การศึกษาหน่วยตามของพยางค์เชิงกลสัทศาสตร์: พื้นฐานสำหรับระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทย

นาย เอกฤทธิ์ มณีน้อย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2546
ISBN 974-17-4171-5

AN ACOUSTIC STUDY OF SYLLABLE RHYMES: A BASIS FOR THAI CONTINUOUS SPEECH RECOGNITION SYSTEM

Mr. Ekkarit Maneenoi

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic year 2003

ISBN 974-17-4171-5

Thesis Title	An Acoustic Study of Syllable Rhymes: A Basis for Thai Continuous Speech Recognition System
By	Mr. Ekkarit Maneenoi
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor Somchai Jitapunkul, Dr.Ing.
Thesis Co-advisor	Assistant Professor Sudaporn Luksaneeyanawin, Ph.D.
	Chularat Tanprasert, Ph.D.

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

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

THESIS COMMITTEE

.....Chairman
(Professor Prasit Prapinmongkolkarn, D.Eng.)

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

.....Thesis Co-advisor
(Assistant Professor Sudaporn Luksaneeyanawin, Ph.D.)

.....Thesis Co-advisor
(Chularat Tanprasert, Ph.D.)

.....Member
(Associate Professor Watit Benjapolakul, D.Eng.)

.....Member
(Assistant Professor Boonserm Kijsirikul, D.Eng.)

เอกฤทธิ์ มณีน้อย, นาย : การศึกษาหน่วยตามของพยางค์เชิงกลสัทศาสตร์: พื้นฐานสำหรับระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทย. (AN ACOUSTIC STUDY OF SYLLABLE RHYMES: A BASIS FOR THAI CONTINUOUS SPEECH RECOGNITION SYSTEM) อ. ที่ปรึกษา : รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล, อ. ที่ปรึกษาร่วม : ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน, ดร. จุพารัตน์ ตันประเสริฐ 250 หน้า. ISBN 974-17-4171-5

วิทยานิพนธ์เล่มนี้มีวัตถุประสงค์ของงานวิจัยเพื่อพัฒนาหน่วยเสียงเชิงกลสัทศาสตร์สำหรับแบบจำลองหน่วยตามพยางค์ภาษาไทย งานวิจัยนี้ทำการศึกษาคุณลักษณะของพยางค์ในภาษาไทยทั้งเชิงกลสัทศาสตร์และระบบเสียงภาษา โครงสร้างของพยางค์ในภาษาไทยมีคุณลักษณะในเชิงกลสัทศาสตร์ที่สระมีผลกระทบอย่างมากต่อความยาวของพยัญชนะตัวสะกด ความสัมพันธ์ดังกล่าวนี้จะเกิดขึ้นระหว่างสระและพยัญชนะตัวสะกดเท่านั้น ส่วนความยาวของพยัญชนะต้นจะไม่มีผลกระทบจากสระ จากคุณลักษณะดังกล่าวสามารถสรุปได้ว่าพยัญชนะตัวสะกดมีความสัมพันธ์กันอย่างมาก ในเชิงระบบเสียงภาษาพยางค์ประกอบด้วยคู่ของหน่วยเริ่มพยางค์และหน่วยตามพยางค์ โดยที่หน่วยเริ่มพยางค์ประกอบด้วยพยัญชนะต้นและส่วนที่เปลี่ยนจากพยัญชนะไปสู่สระ ส่วนหน่วยตามพยางค์ประกอบด้วยสระ พยัญชนะตัวสะกดและวรรณยุกต์ หน่วยเริ่มพยางค์และหน่วยตามพยางค์นอกจากจะมีข้อมูลเชิงบริบทของสระแล้วยังมีการจำลองเชิงภาษาไว้ในระดับพยางค์อีกด้วย ดังนั้นการจำแนกพยางค์ออกเป็นสองส่วนคือหน่วยเริ่มพยางค์และหน่วยตามพยางค์จึงมีความเหมาะสมสำหรับภาษาไทย เนื่องจากงานวิจัยนี้มีวัตถุประสงค์ในการเปรียบเทียบประสิทธิภาพของแบบจำลองหน่วยเสียงประเภทต่างๆ ดังนั้นงานวิจัยนี้จึงไม่พัฒนาระบบรู้จำวรรณยุกต์ด้วย

หน่วยเสียงประเภทต่างๆที่ใช้ในระบบรู้จำเสียงพูดถูกนำมาประเมินผลเปรียบเทียบกับหน่วยเสียงที่นำเสนอในงานวิจัยนี้มีการทดลองจำนวนมากเพื่อที่จะค้นหาหน่วยเสียงที่สามารถจำลองคุณลักษณะเชิงกลสัทศาสตร์ได้เหมาะสมและให้ผลการรู้จำที่ดีที่สุดโดยผลการรู้จำเสียงพูดจากหน่วยเสียงประเภทต่างๆ จะถูกนำมาเสนอและเปรียบเทียบ ฐานข้อมูลเสียงพูดสำหรับใช้ฝึกฝนในงานวิจัยนี้บันทึกจากผู้พูดเพศชายจำนวน 9 คนและเพศหญิงจำนวน 11 คน กลุ่มผู้พูดชุดนี้จะบันทึกเสียงพูดทดสอบแบบขึ้นกับผู้พูดอีกด้วย สำหรับเสียงพูดทดสอบแบบไม่ขึ้นกับผู้พูดจะได้จากการบันทึกเสียงพูดของผู้พูดเพศชายจำนวน 5 คนและเพศหญิงจำนวน 5 คนอีกกลุ่มหนึ่ง ฐานข้อมูลเสียงพูดดังกล่าวนี้ได้รับการออกแบบให้ครอบคลุมหน่วยเริ่มพยางค์และหน่วยตามพยางค์ทั้งหมดที่มีอยู่ในภาษาไทย

จากผลการทดลองแสดงให้เห็นว่าแบบจำลองหน่วยเริ่มพยางค์และหน่วยตามพยางค์มีประสิทธิภาพที่ดีกว่าหน่วยเสียงประเภทอื่นๆ อัตราการรู้จำของแบบจำลองหน่วยเริ่มพยางค์และหน่วยตามพยางค์มีประสิทธิภาพที่ดีกว่าแบบจำลองพื้นฐาน monophone, inter-syllable triphone และ context-dependent Initial-Final ร้อยละ 26.2, 6.4 และ 4.2 สำหรับระบบรู้จำแบบขึ้นกับผู้พูดโดยใช้การจำลองเชิงกลสัทศาสตร์เท่านั้น และร้อยละ 29.7, 6.0 และ 4.2 สำหรับระบบรู้จำแบบขึ้นกับผู้พูดโดยใช้การจำลองเชิงกลสัทศาสตร์และการจำลองเชิงภาษา การใช้การจำลองเชิงภาษาทำให้อัตราการรู้จำของหน่วยเริ่มสูงขึ้นประมาณร้อยละ 16-21 สำหรับระบบรู้จำแบบขึ้นกับผู้พูดและแบบไม่ขึ้นกับผู้พูด นอกจากนี้อัตราการรู้จำของหน่วยตามพยางค์ถูกปรับปรุงขึ้นอย่างมากประมาณร้อยละ 45-47 สำหรับระบบรู้จำแบบขึ้นกับผู้พูดและแบบไม่ขึ้นกับผู้พูดเมื่อมีการใช้การจำลองเชิงภาษา ผลการทดลองแสดงให้เห็นว่าแบบจำลองหน่วยเริ่มพยางค์และหน่วยตามพยางค์มีอัตราการรู้จำที่สูงมาก นอกจากนี้แบบจำลองหน่วยเสียงดังกล่าวยังมีประสิทธิภาพที่ดีในด้านความซับซ้อนอีกด้วย

ภาควิชา	วิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต.....
สาขาวิชา	วิศวกรรมไฟฟ้า	ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา	2546	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....
		ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

##4271830521 : MAJOR ELECTRICAL ENGINEERING
KEYWORD: ONSETS / RHYMES / CONTINUOUS SPEECH RECOGNITION/ THAI SPEECH ANALYSIS / ACOUSTIC MODELING / TONAL LANGUAGE

EKKARIT MANEENOI : THESIS TITLE (AN ACOUSTIC STUDY OF SYLLABLE RHYMES: A BASIS FOR THAI CONTINUOUS SPEECH RECOGNITION SYSTEM) THESIS ADVISOR: ASSOC. PROF. SOMCHAI JITAPUNKUL, Dr.Ing., THESIS COADVISOR: ASST. PROF. SUDAPORN LUKSANEEYANAWIN, Ph.D., CHULARAT TANPRASERT, Ph.D., 250 pp. ISBN 974-17-4171-5.

The objective of this dissertation is to develop a new speech unit on acoustic modeling of the Thai language. The Thai syllables were studied in both acoustical and phonological properties. From the acoustical point of view, in the syllable structure, the final consonant is strongly influenced by the vowel duration. This relationship occurs only between the vowel and the final consonant. In contrast, the initial consonant is not affected by the duration of the vowel. Hence, the vowel and the final consonant are tightly tied while an initial consonant is loosely tied with the vowel in the syllable. From a phonological point of view, a syllable is composed of a pair of an onset and a rhyme unit. The onset consists of an initial consonant and its transition towards the following vowel. Along with the onset, the rhyme is composed of a vowel, a final consonant, and a tone. The onset-rhyme not only includes its contextual information, but also embeds the language modeling at the syllable level. Consequently, the decomposition of the syllable into an onset and rhyme is appropriate to the Thai language. The whole set of Thai syllables can be recognized by identifying onsets and rhymes. This research has the objective to compare the efficiency of the units. Therefore, a tone recognition system is not implemented in this research.

To evaluate the effectiveness of the proposed acoustic model, various conventional speech units used in speech recognition systems have been investigated. Several experiments have been carried out to find the proper speech unit that can accurately create acoustic model and give a higher recognition rate. Results of recognition rates under different acoustic models are given and compared. The speech corpus used for training in this experiment was recorded from 9 male and 11 female speakers. This group of speakers also produced the speaker-dependent test set. In addition, the speaker-independent test set was produced from other 5 male and 5 female speakers. This speech corpus was designed to cover all onset-rhyme units in the Thai language.

Experimental results show that the onset-rhyme model improves on the efficiency of other speech units. The onset-rhyme model improves on the accuracy of the baseline monophone model, the inter-syllable triphone model, and the context-dependent Initial-Final model by nearly 26.2 %, 6.4%, and 4.2 % for the speaker-dependent systems using only an acoustic model, and 29.7 %, 6.0 %, and 4.2 % for the speaker-dependent systems using both acoustic and language model respectively. Using the language model, the onset accuracy is increased by around 16-21 % for both SD and SI systems. In addition, the accuracy of the rhyme is substantially improved by nearly 45-47 % for the SD and SI systems when the language model is applied. The results show that the onset-rhyme models attain a high recognition rate. Moreover, they also give more efficiency in terms of system complexity.

Department	Electrical Engineering	Student's signature.....
Field of study	Electrical Engineering	Advisor's signature.....
Academic year	2003	Co-advisor's signature .....
		Co-advisor's signature .....

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 my study. I would like to express my appreciation to my co-advisor, Asst. Prof. Dr. Sudaporn Luksaneeyanawin. She has provided me interdisciplinary knowledge in the research on speech technology. Also, Dr. Chularat Tanprasert, one of my co-advisors, gives me some useful suggestions and provides information about the scholarship.

Secondly, I would like to show my thankfulness to all of my thesis committee. The chairman, Prof. Dr. Prasit Prapinmongkolkarn, and the other committees, Assoc. Prof. Dr. Watit Benjapolakul and Asst. Prof. Dr. Boonserm Kijsirikul, have given me some valuable comments regarding my work.

Thirdly, I would like to acknowledge Thai Graduate Institute of Science and Technology (TGIST) and Government Research Grant in Research and Development Cooperative Project between Electrical Engineering Department and Private Sector for financial support for the research to me. Also, the Graduate School of Chulalongkorn University provided some grants for my research.

In addition, I would like to thank all of my colleagues and friends at the Center of Excellence in Telecommunication Engineering. They assist and support me in many ways during my study. A very special thankfulness was gave to Dr. Visarut Ahkuputra, who taught me many things in both academic knowledge and social activity.

Finally, I would like to express my deepest gratefulness to my family, who have entirely supported, encouraged, and believed in me without any doubts throughout my study.

Table of Contents

Table of Contents (cont.)

Table of Contents (cont.)

Table of Contents (cont.)

Table of Contents (cont.)

Table of Contents (cont.)

List of Tables

List of Tables (cont.)

Page

List of Tables (cont.)

List of Tables (cont.)

Page

List of Tables (cont.)

Page

List of Tables (cont.)

List of Tables (cont.)

List of Tables (cont.)

List of Figures

List of Figures (cont.)

Page

List of Figures (cont.)

Page

List of Figures (cont.)

Page

List of Figures (cont.)

Page

Chapter 1

Introduction

Speech is one of the most natural ways for human communication. 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.

The research of speech recognition has been continuously developed for the last half century. A number of significant advances in the past two decades 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, psychology, physiology, pattern recognition, computer science, communication, and information technology (Rabiner and Juang, 1993).

Primarily, the speech recognizer processes the input utterance in bottom-up direction. According to the hierarchical model of speech recognition, acoustic features are extracted in acoustic processor from input speech and converted into a phoneme sequence by means of segmentation and pattern recognition (Furui, 2001). In the acoustic matching unit,

performance of the phone recognition relied on selecting types of speech unit. Choice of speech units depends on the type of recognition and on the size of the vocabulary.

Initially, speech recognition system utilized a simple pattern matching technique to recognize isolated utterances. The reference templates were created based upon the word model. Although the word-based approach can handle coarticulatory effect in the model by treating each utterance as a whole, segment boundaries between words in fluent speech are difficult to detect. Moreover, the 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. Compared to word models, subword units reduce the number of parameters, enable cross-word modeling, and facilitate adding new vocabulary. Various types of phone models have been investigated from an independent phone context, a single phone context (left or right context), left and right context (triphones), and generalized triphones. The choice of speech unit is dependent 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. In spite of using traditional context-dependent units such as diphone or triphone employed in the English speech recognizer, Initials and Finals are utilized as a fundamental unit in a Mandarin Chinese dictation machine (Lee, et al., 1997). The Initial comprises the initial consonant of the syllable while the Final consists of the vowel part including possible medial or nasal ending (Lee, et al., 1993). Different syllable structures of English and Chinese Mandarin will result in using different speech units. Thai syllable structure is different from the others, and therefore, it is vital to figure out the proper speech units used in speech recognition for Thai language.

1.1 Continuous Speech Recognition System

The beginning speech recognition system was based on template matching. The simple pattern matching techniques are not applicable to recognition of fluent speech because segment boundaries are difficult to detect. In normal speech, 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. Since a limited number of the reference templates are used in the recognizer, this method is suitable for a small vocabulary speech recognition application.

Word reference template techniques are limited by the number of templates and by capability of handling variability of speech. Changing from using the whole word unit to the subword unit and exploiting statistical technique can overcome disadvantages resulting from a simple matching technique. 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 the natural speech. Each acoustic model can be concatenated in a series to generate a composite model of a continuous speech utterance. A small number of acoustic models and a dictionary are used to construct a compound model for the word. This approach reduces the number of data required to cover a vocabulary set by using a dictionary.

In large-vocabulary continuous speech recognition, input utterance is processed with many kinds of information including lexicon, syntax, semantics, pragmatics, context, and prosodic (Furui, 2001). The lexicon indicates the phoneme structure of words, syntax expresses the grammatical structure, semantics defines the relationship between words as well as the attributes of each word, pragmatics expresses general knowledge concerning the current topics of conversation, context concerns the contextual information, and prosodic represents accent and intonation. These knowledge sources are combined together as shown in Figure 1.1 (Deller, et al., 1993). The system performance depends on what kinds of these knowledge sources are used and how these knowledge sources are rapidly combined to produce the most probable recognition.

The structure of a typical large-vocabulary continuous speech recognition system is shown in Figure 1.2. Initially, a speech signal is converted into a time series of feature parameters in the spectral analysis part. The system predicts a sentence hypothesis based on the current topic, the meaning of words, and the language grammar, and represents a sentence as a sequence of words. This word sequence is then converted into phoneme sequence models, which were typically represented by HMM models. The likelihood or probability of producing the time series of feature parameters from the sequence of phoneme models is computed. Combined with the linguistic likelihood of the hypothesized sequence, the overall likelihood of the uttered sentence was calculated. The likelihood is computed for the other sentence hypotheses, and the sentence with the highest likelihood score is selected as the recognition sentence.

Many state-of-the-art speech recognizers make use of continuous density HMM with Gaussian mixtures for acoustic modeling (Picone, 1996). Other approaches include segmental-based models and neural networks to estimate the acoustic observation likelihoods (Glass, et al., 1999; Hochberg, et al., 1994). The main advantage of continuous density modeling over observation density is that the number of parameters used to model an observation distribution can easily be adapted to the number of available training data (Gauvain and Lamel, 1998). Another disadvantage of discrete and semi-continuous hidden Markov models is that both systems still employ vector quantization technique, which produces the quantization error (Huang, et al., 2001).

1.2 Researches in Thai Speech Recognition

Speech recognition research in Thailand has been conducted for a decade. These works are based on word-based approach and phoneme-based approach. Various techniques, distinctive feature, dynamic time warping, hidden Markov model, neural network, and fuzzy-neural network, were utilized in the researches. A wide variety of vocabulary sets, isolated Thai numerals, isolated Thai words, and polysyllabic Thai words were recognized with these techniques.

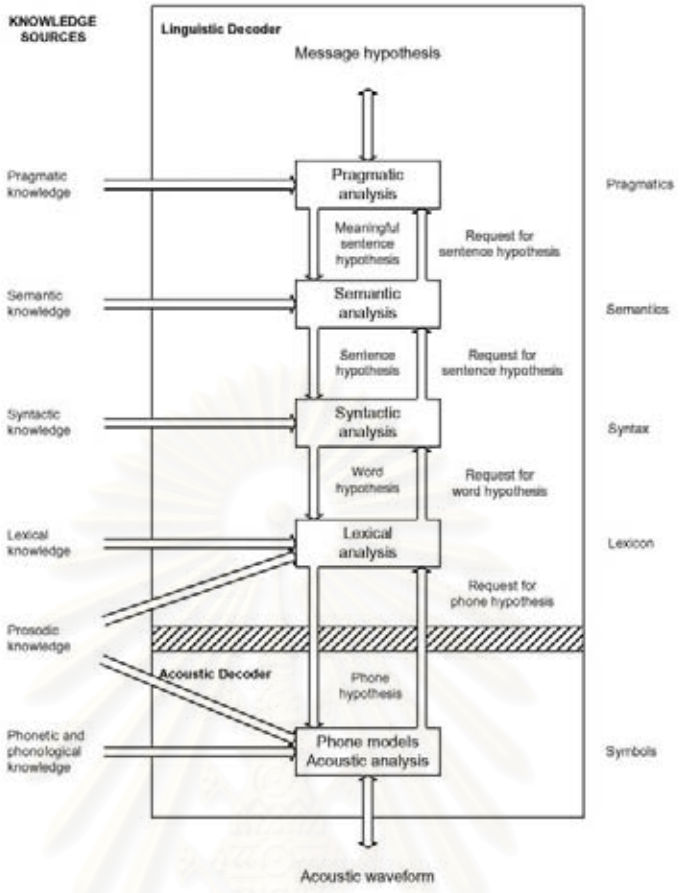


Figure 1.1 The acoustic and linguistic processors (Deller, et al., 1993).

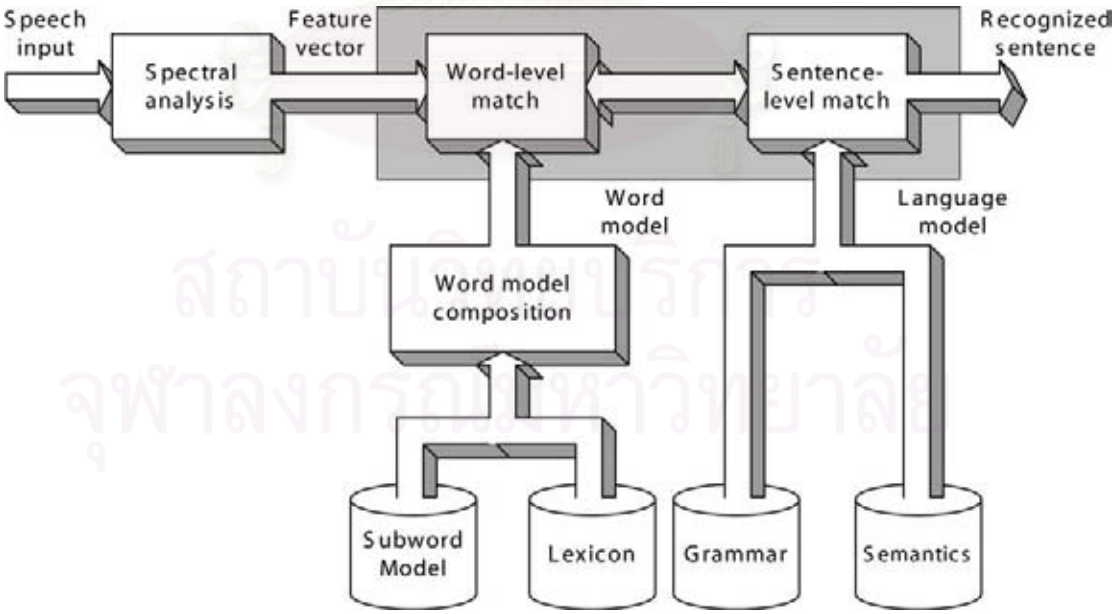


Figure 1.2 A typical continuous speech recognition system (Furui, 2001).

Recognizing unit smaller than word, phoneme-based speech recognition classifies consonants and vowels using acoustic-phonetic features. A study of acoustic characteristics of the vowels /i,a,u/ in Thai and its use in speaker identification (Leelasiriwong, 1991), a Thai speech recognition system based on phonemic distinctive features (Thubthong, 1995), speaker-independent isolated Thai spoken vowel recognition by using spectrum distance measurement and dynamic time warping (Phatrapornnant, 1995), and Thai vowel phoneme recognition using artificial neural networks and hidden Markov models (Maneenoi, 1998; Maneenoi, et al., 1998; Maneenoi, et al., 1999; Maneenoi, et al., 2000) are phoneme-based speech recognition systems. Leelasiriwong (1991) used the first three formant frequencies and the fundamental frequency of the vowels /i,a,u/ for speaker identification. These frequencies, measured from power spectrum, were statistically modeled. The experimental results showed that the fundamental frequency and the formant frequencies are significantly dependent on the sex of speakers. Thubthong (1995) employed the phonemic distinctive feature technique to classify Thai phonemes by their acoustic-phonetic features. Three sets of hypothetical words created from a selected set of phonemes were used in the research. This system is a speaker-dependent system, which was not tested with another group of speakers other than the training speakers. Hence, higher recognition accuracy was obtained by using the same set of training and testing speakers. However, this technique is not practically implemented as a recognition engine because of its inability to cope with varying patterns of speech utterances. Then, it should be used as a post-processing of a recognition system in order to improve the recognition performance. Phatrapornnant (1995) used spectrum distance measurement computed from fast Fourier transform together with dynamic time warping technique to classify 24 Thai vowels uttered in isolated manner. In addition, this system can discriminate 5 Thai tones from 3 Thai vowels /a:/, /i:/, and /u:/ by calculating mean square error with tone's reference formula. Maneenoi (1998) utilized artificial neural networks to classify Thai monophthongs. Nine Thai vowel phonemes are recognized using various feature for example, linear prediction coefficient, LPC derived cepstral coefficient, formant frequency, and spectral intensity. Fundamental frequency and energy were used as feature for extracting a stable voiced portion in the central region of syllable from the entire speech

waveform at which vowel is located. Moreover, spectral characteristics of vowels and vowel durations were study in this research.

Word-based speech recognition models whole word as a single reference template. Thai speech recognition using syllable units is a speaker-dependent system, which was not tested with another group of speakers other than the training speakers (Prathumthan, 1986). Hence, higher recognition accuracy was obtained by using the same set of training and testing speakers. Contrary to the speaker-dependent system, the speaker-independent system is tested with different group of speakers other than group of training speakers. The researches, multispeaker speech recognition system (Thumpothong, 1989) and speaker-independent Thai numeral voice recognition by using dynamic time warping (Pensiri, 1995) employed the dynamic time warping technique. A powerful statistical technique, hidden Markov model, was used in a speaker-independent Thai numeral speech recognition system by hidden Markov model and vector quantization (Areepongsa, 1995) and speaker-independent Thai polysyllabic word recognition system using hidden Markov model (Ahkuputra, 1996). A highly parallel computational technique, neural network, was utilized in speaker-independent Thai numeral speech recognition using LPC and the back propagation neural network (Pornsukjantra, 1996), a modified back propagation algorithm for neural networks (Maneenoi, et al., 1997), and speaker independent Thai polysyllabic word recognition using fuzzy-technique and neural network (Wutiwiwatchai, 1997). All of these researches are speaker-independent system.

Thumpothong (1989) exploited the dynamic programming technique for computing the distance between the test and the reference patterns and employed the K-nearest neighbour (KNN) technique for decision rules in the recognition stage. Pensiri (1995) employed dynamic time warping technique to recognize Thai numeral. Speech parameters were obtained from applying discrete Hartley transform to every frame of voice. The experimental result showed that the recognition accuracy drops when the classified patterns increase. Thus, this technique is inappropriate for recognition a large set of vocabulary. Areepongsa (1995) proposed to use hidden Markov model (HMM) and vector quantization to recognize Thai numeral. This research studied the relationship between the accuracy of speech recognition versus the number of training sets. From the experimental result, the accuracy rate increases

along with the increment of the training data. Ahkuputra (1996) developed an algorithm for speaker independent Thai polysyllabic word recognition by using the hidden Markov model in conjunction with the vector quantization algorithm and the endpoint detection algorithm for syllable endpoint detection and separation. The 70-word vocabulary set including Thai numeral was recognized by these algorithms. Pornsukjantra (1996) conducted a research on recognition of Thai numeral using LPC and the back propagation neural network. A set of single syllable Thai numeral from 0 to 9 and a set of two and three numeral syllables were classified. Time normalization algorithm is required to adjust unequal duration of input data to fit an input of neural network. Wutiwiwatchai (1997) integrated the fuzzy-technique into the conventional neural network to enhance training data. Instead of using fuzzy membership input data and class membership desired-output data during training, the fuzzy membership input data and binary desired-output were used in this research. The syllable detection and tone detection algorithms are used for vocabulary pre-classification in order to reduce the number of vocabularies to be fed to the neural network. The recognition accuracy of the modified input data was improved compared to the recognition using only LPC input data.

The syllable segmentation algorithm was used for segmenting a Thai spoken sentence into a set of syllables. Syllable segmentation algorithm for Thai connected speech employed the energy, band crossing rate, fundamental frequency, and duration of speech signal to detect the syllable boundary (Jittiwarangkul, 1998). The average accuracy of the algorithm is 90 percent tested on a set of ambiguous syllable boundaries. This technique could be applied to a syllable-based speech recognition system.

According to the researches described above, Thai researchers have been accumulated a lot of expertise from isolated word recognition technique, and they can implement some simple speech recognition applications. Voiced astronomical encyclopedia retrieval, BTS sky-train ticketing system (Charnvivit, et al., 1999), and voice-activated web browser (Udompisit and Sothipunchai, 2000) are speech recognition systems based on isolated word model.

Designed for the specific task as described above, the prototype speech recognition systems get speech signal directly from microphone, process the input speech by signal processing algorithm, employ pattern

recognition technique to acoustic features, and respond the result back to users. However, there are still many major drawbacks in these systems that cause unsatisfactory recognition accuracy. Confusing words may produce irresolvable error, which deteriorate the system efficiency. Containing a small number of vocabularies, the isolated speech recognition has a limited capability to deal only with some particular applications. Another problem is that an enormous resource is required for storing and processing the isolated word models when vocabulary size becomes large. In addition, utterance in isolated manner is unsuitable for a wide variety of applications, for example, a friendly user interface in spoken dialogue application. Thus, continuous speech recognition seems to be a good solution for resolving the disadvantages of an isolated word speech recognition system.

The acoustic modeling of onsets was proposed as an acoustic unit for Thai continuous speech recognition (Ahkuputra, 2002). The onset is a subsyllable unit comprising the initial consonant and its transition to the following vowel. To implement a complete large vocabulary speech recognition system for continuous speech, the subsyllable rhyme unit is proposed in this research. The rhyme is composed of the vowel and the optional final consonant. These two subsyllable units make a complete syllable in the speech recogntion system. To evaluate the efficiency of the proposed speech unit, various speech units will be utilized in the continuous speech recognition system. The criteria used to evaluate performance of a speech unit will be elaborated in the next section.

1.3 Selection of Speech Unit

One important issue in developing a speech recognition system is the selection of the speech unit. The choice of speech unit usually is dependent on the size of vocabulary to be recognized and the availability of sufficient training data for constructing effective reference models. Furthermore, efficiency of speech recognition system is relied on the number of speech units. Three criteria, accurate, trainable, and generalized, must be considered in choosing appropriate speech units (Huang, et al., 2001). Firstly, the speech unit should be accurate to represent the acoustic realization that appears in different contexts. Secondly, the unit should be trainable to estimate the parameters of the unit with sufficient data. Thirdly, the unit should be generalized, so that any new word can be derived from a

predefined unit inventory for task-independent speech recognition. A practical challenge is how to select a speech unit, which meets these criteria. Therefore, it is important to select an appropriate speech unit for the Thai language.

1.4 Scope of the Dissertation

In order to develop appropriate recognition units for Thai continuous speech recognition, acoustical properties of the Thai continuous speech have been thoroughly analyzed. This research focuses on the modeling of syllable rhymes consisting of the vowel and the final consonant or the codas. Followings are scopes and goals of this dissertation

- To model acoustic characteristics of Thai syllable rhymes for Thai speech recognition.
- To develop more appropriate recognition units for the modeling of Thai speech recognition in terms of accuracy, trainability, and generalization.
- To provide basic acoustic knowledge for Thai continuous speech recognition.

1.5 Dissertation Outline

Chapter 2 provides a concise introduction to the theory and application of fundamental techniques for speech recognition. In chapter 3, phonological properties and acoustic analysis of the Thai language are described in details to provide basic knowledge of the Thai language. Chapter 4 describes the proposed onset-rhyme model. The acoustic modeling of various speech units with hidden Markov models is also explained in this chapter. Moreover, the construction of the Thai continuous speech recognition system is presented. In chapter 5, system configurations and experimental results are elaborated. In addition, performance comparison between the proposed onset-rhyme model and other speech units in terms of accuracy and complexity is described in this chapter. Finally, chapter 6 discusses and concludes all experimental results. Contributions and suggestions of future works on Thai continuous speech recognition are also detailed in this chapter.

Chapter 2

Fundamental Techniques for Speech Recognition

This chapter provides a concise introduction to the theory and application of fundamental techniques for speech recognition. Signal processing for speech recognition will be described to understand the characteristics of speech signals. Then, theory of the hidden Markov model will be elaborated. Finally, details of the large vocabulary continuous speech recognition system will be explained.

2.1 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 for speech recognition have been reported.

2.1.1 Short-Time Fourier Analysis of Speech Signals

There are two important reasons for analyzing speech signal in the frequency domain (Furui, 2001). The first reason is that speech wave is considered to be reproducible by summing the sinusoidal waves, the amplitude, and phase of which change slowly. The other reason is that the critical features for perceiving speech by the human ears are mainly relied on the spectral information, with the phase information does not usually play a key role.

In order to study spectral properties of speech signal, the concept of short-time Fourier analysis of a signal will be introduced. The standard Fourier representation that is appropriate for periodic, transient, or stationary random signal, is not applicable to the representation of speech signal, whose properties change markedly as a function of time (Rabiner and Schafer, 1978). However, the short-time analysis principle is a valid approach to speech processing. A time interval on the order of 10 to 30

millisecond, which is assumed to change relatively slowly with time, is suitable for applying the short-time analysis. Furthermore, short-time Fourier analysis 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.

2.1.2 LPC Analysis for Speech Signals

Linear Predictive Coding (LPC) can provide a complete description for a speech prediction model at the vocal tract level. The basic idea underlying LPC is that each speech sample, x_t, 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_t, which is defined by

$$e_t = x_t - \sum_{i=1}^{p} a_i x_{t-i} \tag{2.1}$$

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

The LPC coefficients, which minimize the mean-square prediction error, can be obtained by setting the partial derivative of the mean-square prediction error, with respect to each a_i, equal to zero. By minimizing the prediction error, the LPC technique models the spectrum as a smooth spectrum of an order-p all-pole filter (Rabiner and Schafer, 1978). 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. The solution of this equation can be achieved with various algorithms (Rabiner and Schafer, 1978; Deller, et al., 1993). However, three different approaches, the covariance method and the autocorrelation method, have been mainly used for this task (Rabiner and Juang, 1993; Huang, et al., 2001).

2.1.3 Cepstral Analysis for Speech Signals

The basic model of speech production can be considered as a vocal tract filter excited by a periodic excitation function for voiced speech or white noise for unvoiced speech (Vuuren, 1998). 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 spectra in the frequency domain. Short-time spectra comprise a slowly varying envelope corresponding to the vocal tract filter and a rapidly varying fine structure corresponding to the periodic excitation frequency and its harmonics (Rabiner and Schafer, 1978). In the frequency domain, the product of the excitation and filter spectra is transformed to the summation of these two spectra 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 (Rabiner and Schafer, 1978).

Cepstral coefficients, which can also be obtained from LPC analysis (Rabiner and Schafer, 1978; Rabiner and Juang, 1993), have been widely used in speech recognition. The cepstral coefficients, c_n, obtained from LPC analysis, can be computed recursively from the LPC coefficients, a_i, as

$$c_n = -a_n - \sum_{k=1}^{n-1} \frac{n-k}{n} \; a_k c_{n-k} \qquad n \geq 1 \tag{2.2}$$

where $a_k = 0$ when $k > p$

A major advantage of the cepstral analysis is that correlation between coefficients is extremely small such that simplified modeling assumption can be applied.

2.1.4 Filterbank Analysis for Speech Signals

Alternatively, the spectral features can be obtained by passing the speech signal through a bank of bandpass filters. One of the main advantages of this approach is that the bandpass can be placed along perceptual frequency scales such as critical band (Dautrich, et al., 1983), bark scale (Ali, et al., 2002), or mel scale (Bu and Church, 2000). The filterbanks are generally triangular, and they are equally spaced along the mel scale, which is defined by

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.3}$$

Obviously, the mel scale is linear 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.

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 as shown in eq. 2.4 , then, the mel-frequency ceptral coefficients (MFCCs) are obtained.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{i\pi}{N}(j-0.5)\right) \tag{2.4}$$

where N is the number of filterbank channels and m_j is the log filterbank amplitude. Since MFCCs give good discrimination, they have been widely used in many speech recognition applications (Mokbel and Chollet, 1995; Vergin, O'Shaughnessy, and Farhat, 1999).

2.1.5 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 (Juang, et al., 1987). 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 is also known as cepstrum liftering. Several weighting functions or lifting windows have been proposed for speech recognition (Juang, et al., 1987; Tokhura, 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

$$w(i) = 1 + \frac{Q}{2}\sin\left(\frac{i\pi}{Q}\right) \qquad i = 1,...,N \tag{2.5}$$

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

$$\hat{c}(i) = c(i)w(i) \qquad i = 1,...,N \tag{2.6}$$

2.1.6 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 (Furui, 1986). 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

$$\frac{\partial}{\partial t}\left[\log\left|S\left(e^{j\omega},t\right)\right|\right]=\sum_{m=-\infty}^{\infty}\frac{\partial c_m(t)}{\partial t}e^{j\omega} \tag{2.7}$$

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 $\partial c_m(t)/\partial t$ is using an orthogonal polynomial fit over a finite length window; that is

$$\frac{\partial c_m(t)}{\partial t}=\Delta c_m(t)\approx\mu\sum_{k=-K}^{K}k c_m(t+k) \tag{2.8}$$

where μ is an appropriate normalization constant and $(2K+1)$ is the number of frames over which the computation is performed (Rabiner and Juang, 1993).

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

$$O_t=\left(\hat{c}_1(t),\hat{c}_2(t),...,\hat{c}_N(t),\Delta c_1(t),\Delta c_2(t),...,\Delta c_N(t)\right) \tag{2.9}$$

where O_t is a vector of $\hat{c}_i(t)$ and $\Delta c_i(t)$ with N components.

2.2 Hidden Markov Model

The hidden Markov model 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. The underlying assumption of the HMM is that the data samples can be well characterized as a parametric random process, and the parameters of the stochastic process can be estimated in a precise and well-defined framework. Speech observation sequences corresponding to an acoustic event can be modeled by traversing an underlying sequence of connected states, each associated with an output distribution. The output distribution and the relative likelihood moving between states are estimated from a number of observation sequences of particular speech unit to be modeled. This is necessary to make speech recognition computationally tractable, and eases the task of decoding a continuous waveform into a discrete set of symbols. The HMM has become one of the most successful statistical methods used in speech recognition, because of few assumptions need to be built into the models, and all model parameters can be efficiently estimated from the training data. Many successful speech recognition systems have employed the HMM approach as a major recognition part. Not only can the HMM be used in speech recognition, but it also can be applied in statistical language modeling, spoken language understanding, machine translation, and so on.

This section briefly outlines the theoretical framework of the HMM by explaining the definition of HMM. Then the essential algorithms needed to estimate the model parameters and decoding are described. All initial discussions are based on the discrete HMM. However, most of the discrete HMM concepts can be extended to the continuous HMM as described succeeding the discrete HMM. This chapter also introduces the terminology, which will be used throughout this thesis.

2.2.1 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.

1) The number of states in the model, 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 $\boldsymbol{S} = \{S_1, S_2, ..., S_N\}$ and q_t respectively.
2) The number of distinct observation symbols per state, M. The observation symbols correspond to the physical output of the system being modeled. The individual symbols is denoted as $\mathbf{V} = \{V_1, V_2, ..., V_M\}$
3) The state transition probability distribution, $\mathbf{A} = \{a_{ij}\}$ where

$$a_{ij} = P\left[q_{t+1} = S_j \middle| q_t = S_i\right], \qquad 1 \le i, j \le N . \tag{2.10}$$

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

$$b_j(k) = P[V_k \text{ at } t | q_t = S_j], \quad 1 \le j \le N \; 1 \le k \le M . \tag{2.11}$$

5) The initial state distribution, $\boldsymbol{\pi} = \{\pi_i\}$, where

$$\pi_i = P[q_1 = S_i], \qquad 1 \le i \le N . \tag{2.12}$$

Since a_{ij}, $b_j(k)$, and π_i are all probabilities, they must satisfy the following properties:

$a_{ij} \ge 0$, $b_j(k) \ge 0$, $\pi_i \ge 0$ for all i, j, k

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2.13}$$

$$\sum_{k=1}^{M} b_j(k) = 1 \tag{2.14}$$

$$\sum_{i=1}^{N} \pi_i = 1 \tag{2.15}$$

Given appropriate value of N, M, $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$, the HMM can generate an observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$, where each observation $\boldsymbol{O}_t$ is one of the symbols from $\mathbf{V}$, 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, $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$. For convenience, the compact notation is used to represent the complete parameter set of the model

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \tag{2.16}$$

In the first-order hidden Markov model, there are two assumptions. The first is the Markov assumption for the Markov chain.

$$P(q_t | q_1^{t-1}) = P(q_t | q_{t-1}) \tag{2.17}$$

where q_1^{t-1} represents the state sequence $q_1, q_2, \ldots, q_{t-1}$. At each observation time t, a new state is entered based on the transitional probability, which only depends on the previous state. The transition may allow the process to remain in the previous state. The second is the output-independence assumption.

$$P(\boldsymbol{O}_t | \boldsymbol{O}_1^{t-1}, q_1^t) = P(\boldsymbol{O}_t | q_t) \tag{2.18}$$

The output-independence states that the probability that a particular symbol is emitted at time t depended only on the state q_t and is conditionally independent of the past observations. Although these assumptions severely limit the memory of the first-order HMM and may lead to model deficiency, in practice, they reduce the number of free parameters need to be estimated. Furthermore, these assumptions make evaluation, decoding, and learning feasible and efficient without significantly affecting the modeling capability.

2.2.2 Observation Density Functions

The observation density functions have to model the distribution of the feature vector for the different parts in data. These distributions are estimated from large amounts of training data. The most frequently distributions are listed below.

2.2.2.1 Discrete Density Functions

This type of density modeling requires that the multidimensional continuous observations be quantized into a number of symbols. Each state now has a discrete distribution that gives the probability of each symbol for that state. The discrete symbols are normally generated by a vector quantizer, which assigns a discrete symbol to each observation vector by choosing the nearest example from a small codebook of reference vector. This is implicitly dealt with the choice of distance metric for the clustering procedure in the vector quantization. The Euclidian distance measure, for instance, is used in the k-means clustering algorithm. In order to reduce the quantization distortion for large observation vectors, the multiple independent codebooks for vector quantization were introduced. All components were assumed independent and their probabilities were simply multiplied to give the probability of the component vector.

2.2.2.2 Continuous Density Functions

In this case, the observation probability distribution in state j, $b_j(\boldsymbol{O}_t)$, is a general parametric distribution of a predetermined form. The generalized method to continuous output density functions requires that the probability density functions be strictly log concave. The re-estimation algorithm can be extended to various types of elliptically symmetric density functions. The rationale of continuous density function is that the continuous observations can be directly modeled without quantization. However, the choice of different density functions to model a given observation largely depends on the characteristics of observations. In addition, a single continuous probability density function associated with each state is usually inadequate to model complicated observations, then, finite mixture components are required.

2.2.3 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:

2.2.3.1 The Evaluation Problem

Given the observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$, and the model $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, how to compute $P(\mathbf{O}|\lambda)$, the probability that the observation sequence was 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.

2.2.3.2 The Decoding Problem

Given the observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$, and the model $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, what the most likely state sequence $Q = q_1, q_2, \ldots, q_T$ 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.

2.2.3.3 The Estimation Problem

Given the observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$, how to adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ to maximize $P(\mathbf{O}|\lambda)$. 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.

Formal mathematical solutions to these problems will be presented in the followings sections. The three problems are closely related under the same probabilistic framework.

2.2.4 Solutions to the Three Basic Problems of HMM

2.2.4.1 Solution to the Evaluation Problem

To calculate the probability of an observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$ given the model λ, $P(\mathbf{O}|\lambda)$, the most straightforward way is to enumerate every possible state sequence of length T (the number of observations). For every fixed state sequence

$$Q = q_1, q_2, \ldots, q_T \tag{2.19}$$

where q_1 is the initial state. The probability of the observation sequence $\mathbf{O}$ for this state sequence is

$$P(\mathbf{O}|Q,\lambda) = \prod_{t=1}^{T} P(\boldsymbol{O}_t|q_t,\lambda) \tag{2.20}$$

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

$$P(\mathbf{O}|Q,\lambda) = b_{q_1}(\boldsymbol{O}_1) b_{q2}(\boldsymbol{O}_2) \cdots b_{qT}(\boldsymbol{O}_T) \tag{2.21}$$

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

$$P(Q|\lambda) = P(q_1|\lambda) \prod_{t=2}^{T} P(q_t|q_{t-1},\lambda) \tag{2.22}$$

$$= \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \tag{2.23}$$

$$= a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T} \tag{2.24}$$

where $a_{q_0 q_1}$ denotes π_{q_1} for simplicity.

The joint probability of $\mathbf{O}$ and Q, which $\mathbf{O}$ and Q occur simultaneously, is simply the product of the above two terms

$$P(\mathbf{O}, Q|\lambda) = P(\mathbf{O}|Q,\lambda) P(Q,\lambda) \tag{2.25}$$

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

$$P(\mathbf{O}|\lambda) = \sum_{all\ Q} P(\mathbf{O}, Q|\lambda) P(Q,\lambda) \tag{2.26}$$

$$= \sum_{all\ Q} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{O}_t) \tag{2.27}$$

The interpretation of the computation in the above equation is the following. A transition starts from an initial state q_1 with probability $a_{q_0 q_1}$, and generates the symbol $\boldsymbol{O}_1$ in this state with probability $b_{q_1}(\boldsymbol{O}_1)$. Then, a transition is made from the initial state $q_1$to state q_2 with transition probability $a_{q_1 q_2}$, and generates the symbol $\boldsymbol{O}_2$ with output probability $b_{q_2}(\boldsymbol{O}_2)$ 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 $b_{q_T}(\boldsymbol{O}_T)$ generating symbol $\boldsymbol{O}_T$ is reached.

The computation of $P(\mathbf{O}|\lambda)$, according to its direct definition () involves on the order of $O(N^T)$ calculations. At every time $t = 1,2,\ldots,T$, 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.

2.2.4.1.1 The Forward Procedure

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

$$\alpha_i(t) = P(\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_t, q_t = S_i|\lambda) \tag{2.28}$$

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

Forward algorithm

Initialization:

$$\alpha_i(1) = \pi_i b_i(\boldsymbol{O}_1), \qquad 1 \le i \le N \tag{2.29}$$

Induction:

$$\alpha_j(t+1) = \left[\sum_{i=1}^{N} \alpha_j(t) a_{ij}\right] b_j(\boldsymbol{O}_{t+1}), \qquad 1 \le t \le T-1$$

$$1 \le j \le N \tag{2.30}$$

Termination:

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_i(T). \tag{2.31}$$

In the first step, the forward probabilities are initiated as the joint probability of state S_i and initial observation $\boldsymbol{O}_1$. The induction step, which is the most important forward calculation, is illustrated in Figure 2.1. This figure shows how state S_j can be reached at time $t+1$ from the N possible states, S_i, $1 \le i \le N$, at time t. Since $\alpha_i(t)$ is the probability of joint event that $\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_t$ are observed, and the state at time t is S_i, the product $\alpha_i(t)a_{ij}$ is then the probability of joint event that $\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_t$ are observed, and state S_j is reached at time $t+1$ via state S_i at time t. Summing this product over all the N possible states S_i, $1 \le i \le N$ at time t results in the probability of S_j at time $t+1$ through all the previous partial observations. By multiplying the summed quantity by the probability $b_j(\boldsymbol{O}_{t+1})$, $\alpha_j(t+1)$, the probability of the new observation sequence $\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_t, \boldsymbol{O}_{t+1}$, is obtained in state j. The computation of the induction step is performed for all state j, $1 \le j \le N$, for a given t. This computation is then iterated for $t = 1,2,\ldots,T-1$. Finally, the termination step gives the desired calculation of $P(\mathbf{O}|\lambda)$ as the sum of the terminal forward variables $\alpha_i(T)$.

The computation in the calculation of $\alpha_j(t)$ 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 2.2. 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 $\alpha_i(1)$, $1 \le i \le N$, is calculated. At time $t = 2,3,\ldots,T$, the only values of $\alpha_j(t)$, $1 \le j \le N$, are needed to compute. Each calculation of $\alpha_j(t)$ involves only N previous values of $\alpha_i(t-1)$, because each of N grid point is reached from the same N grid points at the previous time slot.

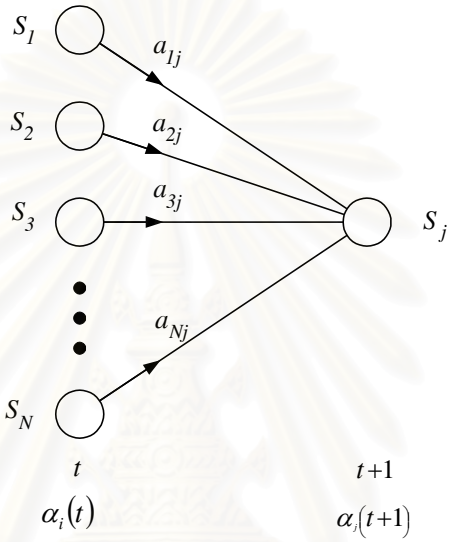


Figure 2.1. The sequence of operations required for the computation of the forward variable $\alpha_j(t+1)$

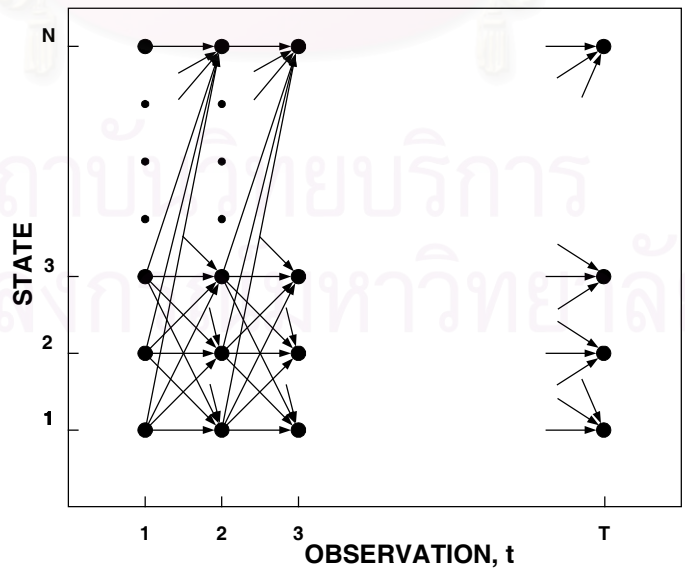


Figure 2.2. Implementation of the computation of $\alpha_i(t)$ in terms of a lattice of observations t and state i

2.2.4.1.2 The Backward Procedure

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

$$\beta_i(t) = P(\boldsymbol{O}_{t+1}, \boldsymbol{O}_{t+2}, \ldots, \boldsymbol{O}_T | q_t = S_i, \lambda) \tag{2.32}$$

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:

The initialization arbitrarily defines $\beta_j(T)$ to be 1 for all i. In order to be in state S_i at time t, and to account for the rest observation sequence, a transition from state 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 $\boldsymbol{O}_{t+1}$ in state S_j (the $b_j(\boldsymbol{O}_{t+1})$ term), and then accounts for the remaining partial observation sequence from state S_j (the $\beta_j(t+1)$ term).

Backward algorithm

Initialization:

$$\beta_i(T) = 1, \qquad 1 \le i \le N \tag{2.33}$$

Induction:

$$\beta_j(t) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{O}_{t+1}) \beta_j(t+1),$$
$$t = T-1, T-2, \ldots, 1, \quad 1 \le j \le N \tag{2.34}$$

The computational complexity of $\beta_j(t)$ is similar to that of $\alpha_i(t)$, which also produces a lattice with observation length and state number. The induction step is illustrated in Figure 2.3.

As mentioned above, both the forward and backward procedures can be applied to compute $P(\mathbf{O}|\lambda)$ 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.

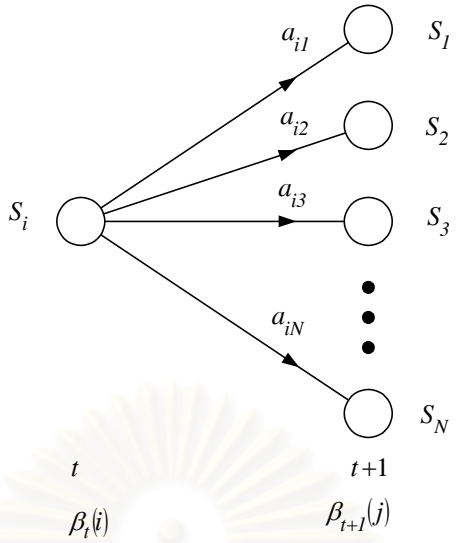


Figure 2.3. The sequence of operations required for the computation of the backward variable $\alpha_i(t)$

2.2.4.2 Solution to the Decoding Problem

The hidden part of HMM, which is the state sequence, cannot be uncovered, but can be interpreted in some meaningful ways. A typical use of the recovered state sequence is to learn about the structure of the model, and to get average statistics within individual states. There are several possible ways to find the optimal state sequence associated with the given observation sequence. One possible optimality criterion is to choose the states q_t, which are in the best path with the highest probability. A formal technique for finding this single best state sequence is called the Viterbi algorithm, which is very similar to the DTW algorithm.

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

$$\gamma_i(t) = P(q_t = S_i | \mathbf{O}, \lambda) \tag{2.35}$$

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

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{P(\mathbf{O}|\lambda)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{i=1}^{N} \alpha_i(t)\beta_i(t)} \tag{2.36}$$

$\alpha_i(t)$ accounts for the partial observation sequence $\boldsymbol{O}_1, \boldsymbol{O}_2, \cdots, \boldsymbol{O}_t$ and the state S_i at time t, while $\beta_i(t)$ accounts for the remainder of the observation

sequence $\boldsymbol{O}_{t+1},\boldsymbol{O}_{t+2},\ldots,\boldsymbol{O}_T$ and the state S_i at time t. The normalization factor, $P(\mathbf{O}|\lambda)$, makes $\gamma_i(t)$, a probability measure so that

$$\sum_{i=1}^{N}\gamma_i(t)=1. \tag{2.37}$$

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

$$q_t=\arg\max_{1\le i\le N}[\gamma_i(t)], \qquad 1\le t\le T \tag{2.38}$$

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

2.2.4.2.1 The Viterbi Algorithm

To find the single best state sequence, $Q=\{q_1,q_2,\ldots,q_T\}$, for the given observation sequence $\mathbf{O}=\{\boldsymbol{O}_1,\boldsymbol{O}_2,\ldots,\boldsymbol{O}_T\}$, the quantity $\delta_t(i)$ is needed to define

$$\delta_i(t)=\max_{q_1,q_2,\ldots,q_{t-1}}P(q_1,q_2,\ldots,q_t=i,\boldsymbol{O}_1,\boldsymbol{O}_2,\ldots,\boldsymbol{O}_t|\lambda) \tag{2.39}$$

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 state S_i. By induction, the Eq. (2.39) becomes

$$\delta_j(t+1)=\left[\max_i\delta_i(t)a_{ij}\right]\cdot b_j(\boldsymbol{O}_{t+1}) \tag{2.40}$$

Viterbi algorithm

Initialization:

$$\delta_i(1) = \pi_i b_i(\boldsymbol{O}_1), \qquad 1 \le i \le N \qquad (2.41)$$

$$\psi_i(1) = 0 \qquad (2.42)$$

Induction:

$$\delta_j(t) = \max_{1 \le i \le N}\left(\delta_i(t-1)a_{ij}\right)b_j(\boldsymbol{O}_t), \qquad 2 \le t \le T, \quad 1 \le j \le N \qquad (2.43)$$

$$\psi_j(t) = \arg\max_{1 \le i \le N}\left(\delta_i(t-1)a_{ij}\right), \qquad 2 \le t \le T, \quad 1 \le j \le N \qquad (2.44)$$

Termination:

$$P^* = \max_{1 \le i \le N}\left[\delta_i(T)\right] \qquad (2.45)$$

$$q_T^* = \arg\max_{1 \le i \le N}\left[\delta_i(T)\right] \qquad (2.46)$$

Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right), \qquad t = T-1, T-2, \ldots, 1 \qquad (2.47)$$

To actually retrieve the state sequence, the array $\psi_j(t)$ is required to keep track of the argument, which maximizes Eq. (2.40) for each t and j. The complete procedure for finding the best state sequence can be started as follows:

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. (2.43), 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.

2.2.4.3 Solution to the Estimation Problem

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

$P(\mathbf{O}|\lambda)$ 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 $\mathcal{P}(X)$ with non-negative coefficients of degree d in variables x_{ij}, $i = 1,2,\ldots,p$, $j = 1,2,\ldots,q_i$, defined over a stochastic domain $\mathcal{D}: x_{ij} \geq 0$, $\sum_{j=1}^{q_i} x_{ij} = 1$, through a series of transformations performed on $\{x_{ij}\}$, was firstly purposed. The transformation is defined as

$$T(x_{ij}) = \frac{x_{ij}\dfrac{\partial \mathcal{P}(X)}{\partial x_{ij}}}{\sum_{j=1}^{q_i} x_{ij}\dfrac{\partial \mathcal{P}(X)}{\partial x_{ij}}} \tag{2.48}$$

and is often referred to a growth transformation of $\mathcal{P}(X)$. A special case of the resstimation 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 reestimated according to Eq. (2.48) in the following. However, the re-estimation formulas for the parameters of a continuous density HMM will be described later.

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

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

$$\xi_{ij}(t) = P\left(q_t = S_i, q_{t+1} = S_j \middle| \mathbf{O}, \lambda\right) \tag{2.49}$$

The sequence of events leading to the conditions required by Eq. (2.49) is illustrated in Figure 2.4. From the definition of the forward and backward variables, $\xi_{ij}(t)$ can be written in the form

$$\xi_{ij}(t) = \frac{\alpha_i(t) a_{ij} b_j(\boldsymbol{O}_{t+1}) \beta_{t+1}(j)}{P(\mathbf{O}|\lambda)} \tag{2.50}$$

$$= \frac{\alpha_i(t) a_{ij} b_j(\boldsymbol{O}_{t+1}) \beta_j(t+1)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i(t) a_{ij} b_j(\boldsymbol{O}_{t+1}) \beta_j(t+1)} \tag{2.51}$$

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

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

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_{ij}(t) \tag{2.52}$$

If $\gamma_i(t)$ 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 state S_i, is obtained. Similarly, summation of $\xi_{ij}(t)$ over t from $t=1$ to $t=T-1$ can be interpreted as the expected number of transitions from state S_i to state S_j. That is

$$\sum_{t=1}^{T-1} \gamma_i(t) = \text{ expected number of transitions from } S_i \tag{2.53}$$

$$\sum_{t=1}^{T-1} \xi_{ij}(t) = \text{ expected number of transitions from } S_i \text{ to state } S_j \tag{2.54}$$

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 $\boldsymbol{\pi}$ are

$$\bar{\pi}_i = \text{expected frequency in state } S_i \text{ at time } \gamma_i(1) \tag{2.55}$$

$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \tag{2.56}$$

$$= \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \tag{2.57}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } S_j \text{ and observing symbol } v_k}{\text{expected number of times in state } S_j} \tag{2.58}$$

$$= \frac{\sum_{t=1}^{T} \gamma_j(t)}{\sum_{t=1}^{T} \gamma_j(t)} \quad \text{s.t. } O_t = v_k \tag{2.59}$$

From Eq. (2.55) to (2.59), 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 $P(\mathbf{O}|\bar{\lambda}) \geq P(\mathbf{O}|\lambda)$.

Thus, if $\bar{\lambda}$ is iteratively used to replace λ and repeats until the above re-estimation calculation, $P(\mathbf{O}|\lambda)$ 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.

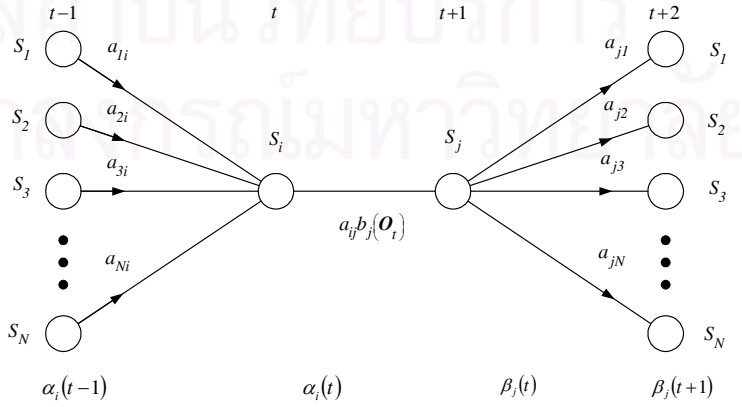


Figure 2.4. The sequence of operations required for the computation of the joint event that the system is in state S_i at time t and state S_j at time $t+1$

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 K observation sequences denoted as

$$\mathbf{O} = \left[\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, \ldots, \mathbf{O}^{(k)}\right] \tag{2.60}$$

where $\mathbf{O}^{(k)} = \left\{\boldsymbol{O}_1^{(k)}, \boldsymbol{O}_2^{(k)}, \ldots, \boldsymbol{O}_{T_k}^{(k)}\right\}$ is the k^{th} observation sequence. Assuming that observation sequences are independent of each other, the parameter estimations of HMM is then based on the maximization of

$$P\left(\mathbf{O}\middle|\lambda\right) = \prod_{k=1}^{K} P\left(\mathbf{O}^{(k)}\middle|\lambda\right) \tag{2.61}$$

$$= \prod_{k=1}^{K} P_k \tag{2.62}$$

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:

$$\overline{a}_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_i^k(t) a_{ij} b_j\left(\boldsymbol{O}_{t+1}^{(k)}\right) \beta_j^k(t+1)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_i^k(t) \beta_i^k(t)} \tag{2.63}$$

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

$$\overline{b}_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{\substack{t=1 \\ s.t.\ O_t = v_l}}^{T_k - 1} \alpha_i^k(t) \beta_j^k(t)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_i^k(t) \beta_i^k(t)} \tag{2.64}$$

2.2.5 Continuous Density Hidden Markov Model

If the observation does not come from a finite set, but from a continuous space, the discrete output distribution discuss in the previous sections can

be extended to the continuous output probability density function. 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 section 3.4.3.1 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 was 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.

2.2.5.1 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

$$\mathcal{N}(\boldsymbol{O};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{O}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{O}-\boldsymbol{\mu})} \qquad (2.65)$$

where n is the dimensionality of the observation vector $\boldsymbol{O}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix respectively. The advantage of normal distributions is that the parameters of Gaussain can be easily and reliably estimated from a large number of data. In order to obtain more accurate approximations, Gaussian mixtures were 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

$$b_j(\boldsymbol{O}_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\boldsymbol{O}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{2.66}$$

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

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad 1 \le j \le N \tag{2.67}$$

$$c_{jm} \ge 0, \qquad 1 \le j \le N,\ 1 \le m \le M \tag{2.68}$$

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

$$P(\mathbf{O}|\lambda) = \sum_{all\ Q} P(\mathbf{O}, Q|\lambda) \tag{2.69}$$

$$= \sum_{all\ Q} P(Q|\lambda) P(\mathbf{O}|Q, \lambda) \tag{2.70}$$

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

$$\mathcal{Q}(\lambda, \bar{\lambda}) = \frac{1}{P(\mathbf{O}|\lambda)} \sum_{all\ S} P(\mathbf{O}, Q|\lambda) \log P(\mathbf{O}, Q|\bar{\lambda}) \tag{2.71}$$

The mathematical derivation of the re-estimation algorithm of the continuous probability density functions is described in Appendix . By using an auxiliary $\mathcal{Q}$-function, reestimated HMM parameters for the multimodal Gaussian distributions are

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{jm}(t)} \tag{2.72}$$

$$\bar{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \cdot \boldsymbol{O}_t}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{2.73}$$

$$\overline{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \cdot (\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{2.74}$$

where prime denotes vector transpose and $\gamma_{jm}(t)$ is the probability of being in state j at time t with the m^{th} mixture component for $\boldsymbol{O}_t$

$$\gamma_{jm}(t) = \left[\frac{\alpha_j(t)\beta_j(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \right] \left[\frac{c_{jm}\mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sum_{m=1}^{M} c_{jm}\mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \right] \tag{2.75}$$

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.

2.2.6 Hidden Markov Model for Speech Recognition

2.2.6.1 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 section 2.2.4.1; 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

$$a_{11} = 0 \text{ and } a_{Ni} = 0 \quad \forall i \tag{2.76}$$

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 call null or tee models (Young, et al., 1999.). A model with entry and exit states is depicted in Figure 2.5 and a tee model is shown in Figure 2.6. Using tee models and non-emitting entry and exit states, a series of HMMs, with tee model between word, 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:

$$\alpha_1^{(q)}(1) = \begin{cases} 1 & q = 1 \\ \alpha_1^{(q)}(1) a_{1N_q}^{(q-1)} & otherwise \end{cases} \tag{2.77}$$

$$\alpha_j^{(q)}(1) = \alpha_1^{(q)}(1) a_{1j}^{(q)} b_1^{(q)}(\boldsymbol{O}_t) \tag{2.78}$$

$$\alpha_{N_q}^{(q)}(1) = \sum_{i=2}^{N_q - 1} \alpha_i^{(q)}(1) a_{iN_q}^{(q)} \tag{2.79}$$

Recursion:

$$\alpha_1^{(q)}(t) = \begin{cases} 0 & q = 1 \\ \alpha_{N_{q-1}}^{(q-1)}(t-1) + \alpha_1^{(q-1)}(t) a_{1N_{q-1}}^{(q-1)} & otherwise \end{cases} \tag{2.80}$$

$$\alpha_j^{(q)}(t) = \left[\alpha_1^{(q)}(t) a_{1j}^{(q)} + \sum_{i=2}^{N_{q-1}} \alpha_i^{(q)}(t-1) a_{ij}^{(q)} \right] b_j^{(q)}(\boldsymbol{O}_t) \tag{2.81}$$

$$\alpha_{N_q}^{(q)}(t) = \sum_{i=2}^{N_q - 1} \alpha_i^{(q)}(t) a_{iN_q}^{(q)} \tag{2.82}$$

The corresponding backward equations are:

Initialization:

$$\beta_{N_q}^{(q)}(T) = \begin{cases} 1 & q = 1 \\ \beta_{N_{q+1}}^{(q+1)}(T) a_{1N_q+1}^{(q+1)} & otherwise \end{cases} \tag{2.83}$$

$$\beta_i^{(q)}(T) = \beta_{N_q}^{(q)}(T) a_{iN_q}^{(q)} \tag{2.84}$$

$$\beta_1^{(q)}(T) = \sum_{j=2}^{N_q-1} \beta_j^{(q)}(T) a_{1j}^{(q)} b_j^q(\boldsymbol{O}_T) \tag{2.85}$$

Recursion:

$$\beta_{N_q}^{(q)}(t) = \begin{cases} 0 & q = 1 \\ \beta_1^{(q+1)}(t+1) + \beta_{N_{q+1}}^{(q+1)}(t) a_{1N_{q+1}}^{(q+1)} & otherwise \end{cases} \tag{2.86}$$

$$\beta_i^{(q)}(t) = \beta_{N_q}^{(q)}(t) a_{iN_q}^{(q)} + \sum_{j=2}^{N_q-1} \beta_j^{(q)}(t+1) a_{ij}^{(q)} b_j^{(q)}(\boldsymbol{O}_{t+1}) \tag{2.87}$$

$$\beta_1^{(q)}(t) = \sum_{j=2}^{N_q-1} \beta_j^{(q)}(t) a_{1j}^{(q)} b_j^{(q)}(\boldsymbol{O}_t) \tag{2.88}$$

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. transition 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:

$$a_{ij}'^{(q)} = \frac{\sum_{t=1}^{T-1} \alpha_i^{(q)}(t) a_{ij}^{(q)} b_j^{(q)}(\boldsymbol{O}_{t+1}) \beta_j^{(q)}(t+1)}{\sum_{t=1}^{T-1} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.89}$$

$$a_{1j}'^{(q)} = \frac{\sum_{t=1}^{T} \alpha_1^{(q)}(t) a_{1j}^{(q)} b_j^{(q)}(\boldsymbol{O}_t) \beta_j^{(q)}(t)}{\sum_{t=1}^{T} \alpha_1^{(q)}(t) \beta_1^{(q)}(t) + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q)}(t)} \tag{2.90}$$

$$a_{iN_q}'^{(q)} = \frac{\sum_{t=1}^{T} \alpha_i^{(q)}(t) a_{iN_q}^{(q)} \beta_{N_q}^{(q)}(t)}{\sum_{t=1}^{T} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.91}$$

$$a_{1N_q}'^{(q)} = \frac{\sum_{t=1}^{T} \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t)}{\sum_{t=1}^{T} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t) \tag{2.92}$$

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. (2.72)-(2.75) do not change once the modifications have been made to the forward and backward probabilities.

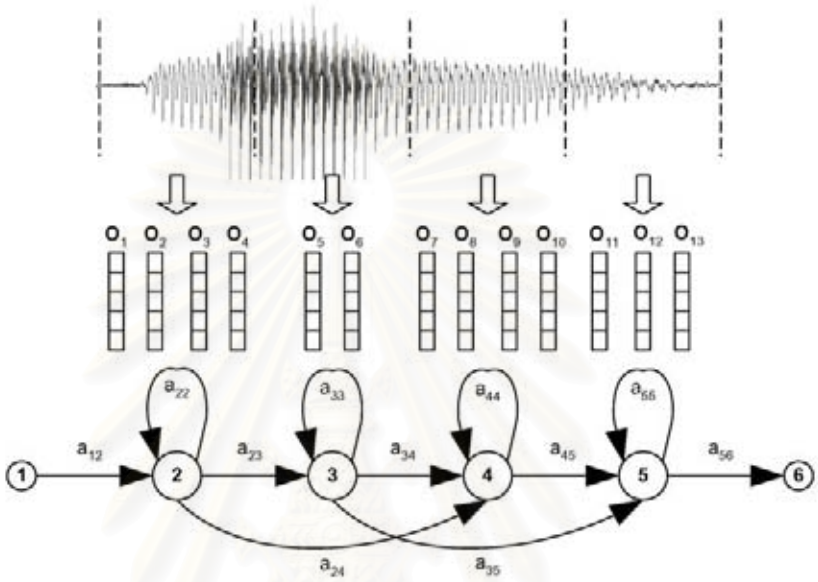


Figure 2.5. HMM with non-emitting entry and exit states

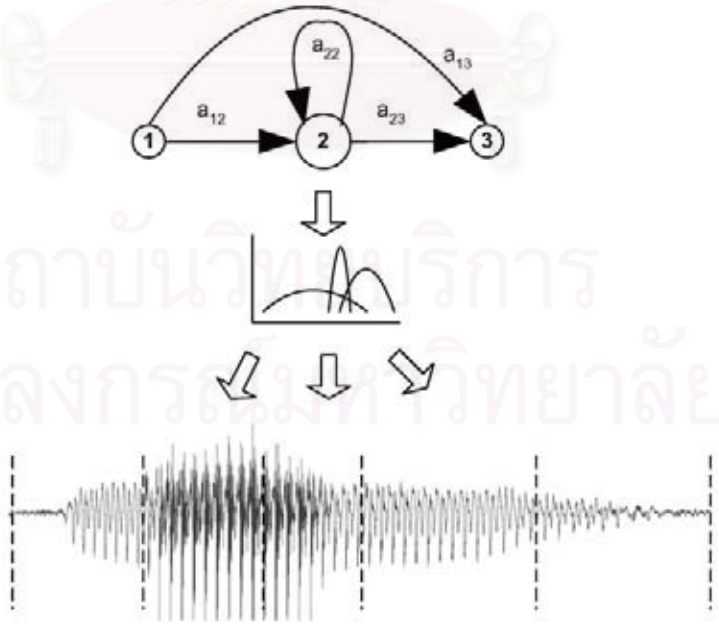


Figure 2.6. Tee model HMM

2.2.6.2 Multiple Observation Sequence

In a complex large vocabulary speech recognition system, there may be literally thousands of models representing context-dependent subword units or segmental subword 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. 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 $\mathbf{P}(\mathbf{O}|\lambda)$ 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

$$a_{ij}'^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(q)}(t)a_{ij}^{(q)}b_j^{(q)}(\boldsymbol{O}_{t+1})\beta_j^{(q)}(t+1)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(q)}(t)\beta_i^{(q)}(t)} \tag{2.93}$$

$$a_{1j}'^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_1^{(q)}(t)a_{1j}^{(q)}b_j^{(q)}(\boldsymbol{O}_t)\beta_j^{(q)}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_1^{(q)}(t)\beta_1^{(q)}(t)+\alpha_1^{(q)}(t)a_{1N_q}^{(q)}\beta_1^{(q)}(t)} \tag{2.94}$$

$$a_{iN_q}'^{(q)} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)}(t)a_{iN_q}^{(q)}\beta_{N_q}^{(q)}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)}(t)\beta_i^{(q)}(t)} \tag{2.95}$$

$$a'^{(q)}_{1N_q} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t) \tag{2.96}$$

$$\gamma_{jm}^{(q)}(t) = \left[\frac{\alpha_j^{(q)}(t) \beta_j^{(q)}(t)}{P_r} \right] \left[\frac{c_{jm} b_{jm}^{(q)}(\boldsymbol{O}_t)}{b_j^{(q)}(\boldsymbol{O}_t)} \right] \tag{2.97}$$

$$c'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{jm}^{(q)}(t)} \tag{2.98}$$

$$\boldsymbol{\mu}'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t) \cdot \boldsymbol{O}_t}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)} \tag{2.99}$$

$$\boldsymbol{\Sigma}'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t) \cdot (\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})'}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)} \tag{2.100}$$

The implementation of these equations can be made with attention to some cancellations within the terms. In particular, the recursion for $\alpha_j^{(q)}(t)$ contains the term $b_j^{(q)}(\boldsymbol{O}_t)$ within it, which is also in the denominator of the formula for $\gamma_{jm}^{(q)}(t)$. The variable $U_j^{(q)}(t)$ is defined as

$$U_j^{(q)}(t) = \begin{cases} \alpha_1^{(q)}(t) a_{1j}^{(q)} & \text{if } t = 1 \\ \alpha_1^{(q)}(t) a_{1j}^{(q)} + \sum_{i=2}^{N_q - 1} \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q)}(t) & \text{otherwise} \end{cases} \tag{2.101}$$

to represent $\alpha_j^{(q)}(t)$ without $b_j^{(q)}(\boldsymbol{O}_t)$ term. The computation of this latter term is cancelled entirely, giving

$$\gamma_{jm}^{(q)}(t) = \frac{1}{P_r} U_j^{(q)}(t) \beta_j^{(q)}(t) c_{jm} b_{jm}^{(q)}(\boldsymbol{O}_t) \tag{2.102}$$

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

$$b'_j(\boldsymbol{O}_t) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_j(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_j(t)} \quad \text{s.t. } s_j \text{emits } \boldsymbol{O}_t \tag{2.103}$$

It is should be noted that this formula has an identical form to the re-estimation equation for mixture weights of Gaussian mixture distributions, if the mixture number m is treated as the index of the emitted observation. Thus, there is a direct correspondence between an M-mixture Gaussian distribution and a discrete distribution of M observation symbols.

2.3 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 view as a problem in communication theory (Shannon, 1948). A spoken of words 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 a is

$$P(w,a) = P(a|w)P(w) = P(w|a)P(a) \tag{2.104}$$

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 2.7 (Furui, 2001).

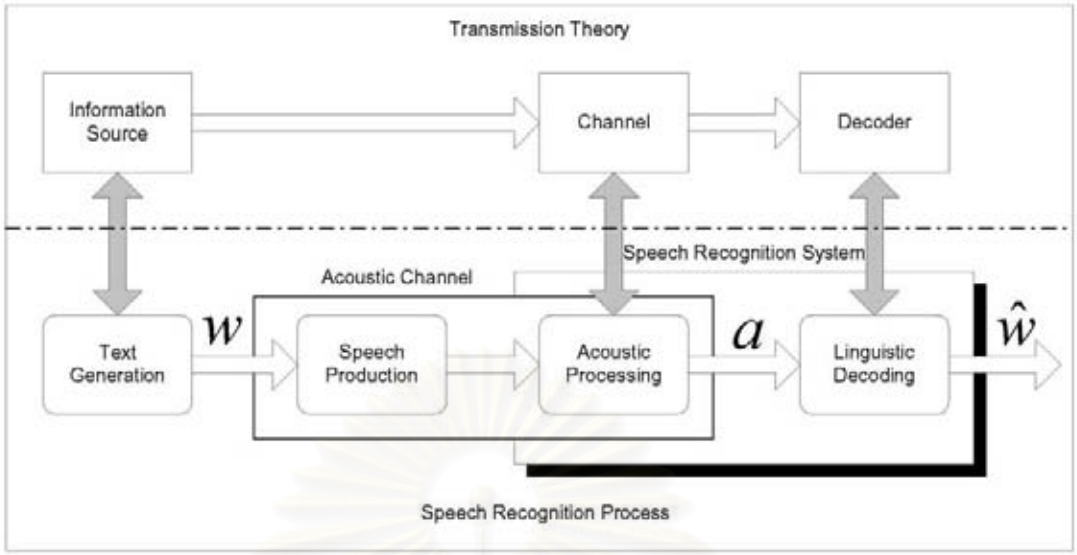


Figure 2.7. Structure of speech recognition according to information theory

The above definition of the speech recognition problem can be viewed as the following practical considerations (Valtchev, 1995):

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 $F(\lambda)$ 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

$$\hat{w} = \arg\max_{w} P(w|a) = \arg\max_{w} P(a|w)\frac{P(w)}{P(a)} \qquad (2.105)$$

2.3.1 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; Steinbiss, et al., 1995; Robinson, 2002; Renals and Hochberg, 1995; Ortmanns and Ney, 2000; Luk and Damper, 1998; Lleida and Rose, 2000; Deshmukh, et al., 1999). Since the work reported in this research is based on the former, the basic principle of Viterbi decoding is reviewed here.

According to the MAP rule, the decoder computes the likelihood of the unknown observation sequence given each acoustic model and choosing the one with the highest likelihood. In general, it is possible to use the forward probability calculation to compute the overall likelihood $P(\mathbf{O}|\lambda)$, and to identify the utterance based on these quantities. However, in practice, the most likely state sequence, which generates the sequence $\mathbf{O}$, is interested in. In addition, in many cases, the decision of choosing w is implicitly incorporated in the model by combining several models in parallel with common initial and final states and, in such case, the maximum likelihood path is an essential outcome of recognition. The Viterbi algorithm is a general dynamic programming technique used to find the most likely path in a trellis of nodes. The likelihood of the path is computed according to Eq. (2.43).

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 (Hajime, 1998; Wakita, et al., 1999; Junqua, et al., 1995; Richardson, et al., 1995; Deshmukh, et al., 1999; Kenny, et. al, 1994). A typical example is a two-pass system, where the first pass generates a list of the N most probable sequence using simplified acoustic models (Austin, et al., 1991; Wilcox and Bush, 1992; Huang, et al., 1994). The second-pass re-scores the list using detailed acoustic models and a language model (Mohri et al., 2002; Sato, et. al, 2002; Johnsen, 1989; Junqua, 1990; Matsunaga and Sakamoto, 1996; Tran, et al., 1996). A Japanese speech recognition system, for example, utilizes a two-pass search algorithm as shown in Figure 2.8 (Furui, 2001). However, a fundamental problem of the multi-pass decoding system is that search errors introduced in early passes are impossible to correct. Therefore, these errors result in degraded performance.

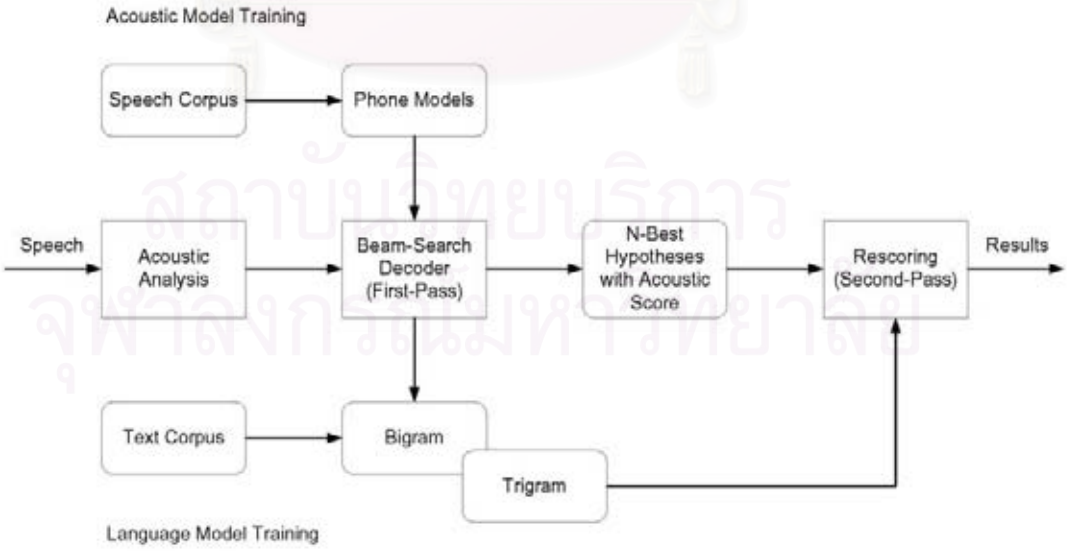


Figure 2.8. Two-pass search structure used in the Japanese broadcast-news transcription system

2.3.2 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). The language model defines the priori probability of a sequence of a word sequence W. The probability of a sentence, a sequence of words $w_1, w_2, ..., w_n$, provided by the language model, is given by

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)P(w_4|w_1, w_2, w_3)\cdots P(w_n|w_1, ..., w_{n-1})$$

$$= \prod_{i=1}^{n} P(w_i|w_1, ..., w_{i-1}) \quad (2.106)$$

where $P(w_i|w_1, ..., w_{i-1})$ indicates the probability that the word w_i was spoken given that the word sequence $w_1, w_2, ..., w_i$ 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). Instead, the language model probability is approximated in the following ways:

2.3.2.1 N-gram Language Models

For a vocabulary of size v, there are v^i different histories of words to specify $P(w_i|w_1, ..., w_{i-1})$ completely, so v^i values would have to be estimated. In reality, the probabilities $P(w_i|w_1, ..., w_{i-1})$ are impossible to estimate for even moderate values of i, since most histories $w_1, w_2, ..., w_i$ are unique or have occurred only a few times. A practical solution to the above problems is that the probability $P(w_i|w_1, ..., w_{i-1})$ is assumed depends only on some equivalence classes. The equivalence class can be simply based on the several previous words $w_{i-N+1}, w_{i-N+2}, ..., w_{i-1}$. This leads to an $n-gram$ language model.

The bigram models approximate the probability of a word depends only on the identity of immediately preceding word. To estimate $P(w_i|w_{i-1})$, the frequency with which the word w_i occurs given that the last word is w_{i-1},

simply count how often the sequence (w_i, w_{i-1}) occurs in some text and normalize the count by the number of times w_{i-1} occurs.

For a trigram model, the probability of a word depends on the two preceding words. The trigram can be estimated by observing the frequencies or counts of word pair $C(w_{i-2}, w_{i-1})$ and triplet $C(w_{i-2}, w_{i-1}, w_i)$ as follows:

$$P(w_i | w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \quad (2.107)$$

In principle, estimates for these probabilities can be directly calculated from text data by a simple frequency count. The estimates can be stored in a look-up table. This type of language model can be easily integrated with the recognition algorithm, and can be implemented as finite state networks. Therefore, the $n-gram$ is the most popular type of stochastic language model (Duchateau, 1998; Potamianos and Jelinek, 1998; Ng, et al., 2000; Clarkson and Robinson, 2001; O'Boyle, et al., 1994; Niesler and Woodland, 1999; Iyer and Ostendorf, 1999).

Deriving trigram and even bigram probabilities is still a sparse estimation, even with very large corpora. Even among the observed trigrams, the vast majority occurred only once. Therefore, straightforward maximum-likelihood estimation of $n-gram$ from counts is not advisable. Instead, various smoothing techniques have been developed. These include discounting the ML estimations (Witten and Bell, 1991), recursively backing off to lower-order $n-gram$ (Katz, 1987; Ney, et al., 1994; Kneser and Ney, 1995), and linearly interpolating $n-gram$ of different order (Jelinek and Mercer, 1980). The variable-length $n-gram$, language model with a longer span larger than n, can be estimated with the same amount of text data. The span varies depending on the word context, the previous words. A longer or shorter context can be preferable as there can be more or less examples for that context in the text data, or as the context is more or less relevant to predict the next world. Long distance grammars are primarily used to rescore n-best hypothesis lists from previous decoding (Rosenfeld, 1994; Blasig, 1999).

Another way to overcome sparseness is by vocabulary clustering. For any given assignment of a word w_i to class c_i, there may be many-to-many mappings. For instance, a word w_i may belong to more than one class, and

a class c_i may contain more than one word. For simplicity, a word w_i is assumed to be uniquely mapped to only one class c_i. The $n-gram$ model can be computed based on the previous $n-1$ classes:

$$P(w_i|c_{i-n+1},...,c_{i-1}) = P(w_i|c_i)P(c_i|c_{i-n+1},...,c_{i-1}) \quad (2.108)$$

where $P(w_i|c_i)$ denotes the probability of word w_i given class c_i in the current position, and $P(c_i|c_{i-n+1},...,c_{i-1})$ denotes the probability of class c_i given the class history. With such a model, the class mapping $w \to c$ can be learned from either text or task knowledge. In general, the class trigram can be express as:

$$P(W) = \sum_{c_1,...,c_n} \prod_i P(w_i|c_i)P(c_i|c_{i-2},c_{i-1}) \quad (2.109)$$

If the classes are nonoverlapping, a word may belong to only one class, then Eq. (2.109) can be simplified as:

$$P(W) = \prod_i P(w_i|c_i)P(c_i|c_{i-2},c_{i-1}) \quad (2.110)$$

As a typical example, the bigram probability of a word given the prior word (class) can be estimated as

$$\begin{aligned} P(w_i|w_{i-1}) &= P(w_i|c_{i-1}) \\ &= P(w_i|c_i)P(c_i|c_{i-1}) \end{aligned} \quad (2.111)$$

Class-based language models have been shown to be effective for rapid adaptation, training on small data sets, and reduced memory requirement for real-time speech application. For general-purpose large vocabulary dictation application, class-based $n-gram$ have not significantly improved recognition accuracy. They are mainly used as a back-off model to complement the lower-order $n-gram$ for better smoothing. Nevertheless, for limited domain speech recognition, the class-based $n-gram$ is very helpful as the class can efficiently encode semantic information for improving keyword spotting and speech understanding accuracy. (Wakita, et al., 1996; Deligne and Sagisaka, 2000; Dagan, et al., 1995; Riccardi, et al., 1996; Gao and Chen, 1997; Ward and Issar, 1996; Palmer, et al., 2000; Whittaker and Woodland, 2001; Yokoyama, et al., 2003; Ueberla, 1995)

2.3.2.2 Decision Tree Models

Decision trees and classification and regression trees algorithms were first applied to language modeling by Bahl and et al (Bahl, et al., 1989). A decision tree can arbitrarily partition the space of histories of words by asking arbitrary questions about the history $w_{i-N+1}, w_{i-N+2}, ..., w_{i-1}$ at each of the internal nodes. The training data at each leaf are then used to construct a probability distribution $P(w_i | w_1, ..., w_{i-1})$ over the next word. To reduce the variance of the estimate, this leaf distribution id interpolated with internal-node distributions found along the path to the root. Usually, trees are grown by greedily selecting, at each node, the most informative question, as judged by reduction in entropy. Pruning and cross validation are used.

Applying this language model is quite a challenge. The space of histories of words is very large, for example, 10^{100} for a 20-word sequence over a 100,000 word vocabulary, and the space of possible questions is even larger ($2^{10^{100}}$). Even if questions are restricted to individual words in the histories, there are still $20 \cdot 2^{10^5}$ such questions.

Theoretically, decision trees represent the ultimate in partition-based models. It is likely that trees exist that significantly outperforms n-grams. But finding them seems difficult for both computational and data sparseness reasons. Therefore, this approach was largely abandoned (Nadas, et al., 1991)

2.3.2.3 Linguistically Motivated Models

While all statistical language models get some inspirations from an intuitive view of language, in most models, actual linguistic content is quite negligible. Several language models, however, are directly derived from grammars commonly used by linguists.

A. Context-free grammar (CFG)

A CFG is a crude-well understood model of natural language. It is defined by a vocabulary, a set of non-terminal symbols, and a set of production of transition rules. Sentences are generated, starting with an initial non-terminal. By repeated application of the transition rules, which transform a non-terminal into a sequence of terminals (words) and non-terminals, until a terminals-only sequence is achieved (Huang, et al., 2001).

A probabilistic context-free grammar puts the probability distribution on the transitions producing from each terminal, thereby inducing a distribution over the set of all sentences. These transition probabilities can be estimated from annotated corpora using various algorithms such as the inside-outside algorithm and the expectation-maximization algorithm (Baker, 1976). However, the likelihood surfaces of these models tend to contain many local maxima. In addition, even if global ML estimation were feasible, it is generally believed that context-sensitive transition probabilities are needed to adequately account for actual behavior of language. Unfortunately, there is still no efficient algorithm for this situation.

Moreover, it is assumed that the expansion of any one non-terminal is dependent of the expansion of other non-terminals. Thus each probabilistic context-free grammar rule probability is multiplied together without considering the location of the node in the parse tree. This is against the intuition of since there is a strong tendency toward the context-dependent expansion. Another problem is the lack of sensitivity to words. The lexical information can only be represented via the probability of pre-terminal nodes, such as verb or noun, to be expanded lexically (Huang, et al., 2001).

B. Link grammar

Link grammar is a lexical grammar proposed by (Sleator and Temperly, 1991). Each word is associated with one or more ordered sets of typed links. Each such link must be connected to a similarly typed link of another word in the sentence. A legal parse consists of satisfying all links in the sentence via a planar graph. Link grammar has the same expressive power as a CFG, but arguably conforms better to human linguistic intuition (Sleator and Temperly, 1991).

2.3.2.4 Adaptive Models

Dynamic adjusting of the language model parameter, such as n-grams probabilities, vocabulary size, and the choice of word in vocabulary, is important since the topic of conversation is highly nonstationary (Iyer, et al., 1994; Jardino, 1996; Mahajan, et al., 1992). For example, in the dictation application, a particular set of words in vocabulary may suddenly burst forth and then become dormant later, based on the current conversation. Because

the topic of conversation may change from time to time, the language model should be dramatically different based on the topic of conversation. The adaptive model approach is introduced that can improve the quality of the language model based on the real usage of the application.

A. Cache language models

To adjust word frequencies observed in the current conversation, a dynamic cache language model is introduced. The basic idea of this technique is to accumulate word n-grams dictated so far in the current document and use these data to create a local dynamic n-grams model. The static n-grams model is used to adapt the probability as $P_s\left(w_i \middle| w_{i-n+1},...,w_{i-1}\right)$. The interpolation weight, λ, can be made to vary with the size of the cache.

$$P_{adaptive}(W) = \lambda P_{static}(W) + (1-\lambda) P_{cache}(W) \qquad (2.112)$$

The cache model is desirable in practice because of its impressive empirical performance improvement. In a dictation system, new words that are not in the static vocabulary have often occurred. The same words also tend to be repeated in the same article. The cache model can address this problem effectively by adjusting the parameters continually as recognition and correction proceed for incrementally improved performance (Huang, et al., 2001).

B. Topic-adaptive models

The topic can change over the time. Such topic or style information plays a critical role in improving the quality of the static language model. For example, the prediction of weather the word following the phrase “a bright” is “green” or “idea” can be improved substantially by knowing weather the topic of discussion is related to “color” or “cleverness”.

Domain or topic-clustered language models split the language model training data according to topic. The training data may be divided using the known category information or using the automatic clustering (Ney, et al., 1994; Popovici and Baggia, 1997; Ney and Essen, 1991). In addition, a given segment of data may be assigned to multiple topics. A topic dependent language model is then built from each cluster of the training data. Topic language models are combined using linear interpolation or maximum entropy as discussed in the next section (Kalai, et al., 1999).

C. Maximum entropy models

The language models as discussed above combine different n-grams models via linear interpolation. A different way to combine sources is the maximum entropy approach. It constructs a single model that attempts to capture all the information provided by the various knowledge sources. Each such knowledge source is reformulated as a set of constraints that the desired distribution should satisfy. These constraints can be, for example, marginal distributions of the combined model. Their intersection, if not empty, should contain a set of probability functions that are consistent with these separate knowledge sources. The maximum entropy principle can be stated as follows:

- Reformulate different information sources as constraints to be satisfied by the target estimate
- Among all probability distributions that satisfy these constraints, choose the one that has the highest entropy

One of the most effective applications of the maximum entropy model is to integrate the cache constraints into the language model directly, instead of interpolating the cache n-grams with the static n-grams. The new constraint is that the marginal distribution of the adapted model is the same as the lower-order n-grams in the cache (Rosenfeld, 1994; Wu and Khudanpur, 2002; Khudanpur and Wu, 1999; Martin, et al., 1999; Rosenfeld, 1996; Zhang, et al., 2000; Martin, et al., 2000; Chen and Rosenfeld, 2000; Rosenfeld, 1997; Wang, et al., 2001; Chen, et al., 1998). In practice, the maximum entropy method has not offered any significant improvement in comparison to the linear interpolation (Huang, et al., 2001).

2.3.2.5 Complexity Measures of Language Models

The choice of the language model in a large vocabulary recognition system heavily influences the difficulty of the recognition task and then the recognition performance. In the construction of the word sequence during recognition, if the language model can easily predict each next word, giving high probabilities for some words and low probabilities for the other words, then the recognition task is easy. Sentences from the text data on which the language model is based are more easily recognized with only little discounting than with more discounting of the probabilities. Conclusively, it

becomes very difficult to recognize a sentence that does not resemble enough the text data on which the language model is based.

The most common metric for evaluating a language model is the word recognition error rate, which requires the participation of a speech recognition system. Alternatively, the probability that the language model assigns to the test word strings can be measured without involving speech recognition systems. This is the derivative measure of cross-entropy known as a test-set perplexity (Huang, et al., 2001). This perplexity is related to the entropy H of the information source that produces the sequence of words of which consists the text data.

Given a language model that assigns probability $P(W)$ to a word sequence W, a compression algorithm that encodes the text W using $-\log_2 P(W)$ bits can be derived. The cross-entropy $H(W)$ of a model $P(w_i|w_{i-n+1},...,w_{i-1})$ on data W, with a sufficient long word sequence, can be simply approximated as

$$H(W) = -\frac{1}{N_W}\log_2 P(W) \tag{2.113}$$

where N_W is the word length of the text W.

The perplexity of a language model $P(W)$ is defined as the reciprocal of the geometric average probability assigned by the model to each word in the test set W. This is a measure, related to cross-entropy, known as test set perplexity:

$$Perplexity = 2^{H(W)} \tag{2.114}$$

The perplexity can be roughly interpreted as the geometric mean of branching factor of the text when presented to the language model (Huang, et al., 2001). The perplexity defined in Eq. (2.114) has two key parameters, a language model and a word sequence. The test-set perplexity evaluates the generalization capability of the language model, whereas the training-set perplexity measures how the language model fits the training data, like the likelihood. Lower perplexity correlates with better recognition performance. This is because the perplexity is a statistically weighted word branching measured on the test set. The higher the perplexity, the more branches the speech recognizer needs to consider statistically. The SPHINX, for example, on the 997-word resource management task, SPHINX attained a word

accuracy of 96% with a grammar (perplexity 60), and 82% without grammar (perplexity 997) (Lee, et al., 1989).

A language with higher perplexity means that the number of word branching from a previous word is larger on average. In this case, the perplexity is an indication of the complexity of the language. The perplexity of a particular language model can change dramatically in terms of vocabulary size, the number of states of grammar rules, and the estimated probabilities (Huang, et al., 2001). A language model with perplexity of X has roughly the same difficulty as another language model in which every word can be followed by X different words with equal probabilities. In the task of connected digit recognition, for example, the perplexity is 10.

2.4 Summary

This chapter reviewed fundamental techniques used in speech recognition. The important issue in speech recognition is acoustic pattern matching, which has a close relation with signal processing and language modeling. Selection of signal processing methods depends on the subsequent distortion measures or probability density function. Cepstral-based analysis is widely used due to its low correlation property. In acoustic pattern matching, hidden Markov model is the currently technique of state-of-the art speech recognition. Language modeling assists acoustic pattern matching since it can be used to impose constraints in acoustic search space. In the large vocabulary speech recognition system section, it should be note that hidden Markov model and language modeling are usually combined in the same computational framework in practical speech recognition system design.

Chapter 3

Phonological and Acoustical Analysis of Thai Language

This chapter is intended to provide the essential knowledge of the Thai language. Since the syllable is principally considered a fundamental unit for acoustic-phonetic analysis, it is important to have a good understanding about Thai syllables. The basic Thai phonetic units will be described. Since Thai language is known of being a tonal language, the five lexical tones and their distinctive linguistic features will be elaborated. Also, major constraints, which combine these phonemic units into syllable, will be explained. In the acoustical point of view, four acoustic parameters, fundamental frequency, formant frequency, intensity, and duration, of Thai syllable will be examined. Furthermore, spectral feature of Thai syllable will be discussed. At the end of this chapter, the acoustic feature extraction techniques will be described. A good understanding on phonological and acoustical properties of Thai language paves the way for creating the appropriate speech unit for Thai speech recognition in subsequent chapters.

3.1 Phonology of Thai Language

This section gives details of Thai language in phonological point of view. Basic phonetic units, consisting of initial consonant, final consonant, and vowel, will be introduced first. Then, details of Thai tone system will be described. Finally, Thai syllable structure and the rule, which combine phonetic units into syllable, will be explained.

3.1.1 Basic Phonetic Units

There are 21 consonantal phonemes, 12 consonant clusters, 18 monophthongs, 6 diphthongs, and 5 tones in Thai language. These phonemic units form totally 26,928 grammatically admissible syllables (Luksaneeyanawin, 1992; Luksaneeyanawin, 1993). Details of each sound unit are described below.

A. Initial Consonant

The Thai language has a total of 33 initial consonants consisting of 21 consonantal phonemes and 12 consonant clusters. A set of 21 consonantal phonemes is categorized by place of articulation and manner of articulation. The place of articulation can be labial, alveolar, palatal, velar, and glottal respectively. The manner of articulation is classified into two major groups, stops and non-stops. The stops are subcategorized into voiceless unaspirated stops, voiceless aspirated stops, and voiced stops. The non-stops are subcategorized into nasals, fricatives, a trill, a lateral, and approximants. The details of Thai consonantal phonemes and consonant clusters are illustrated in Tables 3.1 and 3.2 respectively. The other set of initial consonants comprises consonant clusters composed of two co-articulated consonants.

B. Final Consonant

There are eight different final consonants in the Thai language. The three stops, [p], [t], and [k], appearing at the final position are acoustically different from the initial consonant, that is, they are not audibly released. Also, two approximants, [j] and [w], can occur at the final position of a syllable. Instead of considering a vowel ending with [j] or [w] as a diphthong, they are treated as a vowel and a final consonant separately, though they have vowel-like spectral features. Finally, a group of nasals, [m], [n], and [ng], can be final consonants.

C. Vowel

The Thai language has a complex vowel system. It consists of 18 monophthongs and 6 diphthongs. The monophthongs are qualitatively 9 different vowels, each of which has two members, short and long. Thai monophthongs are categorized according to the tongue position, tongue advancement and tongue height. Tongue advancement relating to the second formant frequency is subdivided into front, central, and back. Tongue height corresponding to the first formant frequency is subdivided into high, middle, and low. The relationship between tongue position and vowels is shown in Table 3.3. Obviously, Thai vowels completely span tongue advancement and tongue height combinations.

Table 3.1 Thai consonantal phonemes

			Place of Articulation				
			Labial	**Alveolar**	**Palatal**	**Velar**	**Glottal**
Manner of Articulation	**Stops**	**Voiceless Unaspirated**	/p/*	/t/*	/c/	/k/*	/?/
		Voiceless Aspirated	/ph/	/th/	/ch/	/kh/	
		Voiced	/b/	/d/			
	Non-stop	**Nasal**	/m/*	/n/*		/ng/*	
		Fricative	/f/	/s/			/h/
		Trill		/r/			
		Lateral		/l/			
		Approximant	/w/*		/j/*		

* These consonants are both releasing consonants and arresting consonants

Table 3.2 Thai consonant clusters

	C_1					
C_2	***p***	***t***	***k***	***ph***	***th***	***kh***
r	/pr/	/tr/	/kr/	/phr/	/thr/	/khr/
l	/pl/		/kl/	/phl/		/khl/
w			/kw/			/khw/

Table 3.3 Thai vowel phonemes

		Tongue Advancement		
		Front	***Central***	***Back***
Tongue Height	***High***	/i, i:/	/v, v:/	/u, u:/
	Medium	/e, e:/	/q, q:/	/o, o:/
	Low	/x, x:/	/a, a:/	/@, @:/
Diphthongs		/ia, i:a/	/va, v:a/	/ua, u:a/

3.1.2 Thai Tones

Basically, a tone is a feature of pitch movement within a syllable. Syllables or words having the same sequence of consonants and vowels but different pitch contours are different lexical entries. In addition to Thai, some

European, African, and Oriental languages are tonal languages. There are 5 tones in the Thai language that are divided into 2 groups corresponding to the change of pitch pattern, static and dynamic tones. The former have either a flat or slightly falling pitch, including the high tone, the mid tone, and the low tone while the latter, characterized by a significant pitch movement during the syllable, consist of the falling tone and the rising tone. Five tonal patterns are depicted in Figure 3.1. Furthermore, Thai syllables are governed by the rules of tone assignments as shown in Table 3.4.

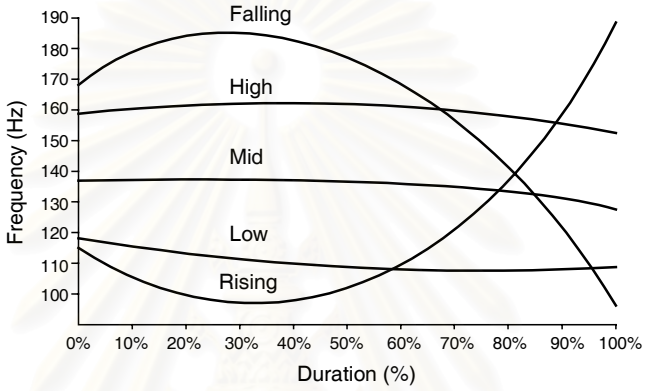


Figure 3.1 Five Thai tones

Table 3.4 Thai tone assignments

Thai syllable	Syllable structure	Possible tone
1 Open syllable		
1.1 Open long syllable	C (V: or V:V)	M, L, F, H, R
1.2 Open short syllable	C (V or VV)	L, F, H
2 Sonorant ending syllable		
2.1 Short syllable ending with sonorant consonant	C (V or VV) C_f	L, F, H, R
2.2 Long syllable ending with sonorant consonant	C (V: or V:V) C_f	M, L, F, H, R
3 Obstruent ending syllable		
3.1 Short syllable ending with obstruent consonant	C (V or VV) C_f	L, F, H
3.2 Long syllable ending with obstruent consonant	C (V: or V:V) C_f	L, F, H

3.1.3 Thai Syllable Structure

Thai syllables are composed of three sound systems, namely consonants, vowels, and tones. The smallest construction of sounds or a syllable in Thai is composed of one monophthong unit or one diphthong, one, two, or three consonants, and a tone (Luksaneeyanawin, 1992; Luksaneeyanawin, 1993). The construction can be represented with the structure illustrated in Figure

3.2. Combinations of these sound units are restricted by the rules shown in Table 3.5.

$$S = c(c)V(V)(C)^{T}$$

Figure 3.2 Thai syllable structure

Table 3.5 Combinations of Thai sound units

Thai Syllable	C_i	V	C_f	T	S	S+T
1 Open syllable						
1.1 Open long syllable	33	12		5	396	1,980
Inadmissible co-occurrences						
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4		5	-8	-40
1.2 Open short syllable	33	12		3	396	1,188
Inadmissible co-occurrences						
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4		3	-8	-24
2 Sonorant ending syllable						
2.1 Short syllable ending with sonorant consonant	33	12	5	5	1,980	9,900
Inadmissible co-occurrences						
Round vowel unit preceding a labialized consonant	33	4	1	5	-132	-660
Front vowel unit preceding a palatalized consonant	33	4	1	5	-132	-660
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4	4	5	-32	-160
2.2 Long syllable ending with sonorant consonant	33	12	5	5	1,980	9,900
Inadmissible co-occurrences						
Round vowel unit preceding a labialized consonant	33	4	1	5	-132	-660
Front vowel unit preceding a palatalized consonant	33	4	1	5	-132	-660
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4	4	5	-32	-160
3 Obstruent ending syllable						
3.1 Short syllable ending with obstruent consonant	33	12	3	3	1,188	3,564
Inadmissible co-occurrences						
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4	3	3	-24	-72
3.2 Long syllable ending with obstruent consonant	33	12	3	3	1,188	3,564
Inadmissible co-occurrences						
Labial consonant clusters /kw, khw/ and 4 round vowels	2	4	3	3	-24	-72
Total					**6,472**	**26,928**

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. However, there are too many syllable units in Thai language, 26,928 units (Luksaneeyanawin, 1993). Thus, the use of syllables as acoustic units for speech recognition in the Thai language is not practicable. As an alternative, the sub-syllable units have to be taken into account.

3.2 Acoustical Analysis of Thai Language

3.2.1 Acoustic Features

The acoustic-phonetic study has produced an extensive understanding of properties of sound. A spoken language is decomposed into elements of linguistically distinctive sounds called phonemes. The continuous sound wave is segmented into discrete regions corresponding to its acoustic properties. Properties of sounds referred to acoustic-phonetic features are employed to classify these phonemes systematically according to their articulatory configurations. Hence, a suitable method of representing the time-varying characteristics of speech signal is via a parameterization of the spectral properties based on the model of speech production. Four acoustic features based on the speech production model, fundamental frequency, formant frequency, energy, and duration, are employed for analysis and recognition. These acoustic features containing crucial information of speech signal are the important cues for distinguishing phoneme units.

A. Formant Frequency

Since the vocal tract is an air tube acting as a resonator, it has certain natural frequencies of vibration. The natural frequencies of the vocal tract are excited by a source or sources located either at the glottis or at some points along the length of the tract (Stevens, 1999). The natural frequencies or resonant frequencies of the vocal tract tube are called formant frequencies

and the resonances are simply called formants (Rabiner and Schafer, 1978). When a speech signal is modified in the vocal tract, and is transmitted toward the lips, the spectrum of the sound emerging from the lips has the peak at the natural frequency of the vocal tract (Denes and Pinson, 1963). Dependent upon the shape and dimensions of the vocal tract, different formant frequencies are formed by varying the shape of the vocal tract. The lowest formant frequency is called the first formant. The next highest frequency is called the second formant, and so forth.

The formant frequencies are estimated from the short time spectrum by the Fourier transform. A general procedure for formant trajectory estimation is based upon linear prediction analysis, with two formant-tracking techniques, solving the roots of the LP polynomial and spectral peak picking (Markel and Gray, Jr., 1980). Solving for the roots assures that the accurate formant frequency and bandwidth will be extracted. Required highly computational expense in solving the roots, the former method is not favorable technique. The spectral peak picking seems to be a practical procedure for formant estimation due to its low computation. Nevertheless, the major disadvantage of the latter method is that closed formant related to closed complex pole pairs may not be extracted from the spectrum.

With a compact representation of the time-varying of speech signal, the formant frequencies have been employed as an acoustic feature for phoneme classification. The first, the second, and the third formant frequency are adequately exploited to identify vowel phonemes. According to vocal tract shape, the formant frequencies are dependent on three factors: the position of the point of maximum constriction in the vocal tract controlled by the backward and forward movement of tongue, the size or cross-sectional area of the maximum constriction controlled by the movements of tongue towards and away from the roof of the mouth and the back of the throat, and the position of lips (Ladefoged, 1962). The first formant associated with tongue height in the second factor is used for classifying vowels into the high, the middle, and the low group. The second formant correlated to tongue advancement in the first factor is utilized to categorize vowels into the front, the central, and the back group. Finally, the third formant, which is dependent on the shape of lips, is employed to define roundness of vowels.

B. Fundamental Frequency

The basic property of a vocal cord sound source is its periodicity expressed by the duration of a complete voice period or by the inverse value of the voice fundamental frequency (Fant, 1970). Related to the number of times the vocal folds open and close per second, the frequency of vocal fold vibration directly determines the lowest frequency of the sound, which is produced (Borden and Harris, 1980). The duration of pitch cycle can always varies 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. The intonation can be imposed on a sentence, a phrase, or a word. English sentences are often characterized by a rising-falling intonation curve (Borden and Harris, 1980). The pitch rises at the first part and falls at the end of a declarative sentence or a question sentence impossible to answer yes or no. Another pattern in English is the end-of-utterance pitch rise appearing in a question sentence to be answered yes or no. 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. According to a speech production model, a periodic glottal excitation waveform is originated from the periodic opening and closure of the vocal cords in the glottis. Air is forced through the glottis from the lung resulting in a train of alternating high and low pressure pulses in vocal tracts (Vuuren, 1998). Only voiced sounds have periodic opening and closure. On the other hand, the air passes through the glottis unrestricted in unvoiced sounds. 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. The modified correlation method and simplified inverse filter tracking (SIFT) algorithm in correlation processing category and the cepstral method in spectrum processing category are the most efficient techniques since they explicitly remove the vocal tract effects (Furui, 2001).

C. Energy

Energy together with other cues, formant and duration, is used to classify both Thai consonants and vowels (Trongdee, 1987; Tarnsakun, 1988; Maneenoi, 1998). In addition, energy, one of the prominent acoustic parameters, is used to detect syllable boundary in Thai connected speech especially in sequence of two consonantal segments and in sequence of two vocalic segments (Sriraksa, 1995; Jittiwarangkul, 1998). Acoustic characteristics of each non-stop consonant are acoustically different both in place of articulation and in manner of articulation. From the acoustic study, the nasals have low second formant energy whereas trill and lateral have high first and second formant energy (Trongdee, 1987). In classification of stops, energy was also employed to distinguish each stop both in place of articulation and in manner of articulation. The classification results show that energy of aspirated stops is higher than unaspirated stops as well as voiced stops and voiceless stops (Tharnsakun, 1988).

D. Duration

One of the important acoustic cues used in classification of Thai phonemes is duration. This acoustic feature was employed to classify both consonants and vowels (Trongdee, 1987; Tarnsakun, 1988; Thubthong, 1995; Maneenoi, 1998). Each non-stop consonant in the same manner of articulation has different durations depended on its structural context (Trongdee, 1987). For stop consonants, the voiceless stops have longer durations than voiced stops. Between voiceless stops, voiceless aspirated stops have longer durations than voiceless unaspirated stops (Tarnsakun, 1988). Two of duration features, noise duration and burst duration, accompanied by other acoustic features were used in categorization of Thai initial consonants (Thubthong, 1995). Since short and long vowels are quantitatively different and Thai vowels appear in both short and long pairs, duration is a predominant acoustic cue used for classification them. Hence, classification of Thai vowels entails duration to distinguish short and long vowels (Thubthong, 1995; Maneenoi, 1998). Additionally, the duration rather than the intensity of the vowel segments can determine which syllable is stress (Denes and Pinson, 1963).

These four acoustic features are the important cues for both speech analysis and speech recognition. Many researches on Thai language have used these acoustical features in their works. Abramson (1960) employed acoustical measurements on the study of the vowels and tones of standard Thai. Trongdee (1987) and Tharnsakun (1988) worked on the analyses of non-stop consonants and stop consonants in Thai respectively. The two works have studied the acoustical characteristics of Thai consonants occurring in monosyllabic words. These studies utilized formant frequency, duration, and intensity of 10 non-stops and 11 stops in Thai with 3 different vowel contexts. Leelasiriwong (1991) studied acoustic characteristics of Thai vowels, /i:, a:, u:/. The first three-formant frequencies and the fundamental frequency of these vowels were statistically modeled used in speaker identification. Thubthong (1995) used the pre-consonantal second formant transition and the other acoustic features for consonantal phoneme classification. Six Thai vowels were classified using two formants and duration as well. Maneenoi (1998) applied the artificial neural network together with the first three-formant frequencies and their energy as an acoustic feature for vowel phoneme recognition. Instead of using linear frequency scale, the non-linear frequency scales, Bark and Mel scale, were applied to the classification of the nine Thai spreading vowels (Ahkuputra, et al., 2003).

3.2.2 Acoustic Feature Extraction for Speech Analysis

A. Formant Frequency Tracking

On the formant frequency estimation, a spectrum envelope of a speech signal is tracked to find a spectral peak as shown in Figure 3.3. The lowest spectral peak is picked and marked as the first formant or F_1. The following spectral peaks are marked as the second F_2, the third F_3, and the forth F_4, respectively.

In order to obtain a spectrum envelope of the power spectrum, the linear predictive coding (LPC) coefficients are analyzed on the speech segment using the Levinson-Durbin recursive algorithm (Rabiner and Juang, 1993; Deller, et al., 1993; Furui, 2001). The LPC coefficients, $a_0, a_1, ..., a_p$, is the coefficients of the all-pole filter with the form in Eq. (3.1) , where p is the order of the LPC coefficients.

$$H(z) = \frac{1}{1 - \sum_{k=0}^{p} a_k z^{-k}} \tag{3.1}$$

The spectrum envelope could be obtained by taking the discrete Fourier transform to evaluate $H(e^{j\omega})$.

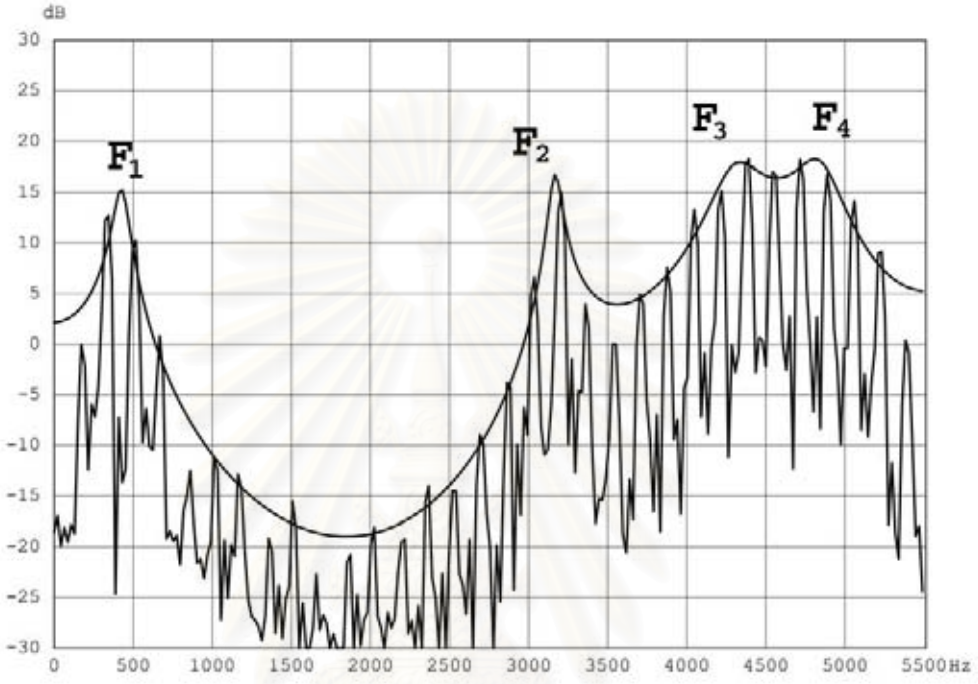


Figure 3.3 Formant frequency

Another method to obtain the formant frequency is the root-solving technique. The complex poles, derived from a set of p-order linear prediction coefficients a_k, are computed by solving roots of a polynomial in Eq. (3.1) (Furui, 2001). Due to a stable all-pole filter model of a linear prediction, the poles solved from roots of the transfer function are located inside a unit circle. Let $z = \mathrm{Re}(z) + j\,\mathrm{Im}(z)$ be a root of a linear prediction polynomial, the formant frequency value, which related to an angle of the complex pole, can be computed as shown in Eq. (3.2). Since the poles occur in complex conjugate pairs then only the upper half of a unit circle is considered to compute their corresponding formant frequencies. In addition, the bandwidth information is estimated from a magnitude of the complex pole as shown in Eq. (3.3), where n is a formant number and f_s is the sampling frequency in Hz. The bandwidth information is additionally considered in order to exclude the undesired frequencies. This root-solving method gives

out much more precise and accurate formant frequency value than the peak picking technique, which depends on resolution of the DFT.

$$F_n = \left[\frac{f_s}{2\pi}\right]\arctan\left[\frac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right] \tag{3.2}$$

$$B_n = \left[\frac{f_s}{\pi}\right]\log|z| \tag{3.3}$$

B. Fundamental Frequency

For the automatic pitch extraction, properties of cepstrum have been utilized to reveal the signal periodicity. The cepstrum is the Fourier transform of the logarithm of the amplitude spectrum of a signal. The resulting independent variable, which is reciprocal frequency, or time, is called "quefrency" (Flanagan, 1972).

The cepstrum is defined as the inverse Fourier transform of a short-time logarithmic amplitude spectrum. The cepstrum analysis is illustrated in Figure 3.4. The quefrency, the independent parameter for the cepstrum, is the time domain parameter resulting from the inverse transform of the frequency domain function (Furui, 2001). Let $x(t)$ is the voiced speech, which is the response of the vocal tract articulation equivalent filter driven by a pseudo-periodic source $g(t)$. Then, $x(t)$ could be given by the convolution of $g(t)$ and the vocal tract impulse response $h(t)$ as follows

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \tag{3.4}$$

$$X(\omega) = G(\omega)H(\omega) \tag{3.5}$$

where $X(\omega)$, $G(\omega)$, and $H(\omega)$ are the Fourier transform of $x(t)$, $g(t)$, and $h(t)$, respectively. By taking the logarithm function and the inverse Fourier transform, the cepstrum $c(\tau)$ is obtained as follows

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \tag{3.6}$$

$$c(\tau) = F^{-1}\left(\log|X(\omega)|\right) \tag{3.7}$$

$$= F^{-1}\left(\log|G(\omega)|\right) + F^{-1}\left(\log|H(\omega)|\right) \tag{3.8}$$

From the right side of Eq. (3.6), the first term represents the spectral fine structure or the periodic pattern and the second term represents the spectrum envelope or the global pattern along the frequency axis. The

fundamental period of the source $g(t)$ could be extracted from the peak at the high quefrency region, that is, the first term indicates the formation of the peak in the high frequency region (Furui, 2001).

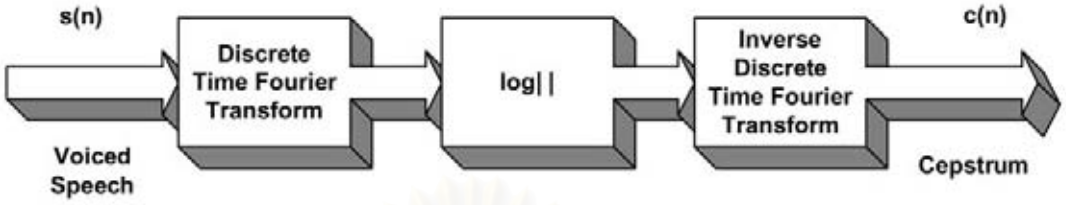


Figure 3.4 Cepstrim analysis

In Figure 3.5, a voiced and unvoiced speech segment are analyzed using spectrum and cepstrum analysis. In voiced speech, the sharp peak occurs in the cepstra plot, which corresponds to the period of pitch. Unlike voiced speech, unvoiced speech has no peak, which results in no fundamental frequency. The example of short-time spectra and cepstra is shown in Figure 3.6. During the voiced speech, a sharp peak occurs in the quefrency domain of the corresponding spectra in the period. The sharp peak disappears in the unvoiced speech portion. The fundamental frequency is directly computed from the location of the peak, which is the reciprocal of the period. The pitch period tracking is shown in Figures 3.5-3.6.

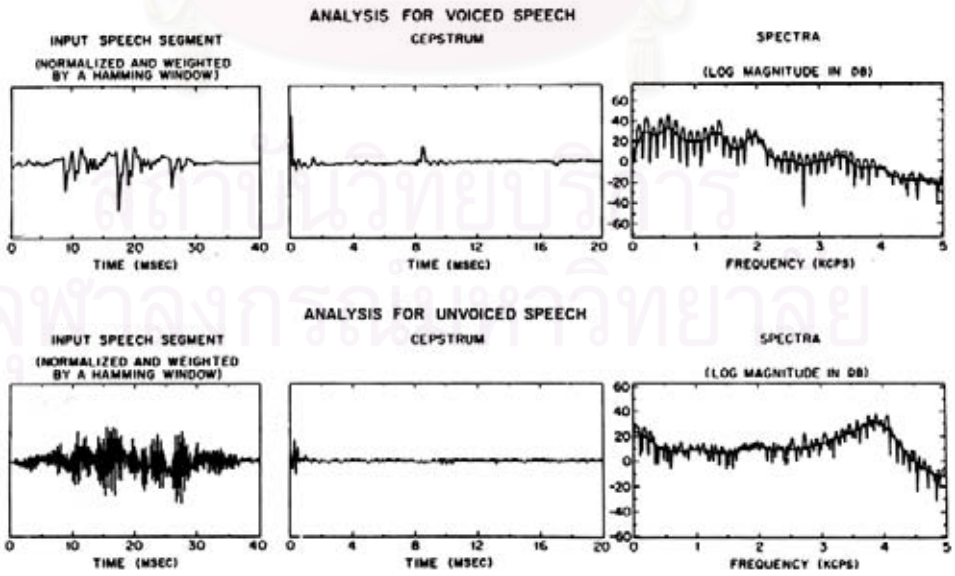


Figure 3.5 Spectrum and cepstrum analysis of voiced and unvoiced speeh sounds (Flanagan, 1972)

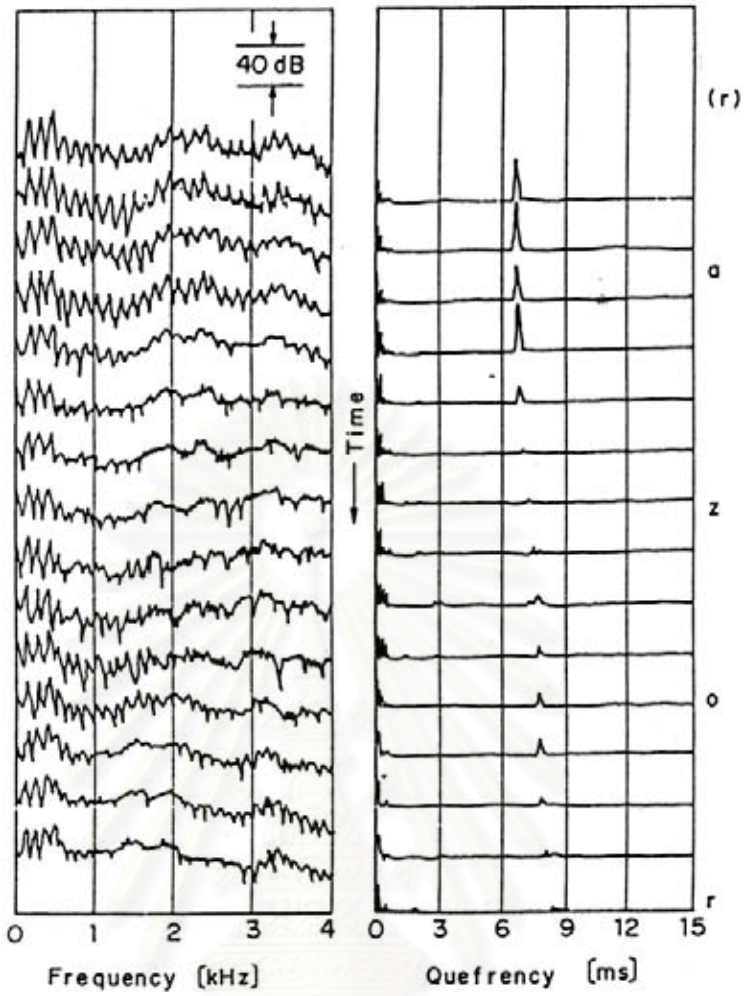


Figure 3.6 Short-time spectra and cepstra for male voice (Furui, 2001)

C. Energy

Amplitude of a speech wave is a peak of a speech waveform. In other words, an 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 position 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 speech wave using the following relation as shown in Eq. (3.9). In Eq. (3.9), $E(m)$ is an absolute energy value of the m^{th} frame, $s(n)$ is an amplitude of the n^{th} sample, and N is the total samples,

$$E(m) = \sum_{n=1}^{N} |s(n)| \tag{3.9}$$

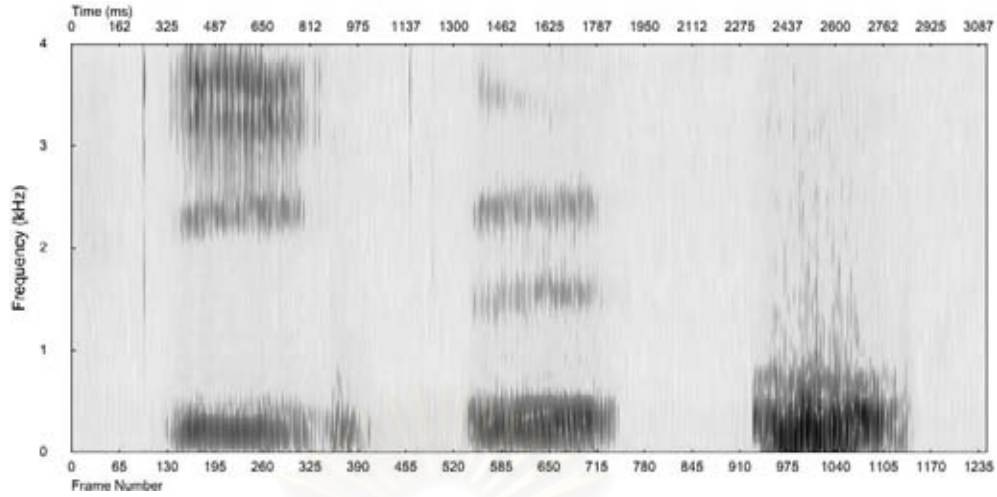
3.2.3 Acoustical Properties of Thai Phonemic Units

A. Vowel

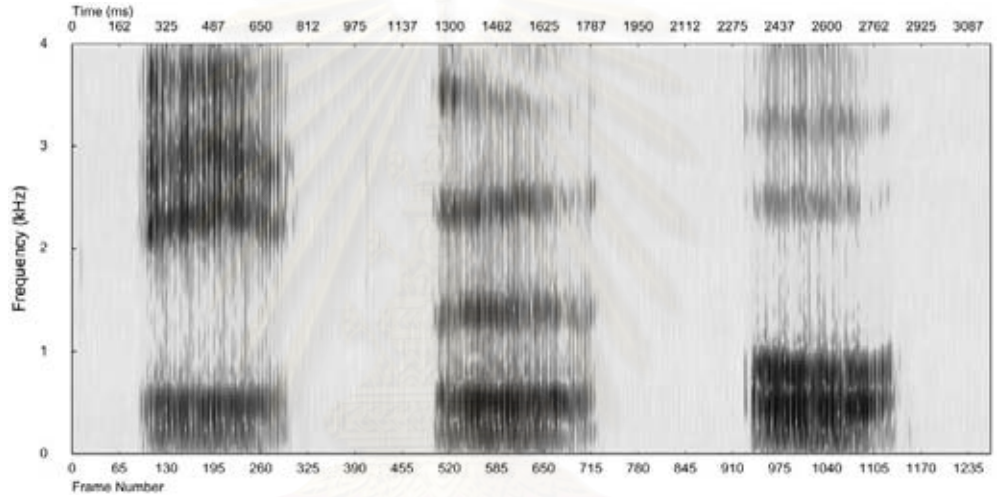
Formant structure is an important indicator to describe vowel sounds in acoustic analysis. It refers to specific resonant frequencies of the vocal tract, which have the greatest energy concentration. Each of the Thai vowels is acoustically analyzed to explore the acoustic characteristics. Spectrums and spectrograms of each vowel are illustrated in Figure 3.7. In Figure 3.7, each Thai vowel shows its unique acoustic characteristics in terms of formants. According to the tongue position, the vowel advancement is represented by the second formant frequency, F_2, whereas the vowel height is represented by the first formant frequency, F_1. The front vowels have the highest F_2 followed by the central and the back vowels, respectively. The low vowels have the highest F_1 followed by the mid and the high vowels, respectively. The vowel triangle, /ii/, /aa/, and /uu/, show distinct characteristics between each other. The vowel triangle is the common set of the vowels existed in every language in the world.

From the acoustic characteristics described above, the Thai vowel distribution in F_2 and F_1 plane is shown in Figure 3.8. Three classification schemes namely, classification by vowel height, classification by vowel advancement, and classification by combined vowel height and advancement, were proposed to classify nine Thai monophthongs using Bayesian classifier (Ahkuputra, et al., 2003). The results show that the use of acoustic features, F_1 and F_2, gives the high accuracy in vowel identification. In addition to the first-two formants, the third formant, F_3, represents the degree of roundness in lip opening. The three dimensional distribution of Thai monophthongs in F_1, F_2 and F_3 plane, is depicted in Figure 3.9. Obviously, Thai vowels completely span tongue advancement and tongue height combinations. From the acoustic analysis, it is generally agreed that the first three formant frequencies are the most informative for vowel perception and discrimination (Maneenoi, 1998; Ahkuputra, 2002; Ahkuputra, et al., 2003).

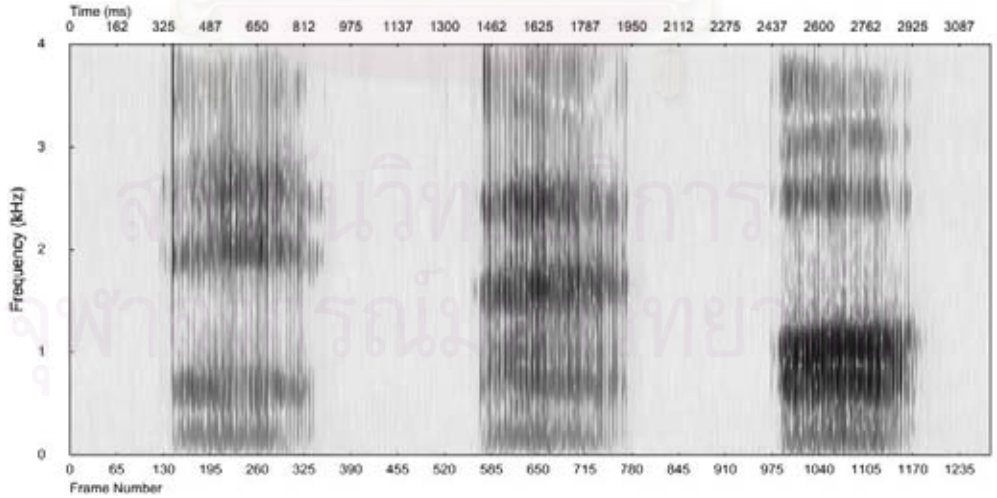
In addition to acoustic characteristics of the Thai vowels, the duration is one of the acoustic cues used to identify the short and long vowels. The durations of the short and long vowels are shown in Figure 3.10.



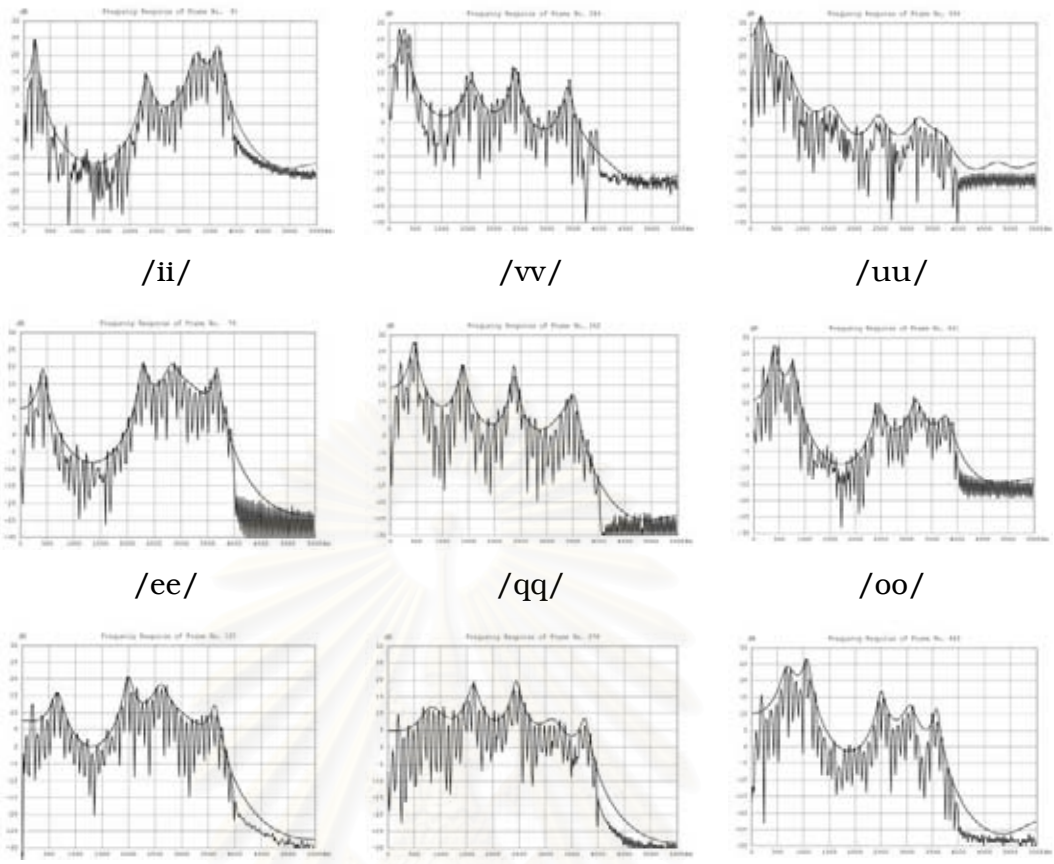
(a) Spectrogram of high vowels /ii/, /vv/, /uu/



(b) Spectrogram of middle vowels /ee/, /qq/, /oo/



(c) Spectrogram of low vowels /xx/, /aa/, /@@/



(d) Short-time spectrum of vowels

Figure 3.7 Spectrogram and spectrum of nine Thai vowels

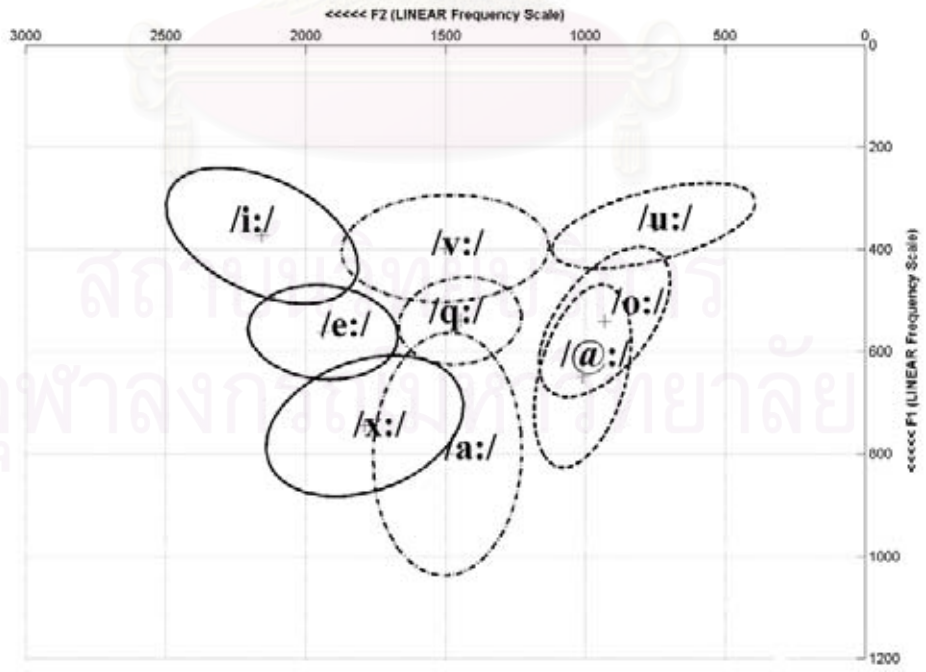
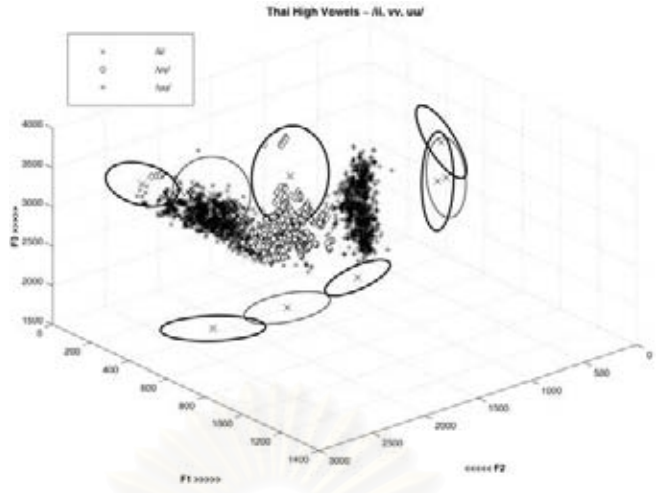
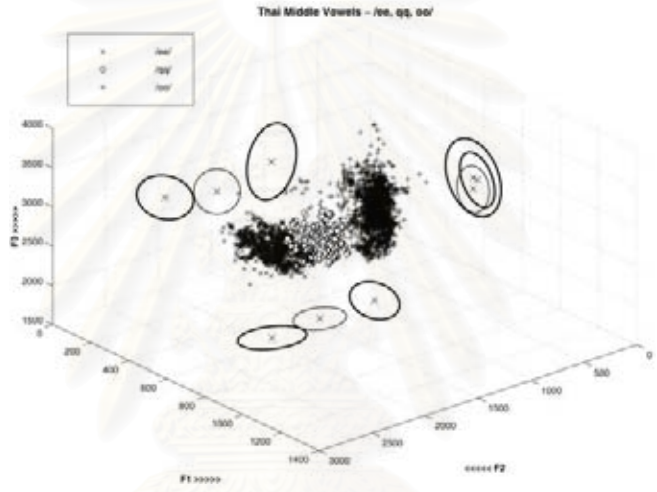


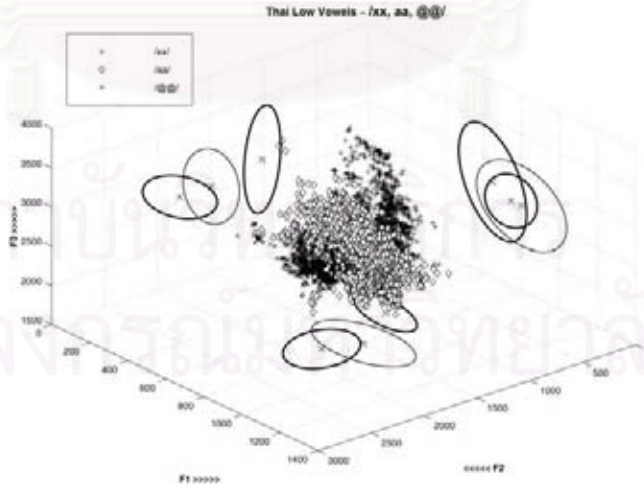
Figure 3.8 Distribution of Thai vowels on F2 and F1 plane



(a) High vowels /ii/, /vv/, /uu/



(b) Middle vowels /ee/, /qq/, /oo/



(c) Low vowels /ee/, /qq/, /oo/

Figure 3.9 Projection of vowel distribution on F1-F2, F1-F3, and F2-F3 plane

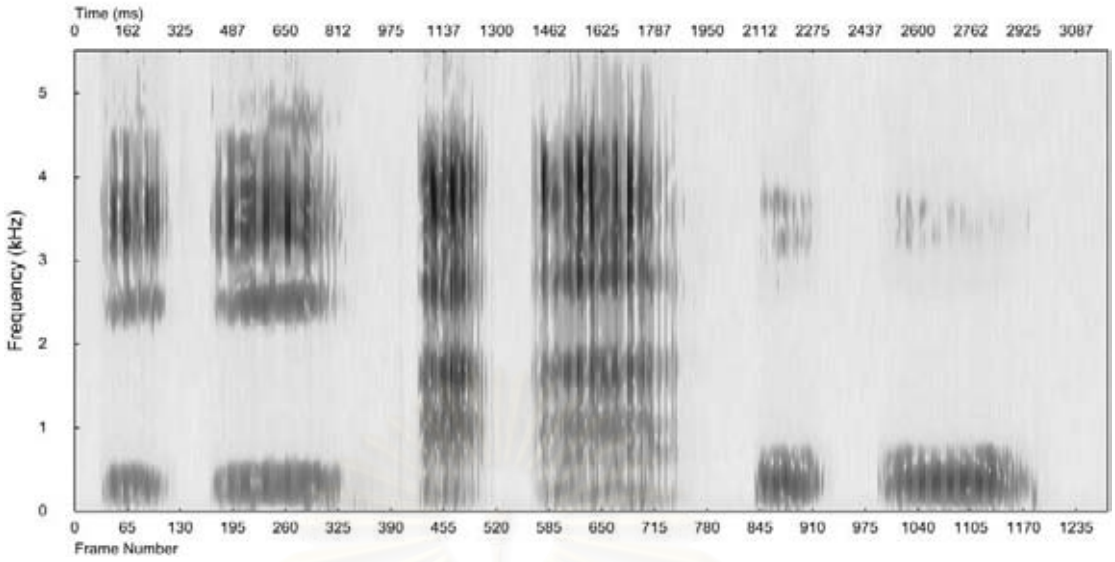


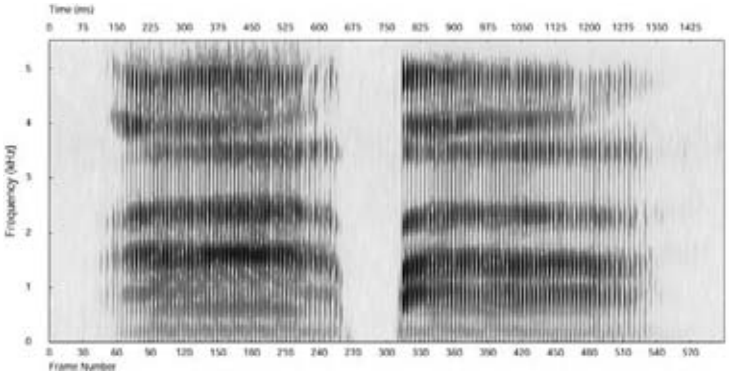
Figure 3.10 Spectrum and duration characteristics of the short-long vowel pairs /i/-/ii/, /a/-/aa/, and /u/-/uu/

B. Consonant

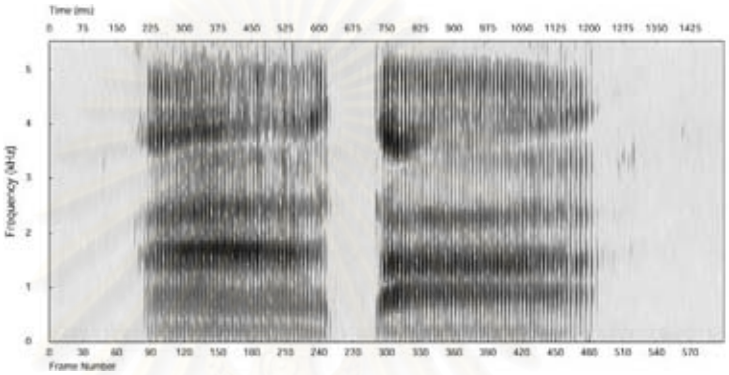
Acoustically, marginal sounds or consonants can be attached along both sides of the syllable nucleus or the vowel. Considering the Thai syllable structure, the left marginal sound is an initial or a releasing consonant and the right marginal sound is a final or an arresting consonant. Examples of releasing consonants are shown in Figure 3.11.

In Figure 3.11, spectrographic information of the syllables /?aa0paa0/, /?aa0taa0/, /?aa0kaa0/, and /?aa0caa0/ are illustrated. These releasing consonants are in the same manners of articulation but different places of articulation. The transitional periods between the releasing consonant and its following vowel are clearly different, according to its locus of each consonant. Thus the transitional period contains the crucial acoustic cues for identification of the releasing consonants.

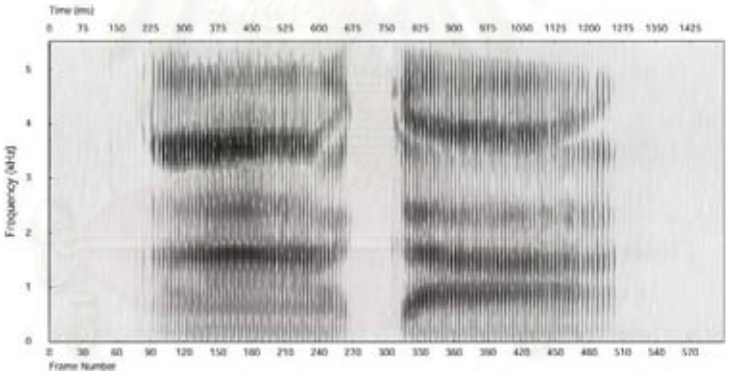
In Figure 3.12, spectrographic information of the syllables /?aa0paa0/, /?aa0phaa0/, and /?aa0baa0/ are shown. These releasing consonants are in the same places of articulation but different manners of articulation. The transitional periods between the releasing consonant and its following vowel are evidently comparable. The phonemes /p/ and /ph/ are unaspirated and aspirated voiceless stops while the phoneme /b/ is voiced stop.



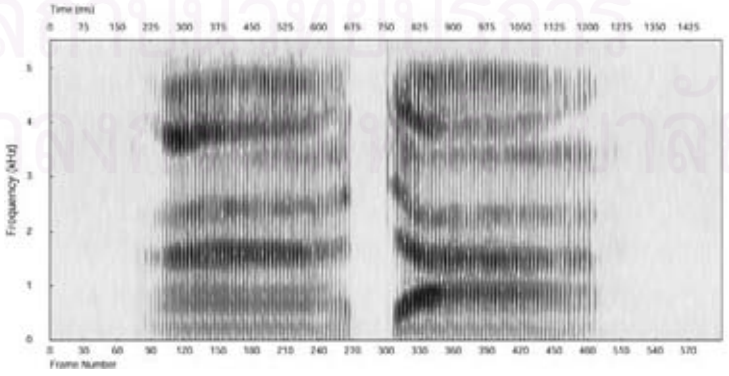
(a) /?aa0paa0/



(b) /?aa0taa0/

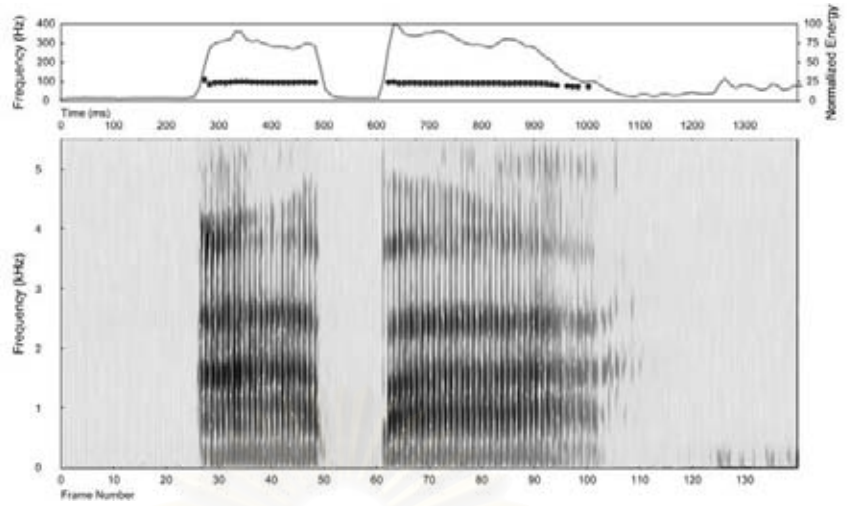


(c) /?aa0kaa0/

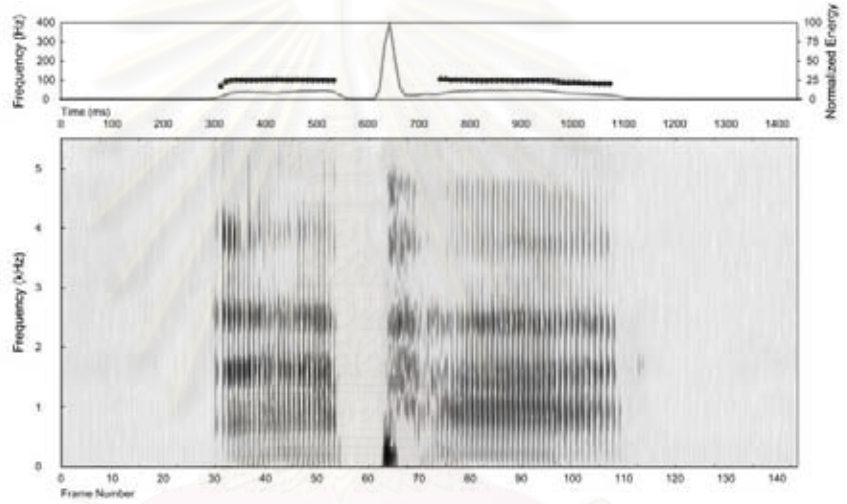


(d) /?aa0caa0/

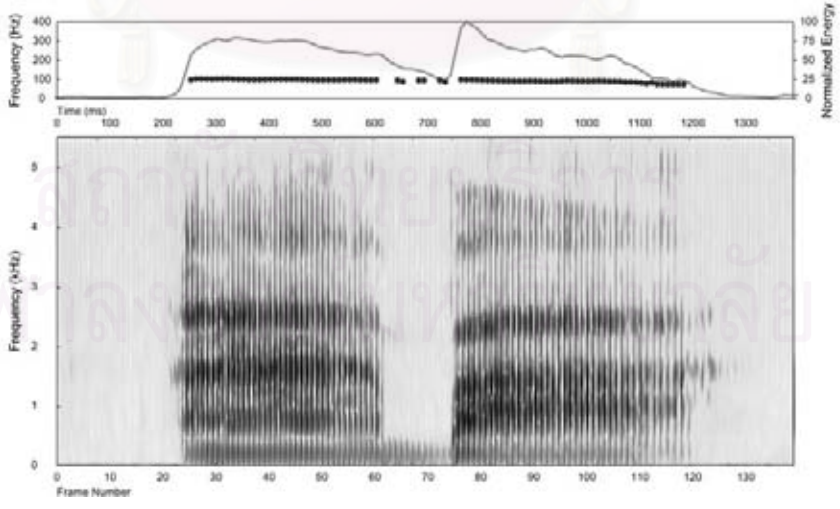
Figure 3.11 Spectrographic information of releasing consonants in the same manners of articulation but in thedifferent places of articulation



(a) /?aa0paa0/



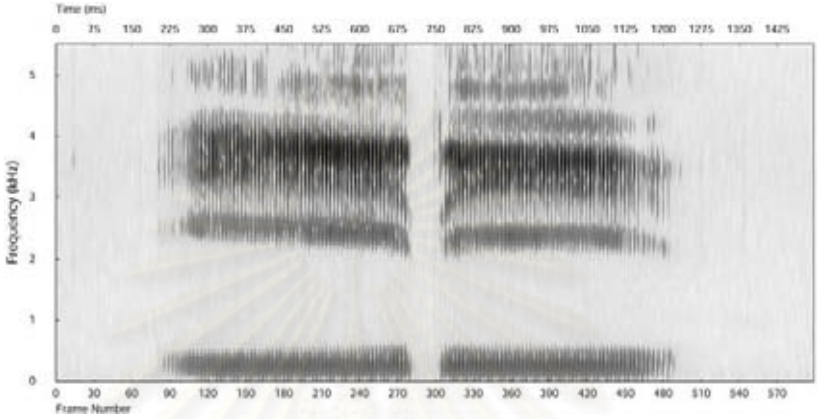
(b) /?aa0phaa0/



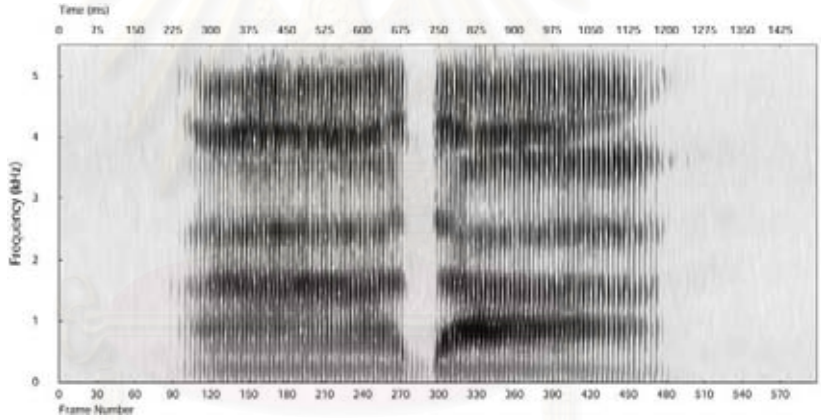
(c) /?aa0baa0/

Figure 3.12 Spectrographic information of the same releasing consonants in the different manners of articulation

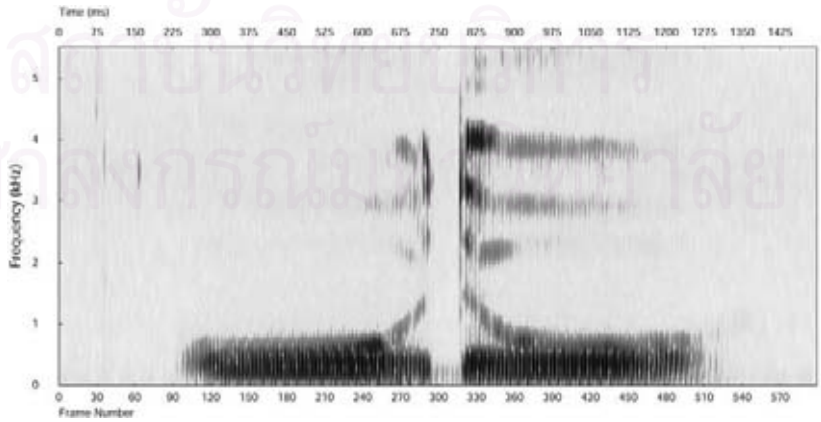
In Figure 3.13, spectrographic information of the syllables /?ii0dii0/, /?aa0daa0/, and /?uu0duu0/ are depicted. These syllables have the same releasing consonants but different vowel context. The transitional periods between the releasing consonant and its following vowel are obviously different, since the formant is moving towards the different vowel from the same locus of the releasing consonant.



(a) /?ii0dii0/



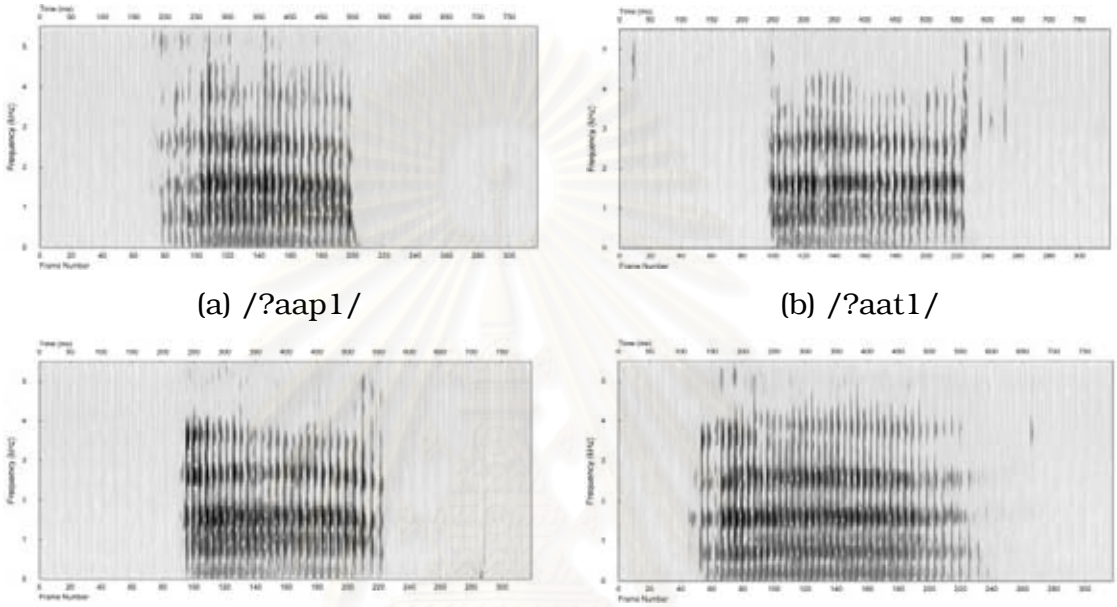
(b) /?aa0daa0/



(c) /?uu0duu0/

Figure 3.13 Spectrographic information of the same releasing consonants in the different vowel context

Apart from the releasing consonant, the three stops, [p], [t], and [k], appearing at the final position are acoustically different from the releasing consonant, that is, they are not audibly released. The spectrograms of these arresting consonants are illustrated in Figure 3.14. In consequence, the transitional period between the marginal sounds and the nucleus provides crucial acoustic information to identify consonants.



(a) /?aap1/ (b) /?aat1/

(c) /?aak1/ (d) /?aa0/

Figure 3.14 Spectrographic information of the arresting consonants

3.3 Summary

In this chapter, the Thai spoken language is described in terms of acoustic-phonetic. Acoustic-phonetic analysis is conducted on the Thai utterances to provide understanding of the Thai spoken language. Several acoustic features and feature extraction techniques are described to study the acoustic features of the Thai utterances. The analysis provides basic acoustic knowledge and solid background of the Thai spoken language. Deep understanding of the characteristics of Thai spoken language leads to the appropriate acoustic modeling of the speech unit, which is described in details in the following chapter.

Chapter 4

The Onset-Rhyme Acoustic Models

This chapter discusses properties of several speech units used in speech recognition. The strength and weakness of speech units according to the criteria of a good speech unit will be pointed out. From the previous chapter, the acoustic-phonetic analysis on the Thai language was conducted. The characteristics of vowels and consonants were thoroughly explored. Not only does the acoustic-phonetic analysis contribute strong knowledge, but it also provides acoustic cues for modeling the appropriate speech unit for the Thai language. The onset-rhyme units are proposed for use as a speech unit in speech recognition of the Thai language. Details of the onset-rhyme will be explained in this chapter. Finally, construction of the Thai continuous speech recognition system will be described.

4.1 General Speech Units

Several different approaches have been proposed for recognition of Western alphabetic languages with very large vocabularies. Based on these approaches, many successful prototype systems gave satisfactory performance (Lee, et al., 1990; Lee, et al., 1993; Lee, et al., 1997; Zue, et al., 1989; Rabiner, et al., 1989). One of the important issues in developing a successful speech recognition system is the selection of the appropriate speech unit. Selection of a set of speech units, usually including phonemes, phone-like-units (PLUs), syllables, subword units, or even smaller or larger units, is dependent on the target language. Apart from the issue of language dependence, the choice of speech units is usually dependent on the size of vocabulary to be recognized and the availability of sufficient training data for constructing effective models. Furthermore, the performance of a speech recognition system depends on the number of speech units. Three criteria, accuracy, trainability, and generalization must be considered in choosing the appropriate speech unit (Huang, et al., 2001). First, the speech unit should accurately represent the acoustic realization that appears in different

contexts. Second, the unit should be trainable for estimating the parameters of the unit with sufficient data. Finally, the unit should be generalized, so that any new word can be derived from a predefined unit inventory for task-independent speech recognition. A practical challenge is how to select a speech unit that meets these criteria. In this section, various speech units are compared, and their strengths and weaknesses in practical applications are pointed out.

4.1.1 Context-Independent Phone Units

In order to share phones across words, the subword unit has to be used. The smallest subword, phoneme or monophone, is a single model representing a phone in all contexts. Since there are merely 57 phonemes in Thai, they can be adequately trained with a few hundred sentences. However, the assumption of phoneme models that a phone in any context is identical to the same phone in any other context is entirely not true. Although each word is intentionally uttered as a concatenation sequence of phoneme, these phonemes are not independently produced because the articulator cannot abruptly move from one position to another. Consequently, the realization of a phoneme is greatly affected by its adjacent phones. The coarticulatory effects on phoneme /d/ in three different contexts are illustrated in Figure 3.13.

4.1.2 Context-Dependent Phone Units

Of the context-dependent phones, the diphone and triphone capture each phone in a particular context. Triphone modeling is much more powerful and consistent than diphone modeling because it can model the most important coarticulatory effect from its neighboring phones. However, too many different triphones need to be modeled for different context dependency on both sides.

Although, the triphone seems to be a good speech unit for acoustic modeling, there are many disadvantages in applying this context-dependent phone unit. Since the triphone is a phone-derivative unit, it inherits some limitations of phone-based approaches, namely the lack of an easy and efficient way for modeling long-term temporal dependencies (Ganapathiraju, et. at., 2001). Triphone unit spans an extremely short time interval. Consequently, integration of spectral and temporal dependences is not easy.

Moreover, since each triphone is modeled with a different context dependency, a large number of triphone patterns will be generated, leading to a great memory requirement and numerous models with poorly estimated parameters. Since the Thai language has a simple syllable structure, decomposing Thai syllables into triphone units produces an excessive number of speech units, which is an inefficient approach.

4.1.3 Words

Words are the most natural units of speech because they are exactly the units to be recognized. By modeling words as fundamental units, the phonological variations can be assimilated because they are able to capture contextual effects within words. Therefore, word models will usually achieve the best performance if there are sufficient training data. Speech recognition research in Thailand has been conducted with the word-based approach for a decade. Several vocabulary sets, isolated Thai numerals, isolated Thai words, and polysyllabic Thai words were recognized with various techniques (Ahkuputra, et al., 1997; Pornsukchandra, et al., 1997; Wutiwiwatchai, et al., 1998, Jitapunkul, et al., 1998; Ahkuputra, et al., 1998). Although a system using a word-based model achieves a high recognition accuracy, the vocabulary size is very limited (Ahkuputra, et al., 1997; Pornsukchandra, et al., 1997; Wutiwiwatchai, et al., 1998, Jitapunkul, et al., 1998; Ahkuputra, et al., 1998). In addition, many ambiguities occurred among the similar sounds, which resulted in incorrect classification.

Using word models in large vocabulary continuous speech recognition causes several severe problems. First, since training data cannot be shared between words, each word has to be trained independently. Many examples of words are required for adequate training data. Therefore, it is nearly impossible to get several repetitions of all the words, which is a major problem in large vocabulary applications. Second, the memory usage increases linearly with the number of words because of no sharing between words. Finally, it would be extremely inconvenient to the user when new words need to be added to the vocabulary, and new words can be easily generated every day. Hence, using word models for large vocabulary continuous speech recognition is not practical.

4.1.4 Syllables

Among the other non-phone-based units, syllables are also used as fundamental acoustic units for continuous speech recognition. Ganapathiraja (Ganapathiraja, et al., 2001) proposed the syllable models, which have many advantages over the phone-based units. First, since a syllable is perceptually defined, acoustical characteristics of a syllable relate to articulation and human perception. Second, a syllable acoustic unit provides compact representation of an utterance. Third, coarticulation effects are integrated within a syllable unit thus, making the unit acoustically stable. Moreover, the longer duration of a syllable simultaneously combines both temporal and spectral variations. These variations are then utilized during recognition. For these reasons, the syllables satisfy the consistency criterion. However, large numbers of syllables are required to cover the whole speech corpus or even the whole language. For the Thai language, the number of syllables that could be grammatically generated from combinations of consonants, vowels, and tones, is 26,928 units (Luksaneeyanawin, 1993). Therefore, the syllable units do not satisfy both trainability and generalization criteria.

4.1.5 Initials and Finals

According to the Mandarin Chinese syllable structure, every syllable is a morpheme, which has its own meaning, and each syllable is an open syllabic structure ending with a vowel or nasal /n/ or /ng/ (Lee, et al., 1993; Lee, et al., 1997; Chen and Liao, 1998; Chen, et al., 1998). Therefore, an Initial followed by a Final is used as the basic acoustic unit in Mandarin speech recognition. The Initial comprises the initial consonant of the syllable while the Final consists of the vowel or diphthong part, including the possible medial or nasal ending (Lee, et al., 1993). A set of 22 Initials and 38 Finals forms the number of 408 phonologically allowed different base syllables of Mandarin Chinese (disregarding tones). In addition, Cantonese is one of the most popular Chinese spoken languages. Similar to Mandarin, it is a bi-syllabic language with multiple tones. Cantonese consists of 20 Initials (including the null initial) and 53 Finals, which form the whole set of 595 syllables (disregarding tones) (Fu, et al., 1996). Because the Initial parts are usually very short compared to Final parts in base syllables and any important difference among the Initial parts of different syllables can be

easily influenced by irrelevant differences among the Final parts of the syllables during the recognition process, these produce a confusing set of Initials (Lee, et al., 1993, Wang, et al., 1997). Therefore, a set of context-dependent Initial models expanded from context-independent Initial models had been proposed to overcome those problems. The error rate was dramatically reduced by using context-dependent Initial models (Wang, et al., 1997). Although the context-dependent Initial is modeled along with its vowel context, the formant transition portion is not included in the model. Missing this important acoustic cue makes the acoustic unit imprecisely modeled.

4.1.6 Whole Word and Sub-word Modeling

Hidden Markov models can be used to model speech at several linguistic levels, ranging from phone, syllable, word etc. Definition of a good speech unit was previously elaborated in section 1.3. The previous section has discussed strength and weakness of the whole-word and the sub-word models corresponding to definition of a good speech unit.

Accurate acoustic models will improve discrimination and overall recognition performance. Trainability will guarantee generalization and better use of model parameters. The accuracy and trainability properties of speech units are illustrated in Figure 4.1.

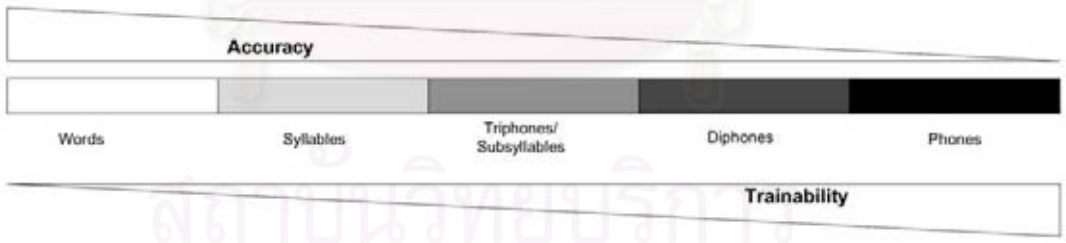


Figure 4.1 Trade-off between accuracy and trainability

Various speech units were reviewed and given details of their advantages and drawbacks for speech recognition. These speech units seem to be unsuitable for the Thai language. An alternate speech unit has to be taken into account. In this paper, the concepts of onset and rhyme are proposed and applied to a Thai speech recognition system. Details of the onset-rhyme model will be described in the next section.

4.2 The Onset-Rhyme Acoustic Models

4.2.1 Acoustical Properties

An onset consists of a releasing consonant and its transition towards the following vowel. Including a transitional portion between consonant and vowel, this onset model provides consonant-vowel (CV) and consonant cluster-vowel (CCV) combinations of Thai syllables. The onset model also provides crucial acoustic cues for classification of each releasing consonant, particularly for stop consonants. Additionally, the onset can deal with the intra-word coarticulatory effects better than the phone model. Along with the onset, a subsyllable rhyme is composed of a steady vowel portion and an arresting consonant. This model contains a whole vowel portion plus the arresting consonant if any. Including a vowel and a final consonant, the rhyme model provides monophthong-consonant (VC) and diphthong-consonant (VVC) combinations of Thai syllables (Maneenoi, et al., 2002; Jitapunkul, et al., 2003; Maneenoi, et al., 2004). The onset-rhyme segment together with other speech units is shown in Figure 4.2. Obviously, the onset covers the transition towards the vowel, which makes the onset precisely modeled.

From acoustical point of view, a pair of onset and rhyme contains an internal syllable juncture within a syllable whereas an external juncture appears between syllables. The internal juncture, which strongly binds the onset and rhyme together, can efficiently handle co-articulation within a syllable. On the other hand, the external juncture provides the crucial acoustic cues between the rhyme and the following onset of the adjacent syllable.

Since the realization of a context-independent phone unit is strongly affected by its neighboring phones, contextual information is needed to model speech units. The recognition accuracy of the speech recognition system using context-dependent speech units is significantly improved. Hence, context-dependent speech units have been widely used for several large-vocabulary speech recognition systems (Lee, et al., 1989; Zue, et al., 1989; Rabiner, et al., 1989; Lee, et, al, 1990; Jelinek, et al., 2001; Chow, et al., 1987). The onset-rhyme is context-dependent modeling. It contains both a releasing consonant and a vowel in the onset and a vowel plus arresting consonant in rhyme. The context-dependent units are able to model the

transitional portion between the releasing consonant and vowel in onset part. In addition, transitional stage between vowel and arresting consonant in rhyme part is modeled. Therefore, the onset-rhyme incorporates both left-context dependent and right-context dependent modeling.

The onset-rhyme should be suitable for representing Thai sound units for several reasons. According to the acoustic properties of Thai syllable, in the syllable structure, the final consonant is strongly influenced by the vowel duration. The duration of a final consonant following a short vowel or a weak vowel is longer than that of a final consonant following a long vowel or a strong vowel as shown in Figure 4.3. This relationship occurs only between the vowel and the final consonant. In contrast, the initial consonant is not affected by the duration of the vowel. Hence, the vowel and the final consonant are tightly tied while an initial consonant is loosely tied with the vowel in the syllable. Consequently, the decomposition of the syllable into an onset and rhyme is appropriate to the Thai language. The whole set of Thai syllables can be recognized by identifying onsets and rhymes.

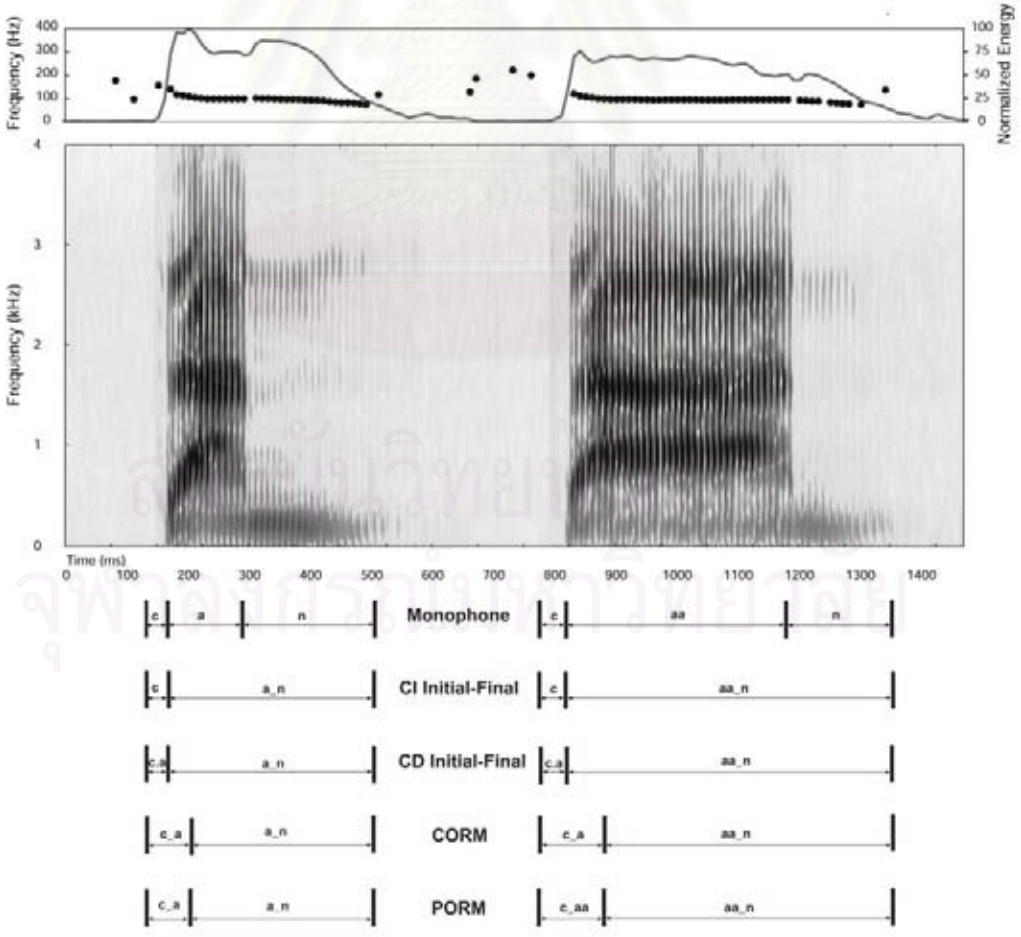


Figure 4.2 Various speech segments

Moreover, unlike other context-dependent phone units, the onset-rhyme is larger than the diphone and triphone. The onset-rhyme is modeled with a consonant, including its transitional stage in onset and the entire vowel along with the final consonant in the rhyme, while the same phone units in diphones or triphones are differently modeled according to their contexts. The use of an acoustic unit with a longer duration facilitates simultaneous exploitation of temporal and spectral variation (Gish and Ng, 1996). Consequently, these onset-rhyme units contain the most variable contextual effects at the beginning portion of the syllable in the onset and at the ending portion of the syllable in the rhyme.

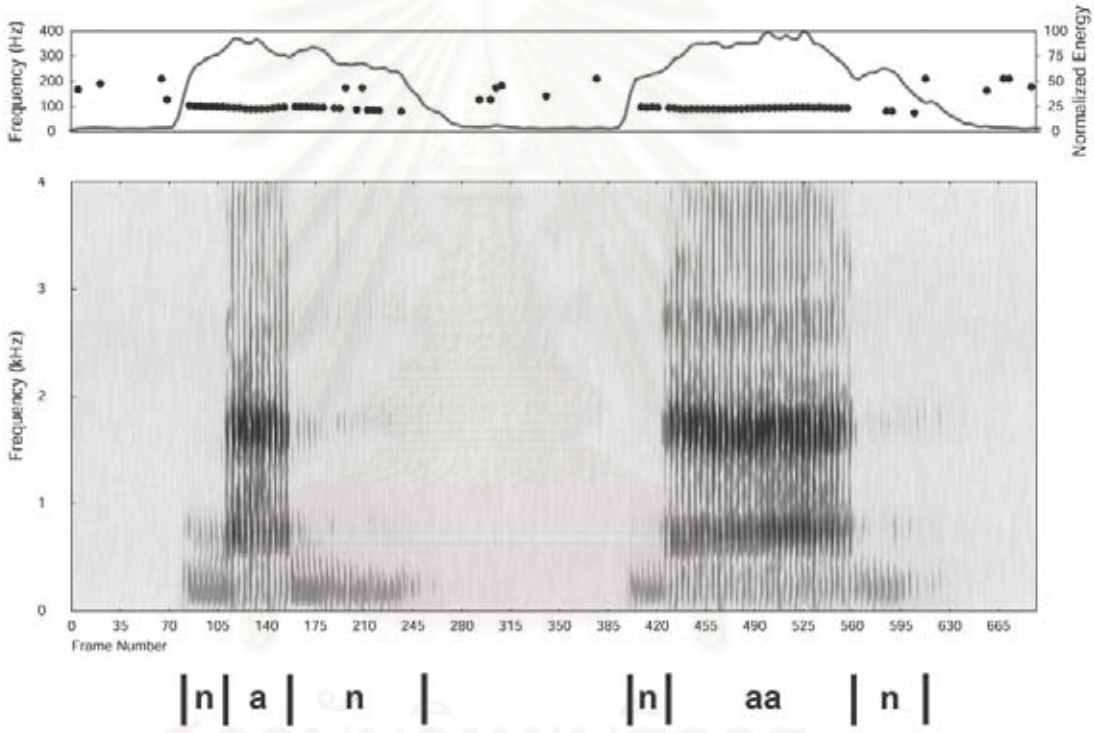


Figure 4.3 Relationship between vowel and final consonant duration

Acoustically, the rhyme contains crucial prosodic information within the segment. The prosodic features that the rhyme carries are tone, stress, accent, intonation, etc. (Luksaneeyanawin, 1992; Luksaneeyanawin, 1993; Thubthong, et al., 2002). The importance of these prosodies varies according to the language. For instance, the rhyme unit in Thai contains tone and stress information while only the stress and accent are provided in English. Tones in Thai are also influenced by the arresting consonant within the rhyme unit (Thubthong, et al., 2002). Although patterns of the same tone in

both obstruent and sonorant endings are different (Luksaneeyanawin, 1992; Luksaneeyanawin, 1993; Thubthong, et al., 2002), their variations are captured within the whole rhyme unit. These are major advantages of the onset-rhyme models over the phone-based models. In the phone-based models, not only the contextual information but also the prosodic information is lost when breaking up the nucleus and coda in the rhyme unit.

Not only does the rhyme contain contextual information, but it also contains prosodic information, including the tone, accent, stress, and intonation. In the Thai language, tones are governed by a vowel and an arresting consonant as shown in Table 3.4. These properties of the rhyme are important in modeling the proper speech units for the tonal languages. The onset-rhyme models have preserved crucial prosodic information within the models. Thubthong (Thubthong, et al., 2002) illustrated the use of tone information within the rhyme unit for tone recognition. The rhyme units provide better results than using the whole syllable or only the vowel segment. Therefore, using only tone information within a vowel segment is not sufficient. Chen (Chen, et al., 2001) used only tone information within a main vowel for tone recognition, which is not adequate for tone recognition since arresting consonants also have large effects on tone patterns. Both the vowel and arresting consonant, making up a rhyme unit, store some prosodic information that is crucial for tone recognition (Thubthong, et al., 2002).

4.2.2 Phonological Properties

From a phonological point of view, a syllable is composed of a pair of an onset and a rhyme unit, where the rhyme comprises a nucleus and coda as shown in Figure 4.4. An onset consists of an initial consonant and its transition towards the following vowel. Along with the onset, the rhyme is composed of a vowel, a final consonant, and a tone. The onset-rhyme not only includes its context information, but also embeds the language modeling at the syllable level. Recognition accuracy can be greatly improved by taking advantage of possible a priori information on the sequences to be recognized. An automatic speech recognition system can successfully use language information if such knowledge is embedded in a language model. Composing the onset and rhyme forms a syllable according to the syllable

structure. The rhyme must have the same vowel as the preceding onset. This indicates that language modeling is embedded in the unit of onset and rhyme.

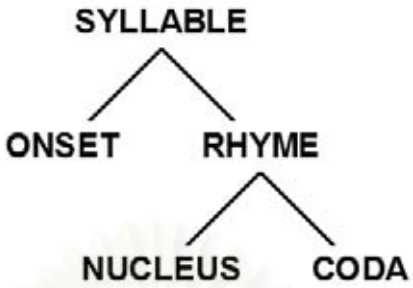


Figure 4.4 Syllable segment

In Figure 4.5, all of the phones, diphones, triphones, and onset-rhyme units are illustrated with regard to their physical and logical representation of speech. A speech signal is assumed to be composed of phone sequences as shown in Figure 4.5 with their physical speech segment and location in the phrase /khaw4 svv3 phaan0 naj0/. Each phone occurs independently without using any contextual information. Thus, context-dependent phones, the diphones and the triphones, are physically similar to the phones but logically differ depending on specific context. For instance, the phone /a/ in /khaw/ and /naj/ are in different context, which are then separately modeled as /kh-a+w/ and /n-a+j/, respectively, as shown in Figure 4.5. However, the context-dependent phones still use the same speech segment as the phones without taking into account on any articulatory effects between each phone. Consequently, the context-dependent phones do not effectively handle any coarticulation between speech segments, which contain crucial acoustic information. Unlike other units, the onset and rhyme units efficiently model coarticulatory effects both within syllables and across syllables. The internal syllable junctures reside within a syllable, tying a pair of onset and rhyme units together and treating coarticulation between releasing consonant and vowel. Within the rhyme unit, the vowel and arresting consonant are tightly tied together to preserve their coarticulation. Also, there are external syllable junctures that consider coarticulation between syllables, which provide acoustic cues between the rhyme and the neighboring onset in the following syllable. These external syllable junctures are syllable boundaries, which are explicitly located and

combined into the models. Examples of both junctures are depicted in Figure 4.5.

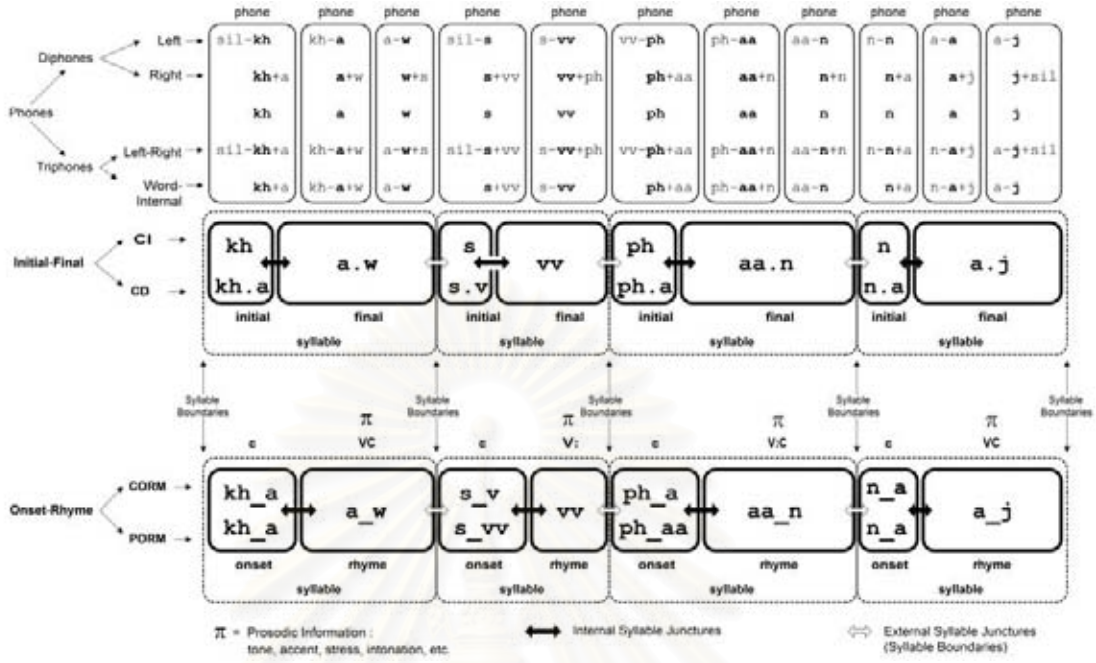


Figure 4.5 Representation of various speech units – phones, initial-final, and onset-rhyme

4.2.3 Types of Onset-Rhyme Models

By considering the duration of the releasing consonant plus its transition preceding different vowel contexts, the onset-rhyme models are defined into two types (Ahkuputra, 2002; Jitapunkul, et al., 2003; Maneenoi, et al., 2004): (1) Phonotactic Onset-Rhyme Model (PORM) and (2) Contextual Onset-Rhyme Model (CORM). These two models are generated from different combinations between the releasing consonant and vowel. According to the duration of the releasing consonant and its transition, the phonotactic onset is created differently for each releasing consonant and each vowel context, even for vowels in the same short-long pair. On the other hand, the contextual onset is modeled similarly for a given releasing consonant an either member of same short-long vowel pair. The number of phonotactic onset and contextual onset units are 792 and 297, respectively, while both models have the same 200 rhyme units as described in Tables 4.1 and 4.2. Based on the onset-rhyme models, a speech recognition system forms syllables using the network of onset and rhyme HMMs. The HMM networks of two onset-rhyme models are depicted in Figures 4.6 and 4.7, respectively. This research will explore both two types of onset-rhyme models in order to

determine which type will be more efficient and suitable for Thai speech recognition systems.

Table 4.1 Number of the rhyme units

Thai Syllable	V	C_f	Units
1 Open syllable			
1.1 Open long syllable	12		12
1.2 Open short syllable	12		12
2 Sonorant ending syllable			
2.1 Short syllable ending with sonorant consonant	12	5	60
Inadmissible co-occurrences			
Round vowel unit preceding a labialized sonorant consonant	4	1	-4
Front vowel unit preceding a palatalized sonorant consonant	4	1	-4
2.2 Long syllable ending with sonorant consonants	12	5	60
Inadmissible co-occurrences			
Round vowel unit preceding a labialized sonorant consonant	4	1	-4
Front vowel unit preceding a palatalized sonorant consonant	4	1	-4
3 Obstruent ending syllable			
3.1 Short syllable ending with obstruent consonant	12	3	36
3.2 Long syllable ending with obstruent consonant	12	3	36
Total			**200**

Table 4.2 Number of the onset units

Onset	**Combination**	**Units**
Contextual onset	$33C_i$ x 9V	297
Phonotactic onset	$33C_i$ x (18V + 6VV)	792

4.2.3.1 Phonotactic Onset-Rhyme Model (PORM)

The onset units of the PORM are generated from combinations of releasing consonants in all possible vowel contexts. Each phonotactic onset is created differently, according to the duration of the releasing consonant and its transition preceding the vowel, even for following vowels is the same short-long pair. Except during the transitional period, the patterns of formant transition of the same releasing consonant with different vowel contexts are similar as indicated in Figures 4.8 and 4.9. Different combinations of the releasing consonants and following vowels lead to 792 possible PORM onsets. Figure 4.10 shows the formant transitions of the releasing consonant [n] occurring in three different vowel contexts [i, ii, iia]. By considering the difference of a releasing consonant plus transitional period in each vowel context, onset units of PORM are individually modeled. For instance, onsets

consisting of the releasing consonant [n] occurring before vowels [i, ii, iia] are separately modeled as [n_i, n_ii, n_iia]. According to their neighboring vowels, the PORM onsets are thoroughly modeled. Consequently, PORM will produce the most accurate onset units due to its completely contextual modeling.

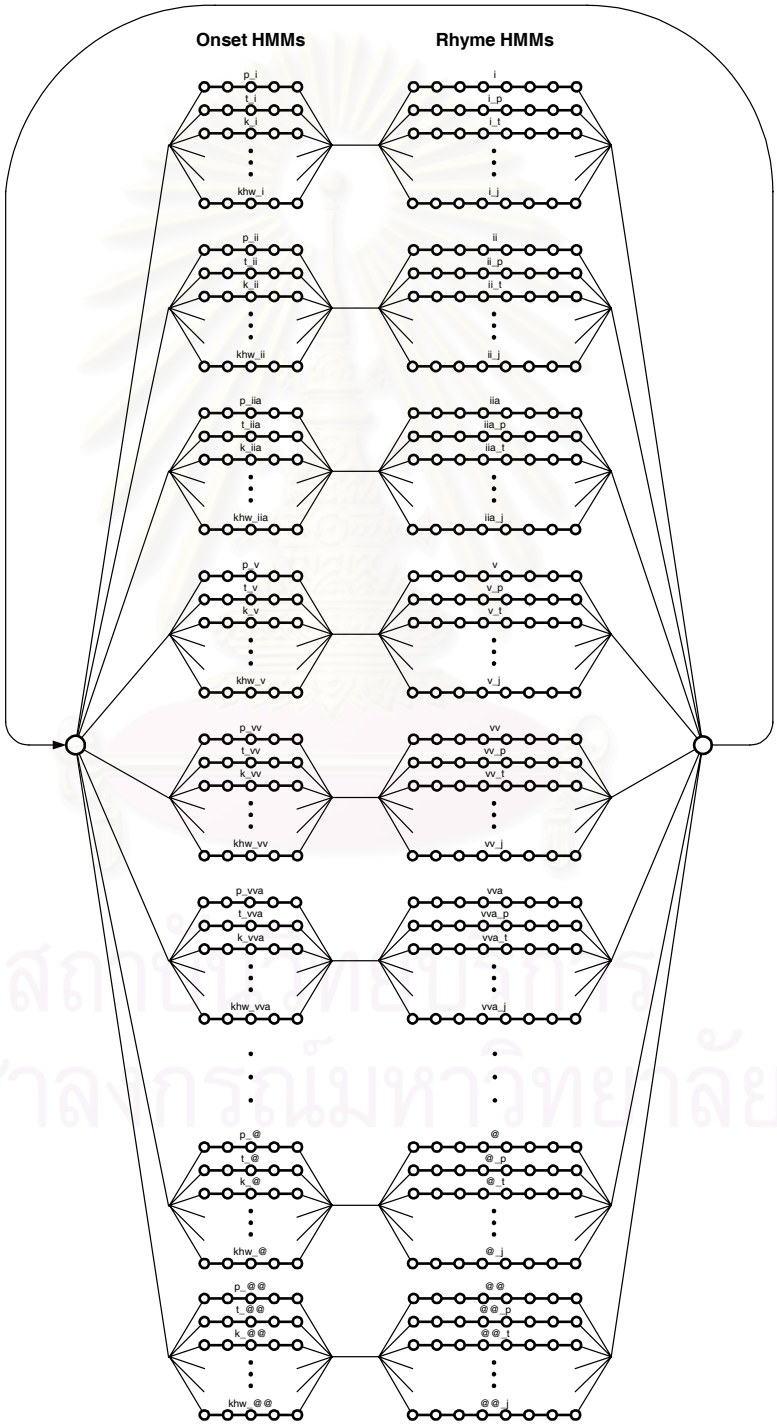


Figure 4.6 Network of phonotactic onset HMMs and rhyme HMMs

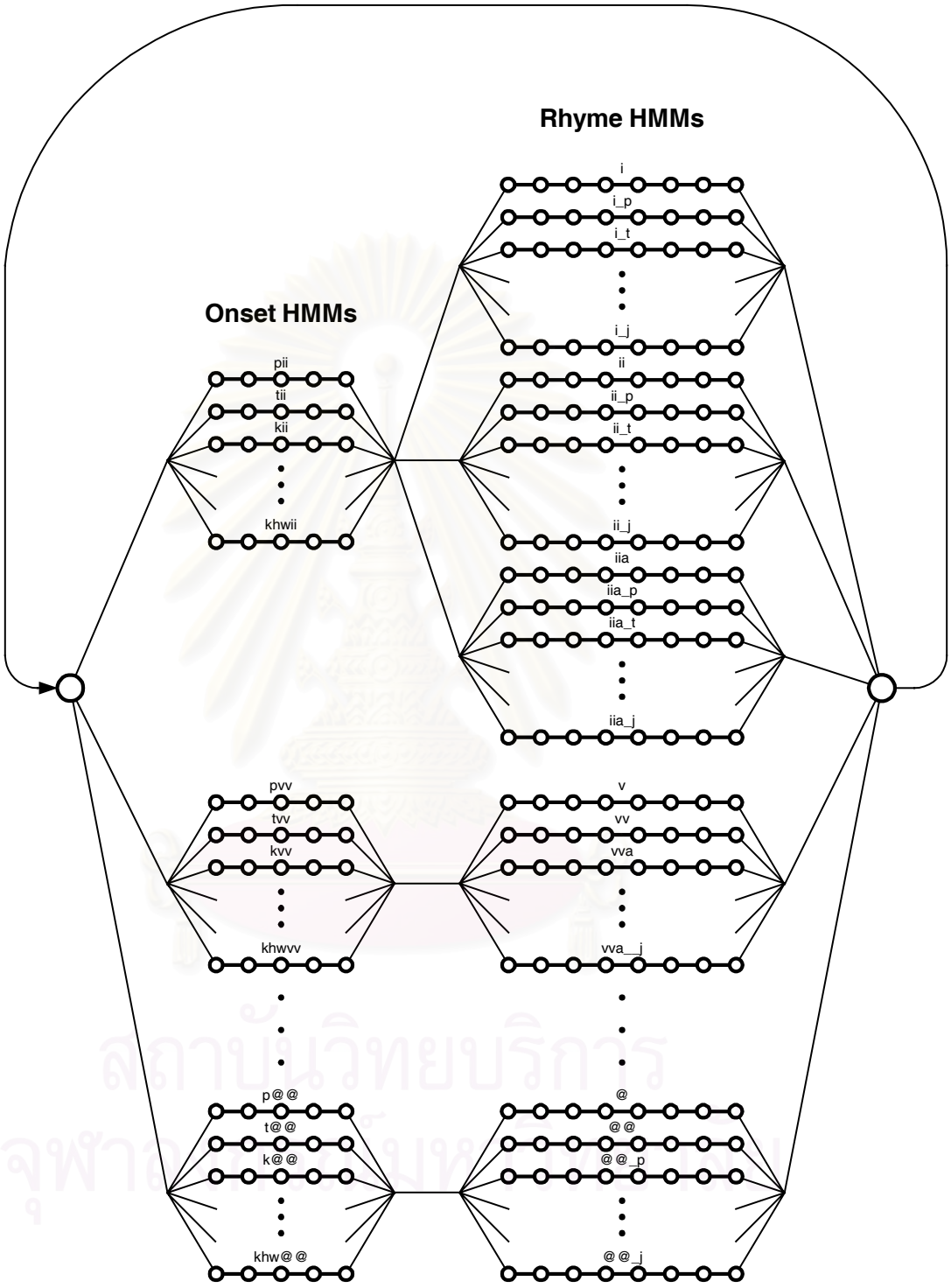


Figure 4.7 Network of contextual onset HMMs and rhyme HMMs

4.2.3.2 Contextual Onset-Rhyme Model (CORM)

The contextual onset-rhyme models are proposed in this paper along with the phonotactic onset-rhyme models. Acoustic analyses conducted on Thai syllables showed similar patterns of formant transitions in particular cases. Apart from the duration of formant transition of male and female speakers as indicated in Figures 4.8 and 4.9, the formant patterns are similar in both short and long vowel contexts for a given releasing consonant. Figure 4.11 shows the formant transitions of the releasing consonant [n] occurring in three different vowel contexts [i, ii, iia]. By disregarding the duration of the formant transition, these onset units can share formant-transitional information. Therefore, combining similar onsets for short-long vowel pairs with the same releasing consonant substantially reduces the number of onset units. The number of these so-called contextual onsets is reduced to 297 in comparison with the 792 phonotactic onsets. CORM gives a lower complexity in terms of search space than PORM, which has a larger number of units. As a result, with fewer onset candidates, the CORM network performs faster in the decoding process while it still produces the same number of syllables as PORM does.

Compared with other context-dependent phone units, the number of onset-rhyme units in the Thai language is smallest. The numbers of possible combination units are summarized in Table 4.3. Consequently, with a small number of onsets and rhymes, a remarkably small database size is required for modeling of the onset-rhyme, compared with those for diphones and triphones. This makes a recognition system more manageable.

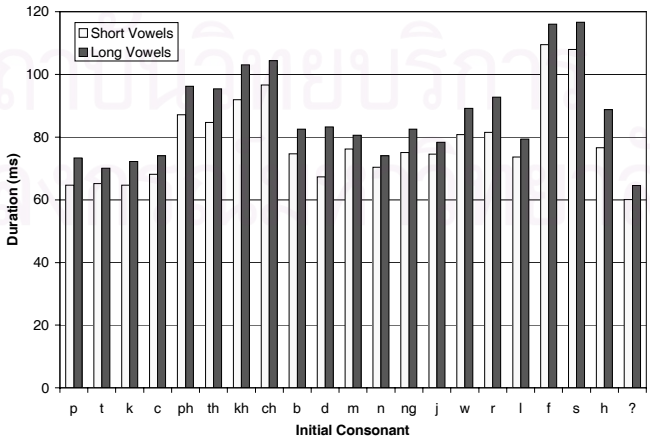


Figure 4.8 Duration of initial consonant preceding short and long vowels for 6 male speakers

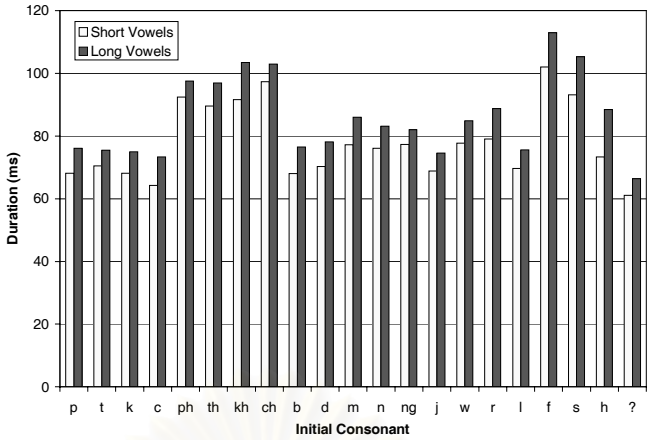
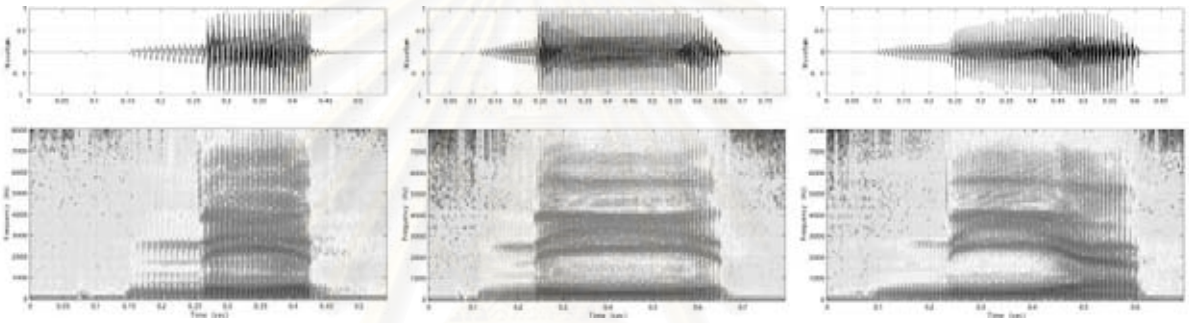


Figure 4.9 Duration of initial consonant preceding short and long vowels for 6 female speakers



(a) /nit3/ (b) /niit2/ (c) /niiat2/

Figure 4.10 Spectrogram of the syllables /nit3/, /niit2/, and /niiat2/

Table 4.3 Numbers of various speech units applying to the Thai language

Speech Unit	Possible	Speech Unit	Possible
monophone	58	CI Initial-Final	33I + 200F
intra-syllable	1, 042 (left) / 1,041	CD Initial-Final	297I + 200F
inter-syllable	1,913	CORM	297O + 200R
intra-syllable	7,769	PORM	792O + 200R
inter-syllable	64,475	syllable	26,928

4.3 Construction of the Thai Continuous Speech Recognition System

4.3.1 Thai Speech Corpus

One of the important issues on construction of a speech recognition system is creating a speech corpus. Since there is no Thai continuous speech corpus available for the research, it is necessary to create the new Thai

continuous speech corpus. This section describes the procedure in creating a Thai continuous speech corpus, beginning with design, then recording, and labeling of the corpus.

4.3.1.1 Design a Thai Continuous Speech Corpus

In this research, the reading speech was selected for less significant coarticulatory effects and pronunciation variations. A Series of Aesop's Fables in Thai was selected because it does not contain any foreign words. Initially, seven Aesop's Fables were analyzed on the distribution of phone and onset-rhyme units. This set of data contains about a hundred sentences. In order to create an initial acoustic model of onset-rhyme, a number of training samples must be sufficient. Therefore, a new set of sentences was composed in order to fulfill the insufficient onset-rhyme units. The Royal Thai Dictionary was additionally used to find out all possible onset-rhyme units existing in the Thai language. Finally, a speech corpus contains a set of 111 sentences from Aesop's Fables, 550 sentences from a new composed set, and 420 sentences from reading paragraphs. There are total 23,790 syllables in the training set.

To evaluate the speech recognition system, a set of test sentences has to be created. In this research, a set of 100 sentences was excerpted from five different reading stories – Thai central geography, Encyclopedia of butterfly, Solar system, Cultivation of rose, and "Doi Suthep" national park. The sentences, excerpted from those stories, are more natural in reading than the composed sentences. There are total 4,985 syllables in these test sentences.

4.3.1.2 Recording of Thai Utterances

Recording of Thai sentences was taken in the quiet laboratory environment. The speech data were recorded with 16 bit resolution and 16 kHz sampling frequency. Two different microphones were used to record simultaneously. The stereo-recorded data were separate into the left and the right channel. This recording gives two different output utterances from one utterance. A complete set of training and test sentences was recorded from 9 male and 11 female speakers. The other 5 male and 5 female speakers recorded only a set of test sentences. The total durations of the speech corpus used in this

experiment for training and testing are approximately 68 hours and 36 hours, respectively

4.3.1.3 Labeling of the Recorded Thai Utterances

The initial set of sentences was label manually according to their transcriptions by "Speech Labeler" program developed by the Thai Speech Processing Research Group, Digital Signal Processing Research Laboratory (Ahkuputra, 2002). User interface of the program is illustrated in Figure 4.11. The output label transcriptions are in phones and onset-rhyme models conforming to the Hidden Markov Model Toolkit (HTK) format (Young, et al., 1999).

The manual label transcriptions were used to create the initial acoustic models. Then, these acoustic models were used to create the automatic labeling system. The automatic labeling system aligns the phonetic transcription of the sentence automatically. The alignment results needed little correction.

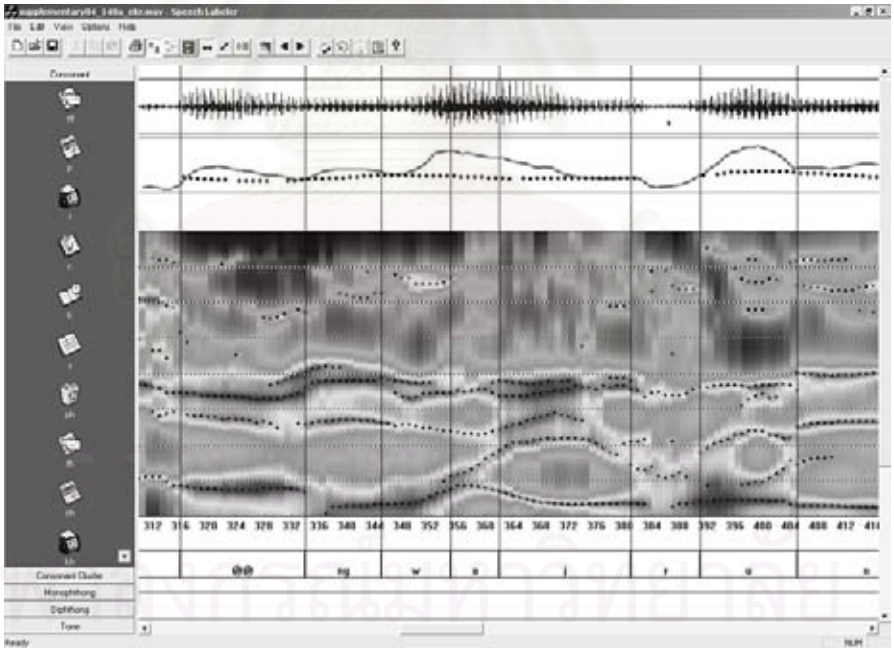


Figure 4.11 "Speech Labeler" program

4.3.2 Speech Signal Processing and Feature Extraction

The speech samples were passed through a signal preprocessing routine consisting of signal pre-emphasis and a smoothing window. In the signal pre-emphasis step, the first-order FIR filter is used for flattening the spectrum (Rabiner and Juang, 1993; Lee, et al., 1989; Lee, et al., 1990;

Juang and Furui, 2001; Furui, 2001). The Hamming window was applied in order to divide the speech signal into frames. Since the MFCCs are the speech parameterization for many speech recognition systems (Mokbel and Chollet, 1995; Vergin, et al., 1999), this research employed the MFCC for representing speech signals. The dynamic feature, the temporal derivative, contributes significantly to improvement of recognition performance. Therefore, the MFCCs were applied and the temporal derivatives were additionally utilized (Maneenoi, et al., 2002; Ahkuputra, 2002; Jitapunkul, et al., 2003; Maneenoi, et al., 2004).

4.3.3 Acoustic Modeling of Speech Units

This section describes the implementation of acoustic modeling of various speech units. This research mainly used four speech units. The context-independent phone, a monophone, was modeled initially, and then this speech unit was used for building the triphone system. The modeling of other speech units, Initial-Final and onset-rhyme, depended on their types. The Initial and the onset were modeled differently, according to their context, while both of the final and the rhyme were modeled as left context-independent units.

Model parameters were initiated and re-estimated using the standard Viterbi alignment process and Baum-Welch algorithm together with the labeled transcriptions. A set of initial acoustic models was then trained with the embedded Baum-Welch algorithm in which a composite model for each complete sentence was used to probabilistically assign observations to states and then update the model parameters with only the unlabeled transcriptions.

To achieve a higher performance, an iterative divide-by-two clustering algorithm was utilized to increase the Gaussian mixture component. The complexity of the models was increased in this mixture incremental. Experimental results will be reported benefit in varying the number of Gaussian mixture components.

The training process of the monophone, Initial-Final, and onset-rhyme models is generally similar as depicted in Figure 4.12. To the trainability problem of the triphone models, the training process is more complex than the others. The acoustic model construction of these speech units will be described in the following section.

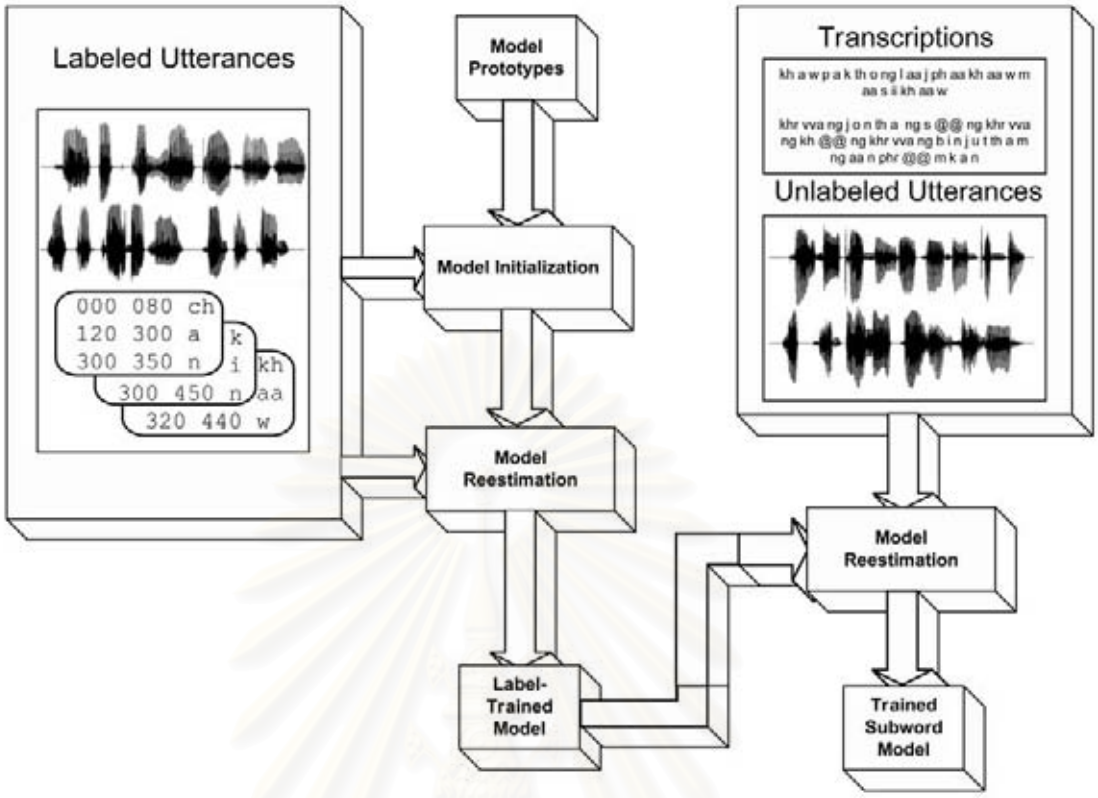


Figure 4.12 General training process

4.3.3.1 Construction of the Context-Independent Phone Model — the Monophone Model

The phone models use standard three state left-to-right topologies with no skip state. Primarily, monophone models were initiated using a single Gaussian observation distribution from the labeled data. The standard Viterbi alignment process and Baum-Welch algorithm were applied to obtain the initial acoustic models. A set of initial acoustic models, was then trained with the embedded Baum-Welch algorithm. To achieve a higher performance, an iterative divide-by-two clustering algorithm was utilized to increase the Gaussian mixture component.

4.3.3.2 Construction of the Context-Dependent Phone Model — the Triphone Model

4.3.3.2.1 Syllable Boundaries

The presence of syllable boundaries in the phone sequences complicates the use of context-dependent phonetic models and this can be dealt with in one of two ways.

A. Intra Syllable Context Dependency

Syllable boundaries represent a distinct context and further expansion of the context across syllable boundaries is blocked.

CHAN KIN KHAAW = **sil ch+a ch-a+n a-n k+i k-i+n i-n kh+aa kh-aa+w aa-w sil**

B. Inter Syllable Context Dependency

Expansion of context can occur into surrounding syllables. The presence of syllable boundaries can be either ignored or used as additional contextual information.

CHAN KIN KHAAW = **sil sil-ch+a ch-a+n a-n+k n-k+i k-i+n i-n+kh n-kh+aa kh-aa+w aa-w+sil sil**

In continuous speech, as opposed to isolated word speech with each word delimited by silence, co-articulatory effects occur across the syllable boundaries since these often have important acoustic significance. However, there are several remarks when using the inter syllable triphones and the intra syllable triphones.

- **Size**

The total number of contexts is much smaller than in the inter syllable case because many contexts will never appear in a corpus. A greater proportion of contexts will be seen in the training data. Moreover, the problem of unseen contexts is less important in this case. The total number of contexts depends on the dictionary but, for modeling the whole Thai language, an intra syllable triphone system needs models for 7,769 distinct contexts while an inter syllable triphone system requires

64,775 models. However, only 10,642 of these appear in the training data.

- **Complexity**

With an intra syllable triphone system, every realization of a syllable is the same and can be taken straightforward from a dictionary. With an inter syllable triphone system, the choice of the first and last models of each syllable depends on the preceding and following syllables. The networks for the decoding process of an intra syllable triphone and an inter syllable triphone systems are shown in Figures 4.13 and 4.14. Obviously, the inter syllable triphone system greatly complicates the decoding process.

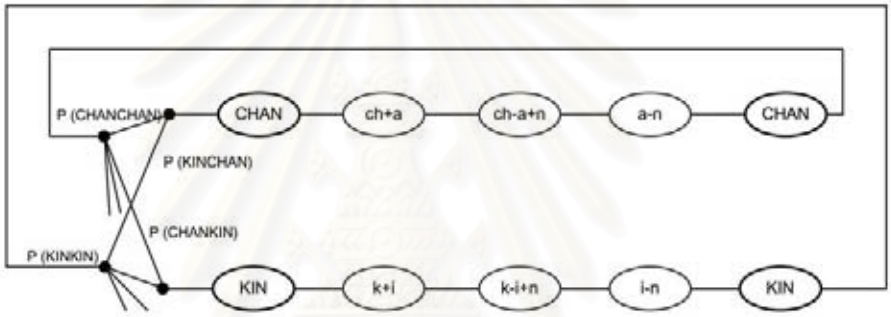


Figure 4.13 Intra-syllable triphone network

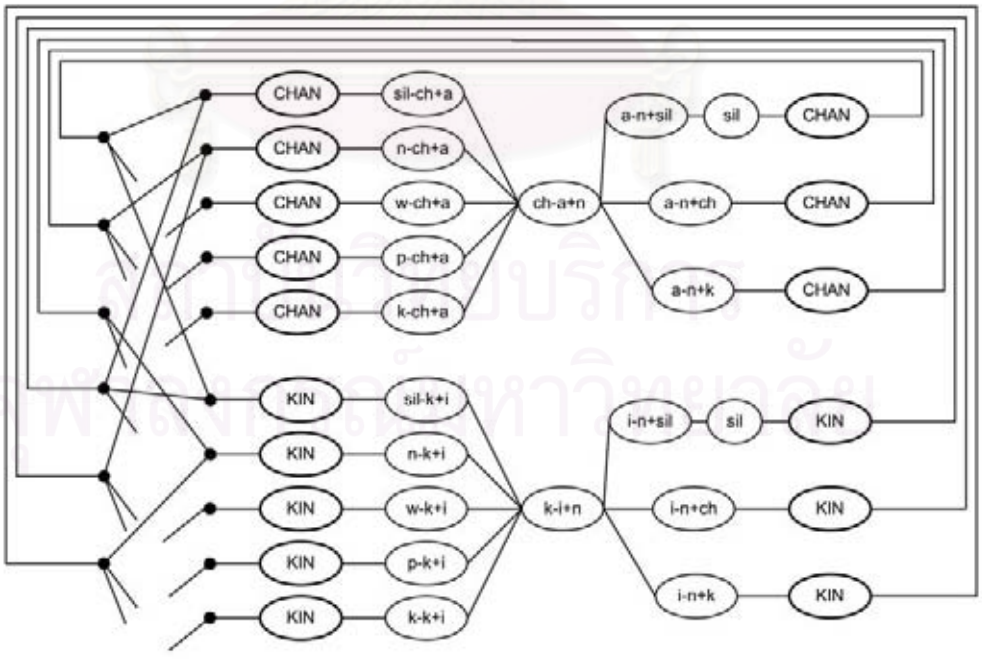


Figure 4.14 Inter-syllable triphone network

4.3.3.2.2 Trainability Problems

Due to the fact that the acoustic realization of the phonemes depends heavily on the phonetic context, it is essential for efficient speech recognition to model this context dependency (Lee and Hon; Lee, et al., 1990). The most commonly used context dependent phoneme model is the phoneme model in a triphone context. Although triphones provide an excellent modeling of context dependency, their exclusive used as acoustic models is prohibitive for vocabulary-independent speech recognition because the set of triphones in the recognition vocabulary often contains triphones that cannot be observed in the training. Another serious problem is that many triphones occur very seldom in the training corpus so the estimation of the models may not be reliable. It is impractical to train a separate model from only a few occurrences especially if mixture Gaussian distributions are used. Since there are many triphone contexts, which occur only a few times in the training corpus and many more than that do not occur at all, special methods must be made to assure that a triphone system is trainable and its parameters can be estimated reliably. This trainability problem becomes even more serious if larger amounts of context are to be taken into account.

There are several ways in which the trainability problem can be relieved (Odell, 1995):

- ***Backing-Off***

  When there is insufficient data to train a given model, it is possible to back-off and use less specific model for which there is enough data. For example, a diphone model could substitute for a triphone, which has only a few examples in the training corpus. If there were few occurrences of that diphone, a monophone model could be used. This guarantee the model used is well trained but it can mean that relatively few models will have full triphone context especially if the training corpus is relatively sparse.

- ***Smoothing***

  In order to maintain a greater degree of context dependency, it is possible to smooth the parameters of more specific model with those of the less specific model. One way in which this can be accomplished is to use interpolation between the less and more specific models with the interpolation weights chosen using deleted

interpolation (Lee and Hon, 1989). This method preserves the context dependency of the unsmoothed models but increases their robustness by effectively sharing training corpus from other contexts to produce more accurate parameters.

- ***Sharing***

  Another method for increasing the robustness of the system is to explicitly share models or parts of models between different contexts (Young, 1992). This method is sensible since the acoustic realizations of a phone occurring in different contexts are often very similar. This method also ensures that all the system parameters are well trained while maintaining the model context dependency.

All these techniques require that a choice is made about which parameters are backed-off to, smoothing with, or shared with others. For a simple backing-off strategy, a simple and obvious hierarchy exists; triphones are more specific and less trainable than biphones, which are more specific and less trainable than monophones. The weakness of this scheme is big jumps in specificity. Similarly, smoothing of parameters must occur through some forms of hierarchy. However, more flexibility is possible since different parts of a model can be smoothed in variable proportions with different models. For instance, initial state can be smoothed with a left diphone to preserve as much left context dependency as possible, while the final state can be smoothed with the corresponding right diphone.

Finally, sharing presents even more possibilities. Parameter sharing between models of the same complexity is possible as well as sharing with models further up the hierarchy. To improve the robustness of the parameter estimation, the emitting probabilities of the triphone states are shared between clusters of states, which are similar according to a distance measure. The training data assigned to the states of one cluster in used to estimate the shared emitting probability of these states. Sharing schemes can be divided into two approaches, bottom-up and top-down.

4.3.3.2.3 Bottom-up Approach

Bottom-up approach to the data insufficiency problem starts by assuming that all contexts are distinct, but to ensure that the parameters of each model can be reliably estimated, some forms of sharing is required. This

method can be accomplished by examining the original models and determining sets that can share parameters. These sets are chosen to ensure that the resulting models can be robustly estimated, and that members of the sets are sufficiently similar to ensure that the model will provide accurate representations.

A. Generalized Triphones

One of the methods to implement parameter sharing is to compare models from different triphone model contexts and merge those similar models. The merged models will be estimated from data. If the more realizations of the triphone in the different contexts are similar, the more accurate models are obtained. The generalized triphones evolved from the discrete distribution show better performance than triphones, which were smoothed with less specific models using deleted interpolation (Lee, et al., 1990; Deng, et al, 1992). However, sharing at the model level may not be appropriate method for triphone models composed of distinct states (Odell, 1995).

B. State Clustering

Sharing distributions at the state level share the output distributions among states. This sharing is constrained so that distributions are specific to a particular state position in a particular phones and they are only shared among the same state occurring in different contexts. The clustering is performed on single Gaussian diagonal covariance models in two stages;

- An iterative merging procedure, which merges the most similar pair of distributions according to the minimum distance between them. This stage terminates when this minimum distance exceeds a predetermined threshold.
- A merging procedure ensures the trainability of the models by ensuring that the occupancy γ of each tied distribution exceeds some thresholds. Each distribution with occupation counts below this threshold is merged with the nearest distribution (with the minimum value of $d(i, j)$).

Data-Driven clustering is performed by placing all states in individual clusters (Anderson, et al., 1994). The pair of the clusters, which when

combined would form the smallest resulting cluster are merged. This process repeats until either the size of the largest cluster reaches the threshold or the total number of clusters has fallen to the specific value. The size of cluster is defined as the greatest distance between any two states. The distance metric depends on the type of state distribution. For single Gaussian distributions, a weighted Euclidean distance between the means is used, and for tied-mixture systems, a Euclidean distance between mixture weights is used. For all other cases, the average probability of each component mean with respect to the other state is used.

The distribution between distributions, which will initially be for a single context and later for a cluster of contexts, i and j is calculated using

$$d(i,j)=\left[\frac{1}{n}\sum_{k=1}^{n}\frac{\left(\mu_{ik}-\mu_{jk}\right)^2}{\sigma_{ik}\sigma_{jk}}\right]^{\frac{1}{2}} \tag{4.1}$$

where n is the dimension of the data, μ_{sk} and σ_{sk} are the mean and variance of the k^{th} dimension of the Gaussian distribution of state s, either i or j. The values γ_s for the untied distributions are calculated during preceding training and then summed to give the occupation counts after merging.

This procedure results in a set of models with Gaussian probability distributions for clusters of contexts with similar acoustic features and ensures that each distribution has enough training samples to accurately estimate its parameters. This procedure reduces the total number of distributions significantly but results in a much smaller reduction in the number of distinct models because different models may share two state distributions and only differ in the final one. This preserves a higher degree of context dependency by allowing for contextual factors that only effect part of a phone. For instance, models with the same right context but different left contexts may have different initial state distributions while sharing those for the final and center states.

The main drawback of the bottom-up approach is that for triphones, which were not observed in the training corpus, no tied model is available. These unseen triphones are modeled by so-called backing-off models (Aubert, et al., 1996). Usually these models are simple generalizations of the triphones such as diphones or monophones. The training of the backing-off

models is performed on the data of the triphones, which were not involved in the clustering process.

4.3.3.2.4 Top-down Approach

The bottom-up approach is limited, as it requires examples of each context to produce initial estimates of the model parameters used in the clustering procedure. It is impossible to use such approach to construct models for contexts that may occur during recognition but do not appear in the training corpus. It is also unreliable for contexts that only occur a few times, since the examples may be unrepresentative and so the parameter estimates used during clustering will be inaccurate.

This problem can be minimized by ensuring that the training data gives adequate coverage of the models needed for recognition. However, this solution seems to be possible only for small vocabularies and system using word-internal context dependency. For large vocabularies and cross-word context dependent systems, it is virtually impossible to ensure that the training data will include examples of every possible context.

Using the top-down approach based on the decision trees avoids the problem of unseen models by using linguistic knowledge together with the training data to decide which contexts, including the unseen ones, are acoustically similar (Odell, et al., 1994; Odell, 1995).

A decision tree for each phoneme selects which of a set of models is used in each context. The model is chosen by traversing the tree, starting from node then selecting the next node depending on the answer to a simple question about the current context. For binary decision trees, these questions will normally be yes/no questions concerning membership of particular sets of phones.

For example, in the decision tree shown in Figure 4.15, the root question is answered by checking if the immediately preceding phone, the left context, is a nasal – m, n, ng. If the actual context is n-i+t, the next question to be asked would concern whether the following phone was an approximant, j and w. Since t is not a member of this set and the answer no results in a terminal node, the model labeled C would be used in the context.

This procedure has several advantages over bottom-up approach

- The hierarchical structure and the form of the questions mean that the tree will find an equally context dependent model for every

context. This method does not require the backing-off technique to less specific models for contexts that have not occurred in the training data.

- Expert knowledge can be incorporated in the form of set of questions that are used to split each node of the tree and this will be used to determine which contexts are similar to any unseen ones.
- The construction procedure can be constrained to ensure that leaf nodes are only generated for sets of contexts that have sufficient examples in the training data to reliably train an accurate model. The clustering does not suffer from the use of under-trained parameters.
- A greater degree of context dependency than triphones can be implemented by extending the types of questions.

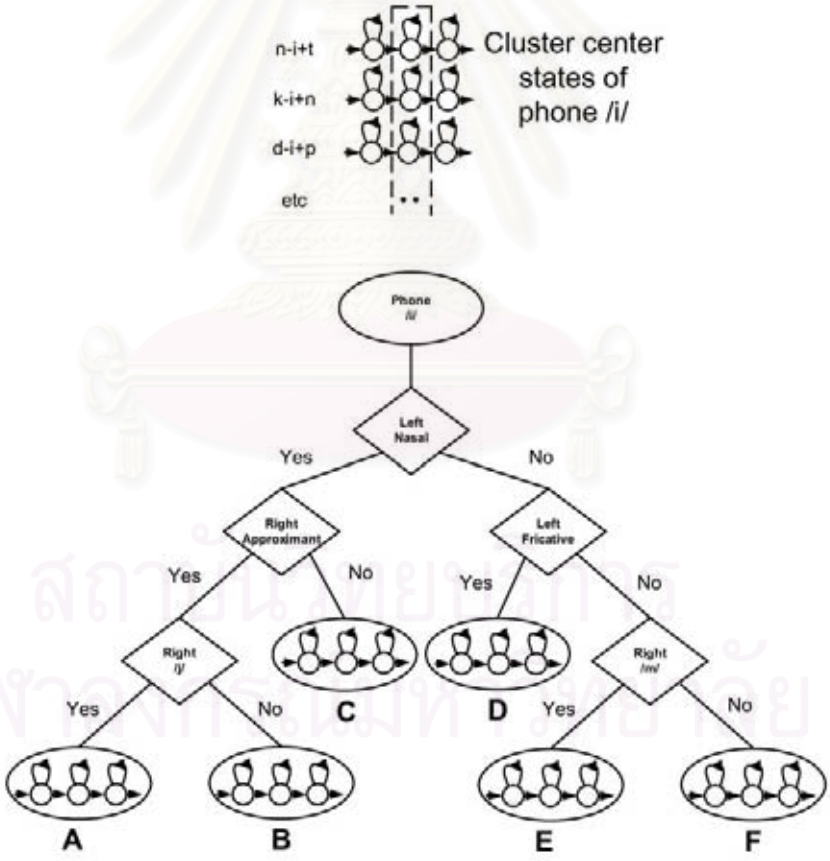


Figure 4.15 A decision tree

A. Decision Tree

To exploit the advantages of the top-down decision tree based approach, it is necessary to be able to automatically construct the trees. The construction procedure should aim to ensure that the resulting set of models provide an accurate and robust estimate of the underlying speech. The trees were constructed in locally optimized fashion starting from a single root node representing all contexts. As each node is created, an optimal question chosen from a finite set is selected to maximize the increase in probability at the resulting terminal nodes generating the training samples. Then the current set of terminal nodes is searched to find the one, which can be split using its optimum question to provide the largest increase in the total probability of the training data. If the increase of probability exceeds a threshold and the number of training samples associated with the node exceeds a threshold, the node is divided using the optimum question and two new terminal nodes were created. When none of the terminal nodes can be split, the procedure terminates and the tree is finished.

Several experiments suggested that sharing at the state distribution rather than at the model level led to improved performance (Hwang and Huang, 1992, Hwang and Huang, 1993). This state-tying approach also has the benefit of simplicity since the underlying model topology is used while the additional alignment technique, such as linear or dynamic time warping, does not need during constructing the trees.

The aim of state-tying is to reduce the number of parameters of the speech recognition system without a significantly degradation in accuracy. The states of the triphones used in training, which are similar according to a distance measure are tied together. First, a suitable triphone list is assembled according to the training corpus. Because this list has to be quite large to achieve an accurate modeling of the acoustic context, simple models are used for emitting probabilities – one Gaussian density with full or diagonal covariance matrix. Using a segmentation of the training data, the mean and the variance of the triphone states are estimated. The triphone states are then subdivided into subsets according to their central phoneme and their position within the phoneme model. Inside these sets, the states are tied together according to a distance measure. In addition, it has to be assured that every model contains a sufficient amount of training data. The resulting models are then re-estimated.

B. Decision Tree Construction

There are a number of criteria for decision tree construction procedure;

- Each leaf must have a minimum number of samples, a minimum occupancy, to ensure that the parameters of the final models can be accurately estimated.
- A finite set of questions can be used to divide each node. These questions constrain the way each node may be divided and also allows the incorporation of expert knowledge needed to predict contextual similarity when little or no data is available to determine which contexts are acoustically similar.
- Hidden Markov models should be able to accurately capture the variability of the terminal nodes. Gaussian mixture probability distributions should be able to accurately represent the training samples at each terminal node.

The first criteria can be satisfied by restricting the choice of question to those models to ensure that any created nodes have a sufficient number of associated samples in the training data. This restricted set of questions is searched in order to maximize the accuracy of the resulting hidden Markov models. Theoretically, this means attempting to minimize the within class variance while maximizing the between class variance. A simple scheme, which attempts to maximize the accuracy of the models with respect to their own class, is a maximum likelihood approach. This approach is very attractive since it is well matched to the way the parameters of models are subsequently estimated.

C. Likelihood Based Decision Criteria

Let S be a set of HMM states and let $L(S)$ be the log likelihood of S generating the set of training frames F under the assumption that all states in S are tied, i.e., they share a common mean $\mu(S)$ and variance $\Sigma(S)$ and that transition probabilities can be ignored. Then, assuming that tying states do not change the frame/state alignment, an approximation for $L(S)$ is given by (Young et al., 1994)

$$L(S) = \sum_{f \in F} \sum_{s \in S} \log\left(P\left(o_f ; \mu(S), \Sigma(S)\right)\right) \gamma_s\left(o_f\right) \tag{4.2}$$

where $\gamma_s\left(o_f\right)$ is a posteriori probability of the observed frame o_f being generated by state s. If the Gaussian probability distribution function is employed, then

$$L(S) = -\frac{1}{2}\left(\log\left[(2\pi)^n \left|\Sigma(S)\right|\right] + n\right) \sum_{s \in S} \sum_{f \in F} \gamma_s\left(o_f\right) \tag{4.3}$$

where n is the dimension of the data. Thus, the log likelihood of the whole data set depends only on the pooled state variance $\Sigma(S)$ and the total state occupancy of the pool, $\sum_{s \in S} \sum_{f \in F} \gamma_s\left(o_f\right)$. The former can be calculated from means and variances of the states in the pool, and the state occupancy counts can be obtained during the preceding Baum-Welch re-estimation. For a given node with state S, which is partitioned into two subsets, $S_y(q)$ and $S_n(q)$, by question q, the node is split using the question q^*, which maximizes

$$\Delta L_q = L\left(S_y(q)\right) + L\left(S_n(q)\right) - L(S) \tag{4.4}$$

provided that both ΔL_{q^*} and the total pooled state occupation counts for both $S_y\left(q^*\right)$ and $S_n\left(q^*\right)$ exceeds their associated thresholds.

As a final stage, the decrease in log likelihood is calculated for merging terminal nodes with differing parents. Any pairs of nodes for which this decrease is less than the threshold used to stop splitting are then merged.

4.3.3.2.5 Triphone Model Construction Procedure

The monophone-based system is used for creating the triphone system according to the following procedure:

1) Manually labeled monophone training: Initiated from a set of single Gaussian distributions, monophone models were generated from the labeled data using the standard Viterbi alignment process. Then, single Gaussian monophone models were reestimated using the Baum-Welch algorithm.

2) Triphone construction: In order to create a triphone system from a trained-monophone system, monophone models were copied and their transition matrices were tied. Then, triphones were initially trained using a forward-backward algorithm. A tree-based clustering was applied to cluster data from a phonetic decision tree. The final processes were state-tying, which merged any identical triphones, and re-training of state-tied triphones.
3) Mixture incrementing: A single Gaussian mixture was split to attain a higher recognition performance. Finally, the splitting mixture models were re-estimated using the forward-backward algorithm.

In order to construct context-dependent triphones, context-independent monophones were cloned and their transition matrices were tied to obtain a much more reliable estimation. Then four passes of the Baum-Welch algorithm were applied to a set of triphones to obtain the triphone parameters. After the initial training of triphones, state-clustering is an efficient technique to form the robust estimation of parameters of mixture distributions. To avoid the unseen triphone problem, a tree-based clustering technique is selected. This technique utilizes a log likelihood criterion and only supports a single-Gaussian continuous density output distribution. The objective of state-tying is to group triphone states into a number of equivalence classes using various linguistic questions concerning the identity of the base phone and the triphone context (Woodland, et al., 1994). Therefore, the total number of mixtures was reduced during the state-tying process. Iterative re-estimation by four passes of the Baum-Welch algorithm and incrementing of mixture components by a mixture-splitting procedure were applied to the tied-state triphones. Finally, the number of Gaussian distributions was increased to sixteen components per state.

4.3.3.3 Construction of the Initial-Final Model

The Initial-Final labels were generated from manual labeling. The Initial segment comprises the initial consonant only while the final consists of the rest of syllable. However, the Initial is contextually modeled in two ways, the context-independent Initial and the context-dependent Initial. The context-independent Initial modeling is independent of its vowel context whereas, the context-dependent Initial is modeled according to its vowel context. The

duration of the Final is longer than that of phone unit. Then the number of HMM states used to model the Final is more than that for a phone. A three state HMM was used for modeling the Initial, whereas the Final is modeled with a six state HMM. The training of the Initial-Final model is started from a single Gaussian observation distribution from the Initial-Final labeled data with the standard Viterbi alignment process and Baum-Welch algorithm. Then the labeled trained models are iteratively re-estimated with the embedded Baum-Welch algorithm. Finally, the number of Gaussian mixture components is increased up to sixteen mixtures per state.

4.3.3.4 Construction of the Onset-Rhyme Model

Initially, onset-rhyme labels were generated from manual labeling. The initial consonant segment and its transition towards the vowel were then converted to the onset segment, corresponding to two types of onset-rhyme models, CORM and PORM. The rhyme segment is converted from the steady vowel portion and the optional final consonant. Since the rhyme consists of the vowel including the optional final consonant, this implies that the rhyme can comprise two connected phones. Therefore, the duration of the rhyme is longer than that for the phone unit. Then the number of HMM states used to model the rhyme is more than that of phone. A three state HMM was used for modeling onset, whereas the rhyme is modeled with a six state of HMM. The onset-rhyme models were initiated starting from a single Gaussian observation distribution from the onset-rhyme labeled data. The standard Viterbi alignment process and Baum-Welch algorithm were applied to obtain the initial acoustic models. A set of initial acoustic models, was then trained with the embedded Baum-Welch algorithm. Iterative re-estimation by four passes of the Baum-Welch algorithm and incrementing of mixture components by a mixture-splitting procedure were employed to train the models. To achieve a higher performance, an iterative divide-by-two clustering algorithm was utilized to increase the number of Gaussian mixture components up to sixteen mixtures per state.

4.3.4 Mixture Component Incrementing

Mixture component incrementing provides an iterative mechanism for building a multiple mixture component system from a single Gaussian system. An output distribution of M mixture components is converted to an

$M+1$ component mixture distribution by cloning the mixture component with the largest weight and then perturbing the mean vectors of the two identical distributions by adding and subtracting 0.2 time of standard deviations respectively (Valtchev, 1995). The new mixture system is then trained using the Baum-Welch algorithm.

Traditionally, mixture densities of HMMs are built using the segmental k-means procedure to initialize the required number of mixture components and then retraining the models using the Baum-Welch algorithm. However, this approach requires exact the number of mixture components prior to building and assessing the performance of the system. The mixture component incrementing approach has been produced the similar results to the k-means clustering procedure, while, at the same time, the advantage of the former is that the number of mixture components can be continuously increased to obtain any desired balance between performance and model complexity (Young and Woodland, 1994).

4.3.5 Architecture of the Recognition System

All recognition systems were based on hidden Markov models using continuous density diagonal covariance mixture Gaussian output probability distributions. The output probability distributions could be shared at the state level but there was no sharing of mixture components that is the models were continuous density rather than tied-mixture or semi-continuous.

4.3.5.1 Word Network

A word network is defined using HTK Standard Lattice Format (SLF) (Young, et al., 1999). The SLF file contains a list of nodes representing syllables and a list of arc representing the transition between syllables. The transition can have probabilities attached to them and these can be used to indicate preferences in a grammar network. The construction of a word level SLF network can be specified by the grammar in form of regular expression. The expressions are constructed from sequences of syllables and metacharacters (Young, et al., 1999). Examples of word networks are shown in Figure 4.16.

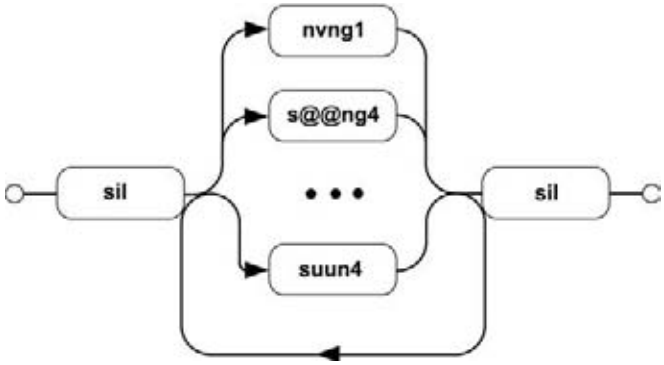


Figure 4.16 Word network for connected digit recognition

4.3.5.2 Language Modeling

In the recognition, two types of models, a no-language model and a bigram language model, are applied. In no-language model, each syllable can connect to all other word with optional silence. This no-language model is employed in order to evaluate the actual performance of acoustic models. The perplexity of the no-language model is equal to the number of words.

For a bigram language model, a word can only connect to words that can legally follow it. This bigram language model is used to evaluate the performance of the recognition system, which is composed of an acoustic model and grammar. A number of text corpus excerpted from various reading paragraphs were used to create a bigram language model. The language model is used to perform the linguistic post-processing and determine the optimal syllable sequence. Building a complex language model would required more study on the syntactic and semantic rules of Thai continuous speech.

4.3.5.3 Vocabulary and Dictionary

The Thai language is found to be relatively discrete in comparison with Western languages. Therefore, recognition of syllable sequences in the utterance is more reasonable than word sequence. The Royal Thai Dictionary is used as the reference to produce an inventory of syllables.

4.3.5.4 Decoding Process

In a continuous speech recognition system, decoding process is controlled by a recognition network compiled from a word-level network, a dictionary, and a set of HMM models. A recognition network ultimately consists of HMM states connected by transitions. However, it can be viewed at three different

levels: word, model, and state as illustrated in Figure 4.17 (Young, et al., 1999).

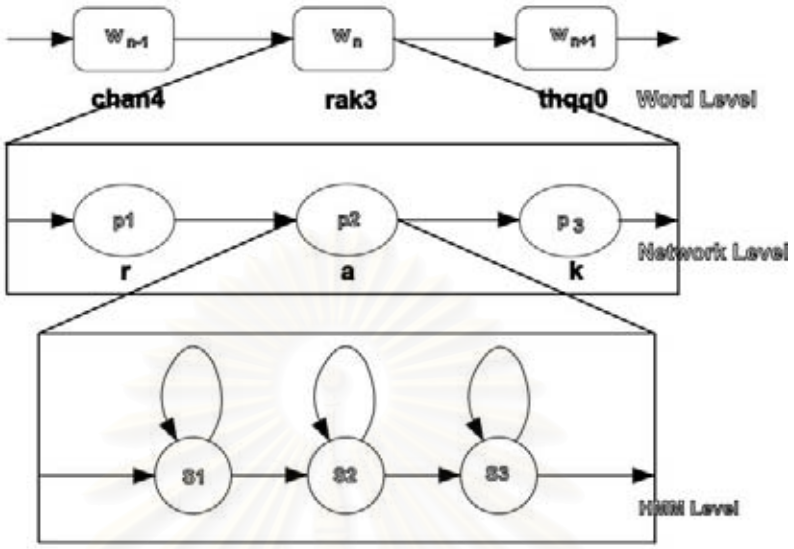


Figure 4.17 Recognition network level

A network describes the sequence of words that can be recognized while a dictionary describes the sequence of HMM models that constitute each word. A word level network will typically represent either a task grammar, which defines all of legal word sequences explicitly or a word loop, which simply puts all words of vocabulary in a loop and allows any word to follow any other word. Word-loop networks are obtained from stochastic language modeling. When a word network is loaded into a recognizer, a dictionary is used to convert each word in the network into a sequence of acoustic units represented by HMM models. Then, a word network is expanded into a HMM level network. Once the HMM network is constructed, it can be input to the decoder module and used to recognize speech input (Young, et al., 1999).

The task of the decoder is to find the paths through the network, which have the highest log probability. A log probability of each path is computed by summing the log probability of each individual transition in the path and the log probability of each emitting state generating the corresponding observation. Within-HMM transitions are determined from the HMM Parameters, between-model transitions are constant, and word-end transitions are determined by the language model likelihood attached to the

word level network. The most likely state sequence through a network can be found using the token passing implementation of the Viterbi algorithm. A token represents a partial path through the network extending from the beginning to the present time. At the end of the utterance, the most likely sequence of words is recovered by trace back through the decisions made about transitions between words. The recognition process is depicted in Figure 4.18.

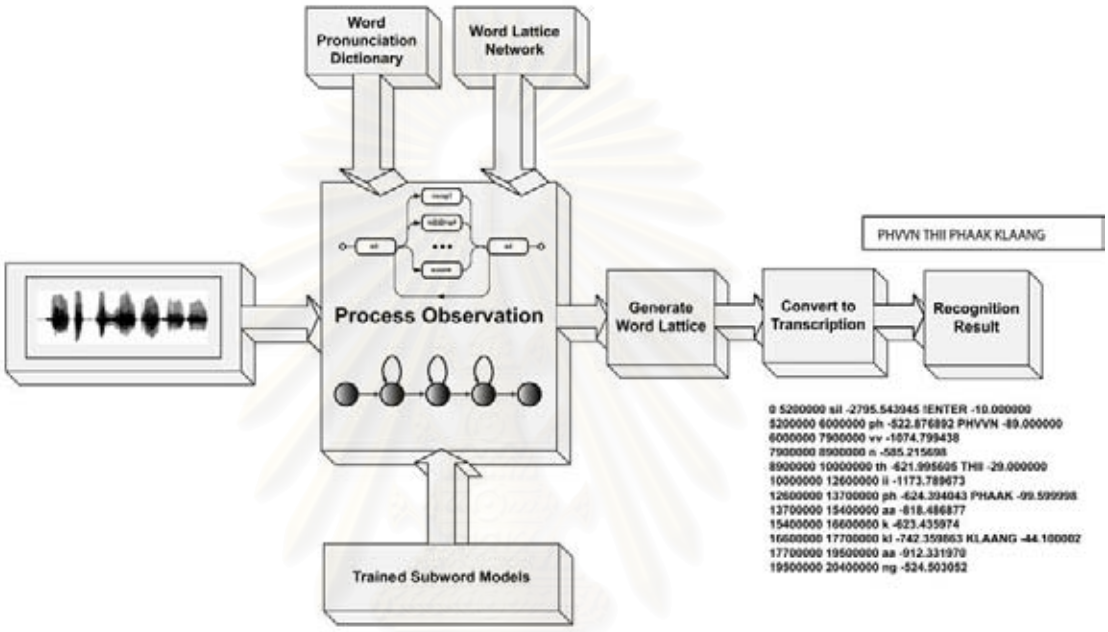


Figure 4.18 Recognition process

4.3.5.4 Evaluating Recognition Results

After the recognizer has processed the test data, the next step is to analyze the results. The transcription output from the recognizer is compared with the original transcription. Using dynamic programming, an optimal string match is obtained from matching each of the recognized and reference label sequences. The optimal string match is label alignment, which has the lowest possible score.

One of the criteria used for evaluating the efficiency of speech units is accuracy. The accuracy of speech units is computed from 2 formulas called “Percentage Correction” and “Percentage Correction” (Young, et al., 1999).

When the optimal alignment has been found, the number of substitution errors (S), deletion errors (D), and insertion errors (I) can be calculated. Then, the percentage correct is

$$\text{Percentage Correct} = \frac{N-D-S}{N} \times 100\% \quad (4.4)$$

where N is the total number of labels in the reference transcriptions. For many purposes, the percentage accuracy defined as

$$\text{Percentage Accuracy} = \frac{N-D-S-I}{N} \times 100\% \quad (4.5)$$

is a more representative figure of recognizer performance. An example of labeled transcription and recognition sentence is shown in Figure 4.19.

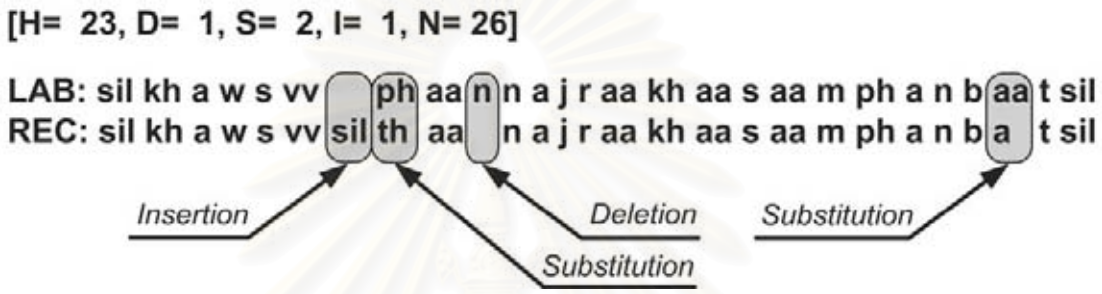


Figure 4.19 Evaluation of recognized sentence

4.4 Summary

This chapter has reviewed general speech units used in speech recognition to find the appropriate speech unit for the Thai language. However, the existing speech units seem to be inappropriate when applying to the Thai language according to the criteria of a good speech unit.

The onset-rhyme model is proposed in this chapter. The model comprises a pair of onset and rhyme units, which makes up a syllable. The onset comprises an initial consonant and its transition towards the following vowel. Together with the onset, the rhyme consists of a steady vowel segment and a final consonant. Two types of the onset-rhyme models, contextual onset-rhyme model (CORM) and phonotactic onset-rhyme model (PORM), are differently modeled according to their contexts.

The construction of the Thai continuous speech recognition system begins from speech corpus design, speech recording, and speech labeling. Then the HMM acoustic modeling is applied to speech units. The triphones have the trainability problem so the additional techniques, model clustering and state-tying, are required to overcome this problem.

Chapter 5

Experimental Results

This chapter describes the configurations of the Thai continuous speech recognition system used in this research. Details of recognition results performed with the Hidden Markov Model Toolkit (HTK) will be elaborated. The experiments were conducted for two main schemes, a recognition system using acoustic modeling only and a recognition system using both acoustic modeling and language modeling. In order to obtain an actual efficiency of the acoustic model, a language model should not be applied. On the other hand, incorporating the language model can boost the performance of a speech recognition system.

The experiment on gender effect will be conducted to study the effect of gender-dependent and gender-independent modeling. In the next experiment, mixture incrementing will be performed on the acoustic models. The objective of this experiment is to observe the improvement of the acoustic model by mixture incrementing technique. The first two experiments are conducted on the least complex acoustic unit, the monophone. The parameters obtained from these experiments will be used throughout the experiments in this dissertation. The other experiments will be carried on the triphone, Initial-Final, and onset-rhyme acoustic units. The final experiment incorporates the language model to the recognition system. This experiment will be conducted on the monophone, triphone, Initial-Final, and onset-rhyme respectively.

Recognition results of the system using the acoustic units—monophone triphone, Initial-Final, and onset-rhyme—are shown and in terms of syllable correction and accuracy. Apart from the syllable accuracy, the recognition results of phone, Initial-Final, and onset-rhyme acoustic units were analyzed. The analysis of the recognition results at the bottom level of the system, the acoustic level, reveals the actual efficiency of the speech unit.

5.1 Thai Continuous Speech Recognition System

This section described the configurations of the Thai continuous speech recognition system used in this research.

5.1.1 Speech Signal Processing and Feature Extraction

The speech samples were passed through a signal preprocessing routine consisting of signal pre-emphasis with a coefficient of 0.97 (Rabiner and Juang, 1993; Lee, et al., 1989; Lee, et al., 1990; Juang and Furui, 2001; Furui, 2001). The 25 ms Hamming window was applied every 10 ms in order to divide the speech signal into frames. This research employed a 12-order MFCC together its temporal derivatives for speech signal representation. (Maneenoi, et al., 2002; Ahkuputra, 2002; Jitapunkul, et al., 2003).

5.1.2 Language Modeling

In the recognition, two types of models, a no-language model and a bigram language model, are applied. The perplexity of the no-language model is 3200. A number of text corpus excerpted from various reading paragraphs, approximately 800,000 syllables, were used to create a bigram language model. The perplexity of this bigram language model is 252.03.

5.1.3 Vocabulary

The pronunciation dictionary is a syllable dictionary composed from monophones, Initial-Final units, or onset-rhyme units. In addition, the triphone network can be generated from the simple monophone dictionary (Young, et al., 1999). This dictionary consists of 3,200 syllables excluding tones, which covers almost 33,000 vocabulary words in the Royal Thai Dictionary.

5.2 Experiments on Gender Effect

Generally, there are different characteristics between male and female speech indicated in the frequency domain. The distinction can be represented by the location of the first three formants for vowels and the fundamental frequency, since men and women have different size of articulatory organ. This experiment intends to study the characteristics of

gender effect. Therefore, the experiment will be performed using systems that were gender-independent and gender-dependent. The monophone model was used in the experiment. For the gender-dependent system, the acoustic models were separately trained for male and female speakers. On the other hands, the acoustic models for the gender-independent system were conjointly trained on male and female speakers.

The experiments will be conducted in 3 schemes:

1. Recognition of male and female speakers on gender-dependent modeling
2. Recognition of male and female speakers on gender-independent modeling
3. Recognition of male speakers with the acoustic model created from female speakers, and recognition of female speakers with the acoustic model created from male speakers

The system configuration is set as follows:

- The standard 3-state left-to-right HMM with no skip state
- 12 order MFCCs with delta coefficients
- A single Gaussian output distribution
- 58 monophone models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

5.2.1 Experimental Results

A. Recognition of male and female speakers on gender-dependent modeling

The syllable recognition results of gender-dependent modeling are shown in Tables 5.2.1 and 5.2.2. The speaker-dependent system of both male and female speakers gives higher recognition results than the speaker-independent system. The results show the correction at 15.6% and 15.9% for the male and female speaker-dependent systems. For the speaker-independent systems, the correction of male and female speakers is 12.2% and 12.7%.

Due to the high insertion rate of the phone model, the accuracy is very low compared to the correction. The accuracy of male and female

speakers in the speaker-dependent system is 4.5% and 5.4%, while the speaker-independent system gives 0.7% and 2.0% of the accuracy for male and female speakers respectively.

Table 5.2.1 Syllable recognition results of male speakers on the system trained with male speakers

System	Correction	Accuracy
Speaker-dependent	15.6%	4.5%
Speaker-independent	12.2%	0.7%

Table 5.2.2 Syllable recognition results of female speakers on the system trained with female speakers

System	Correction	Accuracy
Speaker-dependent	15.9%	5.4%
Speaker-independent	12.7%	2.0%

B. Recognition of male and female speakers on gender-independent modeling

The syllable recognition results of gender-independent modeling are shown in Tables 5.2.3 and 5.2.4. The results show the correction at 13.0% and 13.2% for the male and female speaker-dependent systems. For the speaker-independent systems, the correction of male and female speakers is 10.3% and 11.0%. The accuracy of male and female speakers in the speaker-dependent system is 2.7% and 3.2%, while the speaker-independent system gives 0.1% and 0.6% of the accuracy for male and female speakers respectively.

Comparing the syllable recognition results to gender-dependent modeling, gender-independent modeling gives lower correction and accuracy of both speaker-dependent and speaker-independent system. The correction and accuracy of gender-independent modeling for male and female SD system is lower than gender-dependent modeling around 2.7% and 2.0%. For the male and female SI system, the performance of gender-independent modeling is worse than the performance of gender-dependent modeling around 1.8% of correction and 1.0% of accuracy respectively.

Table 5.2.3 Syllable recognition results of male speakers on the system trained with male and female speakers

System	Correction	Accuracy
Speaker-dependent	13.0%	2.7%
Speaker-independent	10.3%	0.1%

Table 5.2.4 Syllable recognition results of female speakers on the system trained with male and female speakers

System	Correction	Accuracy
Speaker-dependent	13.2%	3.2%
Speaker-independent	11.0%	0.6%

C. Recognition of male speakers with the acoustic model created from female speakers, and recognition of female speakers with the acoustic model created from male speakers

When using the automatic gender classification, the misclassification may be occurred. In this severe condition, the speakers are tested with the acoustic model created from a different gender. The recognition results show very poor performance in this case. The syllable recognition results of gender-independent modeling are shown in Tables 5.2.5 and 5.2.6.

Comparing the syllable recognition results to gender-dependent and gender-independent modeling, the correction and accuracy of male and female SD system is lower than gender-dependent modeling around 10.6% and 17.7%, and around 8.0% and 15.7% for gender-independent modeling. For the male and female SI system, the performance also degrades from gender-dependent and gender-independent modeling around 9.7% of correction and 15.1% of accuracy, and around 7.9% of correction and 14.1% of accuracy respectively.

Table 5.2.5 Syllable recognition results of male speakers on the system trained with female speakers

System	Correction	Accuracy
Speaker-dependent	4.8%	-10.8%
Speaker-independent	2.7%	-14.7%

Table 5.2.6 Syllable recognition results of female speakers on the system trained with male speakers

System	Correction	Accuracy
Speaker-dependent	5.5%	-8.5%
Speaker-independent	2.9%	-12.7%

5.2.2 Discussion

Most parametric representations of speech are highly dependent on a group of speakers, and probability distributions suitable for a certain group of speakers may not be suitable for other group of speakers. Speaker attributed variability is undesirable in speaker-independent speech recognition system. Especially, the gender of the speaker is one of the influential sources of this variability. The recognition results the system based on gender-dependent modeling are higher than those of gender-independent modeling. When the speakers are tested with the acoustic model created from a different gender, the performance of the system would be highly degraded. These experimental results support the use of the gender specific model throughout this dissertation.

5.3 Experiments on Mixture Incrementing

To improve the accuracy of the recognizer, the number of Gausian mixture components per state was increased. Increasing the number of Gaussian mixture components per state can improve the recognizer's performance up to a point where not enough information is available to fit the actual shape of the probability contour.

Though initialization of the output distribution can start from any number of mixtures, it has been found to be more effective to train mixtures incrementally. We first create one mixture component per state, train it and

then build two mixture components per state, etc. This process can be continued until the required number of mixture components is trained. The number of re-estimations to perform before incrementing mixtures is usually chosen to be four or more. This process is shown in Figure 5.3.1.

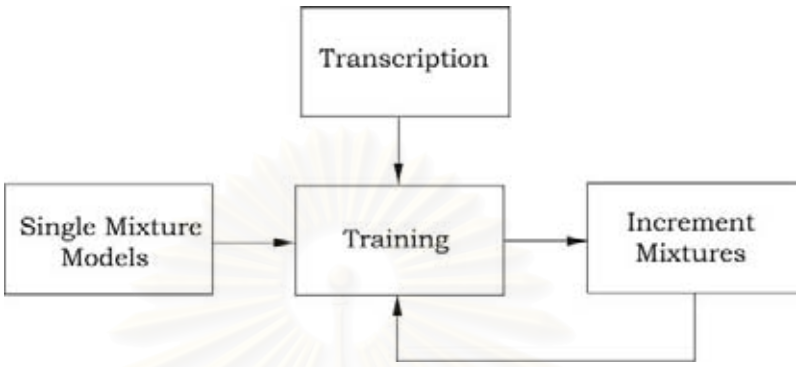


Figure 5.3.1 The process of mixture incrementing and training

In order to determine the improvement of the acoustic models by incrementing the number of mixtures, the experiments are performed beginning at one mixture. As already mentioned, in the training iterations, the number of Gaussian mixture components was increased at a time up to the number of 256, with 2^n at a time thereafter.

The system configuration is set as follows:

- The standard 3-state left-to-right HMM with no skip state
- 12 order MFCCs with delta coefficients
- 58 monophone models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

5.3.1 Experimental Results

Figures 5.3.2-5.3.9 illustrate the correction and the accuracy, as the number of mixtures is increased. As evident from Figures 5.3.2-5.3.5, the syllable recognition rates were increased when the number of mixture components was increased. The syllable recognition results of speaker-dependent systems tend to continually increase when the number of mixture components was increased. On the other hands, the syllable recognition

results of speaker-independent systems tend to be gradually increased. At 256 mixture components per state, male and female SD systems produce 46.1 % and 44.3 % of syllable correction, and 37.2 % and 38.3 % of syllable accuracy. For male and female SI systems, the syllable correction rates are 20.7 % and 25.6 %, and the syllable accuracy rates are 8.8 % and 15.3 %.

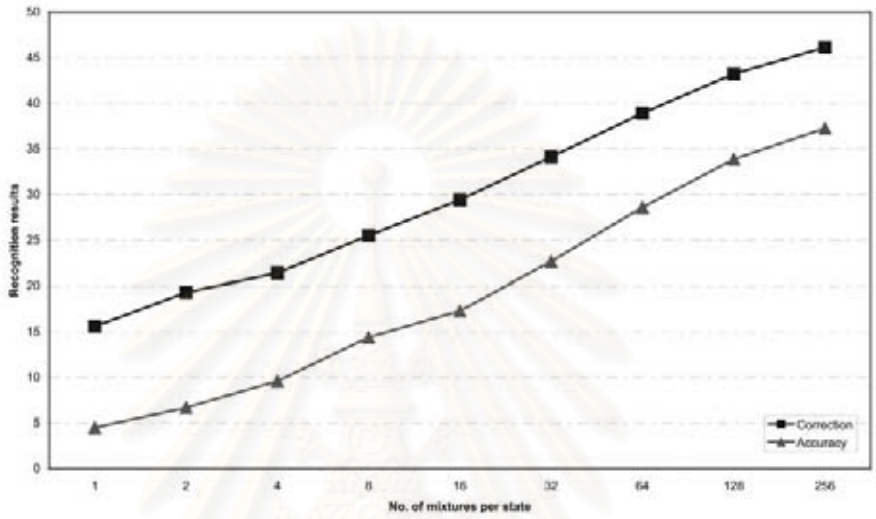


Figure 5.3.2 Syllable recognition results of male SD system

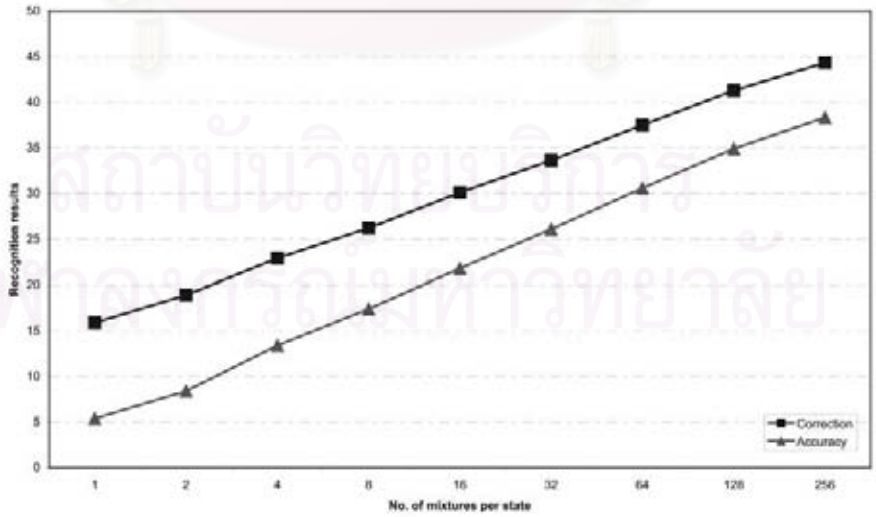


Figure 5.3.3 Syllable recognition results of female SD system

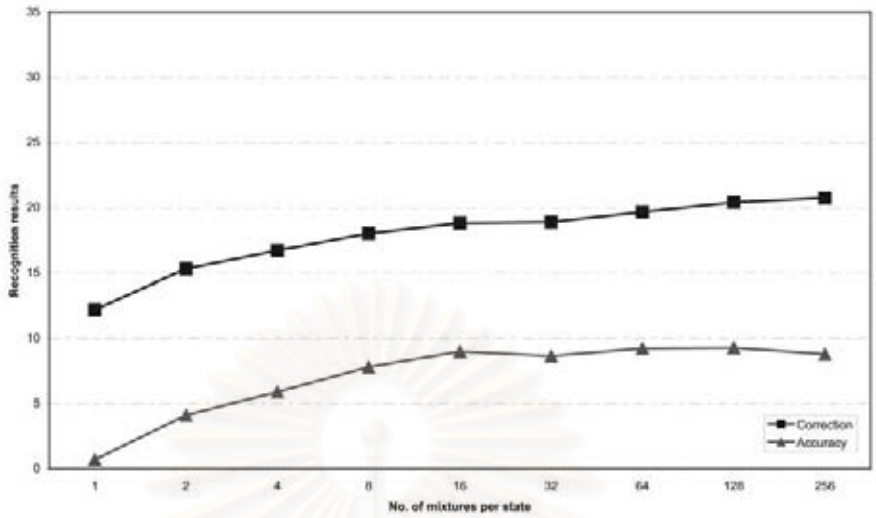


Figure 5.3.4 Syllable recognition results of male SI system

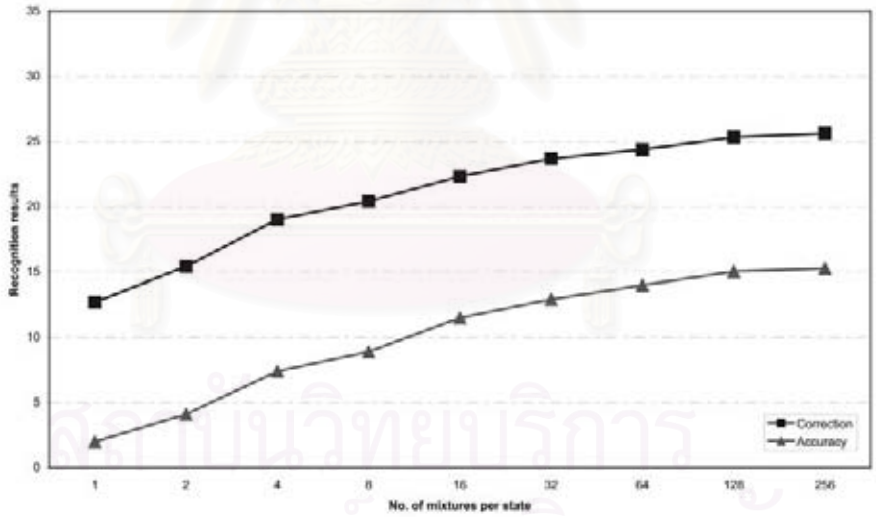


Figure 5.3.5 Syllable recognition results of female SI system

Recognition rate of phone unit in SD and SI systems exhibit the similar way to the syllable recognition results. Phone correction and accuracy rates at 256 mixture components per state are 69.5 % and 59.0 %, and 69.3 % and 61.1 % for male and female SD systems respectively. The

male and female SI systems give the syllable correction rates at 42.5 % and 51.4 %, and the syllable accuracy rates at 28.6 % and 39.5 % respectively. The phone recognition results are shown in Figures 5.3.6-5.3.9.

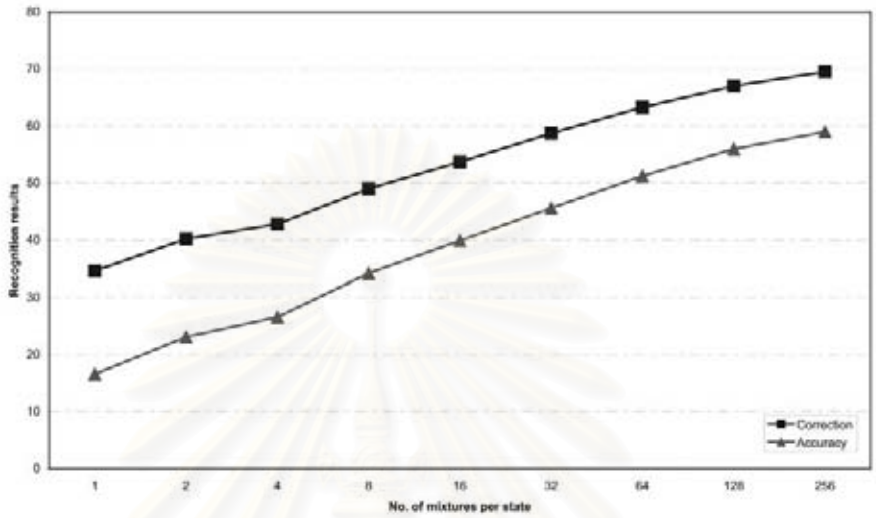


Figure 5.3.6 Phone recognition results of male SD system

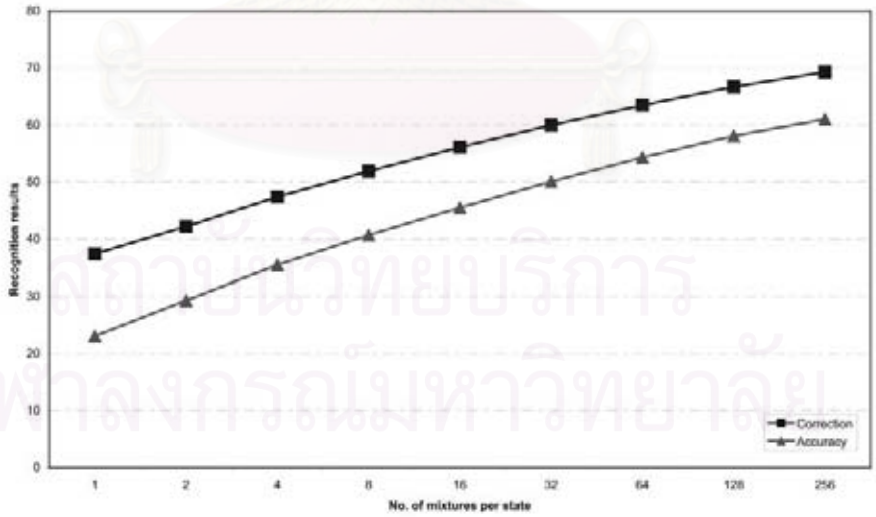


Figure 5.3.7 Phone recognition results of female SD system

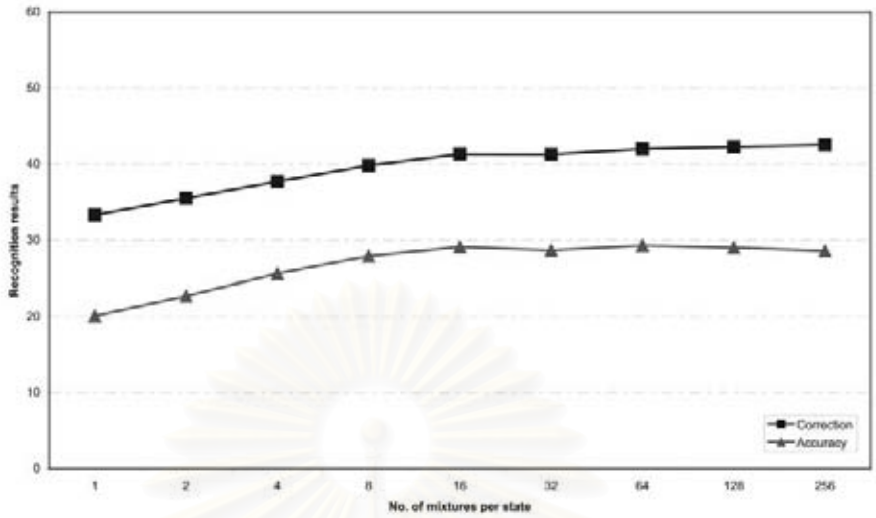


Figure 5.3.8 Phone recognition results of male SI system

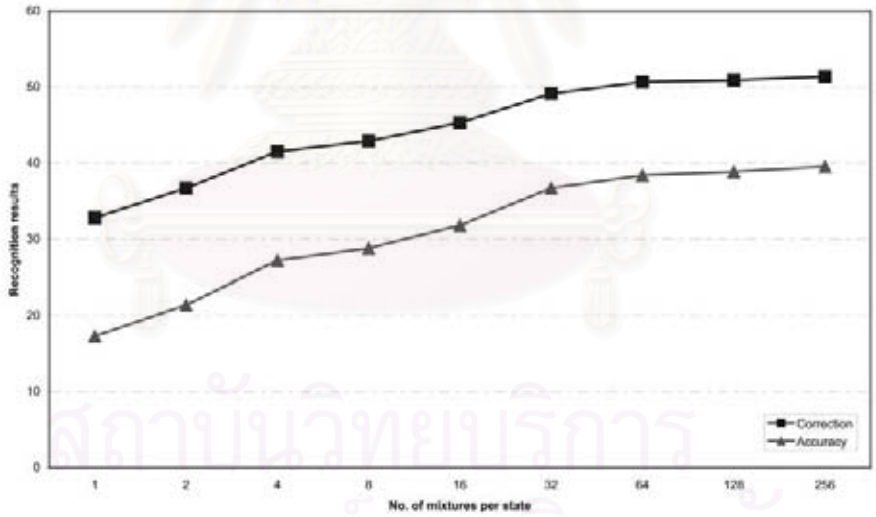


Figure 5.3.9 Phone recognition results of female SI system

5.3.2 Discussion

Mixture component incrementing provides an iterative mechanism for building a multiple mixture component system from a single Gaussian system. The number of mixture components can be continuously increased

to obtain any desired balance between performance and model complexity. The recognition results show that both SD and SI systems achieve a higher performance when the number of mixture components is increased. However, the recognition results of SD systems is likely to rise continually, while the recognition results of SI systems tend to be saturated at a high number of mixture components. In the SD system, incrementing the large number mixture components yields the tightly fit acoustic model. This fitting acoustic model gives out a high recognition result for the SD system. On the contrary to the SD system, the recognition results of the SI system rise slightly when then number of mixture components is increased. The acoustic model trained with a large number of mixture components is very fit to the group of training speakers. When employing this fitting acoustic model to another group of speakers or the SI system, the recognition results increase in dissimilar way compared with those of the SD system. The acoustic model trained with a large number mixture components is not suitable for using in the SI system. Moreover, the acoustic model with a large number mixture components requires extensive computation. Therefore, choosing a suitable number of mixtures for the acoustic model yields a good efficiency in terms of performance and complexity.

5.4 Experiments on Tied-State Triphone Modeling

This experiment creates the acoustic models of the intra and inter syllable triphones and evaluates their performance. There are 2 main steps for creating the triphones.

1. **Creating triphones from monophones**. In order to create a triphone system from a trained-monophone system, monophone models were copied and their transition matrices were tied. Then, triphones were initially trained using a forward-backward algorithm.
2. **State clustering and parameter tying:** After the initial training of triphones, a tree-based clustering technique is employed to form the robust estimation of parameters of mixture distributions by utilizing a log likelihood criterion. The final process was state-tying, which merged any identical triphone states into a number of equivalence classes using various linguistic questions concerning the identity of the base phone and the triphone context (Woodland, et al., 1994).

Consequently, the total number of mixtures was reduced during the state-tying process.

Since the performance of the triphones usually varies depending on the size of the tree, it is necessary to control the size of the decision tree. The number of leaf nodes or tied-states was controlled by the log likelihood criterion. Several numbers of tied-states that were made, when the triphone models were created by tree-based clustering, were tested to obtain the suitable parameters for the model. The suitable parameters of the triphones from state tying process will be used throughout all experiments in this dissertation.

After state tying process, the tied-state triphones will be trained using the Baum-Welch algorithm. Then, the mixture incrementing will be applied to increase the number of mixture components until it reaches the required number.

5.4.1 Experimental Results on Creating the Tied State Triphones

The trainability problem becomes more serious when larger amounts of context are to be taken into account. There are many triphone contexts, which occur only a few times in the training corpus and many more than that do not occur at all. State tying is one of the methods to overcome this trainability problem and ensure the triphone parameters can be estimated reliably.

The first step of this experiment is to vary the log likelihood to control the size of the decision tree. The log likelihood is adjusted to obtain the desired number of tied state triphones. The relation between the log likelihood and the number of tied state triphones is shown in Figures 5.4.1 to 5.4.4.

The system configuration is set as follows:

- The standard 3-state left-to-right HMM with no skip state
- A single component of Gaussian distribution
- 12 order MFCCs with delta coefficients
- 7,769 logical intra-syllable triphone models and 23,307 states
- 64,775 logical inter-syllable triphone models and 193,425 states
- Training speakers – 9 male speaker and 11 female speakers

- Evaluation speakers – 3 speaker-dependent males and 3 speaker-dependent females

Any pairs of nodes for which the decrease in log likelihood is less than the threshold used to stop splitting are then merged at the final stage of the state tying. According to the log likelihood based decision criteria in eq (4.2)-(4.4), the number of tied-state triphones is dropped when the log likelihood is raised. Increasing the log likelihood threshold results in reduction of the number of physical triphone models as well. Both numbers of tied-state triphones and physical triphone models decline exponentially.

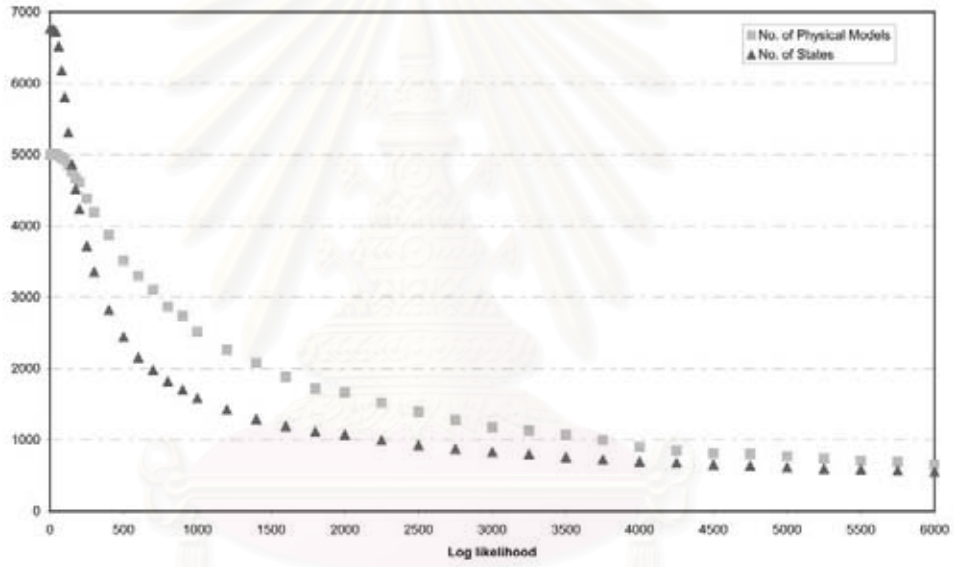


Figure 5.4.1 Relation between log likelihood and the number of tied state intra-syllable triphones of male speakers

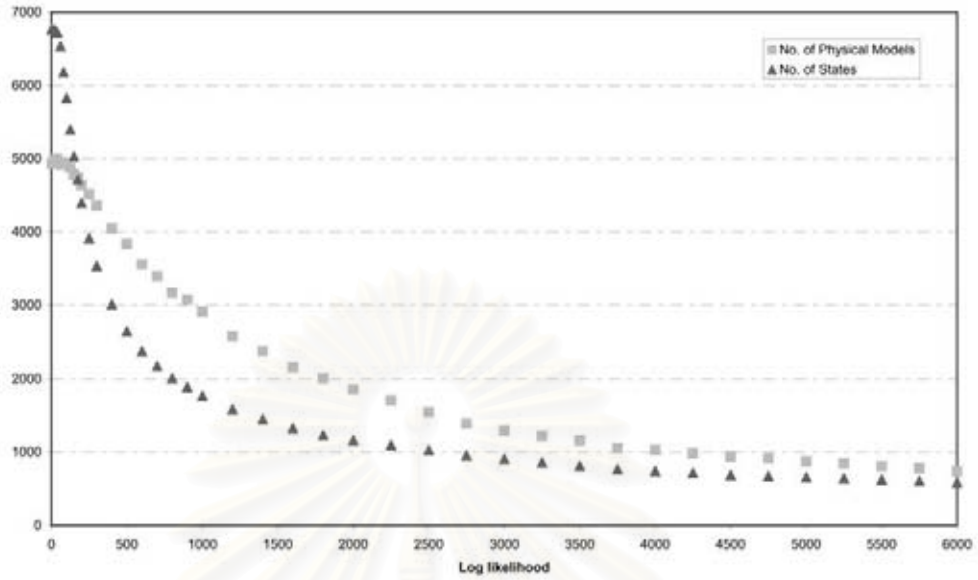


Figure 5.4.2 Relation between log likelihood and the number of tied state intra-syllable triphones of female speakers

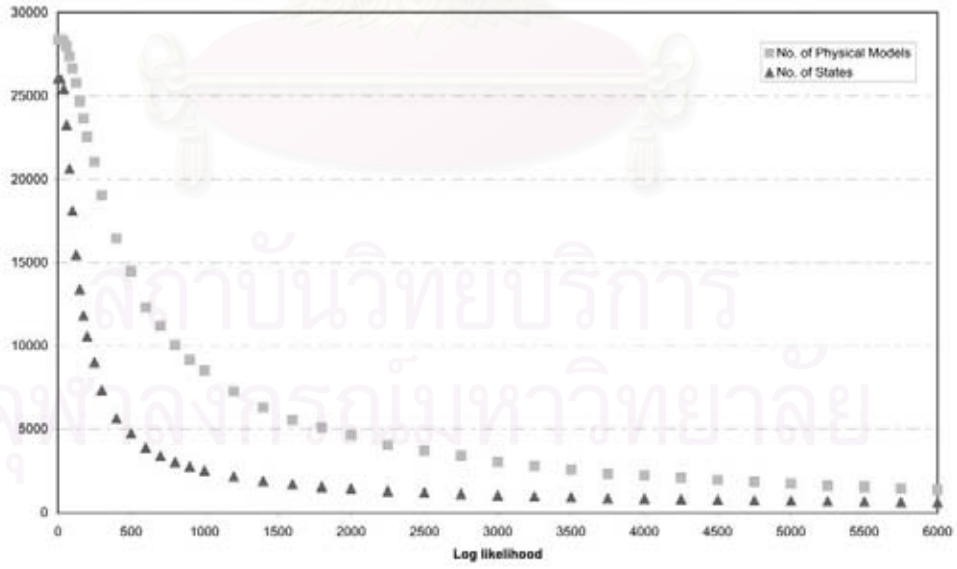


Figure 5.4.3 Relation between log likelihood and the number of tied state inter-syllable triphones of male speakers

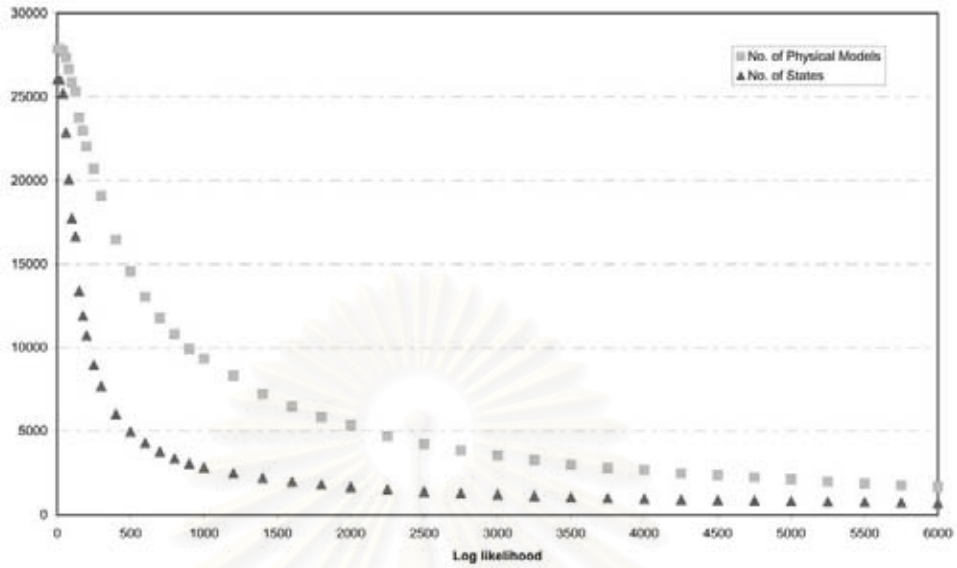


Figure 5.4.4 Relation between log likelihood and the number of tied state inter-syllable triphones of female speakers

Since the decoding process of a triphone system is very complex and it takes much time in the recognition stage, only 3 male and 3 female speakers were selected to evaluate the efficiency of the acoustic models. The triphone parameters of a triphone system, which has the minimum word error rate (WER), will be used in the further experiment. The syllable recognition results of both intra and inter syllable triphone systems are shown in Figures 5.4.5-5.4.8.

Syllable correction rates decrease continually if the number of tied states is dropped. Substitution and deletion errors become higher when the number of tied states is declined. On the contrary, insertion error decreases in the same way as the number of tied states is lessening. Since the inter-syllable triphones provide cross-context acoustic modeling, the correction rates of the inter-syllable triphone system are better than those of the intra-syllable triphone system. The male and female intra-syllable triphone system produces the highest syllable correction rates at 44.4 % and 38.1 %. The highest syllable correction rates of the male and female inter-syllable triphone system are 49.6 % and 48.5 % respectively.

The highest syllable accuracy is used to determine the suitable triphone parameters. The male and female intra-syllable triphone system produces the highest syllable accuracy rates at 20.2 % and 22.1 %. The highest syllable accuracy rates of the male and female inter-syllable triphone system are 25.4 % and 26.1 % respectively. Obviously, the number of states at the highest accuracy is different from he number of states that produces the highest correction. The numbers of tied states, which give the highest accuracy, are summarized in Table 5.4.1.

Phone recognition results with disregarding contexts are illustrated in Figures 5.4.9-5.4.12. The triphone systems show correction and accuracy rates of phone in the similar way to those of the syllable.

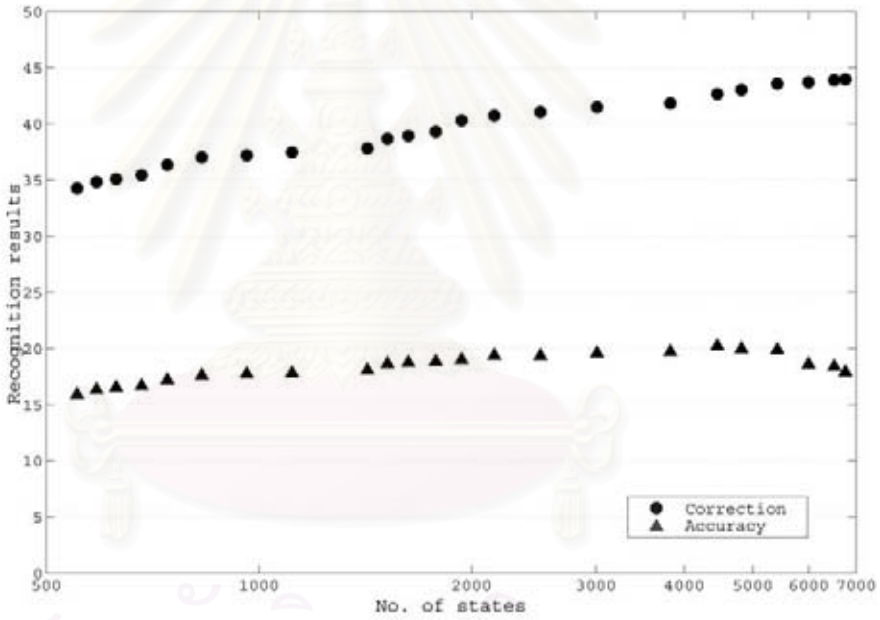


Figure 5.4.5 Syllable recognition results of intra-syllable triphone (male speaker-dependent system)

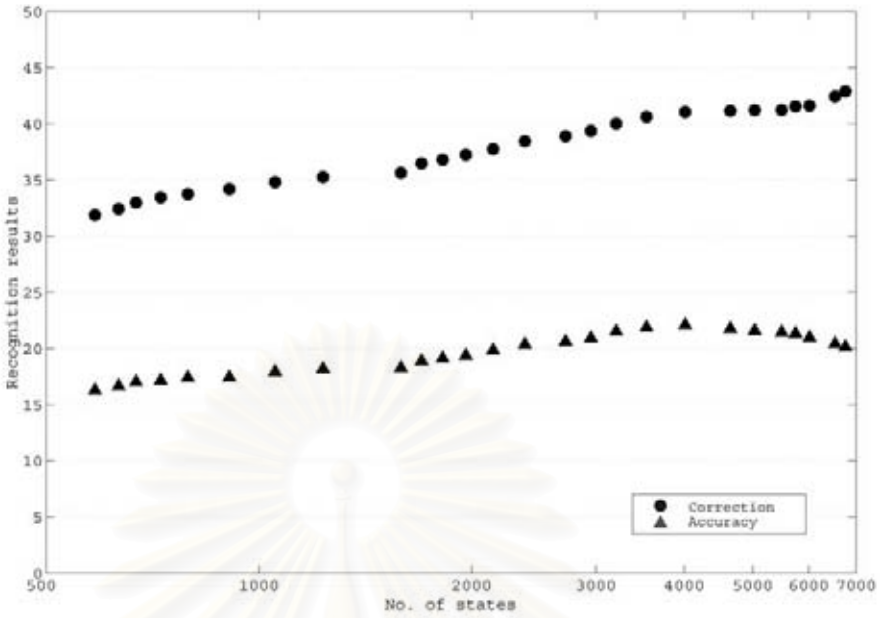


Figure 5.4.6 Syllable recognition results of intra-syllable triphone (female speaker-dependent system)

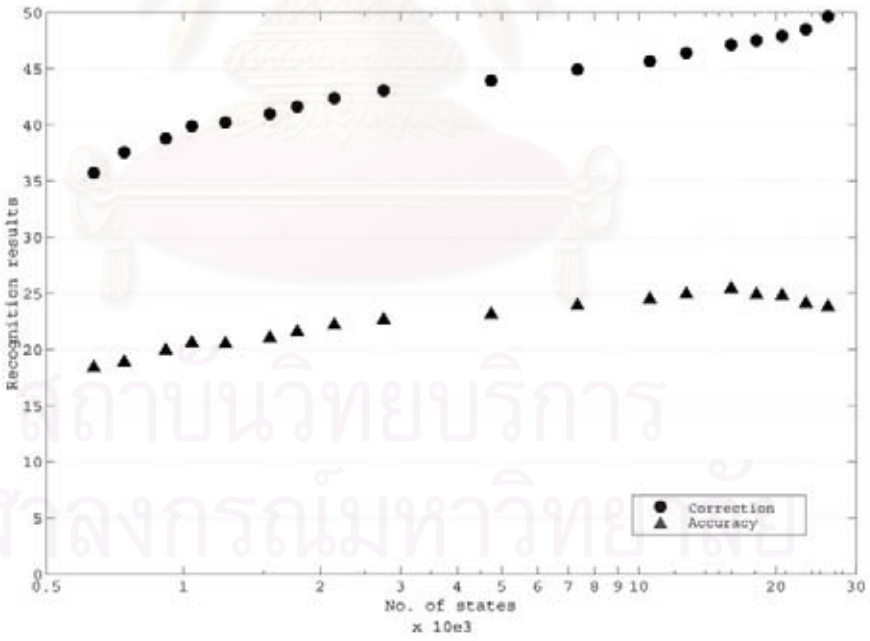


Figure 5.4.7 Syllable recognition results of inter-syllable triphone (male speaker-dependent system)

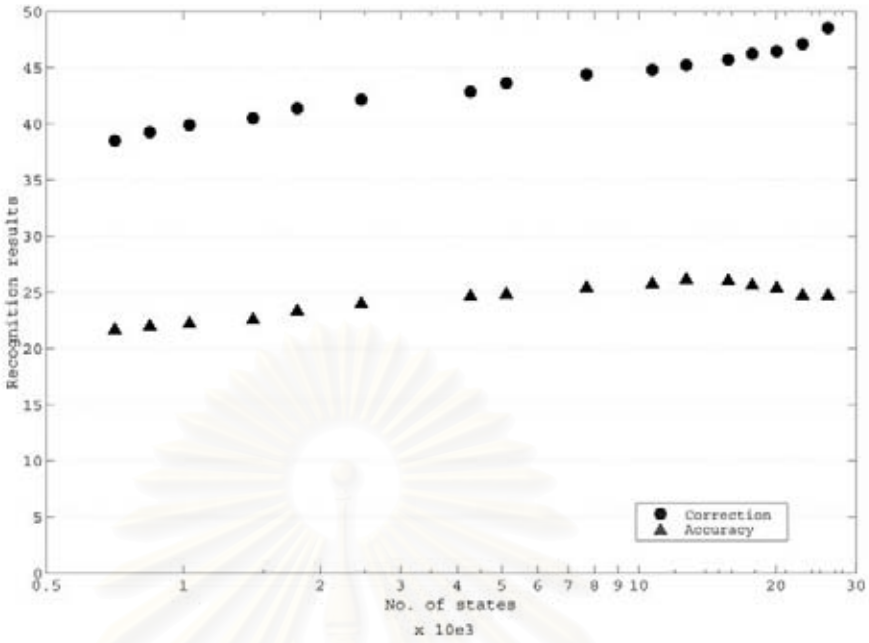


Figure 5.4.8 Syllable recognition results of inter-syllable triphone (female speaker-dependent system)

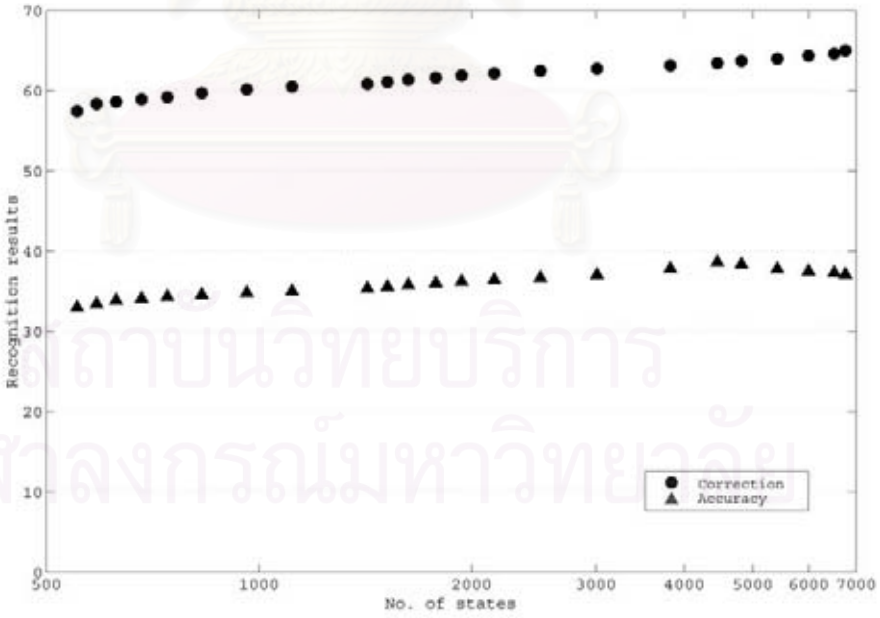


Figure 5.4.9 Phone recognition results of intra-syllable triphone (male speaker-dependent system)

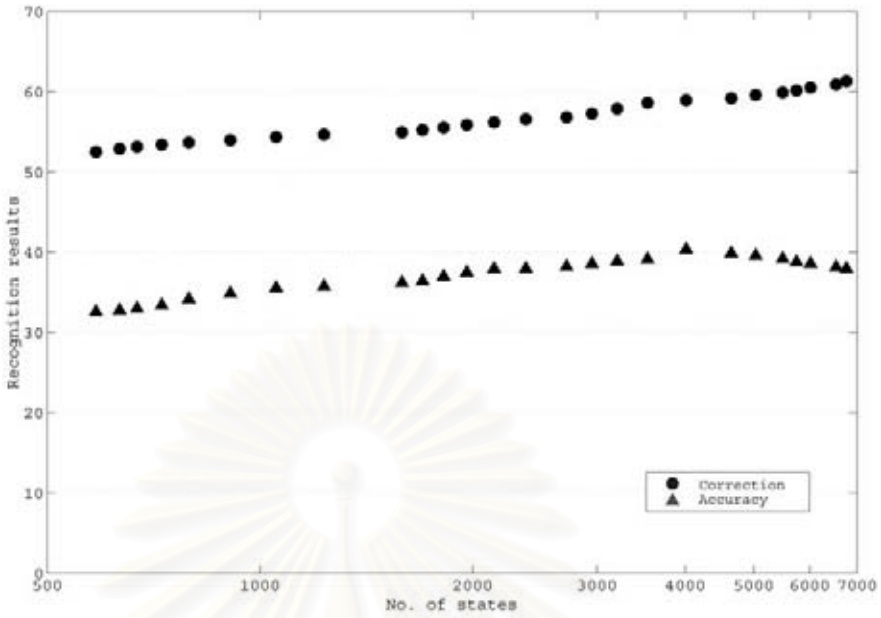


Figure 5.4.10 Phone recognition results of intra-syllable triphone (female speaker-dependent system)

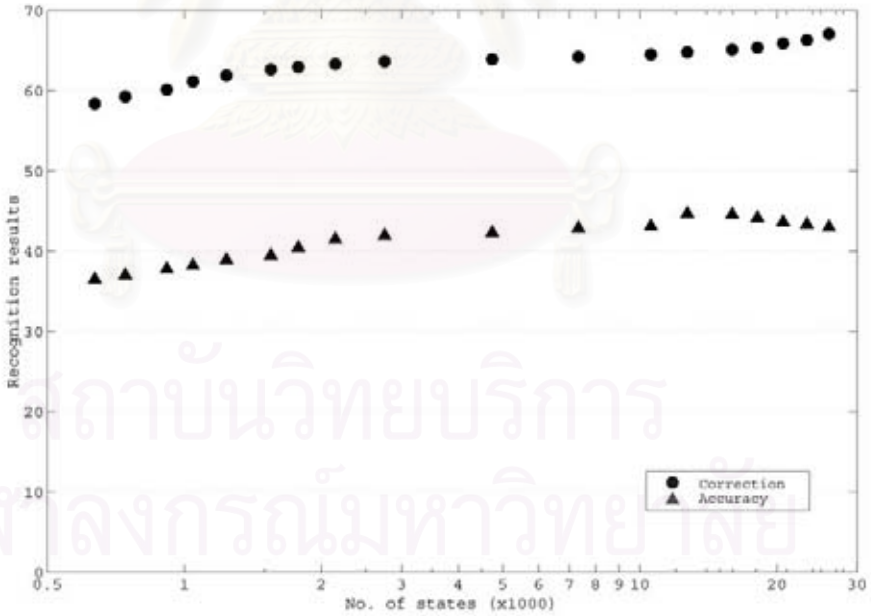


Figure 5.4.11 Phone recognition results of inter-syllable triphone (male speaker-dependent system)

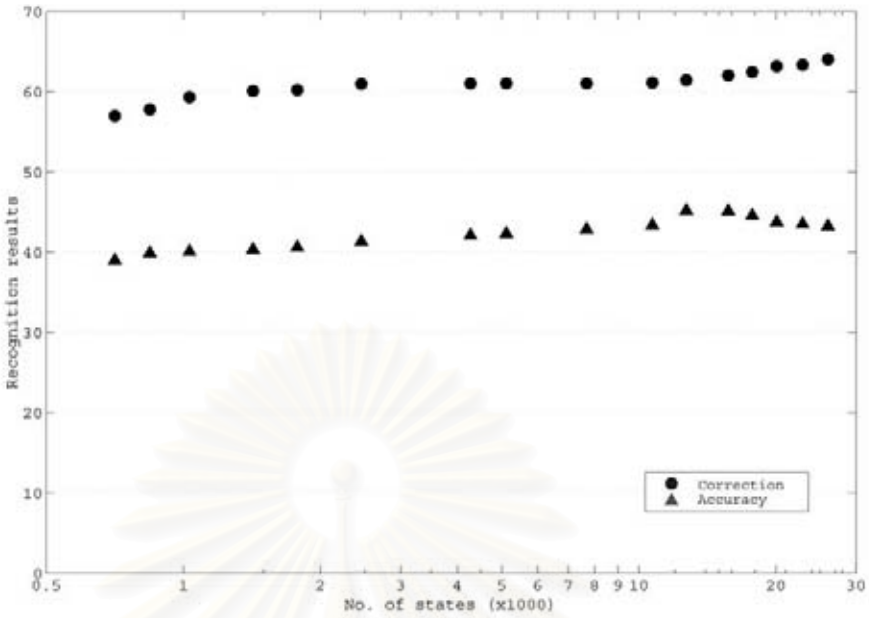


Figure 5.4.12 Phone recognition results of inter-syllable triphone (female speaker-dependent system)

Table 5.4.1 The number of tied states that produces the highest accuracy

Triphone system		No. of tied states	Correction	Accuracy
Intra-syllable	male	4,455	42.7	20.2
	female	4,011	41.0	22.1
Inter-syllable	male	15,949	47.1	25.4
	female	12,718	45.2	26.1

5.4.2 Experimental Results on Mixture Incrementing

The next step after achieving the suitable tied state triphone parameters is to increase the number of Gaussian mixture components. The number of Gaussian mixture components was increased, starting from a single component of Gaussain distribution per state, up to 8 mixture components per state.

The system configuration is set as follows:

- The standard 3-state left-to-right HMM with no skip state
- 12 order MFCCs with delta coefficients
- 4,642 logical intra-syllable triphone models and 4,455 states for male system

- 4,510 logical intra-syllable triphone models and 4,011 states for female system
- 25,776 logical inter-syllable triphone models and 15,949 states for male system
- 23,388 logical inter-syllable triphone models and 12,718 states for female system
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

The syllable recognition results of both intra and inter syllable triphone systems are shown in Tables 5.4.2-5.4.9. Furthermore, phone recognition results with disregarding contexts are shown in Tables 5.4.10-5.4.17. From the experimental results, the performance is boosted when the number of mixture components is increased. The inter-syllable triphone systems have higher syllable accuracy than the intra-syllable triphone systems. At eight mixture components per states, the inter-syllable triphone system gives out the syllable accuracy at 35.5 % and 36.4 %, while the intra-syllable triphone system produces the syllable accuracy at 28.0 % and 30.4 % for male and female SD systems. For the male and female SI systems, the syllable accuracy of the inter-syllable triphone system is 22.3 % and 23.4 %, while the syllable accuracy of the inter-syllable triphone system is 18.7 % and 18.9.

Table 5.4.2 Syllable recognition results of intra-syllable triphone (male speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	30.9	12.7
2	34.5	17.8
4	39.4	22.1
8	43.3	28.0

Table 5.4.3 Syllable recognition results of intra-syllable triphone (female speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	31.5	14.6
2	35.9	21.3
4	40.2	24.5
8	44.4	30.4

Table 5.4.4 Syllable recognition results of intra-syllable triphone (male speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	29.7	14.2
2	31.3	16.0
4	34.1	18.5
8	34.5	18.7

Table 5.4.5 Syllable recognition results of intra-syllable triphone (female speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	29.0	14.4
2	32.9	16.5
4	35.9	17.2
8	36.3	18.9

Table 5.4.6 Syllable recognition results of inter-syllable triphone (male speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	37.4	17.5
2	40.2	24.6
4	43.9	30.1
8	47.3	35.5

Table 5.4.7 Syllable recognition results of inter-syllable triphone (female speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	36.6	18.9
2	41.4	25.2
4	44.1	29.7
8	48.2	36.4

Table 5.4.8 Syllable recognition results of inter-syllable triphone (male speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	33.4	15.7
2	35.8	18.0
4	37.2	21.5
8	39.7	22.3

Table 5.4.9 Syllable recognition results of inter-syllable triphone (female speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	33.6	15.3
2	35.4	17.8
4	36.9	21.0
8	39.9	23.4

The inter-syllable triphone systems also have the higher phone accuracy than the intra-syllable triphone systems. The phone accuracy of the inter-syllable triphone system is 49.6 % and 50.5 % for male and female SD systems, better than 44.0 % and 47.3 % of the intra-syllable triphone system. The inter-syllable triphone surpass the intra-syllable triphone for the SI systems as well. The phone accuracy of the male and female SI inter-syllable triphone system is 41.7 % and 43.8 %, and 34.4 % and 34.8 % of the intra-syllable triphone system.

Table 5.4.10 Phone recognition results of intra-syllable triphone (male speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	56.6	28.6
2	61.2	33.9
4	64.7	38.6
8	67.8	44.0

Table 5.4.11 Phone recognition results of intra-syllable triphone (female speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	56.3	29.4
2	61.2	34.3
4	65.1	40.9
8	68.9	47.3

Table 5.4.12 Phone recognition results of intra-syllable triphone (male speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	52.5	30.6
2	53.9	32.7
4	54.0	33.2
8	54.8	34.4

Table 5.4.13 Phone recognition results of intra-syllable triphone (female speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	49.8	28.9
2	53.8	30.0
4	56.1	31.9
8	58.5	34.8

Table 5.4.14 Phone recognition results of inter-syllable triphone (male speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	60.0	33.7
2	64.2	38.3
4	68.8	43.7
8	72.4	49.6

Table 5.4.15 Phone recognition results of inter-syllable triphone (female speaker-dependent system)

No. of mixtures per state	Correction	Accuracy
1	61.9	41.3
2	65.1	44.6
4	69.8	46.2
8	73.3	50.5

Table 5.4.16 Phone recognition results of inter-syllable triphone (male speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	56.2	36.1
2	57.4	38.3
4	60.9	39.8
8	62.2	41.7

Table 5.4.17 Phone recognition results of inter-syllable triphone (female speaker-independent system)

No. of mixtures per state	Correction	Accuracy
1	58.6	38.9
2	60.4	40.2
4	62.7	41.5
8	63.0	43.8

5.4.3 Discussion

There are several steps in building a triphone system as described in Chapter 4. State tying is the important technique used to overcome the trainability problem of the triphones. The log likelihood threshold in the state tying process controls the number of tied state triphones, which relates to the performance of a triphone system. It is necessary to determine the suitable number of tied state triphones, which give the highest accuracy. This step takes time to adjust the log likelihood threshold, train the models, and evaluate the results, especially for the inter-syllable triphone system, which has a complex decoding process.

Though the process in building a triphone system is quite complex, the performance of a triphone system is much better than the monophone system. The triphones were modeled according their contexts resulting in the accurate acoustic model, comparing to the monophones. The inter-syllable triphone system outperforms the intra-syllable triphone system because the speech units were modeled the co-articulation across the syllable.

The inter-syllable triphone seems to be better than the intra-syllable triphone. However, the numbers of tied states of the inter-syllable triphone system is more than that of the intra-syllable triphone system. The inter-syllable triphone system consumes greater memory than the intra-syllable triphone system. Moreover, the decoding process of the inter-syllable triphone system is much more complex than the intra-syllable triphone system as depicted in Figures 4.13 and 4.14. Recognition of the inter-syllable triphone system takes much more time in the decoding process than the intra-syllable triphone system. This is the weak point of the inter-syllable triphone system.

In the recognition result analysis, the triphone systems have a high correction rate. However, the accuracy is considerable lower than the correction. There are many deletions in the triphone system due to the short duration of the units. The alternative acoustic model will be figured out to overcome the problem of the triphones.

5.5 Experiments on Onset-Rhyme Modeling

This experiment contains three studies on onset-rhyme modeling. The first study will focus on determining the suitable number of states for modeling the onset-rhyme acoustic units. The number of states, tightly linked to the acoustical properties of the speech units, has affected to the performance of the onset-rhyme models. The second study will examine the efficiency of the two onset-rhyme models, CORM and PORM. These two models are generated from different combinations between the releasing consonant and vowel. The number of phonotactic onset and contextual onset units are different, while both models have the same number rhyme units. The syllable recognition results of CORM and PORM will be compared. Apart from the syllable level, the analysis is carried on the acoustic level, the onset-rhyme units. Analysis of recognition results at the bottom level of the system reveals the actual efficiency of the speech unit. The final study will investigate the performance of the system when the number of mixture components is increased.

5.5.1 Experimental Results on Determining the Number of States

In order to determine the appropriate number of states for acoustic modeling of the onset-rhyme models, the number of states should be initially set equal to that of the phone unit. The syllable recognition results will be compared to the phone units and the error will be analyzed. Then, the number of states will be adjusted, corresponding to the acoustical properties of the onset-rhyme models, relating to hidden Markov models.

In this experiment, a smaller amount of onset-rhyme models, CORM, will be deployed to determine the number of states. The smaller number of units, the faster computation required. The standard 3-state left-to-right HMM with no skip state is used for both onset and rhyme modeling. The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- A single component of Gaussian distribution
- 275 onset models and 159 rhyme models
- Training speakers – 9 male speaker and 11 female speakers

- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

Table 5.5.1 Syllable recognition results of speech units modeling with 3-state HMM for male speaker-dependent system

Speech unit	Correction	Accuracy
monophone	15.6	4.5
intra-syllable triphone	30.9	12.7
inter-syllable triphone	37.4	17.5
CORM (o3r3)	22.4	8.0

Table 5.5.2 Syllable recognition results of speech units modeling with 3-state HMM for female speaker-dependent system

Speech unit	Correction	Accuracy
monophone	15.9	5.4
intra-syllable triphone	31.5	14.6
inter-syllable triphone	36.6	18.9
CORM (o3r3)	21.9	7.7

Table 5.5.3 Syllable recognition results of speech units modeling with 3-state HMM for male speaker-independent system

Speech unit	Correction	Accuracy
monophone	12.2	0.7
intra-syllable triphone	29.7	14.2
inter-syllable triphone	33.4	15.7
CORM (o3r3)	19.7	6.8

Table 5.5.4 Syllable recognition results of speech units modeling with 3-state HMM for female speaker-independent system

Speech unit	Correction	Accuracy
monophone	12.7	2.0
intra-syllable triphone	29.0	14.4
inter-syllable triphone	33.6	15.3
CORM (o3r3)	19.4	7.1

Table 5.5.5 Onset - rhyme recognition results of CORM for speaker-dependent system modeling onset and rhyme with 3-state HMM

Gender	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
male	43.4	31.1	27.1	17.4
female	42.9	30.0	26.1	16.6

Table 5.5.6 Onset - rhyme recognition results of CORM for speaker-independent system modeling onset and rhyme with 3-state HMM

Gender	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
male	41.1	25.5	24.7	11.0
female	39.4	22.3	23.5	10.2

The initial experiment, modeling of each onset-rhyme, using the number of states equivalent to the phone models, shows the efficiency of the onset-rhyme models worse than the phone models. From the experimental results shown in Tables 5.5.1-5.5.6, the syllable recognition rates of CORM, modeling with a 3-state HMM, are lower than those of the triphone models. Recognition result of rhyme units is deteriorated compared to onset units. Modeling the rhyme unit with a 3-state HMM cannot capture acoustic information containing in the unit sufficiently. From the acoustic analysis, the duration of rhymes is longer than onsets. A longer state HMM should be employed for modeling the rhyme units. Since rhyme unit appears to be a concatenation of two acoustic units, a vowel and a final consonant, a 6-state is chosen to model the rhyme units.

Table 5.5.7 Syllable recognition results for male SD system

Speech unit	Correction	Accuracy
CORM (o3r3)	22.4	8.0
CORM (o3r6)	32.6	21.5

Table 5.5.8 Syllable recognition results for female SD system

Speech unit	Correction	Accuracy
CORM (o3r3)	21.9	7.7
CORM (o3r6)	31.9	24.0

Table 5.5.9 Syllable recognition results for male SI system

Speech unit	Correction	Accuracy
CORM (o3r3)	19.7	6.8
CORM (o3r6)	28.4	18.2

Table 5.5.10 Syllable recognition results for female SI system

Speech unit	Correction	Accuracy
CORM (o3r3)	19.4	7.1
CORM (o3r6)	27.7	16.9

Table 5.5.11 Onset-rhyme recognition results of CORM for male SD system

CORM	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
o3r3	43.4	31.1	27.1	17.4
o3r6	50.1	43.9	41.8	31.6

Table 5.5.12 Onset-rhyme recognition results of CORM for female SD system

CORM	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
o3r3	42.9	30.0	26.1	16.6
o3r6	52.4	45.1	40.0	32.7

Table 5.5.13 Onset-rhyme recognition results of CORM for male SI system

CORM	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
o3r3	41.1	25.5	24.7	11.0
o3r6	47.2	40.6	38.0	31.4

Table 5.5.14 Onset-rhyme recognition results of CORM for female SI system

CORM	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
o3r3	39.4	22.3	23.5	10.2
o3r6	46.8	39.8	37.6	32.2

Recognition results of both syllable and onset-rhyme levels were significantly improved when modeling the rhyme units with a 6-state HMM (o3r6) as shown in Tables 5.5.7-5.5.14. Syllable accuracy rates increase around 13.5 % and 16.3 % for male and female SD system, and 11.4 % and 9.8 % for male and female SI system. In the acoustic unit level, the system modeling rhyme units with a 6-state HMM (o3r6) provides a better percentage correction and accuracy of the onset-rhyme units compared to the system modeling rhyme units with a 3-state HMM (o3r3). A substantial improvement of the rhyme units for male and female SD system is 14.2 % and 16.1 %, and 20.4 % and 20.0 % for male and female SI system. Conclusively, a 3-state HMM and 6-state HMM seems to be appropriate for acoustic modeling of onset and rhyme and will be used throughout the dissertation.

5.5.2 Experimental Results on Types of Onset-Rhyme Models

By considering the combination between the releasing consonant and vowel, there are two types of the onset-rhyme models, Phonotactic Onset-Rhyme Model (PORM) and Contextual Onset-Rhyme Model (CORM). These two types of onset-rhyme models have different onsets while their rhymes are still identical. The phonotactic onset is created differently for each releasing consonant and each vowel context, even for vowels in the same short-long pair. On the contrary, the contextual onset is modeled similarly for a given releasing consonant an either member of same short-long vowel pair. This experiment will explore both two types of onset-rhyme models in order to determine which type will be more efficient and suitable for Thai speech recognition systems. The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- A single component of Gaussian distribution

- The standard 3-state left-to-right HMM with no skip state for onset modeling
- The standard 6-state left-to-right HMM with no skip state for rhyme modeling
- CORM 275 onset models and 159 rhyme models
- PORM 621 onset models and 159 rhyme models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

Table 5.5.15 Syllable recognition results for male SD system

Speech unit	Correction	Accuracy
CORM (o3r6)	32.6	21.5
PORM (o3r6)	34.1	24.3

Table 5.5.16 Syllable recognition results for female SD system

Speech unit	Correction	Accuracy
CORM (o3r6)	31.9	24.0
PORM (o3r6)	33.7	26.1

Table 5.5.17 Syllable recognition results for male SI system

Speech unit	Correction	Accuracy
CORM (o3r6)	28.4	18.2
PORM (o3r6)	30.3	19.6

Table 5.5.18 Syllable recognition results for female SI system

Speech unit	Correction	Accuracy
CORM (o3r6)	27.7	16.9
PORM (o3r6)	29.8	17.5

Recognition results gave accuracies at 22.8 % and 17.6 % for SD and SI systems using the CORM and at 25.2 % and 18.6 % for SD and SI systems using the PORM.

Table 5.5.19 Onset-rhyme recognition results for male SD system

Speech unit	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
CORM (o3r6)	50.1	39.9	41.8	31.6
PORM (o3r6)	54.7	43.0	42.4	31.8

Table 5.5.20 Onset-rhyme recognition results for female SD system

Speech unit	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
CORM (o3r6)	52.4	45.1	40.0	32.7
PORM (o3r6)	55.3	48.2	41.7	33.6

Table 5.5.21 Onset-rhyme recognition results for male SI system

Speech unit	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
CORM (o3r6)	47.2	40.6	38.0	31.4
PORM (o3r6)	49.9	43.5	39.8	33.1

Table 5.5.22 Onset-rhyme recognition results for female SI system

Speech unit	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
CORM (o3r6)	46.8	39.8	37.6	32.2
PORM (o3r6)	49.4	42.7	39.1	32.9

On the result analysis of the onset-rhyme units, the correction and accuracy of the phonotactic onset are better than the contextual onset as shown in Tables 5.5.19-5.5.22. The accuracy of the PORM onset shows a modest improvement 7.4 % and 7.2 % over the CORM onset for speaker-dependent and speaker-independent systems. Since the rhyme units of both CORM and PORM are identical, the recognition results of the rhyme units are slightly different.

5.5.3 Experimental Results on Mixture Incrementing

To achieve higher performance, similar to the previous experiments, mixture incrementing is applied. The number of Gaussian mixture components was increased up to 8 mixture components per state.

The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- The standard 3-state left-to-right HMM with no skip state for onset modeling
- The standard 6-state left-to-right HMM with no skip state for rhyme modeling
- CORM 275 onset models and 159 rhyme models
- PORM 621 onset models and 159 rhyme models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

At higher mixture components, the recognition results are improved as shown in Tables 5.5.23-5.5.38. The PORM improves the syllable accuracy of the CORM by nearly 2.2 % and 1.9 % for SD and SI systems at eight mixture components per state. At the acoustic unit level, the PORM onset is better than the CORM onset in terms of accuracy. The improvement nearly 3.0 % and 4.1 % for SD and SI systems over the CORM onset shows the advantages of the PORM onset. However, the performance of the CORM and the PORM rhymes is insignificantly different.

Table 5.5.23 Syllable recognition results of CORM male SD system

No. of mixtures per state	Correction	Accuracy
1	32.6	21.5
2	38.9	27.6
4	42.7	33.2
8	47.8	38.9

Table 5.5.24 Syllable recognition results of CORM female SD system

No. of mixtures per state	Correction	Accuracy
1	31.9	24.0
2	36.6	29.5
4	42.1	35.4
8	49.1	42.5

Table 5.5.25 Syllable recognition results of CORM male SI system

No. of mixtures per state	Correction	Accuracy
1	28.4	18.2
2	33.5	22.7
4	37.5	26.1
8	41.3	29.8

Table 5.5.26 Syllable recognition results of CORM female SI system

No. of mixtures per state	Correction	Accuracy
1	27.7	16.9
2	32.6	23.5
4	38.7	26.8
8	43.4	31.2

Table 5.5.27 Syllable recognition results of PORM male SD system

No. of mixtures per state	Correction	Accuracy
1	34.1	24.3
2	41.9	30.4
4	44.8	36.0
8	50.5	41.7

Table 5.5.28 Syllable recognition results of PORM female SD system

No. of mixtures per state	Correction	Accuracy
1	33.7	26.1
2	40.4	33.3
4	46.2	38.6
8	52.7	44.0

Table 5.5.29 Syllable recognition results of PORM male SI system

No. of mixtures per state	Correction	Accuracy
1	30.3	19.6
2	34.4	24.1
4	38.0	27.9
8	42.9	31.3

Table 5.5.30 Syllable recognition results of PORM female SI system

No. of mixtures per state	Correction	Accuracy
1	29.8	17.5
2	34.7	25.6
4	39.6	29.3
8	45.3	33.5

Table 5.5.31 Onset-rhyme recognition results of CORM male SD system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	50.1	39.9	41.8	31.6
2	55.5	46.2	46.2	36.9
4	60.9	52.4	49.7	41.2
8	66.0	58.3	53.6	45.9

Table 5.5.32 Onset-rhyme recognition results of CORM female SD system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	52.4	45.1	40.0	32.7
2	58.3	51.9	44.5	38.1
4	62.8	57.0	47.5	41.7
8	69.4	66.3	52.6	49.6

Table 5.5.33 Onset-rhyme recognition results of CORM male SI system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	47.2	40.6	38.0	31.4
2	51.2	44.2	42.1	35.2
4	54.0	47.4	45.0	39.4
8	59.6	52.0	48.6	43.0

Table 5.5.34 Onset-rhyme recognition results of CORM female SI system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	46.8	39.8	37.6	32.2
2	52.2	44.8	42.9	34.5
4	56.8	48.7	46.9	39.8
8	60.5	53.1	49.4	44.4

Table 5.5.35 Onset-rhyme recognition results of PORM male SD system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	54.7	43.0	42.4	31.8
2	59.9	49.3	47.6	37.7
4	65.6	55.4	51.3	42.8
8	71.2	60.7	54.3	47.2

Table 5.5.36 Onset-rhyme recognition results of PORM female SD system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	55.3	48.2	41.7	33.6
2	62.1	56.9	45.6	39.8
4	67.3	63.4	49.1	42.0
8	74.7	69.8	54.0	51.5

Table 5.5.37 Onset-rhyme recognition results of PORM male SI system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	49.9	43.5	39.8	33.1
2	54.3	47.0	43.6	36.0
4	58.5	49.8	46.3	40.2
8	63.7	54.9	49.4	43.5

Table 5.5.38 Onset-rhyme recognition results of PORM female SI system

No. of mixtures per state	onset		rhyme	
	Correction	Accuracy	Correction	Accuracy
1	49.4	42.7	39.1	32.9
2	55.7	47.3	44.3	34.9
4	60.2	52.5	50.6	45.1
8	65.8	58.3	53.9	45.2

5.5.4 Discussion

The number of HMM states has an effect on the performance of the acoustic model. The onset-rhyme system using a 3-state HMM for modeling both onset and rhyme has a very low recognition results, comparing to the triphone systems. Recognition results in the acoustic level show that accuracy of the rhyme units is deteriorated compared to the onset units. Acoustic analysis reveals that the duration of rhymes is longer than onsets. A longer state HMM should be employed for modeling the rhyme units. The rhyme unit appears to be a concatenation of two acoustic units, a vowel and a final consonant. Therefore, a 6-state HMM is chosen to model the rhyme units. The recognition results of the rhyme units were substantially improved around 15.0 % and 20.2 % for SD and SI systems. Hence, a 3-state HMM and 6-state HMM seems to be appropriate for acoustic modeling of onset and rhyme.

The phonotactic onset units of the PORM model each releasing consonant in every possible vowel context in Thai syllables. The contextual onset units of the CORM treat a given releasing consonant the same way in contexts of vowels belonging to the same short-long pair. From speech signal characteristics, formant transition patterns of the same releasing consonant

are comparable within a short and long monophthong pair, including diphthongs.

The CORM then has a smaller number of onset units than the PORM. Both PORM and CORM share the same rhyme units, which cover every possible combination of a vowel and arresting consonant in Thai syllables. Since the onset of PORM completely models the initial consonant along with its transitional portion towards the vowel in every context, the PORM is more accurate than the CORM. Though the PORM has a higher accuracy than the CORM, the PORM system is more complex than the CORM due to the larger units of the PORM.

5.6 Experiments on Initial-Final Modeling

In phonological point of view, the context-dependent Initial-Final model is similar to the CORM. However, The difference between the CORM and the context-dependent Initial-Final model is the segmentation. The Initial-Final models are investigated the effect of the difference in this segmentation. The Initial followed by a Final has been used as the basic acoustic unit in Chinese speech recognition. There are two types of the Initial-Final models, context-independent Initial-Final and context-dependent Initial-Final. For the context-independent Initial-Final, the Initial comprises an initial consonant of the syllable while the Final consists of a vowel or diphthong part, including a possible medial or nasal ending. On the other hands, context-dependent Initial models are expanded from context-independent Initial models according to its following Final. The two types of Initial-Final models will be applied to the Thai language and its results will be analyzed. Finally, the number of Gaussian mixture components is increased up to eight mixtures per state.

5.6.1 Experimental Results on Types of Initial-Final Models

In this section, the first experiment is conducted to compare the performance of the context-independent Initial-Final and the context-dependent Initial-Final. The Initial-Final models are comparable to the onset-rhyme models in both acoustical and phonological points of views. To compare the Initial-Final models with the onset-rhyme models, the same parameters would be applied to both models. From the previous section, the

appropriate number of HMM states for onset-rhyme modeling was determined. This parameter is applied to Initial-Final modeling as well.

The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- A single component of Gaussian distribution
- The standard 3-state left-to-right HMM with no skip state for Initial modeling
- The standard 6-state left-to-right HMM with no skip state for Final modeling
- 33 context-independent Initial models and 159 Final models
- 279 context-dependent Initial models and 159 Final models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

The initial experiment was conducted to compare the efficiency the context-independent Initial-Final model and the context-independent Initial-Final model. The recognition results show clearly that the context-dependent Initial-Final model outperforms the context-independent Initial-Final model as indicated in Tables 5.6.1-5.6.4. The acoustic model incorporating contextual information improves the syllable accuracy approximately % and % for SD and SI systems.

Table 5.6.1 Syllable recognition results for male SD system

Initial-Final	Correction	Accuracy
Context-independent	24.6	15.6
Context-dependent	29.2	18.5

Table 5.6.2 Syllable recognition results for female SD system

Initial-Final	Correction	Accuracy
Context-independent	24.3	16.5
Context-dependent	29.7	19.4

Table 5.6.3 Syllable recognition results for male SI system

Initial-Final	Correction	Accuracy
Context-independent	21.7	11.2
Context-dependent	25.4	15.9

Table 5.6.4 Syllable recognition results for female SI system

Initial-Final	Correction	Accuracy
Context-independent	20.9	11.8
Context-dependent	25.0	15.6

The recognition results of acoustic units are shown in Tables 5.6.5-5.6.8. The context-dependent Initial shows better results in both SD and SI systems. A substantial improvement of the context-dependent Initial is 7.1 % and 8.7 % for SD and SI systems. On the other hands, the Final of the two models are identical. Exploiting contextual information in the context-dependent Initial does not affect the efficiency of the Final. The results show a slightly different accuracy of the Final model.

Table 5.6.5 Initial-Final recognition results for male SD system

Initial-Final	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
Context-independent	38.8	32.4	36.3	27.5
Context-dependent	45.3	39.1	38.2	29.0

Table 5.6.6 Initial-Final recognition results for female SD system

Initial-Final	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
Context-independent	41.2	36.8	35.6	30.9
Context-dependent	49.5	44.3	37.8	31.6

Table 5.6.7 Initial-Final recognition results for male SI system

Initial-Final	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
Context-independent	35.7	29.2	34.9	29.5
Context-dependent	44.3	39.1	35.9	30.7

Table 5.6.8 Initial-Final recognition results for female SI system

Initial-Final	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
Context-independent	34.4	28.7	32.2	27.5
Context-dependent	42.6	36.1	33.1	28.6

5.6.2 Experimental Results on Mixture Incrementing

A single Gaussian mixture was split to attain a higher recognition performance. The splitting mixture models were re-estimated using the forward-backward algorithm. Finally, the number of Gaussian distributions was increased to eight mixture components per state.

The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- The standard 3-state left-to-right HMM with no skip state for Initial modeling
- The standard 6-state left-to-right HMM with no skip state for Final modeling
- 33 context-independent Initial models and 159 Final models
- 279 context-dependent Initial models and 159 Final models
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

The recognition results show the improvement when the number of mixture components is increased. The context-dependent Initial-Final model highly surpasses the context-independent Initial-Final model in terms of recognition rates. From the syllable recognition results in Tables 5.6.9-5.6.16, the context-dependent Initial-Final model outperforms the context-independent Initial-Final model by 5.0 % and 4.7 % for SD and SI systems at eight mixture components per state.

Table 5.6.9 Syllable recognition results of context-independent Initial-Final (male SD system)

No. of mixtures per state	Correction	Accuracy
1	24.6	15.6
2	30.2	21.6
4	34.4	26.7
8	39.4	31.8

Table 5.6.10 Syllable recognition results of context-independent Initial-Final (female SD system)

No. of mixtures per state	Correction	Accuracy
1	24.3	16.5
2	31.2	25.1
4	35.2	29.4
8	39.6	34.0

Table 5.6.11 Syllable recognition results of context-independent Initial-Final (male SI system)

No. of mixtures per state	Correction	Accuracy
1	21.7	11.2
2	23.5	14.4
4	25.1	18.0
8	27.7	21.2

Table 5.6.12 Syllable recognition results of context-independent Initial-Final (female SI system)

No. of mixtures per state	Correction	Accuracy
1	20.9	11.8
2	25.7	17.0
4	28.6	20.2
8	31.7	24.4

Table 5.6.13 Syllable recognition results of context-dependent Initial-Final (male SD system)

No. of mixtures per state	Correction	Accuracy
1	29.2	18.5
2	33.4	24.0
4	39.9	30.1
8	45.5	36.4

Table 5.6.14 Syllable recognition results of context-dependent (female SD system)

No. of mixtures per state	Correction	Accuracy
1	29.7	19.4
2	34.5	24.6
4	39.4	31.2
8	46.8	39.3

Table 5.6.15 Syllable recognition results of context-dependent Initial-Final (male SI system)

No. of mixtures per state	Correction	Accuracy
1	25.4	15.9
2	31.7	19.3
4	34.6	22.8
8	38.4	26.5

Table 5.6.16 Syllable recognition results of context-dependent Initial-Final (female SI system)

No. of mixtures per state	Correction	Accuracy
1	25.0	15.6
2	30.3	19.7
4	34.9	23.8
8	41.8	28.4

Analysis of recognition results at the acoustic unit level reveals the actual efficiency of the units clearly. The recognition results of the Initial-Final model are shown in Tables 5.6.17-5.6.24. Using the context-dependent

Initial model can improve the accuracy by 5.4 % and 4.6 % over the context-independent Initial model for SD and SI systems at eight mixture components per state.

Table 5.6.17 Context-independent Initial-Final recognition results of male SD system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	38.8	32.4	36.3	27.5
2	44.6	38.6	41.0	33.1
4	49.6	44.5	44.1	39.3
8	55.4	50.6	48.4	42.7

Table 5.6.18 Context-independent Initial-Final recognition results of female SD system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	41.2	36.8	35.6	30.9
2	47.6	44.0	39.9	35.3
4	53.0	50.7	43.2	38.9
8	60.4	56.3	48.9	44.8

Table 5.6.19 Context-independent Initial-Final recognition results of male SI system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	35.7	29.2	34.9	29.5
2	41.4	34.6	37.5	32.7
4	47.0	40.4	40.9	35.3
8	52.1	46.4	43.4	39.7

Table 5.6.20 Context-independent Initial-Final recognition results of female SI system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	34.4	28.7	32.2	27.5
2	39.4	35.2	35.0	32.9
4	44.7	39.0	39.9	35.1
8	50.4	45.7	42.3	38.6

Table 5.6.21 Context-dependent Initial-Final recognition results of male SD system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	45.3	39.1	38.2	29.0
2	51.9	42.5	43.4	35.8
4	57.7	49.2	47.9	40.6
8	63.4	55.7	52.5	45.1

Table 5.6.22 Context-dependent Initial-Final recognition results of female SD system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	49.5	44.3	37.8	31.6
2	56.7	49.5	41.2	36.9
4	59.3	54.4	45.7	39.3
8	65.2	61.9	51.0	47.7

Table 5.6.23 Context-dependent Initial-Final recognition results of male SI system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	44.3	39.1	35.9	30.7
2	48.8	42.7	40.3	33.9
4	52.6	45.9	43.8	37.6
8	57.4	50.5	46.2	42.1

Table 5.6.24 Context-dependent Initial-Final recognition results of female SI system

No. of mixtures per state	Initial		Final	
	Correction	Accuracy	Correction	Accuracy
1	42.6	36.1	33.1	28.6
2	47.3	40.4	38.2	33.7
4	53.7	45.6	43.9	37.5
8	58.4	50.7	47.8	42.9

5.6.3 Discussion

Like a phone unit, context-dependent acoustic modeling provides a better result than context-independent acoustic modeling. The context-dependent Initial-Final yields a superior accuracy over the context-independent Initial-Final. This indicates that context-dependent acoustic modeling contributes an accurate acoustic unit. At eight mixture components per state, the accuracy of the context-dependent Initial is better than that of the context-independent Initial around 5.4 % and 4.6 %, while the Final of the context-dependent Initial-Final provides a higher accuracy around 2.7 % and 3.4 % for the SD and SI systems, respectively.

The context-dependent Initial-Final seems to be similar to the CORM. The number of units of these two models is identical. However, The difference between the CORM and the context-dependent Initial-Final model is the segmentation. The CORM onset consists of the releasing consonant and the transitional stage to its following vowel, whereas the context-dependent Initial contains the releasing consonant only. The recognition results show that the accuracy of the CORM onset is higher than the context-dependent Initial. However, in the other part of the syllable, the Final and the rhyme are comparable. The recognition results of the Final and the rhyme are insignificant different.

5.7 Experiments on Speech Recognition System using Acoustic Modeling Only

In order to obtain the actual efficiency of an acoustic model, a language model should not be applied. All kinds of speech units will be evaluated and compared their efficiency. The efficiency of all speech units will be compared by syllable recognition results since each speech unit has the equal number of syllables. On the other hands, the numbers of acoustic units of the phone, monophone and triphone, and the subsyllable, Initial-Final and onset-rhyme, are different. The results of acoustic unit will be analyzed and compared within the groups of phone and subsyllable units.

All speech units were trained and increased their mixture components up to 16 mixture components per state. Some speech units cannot be increased their mixture components over 16 mixture components per state because there are a few samples of those units, especially the triphones. A large number of mixtures cannot be fit to the model. The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- 16 Gaussian mixture components per state
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

5.7.1 Experimental Results

Syllable recognition rates for each speech unit are shown in Tables 5.7.1-5.7.4. Recognition results show that the accuracy of monophones and triphones, are lower than those for speech units larger than phones due to the high insertion. The inter-syllable triphones performs better accuracy than monophones and intra-syllable triphones. The inter-syllable triphone system performs at 39.0 % and 24.8 % accuracy for male SD and SI systems, and 41.7 %, and 28.8 % accuracy for female SD and SI systems. The PORM gives the highest accuracy at 45.4 % for male SD system and 48.1 % for female SD system. For the SI systems, the PORM also gives the highest accuracy. The accuracy of the CD Initial-Final is 40.9 % and 29.3 % for male SD and SI systems and 44.2 % and 30.4 % for female SD and SI

systems. This is better than that for phone modeling units and the CI Initial-Final model, and worse than those for either PORM or CORM.

Table 5.7.1 Syllable recognition results for male SD system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	29.4	17.3
Intra-syllable triphone	48.0	31.8
Inter-syllable triphone	52.2	39.0
CI Initial-Final	44.8	37.7
CD Initial-Final	50.7	40.9
CORM	52.8	42.9
PORM	55.3	45.4

Table 5.7.2 Syllable recognition results for female SD system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	30.1	21.8
Intra-syllable triphone	47.7	37.7
Inter-syllable triphone	53.1	41.7
CI Initial-Final	44.4	39.0
CD Initial-Final	50.8	44.2
CORM	54.6	46.3
PORM	56.1	48.1

Table 5.7.3 Syllable recognition results for male SI system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	18.8	9.0
Intra-syllable triphone	35.2	19.2
Inter-syllable triphone	41.4	24.8
CI Initial-Final	31.7	25.6
CD Initial-Final	41.5	29.3
CORM	44.3	32.7
PORM	45.9	34.1

Table 5.7.4 Syllable recognition results for female SI system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	22.3	11.5
Intra-syllable triphone	39.2	20.0
Inter-syllable triphone	43.8	28.8
CI Initial-Final	32.9	26.6
CD Initial-Final	45.8	30.4
CORM	47.3	34.2
PORM	49.5	36.7

Apart from the syllable accuracy, the recognition results of Initial-Final and onset-rhyme acoustic units were analyzed. The analysis of the recognition results at the bottom level of the system, the acoustic level, reveals the actual efficiency of the speech unit. Since the context-dependent Initials and the onsets of PORM and CORM are context-dependent, they give better accuracy than the context-independent Initial as shown in Tables 5.7.9-5.7.12. Although the context-dependent Initial and the onset are the right context-dependent units, the onset outperforms the context-dependent Initial. The transitional portion towards the vowel, included in the onset, has contributed substantially to the precise modeling of the initial consonant segment. The accuracy of the onset in CORM is higher than the accuracy of the context-dependent Initial by 2.5 % and 2.1 % for male SD and SI systems, and 3.2 % and 3.0 % for female SD and SI systems. The onset of PORM provides better accuracy than the onset of CORM nearly 3.8 % and 4.2 % for SD and SI systems, and achieves the highest accuracy at 67.3 % and 59.4 % for male SD and SI systems, and 74.4 % and 62.7 % for female SD and SI systems. At the other part of the syllable, the accuracies of the Final and the rhyme are slightly different as shown in Tables 5.7.9-5.7.12. The Final and rhyme units give comparable accuracies due to the similar modeling of these units. The examples of aligned transcription of the recognized result and the reference label are shown in Figures 5.7.1-5.7.4.

Table 5.7.5 Phone recognition results of monophone and triphone for male SD system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	53.7	39.9
Intra-syllable triphone	69.5	47.4
Inter-syllable triphone	76.3	54.7

Table 5.7.6 Phone recognition results of monophone and triphone for female SD system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	56.1	45.5
Intra-syllable triphone	71.0	53.4
Inter-syllable triphone	77.8	59.8

Table 5.7.7 Phone recognition results of monophone and triphone for male SI system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	41.3	29.1
Intra-syllable triphone	56.2	36.1
Inter-syllable triphone	66.2	45.4

Table 5.7.8 Phone recognition results of monophone and triphone for female SI system using acoustic modeling only

Speech unit	Correction	Accuracy
monophone	45.3	31.8
Intra-syllable triphone	60.5	37.8
Inter-syllable triphone	66.0	47.9

Table 5.7.9 Recognition results of Initial-Final and Onset-Rhyme units for male SD system using acoustic modeling only

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	59.3	54.6	52.9	45.4
CD Initial-Final	67.8	61.7	55.8	49.7
CORM	70.9	64.2	57.9	51.2
PORM	74.1	67.3	59.4	51.8

Table 5.7.10 Recognition results of Initial-Final and Onset-Rhyme units for female SD system using acoustic modeling only

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	64.9	60.7	54.4	50.2
CD Initial-Final	70.8	65.9	56.0	53.0
CORM	73.5	69.1	57.3	55.9
PORM	78.3	74.4	58.7	55.6

Table 5.7.11 Recognition results of INTIAL-Final and Onset-Rhyme units for male SI system using acoustic modeling only

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	56.8	50.2	45.8	42.2
CD Initial-Final	61.5	54.9	49.3	45.1
CORM	64.5	57.0	51.9	47.4
PORM	66.7	59.4	52.6	47.6

Table 5.7.12 Recognition results of Initial-Final and Onset-Rhyme units for female SI system using acoustic modeling only

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	55.3	48.6	46.2	41.5
CD Initial-Final	60.2	54.5	51.5	45.7
CORM	64.9	57.5	53.6	46.2
PORM	68.3	62.7	57.4	47.8

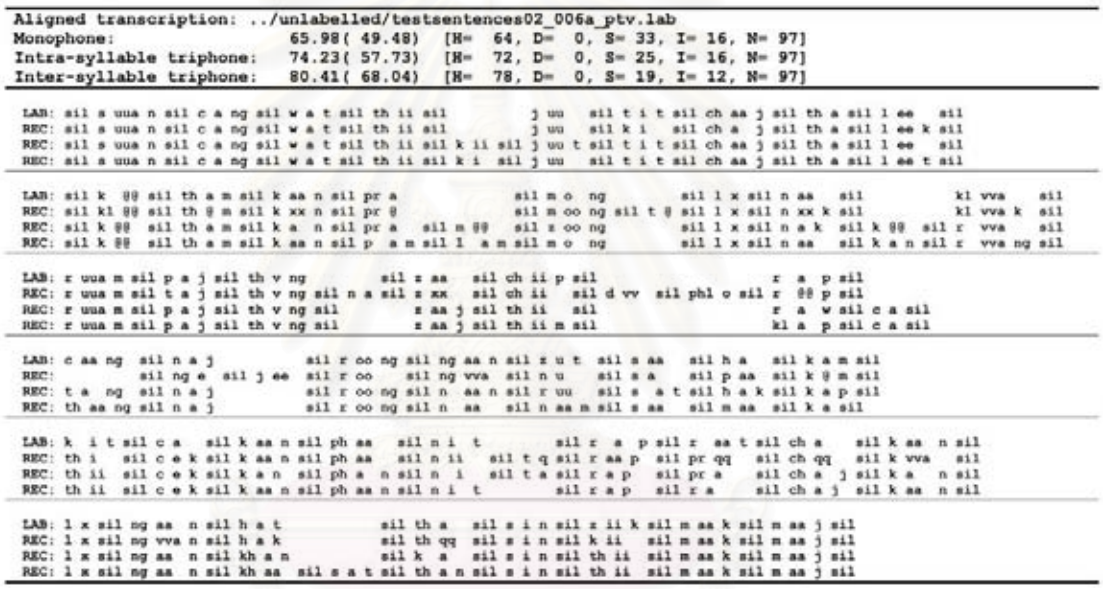


Figure 4.7.1 Alignment of reference vs recognition transcription (monophone, intra-syllable triphone, and inter-syllable triphone)

```
Aligned transcription: ../unlabelled/testsentences02_001_spk.lab
CI Initial-Final:    54.26( 50.00)  [H= 51, D=  0, S= 43, I=  4, N=94]
CD Initial-Final:    65.96( 63.83)  [H= 62, D=  0, S= 32, I=  2, N=94]

CI Initial-Final
LAB: sil ph vv_n th ii ph aa_k kl aa_ng pr a   k @@_p d uua_j j ii s i_p r e_t  c a_ng w a_t l x
REC: sil kh v_n  th ii f  a_k  kl aa_ng kr a_k k @    f uu    r ii s i_p r ee_k c a_ng w a_t t x

CD Initial-Final
LAB: sil ph_v vv_n th_i ii ph_a aa_k kl_a aa_ng pr_a a   k_@ @@_p d_u uua_j j_i ii s_i i_p r_e e_t c_a a_ng w_a a_t l_x x
REC: sil ph_v vv_n th_i ii f_a  a_k  kl_a aa_ng pr_a a_k k_@ @@_p r_u uua_j j_i ii s_i i_p r_e e_k c_a a_ng w_a a_t l_x x

CI Initial-Final
LAB: n v_ng kh ee_t k aa_n p o_k khr @@_ng ph i s ee_t d aa_j k xx c a_ng w a_t
REC: n v_ng th ee_t k a_t  k o_k khr @    f  i s ee   d e_k  k xx c @_ng w a_t

CD Initial-Final
LAB: n_v v_ng kh_e ee_t k_a aa_n p_o o_k khr_@ @@_ng ph_i i s_e ee_t d_a aa_j k_x xx  c_a a_ng w_a a_t
REC: n_v v_ng kh_e e_t  k_a a_t  p_o o_k khr_@ @@_p  ph_i i s_e ee  d_a a_j  k_x xx_k c_a a_w  w_a a_t

CI Initial-Final
LAB: k aa_n c a    n a    b u    r ii   ch a ch qq_ng s a_w      ch a_j n aa_t
REC: k a_k  c ee_t n xx_w b uu_t r ii_k ch x ch qq_ng s aa_w k uu ch a_j n aa_k

CD Initial-Final
LAB: k_a aa_n c_a a   n_a a b_u u    r_i ii   ch_a a ch_q qq_ng s_a a_w         ch_a a_j n_a aa_t
REC: k_a a_k  c_e e_n d_a a b_u uu_t r_i ii_t ch_a a ch_q qq_ng s_a aa_w w_u u_t ch_a a_j n_a aa_k

CI Initial-Final
LAB: n a   kh @@_n n aa   j o_k  n a   kh  @@_n p  a   th o_m n o_n  th a    b u    r ii          sil
REC: n a_k k  @_n  n aa_j j oo_k n a_k khw a_m  pr a_t th o_m n oo_n th xx_w m uu_t r ii_m h uua_j sil

CD Initial-Final
LAB: n_a a kh_@ @@_n n_a aa   j_o o_k n_a a   kh_@  @@_n p_a  a   th_o o_m n_o o_n th_a a    b_u u    r_i ii   sil
REC: n_a a kh_@ @_n  n_a aa_j j_o o_k n_a a_k khw_a a_m  pr_a a_t th_o o_m n_o o_n th_x xx_w h_u uu_t r_i ii_n sil
```

Figure 4.7.2 Alignment of reference vs recognition transcription (CI Initial-Final and CD Initial-Final)

```
Aligned transcription: ../unlabelled/testsentences02_006a_htr.lab
CORM:    86.21( 84.14)  [H= 125, D=  3, S= 17, I=  3, N=145]
PORM:    88.28( 86.21)  [H= 128, D=  0, S= 17, I=  3, N=145]

CORM
LAB: sil s_u uua_n sil c_a a_ng sil w_a a_t  sil th_i ii sil j_u uu sil t_i i_t sil ch_a aa_j sil th_a a sil l_e ee sil
REC: sil s_u uua_n sil c_a a_ng sil w_a aa_t sil th_i ii sil j_u uu sil t_i i_t sil ch_a aa_j           sil l_e ee sil

PORM
LAB: sil s_uua uua_n sil c_a a_ng sil w_a a_t sil th_ii ii sil j_uu uu sil t_i i_t sil ch_aa aa_j sil th_a a sil l_ee ee sil
REC: sil s_uua uua_n sil c_a a_ng sil w_a a_t sil th_ii ii sil j_uu uu sil t_i i_t sil ch_aa aa_j sil th_a a sil l_ee ee sil

CORM
LAB: k_@ @@ sil th_a a_m sil k_a aa_n sil pr_a a   sil m_o o_ng sil l_x x sil n_a aa          sil kl_v vva sil r_u uua_m
REC: k_@ @@ sil th_a a_m sil k_a a_n  sil pr_a a_m sil w_o o_ng sil l_x x sil n_a aa sil t_i i sil z_v  vva sil r_u uua_m

PORM
LAB: k_@@ @@ sil th_a a_m sil k_aa aa_n sil pr_a a sil m_o o_ng sil l_x x sil n_aa aa          sil kl_vva vva sil r_uua uua_m
REC: k_@@ @@ sil th_a a_m sil k_aa aa_n sil pr_a a sil m_o o_ng sil l_x x sil n_aa aa sil t_i i sil l_vva  vva sil r_uua uua_m

CORM
LAB: p_a a_j sil th_v v_ng sil z_a aa   sil ch_i ii_p sil r_a a_p sil c_a aa_ng sil n_a a_j sil r_o oo_ng sil ng_a aa_n sil
REC: p_a a_j sil th_v v_ng sil z_a aa_n sil th_i ii   sil r_a a_p sil c_a aa_ng sil n_a a_j sil r_o oo_ng sil ng_a aa_n sil

PORM
LAB: p_a a_j sil th_v v_ng sil z_aa aa   sil ch_ii ii_p sil r_a a_p sil c_aa aa_ng sil n_a a_j sil r_oo oo_ng sil ng_aa aa_n sil
REC: p_a a_j sil th_v v_ng sil z_aa aa_n sil ch_ii i_p  sil r_a a_p sil c_aa aa_ng sil n_a a_j sil r_oo oo_ng sil ng_aa a_n  sil

CORM
LAB: z_u u_t sil s_aa aa sil h_a a sil k_a a_m sil k_i i_t sil c_a a   sil k_aa aa_n sil ph_aa aa sil n_i i_t sil r_a  a_p sil
REC: z_u u_t sil s_aa aa sil h_a a sil k_a a_m sil k_i i_t sil c_a a_t sil k_aa aa_n sil ph_aa aa sil n_i i_t sil pr_a a_p sil

PORM
LAB: z_u  u_t sil s_a aa sil h_a a sil k_a a_m  sil k_i i_t sil c_a a   sil k_a aa_n sil ph_a aa sil n_i i_t sil r_a  a_p sil
REC: ph_u u_t sil s_a aa sil h_a a sil k_a aa_m sil k_i i_t sil c_e e_k sil k_a aa_n sil ph_a aa sil n_i i_t sil pr_a a_p sil

CORM
LAB: sil r_a  aa_t sil ch_a a k_a aa_n sil l_x x ng_a aa_n h_a a_t th_a a s_i i_n  sil z_i ii_k sil m_a aa_k m_a aa_j sil
REC: sil pr_a aa   sil ch_i i k_a aa_n sil l_x x ng_a aa_n h_a a_t th_a a s_i i_ng sil z_i ii_k sil m_a aa_k m_a aa_j sil

PORM
LAB: sil r_aa aa_t sil ch_a a k_aa aa_n sil l_x x ng_aa aa_n h_a a_t th_a a s_i i_n  z_ii ii_k sil m_aa aa_k m_aa aa_j sil
REC: sil r_aa aa   sil ch_a a k_aa aa_n sil l_x x ng_aa aa_n h_a a_t th_a a s_i i_ng z_ii ii_k sil m_aa aa_k m_aa aa_j sil
```

Figure 4.7.3 Alignment of reference vs recognition transcription (CORM and PORM)

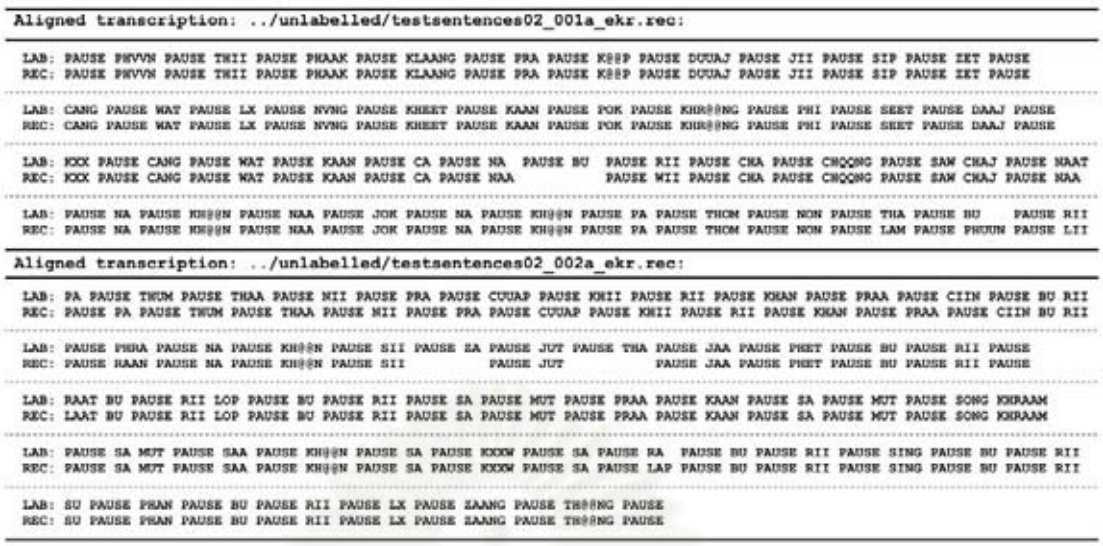


Figure 4.7.4 Alignment of reference vs recognition transcription in the syllable level

5.7.2 Discussion

Considering the training data, the onset-rhyme units seem to have a poorer scalability for the small amount of training data. The triphone system has the additional techniques to overcome the trainability problem. However, when the training data are large enough, the complexity of the onset-rhyme system is lower than the comparable triphone system when state-tying and context-clustering are not employed. Additionally, decoding process of the context-dependent phone model, especially inter-syllable triphone, is much more complex than the onset-rhyme models.

Compared with the context-independent Initial-Final model, the onset-rhyme models provide better modeling of speech segments as follows. The onset explicitly models internal coarticulatory effects within a syllable while the context-independent Initial implies a single model of a given initial consonant occurring in every vowel context. This makes the onset more precisely modeled than the context-independent Initial. In addition, a language model is embedded into the onset-rhyme model by means of phonological rules of composing this model into a syllable. Although the context-dependent Initial models the initial consonant differently depending on its following vowel, it cannot capture contextual information adequately. The Initial does not include the transitional portion towards the vowel within the model, which is an important acoustic cue. As a result, the onset-rhyme model gives a better accuracy than context-dependent Initial-Final model.

The recognition results indicated that the onset-rhyme outperforms the context-dependent Initial-Final model.

The phonotactic onset units of the PORM model each releasing consonant in every possible vowel context in Thai syllables. The contextual onset units of the CORM treat a given releasing consonant the same way in contexts of vowels belonging to the same short-long pair. From speech signal characteristics, formant transition patterns of the same releasing consonant are comparable within a short and long monophthong pair, including diphthongs. Both PORM and CORM share the same rhyme units, which cover every possible combination of a vowel and arresting consonant in Thai syllables. Since the onset of PORM completely models the initial consonant along with its transitional portion towards the vowel in every context, the PORM is more accurate than the CORM. However, The PORM improves the syllable accuracy of the CORM only 2.2 % and 2.0 % for the SD and SI system. Moreover, the CORM then has a smaller number of onset units than the PORM. From the experimental results, the CORM has the efficiency in terms of accuracy and complexity. Therefore, should be appropriate for acoustic modeling of the Thai language.

5.8 Experiments on Speech Recognition System using Acoustic Modeling and Language Modeling

Incorporating the language model can boost the performance of a speech recognition system. A bigram language model model is selected in this research while a complex language model would require more study on the syntactic and semantic rules. For a bigram language model, a syllable can only connect to syllables that can legally follow it. This bigram language model is used to evaluate the performance of the recognition system, which is composed of an acoustic model and grammar. The perplexity of this bigram language model is 252.03. The language model is used to perform the linguistic post-processing and determine the optimal syllable sequence.

The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- 16 Gaussian mixture components per state
- Training speakers – 9 male speaker and 11 female speakers

- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females

5.8.1 Experimental Results

The recognition results of systems using both acoustic modeling and language modeling are shown in Tables 5.8.1-5.8.4. As was the case for the systems using acoustic modeling only, the PORM attains the highest syllable accuracy at 75.2 % for male SD and 75.6 % for female SD systems. For the SI systems, the PORM also achieves the highest syllable accuracy at 62.8 % for male speakers and 64.8 % for female speakers. The system using a CD phone unit, inter-syllable triphone, performs at 68.8 % and 53.7 % accuracy for male SD and SI systems, and 70.1 % and 54.9 % accuracy for female SD and SI systems, better than the system with monophone and intra-syllable triphone units, and worse than the systems utilizing speech units larger than phones. It is noticeable that the language modeling, incorporated in speech recognition system, increases the performance of the system. The accuracy of PORM is boosted from 45.4 % to 75.2 % for male SD system and from 48.1 % to 75.6 % for female SD system, compared with a system using acoustic modeling only. The accuracy of PORM in the SI systems, as well as in the SD systems, is increased from 34.1 % to 62.8 % for male speakers and from 36.7 % to 64.8 % for female speakers. Figures 5.8.1-5.8.4 show the comparison of the syllable accuracy between the system using acoustic model only and the system using both acoustic model and language model.

Table 5.8.1 Syllable recognition results for male SD system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	50.1	42.9
Intra-syllable triphone	67.7	62.6
Inter-syllable triphone	73.4	68.8
CI Initial-Final	66.2	64.1
CD Initial-Final	72.9	71.0
CORM	75.9	73.3
PORM	77.4	75.2

Table 5.8.2 Syllable recognition results for female SD system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	51.7	42.7
Intra-syllable triphone	68.1	61.2
Inter-syllable triphone	74.7	70.1
CI Initial-Final	67.9	65.7
CD Initial-Final	73.6	71.4
CORM	76.1	73.8
PORM	78.5	75.6

Table 5.8.3 Syllable recognition results for male SI system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	42.3	35.8
Intra-syllable triphone	53.1	46.3
Inter-syllable triphone	58.5	53.7
CI Initial-Final	49.4	47.6
CD Initial-Final	60.2	57.9
CORM	63.7	61.7
PORM	64.4	62.8

Table 5.8.4 Syllable recognition results for female SI system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	44.3	38.2
Intra-syllable triphone	55.7	49.0
Inter-syllable triphone	60.4	54.9
CI Initial-Final	50.2	47.7
CD Initial-Final	59.1	56.3
CORM	64.2	61.9
PORM	66.5	64.8

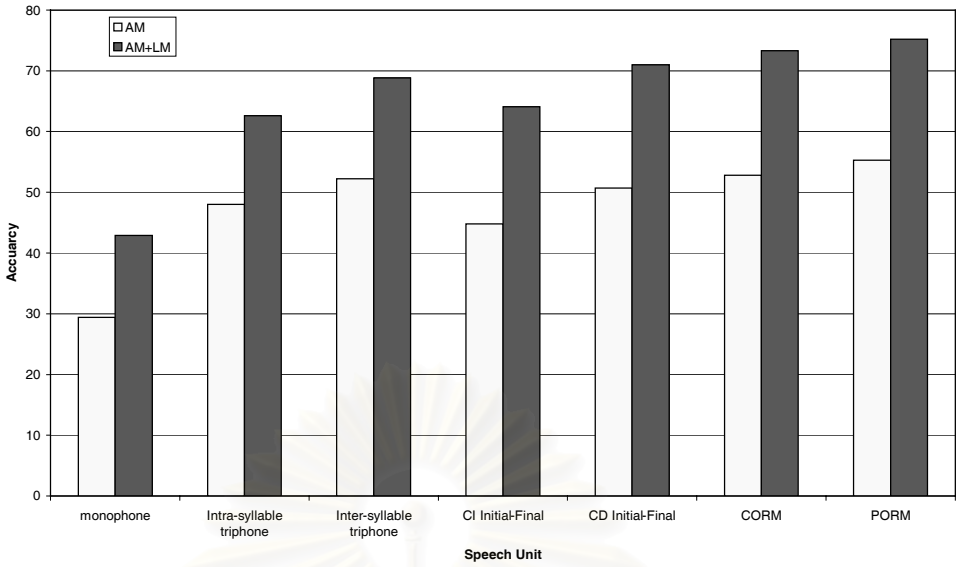


Figure 5.8.1 Syllable accuracy of the male SD system using acoustic model only and the male SD system using acoustic model and language model

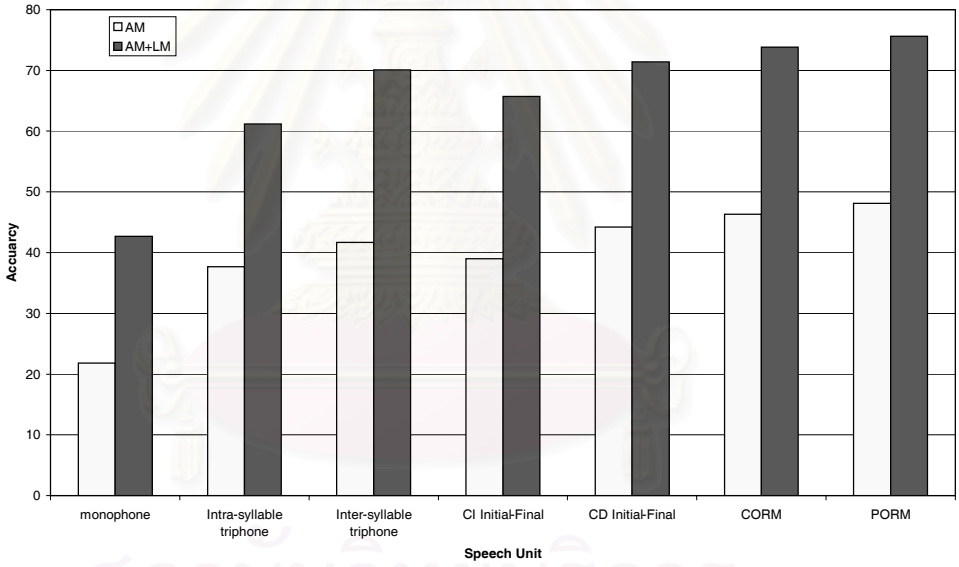


Figure 5.8.2 Syllable accuracy of the female SD system using acoustic model only and the female SD system using acoustic model and language model

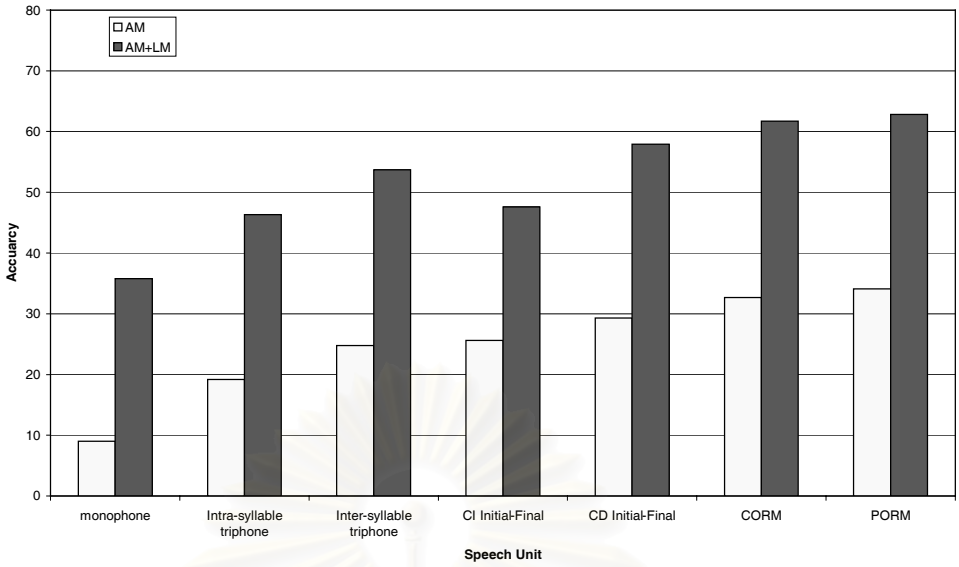


Figure 5.8.3 Syllable accuracy of the male SI system using acoustic model only and the male SI system using acoustic model and language model

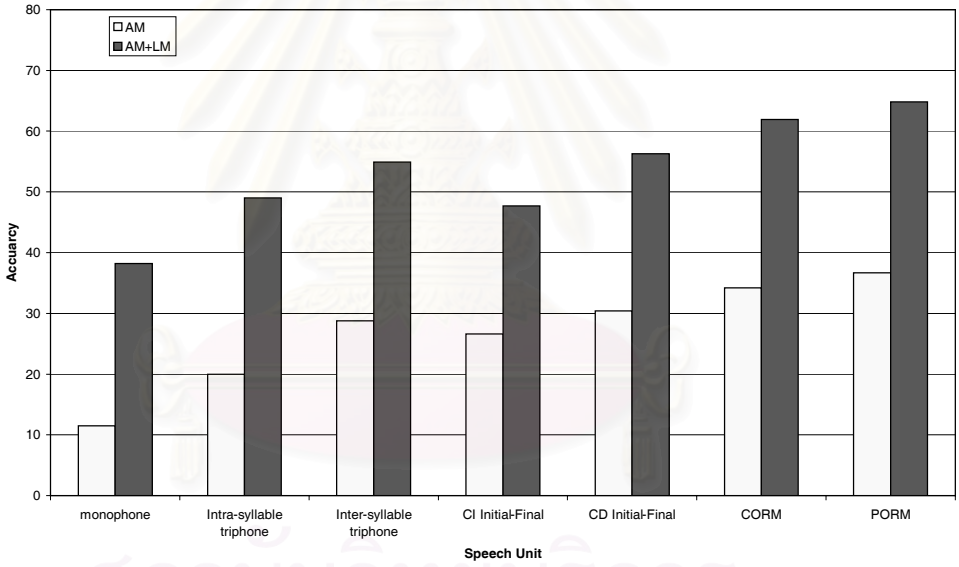


Figure 5.8.4 Syllable accuracy of the female SI system using acoustic model only and the female SI system using acoustic model and language model

As in the previous experiment, the recognition results of acoustic units of Initial-Final and onset-rhyme models were also analyzed. These results are shown in Tables 5.8.9-5.8.12. In this language model-combined system, the accuracies of onsets of both PORM and CORM still exceed that for the context-dependent Initial, and the accuracy of the PORM onset is higher than that of the CORM onset. Using the language model, the Initial and the onset accuracies are increased by around 18-25 % and 16-21 % for

both SD and SI systems. In addition, the accuracies of the Final and the rhyme are substantially improved by nearly 43-46 % and 45-47 % for both SD and SI systems when the language model is applied. Figures 5.8.5-5.8.12 show the comparison of the syllable accuracy between the system using acoustic model only and the system using both acoustic model and language model. The examples of aligned transcription of the recognized result and the reference label are shown in Figures 5.8.13-5.8.14.

Table 5.8.5 Phone recognition results of monophone and triphone for male SD system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	68.7	57.4
Intra-syllable triphone	72.2	63.7
Inter-syllable triphone	83.6	72.1

Table 5.8.6 Phone recognition results of monophone and triphone for female SD system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	69.4	62.5
Intra-syllable triphone	75.2	64.1
Inter-syllable triphone	84.6	73.2

Table 5.8.7 Phone recognition results of monophone and triphone for male SI system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	50.9	42.6
Intra-syllable triphone	62.6	58.2
Inter-syllable triphone	70.7	64.4

Table 5.8.8 Phone recognition results of monophone and triphone for female SI system using acoustic modeling and language modeling

Speech unit	Correction	Accuracy
monophone	54.8	47.5
Intra-syllable triphone	64.2	57.8
Inter-syllable triphone	71.3	65.3

Table 5.8.9 Recognition results of Initial-Final and Onset-Rhyme units for male SD system using acoustic modeling and language modeling

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	69.4	67.2	68.6	66.3
CD Initial-Final	75.9	74.5	74.5	73.1
CORM	79.0	77.8	79.5	77.3
PORM	82.1	79.7	80.9	78.2

Table 5.8.10 Recognition results of Initial-Final and Onset-Rhyme units for female SD system using acoustic modeling and language modeling

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	71.4	68.0	72.5	69.3
CD Initial-Final	77.8	75.1	76.3	73.7
CORM	81.9	78.7	80.5	77.2
PORM	83.5	80.6	80.7	78.1

Table 5.8.11 Recognition results of Initial-Final and Onset-Rhyme units for male SI system using acoustic modeling and language modeling

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	64.2	61.2	63.7	61.7
CD Initial-Final	69.1	67.5	67.6	65.3
CORM	72.5	69.8	70.4	68.6
PORM	73.4	70.3	71.8	69.5

Table 5.8.12 Recognition results of Initial-Final and Onset-Rhyme units for female SI system using acoustic modeling and language modeling

Speech unit	Initial/Onset		Final/Rhyme	
	Correction	Accuracy	Correction	Accuracy
CI Initial-Final	67.2	64.0	65.1	61.9
CD Initial-Final	70.7	68.0	68.6	65.9
CORM	74.5	71.6	71.6	69.7
PORM	76.3	73.8	72.8	70.2

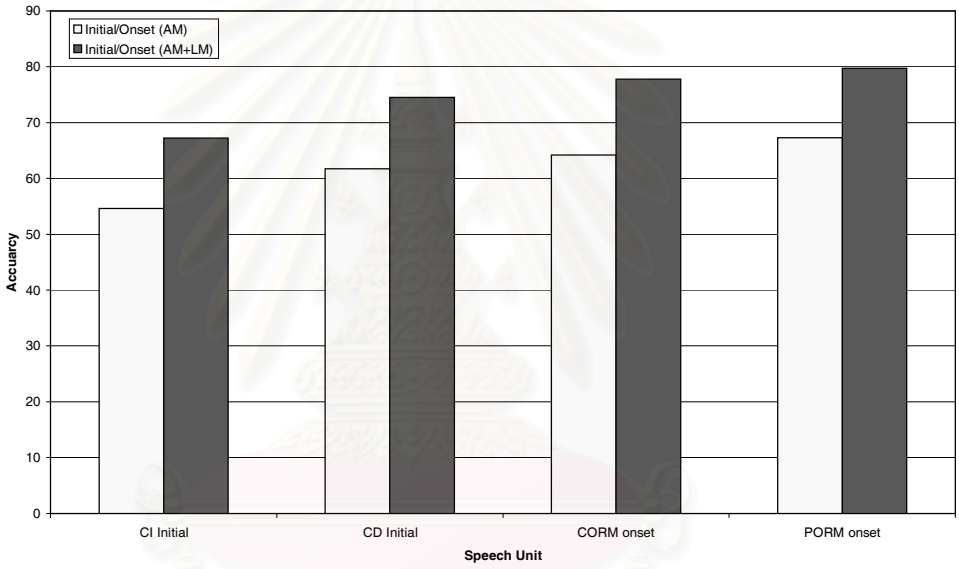


Figure 5.8.5 Initial/Onset accuracy of the male SD system using acoustic model only and the male SD system using acoustic model and language model

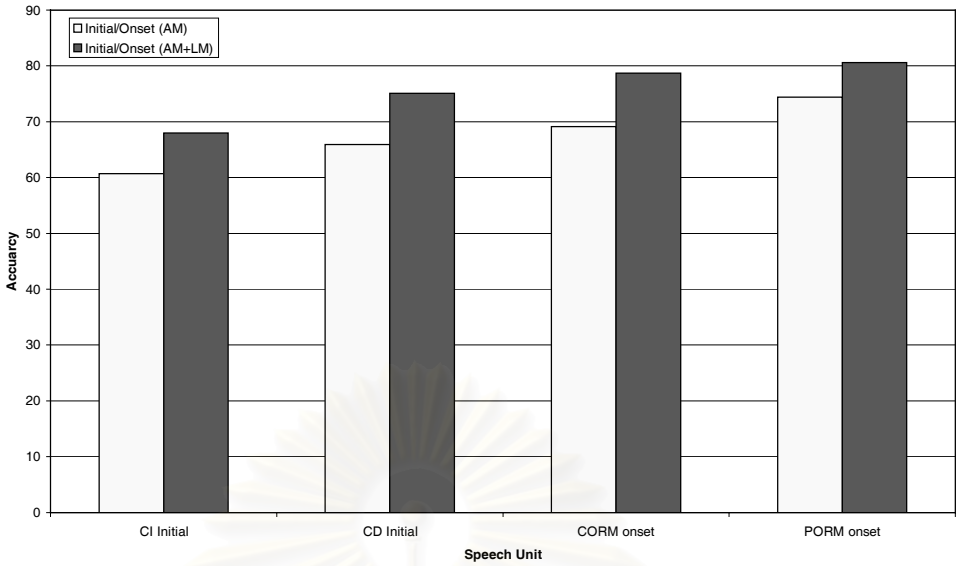


Figure 5.8.6 Initial/Onset accuracy of the female SD system using acoustic model only and the female SD system using acoustic model and language model

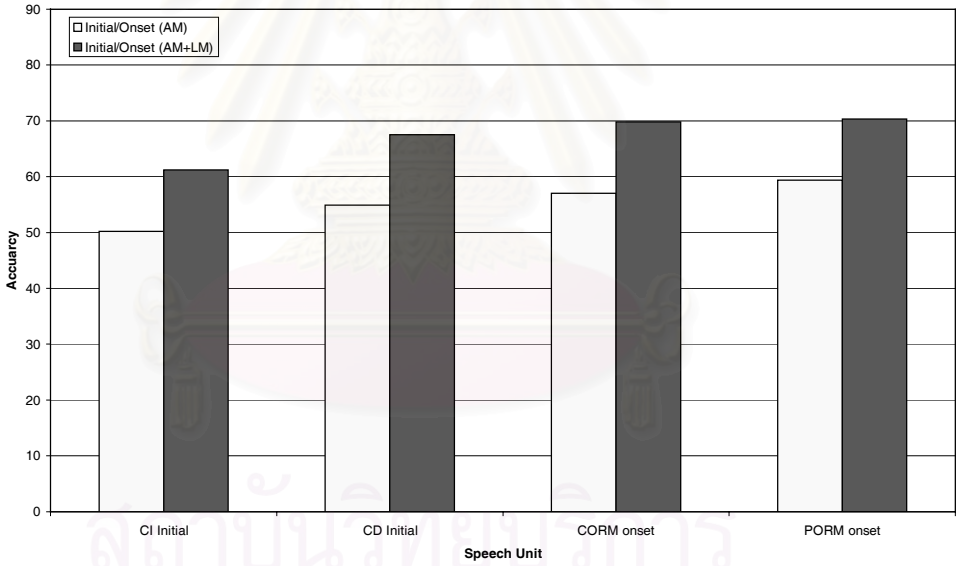


Figure 5.8.7 Initial/Onset accuracy of the male SI system using acoustic model only and the male SI system using acoustic model and language model

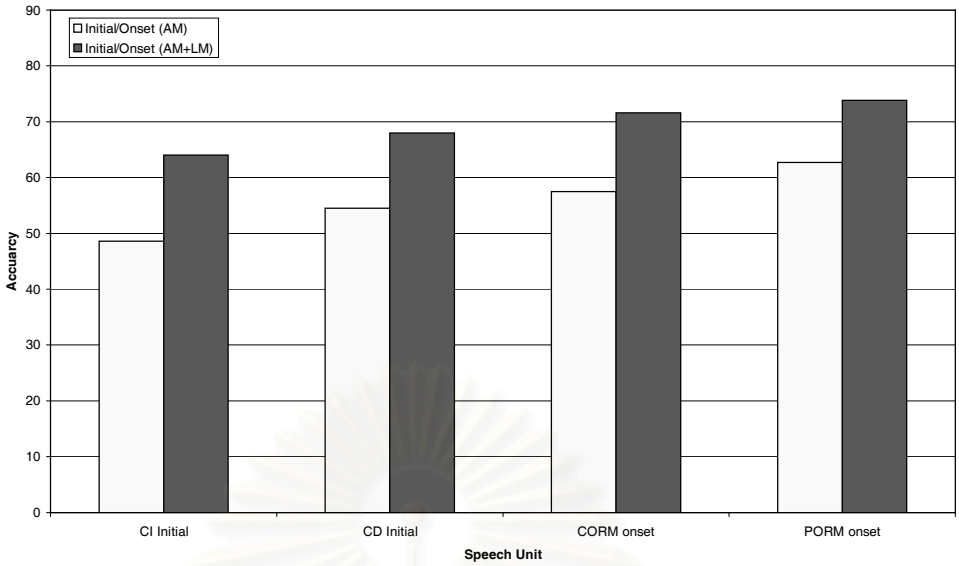


Figure 5.8.8 Initial/Onset accuracy of the female SI system using acoustic model only and the female SI system using acoustic model and language model

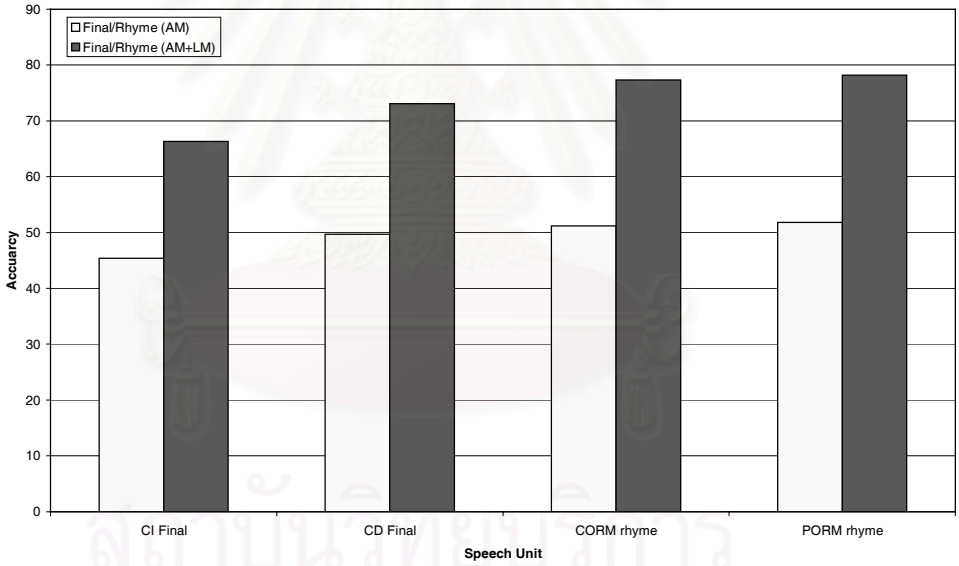


Figure 5.8.9 Final/Rhyme accuracy of the male SD system using acoustic model only and the male SD system using acoustic model and language model

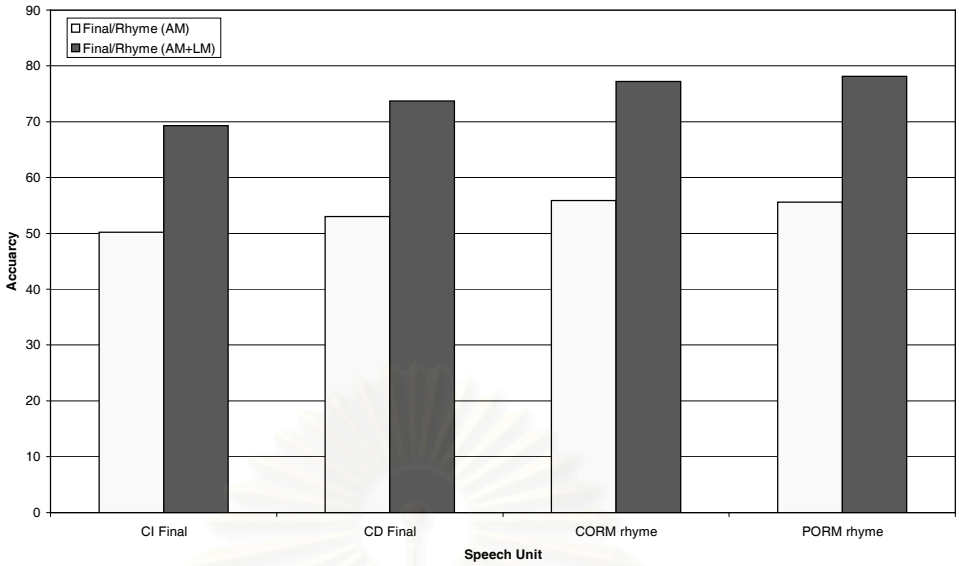


Figure 5.8.10 Final/Rhyme accuracy of the female SD system using acoustic model only and the female SD system using acoustic model and language model

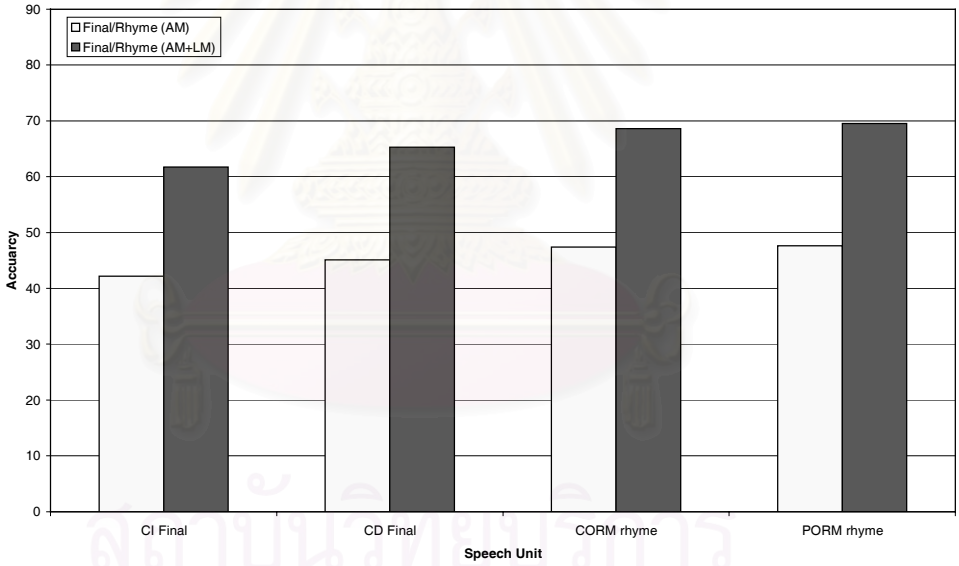


Figure 5.8.11 Final/Rhyme accuracy of the male SI system using acoustic model only and the male SI system using acoustic model and language model

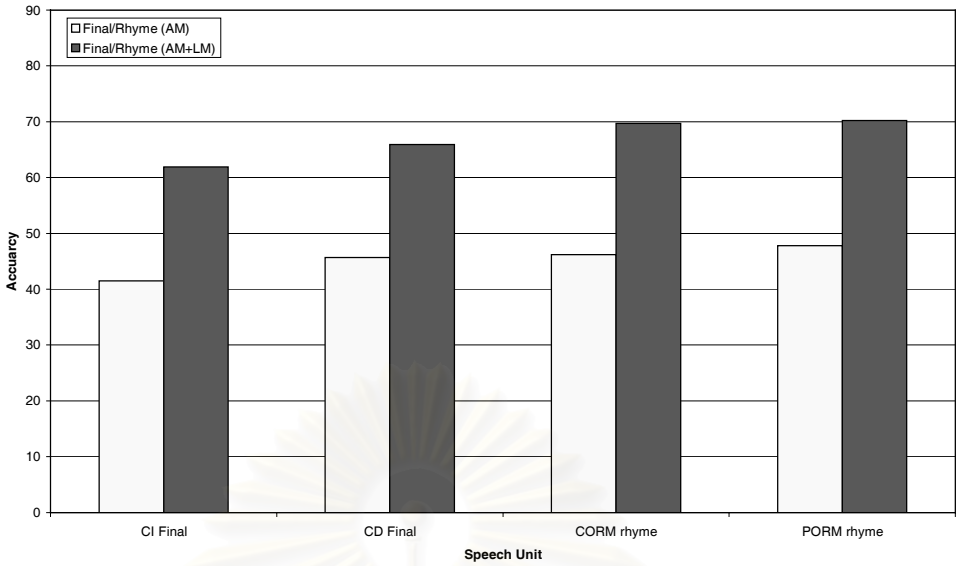


Figure 5.8.12 Final/Rhyme accuracy of the female SI system using acoustic model only and the female SI system using acoustic model and language model

```
Aligned transcription: ../unlabelled/testsentences02a_005_htr.lab
AM:       73.79( 69.66)  [H= 107, D=  3, S= 35, I=  6, N=145]
AM+LM:    93.10( 93.10)  [H= 135, D=  3, S=  7, I=  0, N=145]

LAB: sil pr_a a sil ch_a aa  sil k_@ @@_n sil s_u uua_n sil j_a a_j sil pr_a  a   sil k_@ @@_p  sil z_a aa sil ch_i ii_p sil
REC: sil kr_a a sil s_a  a_k sil k_@ @@   sil s_u uua_n sil j_@ @_j sil khr_a a_k sil k_@ @@_ng sil m_a a  sil ch_i ii_p sil
REC: sil pr_a a sil ch_a aa  sil k_@ @@_n sil s_u uua_n sil j_a a_j sil pr_a a    sil k_@ @@_p  sil z_a aa sil ch_i ii_p sil

LAB: sil k_a a sil s_e ee_t sil tr_a a   sil k_a a_m sil th_a a_m sil n_a aa   sil pl_u uu_k sil kh_a aa_w sil m_i ii sil
REC: sil k_a a sil s_e e_t  sil k_a  a_k sil t_@ @_m sil th_a a_m sil n_a aa_m sil pl_u u_k  sil kh_a a_w  sil m_i ii sil
REC: sil k_a a sil s_e ee_t sil tr_a a   sil k_a a_m sil th_a a_m sil n_a aa   sil pl_u uu_k sil kh_a aa_w sil m_i ii sil

LAB: s_u uua_n sil ph_o o_n sil l_a a   sil m_a aa_j sil l_x x sil th_a a_m sil r_a a_j  sil n_@ @@_k sil c_a aa_k sil n_i ii
REC: s_u uua_n sil ph_o o_n sil l_a a_m sil m_a a_j  sil l_x x sil th_a a_m sil r_a a_j  sil n_@ @@   sil s_a aa_k sil m_i iia_w
REC: s_u uua_n             sil r_a a   sil m_a aa_j sil l_x x sil th_a a_m sil r_a aa_j sil n_@ @@_k sil c_a aa_k sil n_i ii

LAB: sil j_a a_ng sil m_i ii sil k_i i_t sil c_a a   sil k_a aa_n sil p_a a sil s_u u sil s_a a_t sil l_i iia_ng sil kh_o oo
REC:              sil m_i ii sil k_i i_t sil c_e e_t sil k_a aa_n sil t_a a sil s_u u sil s_@ @   sil l_i iia_ng sil kh_o oo
REC:              sil m_i ii sil k_i i_t sil c_a a   sil k_a aa_n sil p_a a sil s_u u sil s_a a_t sil l_i iia_ng sil kh_o o_p

LAB: sil ph_a a_n sil n_v vva            sil ph_a a_n sil n_o o_m          sil m_i ii sil f_a aa_m sil k_a a_j  sil l_x x
REC: sil ph_a a   sil n_v v_ng sil s_a a_p sil ph_a a_n sil n_o o_m sil ph_o o_p sil m_i ii sil f_a aa_m sil k_a aa_j sil l_x x
REC:              sil r_v vva            sil ph_a a   sil n_o o_m          sil m_i ii sil f_a aa_m sil k_a aa_j sil l_x x

LAB: sil l_i iia_ng sil pl_a aa sil
REC: sil l_i iia_ng sil l_@  @@ sil
REC: sil l_i iia_ng sil pl_a aa sil
```

Figure 5.8.13 Alignment of reference vs recognition transcription (acoustic modeling only and acoustic modeling +language modeling)

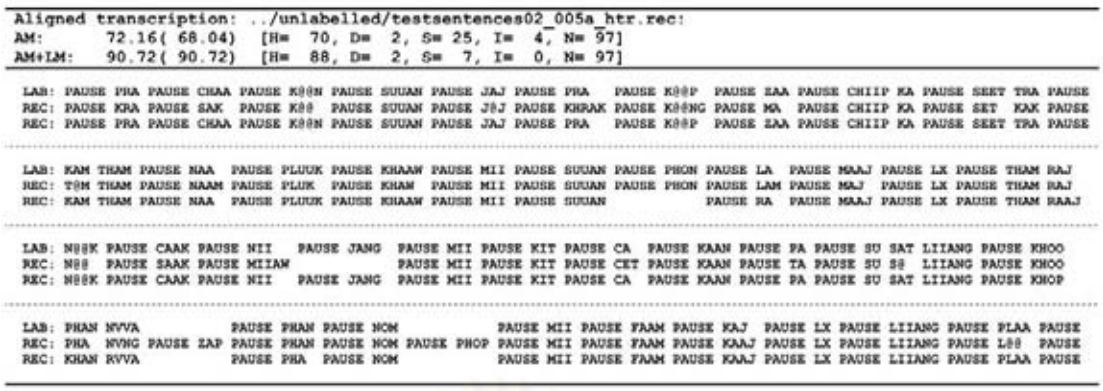


Figure 5.8.14 Alignment of reference vs recognition transcription in the syllable level
(acoustic modeling only and acoustic modeling+language modeling)

5.8.2 Discussion

Incorporating the language model in a large vocabulary speech recognition system reduces the ambiguities between the large set of alternative confusable words that might be hypothesized during the recognition. The recognition results of the systems using acoustic model and language model are clearly higher than those of the systems using acoustic model only. When the language model is applied, the substitutions, especially the stop-final consonant, are substantially diminished in the final/rhyme unit. A substantial improvement of final/rhyme units results in a high recognition rate of the system. The language model also enhances the accuracy of the Initial and onset. However, the improvement in syllable accuracy of he Initial and onset is much lower than the final/rhyme.

5.9 Experiments on Speech Recognition System using Different Test Sets

This experiment aims to evaluate the efficiency of the speech recognition system of both using acoustic modeling only and using acoustic modeling together with language modeling by two different test sets. The system was already tested with the first test set in the experiment 5.7 and 5.8. The second test set used in Visarut' s work (Ahkuputra, 2002) will be tested in this experiment. The results of these 2 test sets will be reported and compared as well.

The system configuration is set as follows:

- 12 order MFCCs with delta coefficients
- 16 Gaussian mixture components per state
- Training speakers – 9 male speaker and 11 female speakers
- Evaluation speakers – 9 speaker-dependent males and 5 speaker-independent males, and 11 speaker-dependent females and 5 speaker-independent females
- Test set I – 100 sentences, 4985 syllables
- Test set II – 30 sentences, 576 syllables

5.9.1 Experimental Results

Compared to the test set I, the test set II has the number of syllables, 576, less than that in the test set I, 4985. The vocabulary in the test set I would probably be complicated than those in the test set II. Figures 5.9.1-5.9.4 show the syllable accuracy of the systems using acoustic model only while Figures 5.0.5-5.9.8 show the syllable accuracy of the systems using acoustic model and language model. The recognition results show that the accuracy of the test set II is higher than the test set I for every speech unit. The recognition results of these two test sets are in the similar way that is the subsyllable units, Initial-Final and onset-rhyme, outperforms the phone units.

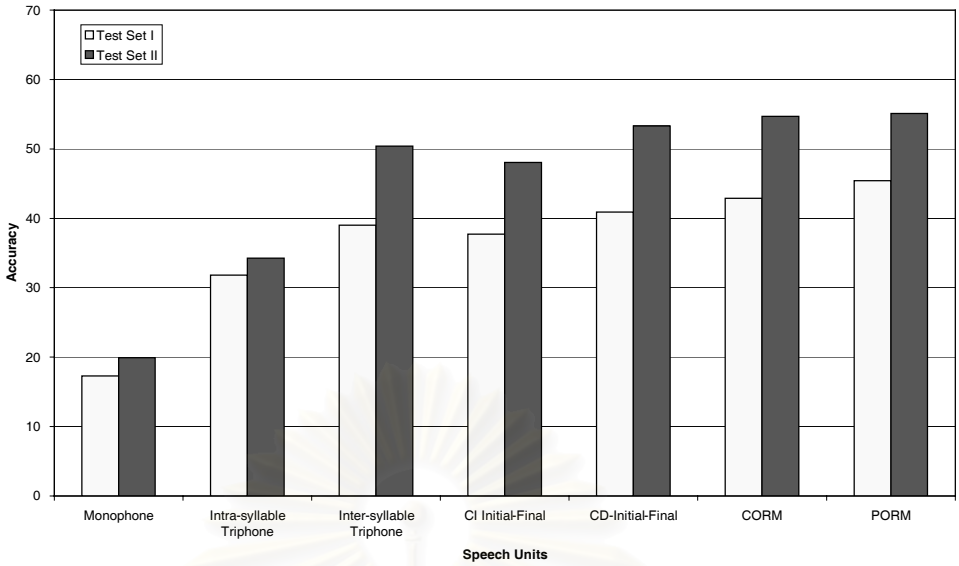


Figure 5.9.1 Syllable accuracy of the male SD system using acoustic model only

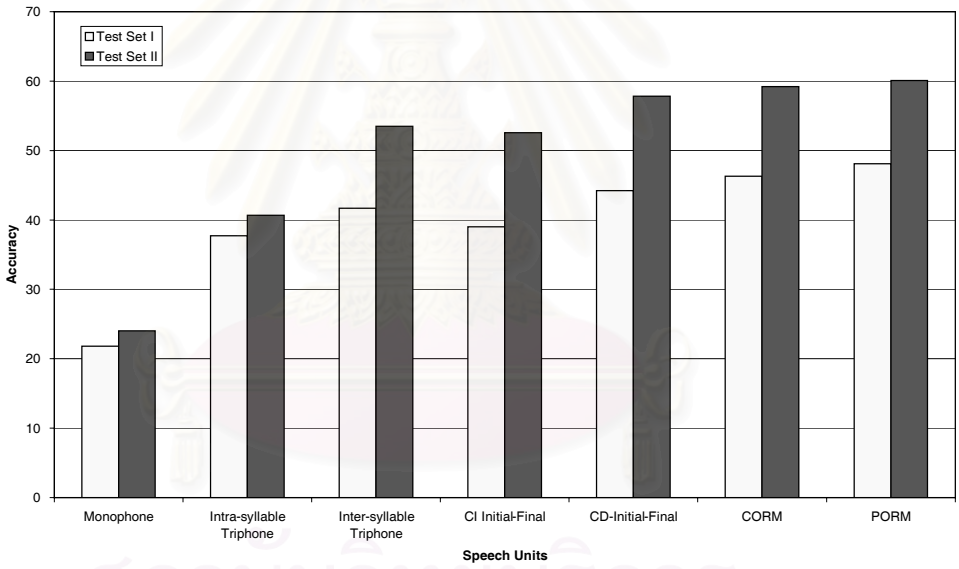


Figure 5.9.2 Syllable accuracy of the female SD system using acoustic model only

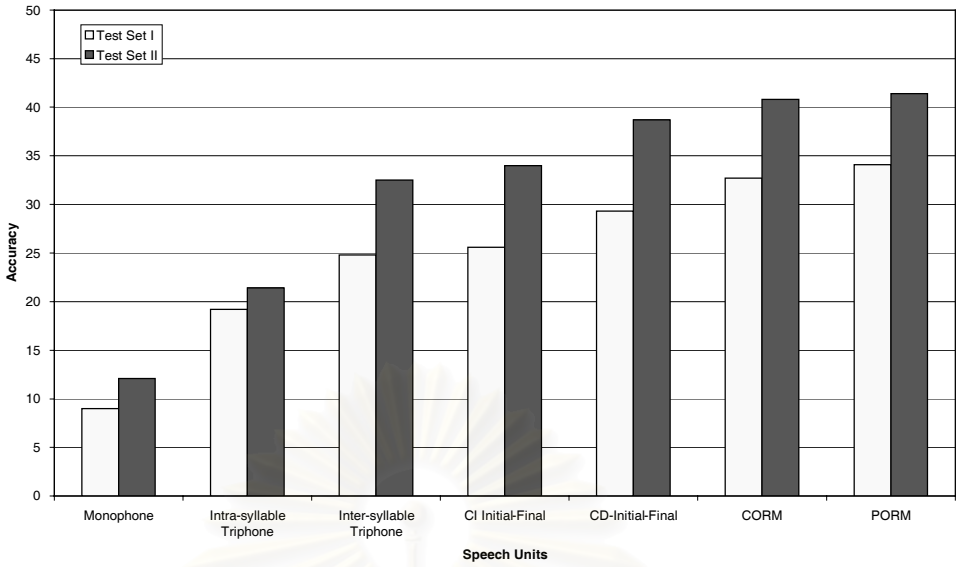


Figure 5.9.3 Syllable accuracy of the male SI system using acoustic model only

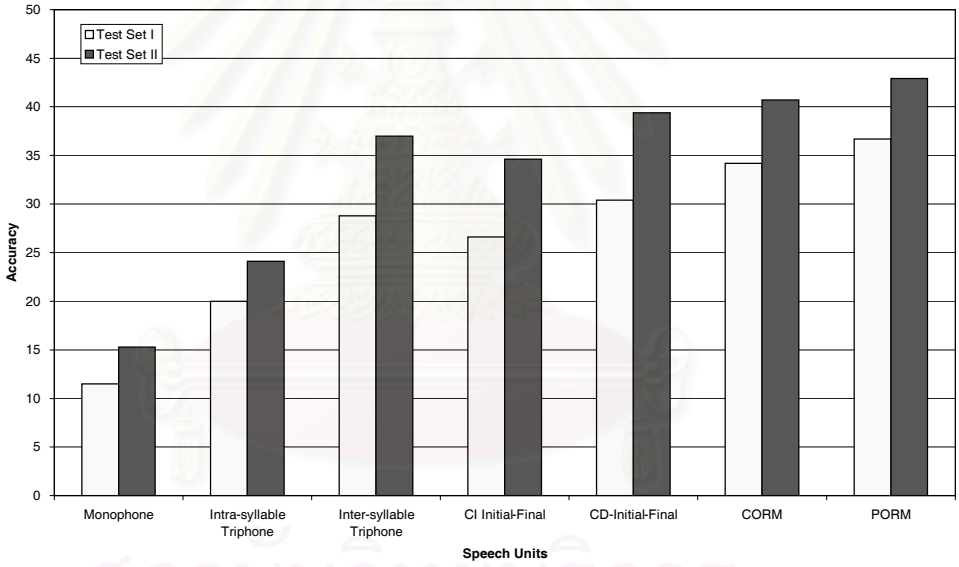


Figure 5.9.4 Syllable accuracy of the female SI system using acoustic model only

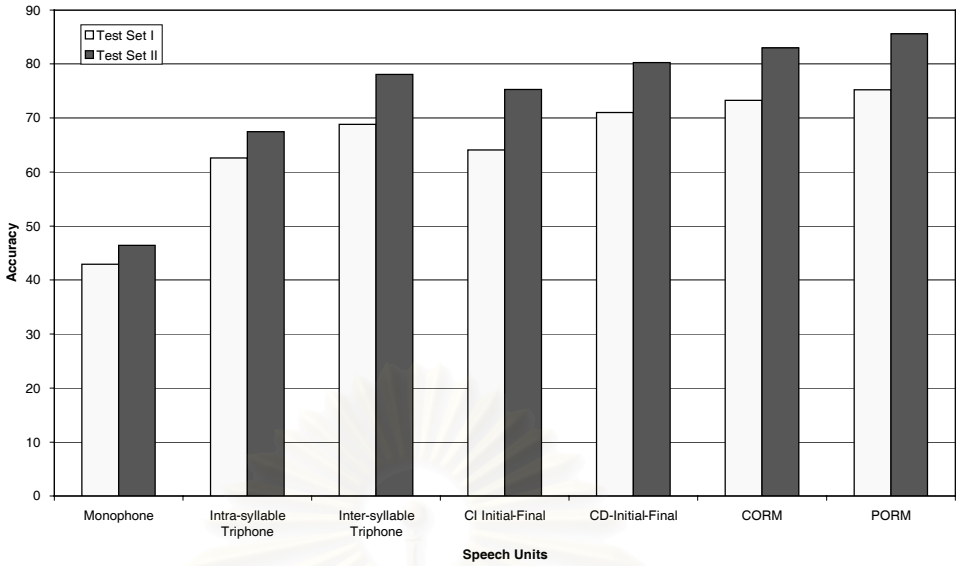


Figure 5.9.5 Syllable accuracy of the male SD system using acoustic model and language model

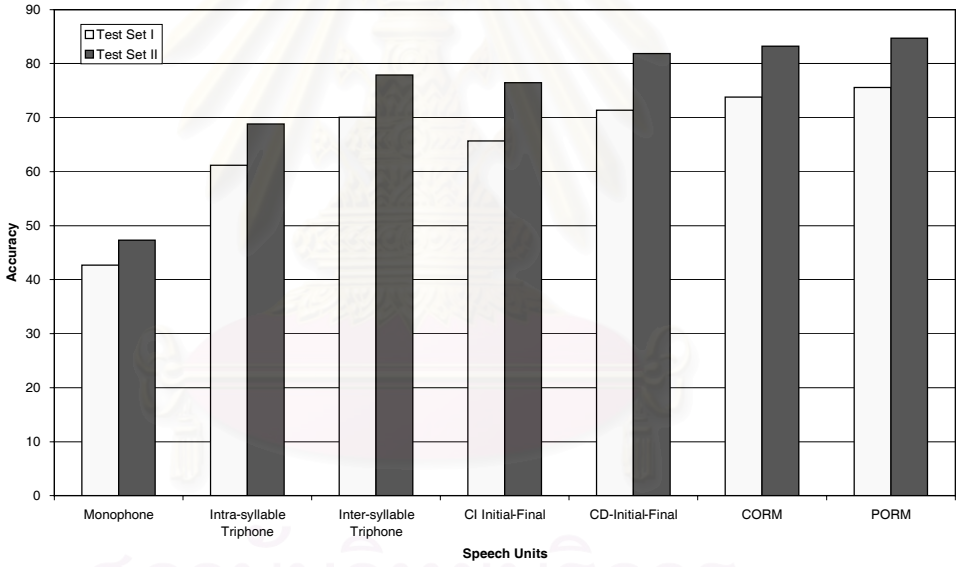


Figure 5.9.6 Syllable accuracy of the female SD system using acoustic model and language model

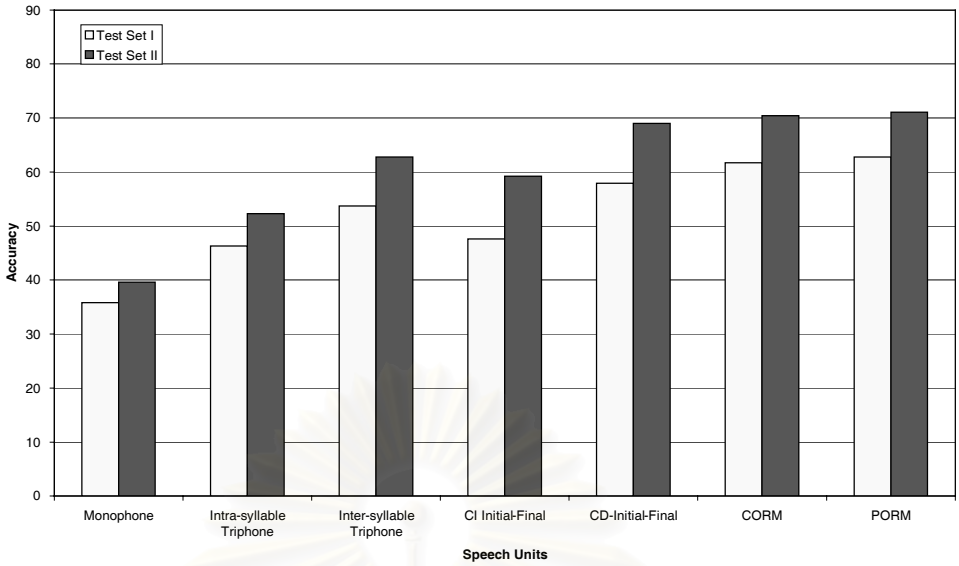


Figure 5.9.7 Syllable accuracy of the male SI system using acoustic model and language model

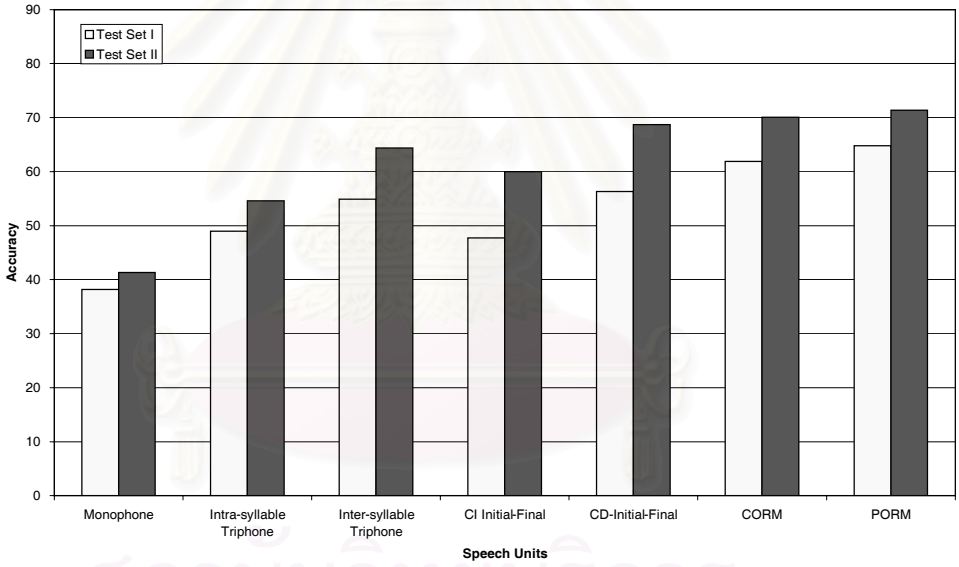


Figure 5.9.8 Syllable accuracy of the female SI system using acoustic model and language model

5.9.2 Discussion

To prove the superiority of the onset-rhyme models, the experiment using more than two kinds of corpora was conducted. Two test sets with the different number of syllables were employed in the experiment. The evaluation was carried on the system using acoustic model only and the system using both acoustic model and language model. The recognition results show the syllable accuracies of these two test sets are in the similar

way that is the subsyllable units, Initial-Final and onset-rhyme, outperforms the phone units. This verifies that the proposed acoustic unit, onset-rhyme, is superior in terms of accuracy.

5.10 Summary

In this chapter, the experiments were conducted on several speech units to study the efficiency of the acoustic model. The speech units used in this experiments are, the monophone, the triphone, the Initial-Final, and the onset-rhyme. Two onset-rhyme models, the Phonotactic Onset-Rhyme Model (PORM) and the Contextual Onset-Rhyme Model (CORM) are applied to Thai in a continuous speech recognition system in order to illustrate their feasibility. The results of all speech units are given and analyzed in details. The recognition results show major improvements over other acoustic units in many ways, indicating better performance of the models while maintaining manageable system complexity.

Chapter 6

Conclusions

This dissertation presents acoustic modeling of the rhyme units in the onset-rhyme models for Thai continuous speech recognition. Several conventional speech units were tested their effectiveness. The performance was analyzed both in terms of the computational complexity and the recognition accuracy. The experiments on recognition system using acoustic modeling only were conducted to obtain an actual efficiency of the acoustic model. On the other hand, the experiments on recognition system, incorporating the language model, were conducted to obtain the overall performance of a speech recognition system.

6.1 Conclusions of the Dissertation

In this dissertation, the onset-rhyme models are proposed and applied to speech recognition of Thai language. The main interest of this dissertation is the rhyme unit. From the acoustical point of view, in the syllable structure, the final consonant is strongly influenced by the vowel duration. This relationship occurs only between the vowel and the final consonant. In contrast, the initial consonant is not affected by the duration of the vowel. Hence, the vowel and the final consonant are tightly tied while an initial consonant is loosely tied with the vowel in the syllable. From a phonological point of view, a syllable is composed of a pair of an onset and a rhyme unit. The onset consists of an initial consonant and its transition towards the following vowel. Along with the onset, the rhyme is composed of a vowel, a final consonant, and a tone. The onset-rhyme not only includes its contextual information, but also embeds the language modeling at the syllable level. Therefore, the syllable should be decomposed into 2 parts, the onset and the rhyme. The Thai syllables can be recognized by identifying the onset and the rhyme.

Various conventional speech units were studied and compared their strengths and weaknesses in practical applications. This research mainly used four speech units. The context-independent phone, a monophone, was

modeled initially, and then this speech unit was used for building the triphone system. The modeling of other speech units, Initial-Final and onset-rhyme, depended on their types. The Initial and the onset were modeled differently, according to their context, while both of the final and the rhyme were modeled as left context-independent units. The three criteria, accuracy, trainability, and generalization must be considered in choosing the appropriate speech unit. Therefore, this research evaluates the efficiency of those speech units based on these criteria.

The hidden Markov models (HMM) were used to create the acoustic models. The left-right topology with no skipping state is selected. The phone units use three states for acoustic modeling. In the triphone modeling, the additional techniques were employed to create the triphone models from the monophone models. The numbers of HMM states were varied for the onset-rhyme according to its characteristics. The onsets use three states for modeling the initial consonant and the vowel transition while the rhymes, considered as two concatenating phone units, use six states for modeling the vowel and final consonant.

To test the gender-dependent effect, the systems were separately and conjointly trained on male and female speakers. The recognition results show that the gender-dependent system gives more accuracy than that of the gender-independent system.

Multiple mixture components provide the accurate acoustic model. The acoustic models can be created from a single mixture Gaussian distribution. Then, an iterative divide-by-two clustering algorithm was utilized to increase the number of Gaussian mixture components to the desired value. The more mixture components were created, the more accuracy the acoustic models have. However, increasing in the number of mixture components requires more computation. The experimental results showed that the recognition rates become higher when the numbers of mixture components are increased.

Recognition results of the system using acoustic models only show that the accuracy of monophones and triphones are lower than those for speech units larger than phones. Although the correction of the inter-syllable triphones is higher than that of the CD Initiai-Final model, its accuracy is lower than that of the CD Initiai-Final model due to the high insertion. The inter-syllable triphones performs better accuracy than

monophones and intra-syllable triphones. The PORM gives the highest accuracy. The performance of the CD Initial-Final model is better than that for phone modeling units and the CI Initial-Final model, and worse than those for either PORM or CORM.

Apart from the syllable accuracy, the recognition results of Initial-Final and onset-rhyme acoustic units were analyzed. Since the context-dependent Initials and the onsets of PORM and CORM are context-dependent, they give better accuracy than the context-independent Initial. Although the context-dependent Initial and the onset are the right context-dependent units, the onset outperforms the context-dependent Initial. The transitional portion towards the vowel, included in the onset, has contributed substantially to the precise modeling of the initial consonant segment. The accuracy of the onset in CORM is higher than the accuracy of the context-dependent Initial but lower than that of the PORM. At the other part of the syllable, the accuracies of the Final and the rhyme are slightly different. The Final and rhyme units give comparable accuracies due to the similar modeling of these units.

Obviously, the recognition results of the rhyme are very poor compared to the onset. The three stops, [p], [t], and [k], appearing at the final position are acoustically different from the initial consonant, that is, they are not audibly released. Most of the errors in the rhyme result from the substitution of the rhyme ending with these stop consonants. Furthermore, the rhyme without a final consonant may be recognized as the rhyme with a final consonant, and vice versa. This serious problem is hard to overcome by the acoustic model only.

Incorporating the language model can boost the performance of a speech recognition system. The recognition results of systems using both acoustic modeling and language modeling are improved ranging from 20.9 % to 30.8 % of syllable accuracy. The PORM attains the highest syllable accuracy at 75.2 % for male SD and 75.6 % for female SD systems. For the SI systems, the PORM also achieves the highest syllable accuracy at 62.8 % for male speakers and 64.8 % for female speakers.

In the language model-combined system, the accuracies of onsets of both PORM and CORM still exceed that for the context-dependent Initial, and the accuracy of the PORM onset is higher than that of the CORM onset. Using the language model, accuracy of the Initial and the onset is increased

by around 18-25 % and 16-21% for both SD and SI systems. In addition, the accuracies of the Final and the rhyme are substantially improved by nearly 43-46 % and 45-47 % for the SD and SI systems when the language model is applied.

The speech units were evaluated with two test sets. The recognition results of these two test sets are in the similar way that is the subsyllable units have a higher accuracy than the phone units. This verifies that the proposed acoustic unit, onset-rhyme, is superior in terms of accuracy.

Speech units used for Thai speech recognition evaluated in this research are onset-rhyme, Initial-Final, triphone, and monophone. The result of the evaluation is summarized in Table 6.1. The units are relatively compared based on three criteria, i.e., accuracy, generalization, and trainability, in using the units for the Thai continuous speech recognition system. The onset-rhyme models have a finite number of speech units that economically represent the all potential speech units of the language. Based on the experiments, the onset-rhyme models also satisfy all the major criteria in selection of good acoustic units. First, the onset-rhyme models are accurate in that each of the onset and rhyme units gives a high recognition performance. Second, the onset and rhyme units are reliably estimated with only a small set of training utterances that satisfy the trainability criterion. Finally, the onset-rhyme models are generalized in that the same onset and rhyme units have similar characteristics for different instances.

The Thai language was selected in this research, since it is a predominantly monosyllabic language with simple syllable structure, i.e., syllable with one or none final consonant. The languages with these two characteristics are spoken in mainland south-east Asia and China. They include languages in Tai Kadai (Thai, Lao, etc.) Mon-Khmer (Cambodian, Vietnamese, etc.), Hmong-Mian (Mao, Yao, etc.), and Sino-Tibetan (Chinese, Tibetian, etc.) language families. In addition, the Thai language also has a complex vowel system, i.e., contrast of short-long vowel pairs.

Considering accuracy, the onset-rhyme and the Initial-Final are better than the triphone and the monophone. However, the onset-rhyme is best in term of accuracy. According to the criteria of generalization, all the candidates have the generalization capability. When taking trainability into account, the worst one is the triphone. The best one is the monophone. However, the triphone and the monophone cannot be considered as good

candidates because the degrees of accuracy for both of them are low. The only two candidates left are the onset-rhyme and the Initial-Final. In term of trainability, they are very close, although the Initial-Final is a little bit better.

Therefore, applying the onset-rhyme model to Thai continuous speech recognition shows major improvements over other acoustic units in many ways, indicating better performance of the models while maintaining system simplicity. Therefore, these make the onset-rhyme models the efficient acoustic speech units for continuous speech recognition in the Thai language and other predominantly monosyllabic languages with simple syllable structure as well.

Table 6.1 Evaluation of various speech units
for Thai continuous speech recognition

Criteria	**Efficiency of speech units**			
	Onset-rhyme	**Initial-Final**	**Triphone**	**Monophone**
Accuracy	4	3	2	1
(considered by % recognition)	CORM (44.6 %) PORM (46.8 %)	CI IF (38.4 %) CD IF (42.6 %)	intra (34.8 %) inter (40.4 %)	19.6 %
Generalization	All speech units have the generalization capability			
Trainability	2	3	1	4
(considered by the no. of units trained)	CORM (497) PORM (992)	CI IF (233) CD IF (497)	intra (7,769) inter (64,475)	58

* 4 is the best score and 1 is the worst score in terms of comparison

6.2 Contributions of the Dissertation

This section summarized the contributions made during doing the research in this dissertation. The works begin from acoustic-phonetic analysis of the Thai language. Several tools were developed for analysis and utilization of text and speech corpus. The onset-rhyme model for Thai continuous speech recognition was emerged from this analysis. Hundreds of sentences were composed for voice recording. The speech corpus for dictation task was constructed according to these sentences. In addition, hundreds thousands of Thai syllables were collected, analyzed, and transcribed for building the language model. The details of the contributions are described as follows:

6.2.1 The Onset-Rhyme Acoustic Model

This dissertation conducted a basic research on acoustic modeling for Thai continuous speech recognition. From the acoustic-phonetic analysis of Thai syllables, the appropriate speech unit for the Thai language is the onset-rhyme model. An onset comprises an initial consonant and its transition towards the following vowel. Together with the onset, the rhyme consists of a steady vowel segment and a final consonant. Using the onset-rhyme model in Thai speech recognition shows major improvements over other acoustic units in many ways, indicating better performance of the model.

6.2.2 Thai Text Corpora and Thai Continuous Speech Corpora

This dissertation provides sets of text corpus used in speech recording for training and testing. The text corpus was designed for recording in reading or dictating style. In addition, the text corpus was created to cover all onsets and rhymes existing in the Royal Thai Dictionary. The set of sentences was carefully composed to contain samples of onset and rhyme adequately for creating the acoustic models. Therefore, a set of 1,081 sentences was created for training the acoustic model while a set of 100 sentences was used to evaluate the efficiency of the acoustic model.

The Thai continuous speech corpus was recorded from 14 male and 16 female speakers uttering in reading style. The total durations of the speech corpus used in this dissertation for training and testing are approximately 68 hours and 36 hours, respectively. Initially, the set of recorded speech was labeled according to the phonetic transcriptions. This process spends a lot of time to label manually. The labeled corpus was used to train the initial acoustic models for the automatic labeling system. The text and speech corpora constructed in this dissertation can be reliably used as reference corpora for further research in Thai speech recognition.

6.2.3 The Language Model

From the experimental results, the speech recognition system, employing the acoustic model only, gives unsatisfactory accuracy. Since the n-gram language model has been among the most successful approaches used for language modeling, particularly for speech recognition, this dissertation

explored the use of n-gram language model in Thai continuous speech recognition.

The n-gram language model can be trained from the text corpus ranging from millions of word to billions of words. The text corpus for building the n-gram model was excerpted from various kinds of reading paragraphs. Nearly millions of syllables were transcribed into phonetic representation. These phonetic transcriptions were then manually checked and rectified. This process consumes a lot of time. The n-gram language model was build from the correct transcriptions by CMU language modeling toolkit. After taking several steps in building the language model, the output from the toolkit is in the ARPA format. Finally, the ARPA format was converted to the lattice network in order to use the HTK decoding module. Experimental results show that the speech recognition system including the language model provides notable performance.

6.2.4 Program Development

Several tools have been developed for the research in Thai continuous speech recognition. The development begins from the analysis tool through the evaluation tool. Details of the tools contributed by this dissertation are listed as follows:

- **Speech analysis tool** – Since acoustic-phonetic analysis of the Thai language is the important part of this dissertation, the speech analysis program was developed in this analysis. This program clearly illustrates acoustical properties of speech signal in both time and frequency domain. Many parameters can be adjusted to study the dominant feature of speech signal. Development of this tool contributes much more understanding of characteristics of speech signal.
- **Thai text analysis tool** – To design and construct the Thai continuous speech corpora, the Thai text analysis tool was written to analyze the distribution of speech units from selected paragraphs. This tool assists the author to compose a set of sentences for building acoustic models with sufficient training samples.
- **Speech-labeling tool** – It is necessary to indicate the boundary of recorded sounds because the speech corpus associated with transcriptions was employed to create the initial acoustic models. The

speech-labeling tool was developed to mark the boundary of recorded speech. This tool displays essential information in graphic mode, and provides some easy use functions.

- **Evaluation tool** – The recognition results were shown in the specific format. It is necessary to analyze these results in form of easy interpretation. The original HTK program cannot provide a suitable tool for phonetic alignment. Then, the evaluation tools were written to handle the results with reliable evaluation.

These programs have contributed many utilities for a basic research in speech analysis and an advance research in speech recognition. They also provide some modules for further development.

6.3 Future Research on Thai Speech Recognition

All works in this dissertation used speech recorded in a quiet laboratory condition. Performance of the recognition systems degrades substantially when source of speech is in the noise and the distortion environment. To implement systems in practical applications, the speech recognizers must be more robust to the background noise and the distortion.

While the recognition system described in this dissertation gives high accuracy, there are many interesting points for further refinement of the acoustic models. Improvement in modeling accuracy has to taken account of variation in stress. Both stress and prosody form important cues for human speech recognition. It may be possible to study such features for acoustic modeling.

The speech corpus used in this dissertation was read from prepared texts. Recognition performance dramatically reduces for spontaneous or conversational speech. The fluent nature of such speech does not match either the acoustic or the language model. In particular, some forms of dialogue modeling are required to interact with the user to obtain clarification.

Due to the limited resources, the speech corpus was recorded from only 14 male and 16 female speakers. However, this corpus can be sufficiently verified the proposed acoustic models. To implement a practical speech recognition system, more speakers are needed to model variation from various speakers.

This dissertation incorporated the language model into a speech recognition system resulting in substantial improvement of the recognition results. The language model is directly estimated from text data. The large size of text corpus produces the accurate language model. Therefore, more text data should be collected and analyzed. Furthermore, various types of language models should be studied.

In this dissertation, the recognition results of the onset-rhyme models are the syllable without tone. Presently, the recognition of Thai complete syllable with tone is not complete. When the separated recognition of base syllable and tone is used, tone recognition system is required for producing a tone sequence. The additional module is needed to integrate and manipulate different sequences of base syllable and tone. On the other hand, using onset-rhyme as an acoustic unit provides the alternative of employing not only separated recognition but also joint recognition of base syllable and tone. Since prosodic information is preserved in rhyme portion, base syllable and tone can be simultaneously recognized to produce a complete tonal syllable. However, there are several advantages and drawbacks of these two schemes needed to study before they will be applied to Thai language.

REFERENCES

Abramson, A. S. The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. Doctoral dissertation, Faculty of Philosophy, Columbia University, 1960.

Ahkuputra, V. A Speaker Independent Thai Polysyllabic Word Recognition System using Hidden Markov Model. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1996.

Ahkuputra, V., Jitapunkul, S., Pornsukchantra, W., and Luksaneeyanawin, S. A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model, Proceedings of the 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, pp. 593-599, 1997.

Ahkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S., and Amornkul, P. Comparison of Different Techniques on Thai Speech Recognition, Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 177-180, 1998.

Ahkuputra, V. An Acoustic Study of syllable Onsets: A Basis for Thai Continuous Speech Recognition System. Doctoral dissertation, Department of Electrical Engineering, Chulalongkorn University, 2002.

Ahkuputra, V. Maneenoi, E., Luksaneeyanawin, S., and, Jitapunkul, S. Acoustic Modelling of Vowel Articulation on the Nine Thai Spreading Vowels. International Journal of Computer Processing of Oriental Languages 16, No. 3 (March-June 2003): 171-195.

Ali, A.M.A., Van der Spiegel, J., and Mueller, P. Robust Auditory-Based Speech Processing Using the Average Localized Synchrony Detection. IEEE Transactions on Speech and Audio Processing 10, No. 5 (July 2002): 279-292.

Anderson, O., Dalsgaard, P., and Barry, W. On the Use of Data-driven Clustering Technique for Identification of Poly- and Mono-phonemes for Four European Languages, Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing, pp. 121-124, 1991.

Areepongsa, S. Speaker Independent Thai Numeral Speech Recognition by Hidden Markov Model and Vector Quantization. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1995.

Aubert, X., Beyerlein, P., and Ullrich, M. A Bottom-up Approach for Handling Unseen Triphones in Large Vocabulary Continuous Speech Recognition, Proceedings of the 1996 International Conference on Spoken Language Processing, Vol. 1, pp. 14-17, 1996.

Austin, S., Schwartz, R., and Placeway, P. The Forward-Backward Search Algorithm, Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 697-700, 1991.

Bahl, L.R., Brown, P.F., de Souza, P.V., and Mercer, R.L. A Tree-based Statistical Language Model for Natural Language Speech Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 37, No. 7 (1989): 1001-1008.

Baker, J.K. Trainable Grammar for Speech Recognition, Proceeding of Conference of Acoustical Society of America, pp. 547-555, 1976.

Blasig, R. Combination of Words and Word Categories in Varigram Histories, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 529 -532, 1999.

Borden, G.J. and Harris, K.S. Speech Science Primer : Physiology, Acoustic, and Perception of Speech, Waverly Press Inc., Baltimore, Maryland, U.S.A., 1980.

Bu, L. and Church, T.D. Perceptual Speech Processing and Phonetic Feature Mapping for Robust Vowel Recognition. IEEE Transactions on Speech and Audio Processing 8, No. 2 (March 2000): 105-114.

Charnvivit, P., Lek-uthai, P., and Teangjai, S. Thai Speech Recognition System. Senior Project Report, Department of Electrical Engineering, Chulalongkorn University, 1999.

Chen, S. and Liao, Y. Modular Recurrent Neural Networks for Mandarin Syllable Recognition, IEEE Transactions on Neural Networks 9, No. 6 (November 1998): 1430-1441.

Chen, S. Liao, Y. Chiang, S., and Chang, S. An RNN-based Preclassification Method for Fast Continuous Mandarin Speech Recognition, IEEE Transactions on Speech and Audio Processing 6, No. 1 (January 1998): 86-90.

Chen, S.F., Seymore, K., and Rosenfeld, R. Topic Adaptation for Language Modeling Using Unnormalized Exponential Models, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 681-684, 1998.

Chen, S.F. and Rosenfeld, R. A Survey of Smoothing Techniques for ME Models. IEEE Transactions on Speech and Audio Processing 8, No. 1 (January 2000): 37-50.

Chen, C. J., Li, H.P., 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, Vol. 1, pp 61-64, 2001.

Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, G., Makhoul, J., Price, P., Roucos, S., and Schwartz, R. BYBLOS: The BBN continuous speech recognition system, Proceedings of the 1987 IEEE

International Conference on Acoustics, Speech, and Signal Processing, Vol. 12, pp. 89-92, 1987.

Clarkson, P. and Robinson, T. Improved Language Modelling through Better Language Model Evaluation Measures. Computer Speech & Language 15, Issue 1 (January 2001): 39-53.

Dagan, I., Marcus, S., and Markovitch, S. Contextual Word Similarity and Estimation from Sparse Data. Computer Speech & Language 9, Issue 2 (April 1995): 123-152.

Dautrich, B., Rabiner, L., and Martin, T. On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 31, No. 4 (August 1983): 793-807.

Deligne, S. and Sagisaka, Y. Statistical Language Modeling with a Class-Based n-multigram Model. Computer Speech & Language 14, Issue 3 (July 2000): 261-279.

Deller Jr., J., Proakis, J., and Hansen J. Discrete –Time Processing of Speech Signals. Macmillan Publishing Company, New York, 1993.

Denes, P.B. and Pinson, E.N. The Speech Chain, Bell Telephone Laboratories Inc. 1963.

Deng, L., Kenny, P., Lennig, M., and Mermelstein, P. Modeling Acoustic Transitions in Speech by State-interpolation Hidden Markov Models, IEEE Transactions on Acoustics, Speech, and Signal Processing 40, No. 2 (February 1992): 265-271.

Deshmukh, N., Ganapathiraju, A., and Picone, J. Hierarchical Search for Large-Vocabulary Conversational Speech Recognition: Working Toward A Solution to The Decoding Problem. IEEE Signal Processing Magazine, 16, No. 5 (September 1999): 84-107.

Duchateau, J. HMM-Based Acoustic Modeling in Large Vocabulary Speech Recognition. Doctoral dissertation, Katholieke Universiteit Leuven, 1998.

Fant, G. Acoustic Theory of Speech Production, Mouton & Co., Printers, Hague, The Netherlands. 1970.

Flanagan, J.L. Speech Analysis, Synthesis, and Perception, Springer-Verlag, 1972.

Fu, S.W.K., Lee, C.H., and Clubb, O.L. A Survey on Chinese Speech Recognition. International Journal of Chinese and Oriental Languages Information Processing 6, No. 1 (1996): 1-17.

Furui, S. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Transactions on Acoustics, Speech, and Signal Processing 34, No. 1 (January 1986): 52-59.

Furui, S. Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, Inc., New York, 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, No. 4 (2001): 358-366.

Gao, J. and Chen, X. Probabilistic Word Classification Based on a Context-Sensitive Binary Tree Method. Computer Speech & Language 11, Issue 4 (October 1997): 307-320.

Gauvain, J. and Lamel, L. Large Vocabulary Continuous Speech Recognition: Advances and Applications. Proceedings of the IEEE 88, No. 8 (August 1998): 1181-1200.

Gish, H. and Ng K., Parameter Trajectory Models for Speech Recognition, Proceedings of the 1996 IEEE International Conference on Speech and Language Processing, pp. 466-469, 1996.

Glass, J.R., Hazen, T.J., and Hetherington, I.L. Real-Time Telephone-Based Speech Recognition in the Jupiter Domain, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 61-64, 1999.

Hajime, T., Yamamoto, H., Takezawa, T., and Sagisaka, Y. Reliable Utterance Segment Recognition by Integrating A Grammar with Statistical Language Constraints. Speech Communication 26, Issue 4 (December 1998): 299-309.

Hochberg, M.M., Renals, S.J., Robinson, A.J., and, Kershaw D.J. Large Vocabulary Continuous Speech Recognition Using A Hybrid Connectionist-HMM System, Proceedings of the 1994 International Conference on Spoken Language Processing, pp. 1499-1502, 1994.

Huang, E.F., Wang, H.C., and Soong, F.K. A Fast Algorithm for Large Vocabulary Keyword Spotting Application. IEEE Transactions on Speech and Audio Processing 2, No. 3 (1994): 449-452.

Huang, X., Acero, A., and Hon, H.W. Spoken Language Processing, Prentice Hall PTR, New Jersey, U.S.A., 2001.

Hwang, M.Y. and Huang, X.D. Subphonetic Modeling for Speech Recognition, Proceedings DARPA Speech and Natural Language Workshop, New York, pp. 174-179, 1992.

Hwang, M.Y. and Huang, X.D. Shared-distribution Hidden Markov Models for Speech Recognition, IEEE Transactions on Speech and Audio Processing 1, No. 4 (1993): 414-420.

Iyer, R. and Ostendorf, M. Relevance Weighting for Combining Multi-domain Data for n-gram Language Modeling. Computer Speech & Language 13, Issue 3 (July 1999): 267-282.

Iyer, R., M. Ostendorf, and Rohlicek, J.R. Language Modeling with Sentence Level Mixtures, Proceeding of the ARPA Human and Language Technology Workshop, pp. 82-86, 1994.

Jardino, M. Multilingual Stochastic N-gram Class Language Models, Proceeding of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 161-163, 1996.

Jelinek, F. and Mercer, R. L. Interpolated Estimation of Markov Source Parameters from Sparse Data, Proceedings of the 1980 Workshop Pattern Recognition in Practice, pp. 381-397, 1980.

Jelinek, B., Zheng, F., Parihar, N., Hamaker, J., and Picone, J. Generalized Hierarchical Search in the ISIP ASR System, Proceedings of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers, Vol. 2, pp. 1553-1556, 2001.

Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Maneenoi, E., Kasuriya, S., and Amornkul, P. Recent Advances of Thai Speech Recognition in Thailand, Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 173-176, 1998.

Jitapunkul, S., Maneenoi, E., Ahkuputra, V., and Luksaneeyanawin, S. Performance Evaluation of Phonotactic and Contextual Onset-Rhyme Models for Speech Recognition of Thai Language, Proceedings of the 8th European Conference on Speech Communication and Technology Eurospeech 2003, Geneva, Switzerland, pp. 1841-1844, 2003.

Jittiwarangkul, N. Syllable Segmentation Algorithm for Thai Connected Speech. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1998.

Johnsen, M.H., A Sub-word Based Speaker Independent Speech Recognizer Using A Two-pass Segmentation Scheme, Proceedings of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 318-321, 1989.

Juang, B.H. and Furui, S. Automatic Recognition and Understanding of Spoken Language – A First Step toward Natural Human-Machine Communication. Proceedings of the IEEE 88, No. 8 (2000): 1142-1165.

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, No. 7 (July 1987): 947-954.

Junqua, J.C. ORION: A Two Pass Hybrid System for Isolated-words Automatic Speech Recognition, Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 41-44, 1990.

Junqua, J.C., Valente, S., Fohr, D., and Mari, J.F. An N-Best Strategy, Dynamic Grammars and Selectively Trained Neural Networks for Real-Time Recognition of Continuously Spelled Names over The Telephone,

Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 852-855, 1995.

Kalai, A., Chen, S., Blum, A., and Rosenfeld, R. On-line Algorithms for Combining Language Models, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 745-748, 1999.

Katz, S., Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35 Issue 3 (March 1987): 400-401.

Kenny, P., Labute, P., Li, Z., and O'Shaughnessy, D. New Graph Search Techniques for Speech Recognition, Proceedings of the 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 553-556, 1994.

Khudanpur, S. and Wu, J. A Maximum Entropy Language Model Integrating n-grams and Topic Dependencies for Conversational Speech Recognition, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 553-556, 1999.

Kneser, R. and Ney, H. Improved Backing-off for M-gram Language Modeling, Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 181-184, 1995.

Ladefoged, P. Elementary of Acoustic Phonetics, The University of Chicago Press. 1962.

Lee, K.F. and Hon, H.W., Speaker-independent Phone Recognition Using Hidden Markov Models. IEEE Transactions on Acoustics, Speech and Signal Processing 37, No.11 (1989): 1641-1648.

Lee, K.F., Hon, H.W., Hwang, M.Y., Mahajan, S., and Reddy, R. The SPHINX Speech Recognition System, Proceedings of the 1989 IEEE International Conference on, Acoustics, Speech, and Signal Processing, pp. 445-448, 1989.

Lee, K.F., Hon, H.W., and Reddy, R. An Overview of the SPHINX Speech Recognition System. IEEE Transactions on Acoustics, Speech, and Signal Processing 38, No. 1 (January 1990): 35-45.

Lee, K.F. Context-dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 38, No. 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, No. 2 (1993): 158-179.

Lee, L.S. Voice Dictation of Mandarin Chinese, IEEE Processing Magazine, 14, No. 4 (1997): 63-101.

Leelasiriwong, W. A Study of Acoustic Characteristics of the Vowels /i,a,u/ in Thai and Its Use in Speaker Identification. Master's thesis, Department of Physics, Chulalongkorn University, 1991.

Lleida, E. and Rose, R.C., Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures. IEEE Transactions on Speech and Audio Processing 8, No. 2 (March 2000): 126-139.

Loizou, P.C. Robust Speaker-Independent Recognition of A Confusable Vocabulary. Doctoral dissertation, Arizona State University, 1995.

Luk, R.W.P. and Damper, R.I., Computational Complexity of A Fast Viterbi Decoding Algorithm for Stochastic Letter-Phoneme Transduction. IEEE Transactions on Speech and Audio Processing 6, No. 3 (May 1998): 217–225.

Luksaneeyanawin, S. A Three Dimensional Phonology: A Historical Implication, Proceedings of the 3rd International Symposium on Language and Linguistics, pp. 75-90, 1992.

Luksaneeyanawin, S. Linguistics Research and Thai Speech Technology, Proceedings of the 5th International Conference on Thai Studies. School of Oriental and African Studies, University of London, pp. 1-29, 1993.

Mahajan, M., Beeferman, D., and Huang, X.D. Improved Topic-Dependent Language Modeling Using Information Retrieval Technique, Proceeding of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 157-160, 1992.

Maneenoi, E., Jitapunkul, S., Ahkuputra, V., and Wutiwiwatchai, C. Modification of BP Algorrithm for Thai Speech Recognition, Proceedings of the 1997 International Symposium on Natural Language Processing: (NLPRS'97), Phuket, Thailand, pp. 287-294, 1997.

Maneenoi, E., Jitapunkul, S., Ahkuputra, V., and Wutiwiwatchai, C. Modification of BP Algorithm for Thai Speech Recognition, Proceedings of 20th Electrical Engineering Conference, Bangkok, Thailand, pp. 343-348, 1997.

Maneenoi, E. Thai Vowel Phoneme Recognition using Artificial Neural Networks. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1998.

Maneenoi, E., Jitapunkul, S., Ahkuputra, V., Thathong, U., and Thampanitchawong, B. Thai Vowel Phoneme Recognition Using Neural Network, Proceedings of 22th Electrical Engineering Conference, Bangkok, Thailand, pp. 493-496, 1999.

Maneenoi, E., Jitapunkul, S., Ahkuputra, V., Thathong, U., Thampanitchawong, B., and Laksaneeyanawin, S., Thai Monophthongs Recognition Using Continuous Density Hidden Markov Model and LPC Cepstral Coefficients, Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 2000.

Maneenoi, E., Jitapunkul, S., and Ahkuputra, V., Thai Monophthongs Classification Using CDHMM, The Journal of the Acoustical Society of America, Vol.108, No.5, Pt.2 of 2, November 2000, pp.2575. [Abstract: 3pSC7] (The Joint Meeting: 140th meeting of the Acoustical Society of America and NOISE-CON 2000, California, U.S.A., December, 3-8, 2000.)

Maneenoi, E., Ahkuputra, V., Luksaneeyanawin, S., and Jitapunkul, S. Acoustic Modeling of Onset-Rhyme for Thai Continuous Speech Recognition, Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, Australia, pp. 462-467, 2002.

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, No. 5, 2004.

Markel, J.D. and Gray Jr. A.H., Linear Prediction of Speech, 1980.

Martin, S.C., Ney, H., and Hamacher, C., Maximum Entropy Language Modeling and the Smoothing Problem. IEEE Transactions on Speech and Audio Processing 8, No. 5 (September 2000): 626-632.

Martin, S.C., Ney, H., and Zaplo, J. Smoothing Methods in Maximum Entropy Language Modeling, Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 545-548, 1999.

Matsunaga, S. and Sakamoto, H. Two-pass Strategy for Continuous Speech Recognition with Detection and Transcription of Unknown Words, Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 538-541, 1996.

Mohri, M., Pereira, F., and Riley, M. Weighted Finite-State Transducers in Speech Recognition. Computer Speech & Language 16, Issue 1 (January 2002): 69-88.

Mokbel, C.E. and Chollet, G.F.A. Automatic Word Recognition in Cars. IEEE Transactions on Speech and Audio Processing 3, No. 5 (September 1995): 346-356.

Nadas, A., Nahamoo, D., Picheny, M.A., and Powell, J. An Iterative 'Flip-Flop' Approximation of the Most Informative Split in the Construction of Decision Trees, Proceedings of the 1991 IEEE International

Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 565-568, 1991.

Ney, H. and Essen, U. On Smoothing Techniques for Bigram-based Natural Language Modelling, Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 825-828, 1991.

Ney, H., Essen, U., and Kneser, R. On Structuring Probabilistic Dependences in Stochastic Language Modelling. Computer Speech & Language 8, Issue 1 (January 1994): 1-38.

Ng, C., Wilkinson, R., and Zobel, J. Experiments in Spoken Document Retrieval Using Phoneme n-grams. Speech Communication 32, Issues 1-2 (September 2000): 61-77.

Niesler, T.R. and Woodland, P.C. Variable-length Category n-gram Language Models. Computer Speech & Language 13, Issue 1 (January 1999): 99-124.

O'Boyle, P., Owens, M., and Smith, F. J. A Weighted Average n-gram Model of Natural Language. Computer Speech & Language 8, Issue 4 (October 1994): 337-349.

Odell, J., Woodland, P., and Young, S. Tree-based State Clustering for Large Vocabulary Speech Recognition, Proceedings of the 1994 International Symposium on Speech, Image Processing and Neural Networks, Vol. 2, pp. 690-693, 1994.

Odell, J. The Use of Context in Large Vocabulary Speech Recognition. Doctoral dissertation, Cambridge University, 1995.

Ortmanns, S. and Ney, H. The Time-conditioned Approach in Dynamic Programming Search for LVCSR. IEEE Transactions on Speech and Audio Processing 8, No. 6 (November 2000): 676-687.

Palmer, D., Ostendorf, M., and Burger, J. Robust Information Extraction from Automatically Generated Speech Transcriptions. Speech Communication 32, Issues 1-2 (September 2000): 95-109.

Pensiri, R. Speaker-Independent Thai Numeral Voice Recognition by using Dynamic Time Warping. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1995.

Phatrapornnant, T. Speaker-Independent Isolated Thai Spoken Vowel Recognition by using Spectrum Distance Measurement and Dynamic Time Warping. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1995.

Picone, J. A. Review of Large Vocabulary Continuous Speech Recognition. IEEE Signal Processing Magazine 13 September (1996): 45-57.

Popovici, C. and Baggia, P. Specialized Language Models Using Dialogue Predictions, Proceedings of the 1997 IEEE International Conference

on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 815-818, 1997.

Pornsukjantra, W. Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1996.

Pornsukchantra, W., Jitapunkul, S., and Ahkuputra, V. Speaker-Independent Thai Numeral Speech Recognition Using LPC and the Back Propagation Neural Network, Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 NLPRS'97, Phuket, pp. 585-588, 1997.

Potamianos, G. and Jelinek, F. A Study of n-gram and Decision Tree Letter Language Modeling Methods. Speech Communication 24, Issue 3 (June 1998): 171-192.

Prathumthan, T. Thai Speech Recognition using Syllable Units. Master's thesis, Department of Computer Engineering, Chulalongkorn University, 1986.

Rabiner, L.R. and Schafer, R.W. Digital Processing of Speech Signals. Prentice-Hall Inc. 1978.

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, No. 8 (1989): 1214-1224.

Rabiner, L.R. and Juang, B.H. Fundamentals of Speech Recognition. Prentice-Hall Inc. 1993.

Ravishankar, M.K. Efficient Algorithm for Speech Recognition. Doctoral dissertation, School of Computer Science, Carnegie Mellon University, 1996.

Renals, S. and Hochberg, M. Efficient Search Using Posterior Phone Probability Estimates, Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 596 –599, 1995.

Riccardi, G., Pieraccini, R., and Bocchieri, E. Stochastic Automata for Language Modeling. Computer Speech & Language 10, Issue 4 (October 1996): 265-293.

Richardson, F., Ostendorf, M., and Rohlicek, J.R. Lattice-Based Search Strategies for Large Vocabulary Speech Recognition, Proceedings pg the 1995 International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 576–579, 1995.

Robinson, A. J. Connectionist speech recognition of Broadcast News. Speech Communication 37, Issues 1-2 (2002): 27-45.

Rosenfeld, R. A Maximum Entropy Approach to Adaptive Statistical Language Modelling. Computer Speech & Language 10, Issue 3 (July 1996): 187-228.

Rosenfeld, R. A Whole Sentence Maximum Entropy Language Model, Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 230-237, 1997.

Rosenfeld, R. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. Doctoral dissertation, School of Computer Science, Carnegie Mellon University, 1994.

Sato, T., Ghulam, M., Fukuda, T., and Nitta, T. Confidence Scoring for Accurate HMM-Based Word Recognition by Using SM-Based Monophone Score Normalization, Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 217-220, 2002.

Shannon, C.E. A Mathematic Theory of Communication, Bell Systems Technical Journal 27 (1948): 379-423.

Sleator, D. and Temperly, D. Parsing English with a Link Grammar, Technical Report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University, 1991.

Sriraksa, U. Acoustic Characteristics Signaling Syllable Boundary in Thai Connected Speech. Master's thesis, Department of Linguistics, Chulalongkorn University, 1995.

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.

Stevens, K.N. Acoustic Phonetics, The MIT Press, Cambridge, Massachusetts. 1999.

Thumpothong, P. Multispeaker Speech Recognition System. Master's thesis, Department of Computer Engineering, Chulalongkorn University, 1989.

Tharnsakun, W. The Acoustic Analysis of Thai Stops. Master's thesis, Department of Linguistics, Chulalongkorn University, 1988.

Thubthong, N. A Thai Speech Recognition System based on Phonemic Distinctive Features. Master's thesis, Department of Computer Engineering, Chulalongkorn University, 1995.

Thubthong, N., Kijsirikul, B., and Luksaneeyanawin, S. An Empirical Study for Constructing Thai Tone Models, Proceedings of SNLP-Oriental COCOSDA 2002, pp. 179-186, 2002.

Tokhura, Y., A Weighted Cepstral Distance Measure for Speech Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 35, No. 10 (October 1987): 1414-1422.

Tran, B.H., Seide, F., and Steinbiss, T. A Word Graph Based N-Best Search in Continuous Speech Recognition, Proceedings of the Fourth International Conference on Spoken Language, Vol. 4, pp. 2127-2130, 1996.

Trongdee, T. The Acoustic Analysis of Thai Non-Stops. Master's thesis, Department of Linguistics, Chulalongkorn University, 1987.

Udompisit, L. and Sothipunchai, S. Voice Activated Web Browser. Senior Project Report, Department of Electrical Engineering, Chulalongkorn University, 2000.

Ueberla, J.P. Analysing Weaknesses of Language Models for Speech Recognition, Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 205 –208, 1995.

Valtchev V. Discriminative Methods in HMM-based Speech Recognition. Doctoral dissertation, University of Cambridge, 1995.

Vergin, R., O'Shaughnessy, D., and Farhat, A. Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-independent Continuous-Speech Recognition. IEEE Transactions on Speech and Audio Processing 7, No. 5 (September 1999): 525-532.

Vuuren, S.V. Pitch Estimation. Technical Report No. CSE-98-05, Antropic Speech Processing Group, Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology, February 1998.

Wakita, Y., Singer, H., and Sagisaka, Y. Multiple Pronunciation Dictionary Using HMM-State Confusion Characteristics. Computer Speech & Language 13, Issue 2 (April 1999): 143-153.

Wakita, Y.; Kawai, J.; Iida, I. An Evaluation of Statistical Language Modeling for Speech Recognition Using a Mixed Category of Both Words and Parts-Of-Speech, Proceedings of the Fourth International Conference on Spoken Language, Vol. 1, pp. 530-533, 1996.

Wang, H.M., Ho, T.H., Yang, R.C., Shen, J.L., Bai, B.R., Hong, J.C., Chen, W.P., Yu, T.L., and Lee, L.S. Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data. IEEE Transaction on Speech and Audio Processing 5, No. 2 (1997): 195-200.

Wang, S., Rosenfeld, R., and Zhao, Y. Latent Maximum Entropy Principle for Statistical Language Modeling, Proceedings of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 182-185, 2001.

Ward, W. and Issar, S. A Class Based Language Model for Speech Recognition, Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 416-418, 1996.

Whittaker, E.W.D. and Woodland, R.C. Efficient Class-based Language Modelling for Very Large Vocabularies, Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 545-548, 2001.

Wilcox, L.D. and Bush, M.A. Training and Search Algorithms for An Interactive Wordspotting System, Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 97-100, 1992.

Witten, I.H. and Bell, T.C. The Zero-frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Transactions on Information Theory 37 Issue 4 (July 1991): 1085-1094.

Woodland, P.C., Odell, J.J., Valtchev, V., and Young, S.J. Large Vocabulary Continuous Speech Recognition using HTK, Proceedings of the 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 125-128, 1994.

Wu, J. and Khudanpur, S. Building a Topic-dependent Maximum Entropy Model for Very Large Corpora, Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 777-780, 2002.

Wutiwiwatchai, C. Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique. Master's thesis, Department of Electrical Engineering, Chulalongkorn University, 1997.

Wutiwiwatchai, C., Jitapunkul, S., Ahkuputra, V., Maneenoi, E., Amornkul, P., and Luksaneeyanawin, S. A New Strategy of Fuzzy-Neural Network for Thai Numeral Speech Recognition, Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems, pp. 157-160, 1998.

Yokoyama, T., Shinozaki, T., Iwano, K., and Furui, S. Unsupervised Class-Based Language Model Adaptation for Spontaneous Speech Recognition, Acoustics, Speech, and Signal Processing, Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 236-239, 2003.

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, pp. 569-572, 1992.

Young, S.J., Odell, J.J., and Woodland, P.C. Tree-based Tying for High Accuracy Acoustic Modelling, Proceedings ARPA Workshop on Human Language Technology, pp. 286-291, 1994.

Young, S.J. and Woodland, P.C. State Clustering in Hidden Markov Model-based Continuous Speech Recognition. Computer Speech and Language 8 (1994): 369-383.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. The HTK Book for HTK Version 3.0, Microsoft Corporation, 1999.

Zhang, R., Black, E., Finch, A., and Sagisaka, Y. Integrating Detailed Information into a Language Model, Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1595 -1598, 2000.

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, Vol. 1, pp. 389-392, 1989.

APPENDICES

APPENDIX A

The Thai Text Training Corpus

Table A1.1 Statistic of the Thai initial consonants in the training corpus

Unit	Amount	Percent
m	1,799	8.095 %
n	1,784	8.027 %
th	1,645	7.402 %
kh	1,584	7.127 %
s	1,401	6.304 %
l	1,368	6.156 %
k	1,289	5.800 %
c	1,248	5.616 %
t	1,211	5.449 %
r	1,174	5.283 %
d	1,019	4.585 %
j	962	4.329 %
ph	908	4.086 %
p	887	3.991 %
ch	696	3.132 %
h	695	3.127 %
z	674	3.033 %
w	668	3.006 %
b	557	2.506 %
f	210	0.945 %
ng	210	0.945 %

Table A1.2 Statistic of the Thai final consonants in the training corpus

Unit	Amount	Percent
n	3,882	23.140 %
ng	3,273	19.510 %
j	3,155	18.807 %
k	1,571	9.365 %
m	1,561	9.305 %
w	1,260	7.511 %
t	1,156	6.891 %
p	918	5.472 %

Table A1.3 Statistic of the Thai consonant clusters in the training corpus

Unit	Amount	Percent
kr	235	13.048 %
kl	204	11.327 %
pr	204	11.327 %
khr	195	10.827 %
phr	189	10.494 %
pl	167	9.273 %
khw	153	8.495 %
tr	132	7.329 %
phl	123	6.830 %
khl	120	6.663 %
kw	68	3.776 %
thr	11	0.611 %

Table A1.4 Statistic of the Thai vowels in the training corpus

Unit	Amount	Percent
a	6,141	25.809 %
aa	4,970	20.888 %
@@	1,650	6.935 %
ii	1,347	5.661 %
o	1,238	5.203 %
i	858	3.606 %
uu	817	3.434 %
xx	784	3.295 %
uua	733	3.081 %
e	716	3.009 %
vva	600	2.522 %
u	594	2.496 %
iia	524	2.202 %
v	505	2.122 %
@	428	1.799 %
x	421	1.769 %
qq	413	1.736 %
oo	373	1.568 %
vv	339	1.425 %
ee	272	1.143 %
q	58	0.244 %
ia	4	0.017 %
ua	3	0.013 %
va	2	0.008 %

Table A1.5 Statistic of the context-independent Initials in the training corpus

Unit	Amount	Percent
m	1,799	7.562 %
n	1,784	7.499 %
th	1,645	6.915 %
kh	1,584	6.658 %
s	1,401	5.889 %
l	1,368	5.750 %
k	1,289	5.418 %
c	1,248	5.246 %
t	1,211	5.090 %
r	1,174	4.935 %
d	1,019	4.283 %
j	962	4.044 %
ph	908	3.817 %
p	887	3.728 %
ch	696	2.926 %
h	695	2.921 %
z	674	2.833 %
w	668	2.808 %
b	557	2.341 %
kr	235	0.988 %
f	210	0.883 %
ng	210	0.883 %
kl	204	0.858 %
pr	204	0.858 %
khr	195	0.820 %
phr	189	0.794 %
pl	167	0.702 %
khw	153	0.643 %
tr	132	0.555 %
phl	123	0.517 %
khl	120	0.504 %
kw	68	0.286 %
thr	11	0.046 %

Table A1.6 Statistic of the context-dependent Initials and CORM onsets in the training corpus

Unit	Amount	Percent	Unit	Amount	Percent
n_a	1097	4.611 %	ph_v	137	0.576 %
m_a	1037	4.359 %	pr_a	135	0.567 %
c_a	689	2.896 %	f_a	134	0.563 %
kh_a	678	2.850 %	l_u	134	0.563 %
th_a	626	2.631 %	s_v	134	0.563 %
k_a	608	2.556 %	t_o	134	0.563 %
th_i	551	2.316 %	k_i	128	0.538 %
s_a	533	2.240 %	s_o	124	0.521 %
w_a	502	2.110 %	kh_v	121	0.509 %
p_a	500	2.102 %	kh_u	120	0.504 %
j_a	451	1.896 %	s_@	120	0.504 %
r_a	449	1.887 %	r_@	119	0.500 %
h_a	427	1.795 %	n_v	118	0.496 %
ch_a	408	1.715 %	kl_a	117	0.492 %
t_a	401	1.686 %	r_v	111	0.467 %
d_a	396	1.665 %	ph_@	109	0.458 %
l_a	392	1.648 %	z_i	108	0.454 %
l_x	373	1.568 %	ng_a	107	0.450 %
k_@	300	1.261 %	n_@	103	0.433 %
z_a	286	1.202 %	khr_a	102	0.429 %
n_i	270	1.135 %	m_x	101	0.425 %
kh_@	269	1.131 %	c_i	97	0.408 %
m_i	257	1.080 %	phr_@	97	0.408 %
p_e	251	1.055 %	l_e	95	0.399 %
ph_a	240	1.009 %	z_@	94	0.395 %
s_i	237	0.996 %	b_o	93	0.391 %
b_a	235	0.988 %	ph_i	89	0.374 %
kh_o	224	0.942 %	th_v	89	0.374 %
j_u	213	0.895 %	r_o	87	0.366 %
th_u	204	0.858 %	c_@	80	0.336 %
t_@	199	0.836 %	w_i	80	0.336 %
d_u	193	0.811 %	n_o	77	0.324 %
t_u	190	0.799 %	k_x	76	0.319 %
s_u	187	0.786 %	ch_i	70	0.294 %
kr_a	182	0.765 %	kh_i	70	0.294 %
r_u	168	0.706 %	d_o	68	0.286 %
ph_u	162	0.681 %	d_q	68	0.286 %
d_i	160	0.673 %	n_u	68	0.286 %
t_x	160	0.673 %	t_i	68	0.286 %
c_v	154	0.647 %	h_e	67	0.282 %
m_v	154	0.647 %	b_i	66	0.277 %
l_o	149	0.626 %	l_v	65	0.273 %
r_i	143	0.601 %	l_q	64	0.269 %
khw_a	142	0.597 %	m_@	64	0.269 %
c_o	139	0.584 %	b_@	63	0.265 %

Unit	Amount	Percent
j_i	63	0.265 %
kw_a	62	0.261 %
j_@	61	0.256 %
ph_o	61	0.256 %
j_v	60	0.252 %
m_o	60	0.252 %
pl_@	59	0.248 %
ch_v	58	0.244 %
th_@	56	0.235 %
m_u	55	0.231 %
m_e	54	0.227 %
d_e	53	0.223 %
j_o	53	0.223 %
z_u	51	0.214 %
k_e	50	0.210 %
l_@	50	0.210 %
ch_@	49	0.206 %
p_i	48	0.202 %
h_x	46	0.193 %
l_i	46	0.193 %
th_x	46	0.193 %
ph_e	45	0.189 %
phr_a	45	0.189 %
z_o	45	0.189 %
s_e	43	0.181 %
k_o	42	0.177 %
kh_x	42	0.177 %
h_u	41	0.172 %
kl_u	41	0.172 %
w_e	41	0.172 %
d_v	40	0.168 %
kh_e	40	0.168 %
r_x	40	0.168 %
k_q	39	0.164 %
b_u	38	0.160 %
h_@	38	0.160 %
khr_u	38	0.160 %
ph_x	38	0.160 %
n_x	37	0.156 %
ch_u	36	0.151 %
pl_a	36	0.151 %
t_v	36	0.151 %
b_x	35	0.147 %
tr_i	35	0.147 %
z_e	35	0.147 %
h_o	34	0.143 %
k_u	34	0.143 %
khl_a	34	0.143 %

Unit	Amount	Percent
c_u	33	0.139 %
j_e	32	0.135 %
phl_a	32	0.135 %
ch_o	31	0.130 %
z_v	31	0.130 %
ng_q	29	0.122 %
r_e	29	0.122 %
th_q	29	0.122 %
r_q	28	0.118 %
tr_u	28	0.118 %
h_i	27	0.113 %
ph_q	27	0.113 %
tr_o	27	0.113 %
c_e	26	0.109 %
khr_v	25	0.105 %
ng_o	24	0.101 %
p_u	24	0.101 %
phl_x	24	0.101 %
d_@	23	0.097 %
c_q	22	0.092 %
j_x	22	0.092 %
khl_v	22	0.092 %
pl_i	22	0.092 %
th_e	22	0.092 %
th_o	22	0.092 %
pr_i	21	0.088 %
ch_e	20	0.084 %
kh_q	20	0.084 %
phl_q	20	0.084 %
pr_@	20	0.084 %
p_@	19	0.080 %
phr_i	19	0.080 %
t_e	19	0.080 %
d_x	18	0.076 %
tr_a	18	0.076 %
z_q	18	0.076 %
f_o	17	0.071 %
f_u	17	0.071 %
m_q	17	0.071 %
pl_x	17	0.071 %
w_o	17	0.071 %
ng_u	16	0.067 %
s_x	16	0.067 %
khl_@	15	0.063 %
p_v	15	0.063 %
ch_q	14	0.059 %
f_v	14	0.059 %
khl_u	14	0.059 %

Unit	Amount	Percent
ng_@	14	0.059 %
pl_o	13	0.055 %
k_v	12	0.050 %
khl_o	12	0.050 %
kl_i	12	0.050 %
kr_u	12	0.050 %
p_o	12	0.050 %
phl_o	12	0.050 %
phl_u	12	0.050 %
b_v	11	0.046 %
kr_o	11	0.046 %
p_x	11	0.046 %
pr_o	11	0.046 %
ch_x	10	0.042 %
kl_@	10	0.042 %
b_q	9	0.038 %
khr_@	9	0.038 %
phl_e	9	0.038 %
c_x	8	0.034 %
h_v	8	0.034 %
khl_e	8	0.034 %
n_q	8	0.034 %
phl_i	8	0.034 %
phr_o	8	0.034 %
phr_x	8	0.034 %
pl_v	8	0.034 %
tr_x	8	0.034 %
w_@	8	0.034 %
w_x	8	0.034 %
b_e	7	0.029 %
f_i	7	0.029 %
f_x	7	0.029 %
h_q	7	0.029 %
j_q	7	0.029 %
kr_@	7	0.029 %
kr_i	7	0.029 %
kr_x	7	0.029 %
p_q	7	0.029 %
pl_u	7	0.029 %
s_q	7	0.029 %
tr_@	7	0.029 %
f_q	6	0.025%
khl_x	6	0.025 %
khr_e	6	0.025 %
khr_o	6	0.025 %

Unit	Amount	Percent
khw_x	6	0.025 %
kl_o	6	0.025 %
kl_x	6	0.025 %
n_e	6	0.025 %
ng_e	6	0.025 %
ng_i	6	0.025 %
tr_e	6	0.025 %
z_x	6	0.025 %
f_@	5	0.021 %
khl_i	5	0.021 %
khr_x	5	0.021 %
kr_q	5	0.021 %
phr_v	5	0.021 %
pl_e	5	0.021 %
w_u	5	0.021 %
khl_q	4	0.017 %
khr_i	4	0.017 %
kl_e	4	0.017 %
kl_q	4	0.017 %
kl_v	4	0.017 %
kr_e	4	0.017 %
ng_v	4	0.017 %
ng_x	4	0.017 %
phr_u	4	0.017 %
pr_u	4	0.017 %
pr_x	4	0.017 %
t_q	4	0.017 %
thr_a	4	0.017 %
thr_i	4	0.017 %
w_v	4	0.017 %
f_e	3	0.013 %
khw_i	3	0.013 %
kw_i	3	0.013 %
kw_x	3	0.013 %
phl_@	3	0.013 %
phl_v	3	0.013 %
phr_q	3	0.013 %
pr_e	3	0.013 %
pr_q	3	0.013 %
pr_v	3	0.013 %
thr_@	3	0.013 %
tr_v	3	0.013 %
w_q	3	0.013 %
khw_e	2	0.008 %

Table A1.7 Statistic of the PORM onsets in the training corpus

Unit	Amount	Percent	Unit	Amount	Percent
n_a	674	2.833%	l_o	126	0.530%
m_a	546	2.295%	m_vva	123	0.517%
m_aa	492	2.068%	c_o	121	0.509%
th_ii	477	2.005%	pr_a	113	0.475%
th_a	425	1.786%	s_o	113	0.475%
n_aa	424	1.782%	ph_uu	111	0.467%
c_a	420	1.765%	s_@@	111	0.467%
kh_a	417	1.753%	d_a	109	0.458%
p_a	398	1.673%	t_@@	109	0.458%
k_a	392	1.648%	ph_a	108	0.454%
s_a	359	1.509%	j_a	107	0.450%
j_aa	348	1.463%	ph_vva	106	0.446%
k_@@	287	1.206%	r_@@	102	0.429%
d_aa	284	1.194%	t_o	102	0.429%
w_aa	273	1.148%	p_aa	101	0.425%
c_aa	269	1.131%	k_i	98	0.412%
kh_@@	268	1.127%	d_ii	97	0.408%
r_a	267	1.122%	d_uua	96	0.404%
kh_aa	263	1.106%	m_xx	96	0.404%
h_a	255	1.072%	ph_@@	95	0.399%
t_aa	253	1.063%	r_uu	94	0.395%
p_e	249	1.047%	s_iia	94	0.395%
m_ii	238	1.000%	b_a	93	0.391%
l_a	236	0.992%	n_v	91	0.383%
w_a	230	0.967%	s_u	91	0.383%
kh_o	221	0.929%	t_@	91	0.383%
l_x	221	0.929%	n_@@	90	0.378%
k_aa	216	0.908%	ng_aa	88	0.370%
ch_a	212	0.891%	kh_v	87	0.366%
th_aa	201	0.845%	f_a	85	0.357%
ch_aa	197	0.828%	d_uu	84	0.353%
r_aa	182	0.765%	n_i	83	0.349%
kr_a	180	0.757%	th_uu	82	0.345%
n_ii	175	0.736%	z_@@	81	0.340%
s_aa	175	0.736%	kh_u	79	0.332%
j_uu	173	0.727%	khr_a	79	0.332%
h_aa	172	0.723%	r_iia	79	0.332%
z_aa	157	0.660%	r_vva	77	0.324%
l_aa	156	0.656%	b_o	76	0.319%
c_v	154	0.647%	s_ii	75	0.315%
l_xx	153	0.643%	w_i	75	0.315%
t_uua	151	0.635%	n_o	74	0.311%
t_a	148	0.622%	c_@@	72	0.303%
b_aa	142	0.597%	c_i	72	0.303%
t_xx	141	0.593%	l_uu	70	0.294%
ph_aa	133	0.559%	phr_@	70	0.294%
z_a	129	0.542%	th_u	69	0.290%
khw_aa	127	0.534%	d_qq	67	0.282%

Unit	Amount	Percent
s_i	67	0.282%
th_v	65	0.273%
k_xx	63	0.265%
kw_aa	62	0.261%
l_e	62	0.261%
m_@@	62	0.261%
j_i	60	0.252%
s_uua	60	0.252%
kl_a	58	0.244%
kl_aa	58	0.244%
s_vv	56	0.235%
l_qq	55	0.231%
z_ii	54	0.227%
b_@@	53	0.223%
d_e	53	0.223%
th_uua	53	0.223%
h_e	51	0.214%
ph_o	51	0.214%
th_@@	51	0.214%
m_e	50	0.210%
d_oo	49	0.206%
m_o	49	0.206%
s_vva	49	0.206%
f_aa	48	0.202%
l_u	48	0.202%
r_o	48	0.202%
r_uua	48	0.202%
l_@@	43	0.181%
r_ii	43	0.181%
ph_i	42	0.177%
l_vva	41	0.172%
t_i	40	0.168%
k_qq	39	0.164%
r_oo	39	0.164%
j_u	38	0.160%
r_xx	38	0.160%
th_i	38	0.160%
ph_uua	36	0.151%
pl_aa	36	0.151%
s_uu	36	0.151%
th_iia	36	0.151%
ph_e	35	0.147%
w_ee	35	0.147%
ch_iia	34	0.143%
ch_vva	34	0.143%
b_iia	33	0.139%
ch_uua	33	0.139%
d_i	33	0.139%

Unit	Amount	Percent
j_@	33	0.139%
l_ee	33	0.139%
kl_uua	32	0.135%
pl_@@	32	0.135%
t_oo	32	0.135%
h_uua	31	0.130%
j_vv	31	0.130%
k_e	31	0.130%
kh_ii	31	0.130%
kh_x	31	0.130%
m_vv	31	0.130%
z_ee	31	0.130%
ch_@@	30	0.126%
d_iia	30	0.126%
j_o	30	0.126%
z_iia	30	0.126%
ph_iia	29	0.122%
s_v	29	0.122%
z_oo	29	0.122%
h_xx	28	0.118%
j_@@	28	0.118%
n_u	28	0.118%
r_e	28	0.118%
r_qq	28	0.118%
r_vv	28	0.118%
t_u	28	0.118%
z_vv	28	0.118%
b_i	27	0.113%
j_vva	27	0.113%
n_vva	27	0.113%
ph_qq	27	0.113%
phr_@@	27	0.113%
pl_@	27	0.113%
tr_o	27	0.113%
b_u	26	0.109%
kh_vv	26	0.109%
r_u	26	0.109%
ch_vv	25	0.105%
j_e	25	0.105%
l_i	25	0.105%
m_uu	25	0.105%
phl_xx	25	0.105%
tr_ii	25	0.105%
z_uua	25	0.105%
c_e	24	0.101%
d_@@	24	0.101%
d_vva	24	0.101%
h_@	24	0.101%

Unit	Amount	Percent
kh_ee	24	0.101%
n_xx	24	0.101%
phr_a	24	0.101%
th_vv	24	0.101%
z_i	24	0.101%
j_oo	23	0.097%
k_o	23	0.097%
kh_uua	23	0.097%
khr_aa	23	0.097%
l_oo	23	0.097%
pr_aa	23	0.097%
th_x	23	0.097%
th_xx	23	0.097%
ng_q	22	0.092%
th_ee	22	0.092%
z_u	22	0.092%
ch_o	21	0.088%
h_o	21	0.088%
h_x	21	0.088%
phl_qq	21	0.088%
phr_aa	21	0.088%
r_i	21	0.088%
s_e	21	0.088%
s_ee	21	0.088%
b_xx	20	0.084%
ch_e	20	0.084%
ch_ii	20	0.084%
kh_i	20	0.084%
kh_qq	20	0.084%
khl_a	20	0.084%
khr_uu	20	0.084%
n_uu	20	0.084%
n_uua	20	0.084%
p_ii	20	0.084%
ph_xx	20	0.084%
ch_@	19	0.080%
d_o	19	0.080%
h_i	19	0.080%
j_xx	19	0.080%
k_ee	19	0.080%
k_oo	19	0.080%
kh_iia	19	0.080%
ng_a	19	0.080%
pr_iia	19	0.080%
t_e	19	0.080%
t_x	19	0.080%
d_xx	18	0.076%
k_uua	18	0.076%

Unit	Amount	Percent
ph_vv	18	0.076%
phr_i	18	0.076%
t_vv	18	0.076%
z_qq	18	0.076%
f_o	17	0.071%
k_iia	17	0.071%
khr_uua	17	0.071%
l_iia	17	0.071%
m_qq	17	0.071%
p_@	17	0.071%
p_i	17	0.071%
ph_ii	17	0.071%
ph_x	17	0.071%
phl_aa	17	0.071%
pl_xx	17	0.071%
r_@	17	0.071%
t_v	17	0.071%
w_o	17	0.071%
b_oo	16	0.067%
ch_i	16	0.067%
h_ee	16	0.067%
k_u	16	0.067%
kh_e	16	0.067%
kh_uu	16	0.067%
khr_vva	16	0.067%
l_uua	16	0.067%
p_uua	16	0.067%
th_qq	16	0.067%
b_x	15	0.063%
c_oo	15	0.063%
c_qq	15	0.063%
d_vv	15	0.063%
f_uu	15	0.063%
khl_@@	15	0.063%
khw_a	15	0.063%
m_u	15	0.063%
m_uua	15	0.063%
ng_oo	15	0.063%
ph_u	15	0.063%
phl_a	15	0.063%
t_iia	15	0.063%
z_@	15	0.063%
z_o	15	0.063%
c_u	14	0.059%
ch_qq	14	0.059%
h_@@	14	0.059%
khl_aa	14	0.059%
khl_u	14	0.059%

Unit	Amount	Percent
n_x	14	0.059%
ph_@	14	0.059%
pl_iia	14	0.059%
s_xx	14	0.059%
t_ii	14	0.059%
th_o	14	0.059%
tr_uu	14	0.059%
tr_uua	14	0.059%
c_ii	13	0.055%
d_u	13	0.055%
f_v	13	0.055%
h_oo	13	0.055%
k_@	13	0.055%
k_ii	13	0.055%
k_x	13	0.055%
l_v	13	0.055%
n_@	13	0.055%
ng_u	13	0.055%
ph_v	13	0.055%
pl_o	13	0.055%
th_q	13	0.055%
b_uu	12	0.050%
c_iia	12	0.050%
c_uua	12	0.050%
k_vva	12	0.050%
khl_oo	12	0.050%
khl_vva	12	0.050%
m_i	12	0.050%
n_iia	12	0.050%
p_o	12	0.050%
p_vva	12	0.050%
t_uu	12	0.050%
b_@	11	0.046%
h_uu	11	0.046%
kh_xx	11	0.046%
kr_o	11	0.046%
l_vv	11	0.046%
m_oo	11	0.046%
ph_ee	11	0.046%
phl_oo	11	0.046%
pr_@@	11	0.046%
s_oo	11	0.046%
b_vva	10	0.042%
ch_oo	10	0.042%
khl_vv	10	0.042%
kl_i	10	0.042%
ph_oo	10	0.042%
phl_u	10	0.042%

Unit	Amount	Percent
tr_aa	10	0.042%
tr_iia	10	0.042%
b_qq	9	0.038%
kl_@@	9	0.038%
kl_u	9	0.038%
kr_uua	9	0.038%
l_q	9	0.038%
phl_ee	9	0.038%
pr_@	9	0.038%
s_@	9	0.038%
c_@	8	0.034%
f_xx	8	0.034%
h_vva	8	0.034%
kh_vva	8	0.034%
khl_ee	8	0.034%
khr_@@	8	0.034%
n_qq	8	0.034%
ng_@	8	0.034%
ng_o	8	0.034%
p_iia	8	0.034%
p_uu	8	0.034%
phl_i	8	0.034%
phr_oo	8	0.034%
pl_vva	8	0.034%
th_oo	8	0.034%
tr_a	8	0.034%
tr_xx	8	0.034%
w_@	8	0.034%
w_xx	8	0.034%
c_q	7	0.029%
c_uu	7	0.029%
h_qq	7	0.029%
j_ee	7	0.029%
kr_@	7	0.029%
kr_x	7	0.029%
l_@	7	0.029%
m_iia	7	0.029%
ng_@@	7	0.029%
ng_qq	7	0.029%
p_qq	7	0.029%
phr_xx	7	0.029%
pl_i	7	0.029%
pr_oo	7	0.029%
s_qq	7	0.029%
tr_@@	7	0.029%
b_ii	6	0.025%
ch_x	6	0.025%
h_ii	6	0.025%

Unit	Amount	Percent
khl_xx	6	0.025%
khr_ee	6	0.025%
khr_oo	6	0.025%
kl_o	6	0.025%
m_x	6	0.025%
ng_ee	6	0.025%
p_x	6	0.025%
r_v	6	0.025%
w_e	6	0.025%
c_xx	5	0.021%
f_ii	5	0.021%
f_qq	5	0.021%
j_qq	5	0.021%
khl_ii	5	0.021%
khr_vv	5	0.021%
khw_xx	5	0.021%
kr_qq	5	0.021%
p_xx	5	0.021%
phr_v	5	0.021%
pl_ee	5	0.021%
pl_uu	5	0.021%
w_ii	5	0.021%
w_u	5	0.021%
b_e	4	0.017%
ch_xx	4	0.017%
khl_qq	4	0.017%
khr_x	4	0.017%
kl_e	4	0.017%
kl_qq	4	0.017%
kr_i	4	0.017%
l_ii	4	0.017%
m_ee	4	0.017%
n_e	4	0.017%
ng_v	4	0.017%
pr_u	4	0.017%
pr_xx	4	0.017%
t_qq	4	0.017%
th_@	4	0.017%
thr_aa	4	0.017%
thr_i	4	0.017%
w_vv	4	0.017%
z_e	4	0.017%
b_ee	3	0.013%
c_x	3	0.013%
f_@@	3	0.013%
f_e	3	0.013%
j_ii	3	0.013%
j_x	3	0.013%

Unit	Amount	Percent
kh_oo	3	0.013%
khr_v	3	0.013%
khw_iia	3	0.013%
kl_vv	3	0.013%
kl_x	3	0.013%
kl_xx	3	0.013%
kr_ee	3	0.013%
kr_ii	3	0.013%
kr_u	3	0.013%
kw_iia	3	0.013%
kw_x	3	0.013%
n_oo	3	0.013%
ng_i	3	0.013%
ng_ii	3	0.013%
phl_v	3	0.013%
phr_qq	3	0.013%
phr_u	3	0.013%
pr_ee	3	0.013%
pr_o	3	0.013%
pr_vv	3	0.013%
thr_@@	3	0.013%
tr_e	3	0.013%
tr_ee	3	0.013%
tr_v	3	0.013%
z_uu	3	0.013%
z_x	3	0.013%
z_xx	3	0.013%
c_ee	2	0.008%
ch_ua	2	0.008%
f_@	2	0.008%
f_i	2	0.008%
f_u	2	0.008%
h_iia	2	0.008%
j_q	2	0.008%
khr_i	2	0.008%
khr_ii	2	0.008%
khw_ee	2	0.008%
kl_ii	2	0.008%
m_@	2	0.008%
n_ee	2	0.008%
ng_uua	2	0.008%
ng_x	2	0.008%
ng_xx	2	0.008%
p_@@	2	0.008%
p_ia	2	0.008%
phl_@@	2	0.008%
phl_uu	2	0.008%
pl_u	2	0.008%

Unit	Amount	Percent
pr_ii	2	0.008%
pr_q	2	0.008%
r_x	2	0.008%
s_x	2	0.008%
w_q	2	0.008%
z_va	2	0.008%
b_v	1	0.004%
ch_u	1	0.004%
d_v	1	0.004%
f_q	1	0.004%
f_vva	1	0.004%
j_v	1	0.004%
kh_@	1	0.004%
khr_@	1	0.004%
khr_u	1	0.004%
khr_xx	1	0.004%
khw_x	1	0.004%

Unit	Amount	Percent
kl_@	1	0.004%
kl_vva	1	0.004%
kr_aa	1	0.004%
kr_e	1	0.004%
ng_uu	1	0.004%
p_v	1	0.004%
p_vv	1	0.004%
ph_ia	1	0.004%
phl_@	1	0.004%
phl_iia	1	0.004%
phr_uu	1	0.004%
phr_x	1	0.004%
r_ee	1	0.004%
z_ia	1	0.004%
z_ua	1	0.004%
z_vva	1	0.004%

Table A1.8 Statistic of the Finals and rhymes in the training corpus

Unit	Amount	Percent
a_j	1,906	8.012%
aa	1,632	6.860%
a	1,215	5.107%
ii	1,173	4.931%
a_n	868	3.649%
aa_j	766	3.220%
aa_ng	617	2.594%
aa_n	606	2.547%
uu	523	2.198%
a_ng	514	2.161%
o_n	513	2.156%
@@	481	2.022%
a_m	468	1.967%
aa_m	455	1.913%
a_w	453	1.904%
e_n	451	1.896%
@@_ng	407	1.711%
aa_w	376	1.580%
v_ng	330	1.387%
xx	329	1.383%
a_p	303	1.274%
vva	273	1.148%
aa_k	272	1.143%
uua	263	1.106%
x	245	1.030%
a_k	228	0.958%
i_ng	226	0.950%
@@_n	222	0.933%
i_n	219	0.921%
o_ng	193	0.811%
@@_k	192	0.807%
aa_t	184	0.773%
a_t	183	0.769%
uu_k	156	0.656%
uua_j	154	0.647%
o_k	153	0.643%
vv	151	0.635%
o_m	150	0.631%
iia_ng	147	0.618%
xx_w	147	0.618%
o_t	144	0.605%
i	140	0.588%
@_ng	139	0.584%
vv_n	136	0.572%
@	135	0.567%
i_t	127	0.534%
u_k	124	0.521%
iia_n	117	0.492%

Unit	Amount	Percent
vva_ng	117	0.492%
u_n	116	0.488%
xx_ng	112	0.471%
x_ng	109	0.458%
@@_p	108	0.454%
uua_n	108	0.454%
u	104	0.437%
qq_n	103	0.433%
iia_w	96	0.404%
qq_j	96	0.404%
vva_n	95	0.399%
@@_m	91	0.383%
oo_ng	91	0.383%
v_n	87	0.366%
o_p	85	0.357%
u_ng	85	0.357%
uua_t	85	0.357%
oo	81	0.340%
u_t	81	0.340%
@@_j	80	0.336%
ee	80	0.336%
e_t	73	0.307%
e_k	70	0.294%
ii_p	69	0.290%
@_j	68	0.286%
qq	68	0.286%
xx_k	68	0.286%
@@_t	63	0.265%
ee_t	63	0.265%
v_k	62	0.261%
aa_p	59	0.248%
oo_n	57	0.240%
ee_ng	55	0.231%
i_m	55	0.231%
iia	54	0.227%
qq_t	49	0.206%
e_p	47	0.198%
qq_m	46	0.193%
uua_ng	46	0.193%
iia_p	45	0.189%
oo_t	44	0.185%
u_m	43	0.181%
uu_t	42	0.177%
uu_ng	41	0.172%
ii_k	40	0.168%
uua_k	40	0.168%
@_m	38	0.160%
i_w	38	0.160%

Unit	Amount	Percent
oo_j	38	0.160%
xx_n	38	0.160%
oo_k	37	0.156%
q	37	0.156%
vva_k	37	0.156%
xx_p	37	0.156%
@_n	34	0.143%
uu_p	34	0.143%
x_n	32	0.135%
i_p	31	0.130%
vv_t	31	0.130%
e_m	30	0.126%
xx_m	29	0.122%
ii_n	28	0.118%
u_j	28	0.118%
iia_m	27	0.113%
vva_m	25	0.105%
uua_m	24	0.101%
iia_t	23	0.097%
qq_k	23	0.097%
qq_ng	23	0.097%
e_ng	21	0.088%
i_k	21	0.088%
q_n	21	0.088%
vva_p	21	0.088%
ii_t	20	0.084%
xx_t	20	0.084%
e_w	19	0.080%
vva_j	19	0.080%
ee_k	18	0.076%
vv_m	18	0.076%
ee_w	17	0.071%

Unit	Amount	Percent
ii_m	17	0.071%
iia_k	17	0.071%
u_p	17	0.071%
uu_n	17	0.071%
ee_m	16	0.067%
vva_t	15	0.063%
ee_n	14	0.059%
oo_p	14	0.059%
oo_m	13	0.055%
v_p	12	0.050%
uua_p	11	0.046%
x_w	10	0.042%
e	9	0.038%
ee_p	8	0.034%
x_m	8	0.034%
x_k	7	0.029%
@_k	6	0.025%
o	6	0.025%
qq_p	6	0.025%
v	6	0.025%
@_p	5	0.021%
v_t	5	0.021%
@_t	4	0.017%
ia	4	0.017%
uu_m	4	0.017%
v_m	4	0.017%
x_t	4	0.017%
ua	3	0.013%
vv_p	3	0.013%
x_p	3	0.013%
va	2	0.008%

APPENDIX B

The Thai Text Testing Corpus

Table B1.1 Statistic of the Thai initial consonants in the test set I

Unit	Amount	Percent
s	424	9.031%
th	395	8.413%
n	385	8.200%
m	382	8.136%
d	305	6.496%
l	295	6.283%
r	261	5.559%
t	260	5.538%
k	258	5.495%
kh	255	5.431%
p	226	4.814%
ph	198	4.217%
c	193	4.111%
j	181	3.855%
z	167	3.557%
ch	136	2.897%
w	127	2.705%
h	116	2.471%
b	91	1.938%
ng	21	0.447%
f	19	0.405%

Table B1.2 Statistic of the Thai final consonants in the test set I

Unit	Amount	Percent
n	744	23.265%
ng	624	19.512%
j	450	14.071%
k	385	12.039%
t	342	10.694%
m	277	8.662%
p	203	6.348%
w	173	5.410%

Table A1.3 Statistic of the Thai consonant clusters in the test set I

Unit	Amount	Percent
kl	64	22.069%
pr	54	18.621%
khr	42	14.483%
phr	24	8.276%
khw	23	7.931%
pl	18	6.207%
phl	17	5.862%
kr	16	5.517%
khl	14	4.828%
kw	11	3.793%
tr	7	2.414%
thr	0	0.000%

Table A1.4 Statistic of the Thai vowels in the test set I

Unit	Amount	Percent
a	1,164	23.345%
aa	1,031	20.678%
ii	429	8.604%
@@	386	7.742%
i	223	4.473%
o	206	4.132%
uua	174	3.490%
e	170	3.410%
xx	170	3.410%
uu	154	3.089%
u	142	2.848%
v	111	2.226%
vva	107	2.146%
x	99	1.986%
ee	92	1.845%
vv	81	1.625%
oo	77	1.544%
iia	66	1.324%
qq	54	1.083%
@	49	0.983%
ia	0	0.000%
q	0	0.000%
ua	0	0.000%
va	0	0.000%

Table B1.5 Statistic of the context-independent Initials in the test set I

Unit	Amount	Percent
s	424	9.031%
th	395	8.413%
n	385	8.200%
m	382	8.136%
d	305	6.496%
l	295	6.283%
r	261	5.559%
t	260	5.538%
k	258	5.495%
kh	255	5.431%
p	226	4.814%
ph	198	4.217%
c	193	4.111%
j	181	3.855%
z	167	3.557%
ch	136	2.897%
w	127	2.705%
h	116	2.471%
b	91	1.938%
kl	64	22.069%
pr	54	18.621%
khr	42	14.483%
phr	24	8.276%
khw	23	7.931%
ng	21	0.447%
f	19	0.405%
pl	18	6.207%
phl	17	5.862%
kr	16	5.517%
khl	14	4.828%
kw	11	3.793%
tr	7	2.414%

Table B1.6 Statistic of the context-dependent Initials and CORM onsets in the test set I

Unit	Amount	Percent	Unit	Amount	Percent
n_a	222	4.453%	s_@	32	0.642%
th_i	167	3.350%	n_o	31	0.622%
th_a	149	2.989%	p_i	30	0.602%
m_a	148	2.969%	r_@	30	0.602%
k_a	144	2.889%	r_v	29	0.582%
c_a	132	2.648%	z_i	28	0.562%
s_a	130	2.608%	m_x	27	0.542%
d_a	116	2.327%	t_@	27	0.542%
l_a	115	2.307%	r_u	25	0.502%
m_i	115	2.307%	z_@	24	0.481%
p_e	101	2.026%	n_@	23	0.461%
j_a	97	1.946%	k_u	22	0.441%
r_a	93	1.866%	k_i	21	0.421%
l_x	88	1.765%	kh_o	21	0.421%
s_u	86	1.725%	khr_@	21	0.421%
w_a	86	1.725%	khw_a	21	0.421%
kh_a	84	1.685%	th_u	21	0.421%
ph_a	81	1.625%	th_v	21	0.421%
s_i	77	1.545%	c_@	20	0.401%
p_a	75	1.505%	ch_i	20	0.401%
t_a	75	1.505%	m_u	20	0.401%
z_a	72	1.444%	d_x	19	0.381%
ch_a	70	1.404%	m_e	19	0.381%
kh_@	69	1.384%	ph_v	19	0.381%
h_a	62	1.244%	z_u	19	0.381%
d_@	59	1.184%	kh_u	18	0.361%
t_o	58	1.163%	s_o	18	0.361%
s_v	56	1.123%	b_@	17	0.341%
n_i	53	1.063%	ng_a	17	0.341%
pr_a	50	1.003%	d_o	16	0.321%
d_u	47	0.943%	j_@	16	0.321%
r_i	47	0.943%	l_e	16	0.321%
kl_a	45	0.903%	phr_a	16	0.321%
j_u	40	0.802%	th_e	16	0.321%
k_@	40	0.802%	w_i	16	0.321%
n_v	38	0.762%	b_u	15	0.301%
kh_v	37	0.742%	f_a	15	0.301%
d_i	36	0.722%	l_u	15	0.301%
b_a	35	0.702%	r_o	15	0.301%
ph_i	34	0.682%	t_e	15	0.301%
ph_u	34	0.682%	ch_v	14	0.281%
m_v	33	0.662%	k_x	14	0.281%
t_u	33	0.662%	khr_a	14	0.281%
t_x	33	0.662%	s_e	14	0.281%
l_o	32	0.642%	th_@	14	0.281%

Unit	Amount	Percent
b_o	13	0.261%
c_v	13	0.261%
h_u	13	0.261%
h_x	13	0.261%
l_i	13	0.261%
c_u	12	0.241%
ch_e	12	0.241%
kr_a	12	0.241%
kw_a	11	0.221%
m_o	11	0.221%
p_x	11	0.221%
s_x	11	0.221%
t_i	11	0.221%
w_e	11	0.221%
w_o	11	0.221%
z_v	11	0.221%
kh_i	10	0.201%
r_q	10	0.201%
ch_o	9	0.181%
h_o	9	0.181%
k_q	9	0.181%
kl_i	9	0.181%
m_@	9	0.181%
ch_u	8	0.160%
j_o	8	0.160%
kh_e	8	0.160%
kh_x	8	0.160%
n_u	8	0.160%
ph_@	8	0.160%
pl_u	8	0.160%
r_x	8	0.160%
t_q	8	0.160%
c_o	7	0.140%
d_q	7	0.140%
h_@	7	0.140%
j_i	7	0.140%
ph_e	7	0.140%
phl_e	7	0.140%
phl_u	7	0.140%
z_e	7	0.140%
j_e	6	0.120%
k_e	6	0.120%
khl_u	6	0.120%
l_@	6	0.120%
l_v	6	0.120%

Unit	Amount	Percent
ph_q	6	0.120%
th_x	6	0.120%
b_i	5	0.100%
h_e	5	0.100%
h_i	5	0.100%
ph_o	5	0.100%
b_x	4	0.080%
c_i	4	0.080%
j_x	4	0.080%
khl_a	4	0.080%
khl_v	4	0.080%
khr_o	4	0.080%
l_q	4	0.080%
n_q	4	0.080%
ph_x	4	0.080%
phr_@	4	0.080%
pl_i	4	0.080%
r_e	4	0.080%
z_o	4	0.080%
c_x	3	0.060%
d_v	3	0.060%
f_u	3	0.060%
j_v	3	0.060%
kl_u	3	0.060%
kl_v	3	0.060%
kr_u	3	0.060%
n_e	3	0.060%
n_x	3	0.060%
p_@	3	0.060%
p_o	3	0.060%
phr_x	3	0.060%
c_e	2	0.040%
ch_@	2	0.040%
d_e	2	0.040%
k_v	2	0.040%
khr_v	2	0.040%
khw_x	2	0.040%
kl_@	2	0.040%
ng_u	2	0.040%
p_q	2	0.040%
phl_q	2	0.040%
pl_v	2	0.040%
pl_x	2	0.040%
pr_o	2	0.040%
th_o	2	0.040%

Unit	Amount	Percent
tr_a	2	0.040%
tr_i	2	0.040%
tr_u	2	0.040%
w_x	2	0.040%
z_x	2	0.040%
b_e	1	0.020%
b_v	1	0.020%
ch_q	1	0.020%
f_o	1	0.020%
h_q	1	0.020%
h_v	1	0.020%
khr_i	1	0.020%
kl_o	1	0.020%

Unit	Amount	Percent
kl_x	1	0.020%
kr_i	1	0.020%
ng_@	1	0.020%
ng_o	1	0.020%
p_v	1	0.020%
phl_a	1	0.020%
phr_i	1	0.020%
pl_@	1	0.020%
pl_a	1	0.020%
pr_i	1	0.020%
pr_x	1	0.020%
tr_o	1	0.020%

Table B1.7 Statistic of the PORM onsets in the test set I

Unit	Amount	Percent	Unit	Amount	Percent
th_ii	134	2.688%	t_a	30	0.602%
n_a	114	2.287%	m_vva	29	0.582%
m_ii	109	2.187%	r_@@	29	0.582%
n_aa	108	2.166%	s_vva	29	0.582%
p_e	101	2.026%	h_a	28	0.562%
th_a	100	2.006%	s_uua	28	0.562%
m_aa	98	1.966%	m_xx	27	0.542%
d_aa	94	1.886%	ph_ii	27	0.542%
s_a	85	1.705%	p_ii	26	0.522%
k_aa	80	1.605%	ph_aa	26	0.522%
c_a	74	1.484%	s_u	26	0.522%
l_x	72	1.444%	th_i	26	0.522%
r_a	70	1.404%	s_v	25	0.502%
kh_@@	67	1.344%	z_a	25	0.502%
k_a	64	1.284%	z_@@	24	0.481%
kh_a	63	1.264%	n_@@	23	0.461%
l_a	62	1.244%	n_v	23	0.461%
d_@@	59	1.184%	r_aa	23	0.461%
c_aa	58	1.163%	ch_aa	22	0.441%
j_a	57	1.143%	d_a	22	0.441%
ph_a	55	1.103%	k_u	22	0.441%
l_aa	53	1.063%	kh_vv	22	0.441%
m_a	50	1.003%	n_i	22	0.441%
w_a	50	1.003%	ph_uu	22	0.441%
ch_a	48	0.963%	t_oo	22	0.441%
pr_a	48	0.963%	kh_aa	21	0.421%
th_aa	48	0.963%	khw_aa	21	0.421%
z_aa	47	0.943%	l_oo	21	0.421%
p_a	45	0.903%	r_i	21	0.421%
s_aa	45	0.903%	th_v	21	0.421%
t_aa	45	0.903%	c_@@	20	0.401%
s_ii	41	0.822%	z_i	20	0.401%
j_aa	40	0.802%	d_xx	19	0.381%
k_@@	39	0.782%	k_i	19	0.381%
d_uua	38	0.762%	m_ee	19	0.381%
j_uu	37	0.742%	b_a	18	0.361%
kl_aa	36	0.722%	s_o	18	0.361%
t_o	36	0.722%	t_@@	18	0.361%
w_aa	36	0.722%	b_aa	17	0.341%
s_i	35	0.702%	ng_aa	17	0.341%
h_aa	34	0.682%	b_@@	16	0.321%
t_xx	33	0.662%	l_xx	16	0.321%
s_@@	32	0.642%	r_ii	16	0.321%
s_uu	32	0.642%	r_vv	16	0.321%
t_uua	32	0.642%	th_ee	16	0.321%
n_ii	30	0.602%	z_u	16	0.321%
n_o	30	0.602%	n_vva	15	0.301%
p_aa	30	0.602%	phr_a	15	0.301%

Unit	Amount	Percent
t_e	15	0.301%
w_i	15	0.301%
b_u	14	0.281%
d_i	14	0.281%
j_@@	14	0.281%
kh_oo	14	0.281%
c_v	13	0.261%
k_xx	13	0.261%
kh_v	13	0.261%
khr_@	13	0.261%
ph_vv	13	0.261%
r_o	13	0.261%
b_o	12	0.241%
ch_e	12	0.241%
ch_vv	12	0.241%
d_iia	12	0.241%
kr_a	12	0.241%
l_e	12	0.241%
r_uua	12	0.241%
th_@@	12	0.241%
kw_aa	11	0.221%
l_o	11	0.221%
p_xx	11	0.221%
r_uu	11	0.221%
t_i	11	0.221%
w_ee	11	0.221%
w_o	11	0.221%
d_ii	10	0.201%
d_oo	10	0.201%
h_x	10	0.201%
m_uua	10	0.201%
r_iia	10	0.201%
r_qq	10	0.201%
s_xx	10	0.201%
z_vv	10	0.201%
h_o	9	0.181%
h_uua	9	0.181%
k_qq	9	0.181%
khr_a	9	0.181%
kl_a	9	0.181%
m_o	9	0.181%
t_@	9	0.181%
c_u	8	0.160%
ch_i	8	0.160%
ch_o	8	0.160%
d_uu	8	0.160%
f_a	8	0.160%
kh_iia	8	0.160%

Unit	Amount	Percent
kh_x	8	0.160%
khr_@@	8	0.160%
kl_ii	8	0.160%
pl_uu	8	0.160%
r_xx	8	0.160%
s_ee	8	0.160%
t_qq	8	0.160%
th_u	8	0.160%
th_uua	8	0.160%
c_o	7	0.140%
ch_ii	7	0.140%
ch_uua	7	0.140%
d_qq	7	0.140%
f_aa	7	0.140%
j_o	7	0.140%
kh_o	7	0.140%
kh_uua	7	0.140%
l_iia	7	0.140%
l_uua	7	0.140%
ph_@	7	0.140%
ph_uua	7	0.140%
phl_ee	7	0.140%
phl_uu	7	0.140%
r_v	7	0.140%
th_iia	7	0.140%
d_o	6	0.120%
h_@@	6	0.120%
j_e	6	0.120%
kh_ee	6	0.120%
kh_uu	6	0.120%
khl_u	6	0.120%
l_@@	6	0.120%
m_@@	6	0.120%
m_u	6	0.120%
n_uua	6	0.120%
ph_qq	6	0.120%
r_vva	6	0.120%
s_e	6	0.120%
th_xx	6	0.120%
z_ii	6	0.120%
ch_iia	5	0.100%
j_ii	5	0.100%
k_ee	5	0.100%
kh_u	5	0.100%
khr_aa	5	0.100%
l_vva	5	0.100%
m_i	5	0.100%
ph_ee	5	0.100%

Unit	Amount	Percent
ph_i	5	0.100%
ph_o	5	0.100%
ph_u	5	0.100%
ph_vva	5	0.100%
th_uu	5	0.100%
b_i	4	0.080%
h_e	4	0.080%
h_i	4	0.080%
h_u	4	0.080%
j_xx	4	0.080%
khl_aa	4	0.080%
l_ee	4	0.080%
l_qq	4	0.080%
l_u	4	0.080%
l_uu	4	0.080%
m_uu	4	0.080%
m_vv	4	0.080%
n_qq	4	0.080%
phr_@	4	0.080%
pl_iia	4	0.080%
z_ee	4	0.080%
z_o	4	0.080%
b_xx	3	0.060%
c_uu	3	0.060%
h_xx	3	0.060%
j_u	3	0.060%
khl_vva	3	0.060%
kl_u	3	0.060%
kr_u	3	0.060%
l_i	3	0.060%
l_ii	3	0.060%
m_@	3	0.060%
n_ee	3	0.060%
p_o	3	0.060%
phr_xx	3	0.060%
r_e	3	0.060%
z_e	3	0.060%
z_uu	3	0.060%
c_i	2	0.040%
c_ii	2	0.040%
c_x	2	0.040%
ch_vva	2	0.040%
d_e	2	0.040%
d_vva	2	0.040%
f_u	2	0.040%
j_@	2	0.040%
j_i	2	0.040%
j_vva	2	0.040%

Unit	Amount	Percent
ph_i	5	0.100%
ph_o	5	0.100%
ph_u	5	0.100%
ph_vva	5	0.100%
th_uu	5	0.100%
b_i	4	0.080%
h_e	4	0.080%
h_i	4	0.080%
h_u	4	0.080%
j_xx	4	0.080%
khl_aa	4	0.080%
l_ee	4	0.080%
l_qq	4	0.080%
l_u	4	0.080%
l_uu	4	0.080%
m_uu	4	0.080%
m_vv	4	0.080%
n_qq	4	0.080%
phr_@	4	0.080%
pl_iia	4	0.080%
z_ee	4	0.080%
z_o	4	0.080%
b_xx	3	0.060%
c_uu	3	0.060%
h_xx	3	0.060%
j_u	3	0.060%
khl_vva	3	0.060%
kl_u	3	0.060%
kr_u	3	0.060%
l_i	3	0.060%
l_ii	3	0.060%
m_@	3	0.060%
n_ee	3	0.060%
p_o	3	0.060%
phr_xx	3	0.060%
r_e	3	0.060%
z_e	3	0.060%
z_uu	3	0.060%
c_i	2	0.040%
c_ii	2	0.040%
c_x	2	0.040%
ch_vva	2	0.040%
d_e	2	0.040%
d_vva	2	0.040%
f_u	2	0.040%
j_@	2	0.040%
j_i	2	0.040%
j_vva	2	0.040%

Unit	Amount	Percent
k_iia	2	0.040%
k_vva	2	0.040%
kh_@	2	0.040%
kh_e	2	0.040%
kh_vva	2	0.040%
khr_o	2	0.040%
khr_oo	2	0.040%
khr_v	2	0.040%
khw_xx	2	0.040%
kl_@@	2	0.040%
kl_vva	2	0.040%
m_oo	2	0.040%
n_u	2	0.040%
n_x	2	0.040%
ng_uua	2	0.040%
p_@	2	0.040%
p_i	2	0.040%
p_iia	2	0.040%
p_qq	2	0.040%
ph_e	2	0.040%
ph_iia	2	0.040%
ph_x	2	0.040%
ph_xx	2	0.040%
phl_qq	2	0.040%
pl_vva	2	0.040%
pl_xx	2	0.040%
pr_aa	2	0.040%
r_oo	2	0.040%
r_u	2	0.040%
s_vv	2	0.040%
th_@	2	0.040%
th_o	2	0.040%
tr_ii	2	0.040%
tr_uu	2	0.040%
w_xx	2	0.040%
z_iia	2	0.040%
z_xx	2	0.040%
b_@	1	0.020%
b_ee	1	0.020%
b_ii	1	0.020%
b_oo	1	0.020%
b_uu	1	0.020%
b_v	1	0.020%
b_x	1	0.020%
c_e	1	0.020%
c_ee	1	0.020%
c_uua	1	0.020%
c_xx	1	0.020%

Unit	Amount	Percent
ch_@	1	0.020%
ch_@@	1	0.020%
ch_oo	1	0.020%
ch_qq	1	0.020%
ch_u	1	0.020%
d_u	1	0.020%
d_v	1	0.020%
f_o	1	0.020%
f_uu	1	0.020%
h_@	1	0.020%
h_ee	1	0.020%
h_iia	1	0.020%
h_qq	1	0.020%
h_v	1	0.020%
j_oo	1	0.020%
j_v	1	0.020%
k_@	1	0.020%
k_e	1	0.020%
k_x	1	0.020%
kh_i	1	0.020%
kh_ii	1	0.020%
khl_vv	1	0.020%
khr_i	1	0.020%
kl_i	1	0.020%
kl_o	1	0.020%
kl_vv	1	0.020%
kl_xx	1	0.020%
kr_i	1	0.020%
l_v	1	0.020%
m_iia	1	0.020%
n_iia	1	0.020%
n_oo	1	0.020%
n_xx	1	0.020%
ng_@@	1	0.020%
ng_o	1	0.020%
p_@@	1	0.020%
p_v	1	0.020%
ph_@@	1	0.020%
ph_v	1	0.020%
phl_a	1	0.020%
phr_aa	1	0.020%
phr_i	1	0.020%
pl_@@	1	0.020%
pl_aa	1	0.020%
pr_iia	1	0.020%
pr_o	1	0.020%
pr_oo	1	0.020%
pr_xx	1	0.020%

Unit	Amount	Percent
r_@	1	0.020%
r_ee	1	0.020%
s_iia	1	0.020%
s_x	1	0.020%
t_u	1	0.020%
tr_a	1	0.020%

Unit	Amount	Percent
tr_aa	1	0.020%
tr_o	1	0.020%
w_@	1	0.020%
w_ii	1	0.020%
z_vva	1	0.020%

Table B1.8 Statistic of the Finals and rhymes in the test set I

Unit	Amount	Percent
a	466	9.348%
ii	394	7.904%
aa	256	5.135%
a_j	224	4.493%
aa_ng	136	2.728%
e_n	133	2.668%
aa_n	131	2.628%
a_n	125	2.508%
aa_j	125	2.508%
@@_ng	104	2.086%
aa_m	104	2.086%
aa_k	92	1.846%
uu	92	1.846%
aa_w	85	1.705%
v_ng	82	1.645%
o_n	79	1.585%
a_m	78	1.565%
xx	73	1.464%
a_ng	72	1.444%
@@_k	68	1.364%
vva	66	1.324%
x	65	1.304%
i	64	1.284%
aa_t	59	1.184%
u	59	1.184%
@@	58	1.163%
i_t	58	1.163%
@@_n	57	1.143%
a_k	55	1.103%
a_t	54	1.083%
@@_j	49	0.983%
ee_t	49	0.983%
i_n	49	0.983%
uua	47	0.943%
oo	46	0.923%
a_w	45	0.903%
a_p	44	0.883%
aa_p	43	0.863%
vv	42	0.843%
o_k	41	0.822%
u_t	40	0.802%
uua_ng	38	0.762%
uua_n	37	0.742%
o_ng	33	0.662%
vv_n	33	0.662%
i_p	28	0.562%
uua_j	28	0.562%
@	26	0.522%

Unit	Amount	Percent
@@_p	24	0.481%
x_ng	23	0.461%
ii_k	22	0.441%
iia_ng	22	0.441%
xx_ng	22	0.441%
iia_w	21	0.421%
xx_k	21	0.421%
o_m	20	0.401%
u_m	20	0.401%
vva_ng	20	0.401%
o_t	19	0.381%
uu_k	18	0.361%
uu_ng	18	0.361%
@@_t	17	0.341%
ee	17	0.341%
xx_n	17	0.341%
@_ng	15	0.301%
i_ng	15	0.301%
qq_t	14	0.281%
uu_n	14	0.281%
xx_t	14	0.281%
e_m	13	0.261%
o_p	13	0.261%
uua_m	13	0.261%
v_n	13	0.261%
xx_w	13	0.261%
qq_n	12	0.241%
ii_p	11	0.221%
oo_k	11	0.221%
vva_n	11	0.221%
e_t	10	0.201%
oo_j	10	0.201%
@@_m	9	0.181%
e_k	9	0.181%
ee_n	9	0.181%
ee_ng	9	0.181%
u_k	9	0.181%
iia_k	8	0.160%
oo_ng	8	0.160%
qq_p	8	0.160%
u_n	8	0.160%
uua_k	8	0.160%
v	8	0.160%
xx_p	8	0.160%
uu_p	7	0.140%
@_j	6	0.120%
iia_p	6	0.120%
qq	6	0.120%

Unit	Amount	Percent
qq_m	6	0.120%
vva_k	6	0.120%
ee_p	5	0.100%
i_k	5	0.100%
iia_n	5	0.100%
qq_j	5	0.100%
i_w	4	0.080%
u_ng	4	0.080%
uu_m	4	0.080%
v_k	4	0.080%
vv_t	4	0.080%
x_k	4	0.080%
x_n	4	0.080%
e_w	3	0.060%
oo_n	3	0.060%
v_m	3	0.060%
@_n	2	0.040%
e_ng	2	0.040%
ee_k	2	0.040%
ii_n	2	0.040%

Unit	Amount	Percent
iia	2	0.040%
qq_k	2	0.040%
u_j	2	0.040%
uua_p	2	0.040%
vv_p	2	0.040%
vva_p	2	0.040%
x_m	2	0.040%
xx_m	2	0.040%
@_m	1	0.020%
ee_w	1	0.020%
iia_m	1	0.020%
iia_t	1	0.020%
qq_ng	1	0.020%
uu_t	1	0.020%
uua_t	1	0.020%
v_t	1	0.020%
vva_j	1	0.020%
vva_m	1	0.020%
x_w	1	0.020%

Table B2.1 Statistic of the Thai initial consonants in the test set II

Unit	Amount	Percent
kh	46	8.630%
m	43	8.068%
n	38	7.129%
th	37	6.942%
s	36	6.754%
k	34	6.379%
l	33	6.191%
p	27	5.066%
j	26	4.878%
t	26	4.878%
r	25	4.690%
ch	24	4.503%
d	23	4.315%
ph	21	3.940%
c	19	3.565%
f	16	3.002%
b	15	2.814%
h	14	2.627%
w	13	2.439%
z	9	1.689%
ng	8	1.501%

Table B2.2 Statistic of the Thai final consonants in the test set II

Unit	Amount	Percent
ng	100	22.321
n	92	20.536
j	88	19.643
t	44	9.821
w	38	8.482
k	34	7.589
m	34	7.589
p	18	4.018

Table B2.3 Statistic of the Thai consonant clusters in the test set II

Unit	Amount	Percent
pr	11	25.581%
kr	8	18.605%
kl	7	16.279%
khr	6	13.953%
phr	5	11.628%
phl	2	4.651%
khl	1	2.326%
khw	1	2.326%
kw	1	2.326%
pl	1	2.326%

Table B2.4 Statistic of the Thai vowels in the test set II

Unit	Amount	Percent
aa	165	28.646%
a	136	23.611%
@@	35	6.076%
o	33	5.729%
i	23	3.993%
ii	23	3.993%
uua	21	3.646%
uu	19	3.299%
xx	15	2.604%
e	14	2.431%
iia	14	2.431%
vva	14	2.431%
oo	11	1.910%
@	9	1.563%
ee	9	1.563%
qq	9	1.563%
u	9	1.563%
x	7	1.215%
vv	5	0.868%
v	4	0.694%
q	1	0.174%

Table B2.5 Statistic of the context-independent Initials in the test set II

Unit	Amount	Percent
kh	46	7.986%
m	43	7.465%
n	38	6.597%
th	37	6.424%
s	36	6.250%
k	34	5.903%
l	33	5.729%
p	27	4.688%
j	26	4.514%
t	26	4.514%
r	25	4.340%
ch	24	4.167%
d	23	3.993%
ph	21	3.646%
c	19	3.299%
f	16	2.778%
b	15	2.604%
h	14	2.431%
w	13	2.257%
pr	11	1.910%
z	9	1.563%
kr	8	1.389%
ng	8	1.389%
kl	7	1.215%
khr	6	1.042%
phr	5	0.868%
phl	2	0.347%
khl	1	0.174%
khw	1	0.174%
kw	1	0.174%
pl	1	0.174%

Table B2.6 Statistic of the context-dependent Initials and CORM onsets in the test set II

Unit	Amount	Percent
n_a	28	4.861%
kh_a	27	4.688%
m_a	23	3.993%
p_a	18	3.125%
th_a	17	2.951%
s_a	16	2.778%
ch_a	15	2.604%
f_a	15	2.604%
k_a	14	2.431%
c_a	13	2.257%
w_a	13	2.257%
r_a	12	2.083%
d_a	10	1.736%
h_a	9	1.563%
j_a	9	1.563%
th_i	9	1.563%
b_a	8	1.389%
kr_a	8	1.389%
l_a	8	1.389%
pr_a	8	1.389%
k_@	7	1.215%
l_x	7	1.215%
p_e	7	1.215%
ph_a	7	1.215%
ph_u	7	1.215%
s_i	7	1.215%
t_a	7	1.215%
t_o	7	1.215%
j_u	6	1.042%
kh_@	6	1.042%
kl_a	6	1.042%
l_u	6	1.042%
t_@	6	1.042%
b_i	5	0.868%
j_i	5	0.868%
kh_o	5	0.868%
m_v	5	0.868%
r_u	5	0.868%
s_v	5	0.868%
z_i	5	0.868%
ch_i	4	0.694%
khr_v	4	0.694%
m_i	4	0.694%
m_u	4	0.694%
m_x	4	0.694%
n_i	4	0.694%
ng_a	4	0.694%
th_u	4	0.694%
ch_@	3	0.521%
d_@	3	0.521%
j_o	3	0.521%
k_e	3	0.521%
k_q	3	0.521%
kh_v	3	0.521%
l_o	3	0.521%
m_o	3	0.521%
n_o	3	0.521%
ng_@	3	0.521%
ph_o	3	0.521%
r_i	3	0.521%
r_o	3	0.521%
s_@	3	0.521%
th_e	3	0.521%
th_o	3	0.521%
b_u	2	0.347%
c_o	2	0.347%
c_v	2	0.347%
ch_u	2	0.347%
d_e	2	0.347%
d_i	2	0.347%
d_o	2	0.347%
d_q	2	0.347%
h_e	2	0.347%
j_@	2	0.347%
k_i	2	0.347%
k_o	2	0.347%
k_u	2	0.347%
kh_u	2	0.347%
khr_@	2	0.347%
l_@	2	0.347%
l_i	2	0.347%
l_q	2	0.347%
l_v	2	0.347%
n_u	2	0.347%
phr_@	2	0.347%
phr_i	2	0.347%
pr_o	2	0.347%
s_u	2	0.347%
t_i	2	0.347%
t_u	2	0.347%

Unit	Amount	Percent
t_x	2	0.347%
z_a	2	0.347%
c_e	1	0.174%
c_u	1	0.174%
d_u	1	0.174%
d_x	1	0.174%
f_o	1	0.174%
h_@	1	0.174%
h_o	1	0.174%
h_x	1	0.174%
j_x	1	0.174%
k_x	1	0.174%
kh_e	1	0.174%
kh_i	1	0.174%
kh_x	1	0.174%
khl_v	1	0.174%
khw_a	1	0.174%
kl_u	1	0.174%
kw_a	1	0.174%
l_e	1	0.174%
n_v	1	0.174%

Unit	Amount	Percent
ng_q	1	0.174%
p_@	1	0.174%
p_i	1	0.174%
ph_@	1	0.174%
ph_e	1	0.174%
ph_i	1	0.174%
ph_x	1	0.174%
phl_e	1	0.174%
phl_q	1	0.174%
phr_a	1	0.174%
pl_a	1	0.174%
pr_i	1	0.174%
r_@	1	0.174%
r_x	1	0.174%
s_e	1	0.174%
s_o	1	0.174%
s_x	1	0.174%
th_x	1	0.174%
z_@	1	0.174%
z_q	1	0.174%

Table B2.7 Statistic of the PORM onsets in the test set II

Unit	Amount	Percent	Unit	Amount	Percent
kh_aa	17	2.951%	m_ii	4	0.694%
n_a	16	2.778%	m_uu	4	0.694%
m_aa	12	2.083%	m_xx	4	0.694%
n_aa	12	2.083%	ng_aa	4	0.694%
s_aa	12	2.083%	r_a	4	0.694%
m_a	11	1.910%	s_a	4	0.694%
p_a	11	1.910%	s_ii	4	0.694%
ch_aa	10	1.736%	th_uua	4	0.694%
kh_a	10	1.736%	d_@@	3	0.521%
th_a	10	1.736%	d_a	3	0.521%
w_a	10	1.736%	j_o	3	0.521%
k_a	9	1.563%	k_e	3	0.521%
b_aa	8	1.389%	k_qq	3	0.521%
f_aa	8	1.389%	l_o	3	0.521%
kr_a	8	1.389%	m_o	3	0.521%
r_aa	8	1.389%	n_o	3	0.521%
th_ii	8	1.389%	ph_oo	3	0.521%
c_a	7	1.215%	r_uua	3	0.521%
d_aa	7	1.215%	s_@@	3	0.521%
f_a	7	1.215%	s_i	3	0.521%
h_aa	7	1.215%	s_vv	3	0.521%
k_@@	7	1.215%	th_ee	3	0.521%
p_aa	7	1.215%	w_aa	3	0.521%
p_e	7	1.215%	b_i	2	0.347%
ph_uu	7	1.215%	b_iia	2	0.347%
pr_a	7	1.215%	b_u	2	0.347%
t_o	7	1.215%	c_o	2	0.347%
th_aa	7	1.215%	c_v	2	0.347%
c_aa	6	1.042%	ch_@@	2	0.347%
kh_@@	6	1.042%	ch_uua	2	0.347%
kl_aa	6	1.042%	d_e	2	0.347%
l_x	6	1.042%	d_oo	2	0.347%
ph_aa	6	1.042%	d_qq	2	0.347%
ch_a	5	0.868%	h_a	2	0.347%
j_aa	5	0.868%	h_ee	2	0.347%
j_i	5	0.868%	j_@	2	0.347%
k_aa	5	0.868%	j_u	2	0.347%
kh_o	5	0.868%	kh_uua	2	0.347%
m_vva	5	0.868%	kh_vv	2	0.347%
t_@@	5	0.868%	khr_@@	2	0.347%
t_aa	5	0.868%	l_qq	2	0.347%
ch_iia	4	0.694%	l_uua	2	0.347%
j_a	4	0.694%	l_vva	2	0.347%
j_uu	4	0.694%	n_ii	2	0.347%
khr_vva	4	0.694%	ng_@	2	0.347%
l_a	4	0.694%	phr_i	2	0.347%
l_aa	4	0.694%	pr_oo	2	0.347%
l_uu	4	0.694%	r_iia	2	0.347%

Unit	Amount	Percent
r_oo	2	0.347%
r_u	2	0.347%
s_uua	2	0.347%
s_vva	2	0.347%
t_a	2	0.347%
t_i	2	0.347%
t_uua	2	0.347%
t_xx	2	0.347%
th_o	2	0.347%
z_ii	2	0.347%
z_iia	2	0.347%
b_ii	1	0.174%
c_e	1	0.174%
c_u	1	0.174%
ch_@	1	0.174%
d_i	1	0.174%
d_ii	1	0.174%
d_uua	1	0.174%
d_xx	1	0.174%
f_o	1	0.174%
h_@@	1	0.174%
h_o	1	0.174%
h_xx	1	0.174%
j_xx	1	0.174%
k_i	1	0.174%
k_iia	1	0.174%
k_o	1	0.174%
k_oo	1	0.174%
k_u	1	0.174%
k_uua	1	0.174%
k_xx	1	0.174%
kh_ee	1	0.174%
kh_i	1	0.174%
kh_v	1	0.174%
kh_x	1	0.174%
khl_vva	1	0.174%
khw_aa	1	0.174%
kl_uua	1	0.174%
kw_aa	1	0.174%
l_@	1	0.174%
l_@@	1	0.174%
l_e	1	0.174%

Unit	Amount	Percent
l_i	1	0.174%
l_iia	1	0.174%
l_xx	1	0.174%
n_i	1	0.174%
n_iia	1	0.174%
n_u	1	0.174%
n_uua	1	0.174%
n_v	1	0.174%
ng_@@	1	0.174%
ng_q	1	0.174%
p_@	1	0.174%
p_ii	1	0.174%
ph_@@	1	0.174%
ph_a	1	0.174%
ph_ee	1	0.174%
ph_i	1	0.174%
ph_xx	1	0.174%
phl_ee	1	0.174%
phl_qq	1	0.174%
phr_@	1	0.174%
phr_@@	1	0.174%
phr_aa	1	0.174%
pl_aa	1	0.174%
pr_aa	1	0.174%
pr_iia	1	0.174%
r_@@	1	0.174%
r_i	1	0.174%
r_o	1	0.174%
r_xx	1	0.174%
s_ee	1	0.174%
s_o	1	0.174%
s_xx	1	0.174%
t_@	1	0.174%
th_i	1	0.174%
th_oo	1	0.174%
th_xx	1	0.174%
z_@@	1	0.174%
z_a	1	0.174%
z_aa	1	0.174%
z_i	1	0.174%
z_qq	1	0.174%

Table B2.8 Statistic of the Finals and rhymes in the test set II

Unit	Amount	Percent
aa_j	41	7.118%
a_j	40	6.944%
aa	32	5.556%
a	23	3.993%
aa_n	23	3.993%
aa_ng	23	3.993%
ii	20	3.472%
aa_w	19	3.299%
a_ng	17	2.951%
o_n	17	2.951%
a_w	16	2.778%
a_n	14	2.431%
uu	13	2.257%
aa_m	12	2.083%
@@_ng	10	1.736%
a_m	10	1.736%
a_t	9	1.563%
i_ng	9	1.563%
@@	8	1.389%
aa_t	8	1.389%
e_n	7	1.215%
iia_ng	7	1.215%
vva_ng	7	1.215%
@@_p	6	1.042%
aa_k	6	1.042%
i_n	6	1.042%
o_k	6	1.042%
uua	6	1.042%
x	6	1.042%
@@_n	5	0.868%
ee_t	5	0.868%
xx	5	0.868%
@	4	0.694%
@@_k	4	0.694%
@_ng	4	0.694%
a_p	4	0.694%
i_t	4	0.694%
oo_n	4	0.694%
u_k	4	0.694%
uu_k	4	0.694%
uua_n	4	0.694%
uua_ng	4	0.694%
uua_t	4	0.694%
vv	4	0.694%
vva	4	0.694%
a_k	3	0.521%
e_k	3	0.521%
e_ng	3	0.521%
o_m	3	0.521%
o_ng	3	0.521%
o_t	3	0.521%
oo_ng	3	0.521%
qq_j	3	0.521%
qq_n	3	0.521%
v_ng	3	0.521%
xx_n	3	0.521%
@@_m	2	0.347%
ee	2	0.347%
iia_n	2	0.347%
iia_p	2	0.347%
iia_w	2	0.347%
oo_t	2	0.347%
qq_t	2	0.347%
u_t	2	0.347%
uua_m	2	0.347%
xx_k	2	0.347%
xx_m	2	0.347%
xx_ng	2	0.347%
@_m	1	0.174%
aa_p	1	0.174%
e_t	1	0.174%
ee_ng	1	0.174%
ee_p	1	0.174%
i_k	1	0.174%
i_m	1	0.174%
i_p	1	0.174%
i_w	1	0.174%
ii_k	1	0.174%
ii_p	1	0.174%
ii_t	1	0.174%
iia_t	1	0.174%
o_p	1	0.174%
oo	1	0.174%
oo_j	1	0.174%
q_n	1	0.174%
qq_ng	1	0.174%
u_j	1	0.174%
u_m	1	0.174%
u_ng	1	0.174%
uu_ng	1	0.174%
uu_t	1	0.174%
uua_j	1	0.174%
v_n	1	0.174%
vv_n	1	0.174%
vva_j	1	0.174%
vva_n	1	0.174%

Unit	Amount	Percent
vva_t	1	0.174%
x_ng	1	0.174%

Unit	Amount	Percent
xx_p	1	0.174%

VITAE

Ekkarit Maneenoi was born in Saraburi, Thailand in 1976. He received the B.Eng. and M.Eng. degrees in Electrical Engineering from Chulalongkorn University, Thailand in 1996 and 1998, respectively . He is currently pursuing the Ph.D. Degree at the same university. In 1996, he joined Digital Signal Processing Research Laboratory, Department of Electrical Engineering, Chulalongkorn University, and has been involved in research work in speech recognition. He is also a research assistant in Research and Development Cooperative Project between Electrical Engineering Department and Private Sector. His research interests include acoustic modeling, speech analysis, speech recognition, and spoken dialogue system.