis and Synthesis, Proceeding of the IEEE 65,1(1977):1558-1564.
- Rabiner L.R. and Juang B.H. An Introduction to Hidden Markov Models, IEEE ASSP Magazine 3,1(1986):4-16.
- Rabiner L.R. and Juang B.H. Fundamentals of Speech Recognition. PTR Englewood Cliffs, New Jersey: Prentice Hall, 1993.
- Rabiner L.R., Wilpon J.G., and Soong F.K. High Performance Connected Digit Recognition Using Hidden Markov Model. IEEE Transactions on Acoustics, Speech, and Signal Processing 37,8(1989):1214-1224.
- Ravishankar M.K. Efficient Algorithm for Speech Recognition, Doctoral dissertation, School of Computer Science, Carnegie Mellon University, 1996.
- Robinson A. J. Connectionist speech recognition of Broadcast News, Speech Communication 37, Issues 1,2(2002):27-45.
- Shannon C.E. A Mathematic Theory of Communication, Bell Systems Technical Journal 27(1948):379-423.
- Simon Ager, A Guide to Written Language. Available from: <http://www.omniglot.com/writing/lao.htm>, 1998-2004.
- Sondhi M.M. New Methods of Pitch Extraction, IEEE Trans. AU-16,2(Jun. 1968):262-266.
- Steinbissa V., Ney H., Essen U., Tran B.H., Aubert X., Dugast C., Kneser R., Meier H.G., Oerder M., Haeb-Umbach R., Geller, D., Höllerbauer W. and Bartosik H. Continuous speech dictation - From theory to practice. Speech Communication 17, Issues 1,2(1995):19-38.

- Thubthong N., Kijirikul B. and Luksaneeyanawin S. Stress and Tone recognition of polysyllabic words in Thai Speech, Proceedings of the Inter. Conference on Intelligent Technologies, (2001):356-364.
- Thubthong N., Pusittrakul A. and Kijirikul B. Tone Recognition of Continuous Thai Speech Under Tonal Assimilation and Declination Effects Using Half-Tone Model, Inter. Journal of Uni., Fuzziness and Knowledge-Based Systems 9,6(2001):815-825.
- Tungthangthum A. Tone Recognition for Thai, Proceeding of the IEEE Trans. (1998):157-160.
- Valtchev V. Discriminative Methods in HMM-based Speech Recognition. Doctoral dissertation, University of Cambridge, 1995.
- Wang H.M. and Lee L.S., Tone Recognition for Continuous Mandarin Speech with Limited Training Data Using Selected Context-Dependent Hidden Markov Models, Jour. Chinese Institute of Engineers 17,6(1994):775-784.
- Young S., Kershaw D., Odell J., Ollason D., Valtchev V. and Woodland P. The HTK Book Version 3.0, Microsoft Corporation, 2002.
- Young S.J., The General Use of Tying in Phoneme-Based HMM Speech Recognizers, Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, (1992):569-572.
- Zhang G., Zheng F. and W.WU., Tone Recognition of Chinese Continuous Speech, ISCSLP, (Oct. 2000):207-210.
- Zue V., Glass J., Philips M. and Seneff S. Acoustic Segmentation and Phonetic Classification in SUMMIT System, Proceedings of the 1989 IEEE International Conference on Acoustic, Speech and Signal Processing 1(1989):389-392.
- Zhang I.S. and Hirose K. A Study On Robust Segmentation And Location Of Tone Nuclei In Chinese Continuous Speech, IEEE Transactions, ICASSP 2004, 0-7803-8484 (2004).



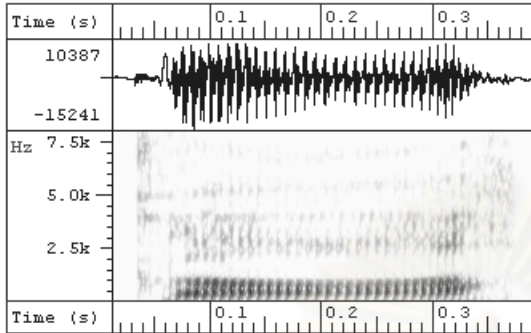
APPENDICES

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

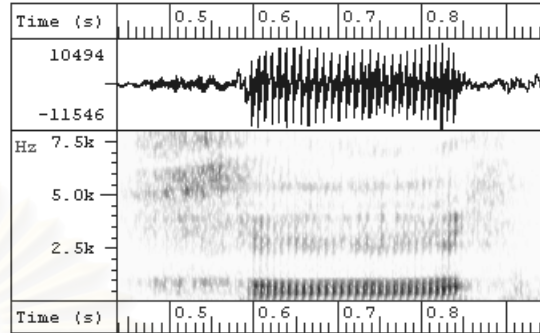
Appendix A

Waveform and Spectrum of Lao Consonants

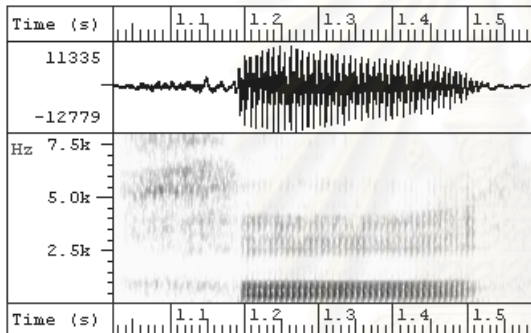
ກ /g/



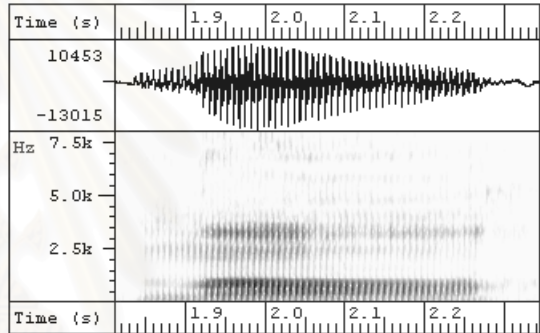
ຂ /k/



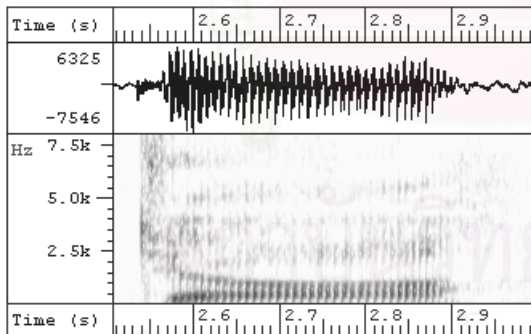
ຄ /k/



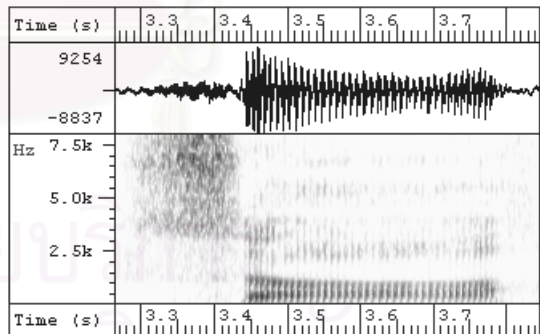
ງ /ng/



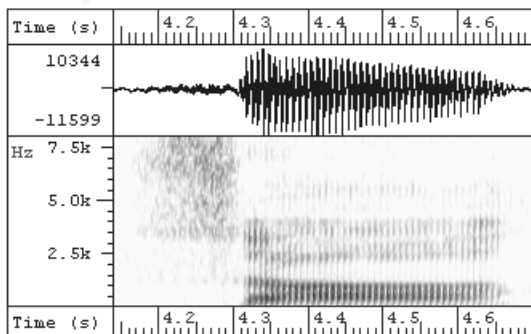
ຈ /j/



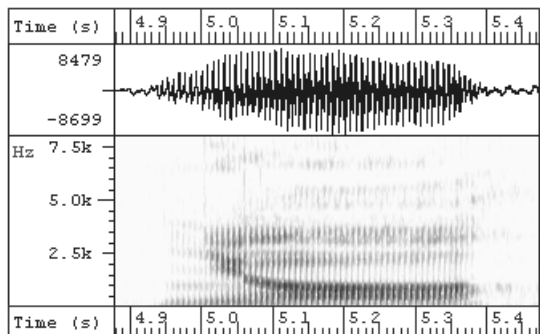
ສ /s/



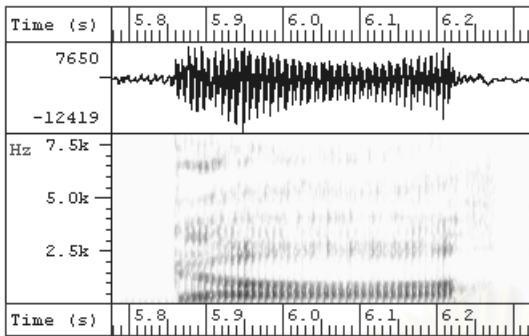
ຊ /s/



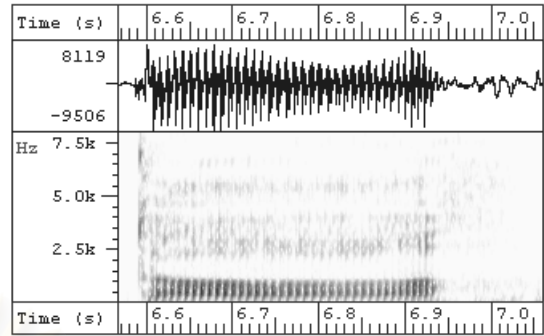
ຢ /ny/



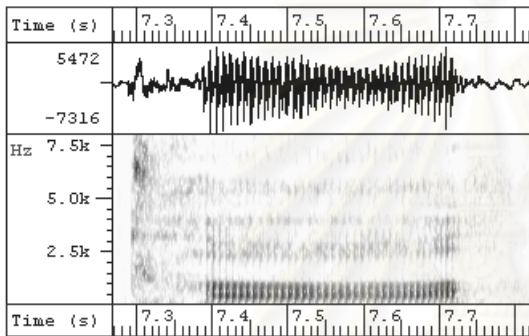
៦ /d/



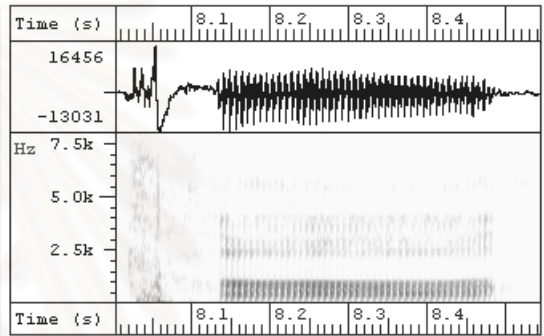
៧ /t/



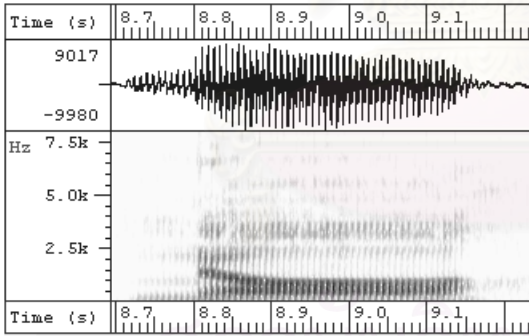
៨ /th/



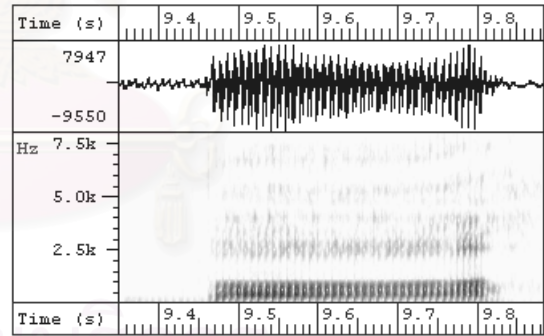
៩ /th/



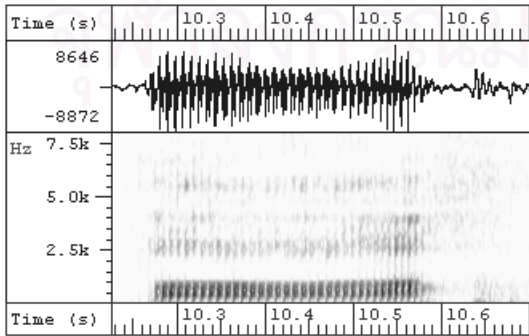
១០ /n/



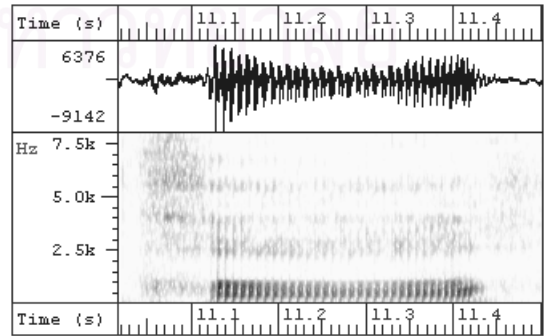
១១ /b/



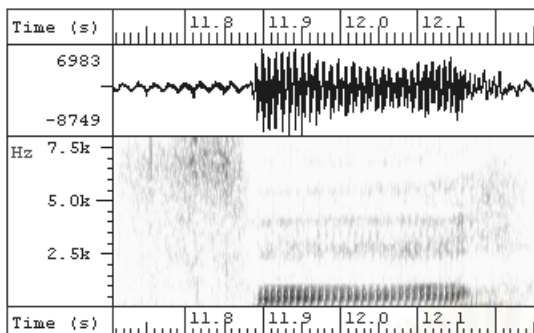
១២ /p/



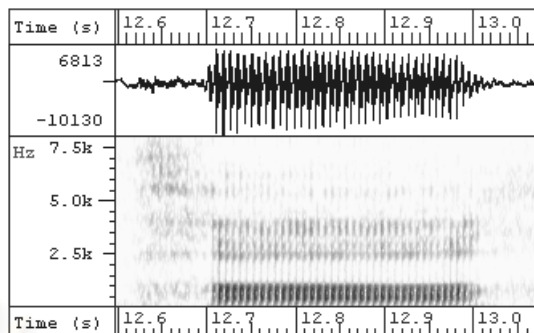
១៣ /ph/



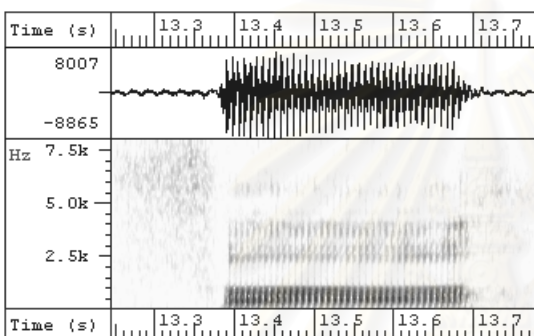
ຝ /f/



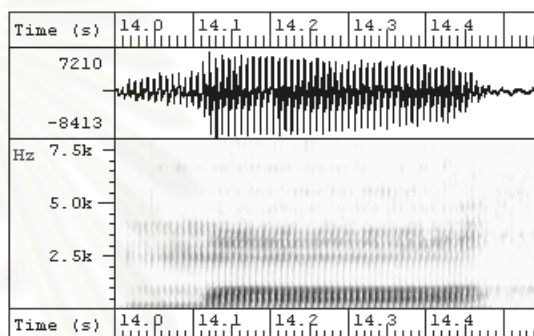
ພ /ph/



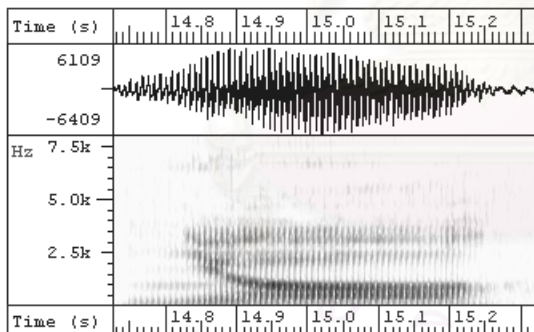
ຟ /f/



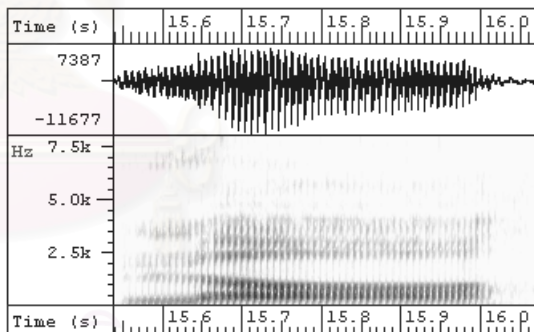
ມ /m/



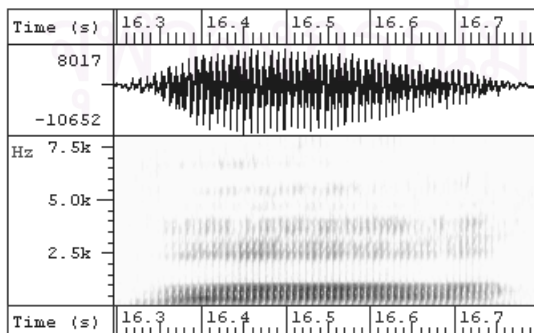
ຢ /y/



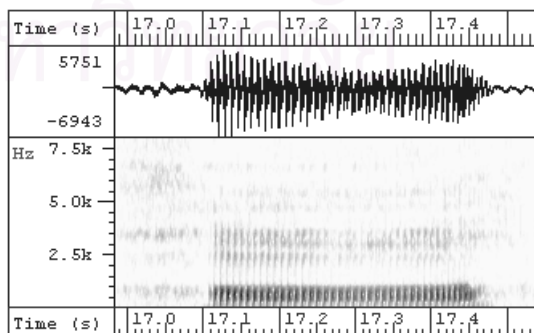
ລ /l/



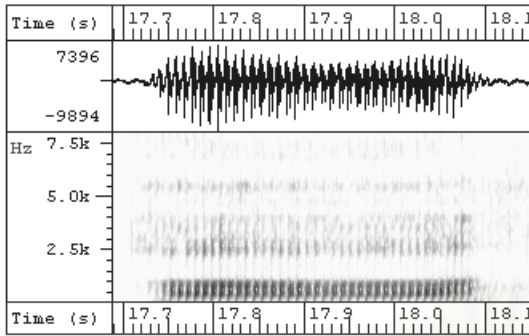
ວ /w/



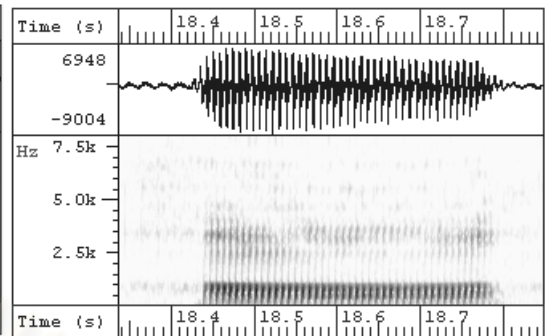
ຫ /h/



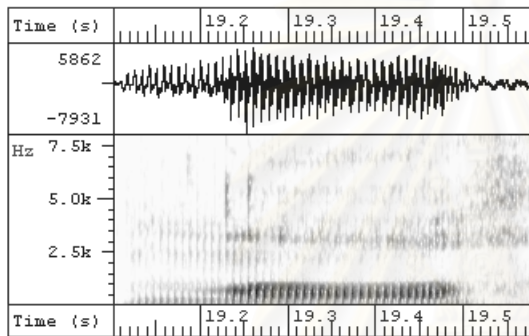
ອ /z/



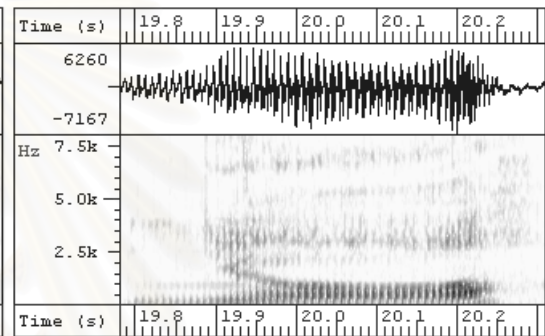
ຮ /h/



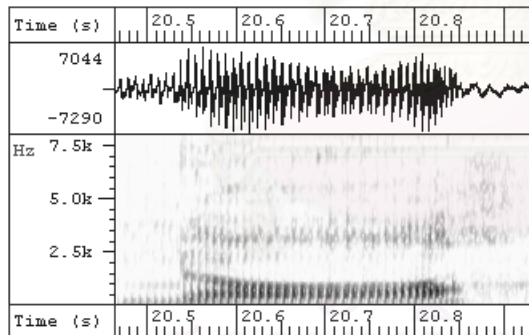
ຫງ /ng/



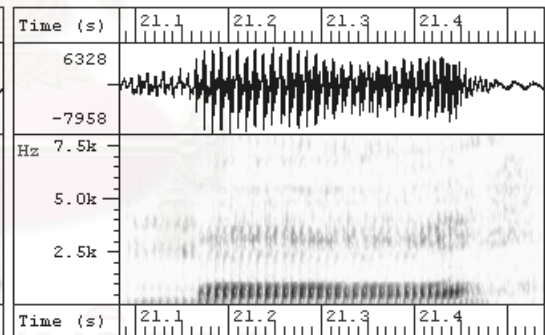
ຫຍ /ny/



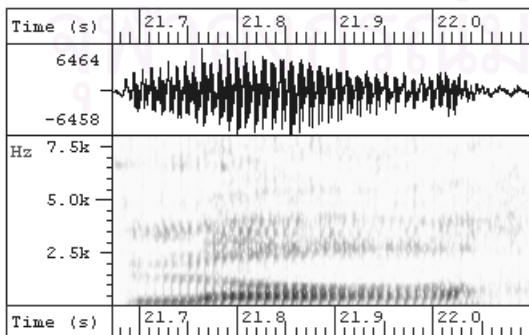
ຫນ (ໜ) /n/



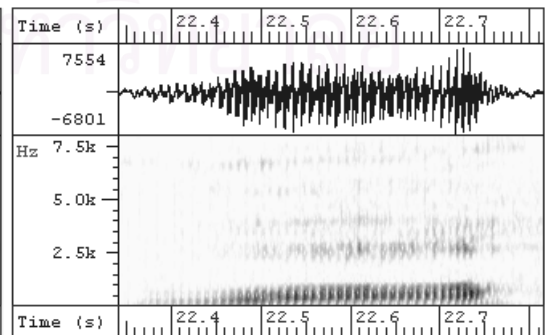
ຫມ (ໝ) /m/



ຫລ (ຫຼ) /l/

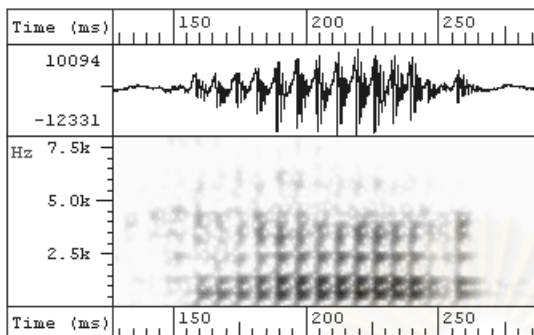


ຫວ /w/

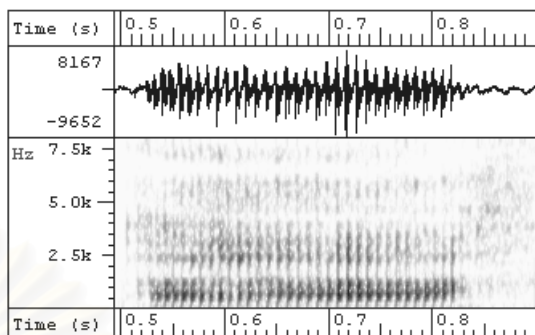


Waveform and Spectrum of Lao Vowels

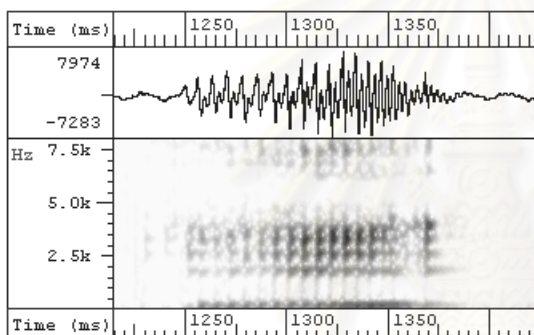
ຂໍ້ /a/



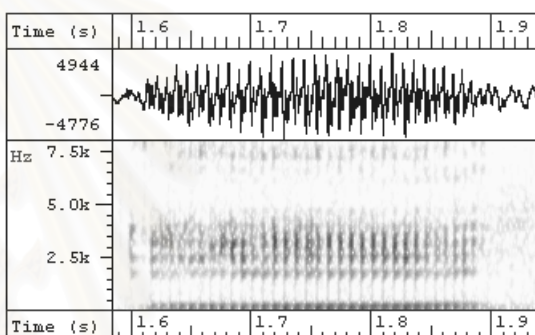
ຂໍ້ /aa/



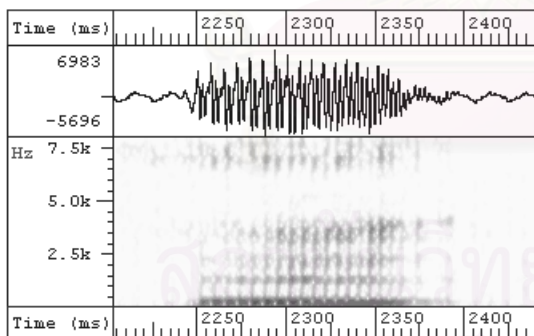
ຂໍ້ /i/



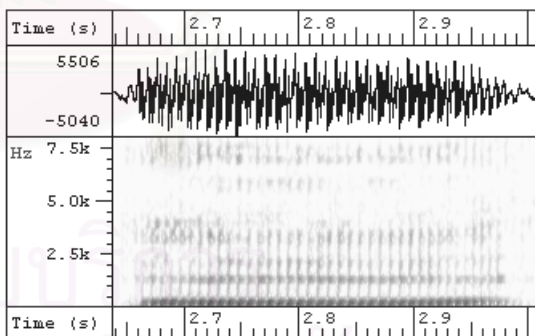
ຂໍ້ /ii/



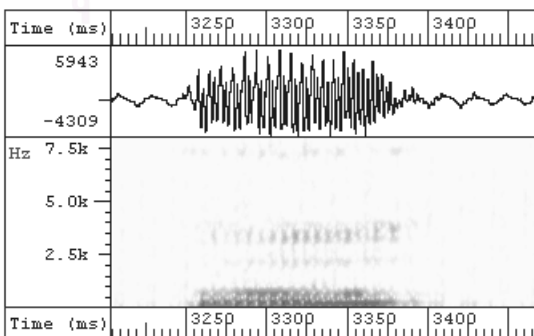
ຂໍ້ /q/



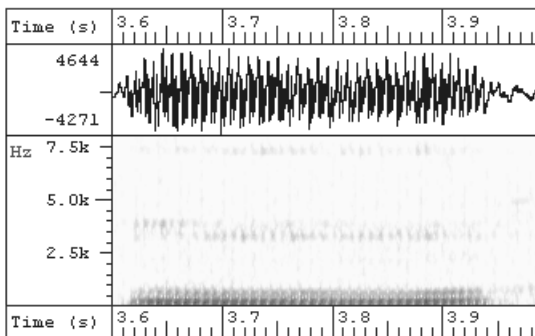
ຂໍ້ /qq/



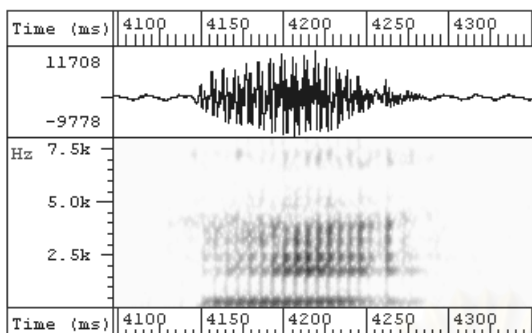
ຂໍ້ /u/



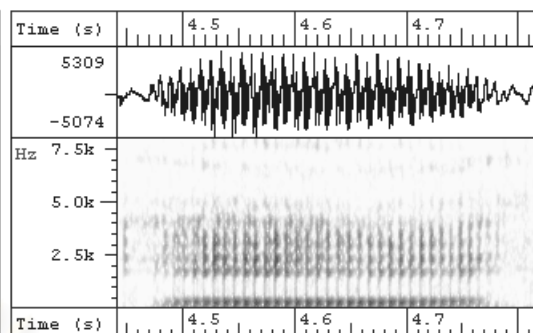
ຂໍ້ /uu/



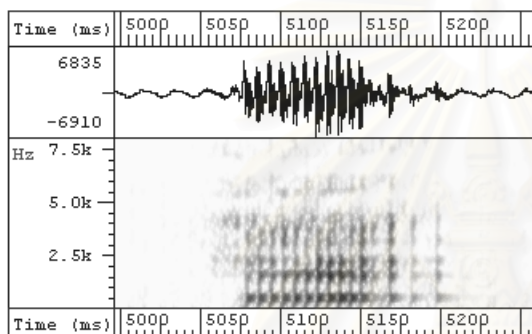
⌘x⌘ /e/



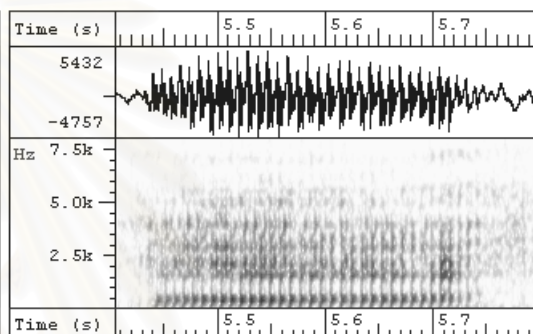
⌘x /ee/



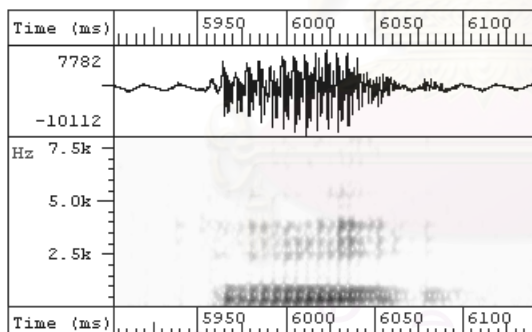
⌘x⌘ /x/



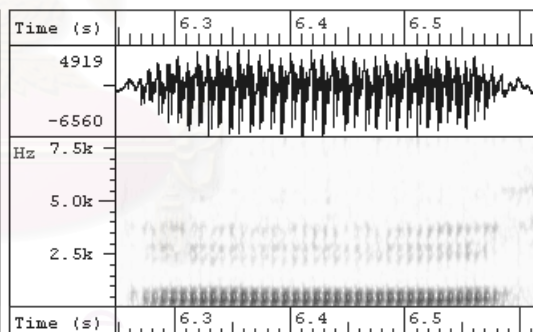
⌘x /xx/



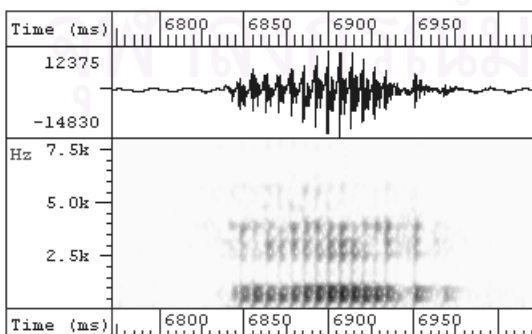
⌘x⌘ /o/



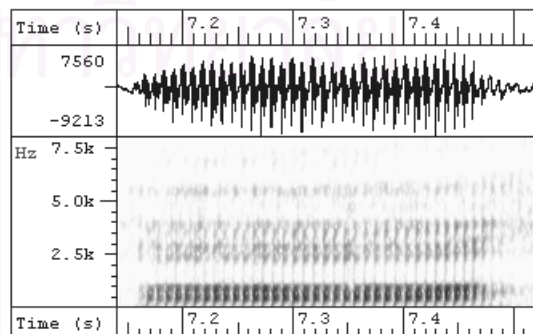
⌘x /oo/



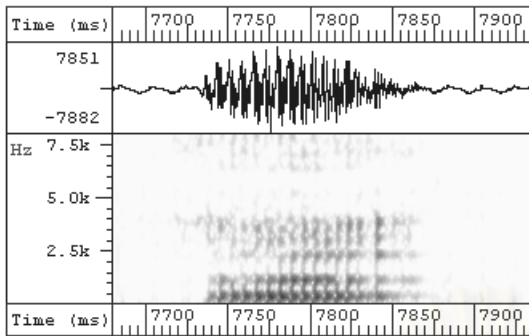
⌘x⌘ /aw/



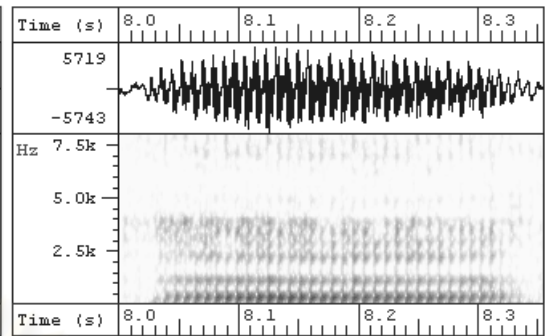
⌘x /aaw/



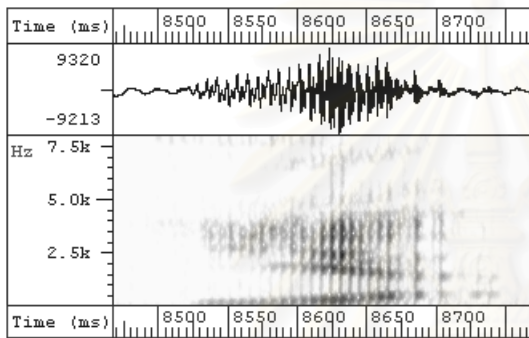
ḿ /qa/



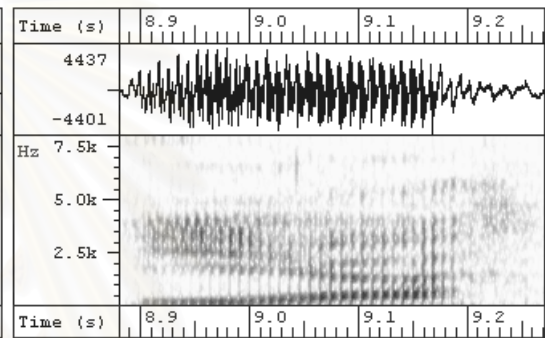
ḿ /qae/



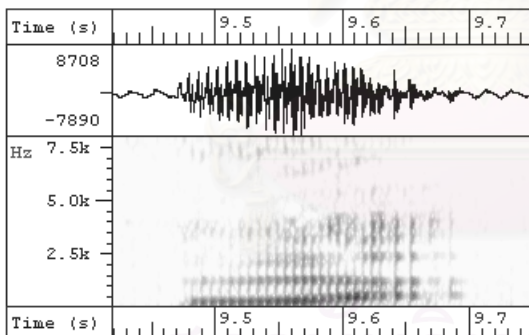
ḿ /ia/



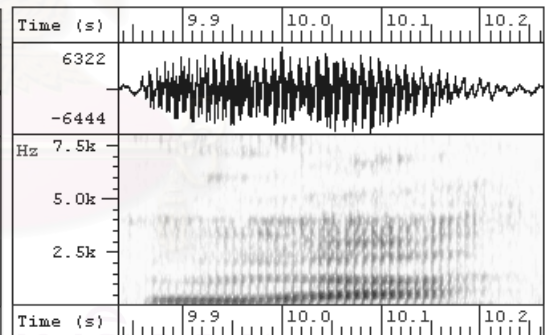
ḿ /iia/



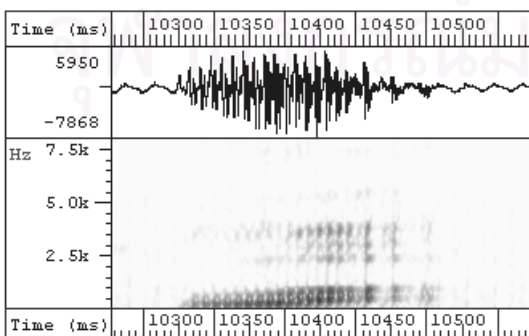
ḿ /qu/



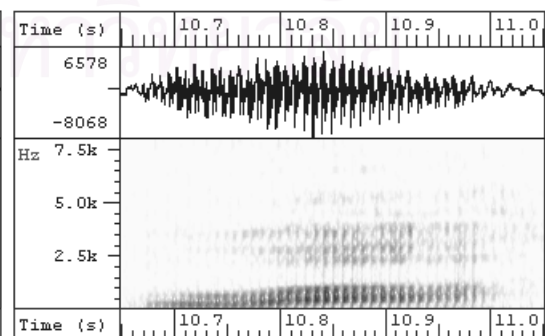
ḿ /qua/



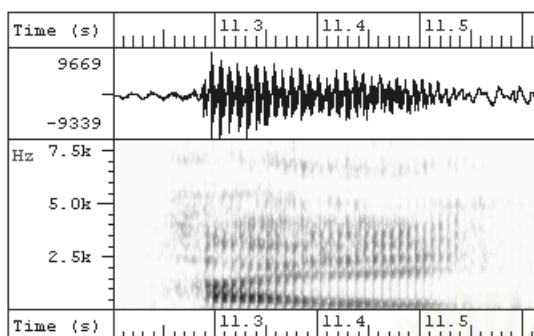
ḿ /ua/



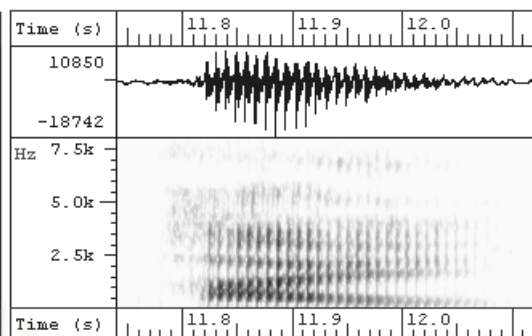
ḿ /uaa/



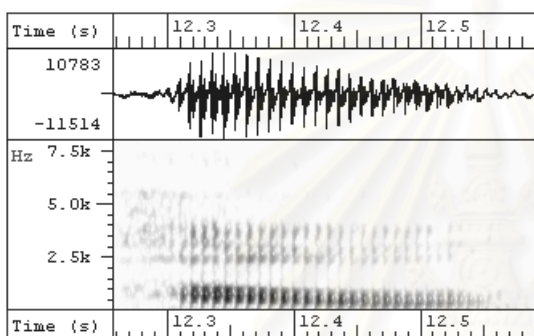
ไอ /ai/



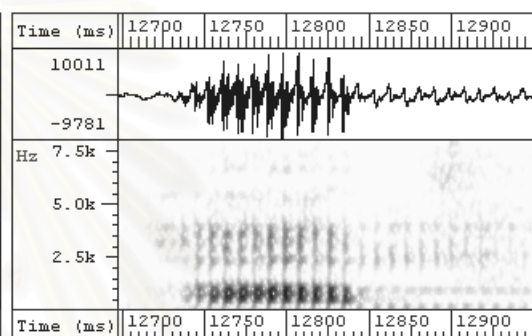
ไอ /ai/



ออ /ao/



อำ /amh/



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

Appendix B

List of Soft-tools

RedHat 9.0 (Linux)	http://www.redhat.com
HTK v. 3.0	http://htk.eng.cam.ac.uk
SFSWin v. 1.5	http://www.phon.ucl.ac.uk/resource/sfs
SFS/WASP v. 1.3	http://www.phon.ucl.ac.uk/resource/sfs/wasp.htm
Praat v. 4.2	http://www.praat.org
Cooledit.pro v. 1.0	http://www.syntrillium.com/cooledit
Winsnoori v. 3.1	http://www.loria.fr/~laprie/winsnoori
Speech Labeler v. 1.0	http://www.eng.chula.ac.th



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

Appendix C

Published Paper

Chanthamenavong S., Maneenoi E. and Jitapunkul S. *Robust Method of Continuous Speech Recognition for A Tonal Language*, ECTI International Conference (ECTI-CON2005) Trans., 2005 (to be publish).



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

Robust Method of Continuous Speech Recognition for A Tonal Language

S. Chanthamenavong*, E. Maneenoi**, and S. Jitapunkul***

DSPRL, Electrical Engineering, Chulalongkorn University, Bangkok, Thailand, 10330.

*senglathamy@hotmail.com, **ekkarit@hotmail.com, ***somchai.j@chula.ac.th

ABSTRACT

This paper proposes a combination of base syllable and tone recognitions technique in order to improve the performance of continuous speech recognition for tonal language. In the experimental result, Lao language was selected to test the proposed technique, the analysis of mel-frequency cepstral coefficients, and normalized pitch contour were mainly applied to extract feature vector, respectively, for base syllable recognition and tone recognition systems. Then, tone recognition was applied immediately when the codebook allowed the target sentence (the result of base syllable recognition) to carry different tones. To improve performance of tonal syllable recognition system specifies language model and tone mapping of a tonal language were used. The results show that, the proposed system can be obtained higher performance in comparison with the baseline system. In term of recognition speed, the proposed system can be improved up to 25% of that of baseline for small codebook.

Keywords: Tonal Syllable Recognition, Continuous Speech Recognition, Lao Recognition.

1. INTRODUCTION

In a tonal language, tones are lexically significant of word meaning. Therefore, tone information is very essential for speech recognition of tonal languages. In a decade, many researchers have been done separately for syllable and tone recognition. However, to create a novel speech recognition system for tonal language, the combining of base syllable recognition and tone recognition are required. Since the success of base syllable recognition is most processing for isolated word recognition. However, continuous speech recognition has been also developed in recent years. Although, continuous speech recognition is a complex system but it can meet the recognition of natural speech target better than isolated word. Almost continuous speech recognition has been processed with applying the syllable structure of a language to design the acoustic model, such as word, monophone, initial-final and onset-rhyme models [1, 5]. Where, the system based on initial-final and onset-rhyme models are leader such as, the research of E. Maneenoi, et al. [1], as their study, the continuous speech recognition system utilizing onset-rhyme model as speech unit has given high performance. Unfortunately, tone recognition has not been considered in their experimentation. Also,

many researches of tone recognition have been developed for tonal language such as Mandarin, Canton, Thai, Japanese and etc. We expect that, those researches will be adapted to Lao language as well.

Various techniques such as dynamic time warping, neural network and hidden Markov models, were studied and used in speech recognition process, such as isolated word, connected word, continuous speech and etc;. However, the Hidden Markov models technique is widely used in continuous speech recognition. Also, the recognition of Lao tones by using HMM has been presented [2]. In their result, Lao tone recognition system is given the high recognition rate in the implementation with isolated word recognition. Presently, there is no Lao speech recognition dealing with continuous speech. Therefore, to implement a natural continuous speech recognition system for Lao speech, it should be exploited as tonal syllable recognition.

In recent years, there are some papers to be already presented the methods for tonal syllable recognition, based on continuous speech such as Joint detection, Sequential detection and Linked detection [3]. The sequential detection method is observed with high performance, and lower computational complexity than that of

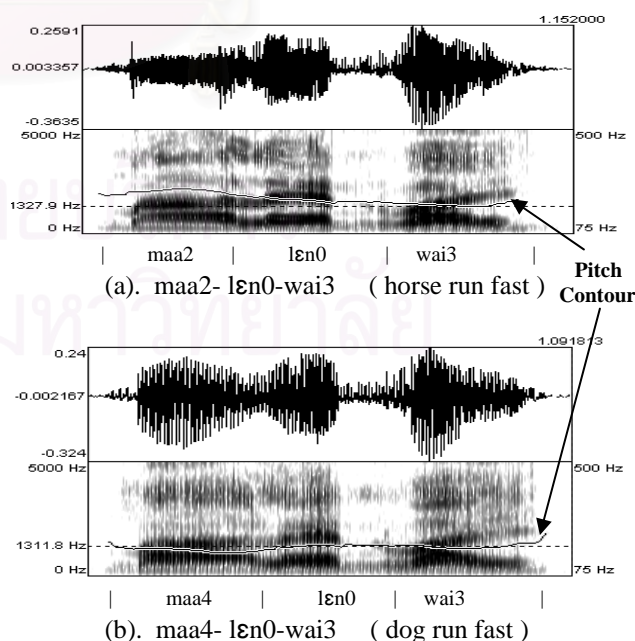


Fig.1: The sample sentences with different tone

Table 1: Tone mapping of Lao syllable

Initial consonant class \ Syllable	Live Syllable			Dead Syllable	
	Inherent Tone	[˩] (low tone mark)	[˨˥] (falling tone mark)	Long Vowel	Short Vowel
High Class [k][s][t][p][f][h]	Low Rising (4)	Mid (0)	Low Falling(1)	Low Falling (1)	Low Rising (4)
Middle Class [g][ch][d][dt][b][bp] [y][ɔ]	Low Rising (4) / Low Falling (1)		High Falling (2)		
Low Class [k][s][t][p][f][h][ng] [ny][n][m][r][l][w]	High Rising(3)		High Falling (2)	Mid (0)	

both linked detection and joint detection methods. In the process of sequential detection, tone recognition will be enabled only when the estimated syllable of base syllable recognition able to carries difference tone. In fact, it's not necessary some time. As specific characteristic of tonal language, especially for Lao language, several Lao syllables can be changed the meaning itself, when it carries different tone [8], but when those syllables are constructed as a sentence, the meaning of that sentence is not always changed with different tones. Therefore, we expect that, the performance of the sequential detection method will be improved and reduce recognition time duration, by recognizing tone of syllabic components in a sentence only when that sentence has a possibility to change the meaning with different tones.

This paper has organized follow as. In section 2, specific characteristic of Lao spoken language, one of tonal language, will be investigated in order to use it as our example. The tonal syllable recognition of continuous speech method will be presented in selection 3. Section 4 will be the evaluation of experimental results, in several aspects. Finally, the conclusions of this paper will be described in section 5.

2. LAO SPOKEN LANGUAGE

Almost Lao spoken words are monosyllabic word, and perform several functions in a sentence. A polysyllabic word is constructed by concatenating each syllable. So, the several combinations of these syllables with tones can produced the several words. In addition, a sentence is formed by a serial construction of these syllables.

There are 27 consonants in Lao alphabetical order and six high consonant clusters, representing 21 sounds, and are divided into three classes, high consonants, middle consonants and low consonants. Lao language has 28 vowels representing 27 sounds and can be divided into three classes, short vowels, long vowels and additional

$$S = C_i(C_c)V(C_f)^T$$

Fig.2: A Lao syllable structure

vowels. Furthermore, the spoken language can be pronounced five tone sounds [8].

Lao syllable structures are composed of three parts of sound, consonants, vowel, and tone. The standard Lao syllable structure can be presented as illustrated in "Fig.2:", where C_i is an initial consonant, C_c is a co-articulated consonant, V is a vowel, C_f is a final consonant and T is a tone. There are three types of Lao syllable, Open-syllable, Live-syllable and Dead-syllable. And each syllable is able to appear with different tone [8, 9], as presented in "Table 1:".

Normally, every Lao syllable consists of a tone, and it always change the meaning itself when it has a different tone. For example: *gai0* (chicken) and *gai1* (far). Also, the meaning of a Lao sentence will be changed, when it consists of syllables with different tones, as shown in "Fig. 1:". However, some Lao sentences do not change the meaning when their tone is changed. For example: *kai1-ka01-gap4-gai0* (I have chicken for a meal), and *kai1-ka00-gap4-gai1* ("It has no meaning in Lao!").

3. PROPOSED METHOD

This paper proposes a tonal syllable recognition method of continuous speech recognition for tonal language. The proposed system will be processed step by step as below (Four step in total), and its chart is presented in "Fig. 3:"

- Feature Extraction: the speech waveform is through signal analyzer to extract phonetic feature for base syllable recognition. While tone feature will be extracted immediately, after the enable of tone recognizer. For the detail can see in section 3.1.
- Base syllable recognition: In this step, phonetic features in HMM of base syllables are implemented based on phonotactic onset-rhyme model in order, to recognize syllabic components of a sentence. The estimated syllable boundaries and its time derivative are obtained as the results of base syllable recognition.
- Automatically, system will check the result of base syllable recognition in codebook that contains all variable sentences associated with tone information. If the checking function returns false, the result of base syllable recognition will be accepted

as final result of system. Otherwise, tone recognition part is implemented immediately.

– If the target sentence has been found in codebook, altogether tone recognizer is enabled. Firstly, tone feature has extracted and segmented by using information of the estimated syllable boundaries and it's time derivative which are the results of base syllable recognizer. Then, HMM of tones associated with tone mapping information (see “Table 1:”) are implemented based on left-tone-dependent model to recognize a possible tone of syllabic component in a sentence.

3.1. Feature Extraction

A hamming window of 30 ms frame size has been applied every 10 ms (frame shift). Then, computation of mel-frequency cepstral coefficients (MFCCs), log-energy and zero-crossing were applied to extract phonetic feature vector for base syllable recognition as shown in upper part of “Fig. 3:”. When tone recognizer is enabled. The computation of log-pitch is used as tone feature vectors,

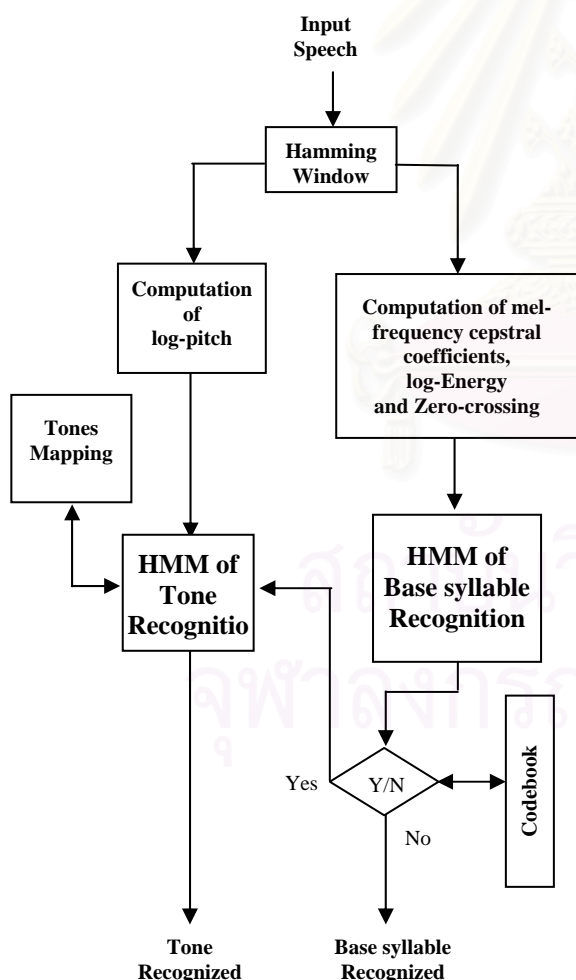


Fig.3: Block diagram of tonal syllable recognition

and is segmented by using information of the estimated syllable boundaries and it's time derivative, which are the result of base syllable recognition. Consequently, mean and variance of pitch contour for the same tone can vary considerably from one speaker to another, especially between male and female. Therefore, the pitch contour should be normalized before applied as tone feature. In paper [6], the pitch was normalized by division with average speaker's pitch. In paper [2], the pitch was normalized by using the difference of pitch and that of left frame, then quantized into three levels (-1, 0, 1). This technique can give high recognition rate for Lao tone recognition. This paper has also adopted the technique of [2] to normalize of tone feature vector.

4. EXPERIMENTS

4.1. Speech Data

The performance of this technique will be evaluated using sample speech data of Lao speech sentence. The speech data will be recorded twice of 10 Lao sentences from both male and female speakers (eight males and seven females) utterance on speed speaking style, in open office environment. The sentences are both tonally and phonetically (five sentences can vary the meaning depend on tone, and the others are not vary in anyways) of short conversation. All the samples were recorded with mono channel, 16 kHz sampling rate and 16 bits resolution. In this experiment, the speech data from five male and five female speakers were used as training set, and other five is used as speaker-independent testing set. For speaker-dependent testing will use the remaindered data from speaker of training set. All the experiments in this paper were used the same sample data on database.

4.2. Recognition Results

The results of each substitution system are shown in “Table 2:”. The computing of tone recognition rate is observed by recognizing tones associated with tone mapping after the codebook allowed a target sentence to carries different tones. Otherwise we will assume that the tone recognition rate is 100%. However, the results of tone recognition that shown in “Table 2:”, are considered only when, tone recognition was enabled. Since, the standard computing recognition rate in [7] is used to obtain the results of base syllable recognition. And the result of tonal syllable recognition (proposed system), are obtained by multiply that of both base syllable and tone recognitions. The tone recognition of sequential detection has been always classified into five tone types (25 left-tone-dependent models) for a syllable. Also, the proposed system has been classified into two tone types only (10 left-tone-dependent models), depending on tone mapping is allowed. While, the joint detection has been recognized very large number of tonal syllable models. In addition, the experiments have also been evaluated the recognition results in cases of applying joint detection, sequential detection, and the proposed method. The result of each case is illustrated in “Table 3:”. As the results, the

proposed system can obtained 84.63% correction for speaker-dependent and 78.17% correction for speaker-independent, which is the higher than that of both joint detection and sequential detection. Where, sequential detection can obtain the recognition rate of, respectively, 84.04% and 77.58% for speaker-dependent and speaker-independent. And joint detection can obtain the recognition rate of, respectively, 72.97% and 68.72%.

The results have shown that, joint detection is obtained lower recognition rate than that of both sequential detection and the proposed system. Consequently, some confusing observation of tone recognition can decrease overall performance of tonal syllable recognition in every times instant. In case of sequential detection and proposed method, the results of tone recognition can be affected to tonal syllable recognition, only when the system allowed to recognizing tone, especially for the proposed method, the probability that tone recognition will be applied during tonal syllable recognition, it is smaller than that of sequential detection method. Although, the results of the proposed method can not be superior to that of sequential detection clearly, but in term of recognition speed it's obtained 25% higher than that of sequential detection. However, the proposed system is required more memories than sequential detection system for a language model.

Table 2: The results of tonal syllable recognition of proposed system

	Speaker-Dependent (% Corr.)	Speaker-Independent (% Corr.)
Tone Recognition	90.52	86.02
Base syllable Recognition	88.79	82.67
Tonal Syllable Recognition	84.63	78.17

Table 3: Comparison the result of proposed system and other baselines system

Recognition Methods	Speaker-Dependent (% Corr.)	Speaker-Independent (% Corr.)
Joint Detection	72.97	68.72
Sequential Detection	84.04	77.58
Proposed Method	84.63	78.17

5. CONCLUSIONS

This paper has introduced a tonal syllable recognition method for continuous speech in tonal language, based on HMM algorithm, by using a language model and specific tone mapping of Lao tonal language to reduce recognition time duration and improve performance of tonal syllable recognition for continuous Lao speech. The result shown that, the performance of proposed system is superior to that of joint detection, while it is similar to that of sequential detection. However, in our experiments, we found that, the recognition speed of proposed system can

be improved up 25% of that of baseline for small codebook.

6. REFERENCES

- [1] E. Maneenoi, V. Ahkuputra, S. Luksaneeyanawin and S. Jitapunkul, "A Study on Acoustic Modeling for Speech Recognition of Predominantly Mono-syllabic Languages", *IEICE Trans. Inf & Syst.*, vol. E87-D, pp. 1146-1163, no.5, May 2004.
- [2] K. Khanthavivone and K. Songwattana, "Tone Recognition Model for Laotian Language Using Pitch Quantization and Hidden Markov Modeling Techniques", *Ladkrabang Engineering Journal*, Vol. 19, No. 4, pp 1-6, December 2002.
- [3] T. Demeechai and K. Makelainen, "Recognition of Syllables in A Tone Language", *Speech Communication*, vol.33, pp.241-254, 2001.
- [4] N. Thubthong and B. Kijisirikul, "Tone Recognition of Continuous Thai Speech Under Tonal Assimilation and Declination Effects Using Half-Tone Model", *Inter. Journal of Uni., Fuzziness and Knowledge-Based Systems*, vol.0, no.0, (1993) 000—000.
- [5] G. Zhang, F. Zheng and W. WU, "Tone Recognition of Chinese Continuous Speech", *ISCSLP*, pp.207-210, 13-25, Beijing, Oct. 2000.
- [6] Lee T., Ching P.C., Chan L.W., Cheng Y.H. and Mak B., "Tone Recognition of Isolated Cantonese syllable", *IEEE Trans, Speech and Audio Process*, 3(3), 204-209, 1995.
- [7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book Version 3.0*, Microsoft Corporation, 1999.
- [8] O. Noulnavong, P. Sisuvhan, B. Paphaphan, B. Sengsuln, S. Sihalaat, and K. Sisawan, *Advance Lao Grammar*, Lao Education Ministry, UNICEF, Copyright ©2003.
- [9] K. Mingbuapha and B. Poomsan Becker, *Lao-English /English-Lao Dictionary*, Copyright ©2003 by P. Publishing, ISBN 1-887521-27-5.

VITAE

Senglathsamy Chanthamenavong was born in Vientiane, Lao P.D.R., in 1979. He received the B.S. Eng. degree in electronic engineering from Department of Electronic Engineering, National University of Laos, Lao P.D.R., in 2001. He is now working toward the M.S. Eng. degree in electrical engineering from Department of Electrical Engineering, Chulalongkorn University, Thailand. His research areas are digital signal processing, speech analysis, speech recognition and spoken dialogue system.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย