การศึกษาหน่วยเริ่มของพยางค์เชิงกลสัทศาสตร์ : พื้นฐานสำหรับระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทย

นาย วิศรุต อาขุบุตร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2545

ISBN 974-17-1358-4

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

Mr. Visarut Ahkuputra

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 2002

ISBN 974-17-1358-4

Thesis Title	An Acoustic Study of Syllable Onsets : A Basis for Thai Continuous Speech Recognition System
By	Mr. Visarut Ahkuputra
Field of study	Electrical Engineering
Thesis Advisor	Associate Professor Somchai Jitapunkul, Dr.Ing.
Thesis Co-advisor	Assistant Professor Sudaporn Luksaneeyanawin, 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 Somsak Panyakeow, D.Eng.)

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.)

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

..... Member
(Chularat Tanprasert, Ph.D.)

วิศรุต อาขุบุตร, นาย : การศึกษาหน่วยเริ่มของพยางค์เชิงกลสัทศาสตร์ : พื้นฐานสำหรับระบบการรู้จำเสียงพูดต่อเนื่องภาษาไทย. (AN ACOUSTIC STUDY OF SYLLABLE ONSETS : A BASIS FOR THAI CONTINUOUS SPEECH RECOGNITION SYSTEM) อ. ที่ปรึกษา : รองศาสตราจารย์ ดร.สมชาย จิตะพันธ์กุล, อ. ที่ปรึกษาร่วม : ผู้ช่วยศาสตราจารย์ ดร.สุดาพร ลักษณียนาวิน 117 หน้า. ISBN 974-17-1358-4.

วิทยานิพนธ์เล่มนี้มีวัตถุประสงค์ของงานวิจัยเพื่อพัฒนาหน่วยเสียงในเชิงกลสัทศาสตร์สำหรับแบบจำลองหน่วยเริ่มพยางค์ภาษาไทย หลักการของหน่วยเริ่มพยางค์และหน่วยตามพยางค์นี้ได้ถูกนำมาประยุกต์ใช้ในการรู้จำเสียงพูดต่อเนื่องภาษาไทย พยางค์ในภาษาไทยได้รับการวิเคราะห์ในเชิงกลสัทศาสตร์และพบว่า พยางค์ประกอบด้วยคู่ของหน่วยเริ่มพยางค์และหน่วยตามพยางค์ โดยหน่วยเริ่มพยางค์นั้นได้รวมส่วนของ Transitional Period ระหว่างพยัญชนะต้นและสระที่อยู่ติดกัน ส่วนหน่วยตามพยางค์นั้นครอบคลุมทั้งส่วนสระและพยัญชนะตัวสะกด ส่วน Transitional Period นั้นมีลักษณะเฉพาะในเชิงกลสัทศาสตร์ตามพยัญชนะต้นและสระซึ่งจำเป็นสำหรับการรู้จำพยัญชนะต้น

งานวิจัยนี้ได้นำเสนอแบบจำลองเชิงกลสัทศาสตร์ของหน่วยเริ่มพยางค์ไว้ ๒ ประเภท ได้แก่ Phonotactic (PORMs) และ Contextual (CORMs) แบบจำลองชนิด PORMs นั้นพิจารณาพยัญชนะต้นในบริบทของสระที่แตกต่างกันว่าเป็นคนละหน่วยเสียง จึงมีจำนวนหน่วยเริ่มพยางค์ ๗๙๒ หน่วยและหน่วยตามพยางค์ ๒๐๐ หน่วย สำหรับแบบจำลองชนิด CORMs จะพิจารณารวมหน่วยเริ่มพยางค์ที่อยู่ในบริบทของคู่สระสั้น-ยาวเดียวกันไว้เป็นหน่วยเสียงเดียวกัน ทำให้หน่วยเริ่มพยางค์ลดลงเหลือเพียง ๒๙๗ หน่วย ในการสร้างแบบจำลองหน่วยเริ่มพยางค์นั้น หน่วยเริ่มพยางค์จะซ้อนทับกับส่วนสระของหน่วยตามพยางค์เพื่อครอบคลุมส่วน Transitional Period ระหว่างพยัญชนะต้นและสระที่อยู่ติดกัน งานวิจัยนี้ได้นำเสนอเทคนิคการซ้อนทับกันของแบบจำลองหน่วยเริ่มพยางค์ทั้งสองไว้ ๒ วิธีการ ได้แก่ การซ้อนทับกันแบบคงที่ (Fixed) และแบบแปรผัน (Variable) โดยการซ้อนทับกันแบบคงที่จะมีระยะซ้อนกับที่ ๑๐ ๒๐ หรือ ๓๐ มิลลิวินาทีเข้าไปยังส่วนสระ สำหรับการซ้อนทับกันแบบแปรผันจะขึ้นกับความยาวของส่วนสระที่ร้อยละ ๕ ๑๐ ๑๕ ๒๐ หรือ ๒๕ ของความยาวส่วนสระ

เนื่องจากมีข้อมูลเสียงพูดผู้ชายเพียงคนเดียว จึงมีหน่วยเริ่มและหน่วยตามพยางค์เพียงบางส่วนที่ถูกจำลองแบบ แบบจำลองชนิด PORMs และ CORMs มีหน่วยเริ่มพยางค์ ๓๘๔ หน่วยและ ๒๑๘ หน่วยตามลำดับโดยมีหน่วยตาม ๑๔๔ หน่วย แบบจำลองของหน่วยเริ่มและหน่วยตามพยางค์นี้ถูกสร้างขึ้นโดยใช้แบบจำลองฮิดเดนมาร์คอฟ อัตรารู้จำหน่วยเริ่มพยางค์ผิดพลาดที่ต่ำที่สุดมีค่าร้อยละ ๑๐.๓๘ เมื่อใช้แบบจำลองชนิด CORMs ที่ระยะซ้อนทับร้อยละ ๒๕ เมื่อพิจารณาถึงแบบจำลองหน่วยเริ่มและหน่วยตาม แบบจำลองชนิด PORMs จะให้อัตรารู้จำคำผิดพลาดร้อยละ ๑๓.๕๒ ที่ระยะซ้อนทับร้อยละ ๒๐ ส่วนแบบจำลองชนิด CORMs จะให้อัตรารู้จำคำผิดพลาดร้อยละ ๑๖.๕๑ ที่ระยะซ้อนทับร้อยละ ๑๕ ส่วนแบบจำลอง phone ให้อัตรารู้จำคำผิดพลาดร้อยละ ๓๗.๑๒ ดังนั้นแบบจำลองชนิด PORMs ให้ผลที่ดีกว่าแบบจำลอง phone โดยลดอัตรารู้จำคำผิดพลาดได้มากถึงร้อยละ ๕๕.๗๒

ภาค	วิชาวิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต .....
สาขา	วิศวกรรมไฟฟ้า	ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา	๒๕๔๕	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

4071809221 : MAJOR ELECTRICAL ENGINEERING
KEY WORD: ACOUSTIC MODELLING / ONSET-RHYME / SPEECH ANALYSIS / THAI CONTINUOUS SPEECH RECOGNITION

VISARUT AHKUPUTRA : THESIS TITLE (AN ACOUSTIC STUDY OF SYLLABLE ONSETS : A BASIS FOR THAI CONTINUOUS SPEECH RECOGNITION SYSTEM) THESIS ADVISOR : ASSOC. PROF. SOMCHAI JITAPUNKUL, Dr.Ing., THESIS COADVISOR : ASST. PROF. SUDAPORN LUKSANEEYANAWIN, Ph.D., 117 pp. ISBN 974-17-1358-4.

The objective of this dissertation is to develop a new acoustic speech units on modelling of the Thai onset units. The concept of onset and rhyme units is applied to Thai continuous speech recognition. Thai syllables are acoustically analysed and found that a syllable is composed of a pair of onset and rhyme units. The onset unit incorporates transitional period existed between releasing consonant and its adjacent vowel. The rhyme unit covers both vowel and arresting consonant. The transitional period has unique acoustic characteristics depending on releasing consonant and vowel which is crucial in recognition of the consonant.

Two acoustic models of the onset-rhyme unit are introduced in this dissertation—Phonotactic Onset-Rhyme Models (PORMs) and Contextual Onset-Rhyme Models (CORMs). The PORMs consider the same releasing consonant in different context as different models. This results in 792 onset units and 200 rhyme units. The CORMs consider the same releasing consonant within similar short-long vowel context as the same models. The number of onset units is then reduced to only 297 units. In modelling of the onset units, the onset unit overlaps over vowel segment of rhyme unit to cover transitional period between releasing consonant and adjacent vowel. Two overlapping techniques are proposed in modelling of the onset units—fixed duration overlap and variable duration overlap. The fixed duration overlap has constant duration at 10, 20, or 30 ms into the vowel segment. The variable duration overlap has variable duration at either 5%, 10%, 15%, 20%, or 25% of the vowel segment.

Due to limited speech data and only one male speaker, only partial set of the onset and rhyme units are modelled in the speaker-dependent recognition system. The PORMs and CORMs contain 384 onset units and 218 onset units, respectively. Both models share the same 144 rhyme units. Acoustic models of these onset and rhyme units are created using Hidden Markov Models. The lowest onset error rate achieved is 10.38% using the CORMs at 25% overlap. Considering both onset and rhyme units, the PORMs provide better word error rate at 13.53% using 20% overlap with no grammar. The CORMs give out 16.51% word error rate at 15% overlap. The phone models give out 37.12% word error rate. Hence, the PORMs outperform the phone models up to 55.76% reduction in word error rate.

Department	Electrical Engineering	Student's signature .....
Field of study	Electrical Engineering	Advisor's signature.....
Academic year	2002	Co-advisor's signature .....

Acknowledgements

Firstly, I would like to express my deepest gratitute to both of my advisors, Assoc. Prof. Dr. Somchai Jitapunkul and Asst. Prof. Dr. Sudaporn Luksaneeyanawin. Both of them have inspired, encouraged, guided, and supported me in every ways throughout my entire studies. I had been their student since the beginning of my graduate study in M.Eng. Especially Assoc. Prof. Dr Somchai Jitapunkul, I have known him for over ten years since my undergraduate study at the Chulalongkorn University.

Secondly, Asst. Prof. Dr. Sudaporn Luksaneeyanawin has taught me many things in both academic and social life. She broadens my aspects in speech technology with her expertise and experience in this particular field. She makes me realize how importance of interdisciplinary research in speech technology. Also, the importance of working with people from various fields of research, who contribute their works to speech technology.

Thirdly, I would like to show my thankfulness to all of my thesis committee especially Prof. Dr. Prasit Prapinmongkolkarn, the chairman. Every time I met him, he shows his concern on the progress of my research. He also encourages and guides me during my study. He also regularly gives me some technical articles related to my research whenever he found. The other committees, Assoc. Prof. Dr. Watit Benjapolakul and Dr. Chularat Tanprasert, please receive my thankfulness for their precious comments and suggestions regarding my work.

In addition, I would like to acknowledge the Telecommunication Consortium Scholarship, National Science and Technology Development Agency (NSTDA). This scholarship provided financial support during my doctoral study. Also, the Graduate School of Chulalongkorn University provided some grants for my research.

Moreover, 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 on the entire years of my study. All of my Ph.D. colleagues, Dheerasak Anantakul, Suphachet Phermphoonwatanasuk, Tanun Jaruwitayakovit, Surapong Suwankawin, and Ekkarit Maneenoi on every kind of their support throughout those years of our study together.

Lastly, I would like to express my deepest gratefulness from my heart to my family. My father, mother, and sister have entirely supported, encouraged, and believed in me without any doubts. Also, every one of my relatives gives me confidence in my study. A very special thankfulness to Miss Montakarn Sriphanlam, who gives me mental support with her entire heart believe in me and my study. These people have given me strength throughout my entire doctoral study. I may not get through this tough time without their supports. With all the support given to me, I believe in myself that I could pass through this tough time with all the strength and encouragements every one gives to me.

Table of Contents

Table of Contents (cont.)

Table of Contents (cont.)

List of Tables

List of Tables (cont.)

List of Figures

List of Figures (cont.)

CHAPTER 1

Introduction

Speech can be considered as the most natural way of human communication and interaction. Humans utilize speech as a communication medium since they were born. As for the human-machine interaction, a conventional method is a keyboard input and screen output as seen in most computer systems. This way of interaction is both inconvenient and inefficient for anyone with less typing skills or people with disabilities. Then, speech input provides an alternative means of human-machine interaction as user-friendly interface which is more natural to human users, less intimidating than a keyboard, and thus requires much less operating skills.

Researches in speech processing are progressing considerably during the past decades up to the present. In the past four decades, research in speech recognition has been considerably progressed since the earliest attempts in the 1950s. (Rabiner and Juang, 1993; Zue, et al., 1995) An interdisciplinary research on speech recognition effectively utilizes knowledge from many sources such as linguistics, psychology, computer science, and engineering. Applications of automatic speech recognition and speech synthesis are incorporated into many tasks such as voice dialing in mobile phones, voice-activated controls, banking, security systems, air traffic information retrieval, weather information retrieval, etc. (Rabiner and Juang, 1993; Zue, et al., 1995)

Spoken language processing as well as computing technology play a major role in rapid advances of spoken language system technology. Several successful speech recognition prototypes have been proposed based on underlying word model (Lee, 1989; Rabiner and Juang, 1993). This word model or word-based approach has already compensated for the coarticulatory effect in the model by treating each utterance as a whole. However, these particular 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. Then, a concept of subword model has been proposed to use a smaller number of units which construct a word or a syllable as a recognition unit, that is, a phonemic unit or a phoneme (Lee and Hon, 1988, 1989; Lee, 1989, 1990; Lee, Hon, and Reddy, 1990; Lee et al., 1990; Rabiner and Juang, 1993).

1.1 Speech Recognition Framework

In recognition of an unknown utterance, each utterance is assumed to comprise a sequence of structured and lingistically meaningful words (Juang and Furui, 2000). Bayes' decision theory have been applied in decoding of a sequence of words as shown in Eq. (1.1).

$$\mathbf{W} = \arg\max_{w} P(W|X) = \arg\max_{w} \frac{P(W)P(X|W)}{P(X)} \quad \text{..... (1.1)}$$

$$\text{where} \quad X = (x_1, x_2, \ldots, x_T) \quad \text{and} \quad W = w_1, w_2, \ldots, w_N; \quad w_i \in V$$

From Eq. (1.1), the sentence or the word sequence **W** is a result of maximum a posteriori on probability of a word sequence W given a possible acoustic realization X in which each word exists in the vocabulary V. The X is an acoustic realization of a sequence of words W. The P(X|W) is related to probabilistic realization of the word sequence. The P(W) defines the probabilistic relationship that exists among words when they appear in sequence (Lee, 1989;

Huang, Acero, and Hon, 2001; Juang and Furui, 2000). The $P(X)$ is probability of acoustic realization sequence X. Applying Bayes' theory, the $P(W)$ and the $P(X|W)$ are referred to language model and acoustic model respectively.

In recognition of continuous speech, various kinds of speech units have been used to handle coarticulatory effects existed in continuous utterances. The complexity of a recognition system is directly related to a number of speech units. Examples of speech units are ranging from words, syllables, phones, etc., where issues in both high acoustic resolution and low estimation reliability, or consistency and trainability, must be considered. Various kinds of speech units currently used in most continuous speech recognition systems are described in the next section.

1.2 Selection of Speech Units

Selection of speech units is one of the most important issue in designing and developing a continuous speech recognizer. Particular speech segments have been used as the basic modelling unit for a continuous speech recognizer which determine the acoustic resolution and estimation reliability of the basic model. Then, the tradeoff between high acoustic resolution and low estimation reliability, or detailed models and limited training data, have to be compromised between the two issues (Lee, 1990; Juang and Furui, 2000). Currently, there are many speech units utilized in speech recognition systems, for examples, word, phone, etc. Summary of evaluation on these speech units is shown in Table 1.1. The details of each speech units are described in this section.

A. Word Model

Word models assimilate phonological variations on within-word contextual effects or coarticulatory effects. Word models are the most natural speech units since a continuous speech recognition system considers a sentence as a sequence of words. Many samples of each word are needed to reliably estimate a word model. Acoustic data of a word are solely used for training of that particular word and is unable to be shared among words. Then, for a large-vocabulary speech recognition system, it is very difficult to collect acoustic data for every new word to be reliably estimated. Moreover, in a continuous speech, there are coarticulatory effects between each word or at word boundaries in which the word models are not be able to model.

B. Context-Independent Phones—Monophone Models

In order to share models across words, common subword models have been used, the phonetic models. Phonetically, the smallest subword units are phonemes or monophones.

Table 1.1 Evaluation of previously proposed units of speech to large vocabulary recognition (Lee, 1990)

Units	Consistency	Trainability
Word model	Yes	No
Phone model	No	Yes
Multi-phone model	Yes	Difficult
Transition model	Yes	Difficult
Word-dependent phone model	Yes	Through Sharing
Context-dependent phone model	Yes	Through Sharing

The sequence of monophone models make up a single word. There are only about 50 phones in English and about 57 phones in Thai, then, the monophone models are sufficiently trained with just a few hundred sentences. However, the monophone models assume that any monophones in different context have similar characteristics, in other words, context-independent. But, in practical, a monophone is strongly affected by its immediate adjacent monophones. Hence, the monophone models is overgeneralize where the word models lack generality (Lee, 1990).

C. Context-Dependent Phones—Diphones and Triphones

Modelling of context-dependent phones is to model phone-in-context which is referred to the immediate left and/or right neighbouring phones. A left-context dependent phone is dependent on the left context while a right-context dependent phone is dependent on the right context. Both of the left-context dependent phones and the right-context dependent phones are the diphones. A triphone considers both the left and right neighbouring phones.

A model of diphone consists of transitional parts of a phone pair : consonant-vowel (CV), vowel-consonant (VC), consonant-consonant (CC), and vowel-vowel (VV). This unit also includes the steady state parts of vowels, nasals, and fricatives (Rosenberg, 1988; Lee, Rabiner, Pieraccini, and Wilpon, 1990). The diphone is a context-dependent model which covers a great deal of phonological variations and contextual effects within the unit and less

Table 1.2 Number of grammatically occurred Thai diphone units

Thai diphone (right context-dependent only)	Combinations	Number of Units
consonant preceding vowel	C + V	33 x 24 = 792
vowel preceding consonant	V + C	24 x 33 = 792
silence preceding consonant	sil + C	1 x 33 = 33
consonant preceding silence	C + sil	8 x 1 = 8
vowel preceding silence	V + sil	24 x 1 = 24
silence	sil	1
Total diphone units in Thai		**1,650**

Table 1.3 Number of grammatically occurred Thai triphone units

Thai Triphones	Combinations	Number of Units
consonant – vowel + consonant	C – V + C	33x24x33 = 26,136
consonant – vowel + silence	C – V + sil	33x24x1 = 792
vowel – consonant + vowel	V – C + V	24x33x24 = 19,008
vowel – consonant + consonant	V – C + C	24x8x33 = 6,336
silence – consonant + vowel	sil – C + V	1x33x24 = 792
vowel – consonant + silence	V – C + sil	24x8x1 = 192
consonant – consonant + vowel	C – C + V	8x33x24 = 6,336
silence	sil	1
Total triphone units in Thai		**59,593**

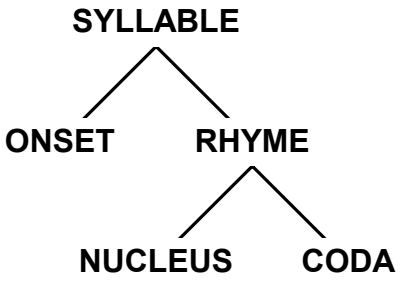


Figure 1.1 Syllable parts.

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

Figure 1.2 Thai syllable structure (Luksaneeyanawin, 1993).

variable than phones. In English, there are combinations of 46x45 = 2,070 diphone units to cover all words in English (Rabiner and Juang, 1993). Phonotactically, there are 1,650 left diphone units, 1,650 right diphone units and 59,593 triphone units grammatically existed in Thai as shown in Table 1.2 and Table 1.3.

A model of triphone is a phone-sized model that considers both left and right neighbouring phones, that is, left-and-right context-dependent phone (Lee, Hon, and Reddy, 1990). The triphone covers the most important coarticulatory effects and is much more sensitive than phone modeling. The numbers of triphone units are listed in Table 1.3. Due to a large number of triphone units, they are very difficult to train using a limited number of training data.

Analysis of syllable structures are shown in Table 1.4. Possible combinations of syllables are described in both English and Thai. Therefore, the triphones are more practical to English than Thai due to the complexity of syllable structure in English which contains many clusters. In Thai, a triphone model is equivalent to word model in syllable structure aspect which is not considered as a subword model. Phonologically speaking, for the Thai syllable, this model does not provide any difference over a word model in recognition. The phonological structure of the Thai syllable is shown in Figure 1.2 (Luksaneeyanawin, 1993). However, application of onset-rhyme models in English might cause a large number of both onset and rhyme units. This is resulted from a large number of clusters in English as shown Table 1.4 in which combinations of English syllables are more complicated than Thai.

Table 1.4 English and Thai syllable structure combinations.

	Structure		Number of Combinations
English	$C_{0\text{-}4}\ V\ C_{0\text{-}3}$		20
	V	$C_{i1}V$	$C_{i1}C_{i2}V$
	VC_{f1}	$C_{i1}VC_{f1}$	$C_{i1}C_{i2}VC_{f1}$
	$VC_{f1}C_{f2}$	$C_{i1}VC_{f1}C_{f2}$	$C_{i1}C_{i2}VC_{f1}C_{f2}$
	$VC_{f1}C_{f2}C_{f3}$	$C_{i1}VC_{f1}C_{f2}C_{f3}$	$C_{i1}C_{i2}VC_{f1}C_{f2}C_{f3}$
	$C_{i1}C_{i2}C_{i3}V$		$C_{i1}C_{i2}C_{i3}C_{i4}V$
	$C_{i1}C_{i2}C_{i3}VC_{f1}$		$C_{i1}C_{i2}C_{i3}C_{i4}VC_{f1}$
	$C_{i1}C_{i2}C_{i3}VC_{f1}C_{f2}$		$C_{i1}C_{i2}C_{i3}C_{i4}VC_{f1}C_{f2}$
	$C_{i1}C_{i2}C_{i3}VC_{f1}C_{f2}C_{f3}$		$C_{i1}C_{i2}C_{i3}C_{i4}VC_{f1}C_{f2}C_{f3}$
Thai	$C_{0\text{-}2}\ V\ C_{0\text{-}1}$		6
	V	$C_{i1}V$	$C_{i1}C_{i2}V$
	VC_{f1}	$C_{i1}VC_{f1}$	$C_{i1}C_{i2}VC_{f1}$

D. Subphonetic Models

There are a number of subphonetic models proposed and applied in many continuous speech recognition systems. In 1987, IBM first proposed the “fenones” as front-end based subphonetic units (Bahl, et al., 1993). The “shared-distribution models” was proposed and applied to the SPHINX II recognizer which was later developed to be the “senonic” models (Huang, et al., 1991; Hwang and Huang, 1992; Hwang, 1993). The shared-distribution models provide generalized triphones which acoustically similar triphones are grouped together into a single model in order to reduce the number of models. However, there are some limitations in this method that lead to over-generalization. Then, the subphonetic model, the senone, was proposed to avoid over-generalization by grouping at the subphonetic level.

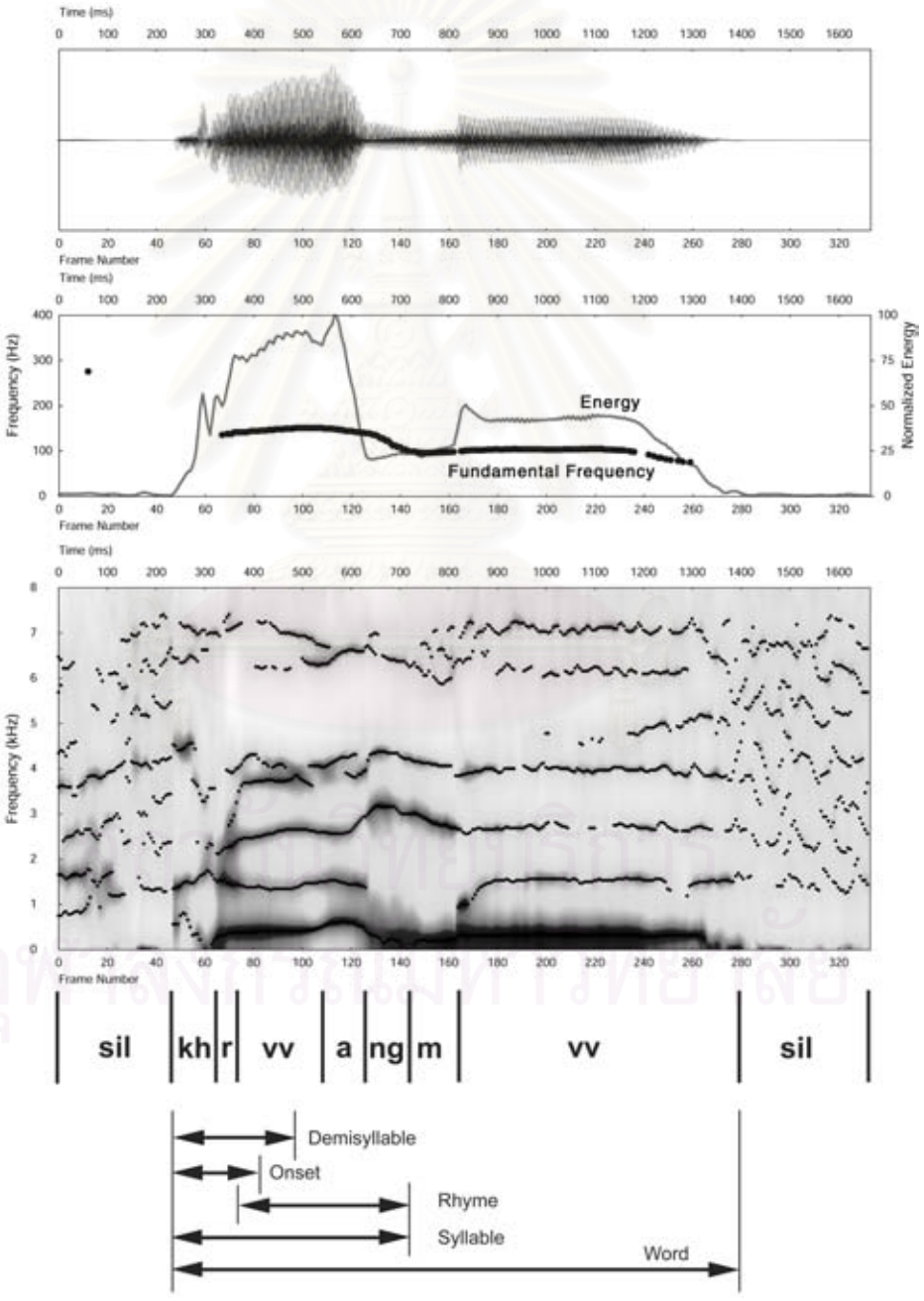


Figure 1.3 Acoustic speech units: word, syllable, demisyllable, and onset-rhyme.

E. Syllable Model

Syllable models have been employed in a syllable-based large-vocabulary continuous speech recognition system (Ganapathiraju, et al., 2001). Syllable models have provided efficient modelling of long-term temporal dependencies. The efficient modelling is resulted from longer duration of the syllable models than phones and triphons. The triphones cover a very short span of a single phone, which is difficult to cover spectral and temporal dependencies. In contrast, there are many advantages of a syllable over the phone-based acoustic units. First, acoustical characteristic of a syllable does relate to articulation and human perception since a syllable is perceptually defined. Second, a syllable acoustic unit provides compact representation of an utterance. Third, coarticulation has been integrated within a syllable acoustic unit thus makes the unit acoustically stable. Moreover, a syllable has longer duration than other units, which simultaneously combined and utilized both temporal and spectral variations. Ganapathiraju, et al. (2001) applied the syllable models to a large-vocabulary continuous speech recognition system which exceeded the performance of a comparable triphone system in both complexity and word error rate. The SWITCHBOARD (SWB) corpus was utilized in training and testing of the systems. The SWB corpus consists of 70,000 words with 9,023 distinct syllables. Using the standard SWB evaluation set, the syllable models gave out only 1% reduction in word error rate over the word-internal triphone system (Ganapathiraju, et al., 2001).

F. Demisyllable Models

The demisyllable is a half syllable unit divided at the center of the syllable nucleus. Splitting the syllable within the vowel creates an initial demisyllable and a final demisyllable (Jennings, Westaway, and Curtis, 1997). These units can be used as concatenating segments in speech synthesizer having the advantage of holding the articulatory information between the phonemes. Furthermore, a few rules were required for smoothing the concatenating segment due to the voicing effect of vowel. This advantage of handling coarticulation of demisyllable in speech synthesis has led to using this unit in speech recognition (Fujimura, Macchi, and Lovins, 1977; Fujimura and Lovins, 1978; Saravari and Satoshi, 1983; Saravari and Satoshi, 1984; Yoshida, Watanabe, and Koga, 1989; Plannerer and Ruske, 1992). Additionally, The number of demisyllable units is much smaller than word, syllable, and triphone units.

However, the demisyllable models divide a syllable at the middle of syllable into two segments. This separation results in loss of prosodic information stored within the whole syllable.

G. Initial and Final Models of Chinese

According to the Mandarin syllable structure, every syllable is a morpheme which has its own meaning, and each syllable is an open syllabic structure ending with vowel or nasal /n/ or /ng/ (Lee, 1997). Therefore, an initial followed by a final is used as the basic acoustic unit in the Mandarin speech recognition. The initial comprises the initial consonant of the syllable while the final consists of the vowel or diphthong part but including 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 tone. 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 null initial) and 53 finals which compose the whole 595 syllables set disregarding tone (Fu, Lee, and Clubb, 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 (Wang et al., 1997; Lee, 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).

Table 1.5 Numbers of the Thai onset units

	Combinations	Units
Theoretical Onset	c(33)	33
Contextual Onset	c(33) x V(9)	297
Phonotactic Onset	c(33) x V(24)	792

Table 1.6 Numbers of the Thai rhyme units

	Combinations	Units
1. Sonorant ending rhyme units		
a. Open syllable rhymes	V(9) + VV(3)	12
b. Short rhyme units with sonorant ending	(V(9) + VV(3)) x C(5)	60
Inadmissible co-occurrences:		
Round vowel units preceding labialized sonorant	(V(3) + VV(1)) x C(1)	-4*
Front vowel units preceding palatalized sonorant	(V(3) + VV(1)) x C(1)	-4*
c. Long rhyme units with sonorant ending	(V(9) + VV(3)) x C(5)	60
Inadmissible co-occurrences:		
Round vowel unit preceding labialized sonorant	(V(3) + VV(1)) x C(1)	-4*
Front vowel unit preceding palatalized sonorant	(V(3) + VV(1)) x C(1)	-4*
2. Obstruent ending rhyme units		
a. Short rhyme units with obstruent ending	(V(9) + VV(3)) x C(4)	48
b. Long rhyme units with obstruent ending	(V(9) + VV(3)) x C(3)	36
Total numbers of Thai rhyme units		**200**

* These rhyme units do not occur grammatically. They are excluded from the sets.

1.3 Onset-Rhyme Acoustic Models

From the previously used speech units, there exists some major disadvantages in applying to the Thai continuous speech recognition system. Considering all of the phone-based models, the models are inefficient in modelling of long-term temporal dependencies. Also, there are a large number of diphone and triphone models with a non-zero probability of occurrence. As a result, the triphone models are inefficient decompositional units and poorly trained (Ganapathiraju, et al., 2001). The number of diphone and triphone models in Thai are listed in Table 1.2 and 1.3. Hence, a larger acoustic unit, a syllable, is a feasible unit for representation of utterances. However, there are a large amount of syllable unit required to cover the whole language. About the demisyllable models, a syllable is divided in the middle of a syllable segment. Prosodic information resides within a syllable segment are lost by the segmentation. Therefore, a new model of acoustic speech unit is proposed, the onset-rhyme models.

The onset and rhyme are phonological units as shown in Figure 1.1. A syllable consists of an onset and a rhyme units. A rhyme unit, which carries prosody, contains nucleus and coda of a syllable. The Thai syllable structure is composed of releasing consonant (c, cc), vowel (V, VV), arresting consonant (C), and tone (T) as depicted in Figure 1.2. Considering the Thai syllable structure, the Thai syllable onset covers releasing consonant while the rhyme covers vowel and arresting consonant respectively. The proposed acoustic speech unit, the onset-rhyme models, then make use of the onset and rhyme units as described. Various speech units are illustrated in Figure 1.3 compared to the onset-rhyme models. There are some

advantages of the onset-rhyme models over other previously proposed speech units. Based on the Thai syllable structure, major advantages of the onset-rhyme models are described as follows.

Firstly, the onset-rhyme models preserve the essential prosodic information within the whole single rhyme unit. Those previously used speech units do not take this into account such as phones and demisyllables. Then, dividing a syllable into demisyllable units are not practical in modelling of acoustic speech units. The onset unit contains a releasing consonant with its transitional period towards its neighbouring vowel nucleus. The rhyme unit covers the whole vowel segment and an arresting consonant. Consequently, the models capture coarticulatory effects over a syllable within the models.

Secondly, the models are consistent in which the same models have similar characteristics across different speech instances. Thirdly, the onset-rhyme models cover a finite set of speech units, which represent all potential speech units of the language. Theoretically, the maximum number of onset-rhyme models are 992 units composed of 792 phonotactic onset units and 200 rhyme units as shown in Table 1.5 and 1.6. Whereas, the diphones and triphones have 1,650 and 59,593 units respectively. Thus, this finite number of units makes the onset-rhyme models sufficiently trained with only a small set of sentences.

Moreover, the onset-rhyme models are context-dependent where phonotactics or phonological rules are embedded into the models in forming syllables. The onset unit is right context-dependent on its adjacent rhyme unit. Whereas, the rhyme unit is left context-dependent on its preceding onset unit. As a result, the onset-rhyme models are context-dependent by nature, which helps reduce complexity of language modelling.

There are many difference between the onset-rhyme models and the initial-final model. Firstly, the initial-final models are context-independent where as the onset-rhyme models are context-dependent by nature. Secondly, the initial-final do not model releasing consonant in every possible syllable context. This issue has made the initial-final models context-independent. Thirdly, the initial-final models do not have internal and external junctures which constitute a pair of initial and final by tying both models together.

1.4 Objectives of the Dissertation

The objective of this dissertation is described as follows.

1. To develop an appropriate speech unit for modeling of Thai syllable onsets.
2. To model acoustic characteristic of Thai syllable onsets.
3. To provide basic acoustic knowledge for Thai continuous speech recognition.

1.5 Scope of the Dissertation

The scope of this dissertation is described as follows.

1. Acoustic-phonetic analysis of the Thai releasing consonants and Thai vowels in syllable onsets.
2. Collect sets of Thai continuous speech of a single speaker in “Stressed Dictation Style or Reading Style” for training and testing of the onset units.
3. Construct acoustic models of the Thai releasing consonants using the onset units.
4. Recognition of Thai releasing consonants using the onset units of the onset-rhyme models on a speaker-dependent Thai continuous speech recognition system.

1.6 Key Words

The key words of this dissertation are shown as follows.

- ACOUSTIC MODELLING
- ONSET AND RHYME
- SPEECH ANALYSIS
- THAI CONTINUOUS SPEECH RECOGNITION

1.7 The Expected Prospects

1. To acquire a knowledge base of the acoustic characteristics of Thai speech units.
2. To acquire a knowledge base of the acoustic features extracted from continuous speech waveform.
3. To provide basic acoustic-phonetic knowledge for Thai continuous speech recognition.

1.8 Research Procedures

1. Feasibility study and literature reviewing of relevant researches in both the same field and others.
2. Study acoustic properties of Thai syllable onsets in continuous speech for each consonants from recorded continuous utterances.
3. Analysis and classification of each consonants from acoustic characteristics of their syllable onsets.
4. Design sets of Thai sentences or dialogs for recording of Thai continuous speech.
5. Record continuous speech corpus from one speaker.
6. Manual labelling of recorded utterances in training databases.
7. Set up a speaker-dependent Thai continuous speech recognition system for training and testing of the models.
8. Training a recognition system using the recorded utterances of a single speaker.
9. Testing and evaluation of the recognition system and its reliability.
10. Analysis of all research results in various aspects.
11. Summarize research results to meet the objectives of this research.

1.9 Summary and Dissertation Outline

The onset-rhyme acoustic models are proposed in this dissertation for Thai continuous speech recognition. This dissertation provides basic research on acoustic modelling of Thai segmentals for continuous speech recognition. This research will focus only at the onset unit of the onset-rhyme models. The onset units cover the whole transitional stage between releasing consonant and vowel nucleus. The transition stage provide crucial acoustic cues for identifying the releasing consonant. Thus, the onset unit provide improved models of releasing consonants especially for the releasing stops. Details of the onset-rhyme models will be thoroughly described later in this dissertation.

In the next chapter, acoustic analysis of Thai utterances are described in details to provide basic acoustic knowledge of Thai language. In Chapter 3, the proposed onset-rhyme models are described in details on acoustic modelling for the Thai continuous speech recognition system using the hidden Markov models. The philosophy and methodologies of creating and using the onset-rhyme models are elaborated in this chapter. Experimental results and discussions are in Chapter 4 with comparison between the phone models and the onset-rhyme models. Finally, Chapter 5 concludes all the experiments and the brief concept of the onset-rhyme models. Contributions and future works on acoustic modelling are also discussed in Chapter 5.

CHAPTER 2

The Acoustic Analysis of Thai Utterances

In the previous chapter, various types of acoustic speech units were briefly described including the onset-rhyme models. In this chapter, acoustic-phonetic analyses are conducted on Thai utterances. The acoustic-phonetic analysis provides both basic acoustic knowledge and phonological understanding of the Thai utterances. The analysis begins from syllable structure of Thai language through its segmental components.

2.1 The Acoustic-Phonetic Analysis

The acoustic-phonetic analysis of speech is the study of acoustic and phonetic properties of speech and their relations. A number of parameters are used in analysing speech waveform, for example, fundamental frequency, formant frequency, amplitude, etc. These parameters have been used to examine speech segments of a speech waveform in order to see temporal changes in utterance. Four acoustic parameters used in acoustic-phonetic analysis are fundamental frequency, formant frequencies, amplitude or intensity, and duration. (Flanagan, 1972; Furui, 2001; Rabiner and Juang, 1993) The four acoustic parameters are employed in psycho-acoustic analysis of human perception comforming to the assumption that human speech perception is based on these parameters.

The details on acoustic studies of Thai language are described in the following section. Specific details of each acoustic parameters, fundamental frequency, formant frequencies, amplitude or intensity, and duration, will be depicted with their application in phoneme recognition.

2.1.1 Acoustic Parameters Analysis

In speech recognition by machine, an acoustic-phonetic approach is one of the recognition methods that have been successfully applied besides the pattern recognition and the artificial intelligence approaches (Rabiner and Juang, 1993). In the acoustic-phonetic approach, the machine attempts to decode the speech signal in a sequential manner based on the observed acoustic features of the signal and the known relations between acoustic features and phonetic symbols. This approach has been in-depth studied for more than four decades. This method is based on the theory of acoustic phonetics which postulates that finite and distinctive phonetic units exist in spoken language. The phonetic units are characterized by its spectrum over time, however, the coarticulation of sounds are highly variable within speakers and neighboring phonetic units. The segmentation and labeling procedure in this approach involves segmenting speech signal into discrete regions corresponding to one phonetic unit with specific acoustic properties. One or more phonetic labels are attached to each segmented region according to their phonetic properties. Then, the sequence of phonetic labels is determined to be a valid recognized word or string of words.

Four acoustic parameters, fundamental frequency, formant frequencies, amplitude or intensity, and duration, are employed as acoustic features for acoustic-phonetic speech recognition. These acoustic cues are essential features for both human perception and computer speech recognition. Basic concept of each acoustic parameters and also analysis details on these parameters has been stated in this section as follows.

2.1.2 Relations between Fundamental Frequency and Formant Frequencies

Both fundamental frequency and formant frequencies are defined from the production level, the fundamental frequency from the periodicity of vocal fold vibrations and the formant frequencies from the vocal tract resonance frequencies (Fant, 1968). The human vocal mechanism is shown in Figure 2.2. The periodicity is a basic property of a vocal cord sound source expressed by the duration T_0 of a complex voice period or by the inverse value of the voice fundamental frequency F_0 as follows (Fant, 1960).

$$F_0 = 1/T_0 \quad \text{..... (2.1)}$$

A voice source is also characterized by its spectrum envelope *S(f)* which is a specification of the amplitudes of the source harmonics as a function of their frequency. The source spectrum envelope identifies personal characteristics of the speakers which varies with voice register, fundamental pitch, and voice intensity (Fant, 1960). In Figure 2.1, a simplified source-filter decomposition of the spectrum of a two-formant voiced sound is illustrated. The waveform of the periodic airflow through the glottis is transformed into a harmonic spectrum *S(f)* which multiplied by the filter characteristics *T(f)* of vocal transmission provides the spectrum *P(f)* of the radiated vowel which is specified by its waveform.

The speech production mechanism is analytically decomposed into the source and filter components, referred to Figure 2.1. The glottis represents a high impedance termination of the vocal tract in which the voice source is defined by the pulsating airflow through the glottis, that is, the saw-toothed periodic time function as shown in Figure 2.1. The transfer functions are introduced by multiplying the amplitude of each harmonic |*S(f)*| of the source spectrum by the value of gain factor |*T(f)*| of the filter function at the frequency f as show in Eq. (2.2). The phase of each harmonic is the sum of the phase of the corresponding source harmonic and the phase of the filter function as shown in Eq. (2.3) as follows.

$$|P(f)| = |S(f)|\,|T(f)| \quad \text{..... (2.2)}$$

$$\angle_{P(f)} = \angle_{S(f)} + \angle_{T(f)} \quad \text{..... (2.3)}$$

The spectral peaks of the sound spectrum |P(f)| are called formants. In Figure 2.1, each resonance has its counterpart in a frequency region of relatively effective transmission

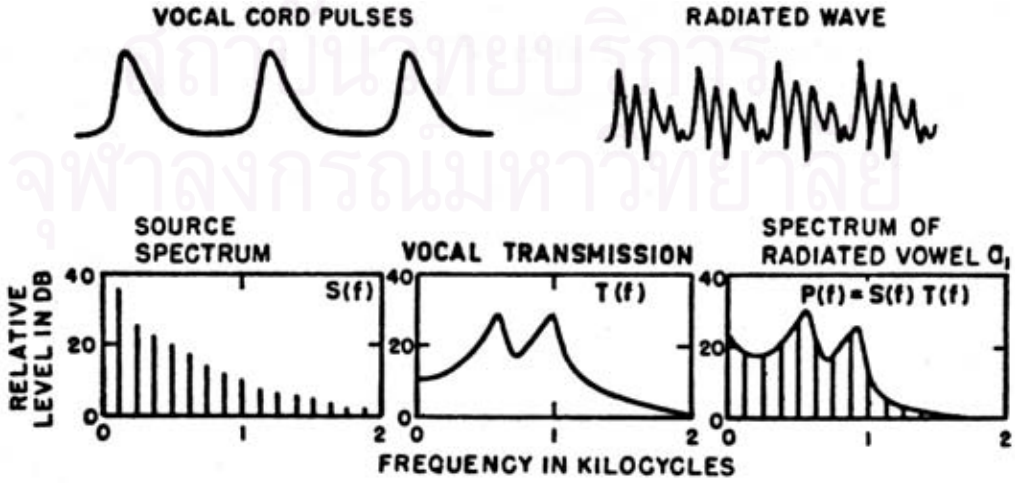


Figure 2.1 Simplified Source-Filter Decomposition of the spectrum of a two-formant voiced sound (Fant, 1960)

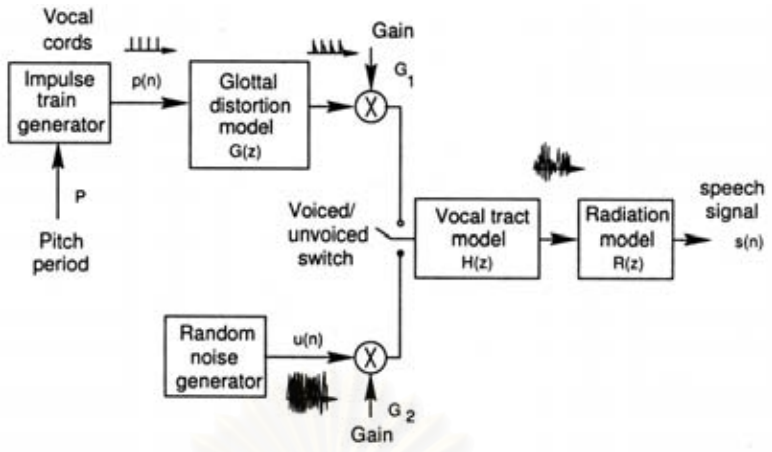


Figure 2.2 A Simple Discrete-Time Model for Speech Production (Vuuren, 1998)

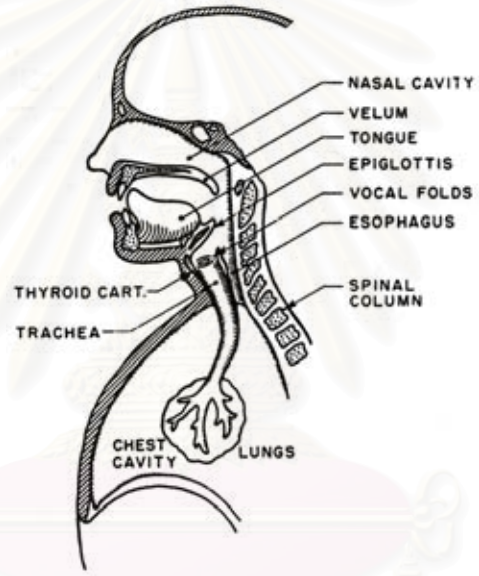


Figure 2.3 Human Vocal Mechanism (Rabiner and Juang, 1993)

through the vocal tract. This selective property of $|T(f)|$ is independent of the source and frequency location at maximum $|T(f)|$ is the resonance frequency which is corresponded to maximum $|P(f)|$ in spectrum of the complete sound. Formants are labeled, F1, F2, …, and so on, in the order of occurrence in the frequency scale. These notations refer to the frequencies of the corresponding vocal tract resonance or the frequencies of the formants. In the analysis of voiced sound, the filter function is independent of the source in a first order approximation. The formant peak coincides with the frequency of a harmonic. The formant frequencies are changed as a result of an articulatory change affecting the dimensions of the various parts of the vocal tract cavity system, that is, the filter function.

2.1.3 Fundamental Frequency

A fundamental frequency (F0) or pitch is a frequency of vocal cords vibration during speech production. A periodic speech wave has a fundamental frequency match to a vocal cord vibration which occurs in a voiced segment of an utterance, that is, a vowel. A fundamental frequency has been used for voiced-unvoiced classification. A vowel segment could be extract from a speech waveform using pitch period as shown in Figure 2.5. Fundamental

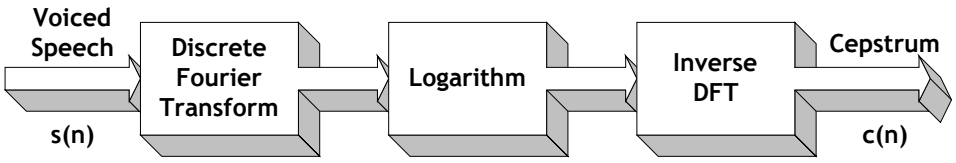


Figure 2.4 Cepstrum Analysis (Furui, 1989; Deller, Proakis, Hansen, 1993)

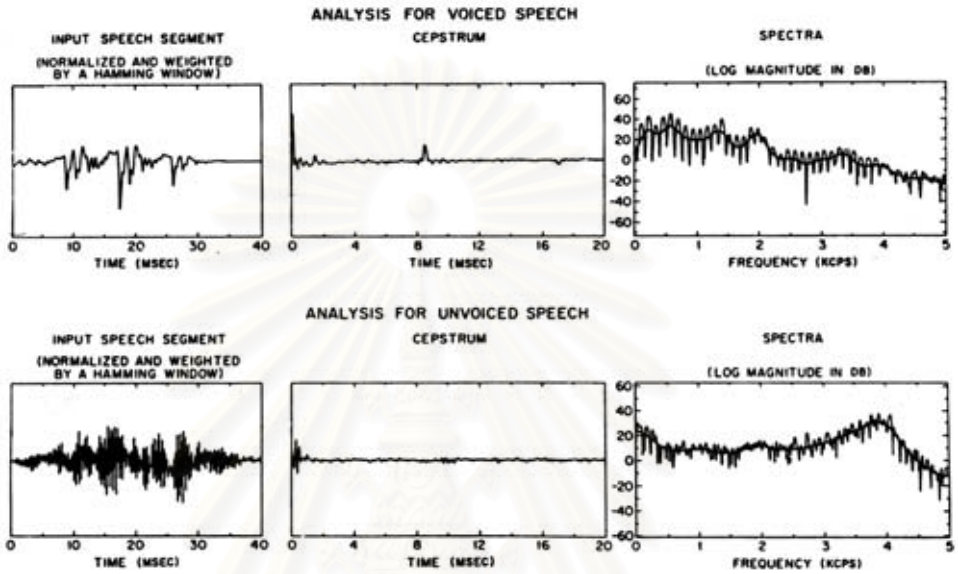


Figure 2.5 Spectrum and Cepstrum Analysis of Voiced and Unvoiced Speech Sounds (Flanagan, 1972)

frequency analysis or pitch extraction has the objective to indicate the epoch of each glottal puff and the measurement of interval between adjacent pulses (Flanagan, 1972). A pitch extraction or pitch estimation is to obtain the period of the glottal excitation waveform that is the result of the periodic opening and closure of the vocal cords in the glottis while air is forced through from the lungs and result in a train of alternating high and low pressure pulses in the vocal tracts (Vuuren, 1998). Only voiced sounds have periodic opening and closure, on the contrary, the air passes through the glottis unrestricted in unvoiced sounds.

In Figure 2.2, the glottal excitation waveform is generated in the same way as generating a voiced sound. These sequences are modified by vocal tract and other speech organs. The output speech signal is modeled as the convolution of the excitation signal with the impulse response of a filter describing the vocal tract and other speech organs. The pitch information of voiced speech is represented as quasi-periodic signal in time domain. The excitation or the vocal cords results in long periods and the resonant cavity of the vocal tract shape results in short periods (Vuuren, 1998). For automatic pitch extraction, properties of the cepstrum have been utilized to reveal signal periodicity. The cepstrum is the Fourier transform of the logarithm of the amplitude spectrum of a signal. Then, 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 the short-time logarithmic amplitude spectrum. The cepstrum analysis is illustrated in Figure 2.4. The quefrency, the independent parameter for the cepstrum, is the time domain parameter results 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

pseudoperiodic 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.

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \quad \text{..... (2.4)}$$

$$X(\omega) = G(\omega)H(\omega) \quad \text{..... (2.5)}$$

Where $X(w)$, $G(w)$, and $H(w)$ are the Fourier transform of $x(t)$, $g(t)$, and $h(t)$ respectively. By taking logarithm and inverse Fourier transform, the cepstrum $c(t)$ is as follows.

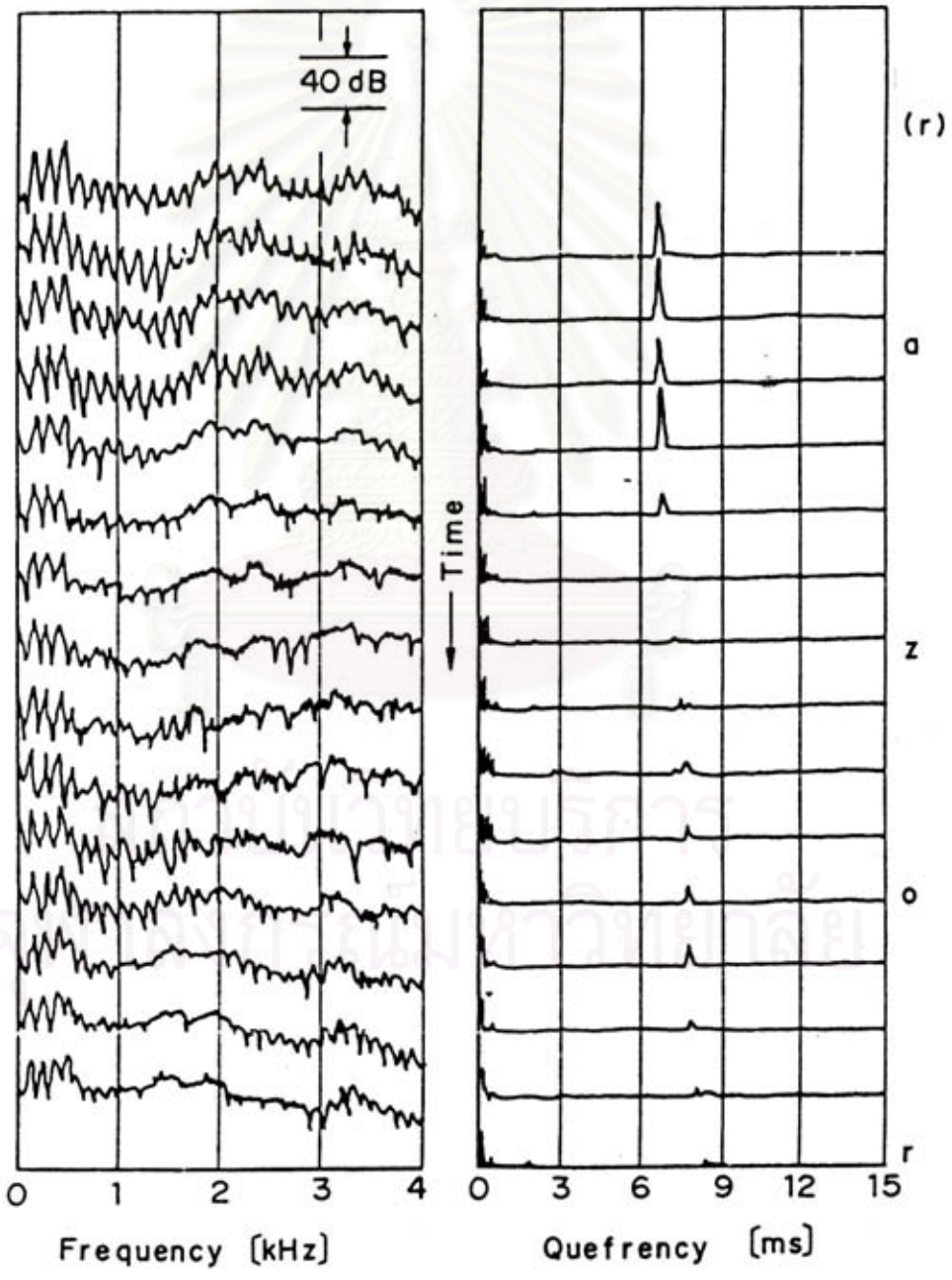


Figure 2.6 Short-time Spectra and Cepstra for male voice (Furui, 2001)

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \quad \text{..... (2.6)}$$

$$c(\tau) = \Im^{-1}\log|X(\omega)| = \Im^{-1}\log|G(\omega)| + \Im^{-1}\log|H(\omega)| \quad \text{..... (2.7)}$$

From the right side of Eq. (2.7), the first term represents the spectral fine structure or the periodic pattern and the second term represents the spectral envelope or the global pattern along the frequency axis. The fundamental period of the source *g(t)* could be extracted from

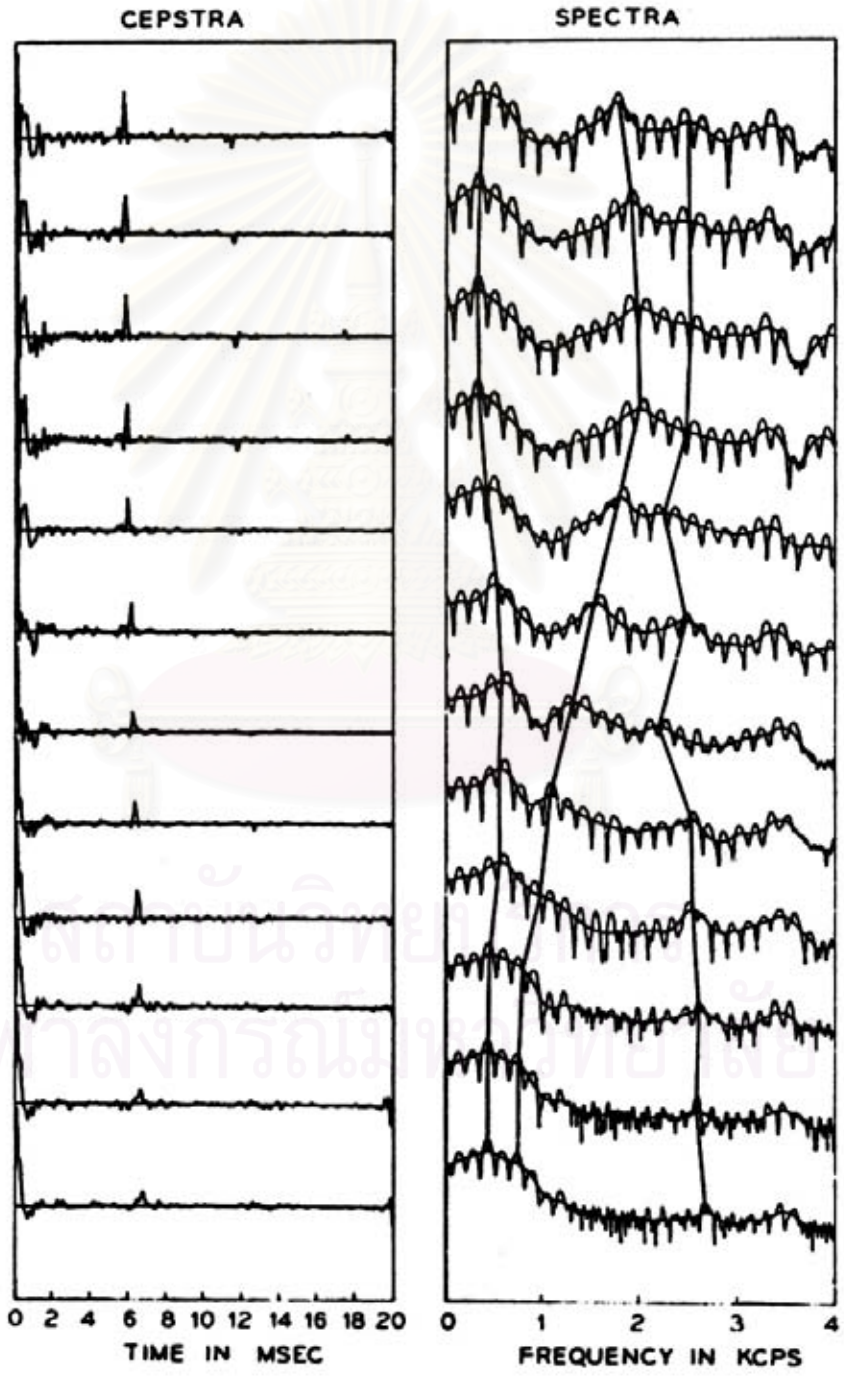


Figure 2.7 Cepstrum Analysis of Continuous Speech (Flanagan, 1972)

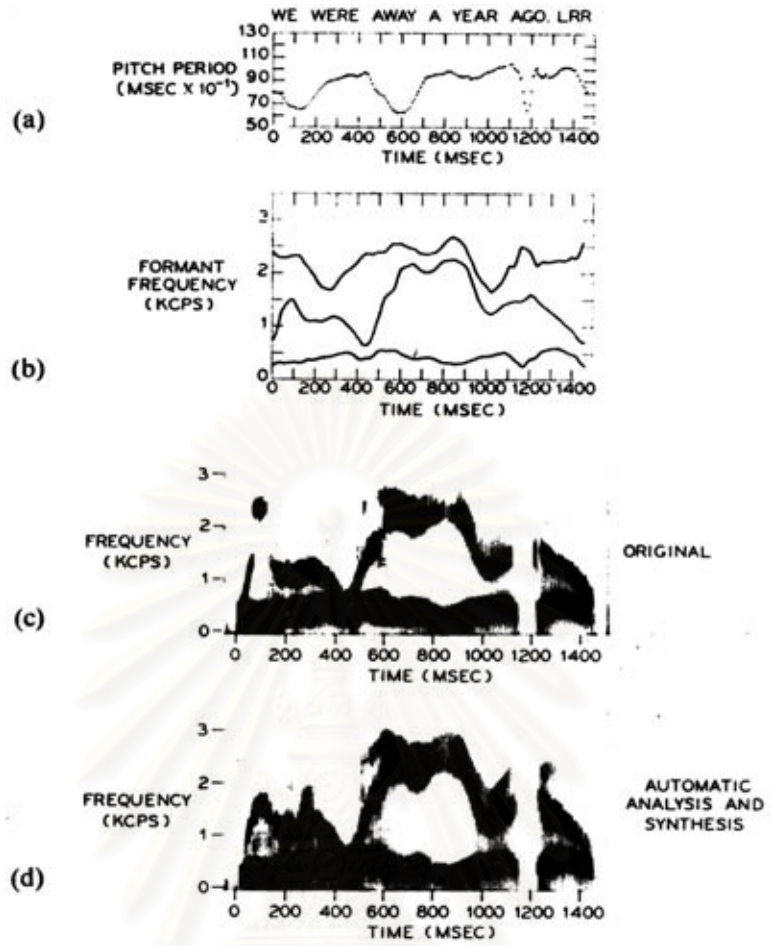


Figure 2.8 Formant Analysis and Synthesis of Speech (Flanagan, 1972)

the peak at the high-quefrency region, that is, the first term which indicates the formation of the peak in the high-quefrency region (Furui, 2001). When the cepstrum value is computed by the discrete Fourier transform (DFT), the equation is shown as follows.

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log\left|X(k)\right| e^{j2\pi kn/N}, \quad 0 \le n \le N-1 \quad \text{..... (2.8)}$$

In Figure 2.5 and 2.6, a voiced and unvoiced speech segment are analysed using spectrum and cepstrum analysis. In voiced speech, the sharp peak occurs in the cepstra plot which correspond to the period of the pitch. Unlike voiced speech, unvoiced speech cepstra has no peak which results in no fundamental frequency in that speech segment and will be classified as unvoiced. In Figure 2.6 and 2.7, the example of short-time spectra and cepstra on the left and the right respectively of male utterance in the word "razor". During the voiced speech or the vowel segment, a sharp peak occurs in quefrency domain of the corresponding cepstra in the period. The sharp peak disappears in the unvoiced speech portion. The existence of a peak during voiced speech segment of a cepstra could be used for voiced-unvoiced classification of speech. The fundamental frequency is computed directly from the location of the peak which is the reciprocal of the period. The pitch period tracking is shown in Figure 2.7 and 2.8. The fundamental frequencies of each speech segment in continuous speech are varied over time during speech production.

The fundamental frequency has been employed as an acoustic cue in many speech recognition research works. The fundamental frequency has been used to distinguish male and female speakers. Since temporal variation in fundamental frequency indicates the mean and standard deviation for females voices are roughly twice those for male voices (Furui, 2001).

For Thai language, the fundamental frequency plays an important role in tone recognition. Thai language has five tones, the mid /0/, the low /1/, the falling /2/, the high /3/, and the rising /4/, as shown in Figure 2.10. There are a number of studies in Thai tone recognition, for example, Potisuk and Harper (1995) and Thubthong (1995). Thubthong (1995) utilized the

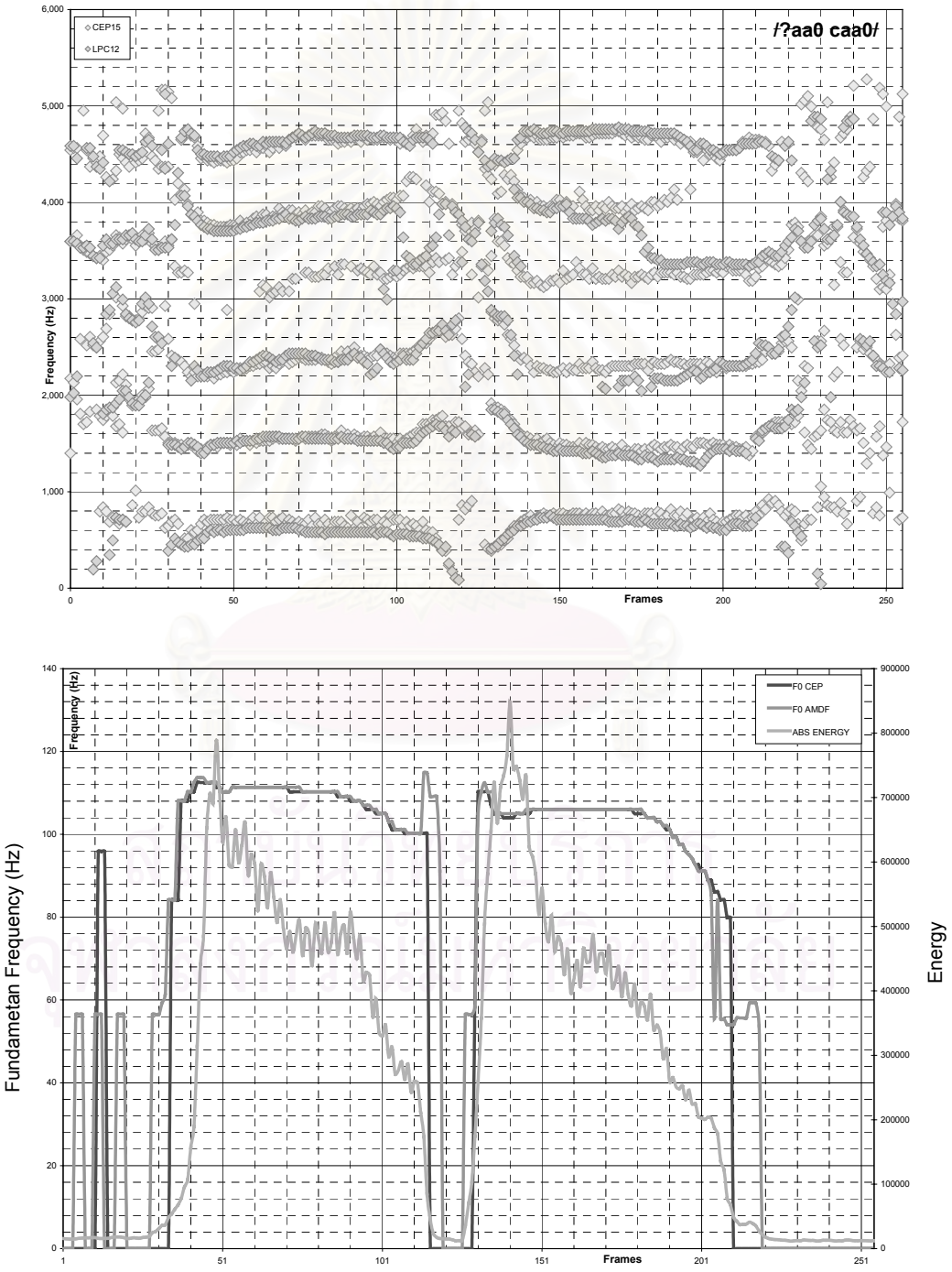


Figure 2.9 Formant Tracking and F0 Estimation of the word /zaa0 caa0/

acoustic-phonetic features, F0 direction and F0 height in tone phoneme recognition. Potisuk and Harper (1995) applied the analysis-synthesis method based on an extension to the Fujisaki model.

2.1.4 Formant Frequencies

Formant frequencies or formants is the resonance frequency of the vocal tract tube in which depend upon the shape and dimension of the vocal tract. The shape of the vocal tract is characterized by a set of formant frequencies. Different sounds are formed by varying the shape of the vocal tract. Then, the spectral properties of the speech signal vary with time as the vocal tract shape varies (Rabiner and Schafer, 1978). Formant frequencies are the dominant frequency components which characterize the phonemes corresponding to the resonant frequency components of the vocal tract (Furui, 2001).

Resonances of the vocal tract are called formants and their frequencies called formant frequencies (Denes and Pinson, 1963). The vocal tract is an air-filled tube that acts as a resonator and has certain natural frequencies of vibration. The vocal resonator emphasize the harmonics of the vocal cord wave at a number of different frequencies and the spectrum of the speech wave will have a peak for each of the natural frequencies of the vocal tract. The value of the natural frequencies of the vocal tract is determined by its shape (Denes and Pinson, 1963). Every vocal tract configuration has unique set of characteristic formant frequencies. The lowest formant frequency is called the first formant (F1). The next highest frequency is called the second formant (F2) and so on.

The formant frequencies are estimated from the spectrum of each speech segment using the Fourier transform. Tracking of the formant frequencies in each speech segment is called formant tracking as shown in Figure 2.7, 2.8, and 2.9. The tracked formants reveal a time-varying property of the vocal tract during speech production which is essential for phoneme recognition. The tracked formants correspond to the frequency peak spectrogram.

The formant frequencies have been utilized as an acoustic cues for phoneme recognition. The first (F1), second (F2), and third formant (F3) have been used to identify vowel phonemes. For Thai language, formant frequencies and formant transition have been used in vowel and consonantal phonemes recognition (Trongdee, 1987; Tarnsakun, 1988; Thubthong, 1995). Trongdee (1987) employed the first, second, and third formant transition to classify stop consonants. Tarnsakun (1988) utilized first and second formant transition in both pre-consonantal and post-consonantal transition to classify non-stop consonants. Thubthong (1995) used pre-consonantal second formant transition with other acoustic features for consonantal phonemes classification.

2.1.5 Amplitude or Intensity

An amplitude of a speech wave is a peak of a speech waveform. In other words, an amplitude is a maximum displacement of vibration of a mass which is displaced from its rest position and moving back and forth between two positions that mark the extreme limits of its motion (Denes and Pinson, 1963). Human perceives sound intensity rather than amplitude of speech wave. The intensity of the sound wave is a power transmitted along the wave through an area of one square centimeter orthogonal to the direction of the sound wave which is the energy available over a small area at the point of measurement (Denes and Pinson, 1963). A sound intensity is measured in watts per square centimeter or in the decibel scale. In speech recognition, an absolute acoustic energy contour could be computed directly from a speech wave using the following relation as shown in Eq. (2.9). In Eq. (2.9), $E(m)$ is an absolute energy value of the m^{th} frame, $s(n)$ is an amplitude of the n^{th} sample, N is the total samples,

$$E(m) = \sum_{n=0}^{N-1} s(n) \quad \text{..... (9)}$$

An intensity is one of the acoustic cues that is used to classify Thai consonants (Trongdee, 1987; Tarnsakun, 1988; Thuthong, 1995). Both th acoustic energy and the intensity of each formant frequency have been used in the classification process. Trongdee (1987) employed intensity of first (F1) and second (F2) formants to categorize Thai non-stop consonants with different manner of articulation. The nasals have low second formant intensity while both trill and lateral has high first and second formant intensity.

Tarnsakun (1988) used intensity to classify the ten Thai stop consonants in both manners and places of articulation. Intensities of different pace of articulation of the stop consonants in releasing phase are ranging from labial, alveolar, alveolar-palatal, and velar. The intensity of aspirated stops is higher than unaspirated stops, same as voiced stops and voiceless stops. The intervocalic non-stop consonants have the highest intensity compared to final and initial respectively (Luksaneeyanawin, 1993).

2.1.6 Duration

Duration is one of the four acoustic cues that have been used in Thai phoneme classification in both vowels and consonants. (Trongdee, 1987; Tarnsakun, 1988; Thubthong, 1995) In Thai vowel phonemes classification, vowel duration is computed from a period of the fundamental frequency or a pitch period of that vowel to classify into short or long vowel as shown in Table 2.1. In Thai consonantal phonemes classification, a duration of marginal sound was employed to categorize each phoneme with the same manner of articulation into appropriate place of articulation as shown in Table 2.2. Trongdee (1987) applied duration to classify non-stop consonants in which duration of each consonant are varied in different structural context, initial, intervocalic, and final.

Tarnsakul (1988) employed duration of three phases of stop consonants, shutting, closure, and releasing, to classify each stop consonants in both manners and places of articulation as shown in Table 1. Voiceless stops have longer duration than the voiced stops and the voiceless aspirated stops have longer duration than the voiceless unaspirated stops. The voiceless plosives, initial consonants /p-, t-, k-/, have longer duration distinctively from the voiceless non-plosives, final consonants /-p, -t, -k/. Thubthong (1995) used a noise duration and burst duration as acoustic parameters to determine the consonants /c/ and /p/ respectively from others.

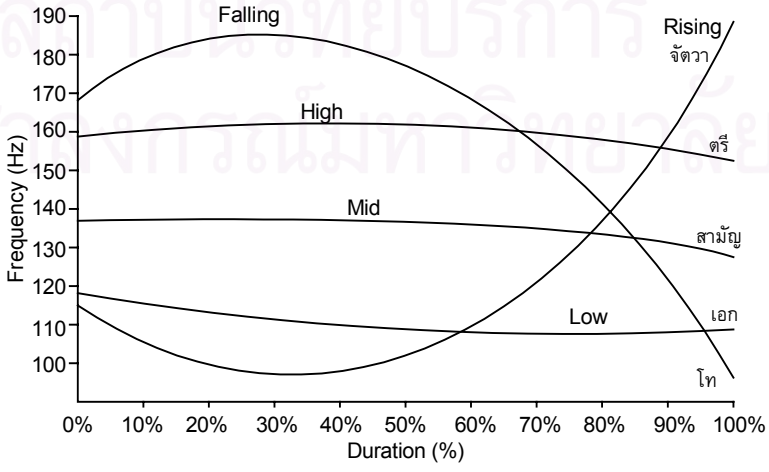


Figure 2.10 Five Thai Tones (Luksaneeyanawin, 1993; Thubthong, 1996)

2.2 Examples of Acoustic Parameters Computation

The absolute energy contour, the fundamental frequency analysis, and the formant tracking of Thai speech /zaa0 caa0/ are shown in Figure 2.9 respectively. The computation of these parameters are shown in this section.

Thai language is a tonal language in which its syllable structure is associated with tones. The syllable structure of Thai language comprises initial consonant (C) or initial consonant cluster (CC), vowel in monophthong (V) or diphthong (VV), final consonant (C), and tones (T) as shown in Figure 1.2. In Thai language, there are 18 monophthongs in short and long pairs with 6 diphthongs in the Thai vowel inventory as shown in Table 2.1. There are 21 consonantal phonemes composed of 11 stops and 10 non-stops as shown in Table 2.2 (Luksaneeyanawin, 1993). Five Thai tones are mid, low, falling, high, and rising respectively as shown in Figure 2.10.

There are various researches on acoustic of Thai segmental units such as Trongdee (1987), Tarnsakun (1988), Leelasiriwong (1991), and Sriraksa (1995). Trongdee (1987) studied the acoustic characteristics of ten Thai non-stop consonants within context of three different vowels, /ii/, /aa/, and /uu/ and also within different structural contexts, initial, intervocalic, and final consonant. Five different classes of consonants, nasals, fricatives, trill, lateral, and approximant, were studied. The other ten Thai stop consonants in three classes, voiceless unaspirated stops, voiceless aspirated stops, and voiced stops, were studied by Tarnsakun (1988) using the same phonetic context scheme as in Trongdee (1987). These two studies thoroughly explored the acoustic characteristics of Thai consonantal phonemes by spectrographic analysis using acoustic parameters, i.e., formant frequencies, formant transition, intensity, duration, etc.

In Figure 2.5 to 2.7, an analysis of a voiced speech segment results in cepstrum and spectrum envelope using the cepstral analysis and discrete Fourier transform have been shown respectively. The cepstrum could be computed using the Eq. (2.4) to (2.8) together with the Figure 2.6 and 2.7 respectively in the previous section. The linear predictive coding (LPC) and the discrete Fourier transform as shown in the Eq. (2.11) have been utilized in spectrum envelope computation. Then, a peak-picking algorithm are employed to pick the spectral peak in the envelope corresponding to formant peaks.

In Figure 2.9, the upper figure is a speech waveform of a word /zaa0 caa0/ recorded using 11.025 KHz sampling frequency, in other words, 11,025 samples in one second. The lower figure is the computed absolute energy, fundamental frequency using cepstral analysis, and fundamental frequency using AMDF analysis as shown in Figure 2.9. The absolute energy value is computed from each 256-sample speech segment using the Eq. (2.10) in the previous section. Also, the fundamental frequency is computed using the cepstral analysis as shown in the previous section and the next section.

2.2.1 Fundamental Frequency Estimation and Tracking

On pitch or fundamental frequency estimation, the cepstral analysis has been employed to separate two convolutionally related properties by transforming the relationship into summation as depicted the previous section and in Figure 2.5. The high quefrency elements are selected to estimate the fundamental frequency of each speech segment. The discrete Fourier transform and the cepstral analysis of the speech segment is shown in Eq. (2.10) and Eq. (2.11) respectively where N is the number of analysis samples.

DFT :
$$X(\omega) = \frac{1}{2\pi N} \sum_{n=0}^{N-1} x(n) e^{-j\omega n}, \qquad \omega = 2\pi f T \quad \text{..... (2.10)}$$

Cepstrum :
$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{j 2\pi k n / N}, \quad 0 \le n \le N-1 \quad \text{..... (2.11)}$$

Table 2.1 Thai vowel system

		Vowel Advancement		
		Front	Central	Back
Vowel Height	High	/i, ii/	/v, vv/	/u, uu/
	Mid	/e, ee/	/q, qq/	/o, oo/
	Low	/x, xx/	/a, aa/	/@, @@/
	Diphthongs	/ia, iia/	/va, vva/	/ua, uua/

Table 2.2 Thai consonants arranged by places of articulation

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

Table 2.3 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/

The result of cepstral analysis on voiced speech segment is shown in Figure 2.5 to 2.7. A voice speech segment is selected from the vowel /aa/ of the word /zaa0 caa0/. In Figure 2.5 to 2.7, there is an explicit peak in the cepstrum plot. The period of the peak in the quefrency domain is correspond to the fundamental period of the glottal excitation. The fundamental frequency value (F0) could be computed using the following equation where F_s is the sampling frequency and L is the period of the cepstral peak in its quefrency domain.

$$F_0 = \frac{F_s}{L} \quad \text{.....(2.12)}$$

For example, the cepstral peak period is at 103 points in the quefrency domain and the speech sampling frequency is 11,025 Hz, then, the fundamental frequency value computed

using the equation 12 results in 11025/103 = 107.039 Hz. For pitch period tracking, the previous procedure is repeated on entire speech segments to obtain the tracking of the pitch as shown in Figure 2.9.

2.2.2 Formant Frequencies Tracking

On formant frequencies estimation, a spectrum envelope of a speech segment is tracked to find a spectral peak as shown in Figure 2.7 using simple peak-picking analysis. The lowest spectral peak is picked and marked as the first formant or F1. The following picked spectral peaks are marked respectively as the second (F2), the third (F3), the fourth formant (F4), and so on.

In order to obtain a spectrum envelope of the power spectrum of each speech segment, the linear predictive coding (LPC) coefficients have been analysed on the speech segment using the Levinson-Durbin recursive algorithm (Rabiner and Juang, 1993; Deller, Proakis, and Hansen, 1993; Furui, 2001). The obtained LPC coefficients, a_0, a_1,…, a_k, …, a_p ,are coefficients of the all-pole filter with the form as follows where p is the number of coefficients of the LPC order.

$$H(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \quad \text{.....(2.13)}$$

The spectrum envelope could be obtained by taking discrete Fourier transform to evaluate $H(e^{jw})$. The spectrum envelope of a voiced speech segment is computed using 12-order LPC coefficients.

2.2.3 Intensity and Duration

The intensity has been employed not only in speech segmentation but also in discrimination of aspirated and unaspirated consonants. The unaspirated /c/ in /zaa0 caa0/ as shown in Figure 2.9 has acoustic silence during vowel-consonantal transition. Unlike the unaspirated /c/, the aspirated /ch/ in /zaa0 chaa0/ produces instantaneous burst due to aspiration during vowel-consonantal transition which occurs explicitly in the energy contour.

The duration are also utilized in discrimination of aspirated and unaspirated consonants besides of the intensity. From the pitch tracking of the word /zaa0 caa0/ compared to the word /zaa0 chaa0/, the duration of the vowel-consonantal transition of the aspirated /ch/ is longer than the unaspirated /c/. This is because of longer duration in releasing phase of an aspirated consonant compared to an unaspirated consonant. From analysis, the duration in vowel-consonantal transition of /c/ and /ch/ are 81.28 ms and 104.49 ms respectively which could be computed directly from the pitch contour.

2.3 Acoustic-Phonetic Analysis on Thai Utterances

The acoustic-phonetic analysis of speech is the study of acoustic and phonetic properties of speech and their relations. A number of parameters are used in analyzing speech waveform, for example, fundamental frequency, formant frequency, amplitude, etc. These parameters have been used to examine speech segments of a speech waveform in order to see temporal changes in utterance. Four acoustic parameters used in acoustic-phonetic analysis are fundamental frequency, formant frequencies, amplitude or intensity, and duration. (Flanagan, 1972; Furui, 2001; Rabiner and Juang, 1993) The four acoustic parameters have been employed in psycho-acoustic analysis of human perception comforming to the assumption that human speech perception is based on these parameters.

The details on acoustic studies of Thai language are described in this section. Specific details of each acoustic parameters, fundamental frequency, formant frequencies, amplitude or intensity, and duration, are depicted with their applications in phoneme recognition.

2.3.1 The Thai Syllables

Phonologically speaking, a syllable is composed of onset and rhyme units where the rhyme comprises nucleus and coda as illustrated in Figure 1.1 and 1.2 in the previous chapter. In Thai syllable, the onset is a releasing consonant (c, cc) while the rhyme contains both vowel and an arresting consonant, (V, V:, VV, VC, V:C, VVC). In acoustic-phonetic analysis, a syllable comprises a nucleus and its marginal sounds. A nucleus of a syllable is vowel (V, V:, VV) in Thai syllable structure. Marginal sounds are a releasing consonant, (c, cc), as left marginal sound and an arresting consonant (C) as right marginal sound of the nucleus respectively.

The Thai language has simple syllable structure as depicted in Figure 2.1 and 2.2. In comparison to the English syllable structure, the Thai syllable structure has only a small amount of syllable combinations while the English syllables are much longer with plenty of clusters as shown in Table 1.4 in Chapter 1. In Table 1.4, the English syllables have much more consonant clusters in both releasing and arrest consonants than the Thai syllables. In consequence, the diphone and triphone models are more practical to the English and the Thai syllable systems as speech units for recognition. This is resulted from complexity of syllable structure in English which contains many clusters. In Thai, a triphone model is equivalent to word model in syllable structure aspect which is not considered as a subword model. Phonologically speaking, for the Thai syllable, this model does not provide any difference over a word model in recognition. The phonological structure of the Thai syllable is shown in Figure 1.2 (Luksaneeyanawin, 1993). However, application of onset-rhyme models in English might cause a large number of both onset and rhyme units. This is resulted from a large number of clusters in English as shown Table 1.4 in which combinations of English syllables are more complicated than Thai.

2.3.2 The Syllable Nucleus—Vowels

An acoustic-phonetic analysis was thoroughly conducted on the Thai vowel system. The Thai vowel system consists of 18 monophthongs in short and long pairs along with 6 diphthongs in short and long pairs as shown in Table 2.1 (Luksaneeyanawin, 1993). In Table 2.1, the Thai vowels are grouped together according to their acoustic characteristics into front, central and back vowel groups by vowel advancement. Also, the Thai vowels are grouped by vowel height into high, mid, low vowel groups.

Articulatorily, these acoustic characteristics are directly related to a speech articulator or a speech production organ, in this case, tongue. The vowel height is height of tongue in high position close to palatal producing small opening cavity, then, mid and low have larger opening respectively. The vowel advancement is position of tonge where front position is close to alveolar producing larger cavity volume, then central and back have smaller cavity volume respectively. This can be illustrated by the human vocal mechanism with human speech production organs as shown in Figure 2.3.

Acoustically, the vowel advancement is represented by the second formant frequency (F2) of a vowel. The vowel height is represented by the first formant frequency (F1) of a vowel. Consequently, the Thai vowel distribution in F2 and F1 plane is shown in Figure 2.11 (Ahkuputra, et al., 2000). In Figure 2.11, normal distribution contour of each Thai monophthong are illustrated in grouping by vowel advancement into front, central, and back with their normal distribution. Ahkuputra, et al. (2000) conducted acoustic analysis and classification of individual Thai monophthong using Bayesian classifier. Three classification schemes were proposed, namely, classification by vowel height, classification by vowel advancement, and classification by combined vowel height and vowel advancement respectively. The results show the use of acoustic-phonetic features, F1 and F2, in vowel identification with high accuracy. In Figure 2.12, three dimensional distribution of Thai vowels

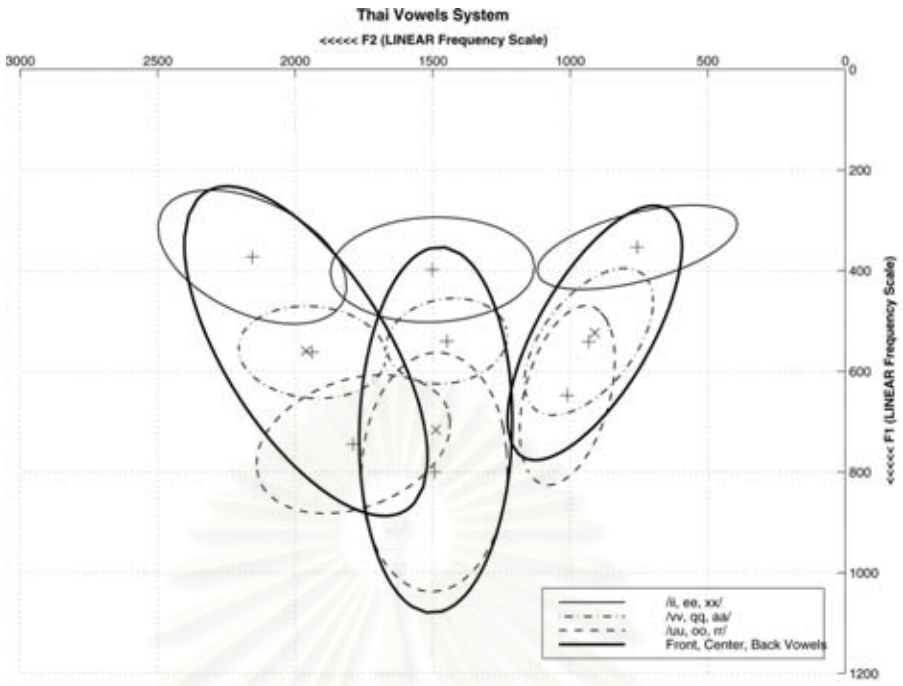


Figure 2.11 Thai vowel distribution on linear F2 and F1 plane

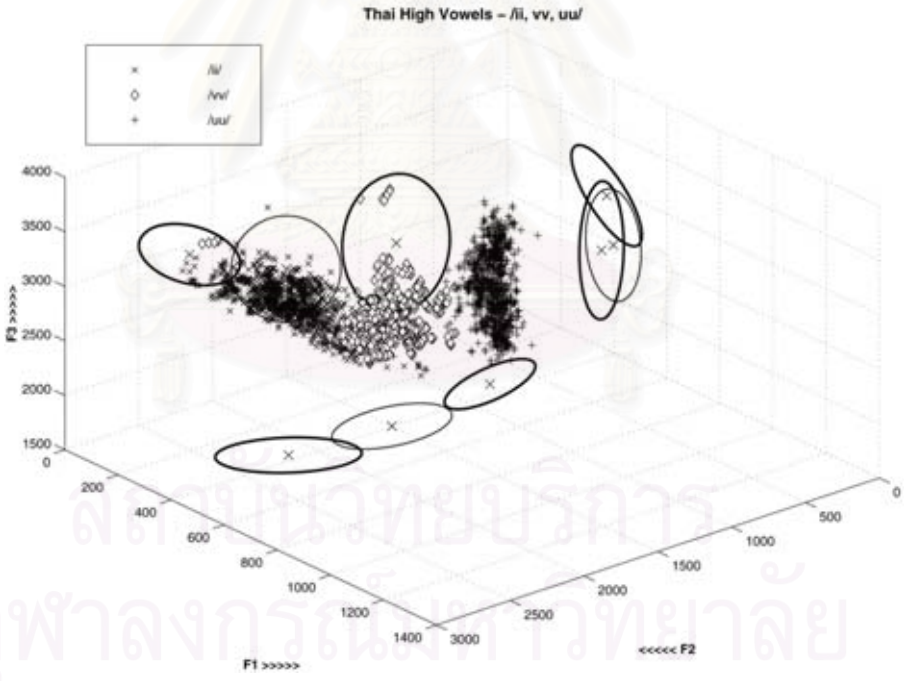


Figure 2.12 Distribution of the Thai high vowels /ii,vv,uu/ on linear F2, F1, and F3 planes

are depicted using F1, F2, and F3. The third formant (F3) represent vowel rounding or the degree of roundness in lip opening.

The Thai vowel system has complete combination of both places and manners of articulation as shown in Table 2.1. The vowels can be grouped by places of articulation using vowel advancement into front, central, and back vowel groups. Also, they can be grouped by

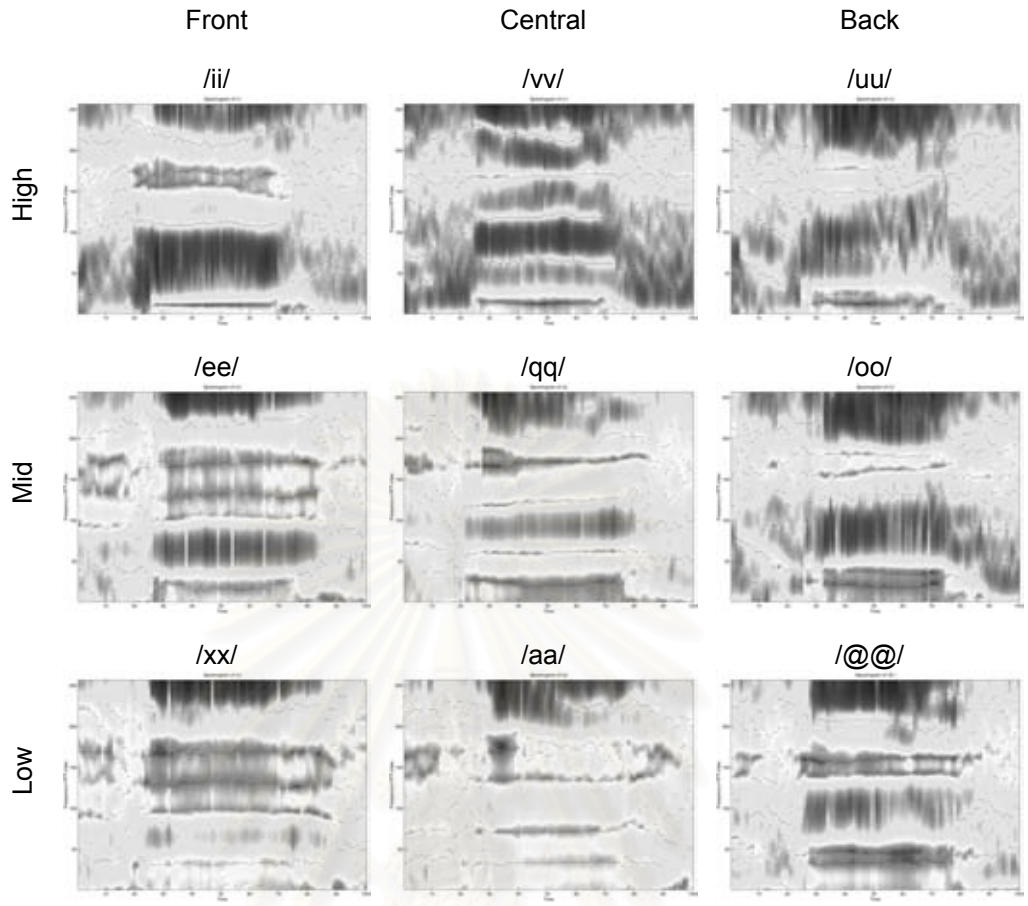


Figure 2.13 Spectrographic Illustration of the Thai vowel system from acoustic analysis

manners of articulation using vowel height into high, mid, and low vowel groups. From acoustic-phonetic analysis, the first formant (F1) represents vowel height and the second formant (F2) represents vowel advancement. Then, the high vowel group has the highest F1 value than the mid and low respectively. The front vowel group has the highest F2 value than the central and back respectively. These relation are shown in the vowel distribution in Figure 2.11, 2.12, and 2.13.

Each of the Thai vowel are acoustically analysed to explore its acoustic characteristics. Spectrographic information of each vowel are illustrated in Figure 2.13 from the acoustic analysis. In Figure 2.13, each Thai vowel shows its unique acoustic-phonetic characteristics in the formants. The front vowels have the highest second formant (F2) followed by central and back vowels respectively. The low vowels have the highest first formant (F1) followed by mid and high vowels respectively. These characteristics correspond to the vowel distribution as depicted in Figure 2.11 and 2.12. The vowel triangle, /ii/, /uu/, /aa/, show distinct characteristics between each other. The vowel triangle is the common set of vowels existed in every language in the world. Then, the vowel triangle are used in analysis of marginal sounds later in the next section.

2.3.3 Marginal Sounds of the Syllable—Consonants

Acoustically, marginal sounds are attached along both sides of the syllable nucleus. Considering the Thai syllable structure, the left marginal sound is a releasing consonant (cc) and the right marginal sound is an arresting consonant (C) relative to the nucleus as depicted in Figure 1.1 and 1.2 in Chapter 1. In consequence, from acoustic analysis, the transitional

period existed between marginal sounds and nucleus has provided crucial acoustic information. These essential acoustic cues have been utilized in identification of consonants.

The Thai consonant system is shown in Table 2.2 and 2.3 arranged by places of articulation. There are 33 consonants composed of 21 consonants and 12 consonant clusters. in Thai as shown in Table 2.2 and 2.3. All of the 33 consonants are releasing consonant but only 8 consonants, /p, t, k, m, n, ng, w, j/, are both releasing and arresting consonants.

Examples of releasing and arresting consonants are shown in Figure 2.14 to 2.16. In Figure 2.14, spectrographic information of the words /paa0/, /taa0/, and /kaa0/ are illustrated. The releasing consonant of each word has different manner of articulation but the same place of articulation. The transition period between releasing consonant and its vowel nucleus clearly differs according to the locus of each consonant. Thus, the transition period contains crucial acoustic cues in identification of releasing consonant

In Figure 2.15, spectrographic information of the words /pii0/, /paa0/, and /puu0/ are shown. Each word has the same releasing consonant but with different vowel context. The figure illustrates variation in context. The transition period of each word is changed according to the vowel context. However, formant transition of each vowel is moving towards the same locus of the releasing consonant In Figure 2.16, spectrographic information of the words /paa0/, /phaa0/, and /baa0/ are shown. Each word has releasing consonant with different manners of articulation. The /p/ and /ph/ are unaspirated and aspirated voiceless stops respectively. The /b/ is voiced stop. Each of the three stops occur in the same vowel context.

In Figure 2.17, spectrographic information of the words /sii4/, /saa4/, and /suu4/ are shown. The /s/ is a fricative but occurs in different vowel context. In Figure 2.18, spectrographic information of the words /kok1/, /kot1/, and /kop1/ are shown. Each word has different arresting stop consonants, /-k/, /-t/, and /-p/ respectively.

2.4 Summary

In this chapter, Thai utterances are acoustically analysed which provide basic knowledge and understanding of Thai utterances. Acoustic-phonetic analysis are thoroughly conducted on Thai utterances. Characteristics of the vowels and marginal sounds are explored in the analyses. The outcome of vowel analysis not only provide solid acoustical background of Thai utterance but also provide acoustic cues for classification of Thai vowels. Details about classification of Thai vowels was written in full article as described in Appendix B

The results of analysis on Thai utterances provide basic acoustic knowledge and understanding of their characteristics. This also provide solid background for modelling of the onset-rhyme models. Acoustic modelling of the onset-rhyme models is described in details in the following chapter.

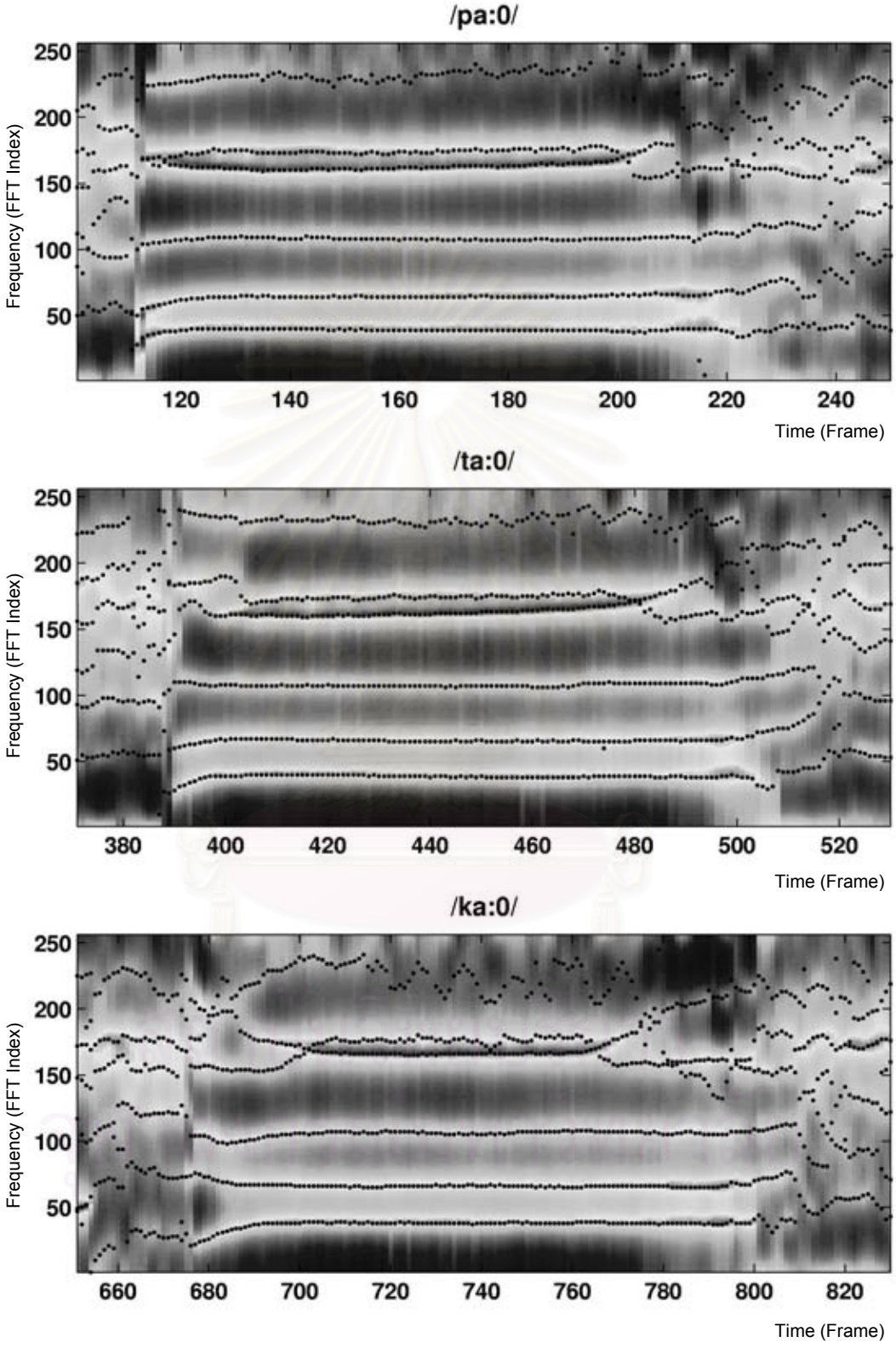


Figure 2.14 Spectograms of the words /paa0/, /taa0/, /kaa0/

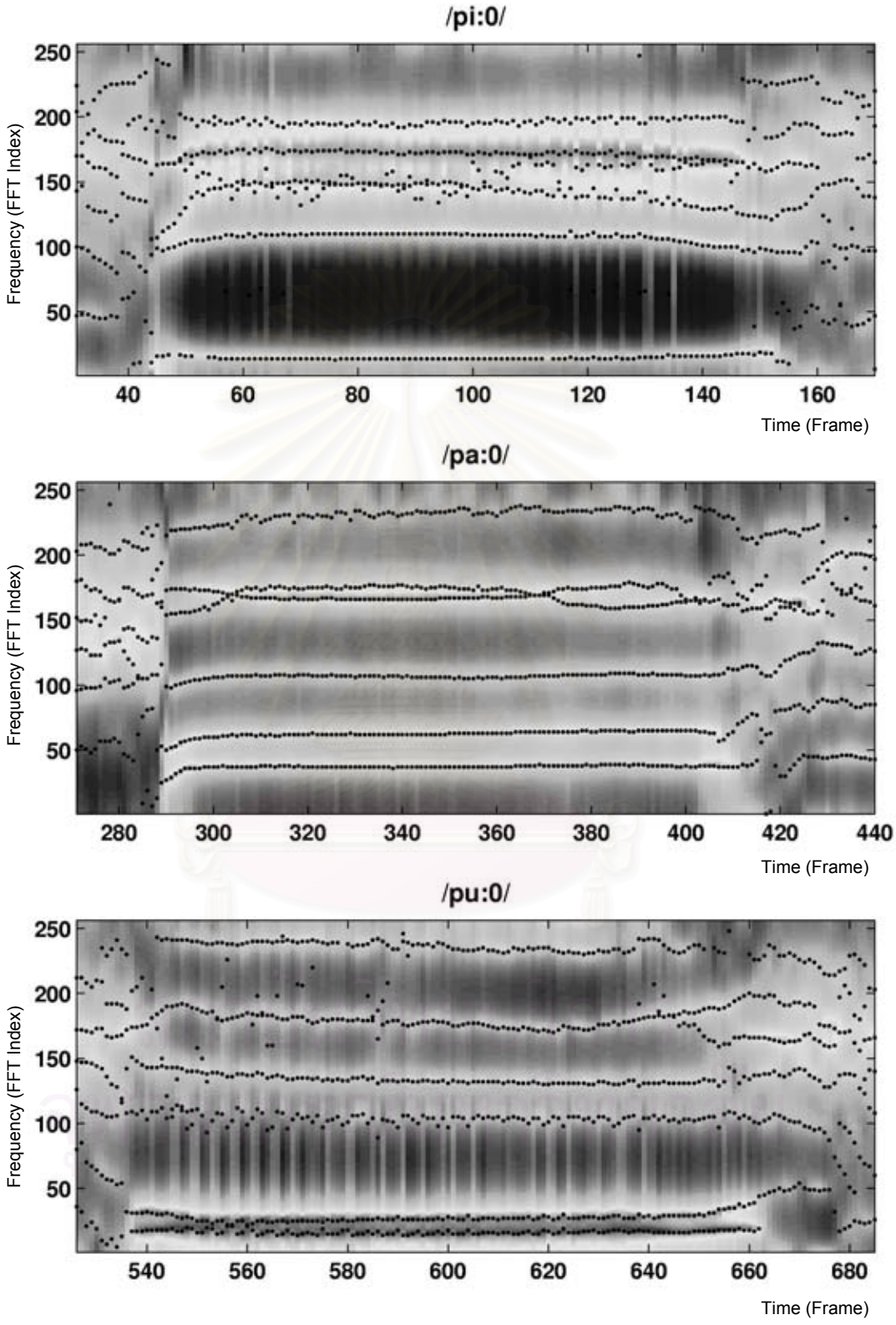


Figure 2.15 Spectograms of the words /pii0/, /paa0/, /puu0/

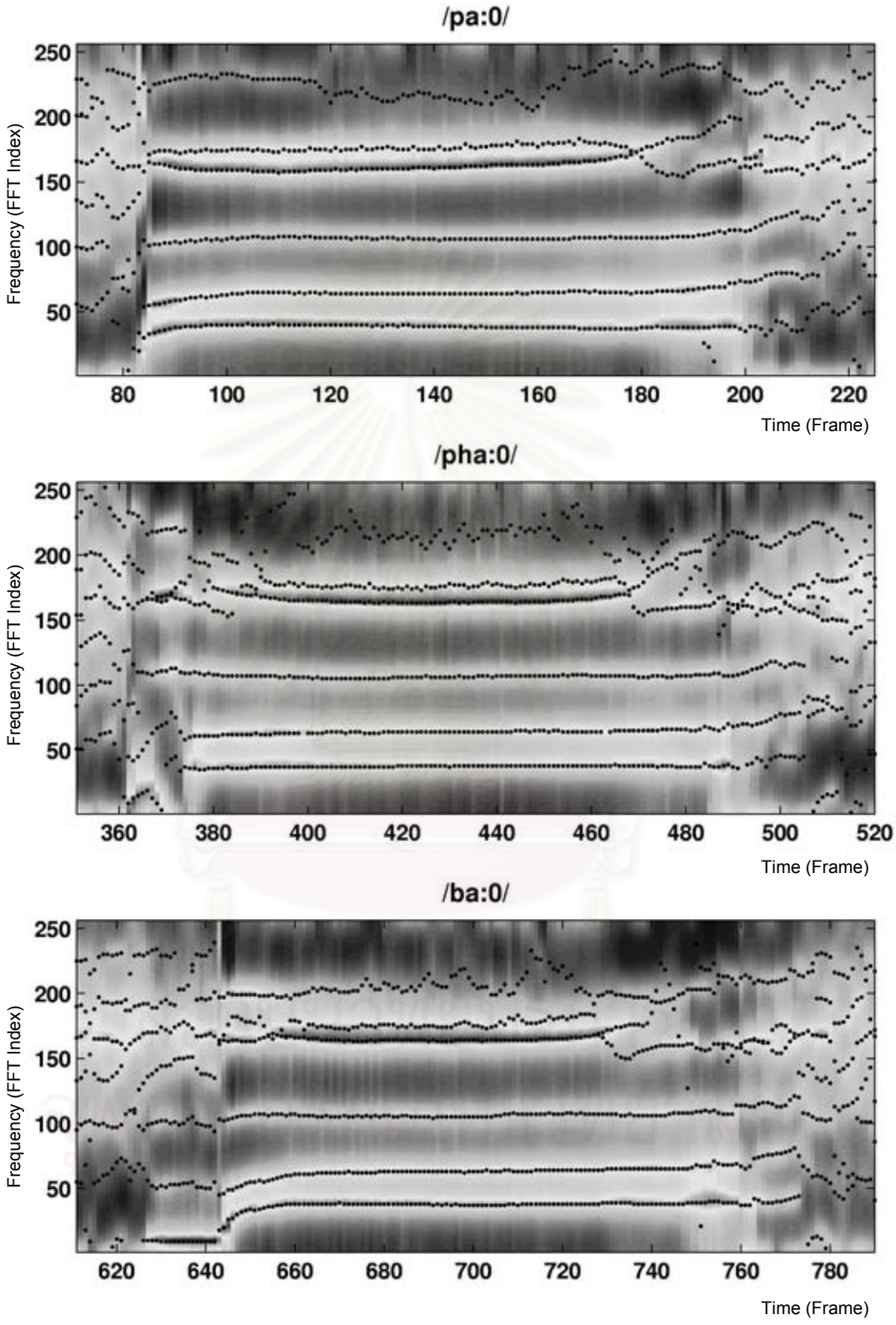


Figure 2.16 Spectograms of the words /paa0/, /phaa0/, /baa0/

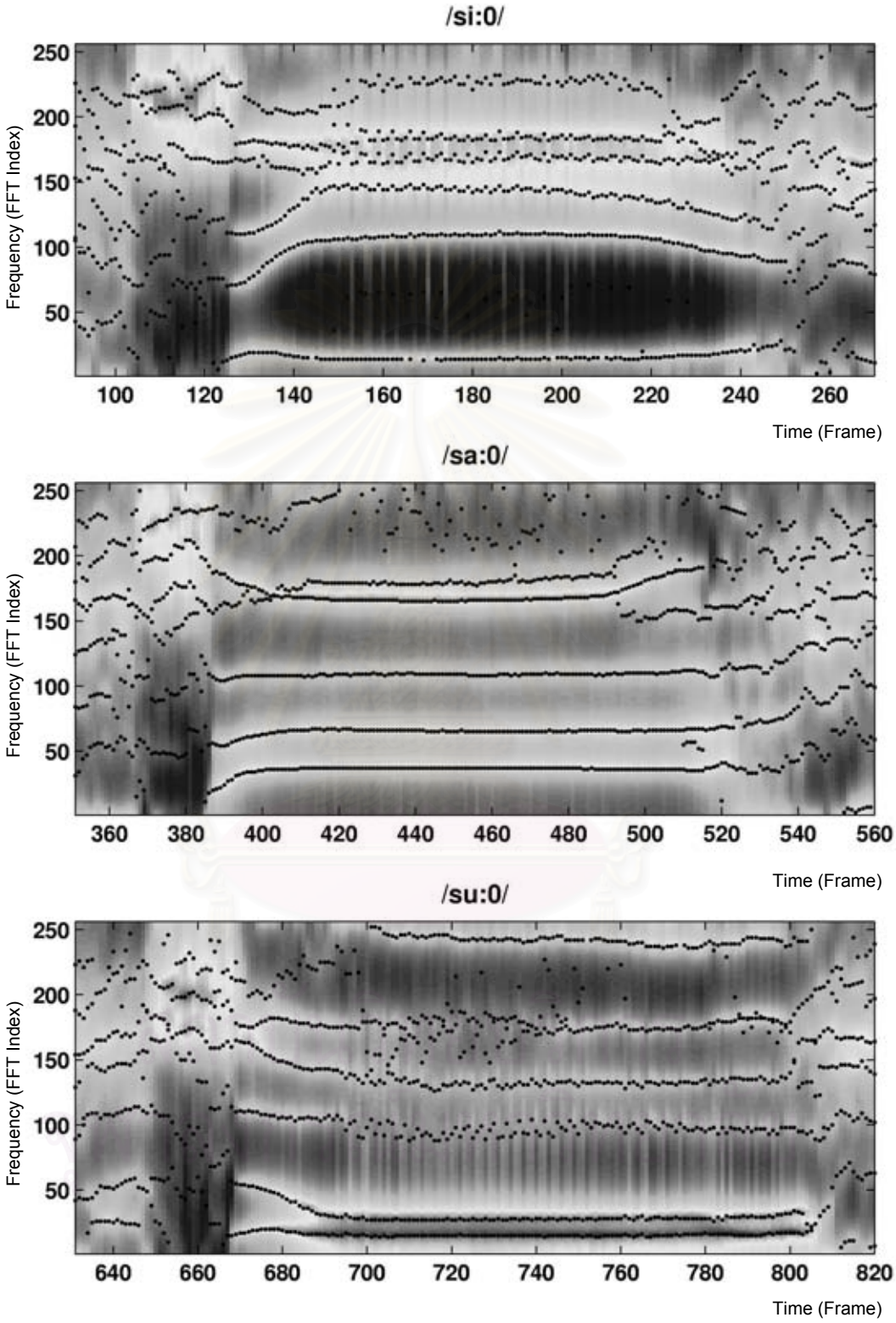


Figure 2.17 Spectograms of the words /sii0/, /saa0/, /suu0/

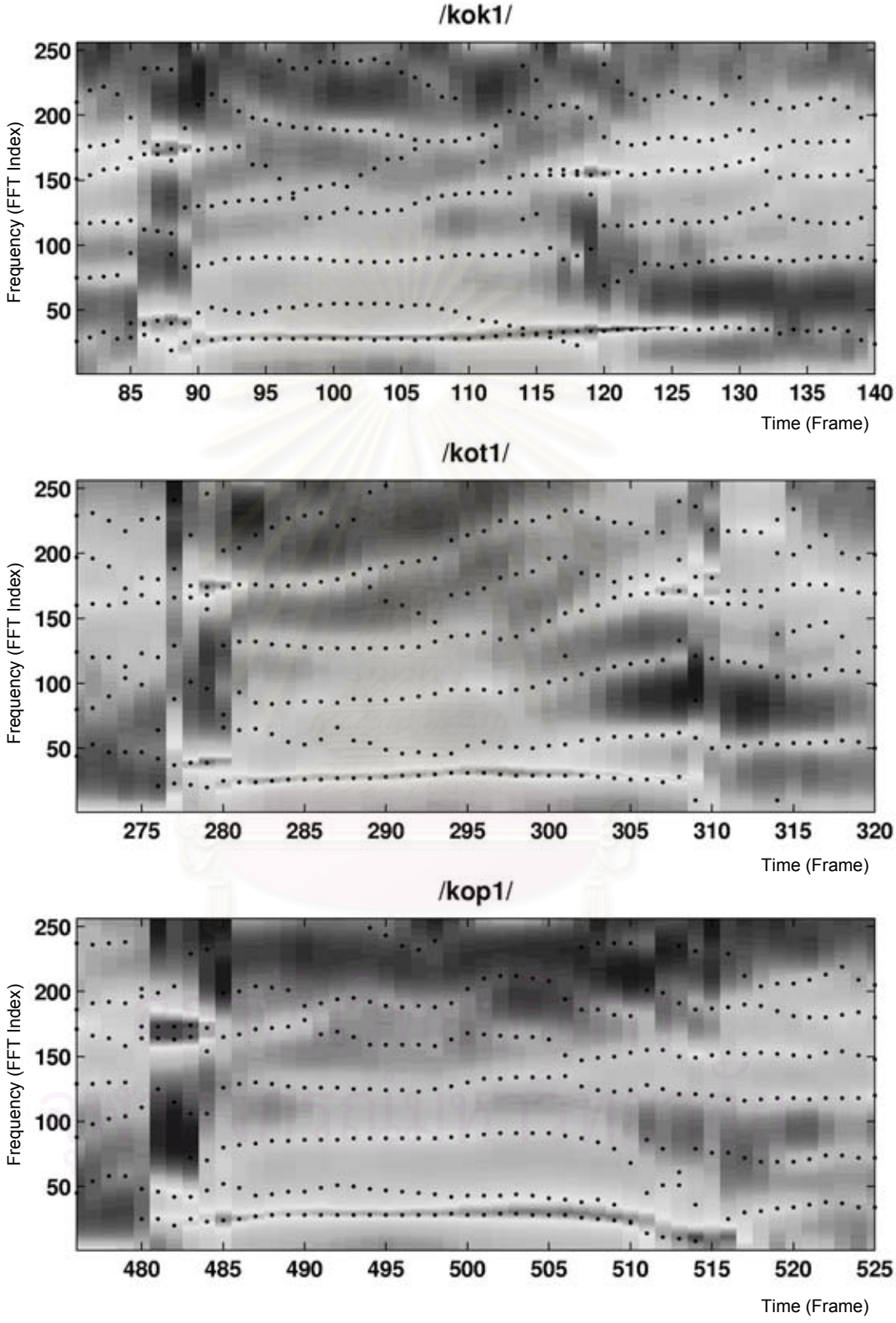


Figure 2.18 Spectograms of the words /kok1/, /kot1/, /kop1/

CHAPTER 3

The Onset-Rhyme Acoustic Models

In the previous chapter, acoustic-phonetic analysis on had been conducted Thai continuous speech. The Thai vowels and consonants were acoustically analysed to explore their acoustic characteristics. The outcome of vowel analysis not only provide solid acoustical background of Thai utterance but also provide acoustic cues for modelling of speech units. Using the acoustic knowledge and understanding, the onset and rhyme units are acoustically modelled as basic recognition units. Concept and details of the onset-rhyme models are explained in this chapter.

3.1 Concept of the Onset-Rhyme Acoustic Models

From phonological point of view, a syllable is composed of an onset and a rhyme where the rhyme comprises nucleus, and coda as illustrated in Figure 1.1 in Chapter 1. The Thai language has simple syllable structure as shown in Figure 1.2 where c is releasing consonant, C is arresting consonant, V is vowel, and T is tone. Hence, the onset covers releasing consonant segment of a syllable and the rhyme covers the rests. The nucleus and the coda of a rhyme represents vowel segment (VV) and arresting consonant (C) of a syllable.

In acoustic modelling, an onset unit consists of a releasing consonant and its transition towards the adjacent vowel nucleus. The onset unit then provides combinations based on both releasing consonant-vowel (cV) and consonant cluster-vowel (ccV) of the Thai syllable. Thus, the onset unit combines crucial acoustic cues, existed in the transitional period, for recognition of the releasing consonant particularly for stop consonants. In other words, the onset unit effectively handles the intra-word coarticulatory effects by the model itself that makes the model context-dependent.

On the other unit, a rhyme unit contains the whole vowel segment and an arresting consonant. Like its unit counterpart, the rhyme unit provides combinations based on both monophthong-consonant (VC) and diphthong-consonant (VVC) of the Thai syllable. Hence, the rhyme unit captures the transitional period between vowel and arresting consonant which makes the unit context-dependent like its counterpart. Physical model of the onset-rhyme model is depicted in Figure 1.3 along with other subword units. In Figure 1.3 and 3.1, phonetic transcriptions of the word /khrvvang2 mvv0/ have been illustrated where the onset unit covers the whole consonant clusters with transitional period and the rhyme unit spans across entire vowel and arresting consonant.

In selection of subword unit, two major criteria must be taken into account for good subword units, consistency and trainability (Lee, 1990), in other words, acoustic resolution and estimation reliability (Juang and Furui, 2000). Good subword units should be consistent and trainable, however, previously used subword units in most large-vocabulary speech recognition systems do not meet both criteria as summarized in Table 1.1 in Chapter 1.

From evaluation of previous subword units, phones are not consistent because different samples of the same phone are not always characteristically similar. A phone is strongly affected by its left and right neighbouring phones. But phone are widely used because they could be sufficiently trained with just a few hundred sentences. Context-dependent phones or triphones are consistent than phones because triphones model coarticulatory effects in both left and right neighbouring phones. However, triphone models are not easily trainable

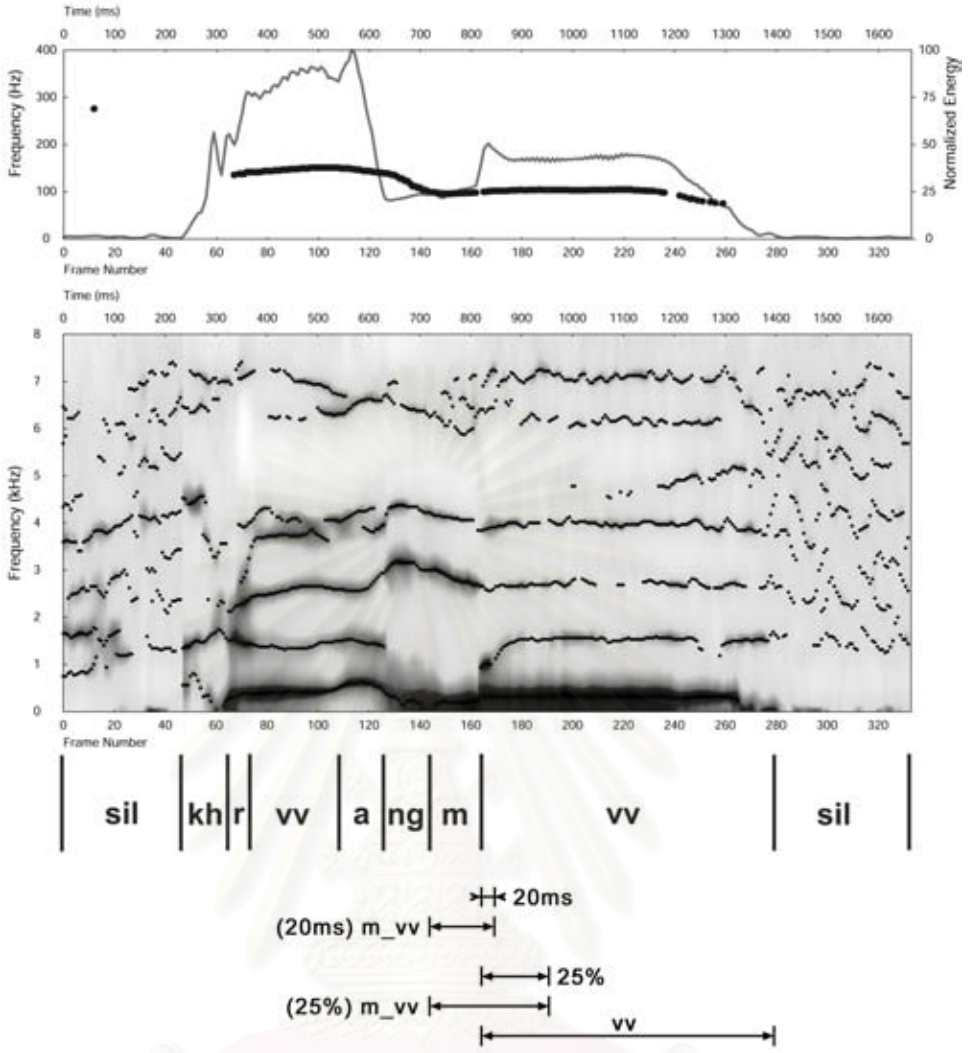


Figure 3.1 Fixed duration and variable duration overlaps of the onset-rhyme models of the word /khrvvang2 mvv0/

because there are a large number of triphone models even in limited training data. Currently, training of triphone models with limited training data has only been done through some techniques, sharing, deleted interpolation, interpolated with context-independent, or generalized triphones, for instance.

Considering the onset-rhyme models, the models are consistent throughout their entire set since the same onset-rhyme models have similar characteristics across different instances. On trainability criterion, since there are a limited number of onset-rhyme models, the models are sufficiently trained with only small set of training sentences. Hence, the onset-rhyme models have met both criteria of consistency and trainability considerations which are major advantanges over other subword units.

The concept of onsets and rhymes was first proposed by Luksaneeyanawin (1992) applied to the Thai speech synthesis system. The subsyllable onset-rhyme models, therefore, have been applied to Thai continuous speech recognition system for several reasons. First, there are only about 26,928 grammatically generated distinct admissible syllables in Thai. (Luksaneeyanawin, 1993) Thai language has 9 monophthongs and 6 diphthongs in short and long pairs, 21 consonants, and 12 consonant clusters as shown in Table 2.1, 2.2, and 2.3 respectively. Among the 21 initial consonants, only 8 consonants, /p, t, k, m, n, ng, j, w/, can

be both releasing and arresting consonants. Then, a number of onset-rhyme models are finite number as shown in Table 1.4 and 1.5. These finite number of onset-rhyme models have represented all potential speech units of the Thai language. Practically, a number of onset-rhyme units should be less since there are some units that do not grammatically exist or have very few occurrences.

Secondly, the Thai language has simple syllable structure as illustrated in Figure 1.2 in which the onset-rhyme are simply applied. The whole Thai syllable set can be recognized in pairs of onset and rhyme models. Thus, the recognizer models input syllables by concatenation of onset and rhyme pairs.

Thirdly, the onset-rhyme models are context-dependent beginning from level of acoustic model up to level of language model. Each of the onset unit contains releasing consonant and its transitional period towards the following vowel. Then, the same releasing consonant followed by different vowel context is individually modelled as a single onset unit. Unlike the phone models, the onset units capture a consonant cluster as a single arresting consonant while the phone models consider as a sequence of consonants. The rhyme units contain the whole vowel and arresting consonant. Like the onset unit, the same vowel followed by different arresting consonant is separately modelled as a single rhyme unit. Hence, a releasing consonant is right context-dependent on its immediate following vowel in an onset unit. Also, an arresting consonant is left context-dependent on its preceding vowel in a rhyme unit. This is a major point of difference to the triphones where a triphone is actually a phone within different context as illustrated in Figure 3.2.

In Figure 3.2, physical segments are illustrated on phones, diphones, triphones, and onset-rhyme units. A speech waveform is described as segments of phones. Using the diphones and triphones, their physical segments are similar to the phones but are logically described according to the context. For example, the /aa/ phone in the syllable /phaa/ has similar characteristics to the diphone /ph-aa/. But the diphone /ph-aa/ is logically defined to have /ph/ as its left context. Using the onset-rhyme models, each syllable is modelled by a pair of onset and rhyme units. For instance, the syllable /khaaw/ is modelled as /khaa/ and /aa_w/.

$$P(W) = P(w_1, w_2, \ldots, w_N) \quad \text{..... (3.1)}$$

$$P(W) = \prod_{t=1}^{N} P(W_t) = P(W_1)P(W_2)\ldots P(W_N) \quad \text{..... (3.2)}$$

$$P(W_t) = P(S_1, S_2, \ldots, S_M) = P(S_1)P(S_2)\cdots P(S_M) \quad \text{..... (3.3)}$$

where
$$P(S_j) = P(O_j, R_j) = P(O_j)P(R_j|O_j) \quad \text{..... (3.4)}$$

$$P(W_t) = \prod_{j=1}^{M} P(S_j)$$

then
$$= P(O_1)P(R_1|O_1)\cdot P(O_2|R_1)P(R_2|O_2) \cdots P(O_j|R_{j-1})P(R_j|O_j) \quad \text{..... (3.5)}$$

$$= P(O_1)P(R_1|O_1) \prod_{j=2}^{M} P(O_j|R_{j-1})P(R_j|O_j)$$

Moreover, the onset-rhyme models also incorporate language modelling into the model at the syllable level. An onset-rhyme model comprises a pair of onset model and rhyme model which makes up a syllable as depicted in Figure 2.5. An onset model is then right context-dependent on the following rhyme and a rhyme is left context-dependent on the preceding onset. Consequently, a sequence of onset and rhyme pairs makes up a sequence of

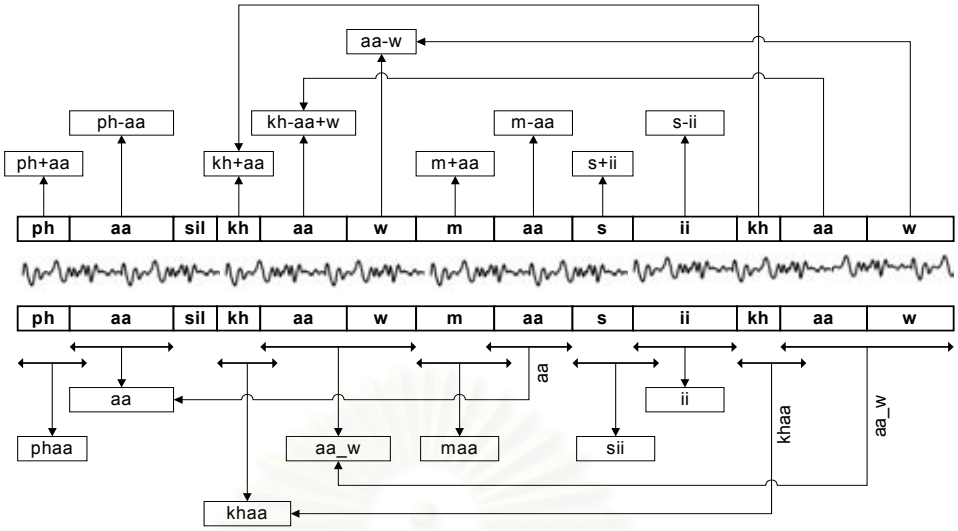


Figure 3.2 Physical speech segments of phones, diphone, triphones, and onset-rhyme units

syllables, sequence of words, and the whole sentence respectively. These bottom-up approach indicates that language modelling is directly embedded into the onset-rhyme models.

Considering the language model $P(W)$ from Eq. (3.1), the probability of a sequence of N words W is stated in Eq. (3.2). Then, probability of the unigram language model, in which each word independently occurs, could be expressed in Eq. (3.2) where N is the number of words. Each word contains a sequence of M syllables, thus, probability of a word W_t is then expressed in Eq. (3.3). Each syllable is modelled as a concatenation of the subsyllable onset-rhyme models.

Since the onset-rhyme model always occurs in a pair of the onset model and the rhyme model, then, a rhyme model depends on its preceding onset. Consequently, the onset unit is only followed by its corresponding rhyme unit as directed in the model network depicted in Figure 3.5 and 3.6. The rhyme unit conditionally depends on its preceding onset, then, the probability of a rhyme unit is described as conditional probability $P(R_j|O_j)$. Hence, the probability of a syllable comprises an onset probability $P(O_j)$ and a rhyme conditional probability $P(R_j|O_j)$ as stated in Eq. (3.5). The $P(O_j)$, $P(O_j|R_{j-1})$, and $P(R_j|O_j)$ In addition, the onset-rhyme models have covered a finite set of speech units that represents all potential speech units of the language comparing to other context-dependent models. As a result, the limited numbers of onset-rhyme models could be sufficiently trained with only a small set of sentences. The models also guarantee that every unit, existed in the language, is modelled. The numbers of onset-rhyme models are shown in Table 1.4 for the onset units and Table 1.5 for the rhyme units.

Finally, the onset-rhyme models have revealed thus provided significant acoustic cues for tone recognition, that is, a syllable boundary as depicted in Figure 3.3. Location of syllable boundaries could be accurately obtained over a pair of onset and rhyme models, that is, at the beginning in front of an onset model and after a rhyme model. Then, tonal information of a syllable is properly extracted and recognised over the whole syllable segment. This is also another major advantages of the onset-rhyme models over other subword models.

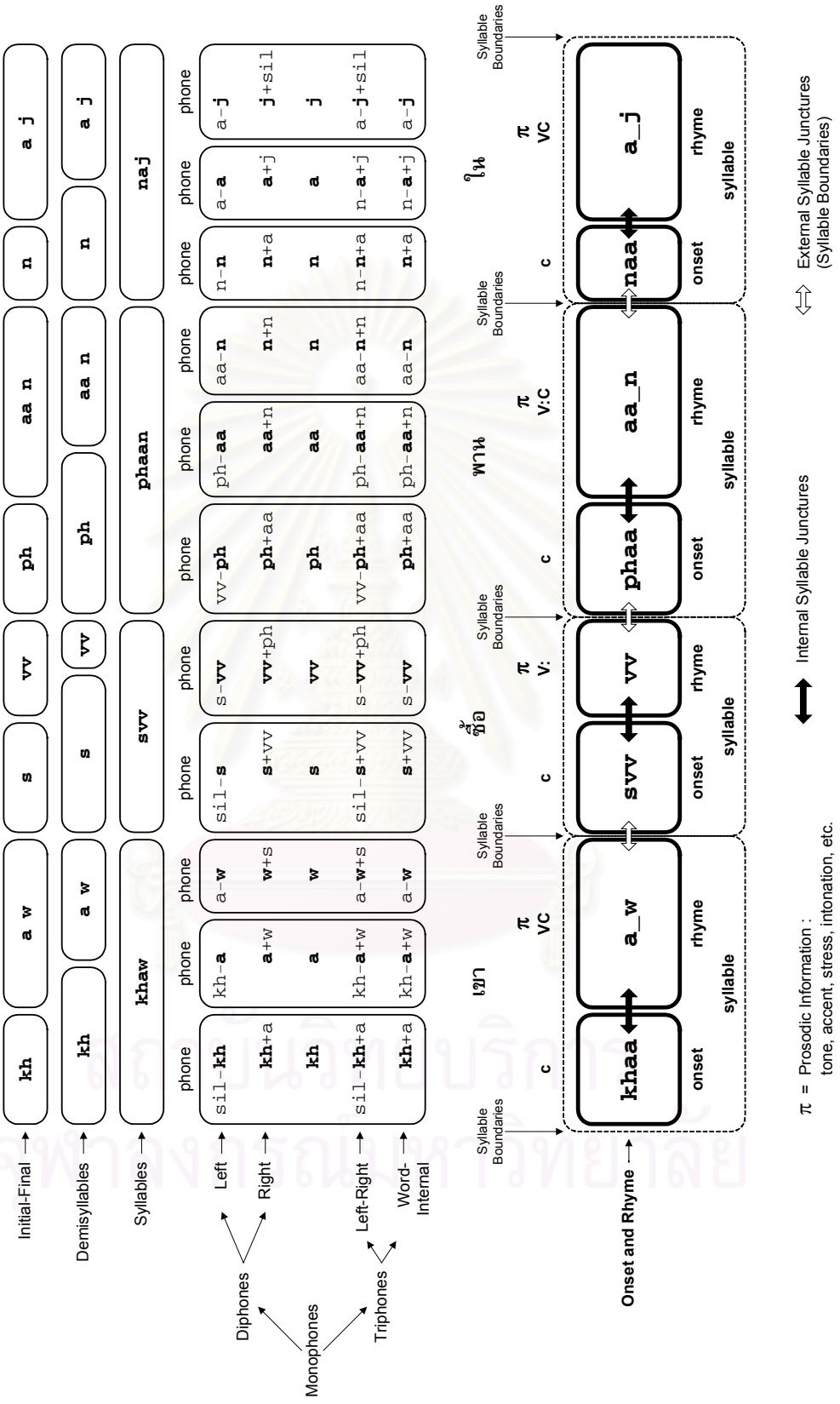
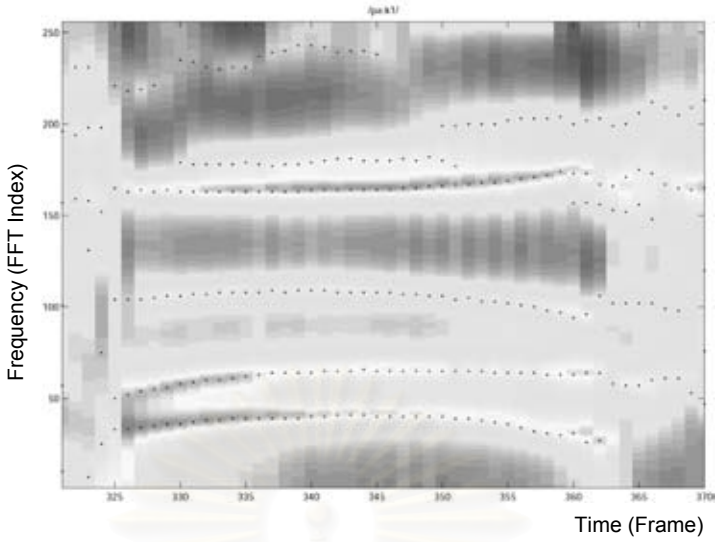
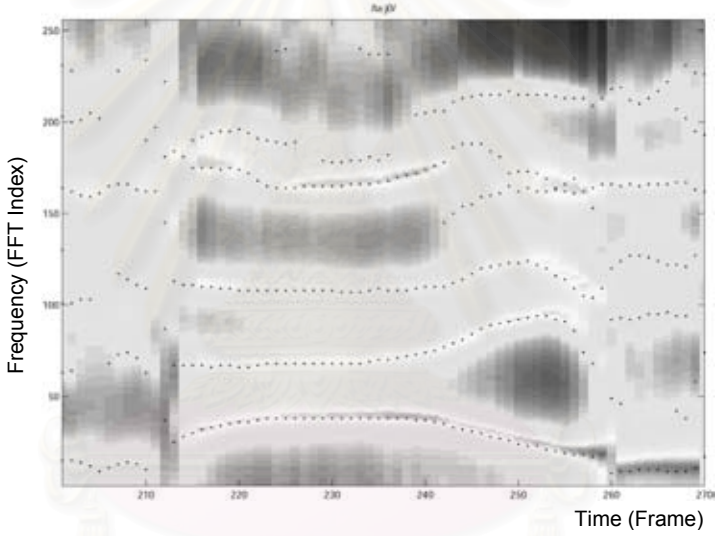


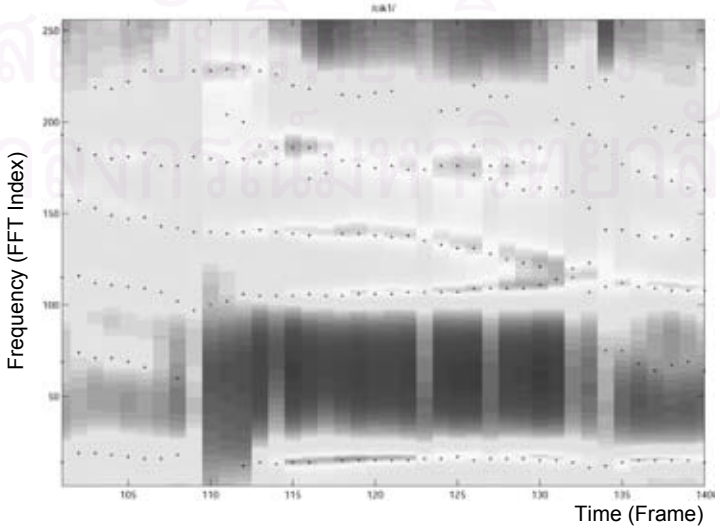
Figure 3.3 Comparison on all speech units : phones, diphones, triphones, syllable, demisyllable, initial-final, and onset-rhyme units.



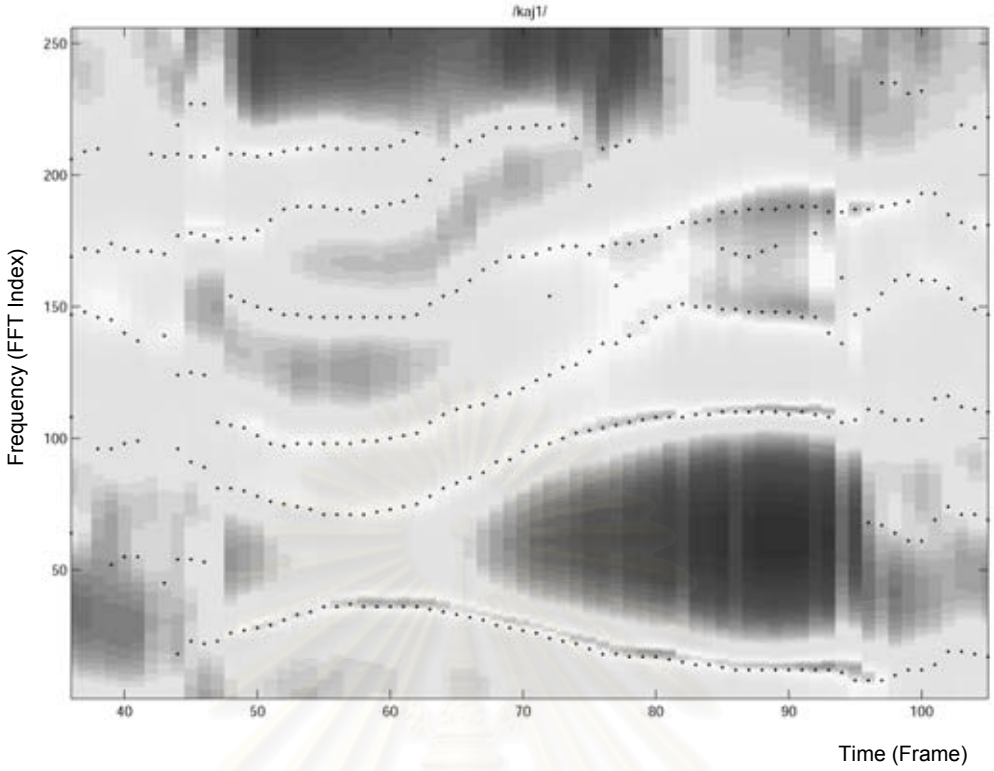
(a) /p/ in the word /paak1/



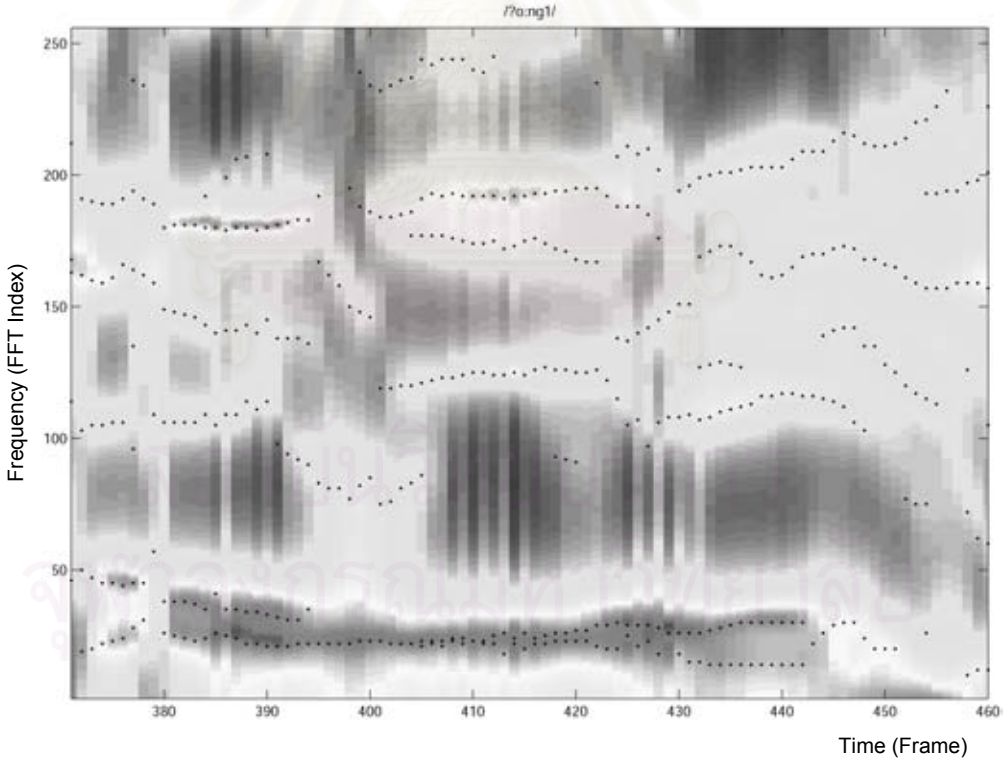
(b) /t/ in the word /taaj0/



(c) /c/ in the word /cik1/

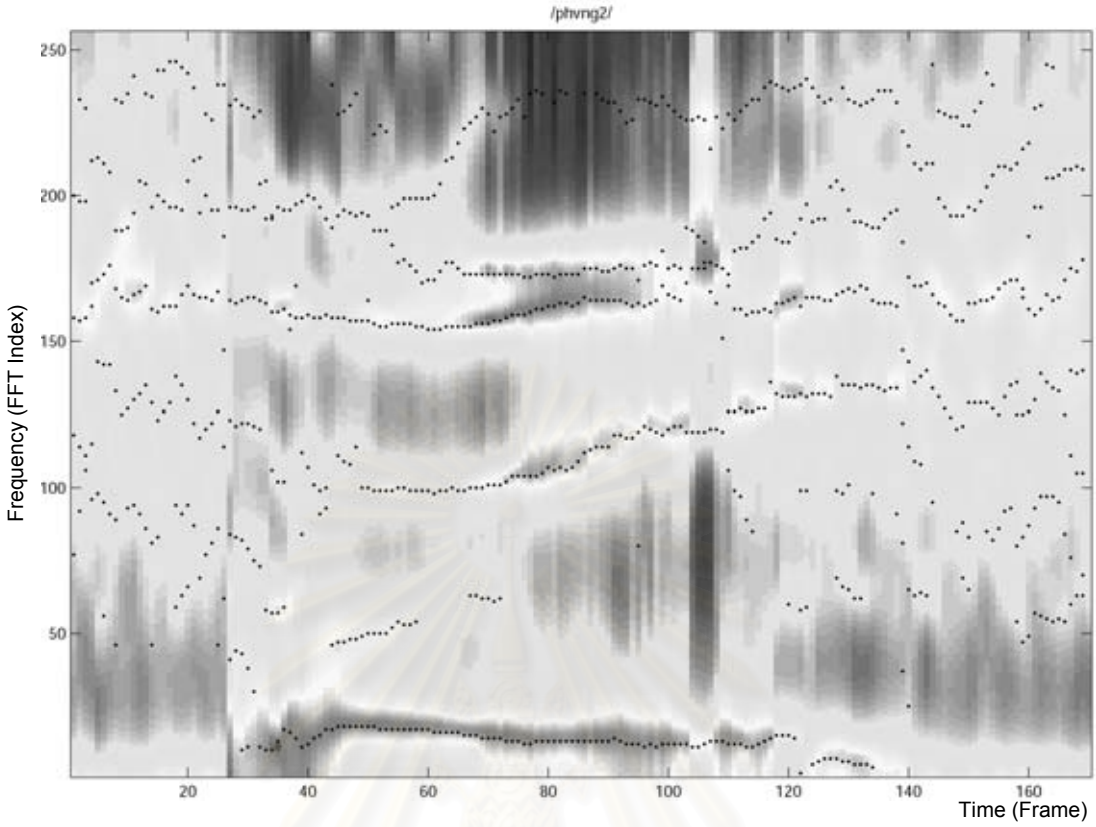


(d) /k/ in the word /kaj1/

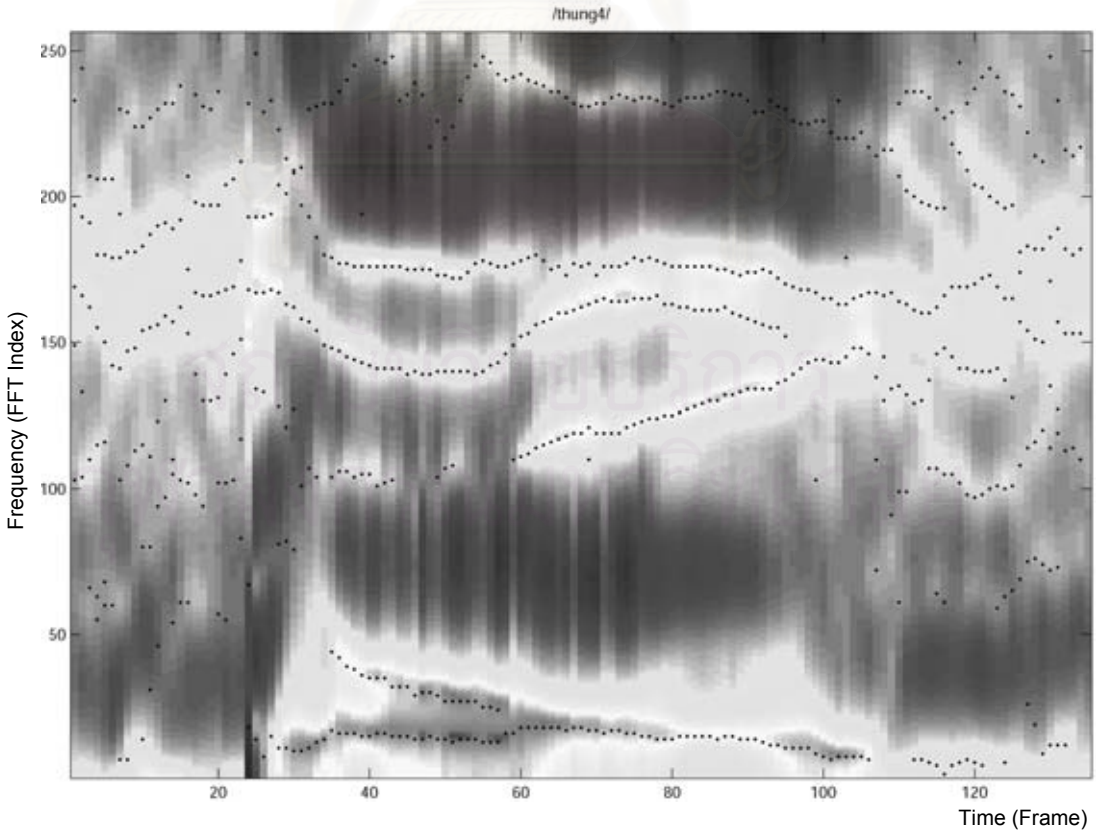


(e) /z/ in the word /zoong1/

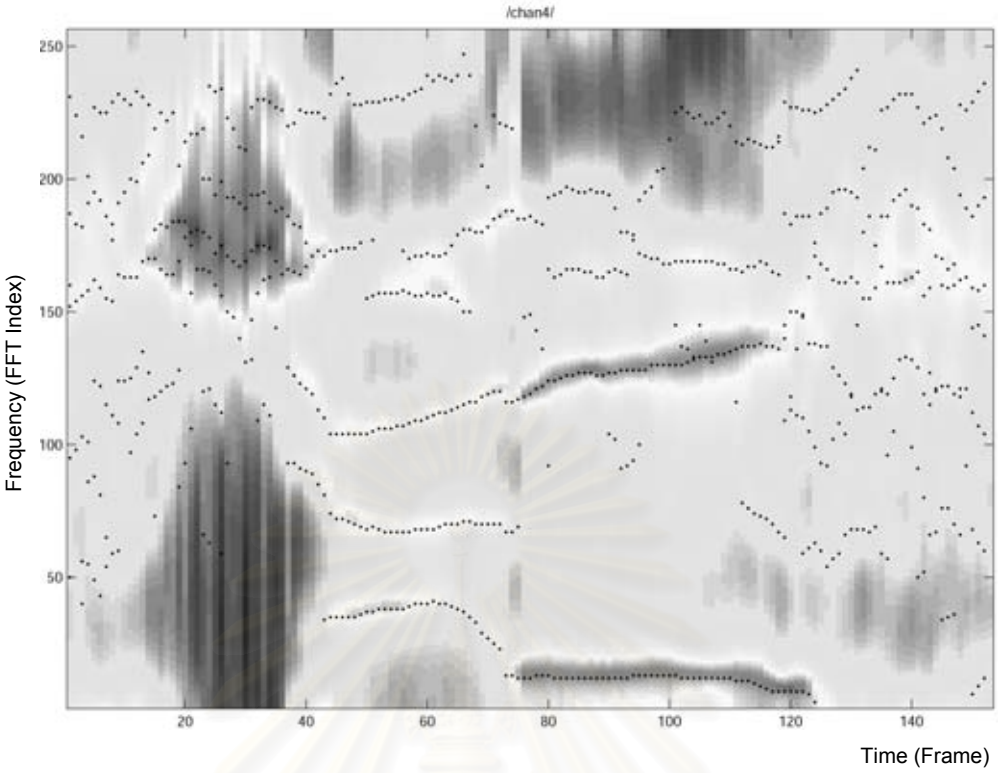
Figure 3.4 Spectrographic illustration of the Thai voiceless unaspirated stop consonants /p, t, c, k, z/ in various syllables.



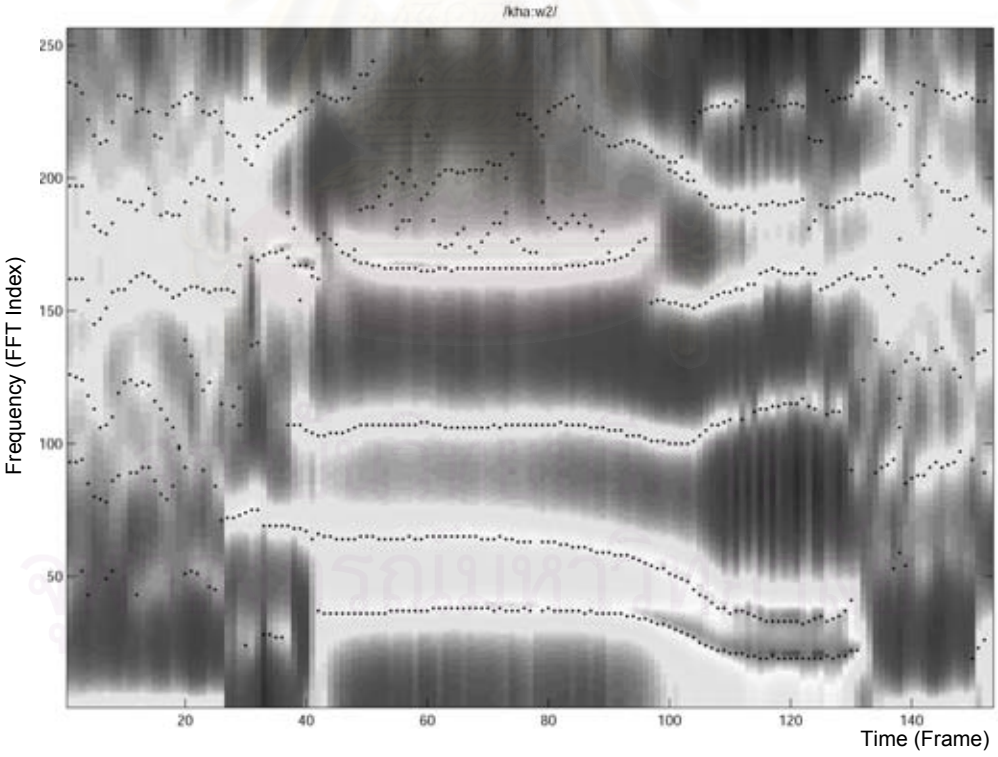
(a) /ph/ in the word /phvng2/



(b) /th/ in the word /thung4/

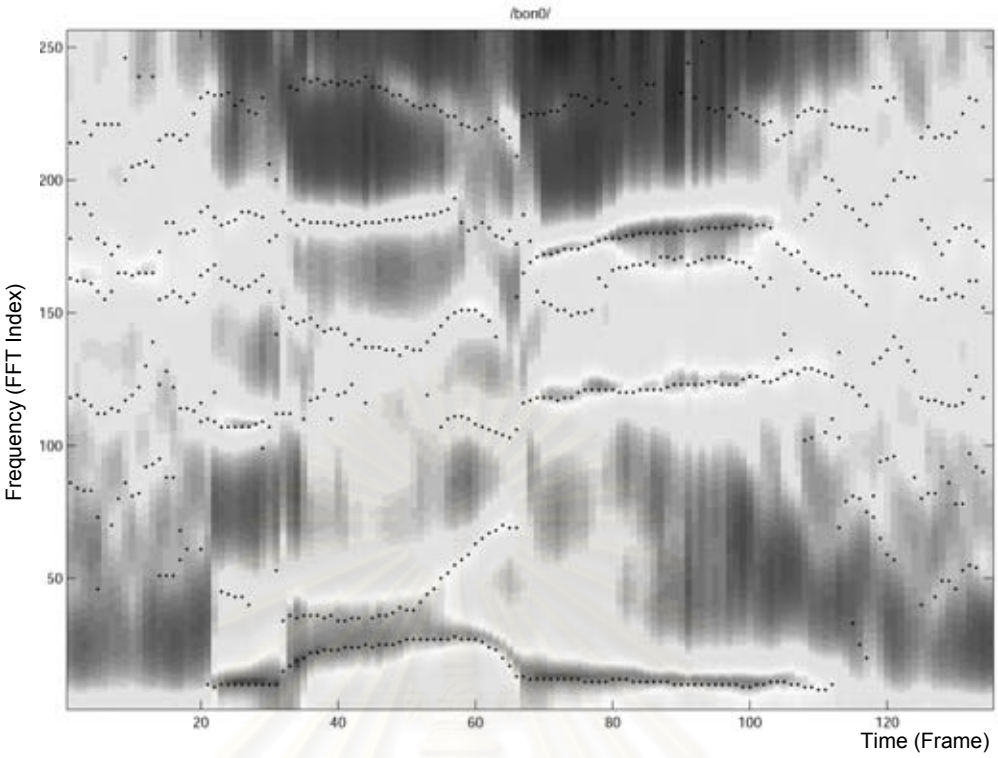


(c) /ch/ in the word /chan4/

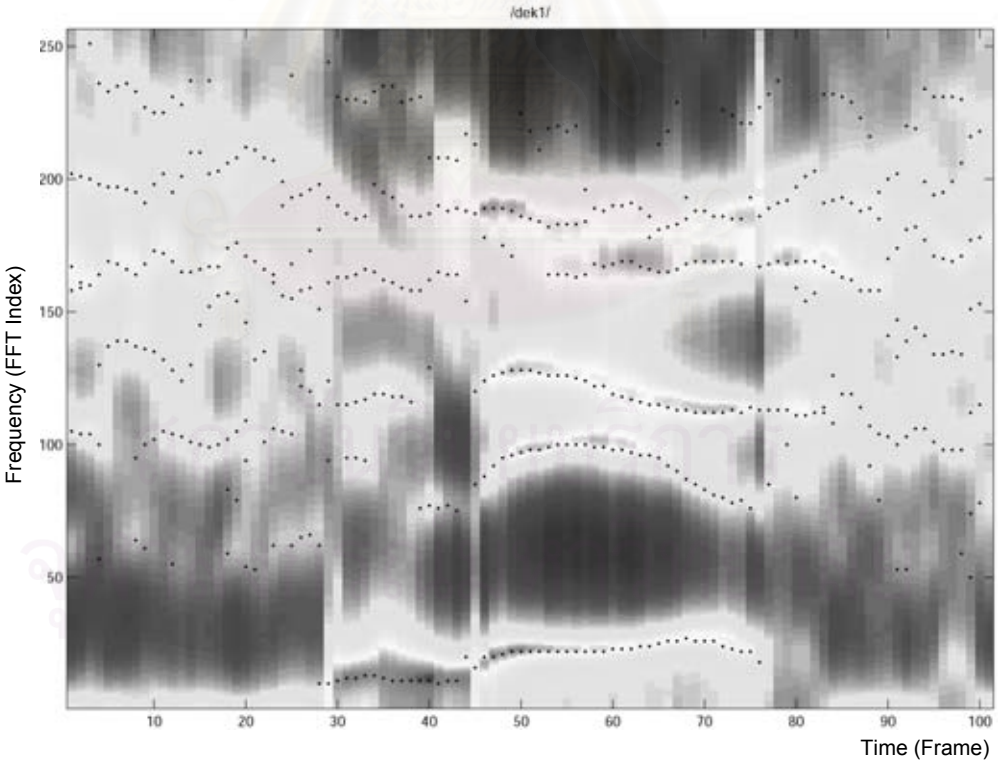


(d) /kh/ in the word /khaaw2/

Figure 3.5 Spectrographic illustration of the Thai voiceless aspirated stop consonants /ph, th, ch, kh/

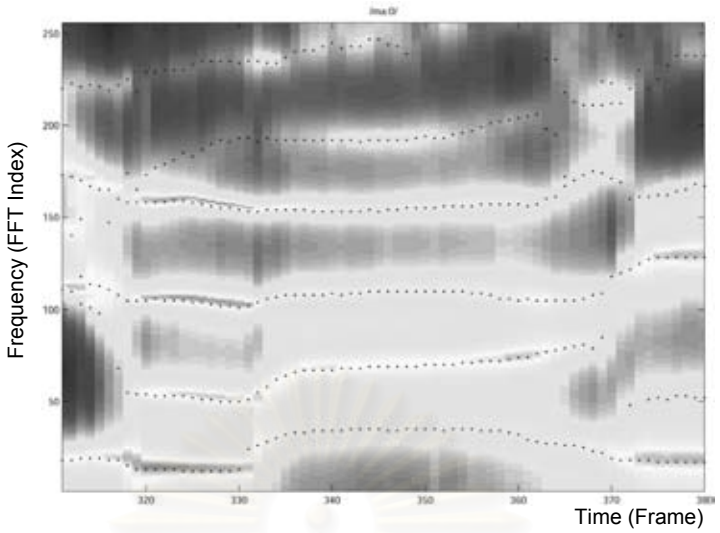


(a) /b/ in the word /bon0/

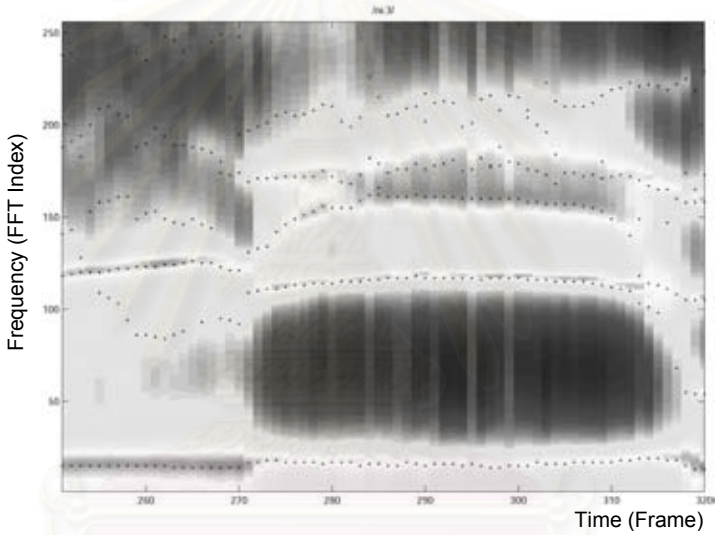


(b) /d/ in the word /dek1/

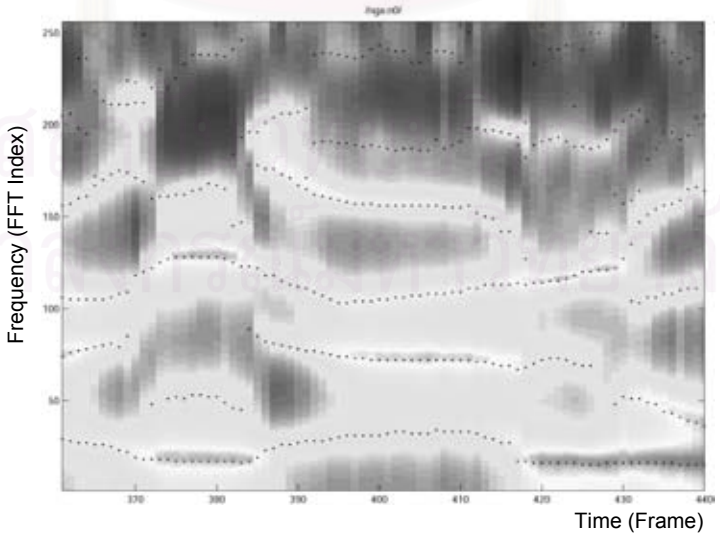
Figure 3.6 Spectrographic illustration of the Thai voiced stop consonants /b, d/



(a) /m/ in the word /maa0/

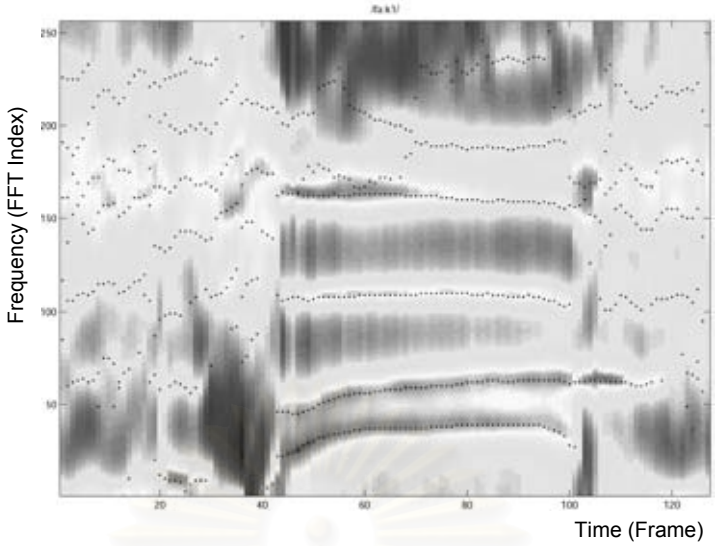


(b) /n/ in the word /nii3/

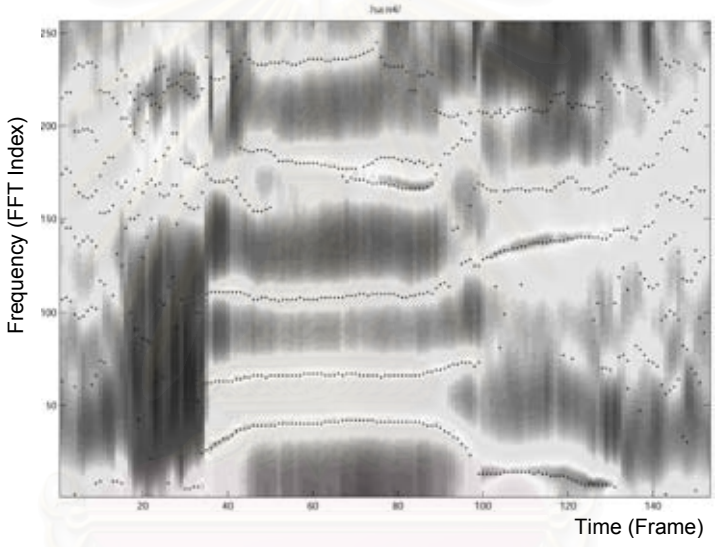


(c) /ng/ in the word /ngaan0/

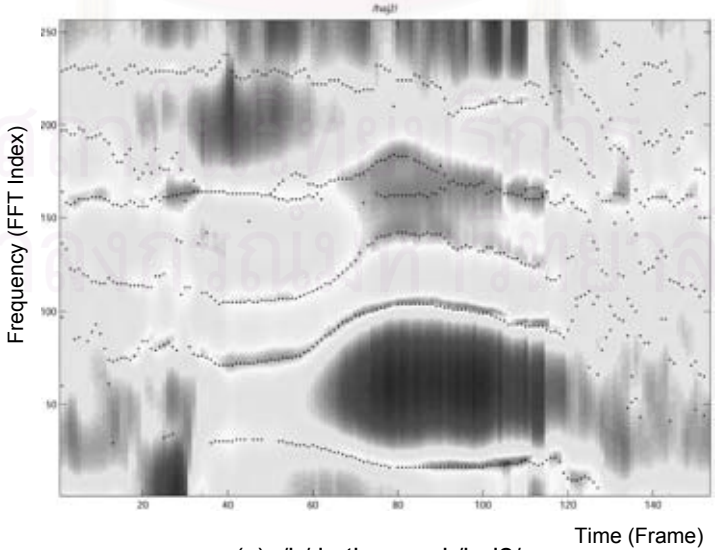
Figure 3.7 Spectrographic illustration of the Thai nasals /m, n, ng/



(a) /f/ in the word /faak1/

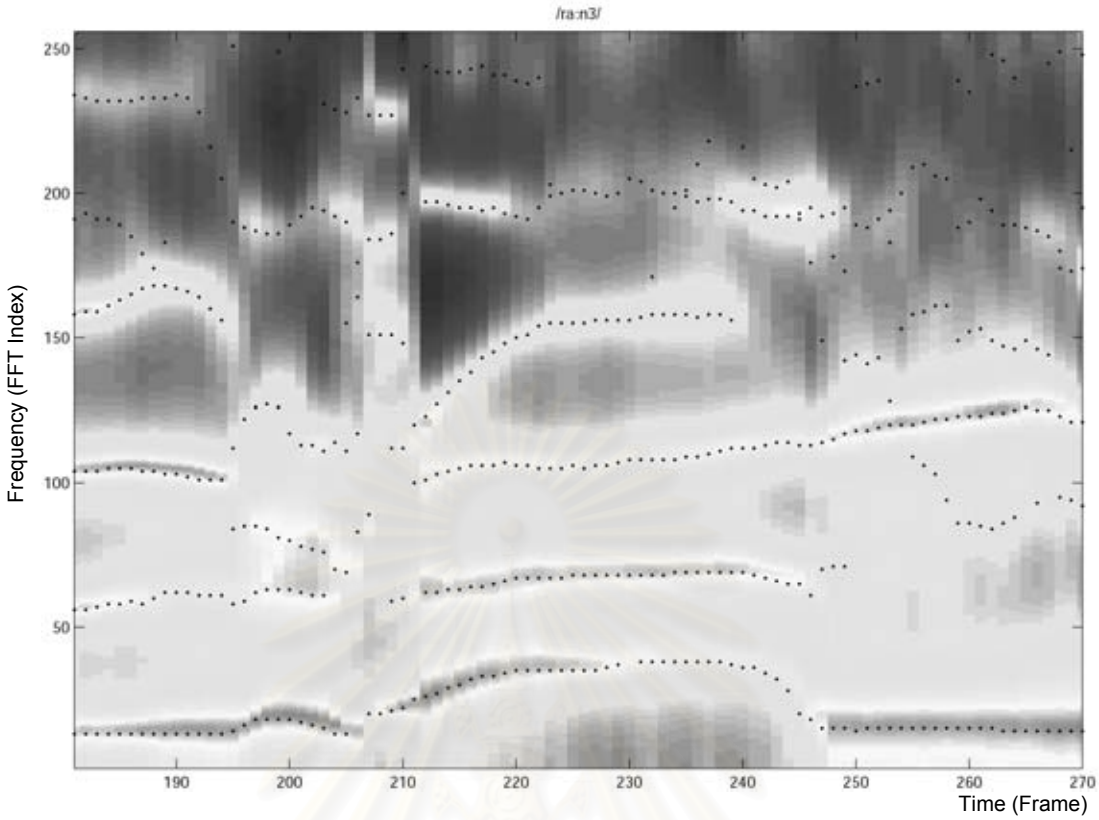


(b) /s/ in the word /saan4/

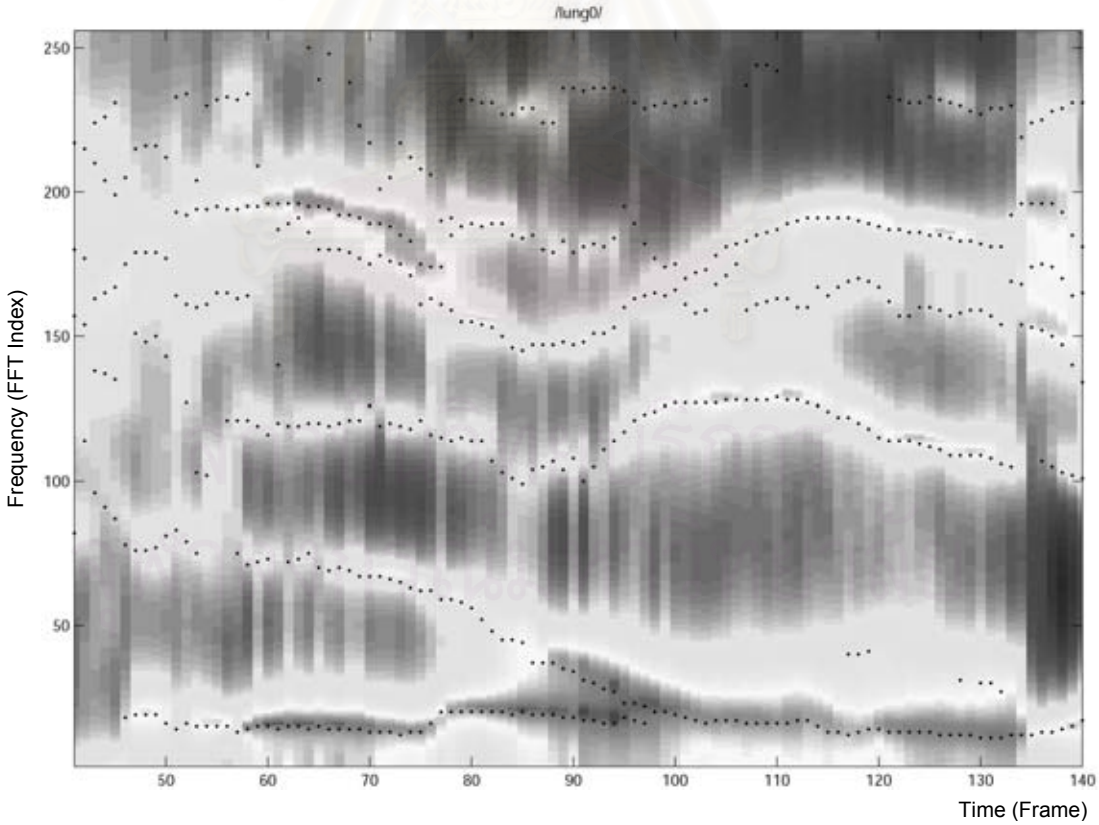


(c) /h/ in the word /haj2/

Figure 3.8 Spectrographic illustration of the Thai fricative /f, s, h/

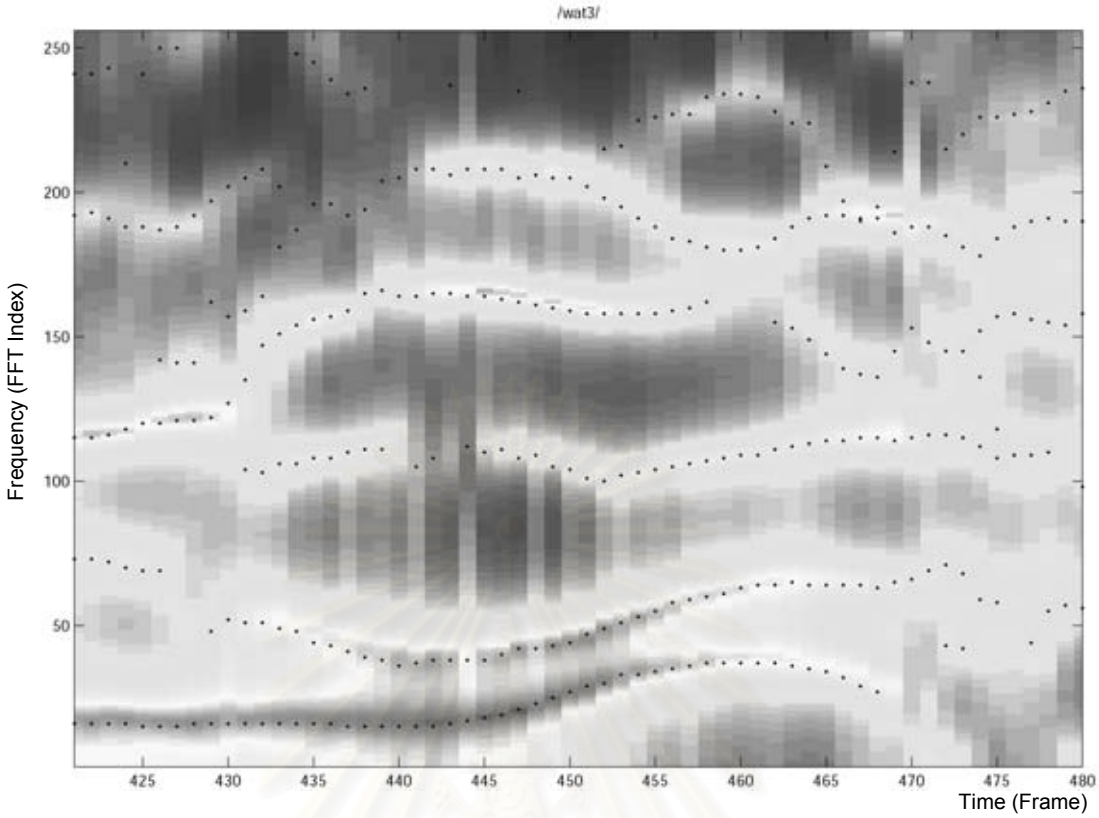


(a) /r/ in the word /raan3/

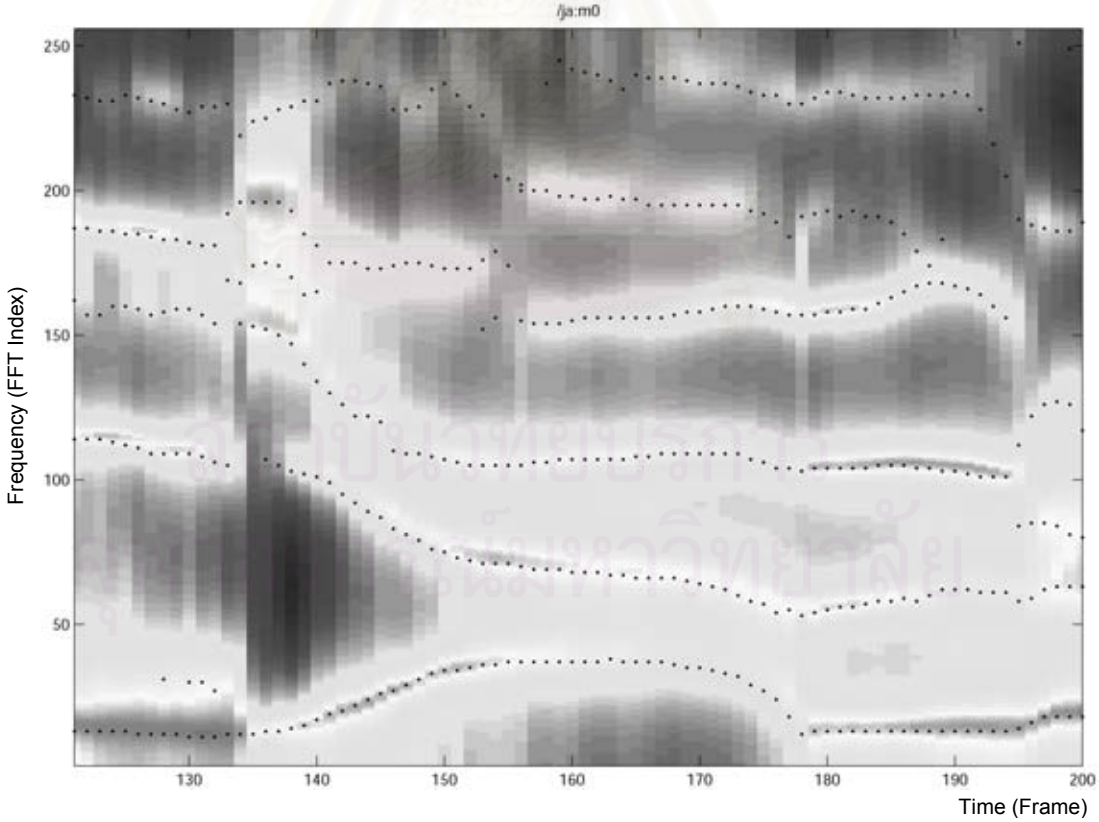


(b) /l/ in the word /lung0/

Figure 3.9 Spectrographic illustration of the Thai trill /r/ and lateral /l/



(a) /w/ in the word /wat3/



(b) /j/ in the word /jaam0/

Figure 3.10 Spectrographic illustration of the Thai approximants /w, j/

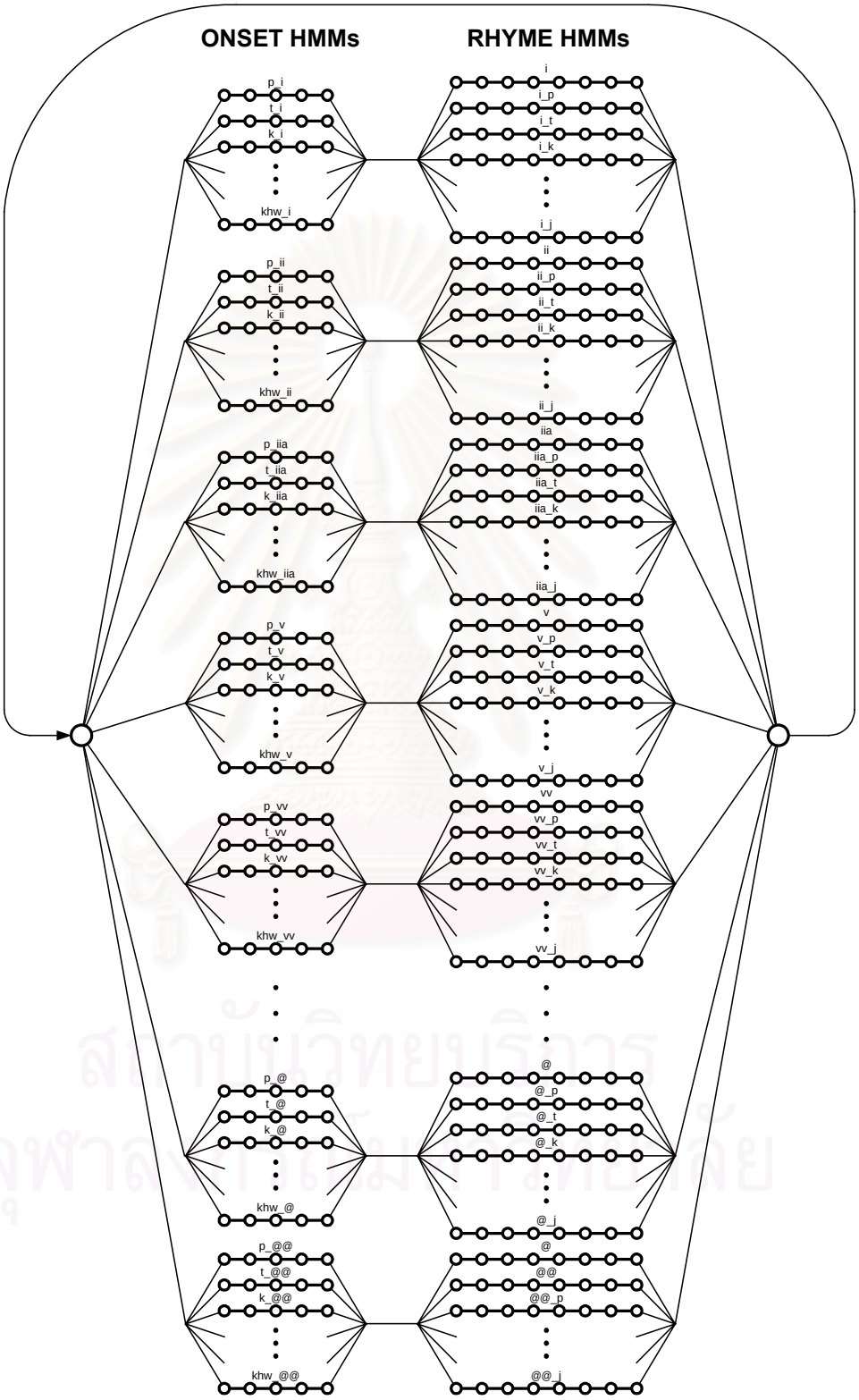


Figure 3.11 Network of the phonotactic onset HMMs and rhyme HMMs in forming syllables

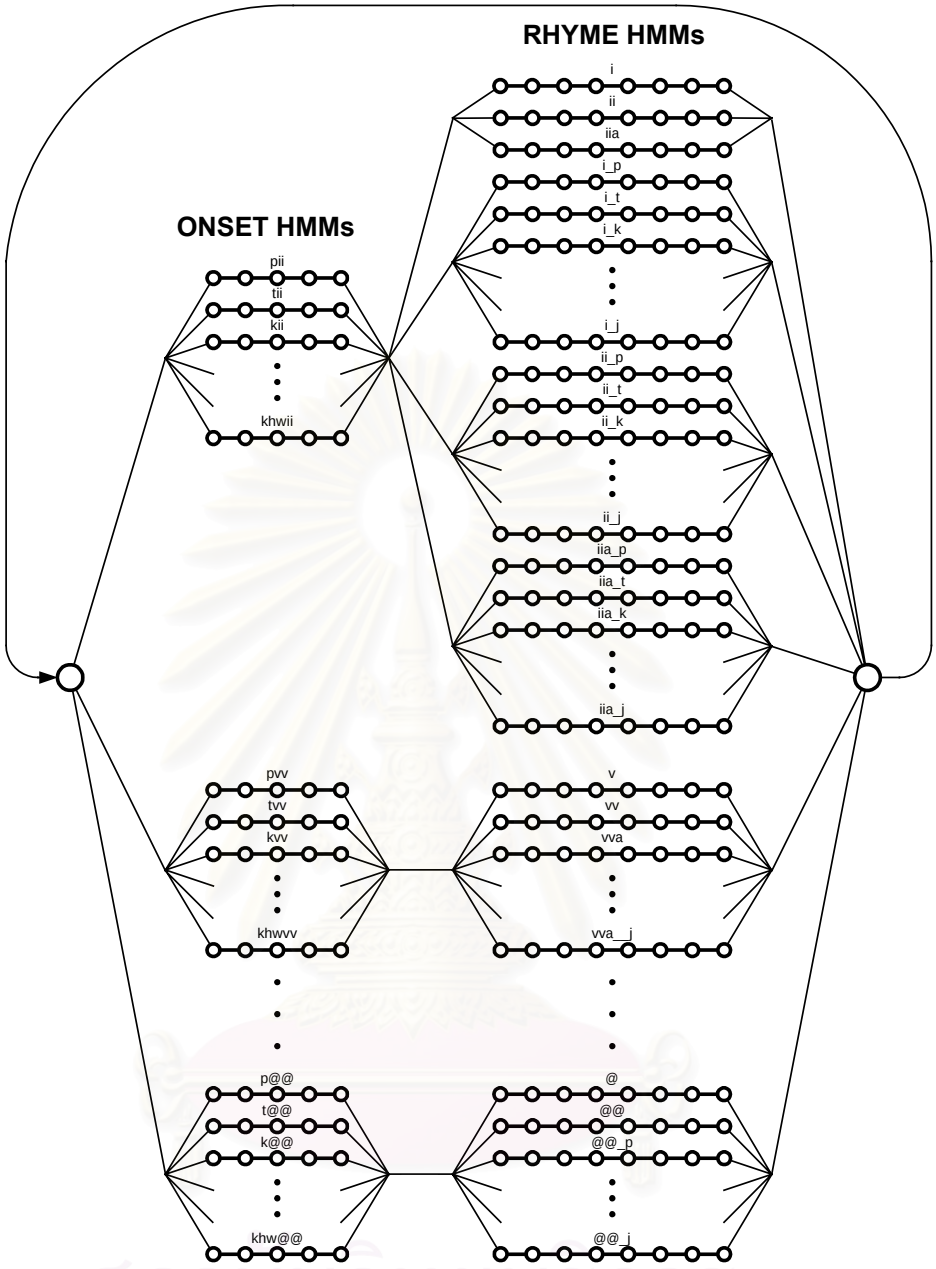


Figure 3.12 Network of the contextual onset HMMs and rhyme HMMs in forming syllables

3.2 Modelling of the Onset-Rhyme Acoustic Models

From the acoustic-phonetic analysis of Thai syllables, each of the Thai releasing consonant is analysed with the results as illustrated in Figure 3.4 to 3.10. Figure 3.4 illustrates acoustic characteristics of the voiceless unaspirated stops /p, t, c, k, z/. In Figure 3.5, acoustic characteristics of the voiceless aspirated stops /ph, th, ch, kh/ are illustrated. Figure 3.6 shows acoustic characteristics of the voiced stops /b, d/. The set of non-stops are shown in Figure 3.7 to 3.10. Figure 3.7 shows acoustic characteristics of the nasals /m, n, ng/. In Figure 3.8, acoustic characteristics are illustrated on the fricatives /f, s, h/. Figure 3.9

illustrates acoustic characteristics of the trill /r/ and the lateral /l/. In Figure 3.10, acoustic characteristics of the approximants /w, j/ are depicted.

In Figure 3.4 to 3.10, each of the Thai releasing consonant shows its unique acoustic characteristics. Transitional period between a releasing consonant and its adjacent vowel contains some acoustic cues in characterizing the releasing consonant. For example, the formant transitions of a vowel in the transitional period are specific to each consonant. The formant movement of a vowel is moving towards the locus of each consonant as illustrated in Figure 2.14 to 2.18 in Chapter 2 and in Figure 3.4 to 3.10. The onset units cover the whole segment of a releasing consonant and its transitional period toward the adjacent vowel. Hence, these acoustic cues are collected within the onset units during modelling.

In modelling of the onset-rhyme models, a number of possible combinations of the model is shown in Table 1.4 and 1.5 in Chapter 1. From analysis of the Thai syllables as illustrated in Figure 3.4 to 3.10, two types of the onset-rhyme models are proposed other than the theoretical onset-rhyme models, that is, the contextual onset-rhyme models, and the phonotactic onset-rhyme models. Details of each type are described as follows.

3.2.1 Types of the Onset-Rhyme Models

A. Theoretical Onset-Rhyme Models

The theoretical onset-rhyme models are basic one-to-one mapping of onset and rhyme units to each segment of the Thai syllable. The onset units are context-independent phone models of all releasing consonants. The rhyme units covers both vowel and arresting consonant. Therefore, both onset and rhyme units are not context-dependent in this theoretical models. The classical onset-rhyme models are then context-independent models. This dissertation does not include this models in analysis and recognition.

B. Phonotactic Onset-Rhyme Models (PORMs)

From the acoustic analysis, transitional period exists between a releasing consonant and adjacent vowel nucleus in a syllable. The transitional period provides crucial acoustic cues in determining the releasing consonant in a syllable. Therefore, an onset unit covers a releasing consonant and the transitional period towards its adjacent vowel nucleus. A rhyme unit covers the whole vowel segment and the releasing consonant. Then, the onset units have partially overlapped over the vowel segment of the rhyme units. In consequence, the onset units have combined crucial acoustic cues, existed in the transitional period, for recognition of the releasing consonants.

The phonotactic onset-rhyme models are proposed in this reseach. The onset units of the phonotactic onset-rhyme models are extended to cover all possible combinations as depicted in Figure 3.12. For example, the /p_i/, which is the /p/ in /i/ context, must be followed by its corresponding rhyme models with the same context, that is, /i_p, i_t, …, i_j/. Consequently, the onset units are thoroughly modelled according to their neighbouring context of the rhyme models. In Figure 3.12, the network configuration of the onset and rhyme pairs are illustrated. The onset units are shown in 5-state HMMs and the rhyme units are shown in 8-state HMMs. The number of HMM states for both units are described in the following section. The PORMs network illustrates complete combinations of the onset and rhyme units in every possible context. For examples, an onset unit is /p_i/ and a rhyme unit is /i_p/ as shown at the bottom of the figure. The /p_i/ is the releasing stop /p/ with transitional period in the /i/ vowel. The /p_i/ must be followed by the rhyme units in same /i/ vowel context as shown in the network. Similar to the CORMs network, each connection between the onset and the rhyme units represents conditional probabilities $P(O_j|R_{j-1})$, and $P(R_j|O_j)$ as stated in Eq. (3.5). These conditional probabilities are determined from pronunciation dictionaries during building the word network.

C. Contextual Onset-Rhyme Models (CORMs)

In order to reduce the number of onset unit in the PORMs, the contextual onset-rhyme models are introduced in this research. The results of acoustic analysis on Thai syllable show similar pattern of formant transitions in some cases. These formant patterns are similar in both short and long vowel context with the same releasing consonant. The examples of these patterns are shown in Figure 2.14 to 2.18 and Figure 3.4 to 3.10. Hence, the onset units can be greatly reduced by combining similar onset units with short-long vowel pairs on the same releasing consonant.

The onset units are always tied with their corresponding rhyme models with the same context included in the whole transitional period. For instance, the /pii/, which is the /p/ in /i,ii,iia/ context, must be followed by its corresponding rhyme units with the same context, that is, /i_p, i_t, …, iia_j/. The network of contextual onset-rhyme hidden Markov acoustic model is shown in Figure 3.11. In Figure 3.11, the network configuration of the onset and rhyme pairs are illustrated. The onset units are shown in 5-state HMMs and the rhyme units are shown in 8-state HMMs. The number of HMM states for both units are described in the following section. For examples, an onset unit is /khw@@/ and a rhyme unit is /@@_j/ as shown at the bottom of the figure. The network shows connections between the onset and rhyme pairs in which the rhyme units must follow the onset units. The CORMs network illustrates sharing of an onset unit with similar rhyme of the same short-long vowel pairs. Each connection between the onset and the rhyme units represents conditional probabilities $P(O_j|R_{j-1})$, and $P(R_j|O_j)$ as stated in Eq. (3.5). These conditional probabilities are determined from pronunciation dictionaries during building the word network. Hence, the contextual onset-rhyme models effectively handle the intra-word and intra-syllable coarticulatory effects by the model themselves that also make the model context-dependent.

3.2.2 Onset Unit Overlapping Schemes

In both of the contextual and phonotactic onset-rhyme models, the onset units have been extended to include transitional period over the vowel segment. As a result, there are two proposed schemes in determining duration of the overlap, that is, the fixed duration overlap and the variable duration overlap. These two schemes have been utilized in acoustic modelling of the onset-rhyme models. The overlapping of the onset units over the vowel segment show explicit modelling of transitional period. Hence, both of the onset models and rhyme models have provided overlapped segment models in acoustic modelling.

A. Fixed Duration Overlap

From acoustic analysis on Thai syllables, the transitional period occurs in a very short duration at the beginning of vowel segment. From speech unit statistics, the minimum length of a short vowel is 30 ms determined from the whole speech corpus. Therefore, the length of overlap should not longer than 30 ms to cover the transitional period. Otherwise, the whole vowel segment will be included in some short vowels. In fixed duration overlap, length of the overlap is predefined at either 10 ms, 20 ms, or 30 ms into the vowel segment of a rhyme unit. Example of the 20-ms fixed duration overlap is depicted in Figure 3.1.

B. Variable Duration Overlap

From acoustic analysis on Thai syllables, duration of the transitional period is proportional to length of the vowel segment. The formant transitions in short vowels tend to move faster than in long vowels. The faster movement makes their duration shorter. Unlike the fixed duration overlap, length of the overlap is varied in percentage of the vowel duration in the variable duration overlap. The length is computed at 5%, 10%, 15%, 20%, and 25% of the vowel duration. Example of the 25% variable duration overlap is depicted in Figure 3.1.

3.3 Task of the Thai Speech Corpus

Since there are no available Thai continuous speech corpus, a new Thai continuous speech corpus was created for building a Thai continuous speech recognition system. The task domain of the corpus is based on some Aesop's fables in Thai. The dictation or reading style is applied throughout the corpus. Thai text data are first collected by typing the Thai text into a computer in plain text. Secondly, the Thai text data are parsed into words and transcribed into phonetic transcriptions. Moreover, the transcribed text are modified by adding new words or new sentences to increase number of occurrences of each onset-rhyme model. The procedures are then repeated until there are sufficient samples of each onset-rhyme model for creating hidden Markov models. The Thai text corpus is then used in recording of speech corpus.

3.3.1 Criteria in Building a Thai Continuous Speech Corpus

In this dissertation, speaking style is controlled to the dictation or reading style. The text data were then collected from a series of Aesop's Fables in Thai for story-telling. The text were selected not to contain any foreign words. A total of seven Aesop's Fables were used and analysed on distribution and amount of onset and rhyme units. The set contains over a hundred sentences. In order to create an initial HMM for each onset and rhyme unit, a number of training samples must be sufficient. Therefore, lists of Thai words that share the same onset units were created. These words were then used in composing sets of sentences to fulfill as much samples as possible on each onset unit.

These sentences were frequently analysed on statistics of each onset and rhyme units. The final sets contain a total of 400 sentences. These sentences are composed of distinct 384 onset units and 144 rhyme units. The pronunciation dictionary contains the total of 2,250 Thai words collected from these sentences. The set of 2,250 words is composed of 1,650 distinct syllables. These syllables contain combinations of both onset units and rhyme units in various context.

3.3.2 Recording of Thai Utterances

Sets of Thai sentences from the Thai text corpus were analysed and used in recording of each Thai sentence. Recording was taken place in the quiet laboratory environment. The resolution of 16-bit at 16 kHz sampling frequency were used in recording of each Thai sentence. Two microphones are used in recording simultaneously into separate left and right stereo channels. Such recording method gives out two utterances in one utter. This way of recording will give out two samples of each utterance at the same time. All of the recorded utterances were recorded in stressed dictation style or reading style. The total of 625 Thai sentences were recorded which contain more then 5 hours of continuous utterances. The 625 sentences are composed of 557 sentences for training and the other 68 sentences for testing. The speech corpus contains continuous utterances of a single male speaker.

3.3.3 Labelling of the Recorded Thai Utterances

All of the recorded utterances were then hand-labelled by their phonetic transcriptions of each sentence. Labelling was done manually by the "Speech Labeller" labelling program created by the Thai speech processing research group at the Digital Signal Processing Research Laboratory. In labelling, understanding about acoustic characteristic of Thai continuous speech is very essential in determining the location and boundary of each phone within a sentence. Labelling of all speech were mostly done by the author and some by members of the Thai speech processing research group. Three output labels are created in phones, in contextual onset-rhyme models, and in phonotactic onset-rhyme models. Format of the output labels are conforming to the Hidden Markov Toolkit (HTK) format (Young, et al., 2000).

3.4 The Thai Continuous Speech Recognition System

In building a Thai continuous speech recognition system, the hidden Markov model toolkit (HTK) is utilized (Young, et al., 2000). The toolkit provides a variety of tools for speech processing, feature extraction, training, and recognition. Pronunciation dictionaries of each word were generated based on the Thai text corpus. The pronunciation dictionaries contain one-to-one mapping of each word into a sequence of recognition units, in this case, phone units, contextual onset-rhyme units, and phonotactic onset-rhyme units respectively.

3.4.1 Speech Signal Processing and Feature Extraction

All the utterances were recorded at 16 kHz sampling frequency and 16-bit resolution. The recorded utterances are preemphasized using the first-order filter with a coefficient of 0.97 (Rabiner and Juang, 1993; Lee, 1989; Lee, et al., 1990; Juang and Furui, 2000; Furui, 2001). The preemphasized speech data are then blocked into 25-ms frame at every 5 ms with the Hamming window applied.

Acoustic features of speech signals are extracted from the preprocessed speech. The Mel-frequency cepstral coefficients (MFCC) are employed as acoustic features representing the speech signal (Lee, 1989; Lee, et al., 1990; Juang and Furui, 2000; Furui, 2001). The MFCC are utilized in many continuous speech recognition (CSR) systems (Lee, 1989; Lee, et al., 1990; Juang and Furui, 2000; Furui, 2001; Huang, Acero, and Hon, 2001), which are then served as standard basic feature for a CSR system (Furui, 2001; Huang, Acero, and Hon, 2001). The MFCC are then employed in this research to provide the same standard configuration as other CSR systems. The 24-order speech feature vectors are computed from every speech frame which composed of 12-order MFCC feature vector and their 12-order time derivatives.

3.4.2 Acoustic Modelling of Speech Units

In building the Thai continuous speech recognition systems, three systems were set up using three different acoustic models in each system. The three acoustic models are phone models, contextual onset-rhyme models, and phonotactic onset-rhyme models. The phone-based system is a baseline system for comparison to the other two onset-rhyme models. Details of the three recognition systems are described in this section.

In determining the length of onset-rhyme hidden Markov models (HMMs), the model length is based on the length of a phone HMM. In the experiments, the phone HMMs were set at 3 active states with other two free connecting states at the beginning and the end of models. These states are illustrated in Figure 3.13. The active states are shown in dark circle with self-loop. For the onset-rhyme HMMs, the onset HMMs has 5 active states and the rhyme HMMs has 8 active states with two mixtures per state, which is called “m2s5s8” configuration. These onset-rhyme HMMs also have free connecting states, one at the beginning and one at the end of each model like the phone HMMs. In the experiments, the length of onset-rhyme HMMs are varied to see the effects of variable model length. Hence, another experiments use 4 active states in the onset HMMs and 6 active states in the rhyme HMMs with three mixtures per state, which is called “m3s4s6” configuration accordingly. The experiments were also set up by different overlap schemes, fixed and variable duration overlaps. Details of these experiments will be described in the next chapter.

In Figure 3.13, the phone HMMs along with the onset and rhyme HMMs are illustrated in the figure. These HMMs are used in recognition of continuous speech, which is treated as concatenation of speech units. In Figure 3.14, the bottom-up approach is depicted in recognition of the phrase /khiian4 tuua0 leek2/ using the onset-rhyme models. For example, the onset HMM “kh_iia” is time-aligned and matched to the speech. A set of rhyme HMM in the same “iia” context are match synchoronously and resulted in the rhyme HMM “iia_n”. Then, a pair of onset unit “kh_iia” and rhyme unit “iia_n” are formed as a syllable “khiian”.This recognition process is then repeated to the entire speech. The pairs of onset and rhyme units

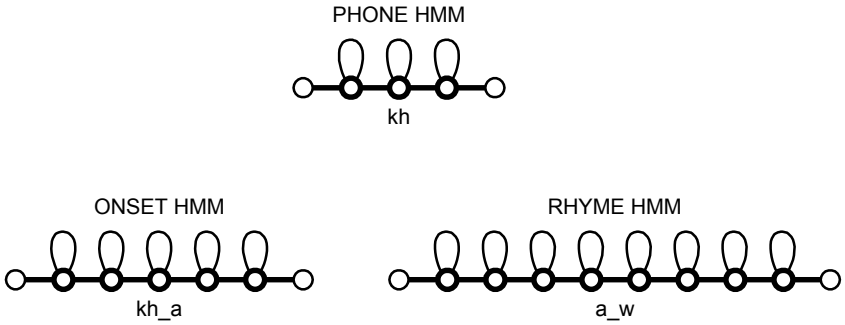


Figure 3.13 HMMs of phone, onset unit, and rhyme unit.

make up syllables, words, and the whole sentence, respectively. This is the bottom-up approach utilized in continuous speech recognition beginning from the smallest segmental units upto the whole sentence.

Modelling of the three acoustic models, phone models, contextual onset-rhyme models (CORMs), and phonotactic onset-rhyme models (PORMs), are described in details as follows. Both of the onset-rhyme models are introduced in this dissertation for modelling of the onset units.

A. Thai Phone Models

In modelling acoustic units using phone models, there are 54 phone models which are composed of 18 monophthongs, 3 diphthongs, 21 consonants, 11 consonant clusters and one silence model. A diphthong is considered as a single-unit model. Also, only the three long-vowel diphthong are modelled, /iia, vva, uua/, resulted from much higher number of occurance. Each phone model is modelled using a 3-state left-right hidden Markov model with three gaussian mixtures for each output probability density function. Initial phone models have been created and reestimated using labelled utterances. Then, embedded training have been applied to reestimate the trained phone models using unlabelled utterances. The list of Thai phone models is shown in Table 2.1, 2.2, and 2.3.

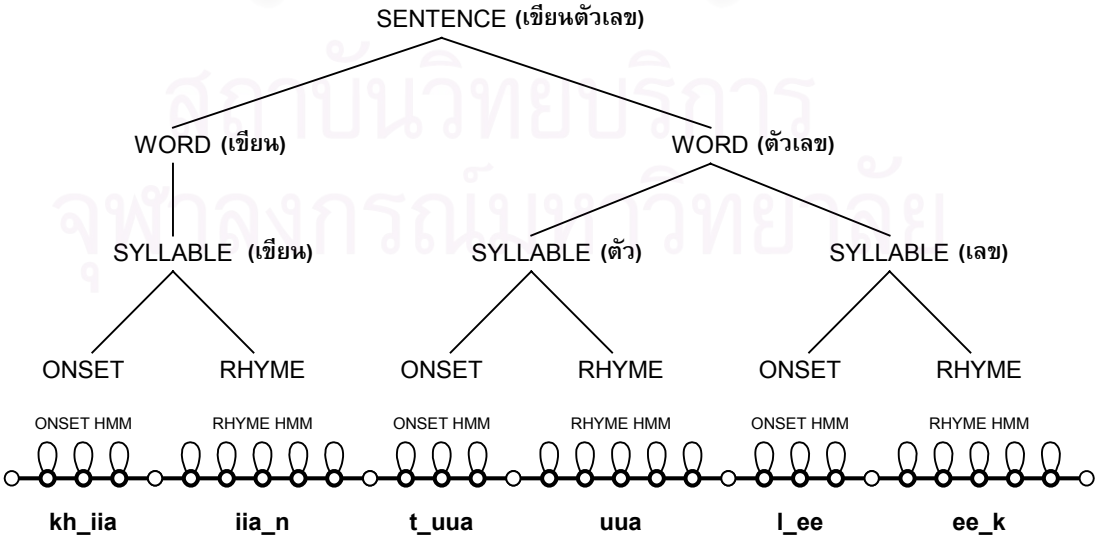


Figure 3.14 The bottom-up approach using the onset-rhyme models on an example phrase /khiian4 tuua0 leek2/ or “เขียนตัวเลข”.

B. Thai Contextual Onset-Rhyme Models (CORMs)

From analysis of Thai syllable structure, there are a total of 497 contextual onset-rhyme models which are composed of 297 onset units and 200 rhyme units. Due to limited data, only a partial set of 363 onset-rhyme models are modelled which contain 218 onset models and 144 rhyme models in this research. Since there are limited training data, the 363 onset-models are analysed from text corpus which are specially created for recording and training of the onset-rhyme model. The speech corpus contains substantial amount of samples sufficient for initialization and training of each model.

In training of the onset-rhyme models, labelled phonetic transcription of each utterance were generated in the CORMs format. An onset unit contains transitional period between a releasing consonant and its adjacent vowel. This resulted in overlapping of an onset unit over the vowel segment of the following rhyme unit. Consequently, two types of overlap are utilized, fixed duration overlap and variable duration overlap, as stated in the previous chapter. Using fixed duration overlap, the length of overlap are preset at either 10ms, 15ms, 20ms, or 30ms respectively. Using variable duration overlap, the length of overlap are determined as percentage of the vowel duration at either 10%, 15%, 20%, or 25% respectively. Then, the labelled onset-rhyme transcription are generated according to types and values of overlap.

In acoustic modelling, the left-right hidden Markov models (HMMs) are used with different number of states between the onset HMMs and the rhyme HMMs. An onset unit and a rhyme unit are modelled by 5-state and 8-state HMMs respectively. Both of the HMMs utilize two Gaussian mixtures for each output probability density function. The left-right HMMs are used

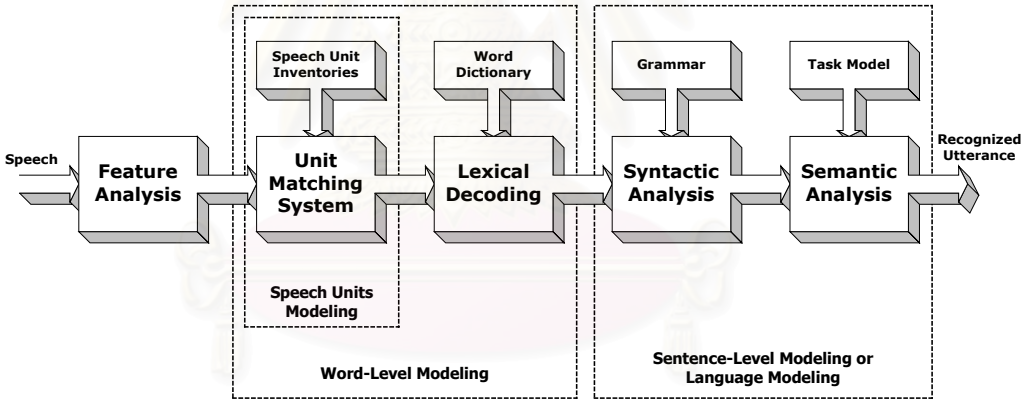


Figure 3.15 The general concept of a continuous speech recognition system.

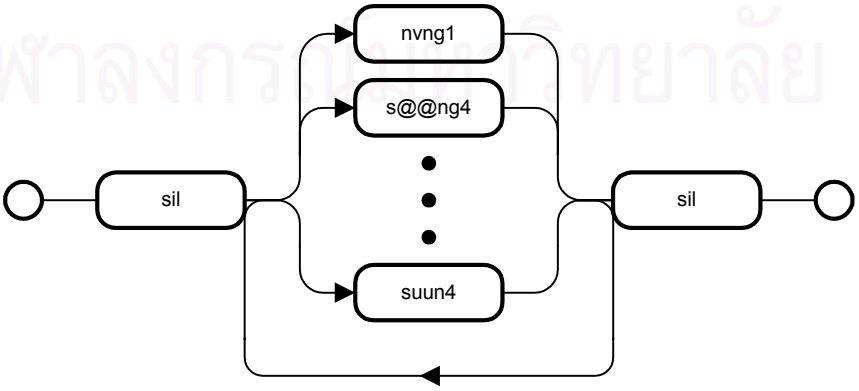


Figure 3.16 Example of a word lattice network.

at 5 states for an onset unit and 8 states for a rhyme units. Both of the onset and rhyme hidden Markov models use two Gaussian mixtures for each output probability density function. The initial onset-rhyme models were created and reestimated using labelled utterances. After labelled training, the trained acoustic models were reestimated using unlabelled utterances by embedded training. From analysis of the Thai text corpus, distribution of the contextual onset-rhyme models are shown in Table A1.2. In Table A1.2, statistics of each model are listed that was designed to have sufficient number of occurrences of each model.

C. Thai Phonotactic Onset-Rhyme Models (PORMs)

In analysis of the models, the Thai phonotactic onset-rhyme models have 992 onset-rhyme models that contain 792 phonotactic onset units and 200 rhyme units. The phonotactic onset-rhyme models have more extensive context-dependent onset units than the contextual onset-rhyme models as shown in Table 1.5 in Chaper 1. This model considers a releasing consonant in different vowel context as separate models. The network of phonotactic onset-rhyme models is shown in Figure 3.12 which shows complete combinations between the onset and the rhyme models.

Due to limited training data, only 528 onset-rhyme units were created and modelled. The 528 onset-rhyme units contain 384 onset units and 144 rhyme units. Like the contextual onset-rhyme models, the labelled phonetic transcriptions are generated in phonotactic onset-rhyme models. A list of all 528 phonotactic onset-rhyme units are shown in Table A1.3 in the Appendix A. In Table A1.3, all 528 phonotactic onset-rhyme models are shown along with their distributions. The Thai text corpus was designed to accommodate all the models with sufficient samples for creating initial models.

3.4.3 Architecture of the Recognition System

The architecture of the recognition system is shown in Figure 3.15, where a general conceptual model of a continuous speech recognition system is illustrated. In Figure 3.17, training of speech units is depicted, using both labelled and unlabelled training data. In Figure 3.18, recognition procedure is illustrated. During recognition, there are no language models or any grammars applied in the decoding process. This means that any words can follow any other words with optional silence as illustrated in Figure 3.16. The unigram model or no grammar means each word has uniform probability of occurance. The word probability of a unigram model is shown in Eq. (3.2) in Section 3.1. These system configurations are described in the HTK manual (Young, et al., 2000).

3.5 Summary

The concept of the onset-rhyme models are described in details along with advantages over other acoustic modelling of the models. Three types of the onset-rhyme models are introduced, namely, theoretical onset-rhyme models, contextual onset-rhyme models, and phonotactic onset-rhyme models. The onset-rhyme models comprise an onset unit and a rhyme unit. The onset unit contains transitional period into the adjacent vowel nucleus, which overlaps into the rhyme unit. In modelling of the onset unit, two types of overlap schemes are proposed, the fixed duration overlap and the variable duration overlap.

In this chapter, modelling of the onset and rhyme units is explained using the hidden Markov models (HMMs). The bottom-up approach used in recognition is illustrated using the onset and rhyme units. This approach describes how a pair of onset and rhyme units forms a syllable, words, and sentence, respectively. The lattice networks of onset and rhyme HMMs are depicted in both contextual and phonotactic onset-rhyme models.

Operating environments of the recognition system are described in details in this chapter. The task domain of the corpus was based on Aesop's fables and other reading-style sentences. All of the recorded utterances were in reading style or dictation style. Utterance of a male speaker were recorded for both training and testing.

Moreover, details of the recognition system is described with acoustic modelling of phones and both onset-rhyme models. Only a partial set of onset-rhyme models were built and utilized in recognition system due to limited training data available. Only 363 models out of 497 contextual onset-rhyme models were selected based on the text corpus. Also, only 528 model out of 992 phonotactic onset-rhyme models were chosen from the text corpus. These onset-rhyme models comprise a set of 2,250 Thai words in the pronunciation dictionary. There are 1,650 distinct syllables within the set of 2,250 Thai words. These words were collected from the text corpus used in recording, training, and testing. Although only partial set of onset-rhyme models were utilized, this could be used in implementation of a small, task-specific Thai continuous speech recognition system. This kind of small system is much easier to optimize to have very high recognition accuracy.

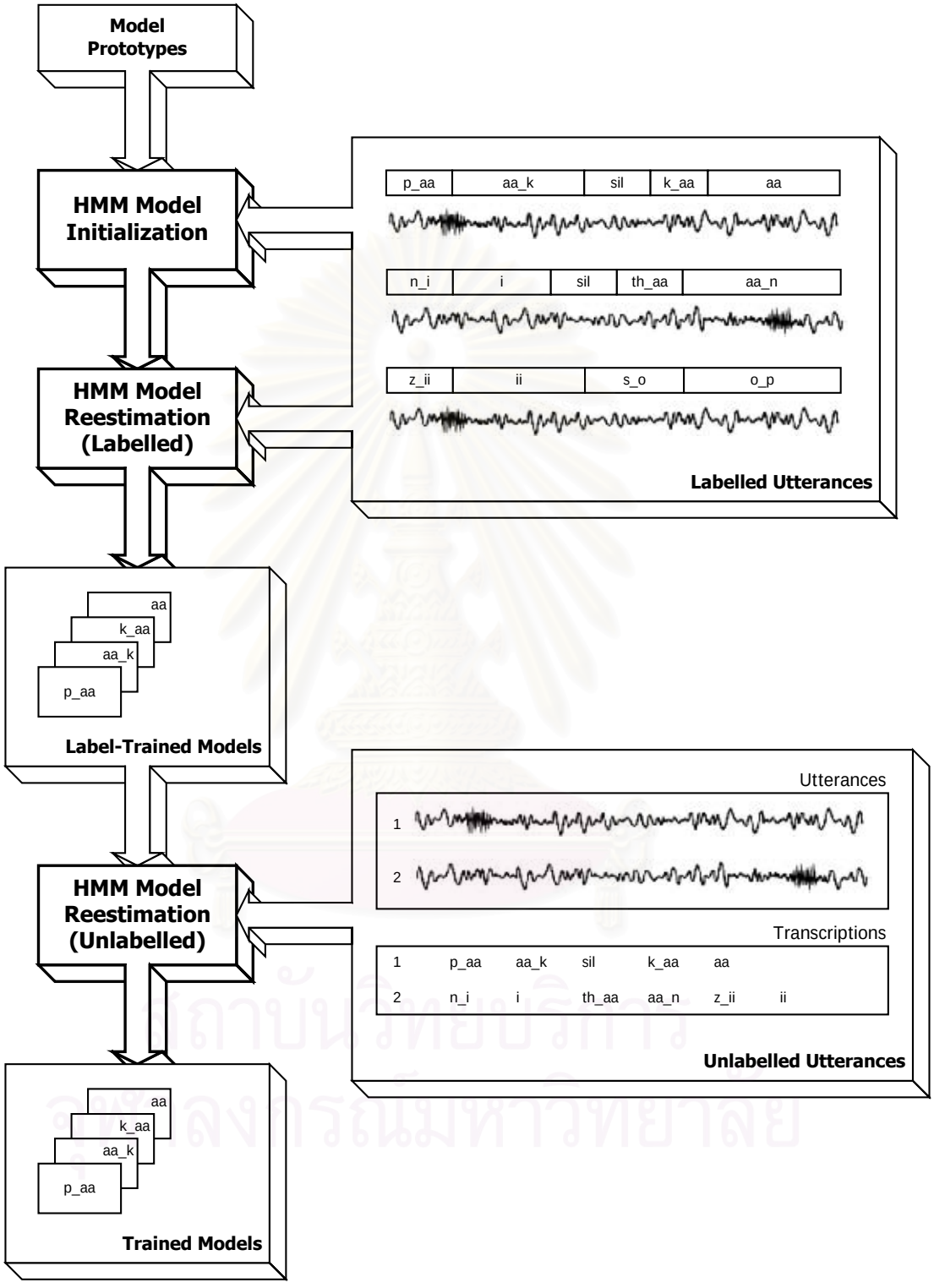


Figure 3.17 Hidden Markov model training process (adapted from Young et al. (2001)).

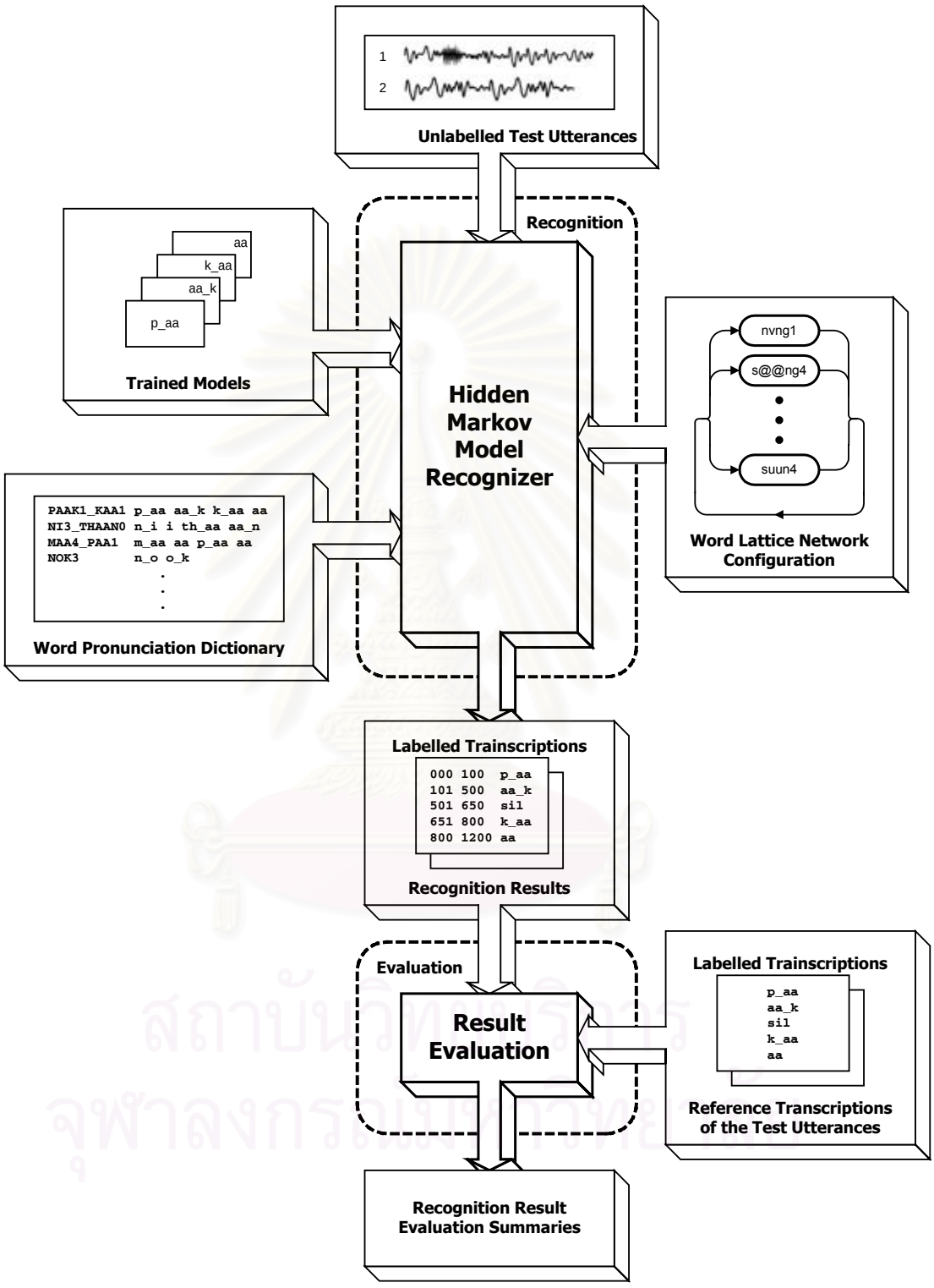


Figure 3.18 Hidden Markov model recognition process (adapted from Young et al. (2001)).

CHAPTER 4

Experimental Results and Discussions

In this chapter, results are collected from series of experiments. The forced alignment and recognition are employed in evaluation of the acoustic models. These acoustic models are phone models, contextual onset-rhyme models, and phonotactic onset-rhyme models. Results from both evaluation methods are compared among the three acoustic models. Result analysis and discussions are given in this chapter.

4.1 Evaluation of Acoustic Models using Forced Alignment

In evaluation of acoustic models, forced alignment are employed in order to determine precision of model boundaries. The forced alignment procedure performs recognition based on a provided word-level transcription of a particular utterance. The word lattice network is then constructed based on the given transcription of word sequence. Pronunciation dictionaries provide description of a word by its composition of acoustic units. For examples, the word "D@@K1 MAAJ3", or flower, is composed of the phones /d @@ k m aa j/ or the contextual onset-rhyme models /d@@ @@_k maa aa_j/. Therefore, each word in the lattice network is expanded based on acoustic unit composition in the pronunciation dictionary and aligned over a proper location on the utterance. The alignment is done using the trained hidden Markov model of each acoustic unit by matching against the most probable speech segments. The procedure is then repeated sequentially on every word in the network. Output of the forced alignment procedure provides time alignment or model boundary information of each acoustic unit on the expanded word lattice network.

In evaluation of acoustic units, the hand-labelled phonetic transcriptions are utilized as reference time alignment. Shifting in syllable boundaries is then computed on each test sentences by comparing to the reference syllable boundaries in both syllable starting and syllable ending. Boundary shifting values are collected and statistically analysed on all acoustic units--phones, contextual onset-rhyme models, and phonotactic onset-rhyme models. The statistical analysis results are shown in Table 4.1 on phone models, in Table 4.2 on contextual onset-rhyme models, and in Table 4.3 on phonotactic onset-rhyme models respectively.

4.1.1 Results and Evaluation of Forced Alignment

Forced alignment results are illustrated in Figure D1.1 to D1.32 in the Appendix D on each test sentences. The figures show hand-labelled time alignment using phone sequence of a test sentence along with the alignment results using the phones, the contextual onset-rhyme models (CORMs), and the phonotactic onset-rhyme models (PORMs). The hand-labelled phonetic transcriptions shown in the figures are used as reference boundary alignment.

Syllable boundaries of each syllable are compared against the hand-labelled syllable boundaries of a particular utterance. Syllable boundary positions from hand-labelled transcriptions are employed as reference points on evaluation of forced alignment. The forced alignment results give out positions of syllable boundaries according to the predefined word transcriptions of a particular utterance. Deviations of syllable boundaries are computed on each corresponding syllable between hand-labelled and forced alignment results. Statistical analyses on amount of deviations or shifts are shown in Table 4.1 using the phone models, Table 4.2 using the contextual onset-rhyme models, and Table 4.3 using the phonotactic onset-rhyme models, respectively.

Table 4.1 Statistical analysis results on shifting in syllable boundaries using phone models in forced alignment

Model Configuration	Shift in syllable beginning (ms)				Shift in syllable ending (ms)			
	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
m3s5	0	860	33.557	70.632	0	725	65.133	95.867

4.1.2 Discussions

In Table 4.1, the phone models show mean boundary shift at 33.56 ms on beginning and at 65.13 ms on syllable ending. In Table 4.2, the statistical analyses on contextual onset-rhyme models (CORMs) are shown on both fixed and variable duration overlap. The 20% variable duration overlap gives out the lowest mean boundary shift at 14.15 ms on beginning and at 32.31 ms on syllable ending. All cases of the CORMs have shown better alignment than the phone models. The mean of syllable shifts in starting and ending for the CORMs are 57.84% and 50.39%, respectively, lower than the phones. Like the CORMs, results of the statistical analyses on the phonotactic onset-rhyme models (PORMs) also show better alignment than the phones as stated in Table 4.3. The 20% variable duration overlap gives out the lowest mean boundary deviation at 15.32 ms on beginning and at 34.41 ms on ending of syllable. The mean deviation of syllable starting and ending for the PORMs are 54.35% and 47.17% respectively better than the phones. The CORMs and PORMs give out comparable results in both cases as shown in Table 4.2 and Table 4.3, respectively. Both cases also illustrate much lower standard deviation (S.D.) in syllable boundary shift than the phones.

Moreover, both cases of fixed and variable duration overlap provide similar results in syllable boundary shifts on forced alignment. From Figure D1.1 to D1.32, the forced alignment results are depicted with all speech units comparing to the hand-labelled data, the phones, the CORMs, and the PORMs. Both of the subsyllable onset-rhyme models show better time alignment than the phones and also have many advantages described as follows.

The onset units show much more precise time alignment than the phones in almost every cases. In the case of an arresting nasal followed by the same releasing nasal, the onset unit performs much more accurate than the phones in boundary alignment between the two adjacent syllables such as /t@@n nang/ as shown in Figure D1.9. This is also existed in the case of two consecutive approximants, /j, w/, as arresting and releasing consonants between the two syllables such as /lqqj jaang/. Boundary shifting in onset units mostly occurs with the releasing voiced stops, /b, d/, especially after an open syllable or an arresting non-stops. These shifts are resulted from acoustic characteristics of the two voiced stops themselves. The voiced stops /b/ and /d/ have voicing or periodic characteristics similar to vowels and non-stops. There are continuity in spectrum and in fundamental frequency of the voiced stops to the preceding vowel in an open syllable or the preceding arresting non-stops. Thus, these lead to more difficult in locating syllable beginning boundary than in other consonants. Examples of this case are illustrated in Figure D1.19 at the syllable /muu baan/, in Figure D1.20 at the syllable /ngaan d@@k maaj/, and in Figure D1.21 at the syllable /kaan buk ruk/ and /chaaj dxxn/.

In the figures, overlapping of the onset unit into the adjacent rhyme unit are clearly depicted with a longer duration of the onset unit. The phone models give out higher average boundary shift because of their context independence especially in both releasing and arresting stops. Acoustically speaking, each phone is effected by its neighbouring phones resulted from coarticulation and contextual effects. Thus, assuming that the same phone is similar across different context leads to an inefficient modellng of utterances. This is resulted in an inaccurate time alignment especially the consonants that can be both releasing and arresting.

Table 4.2 Statistical analysis results on shifting in syllable boundaries using contextual onset-rhyme models in forced alignment

Model Config	Overlap Type	Shift in syllable beginning (ms)				Shift in syllable ending (ms)			
		Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
m3s6s8	10ms	0	335	15.066	24.892	0	410	34.519	44.812
m3s6s8	20ms	0	340	16.368	25.707	0	555	40.365	51.464
m3s6s8	30ms	0	340	14.925	24.809	0	555	35.008	47.242
m2s7s10	10ms	0	330	17.156	27.124	0	540	44.013	53.186
m2s7s10	20ms	0	335	16.012	25.076	0	540	43.192	53.274
m2s7s10	30ms	0	335	17.040	26.867	0	555	44.212	52.714
m3s6s8	25pct	0	340	17.007	28.130	0	225	37.960	41.995
m2s6s10	05pct	0	330	16.202	27.885	0	225	35.920	41.229
m2s6s10	10pct	0	335	16.153	27.067	0	225	36.882	41.464
m2s7s10	15pct	0	300	15.680	25.911	0	560	37.313	48.358
m2s7s10	20pct	0	305	14.146	25.127	0	495	32.305	43.446
m2s7s10	25pct	0	335	15.207	26.678	0	420	35.133	43.470

Table 4.3 Statistical analysis results on shifting in syllable boundaries using phonotactic onset-rhyme models in forced alignment

Model Config	Overlap Type	Shift in syllable beginning (ms)				Shift in syllable ending (ms)			
		Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
m3s6s8	10ms	0	335	17.960	27.519	0	540	43.483	53.650
m3s6s8	20ms	0	340	17.247	26.819	0	260	42.836	49.355
m3s6s8	30ms	0	340	17.222	27.942	0	540	44.842	55.813
m2s7s10	10ms	0	330	17.529	27.524	0	555	44.303	52.543
m2s7s10	20ms	0	335	17.040	27.072	0	410	44.478	51.897
m2s7s10	30ms	0	335	16.849	26.624	0	540	43.682	53.960
m3s6s8	25pct	0	340	16.161	26.579	0	215	32.313	38.105
m2s6s10	05pct	0	335	17.371	27.900	0	220	38.018	41.769
m2s6s10	10pct	0	340	17.164	27.800	0	220	38.300	41.692
m2s7s10	15pct	0	335	16.352	27.466	0	420	37.512	44.774
m2s7s10	20pct	0	335	15.323	25.927	0	215	34.411	40.331
m2s7s10	25pct	0	335	15.879	26.760	0	230	36.816	42.935

In summary, both of the onset-rhyme models provide accurate thus precise time alignment of syllable boundaries. The onset-rhyme models efficiently reduce boundary shift at more than 50% compared to the phone models. Therefore, accurate syllable boundary information also help improving tone recognition by providing precise location of a syllable.

4.2 Evaluation of Acoustic Models by Recognition

In evaluation of acoustic models, the onset-rhyme models are applied to the Thai continuous speech recognition. Then, analysis on the recognition results are conducted in order to explore any improvement in word error rate. The acoustic models, phone models, contextual onset-rhyme models, and phonotactic onset-rhyme models, were used in the recognition. The word error rate is an index, which describes an amount of incorrectly recognized words. Confusion matrices of each speech unit, phone and onset units, show accuracy of speech units in forming a word or a syllable. In evaluation of speech units, the resulting speech unit sequences are compared against correct speech unit sequence of each sentence.

The onset-rhyme models utilise overlapping between an onset unit and a rhyme unit within a syllable. The overlapping of an onset unit covers transitional stage between a releasing consonant and its adjacent vowel, which is the nucleus of a syllable. In determining an amount overlap, two overlapping methodologies are proposed, fixed duration overlap and variable duration overlap. The fixed duration overlap uses a set predefined length of overlap at 10 ms, 20 ms, or 30 ms into the adjacent rhyme unit. On the contrary, the variable duration overlap varies a length of overlap to 5%, 10%, 15%, 20%, or 25% over duration of the adjacent vowel. Therefore, both contextual and phonotactic onset-rhyme models were utilized in a series of experiments covering the fixed and variable duration overlaps. The recognition results of these experiments are illustrated in Table 4.4 to 4.8.

In the fixed duration overlap, the set of predefined length was selected based on length of vowel. Length of speech units are collected and statistically analysed from the speech corpus. The results showed only 30 ms on minimum length of a short vowel. Therefore, the overlap duration may include some parts of a coda if longer than 30 ms. As stated in the Section 3.2.2 in Chapter 3, the fixed duration is then set to be less than 30 ms at 10 ms, 20, or 30 ms, respectively.

In order to find an appropriate hidden Markov model configurations, two sets of experiments were set up. In the first experiment, hidden Markov model configurations are set to 4 states for an onset unit and 6 states for a rhyme unit with three mixtures per state. In the second experiment, a longer hidden Markov model configurations are set to 5 states for an onset unit and 8 states for a rhyme unit with two mixtures per state. Then, these two HMM configurations are called “m3 s4 s6” and “m2 s5 s8”, respectively, throughout this dissertation.

4.2.1 Evaluation of Recognition Results

This section contains details about all the recognition result of all the acoustic models. The acoustic models used in the experiments are phone models, contextual onset-rhyme models (CORMs), and phonotactic onset-rhyme models (PORMs). Evaluation details of recognition results are shown in this section on each acoustic models. Each HMM configuration is used in series of experiments on fixed and variable duration overlap. In the “m3s4s6” case, there are only 20% and 25% overlap because many models of both onset and rhyme units could not be created as shown in Table 4.6 and 4.8. This is due to shorter duration in lower overlap percentage which do not provide substantial amount of data for training. The concept of selection an amount of overlap is stated in Section 3.2.2 in Chapter 3 on both fixed and variable duration overlap.

Table 4.4 Best word error rate achieved using different acoustic models

Acoustic Units	Word Error Rate (%)
Phone model	37.12
Contextual onset-rhyme models at 15% overlap	16.518
Phonotactic onset-rhyme models at 20% overlap	13.529

A. Phone Models

Recognition result of the phone model is shown in Table 4.4. The result shows accuracy at 62.88% or at 37.12% word error rate using the phone models on the 68 test sentences. Recognized phone sequence of each test sentence are evaluated with their correct phonetic transcriptions. Evaluation results are then analysed based on three types of recognition errors—insertion error, deletion error, and substitution error.

Table 4.9 illustrates three insertion errors with examples, insertion errors on vowel phonemes, on initial consonants, and on initial nasals. In cases of insertion errors, insertion of another vowel phonemes results in two subsequent vowel phonemes as shown in the first case of Table 4.9. Besides vowel, insertion of another consonants also occurs which resulted in two subsequent arresting and releasing consonants as shown in the second case of Table 4.9. Moreover, these are also insertion of a voiced stop follows by a releasing nasals which results in two subsequent arresting and releasing consonants as shown in the third case of Table 4.9.

Two different deletion errors are shown in Table 4.10 with examples, deletion errors on arresting stops, and, on two adjacent nasals. Both cases of deletion error result from different causes. Deletion of an arresting consonant is caused by phone modelling which considers both initial and final stops identical. In the case of two adjacent non-stops, there are treated as single non-stop which results from continuity of the two consonants with no explicit word boundary or syllable boundary. Table 4.11 shows five substitution errors with examples, substitution errors between short and long vowels, between voiced stop and nasal, within a group of voiceless stops, within a group of nasals, and on consonant clusters. First, short and long vowel pairs are incorrectly recognised between their counterparts, for example, the word /kan0/ is recognised as /kaan0/ as shown in the first case of Table 4.11. Secondly, some initial nasals are incorrectly recognized as voiced stops such as /ma3 naaw0/ is recognised as /baan laaw/. There are also some recognition errors within the same group of voiceless unaspirated stops or nasals. Moreover, some consonant clusters are incorrectly recognised as shown in the last case of Table 4.11.

The evaluation of each test sentences is shown in the Appendix C. From the results, many kinds of errors are found in both confusion matrix and sentence evaluation. From the confusion matrix, various kinds of recognition errors exist as follows. Firstly, recognition error exists between a pair of short and long vowels such as /i/-/ii/ and /a/-/aa/. There are also some vowel pairs that are misrecognized between each other such as /e/-/ee/, /@/-/@@/, /u/-/uu/, and /o/-/oo/. Secondly, most recognition errors exist within the same group of phones such as stops, etc. In a group of stops, the /c/ stop is incorrectly recognized as /t/ or /k/; the /t/ is incorrectly recognized as /p/, /t/, or /c/; and the /th/ is incorrecly recognized as /t/, /kh/, /ph/, /ch/. This kind of error indicates incorrect recognition within the same manner but different places of articulation. There are also incorrect recognition to different manner of articulations such as /k/ to /kh/, /p/ to /th/, and /th/ to /t/. In a group of non-stops, there are recognition errors within the same group such as in the /m, n, ng/ group of the same manners but different places of articulation. Thirdly, some of the consonant clusters are incorrectly

Table 4.5 Average word error rate of onset-rhyme models using fixed-duration overlap on different state sizes.

Acoustic Units		m3 s4 s6 (1)			m2 s5 s8 (2)		
		10 ms	20 ms	30 ms	10 ms	20 ms	30 ms
Contextual models	onset-rhyme	35.394	35.113	34.957	27.572	27.591	27.785
Phonotactic models	onset-rhyme	37.946	37.190	36.675	27.019	26.213	25.427

Remarks : (1) – 4-state onset HMM and 6-state rhyme HMM at 3 mixtures per state.
(2) – 5-state onset HMM and 8-state rhyme HMM at 2 mixtures per state.

Table 4.6 Average word error rate of onset-rhyme models using variable-duration overlap on different state sizes.

Acoustic Units	m3 s4 s6		m2 s5 s8				
	20%	25%	5%	10%	15%	20%	25%
Contextual onset-rhyme models	20.080	20.487	17.623	17.411	16.518	16.770	16.790
Phonotactic onset-rhyme models	-	18.100	17.226	16.431	13.985	13.529	14.334

Remarks : (1) – 4-state onset HMM and 6-state rhyme HMM at 3 mixtures per state.
(2) – 5-state onset HMM and 8-state rhyme HMM at 2 mixtures per state.

recognized such as /khr, kr/ into /r/ and /pr/ into /r/. Additionally, a voiced stop /b/ is incorrectly recognized as /d/ and the nasal /n/.

B. Contextual Onset-Rhyme Models (CORMs)

The model network is illustrated in Figure 3.11 in Chapter 3. The network is used in creating sequence of onset models and rhyme models to form syllables, words, and a sentence respectively. An example of the onset-rhyme model alignment from recognition is depicted in the Appendix D. Recognition results using the onset-rhyme models is shown in Table 4.5 to 4.8. The result shows unit accuracy at 83.482% or at 16.518% error rate. The onset-rhyme models reduce error rate up to 55.50% compared to the phone model. Recognized sequences of each test sentence are evaluated with their correct onset-rhyme transcriptions. Evaluation results are then analysed on three types of recognition errors—insertion error, deletion error, and substitution error. There are no deletion errors occurred using the onset-rhyme models. These errors will be discussed later in the next section.

In Table 4.12 and 4.13, various types of onset recognition errors are shown with examples from the recognition results in the Appendix C. Firstly, an onset with releasing voiceless unaspirated stop is incorrectly recognized to aspirated stop in the same place of articulation. Secondly, an onset with releasing voiceless stop is incorrectly recognized to another voiceless stop with the same manners. Thirdly, an error occurs within a group of voiced stop /b, d/ and also between a group of nasals /m, n, ng/. Moreover, a consonant cluster is misrecognized into either its aspirated-unaspirated pair of stops or its secondary clusters /r, l, w/. Additionally, there are very few occurrence on removal of an onset unit in the case of diphthongs, which are only found in contextual onset-rhyme models.

Table 4.7 Average error rate of the onset units using fixed-duration overlap.

Acoustic Units	m3 s4 s6 (1)			m2 s5 s8 (2)		
	10 ms	20 ms	30 ms	10 ms	20 ms	30 ms
Contextual onset-rhyme models	24.601	24.115	23.862	17.992	17.871	18.492
Phonotactic onset-rhyme models	34.769	34.259	33.690	23.993	23.357	22.608

Remarks : (1) – 4-state onset HMM and 6-state rhyme HMM at 3 mixtures per state.
(2) – 5-state onset HMM and 8-state rhyme HMM at 2 mixtures per state.

Table 4.8 Average error rate of the onset units using variable-duration overlap.

Acoustic Units	m3 s4 s6 (1)		m2 s5 s8 (2)				
	20%	25%	5%	10%	15%	20%	25%
Contextual onset-rhyme models	12.941	13.732	11.567	11.264	10.523	10.833	10.387
Phonotactic onset-rhyme models	-	16.129	14.811	14.211	12.140	11.753	13.074

Remarks : (1) – 4-state onset HMM and 6-state rhyme HMM at 3 mixtures per state.
(2) – 5-state onset HMM and 8-state rhyme HMM at 2 mixtures per state.

In Table C2.1 to C2.32 in the Appendix C, the recognition results of each test sentence are shown in sequences of contextual onset-rhyme models. Recognized sequence of each test sentence (REC) is then evaluated by comparing to their correct transcription (LAB). The results of evaluation show many recognition errors as summarized in Table 4.12 and 4.13. Coarticulatory effects are major cause of these recognition errors. For example, adjacent syllables within the word /ma3 naaw0/ in the “testsentence01_016a_vis” was incorrectly recognized as /maa a_n zaa aa_w/ instead of /maa a naa aa_w/. The releasing nasal /n/ was treated as arresting nasal in the rhyme of its preceding syllable. In addition, many diphthongs were incorrectly recognized by insertion of a syllable such as /khrvvang2/ was recognized to /rvv vv daa a_ng/ instead of /khrvv vvang/, which shown in the “testsentence01_004b_vis”.

(1) Fixed-duration overlap

The recognition results of the contextual onset-rhyme models using fixed-duration overlap are shown in Table 4.5. In Table 4.5, recognition results of the two experiments are shown where each experiment conducts three different overlap length at 10 ms, 20 ms, or 30 ms.

From the results in Table 4.5, the 10-ms overlap of m2s5s8 configuration shows better result at 27.572% word error rate than any other configurations. All cases of the m2s5s8 configuration show significant decreasing in word error rate compared to the m3s4s6 configuration. The reduction in word error rate is resulted from a longer HMM states in modelling of both onset and rhyme units.

(2) Variable-duration overlap

In Table 4.6, recognition results of the contextual onset-rhyme models using variable-duration overlap are shown. Two experimental results of m3s4s6 and m2s5s8 configurations are

shown in Table 4.6. In Table 4.6, the recognition result lower than 20% overlap in the m3s4s6 configuration is not available because of very short duration of each speech segment and limited training data. Creating an initial HMM model using three mixtures per state requires substantial amount of data and duration of each speech segment must be sufficient for training.

Comparing both m3s4s6 and m2s5s8 configurations, the second configuration shows significantly better results as shown in Table 4.6. Increasing percentage of overlap effects only small reduction in word error rates as shown in Table 4.6. However, there are significant decreasing in word error rates using m2s5s8 compared m3s4s6 configuration as shown in Table 4.6. The reduction in word error rate not only resulted from a longer HMM states but also from amount of acoustic information captured from the transition period.

C. Phonotactic Onset-Rhyme Models

The model network illustrated in Figure 3.12 in Chapter 3 is employed in generating sequences of phonotactic onset and rhyme units. The generated sequences of phonotactic onset and rhyme units form to be syllables, words, and sentence respectively. Recognition results using the phonotactic onset-rhyme models is shown in Table 4.5 to 4.8. The best recognition result is at 13.529% word error rate. The phonotactic onset-rhyme models reduce word error rate up to 63.553% compared to the phone models or up to 18.095% compared to the contextual onset-rhyme models.

In Table 4.12 and 4.13, many kinds of recognition errors are summarised with some examples from the recognition results in Table C3.1 to C3.32 in the Appendix C. Considering only the onset unit, voiceless stops are incorrectly recognized within their group of the same places of articulation, i.e., /k_@@/ to /kh_@@/. They are also misrecognized to other stops with the same manners of articulation, i.e., /ph_aa/ to /kh_aa/. There are some errors on voiced stops /b, d/, which are incorrectly recognized not only with their counterparts but also with nasals. Moreover, recognition of consonant clusters are also substituted with their secondary consonants, /r, w/, for example, /khr/ to /r/, and, /khw/ to /w/. In addition, there are few insertions occurred in some diphthongs such as /th_uua uua/ is recognized to /th_uu uu_k w_aa aa/ in some results as shown in Table 4.13.

The recognition results using phonotactic onset-rhyme models are shown in Table C3.1 to C3.32 in the Appendix C. Output of the result are shown in sequences of phonotactic onset and rhyme units on each test sentence. The results (REC) are then evaluated with the correct transcription (LAB) of each sentence. The outputs of evaluation show word error rate and confusion matrices using the phonotactic onset and rhyme units. The word error rates are shown in Table 4.5 to 4.8 on the entire experiments. These experiments are conducted based on two types of overlap, that is, the fixed-duration and variable-duration overlap. Also, the experiments are conducted on two different configurations of hidden Markov models, m3s4s6 and m2s5s8, as previously described at the beginning of this section.

(1) Fixed-duration overlap

In Table 4.5, recognition results of the phonotactic onset-rhyme models are shown using the fixed-duration overlap. The results are grouped by configurations of hidden Markov models. The best recognition results is at 25.427% word error rate using 30-ms overlap. The word error rates are significantly decreased comparing between the m3s4s6 and m2s5s8 configurations. At each overlap, the word error rates are reduced by 28.796% at 10-ms, by 29.516% at 20-ms, and by 30.669% at 30-ms overlap, respectively. These percentage of reduction illustrated major improvements as much as 30% using longer hidden Markov model states in both onset and rhyme units.

(2) Variable-duration overlap

In Table 4.6, recognition results using variable-duration overlap are shown. The configurations of m3s4s6 and m2s5s8 are shown. From both results, the best word error rate achieved is at 13.529% using 20% overlap. Like the fixed-duration overlap, the word error rate is significantly reduced using longer state in m2s5s8 than in m3s4s6. The word error rate is reduced by 20.807% using 25% overlap.

4.3 Discussions

Eventhough only utterances of a single speaker were used in training and recognition, higher recognition results were achieved with major improvements over the phones. From Table 4.4, comparison of recognition results between the phones and the onset-rhyme models has shown a large amount of reduction in error rate at 55.76% over the phones without any extra techniques or grammars. The onset-rhyme models always occur in pairs of the onset unit and the rhyme unit which makes up syllables, words, and sentence, respectively.

Comparing between the phones and both onset-rhyme models, the phones are more errornous than the onset-rhyme models. The evaluation results of the phones have a large number of substitution errors and insertion errors. Substitution errors of the phones are mostly occurred between short vowels and long vowels, initial stop consonants, and consonant clusters.

Recognized sequences of the onset-rhyme models compared to sequences of the phones illustrate some major point of improvements as shown in Table 4.9 to Table 4.13. Firstly, sequences of the phones are difficult to distinguish between releasing consonants and arresting consonants while this is not the case with the onset-rhyme models. Secondly, deletion of an arresting stop frequently occurs in the phones but not occur in the onset-rhyme models. Thirdly, syllables or even words could be simply determined from sequences of the onset and rhyme pairs as illustrated in Table 4.15 and Table 4.16 for contextual and phonotactic onset-rhyme models, respectively.

Also, there are some common errors between the phones and the onset-rhyme models. For instance, some short and long vowels are incorrectly recognized as shown in Table 4.12 for the phones and in Table 4.13 for the onset-rhyme models. Additionally, an open syllable was incorrectly recognized to have obstruent ending or arresting consonant which is a releasing consonant of the following syllable. Examples of this error are shown in Table 4.12 for the phones and in Table 4.13 for the onset-rhyme models. Moreover, a sonorant-ending syllable with similar arresting and following initial consonant is incorrectly recognized as an open syllable. Examples of this error are shown in Table 4.12 for the phones and in Table 4.13 for the onset-rhyme models.

In comparison, the theoretical onset-rhyme models were not employed in recognition. This is due to context independency of the models. The theoretical onset-rhyme models are context independent models. The onset units are the similar to the context-independent phone models. In addition, the onset units do not include the transitional period between releasing consonant and neighbouring vowels.

4.3.1 Phone Models

The phones have many errors of insertion, deletion, and substitution compared the phones to the two onset-rhyme models. There are plenty of insertion error on releasing consonants. These insertions result in ambiguities to be either arresting consonant of the prior syllable or releasing consonant of the following syllable. These insertions mostly occur in an open syllable within a word. There are also two repeated vowels in the recognized phone sequences as shown in Table 4.9, which do not exist in the onset-rhyme sequences.

The phones also show deletion of arresting stops in many recognition sequences as shown in Table 4.10. The absence of arresting stops is one of the major disadvantages. This type of error is resulted from acoustic characteristic of the Thai arresting stops themselves, which differ from other languages. Moreover, the phones could not recognize any two consecutive nasals or approximants, which occur as arresting and releasing consonants respectively. Examples of these error are shown in Table 4.10 such as the word /naam3 nak1/.

Besides the above errors, the phones have substitution errors between short and long vowels, between voiced stops and nasals, and between nasals. Examples of these errors are shown in Table 4.11. Considering the resulting phone sequences, the phones do not provide any information in forming syllables or words. The resulting phone boundaries or time alignment of each phone does not have any relation between each phone. This is also another major disadvantages of the phones.

4.3.2 Contextual Onset-Rhyme Models (CORMs)

Comparing to the phones, the contextual onset-rhyme models (CORMs) do not have any deletion errors on any arresting stops. These errors are substitution errors in the CORMs. For examples, the words /phaan0 naj0/ is composed of the phones /ph aa n n a j/ and the CORMs /phaa aa_n n_a a_j/. The system recognises the words as /ph aa n a j/ in phones where the arresting nasal /n/ was deleted. Using the CORMs, the words were recognised as /phaa aa n_a a_j/ in which the rhyme /aa_n/ was substitued by /aa/.

The CORMs performs recognition much better than the phones in many ways. Firstly, there are very few errors on the onset units, which contain releasing consonant and its transition. In many cases, the phones are unable to point out whether the consonant is releasing or arresting. For example, considering the evaluation results “testsentences01_002a_vis” in Table C1.1 for the phones and in Table C2.1 for the contextual, there are plenty of insertion errors in the phone sequence while the contextual have none. The recognized CORMs sequences illustrate pairs of onset and rhyme units that make up a syllable as shown in Table C2.1 in the Appendix C. These onset-rhyme pairs give out syllable boundary information, which is the most valuable information for tone recognition.

Comparing to the phonotactic onset-rhyme models (PORMs), the contextual onset-rhyme models give out about 22% higher word error rate than the PORMs as shown in Table 4.4. In Table 4.5 and Table 4.6, the CORMs provide higher word error rate in every cases than the PORMs. However, in Table 4.7 and Table 4.8, the error rate of the onset units in the CORMs are lower than in the PORMs on both fixed and variable duration overlap. The error rate of the CORMs onset units are 20.953% lower than the PORMs in fixed-duration overlap and 11.623% lower in the variable-duration overlap. The lower onset error rates are resulted from more compact models of each onset unit. The CORMs have much lower number of onset units than the PORMs.

4.3.3 Phonotactic Onset-Rhyme Models (PORMs)

Comparing to the phones, the phonotactic onset-rhyme models (PORMs) have many advantages over the phones like their counterparts, the CORMs. For examples, comparing the evaluation results of the “testsentences01_006b_vis” in Table C1.1 for the phones and Table C3.1 in the Appendix C for the PROMs, the phones show many insertion and deletion errors. The recognized PORMs sequence combines syllable boundary information within each onset-rhyme pair. On the other hand, sequences of the phones do not provide any acoustic information as illustrated in Fig. D1.6 in the Appendix D for time alignment of the “testsentences01_006b_vis”.

Comparing recognition errors to the CORMs, the PORMs do not have errors on removal of releasing consonants like the CORMs. The PORMs provides better word error rate at 18.095% lower than the CORMs as shown in Table 4.4. In Table 4.5 and Table 4.6, the PORMs give out lower word error rate than the CORMs in every cases. However, error rates of the onset units using the PORMs are higher than the CORMs in every cases of both fixed

and variable duration overlap as shown in Table 4.7 and Table 4.8. These are resulted from a large number of onset units in the PORMs than the CORMs. The higher amount of onset units in the PORMs makes the model network more complex than the CORMs.

4.4 Summary

In this chaper, all the results of both forced alignment and recognition are described and analysed in details. Evaluation of the onset-rhyme models are conducted in the aspects of forced alignment and recognition. The forced alignment evaluates precision of model boundaries. On the other hand, the recognition evaluates accuracy of the models in modelling speech segments. Both of the contextual and phonotactic onset-rhyme models outperform the phone models in both forced alignment and recognition. The contextual and phonotactic onset units illustrate better alignment and recognition of releasing consonant. The onset units of both onset-rhyme models show significant improvement in recognition of every kinds of releasing consonants.

Table 4.9 Various types of insertion error using phone models

Insertion errors on vowel phonemes

LAB: sil **p a w** m aa j kh @@ ng ph uu k @@ k aa n r aa j kh vv c a p ph uu kh a w
REC: sil t xx sil **p aa a w** m aa j kh @@ ng ph uu k kh @@ k aa n r aa j kh vv sil c a t ph uu k kh aa w

LAB: sil **p a w** m aa j kh @@ ng ph uu k @@ k aa n r aa j kh vv c a p ph uu kh a
REC: sil **t aa a w** m a j t e kh @@ ng t o ph uu k kh @@ k @ k aa n r aa j kh vv sil c a p aa ph uu k u kh aa w

Insertion errors on releasing consonants

LAB: sil p a w m aa j kh @@ ng ph uu **k @@** k aa n r aa j kh vv c a p ph uu **kh a w**
REC: sil t xx sil p aa a w m aa j kh @@ ng ph uu **k kh @@** k aa n r aa j kh vv sil c a t ph uu **k kh aa w**

Insertion errors on releasing nasals

LAB: p a j s a j phr i k b ii p **m a n aa w** k i n p e n z aa h aa n kl aa ng w a n sil
REC: p a j s a j sil phr i k j xx sil b ii p **b a n d aa w** a t th i n t e n z aa h aa z vv t kl aa ng w a n sil

Table 4.10 Various types of deletion error using phone models

Deletion error of a releasing stop

LAB: sil r oo ng r iia n t a ng j uu th aa m kl aa ng **m xx k m aa j** r i m m xx n aa m
REC: sil r oo ng r iia n t aa ng j uu th aa m kl aa ng **m xx m aa j** sil r i m m xx d aa m

LAB: sil kh a w d oo n **m ii t k oo n** b aa t th ii kh aa ng k xx m k o n c o n **l vva t** s aa t t @@ n n a ng
REC: sil kh @@ k u d oo n **m ii k o n** b aa t th ii kh aa ng k xx m k o n t o n **l vva** s aa k t @@ n aa ng

LAB: sil j i ng s aa w w aa ng c xx k a n th a t p a j **c aa k** th aa t l x kr a ch a w **d @@ k m aa j**
REC: sil th i ng s aa w w aa n k xx k aa n t xx th aa t p a j **c aa** t aa t l x sil kr a ch a w **b @@ m aa j**

LAB: p a j s a j phr i k **b ii p** m a n aa w k i n p e n z aa h aa n kl aa ng w a n sil
REC: p a j s a j phr i k j xx sil **d ii** b aa n l aa w c i ng t e n z aa h aa n v ng kl aa ng w aa n sil

Deletion error two adjacent nasals—arresting followed by releasing nasals

LAB: sil kh a w d oo n m ii t k oo n b aa t th ii kh aa ng k xx m k o n c o n l vva t s aa t **t @@ n n a ng**
REC: sil kh @@ k u d oo n m ii k o n b aa t th ii kh aa ng k xx m k o n t o n l vva s aa k **t @@ n aa ng**

LAB: l aa s a t p aa l qq j j aa ng t xx k t @ ng p a **k a n n a j** p aa sil
REC: l aa k s a t aa t p aa p aa sil l qq j j aa ng t xx sil t @ ng t aa t **k aa n a j** p aa p aa sil

LAB: khr vva ng b i n j u t th a m ng aa n phr @@ m k a n m vva **n aa m m a n** m o t sil
REC: khr vva ng b i n sil j u t th aa m ng aa n t xx phr @@ m k aa n sil m vva **n aa m a n** m @@ t sil

LAB: j o k **n aa m n a k** t @@ n k qq t phl qq ng m a j **d aa j j aa ng** j xx p j o n sil
REC: j o k @@ k **n aa m a t** sil t @@ n v k k qq t phl qq ng m a j sil **d aa j aa ng** j xx t aa j oo n sil

Table 4.11 Various types of substitution error using phone models

Substitution errors over short and long vowels

LAB: khr vva ng b i n j u t th a m ng aa n **phr @@ m k a n** m vva **n aa m m a n** m o t sil
REC: khr vva ng b i n sil j u t th a m ng aa n t x **phr @ m** vva **k aa n** sil m vva z aa **n aa m aa n** m @@ t k @ sil

Substitution errors between voiced stop and nasal

LAB: p a j s a j phr i k b ii p **m a n aa w** k i n p e n z aa h aa n kl aa ng w a n sil
REC: p a j s a j phr i k j xx sil d ii **b aa n l aa w** c i ng t e n z aa h aa n v ng kl aa ng w aa n sil

LAB: p a j s a j phr i k b ii p **m a n aa w** k i n p e n z aa h aa n kl aa ng w a n sil
REC: p a j s a j sil phr i k j xx sil b ii p **b a n d aa w a t** th i n t e n z aa h aa z vv t kl aa ng w a n sil

Substitution errors within a group of voiceless unaspirated stops

LAB: p a j s a j phr i k b ii p m a n aa w k i n **p e n** z aa h aa n kl aa ng w a n sil
REC: p a j s a j phr i k j xx sil d ii b aa n l aa w c i ng **t e n** z aa h aa n v ng kl aa ng w aa n sil

Substitution errors within a group of nasals

LAB: sil j i ng s aa w **w aa ng** c xx k a n th a t p a j c aa k th aa t l x kr a ch a w d @@ k m aa j
REC: sil th i ng s aa w **w aa n** k xx k aa n t xx th aa t p a j c aa t aa t l x sil kr a ch a w b @@ m aa j

Substitution errors on consonant clusters

LAB: sil kh o n r aa j b u k r u k kh a w **khr @@ p khr @@ ng s aa n** kh ee m ii
REC: sil kh o n r aa j b u k phl u r uua sil kh a w k @ **phr @ b @ r @@ ng s aa m** kh ee m ii

Table 4.12 Substitution errors on the onset units using the contextual and phonotactic onset-rhyme models.

Onset substitution errors on voiceless stops with the same places of articulation

LAB: sil p_a a_w m_aa aa_j kh_@@ @@_ng ph_uu uu **k_@@ @@** k_aa aa_n r_aa aa_j
REC: sil sil sil p_a a_w m_aa aa_j kh_@@ @@_ng sil ph_uu uu sil **kh_@@ @@** z_@@ @@_k k_aa aa_n r_aa aa_j sil

LAB: sil kh_a a_w d_oo oo_n m_ii ii_t **k_oo oo_n** b_aa aa_t th_ii ii kh_aa aa_ng k_xx xx_m
REC: sil kh_a a_w d_oo oo_n m_ii ii sil **kh_o o_n** b_aa aa_t th_ii ii kh_aa aa_ng sil k_xx xx_m sil

Onset substitution errors on voiceless stops with the same manners of articulation

LAB: sil khaa a_w paa a_k **thoo o_ng** laa aa_j phaa aa khaa aa_w maa aa sii ii khaa aa_w waa a_j
REC: sil khaa a_w paa a_k **khoo o_ng** klaa aa_j phaa aa_k khaa a_w maa aa sii ii khaa a_w sil waa a_j

LAB: thii ii **paa aa_k** thaa aa_ng khaa a_w baa aa_n sil
REC: thii ii **taa a_k** thaa aa_ng khaa a_w baa aa_n sil

LAB: sil kh_a a_w s_vv vv **ph_aa aa_n** n_a a_j r_aa aa kh_aa aa s_aa aa_m ph_a a_n b_aa aa_t sil
REC: sil sil kh_a a_w s_vv vv sil **kh_aa aa** n_a a_j r_aa aa kh_aa aa s_aa aa_m sil ph_a a_n b_aa aa_t sil

LAB: **ph_xx xx_n** k_aa aa_n p_@ @_ng k_a a_n h_ee ee_t r_aa aa_j n_a a_j
REC: **th_xx xx** th_v v_ng sil k_aa aa_n sil p_@ @_ng k_a a_n h_ee ee_t sil r_aa aa_j sil n_a a_j sil

Onset substitution errors on consonant clusters

LAB: sil kh_o o_n r_aa aa_j b_u u_k r_u u_k kh_a a_w **khr_@@ @@_p khr_@@ @@_ng**
REC: sil sil kh_o o_n r_aa aa_j j_u u_k r_uu uu sil kh_a a_w sil **phr_@@ @@_p sil r_@@ @@_ng**

LAB: pr_a a m_o o_ng **khw_aa aa_ng** pr_a a ph_ee ee n_ii ii p_a a_n h_aa aa
REC: pr_a a m_o o_ng sil **kw_aa aa** sil sil sil pr_aa aa_t sil ph_e e_n n_ii ii sil p_a a_n h_aa aa sil

LAB: foo o_n k@@ @@ too o_k proo oo_j **praa aa_j** loo o_ng maa aa ph@@ @@ dii ii sil
REC: sil foo o_n sil k@@ @@ too o_k sil proo oo_t sil **raa aa_j** loo o_ng maa aa ph@@ @@_n dii ii sil sil

Table 4.12 Substitution errors on the onset units using the contextual and phonotactic onset-rhyme models.

Onset substitution errors within a group of voiced stops

LAB: z_ii ii_k kh_o o_n p_a a_j s_a a_j phr_i i_k **b_ii ii_p** m_a a n_aa aa_w
REC: d_ii ii sil kh_o o_n sil p_a a_j s_a a_j sil phr_i i_k sil **d_ii ii** b_a a_n d_aa aa kh_a a_w sil

Onset substitution errors between releasing voiced stop and nasal

LAB: sil m_aa aa_j r_i i_m m_xx xx **n_aa aa_m** kl_aa aa_ng m_vva vva_ng ch_iia iia_ng m_a a_j sil
REC: sil m_aa aa_j sil r_i i_m m_xx xx **d_aa aa_m** sil kl_aa aa_ng m_vva vva_ng ch_iia iia_ng m_a a_j sil

LAB: s_aa aa m_aa aa_t h_aa aa **n_uua uua_j** ng_aa aa_n ph_uu uu r_a a_p ph_i i_t ch_@@ @@_p sil
REC: s_aa aa_p m_aa aa sil h_aa aa **d_uua uua_j** ng_aa aa_n sil ph_uu uu r_a a_p sil ph_i i_t ch_@@ @@_p sil

LAB: khii i_ng haa aa khoo o_n kee e_ng hoo o_k **ngqq q_n** cee e_t sil
REC: khii i_ng sil thaa aa sil khoo o_n kee e_ng sil sil hoo o_k sil **naa a_n** sil sil cee e_t sil

LAB: z_ii ii_k kh_o o_n p_a a_j s_a a_j phr_i i_k b_ii ii_p **m_a a n_aa aa_w**
REC: d_ii ii sil kh_o o_n sil p_a a_j s_a a_j sil phr_i i_k sil d_ii ii **b_a a_n d_aa aa** kh_a a_w sil

LAB: paa a_j saa a_j phrii i_k bii ii_p **maa a naa aa_w** kii i_n pee e_n
REC: sil paa a_j saa a_j sil phrii i_k sil bii ii_p **maa a_n zaa aa_w** kii i_n sil pee e_n

Removal of releasing consonant in the onset unit

LAB: khaa a_w praa a **kuu uua_t** joo o_k naa aa_m naa a_k t@@ @@_n kqq qq_t phlqq qq_ng maa a_j
REC: khaa a_w praa a **uu uua_t** joo o_k naa aa_m naa a_k sil t@@ @@_n sil kqq qq_t phlqq qq_ng maa a_j sil

LAB: sil thaa a_ng dee e_k chaa aa_j lxx x dee e_k jii i_ng daa aa_j pee e_n **tuu uua** thxx xx_n kh@@ @@_ng
REC: sil thaa a_ng dee e_k chaa aa_j lxx x dee e_k jii i_ng sil daa aa_j pee e_n **uu uua** thxx xx kh@@ @@_p

LAB: **kluu** uua maa a lxx xx_ng saa aa_p cvv v_ng thaa a_m faa a_j maa a_j svv vva
REC: **uu** uua maa a lxx xx_ng saa aa_p sil cvv v_ng thaa a_m faa a_j maa a_j svv vv daa a_p

CHAPTER 5

Conclusions

This dissertation presents acoustic modelling techniques for the onset units in the onset-rhyme models. The proposed onset-rhyme models and onset overlapping techniques are summarized in this chapter. Conclusions on the research are summarized including the experimental results. Contributions of this dissertation are given in this chapter along with future research directions on acoustic modelling of the onset-rhyme models.

5.1 Conclusions of the Dissertation

In this dissertation, the novel onset-rhyme acoustic models are proposed and applied to Thai language. The two proposed onset-rhyme models are contextual and phonotactic onset-rhyme models. The primary focus of this dissertation is at the onset unit of the onset-rhyme models. Therefore, both the contextual and the phonotactic onset-rhyme models have different characteristic of the onset units but share the same rhyme units. The onset-rhyme models are composed of pairs of an onset unit and a rhyme unit. The models contain overlapped segments of the onset unit over the rhyme unit. This overlapped segment models is one major advantage of this model, which provide better modelling of a releasing consonant for the onset units.

Phonologically, a syllable comprises an onset and a rhyme segment. The rhyme segment comprises nucleus and coda of a syllable. Considering the Thai syllable structure, an onset segment covers a releasing consonant of a syllable while the rhyme segment covers a vowel and an arresting consonant. Acoustically, the vowel is a nucleus of a syllable that covers most of the whole syllable segment. Whereas, the consonants are considered as marginal sounds attached to both left and right sides of the nucleus, in this case, releasing and arresting consonants.

From the acoustic-phonetic analysis on Thai syllables, transitional period between a nucleus and its marginal sounds contains some encoded acoustic and articulatory information. The formant transitions are varied according to consonant and vowel context. However, each of the releasing consonant has specific acoustic characteristic that provide predictable formant transitions across different vowel context. These informations provide crucial acoustic cues in determining releasing consonants. Hence, modelling of an onset unit includes the transitional period between a releasing consonant and its adjacent vowel.

Comparing to the initial-final models, there are many difference between the onset-rhyme models and the initial-final models. Firstly, the initial-final models are context-independent where as the onset-rhyme models are context-dependent by nature. Secondly, the initial-final do not model releasing consonant in every possible syllable context. This issue has made the initial-final models context-independent. Thirdly, the initial-final models do not have internal and external junctures which constitute a pair of initial and final by tying both models together.

Two onset-rhyme models are introduced in this dissertation—the contextual and phonotactic onset-rhyme models. The two models have different modelling of the onset units. The phonotactic onset-rhyme models (PORMs) consider every different releasing consonant and vowel context as a separate onset unit. The onset units of this models provide complete combinations of releasing consonants in every vowel context. There are a total of 992 PORMs grammatically existed, which comprises 792 onset units and 200 rhyme units.

Similarly, the contextual onset-rhyme models (CORMs) combine some onset units with the same short and long vowel pairs as a single onset unit. From acoustic analysis, formant transitions in the transitional period have similar characteristics in both short and long vowel pairs with the same releasing consonant. Therefore, combining the onset units with similar vowel context help reduce the models to 479 CORMs, which composed of 297 onset units and 200 rhyme units. Both of the PORMs and CORMs share the same rhyme units. The rhyme units are complete grammatically existed combinations of vowels and arresting consonants in Thai.

The hidden Markov models (HMMs) are employed in modelling the onset-rhyme acoustic models. The left-right topologies with no skipping state are selected in modelling using the hidden Markov models. The number of Markov states in each model are varied according to their characteristics. The phone models use three Markov states in each model. Based on the phone models, the onset units are set to use five Markov states, which covers the whole phone of a releasing consonant with its transitional period. The rhyme units are set to use eight Markov states, which covers the whole phones of both vowel and arresting consonant. In the experiments, the number of HMMs states of the onset and rhyme units are varied in two configurations; 5-state onset with 8-state rhyme units and 4-state onset with 6-state rhyme units.

The onset units cover an arresting consonant and transitional period of its adjacent vowel. Then, the onset units overlap into the vowel segment of the rhyme units. In modelling of the onset units, two schemes are proposed in determining an amount of overlap, the fixed and variable duration overlap. The fixed duration overlap provides predefined length of overlap at 10ms, 20ms, or 30ms into the rhyme units. The variable duration overlap provides varying overlap length according to duration of adjacent vowel. The overlap length is set at 5%, 10%, 15%, 20%, or 25% of the vowel duration. A series of experiments were conducted on all overlap schemes to see the effects of overlap length.

The best error rate for the onset units is 10.387% using the CORMs at 25% overlap and 5-state onset HMMs. Using the PORMs, the onset unit error rate is at 11.753% at 20% overlap and 5-state onset HMMs. On the entire onset-rhyme models, the best word error rate (WER) is at 13.529% using the PORMs at 20% overlap with 5-state onset and 8-state rhyme HMMs. The CORMs provide 16.518% word error rate at 15% overlap with 5-state onset and 8-state rhyme HMMs. The phone models give out only 37.12% word error rate. The PORMs reduce word error rate up to 55.76% over the phone models. From experimental results, the variable duration overlap offers significantly better error rates over the fixed duration overlap in all cases. Moreover, the longer Markov states also provides better accuracy in every case.

The onset-rhyme models prove themselves to provide better modelling of speech than the conventional phone models. The onset-rhyme models incorporate language modelling into the models through the pairs of onset and rhyme units. Then, the models are context-dependent where phonotactics are embedded into the models in forming a syllable. The models are consistent in which the same models have similar characteristics across different speech instances. The models cover a finite set of speech units, which represent all potential speech units of the language. The models also capture coarticulatory effects over the entire syllable. The effects are handled by overlapping over the transitional period in an onset unit and by covering the whole nucleus and coda in a rhyme unit. These characteristics of the onset-rhyme models provide better acoustic modelling of speech than other models.

Although only partial set of onset-rhyme models were utilized, this is possible in implementation of a small, task-specific Thai continuous speech recognition system. This kind of small system is much easier to optimize to have very high recognition accuracy.

In selection of speech units for recognition, two criteria were used—consistency and trainability (Lee, 1990). In this dissertation, three new criteria are introduced, that is, economy, workability, and practicality. Various speech units are then compared based on these five criteria. Results of comparison are summarized in Table 5.1 in application to the Thai continuous speech recognition. Details about each criterion is shown as follows.

- **Consistency**—the consistency criterion concerns about acoustic resolution of a speech unit in which the same unit is consistent in every speech instance.

- **Trainability**—the trainability criterion considers estimation reliability of each speech unit. Estimation of each speech unit should be reliable with certain amount of data.

- **Economy**—the models should have finite number of speech units which could be easily and reliably estimated.

- **Workability**—the workability criterion considers ability of the speech units to use in different environments. Also, the speech units should be speaker-independent.

- **Practicality**—the practicality criterion concerns about applying the speech units into actual practice.

Currently used speech units are analysed based on the above criteria including the onset-rhyme models. Summary of the speech units is shown in Table 5.1 based on the five criteria. The onset-rhyme models satisfy all the criteria of consistency, trainability, economy, workability, and practicality. The onset-rhyme models are consistent in which the same onset and rhyme units are characteristically similar across different instances. The models have finite set of speech units that could be trained with a small set of sentences. The finite number of units satisfy both trainability and economy. All the onset and rhyme units cover the whole potential speech unit in every context. On the workability criterion, the onset-rhyme models could be used in various environments, i.e., clean and noisy. On the practicality criterion, the onset-rhyme models could be easily applied to any tone languages resulted from the finite amount of speech units.

Comparing to the other speech units, the onset-rhyme models provide some significant advantages over other acoustic models. Firstly, the onset-rhyme models cover the whole potential speech units of the language. Every combination of consonant-vowel context has been modelled. Secondly, the onset-rhyme models are able to handle any new unknown syllables or words. The context-dependent phone models, diphones and triphones, are unable to cope with new unseen triphones of new words. For these reasons, the onset-rhyme models satisfy both criteria of workability and practicality as described.

Table 5.1 Evaluation of various acoustic speech units for Thai continuous speech recognition.

Speech Units	Consistency	Trainability	Economy	Workability	Practicality
Word models	Yes	No	No	Poor	Poor
Phone models	No	Yes	Yes	Good	Fair
Multi-phone models	Yes	Difficult	No	Good	Fair
Transition models	Yes	Difficult	No	Fair	Fair
Word-dependent phone models	Yes	Through Sharing	No	Poor	Fair
Context-dependent phone models	Yes	Through Sharing	No	Fair	Poor
Initial-Final Models	Final Models Only	Yes	Yes	Fair	Fair
Onset-Rhyme Models	**Yes**	**Yes**	**Yes**	**Very Good**	**Very Good**

5.2 Contributions of the Dissertation

This is the summary of contributions made by the research in this dissertation. These contributions are significant parts, which make up this dissertation. All of the contributions are summarized as follows.

A. The Onset-Rhyme Acoustic Models

This dissertation conducted basic researches on acoustic models for Thai continuous speech recognition. The two novel acoustic models are introduced in this dissertation—contextual and phonotactic onset-rhyme models. The two models are utilized in Thai continuous speech recognition system. The concept of onset and rhyme was applied to a Thai speech synthesis system in 1992 by Luksaneeyanawin (1992a). In the two proposed models, the onset units cover transitional period between releasing consonant and adjacent vowel. An amount of coverage over the transional stage could be determined by the two proposed methods—the fixed and variable duration overlap.

B. The Contextual and Phonotactic Onset-Rhyme Models

The most significant contribution of this dissertation is the introduction of the contextual and phonotactic onset-rhyme models. This dissertation focuses only at the onset unit. These two models have different characteristics of the onset units but share the same rhyme models. Both models have illustrated better modelling of speech than the phone models without applying any language modelling or any other techniques. The advantages of the two models are described as follows.

- The phonotactic onset-rhyme models provide complete modelling of the onset units in every possible context. The onset units contain releasing consonant with transitional period toward its adjacent vowel. Each releasing consonant is modelled in every vowel context existed in the language. This kind of modelling has made the models context-dependent. The model has 792 onset units, which are grammatically occurred in Thai language.

- In the contextual onset-rhyme models, a releasing consonant within context of the same short-long vowel pairs share a single onset unit. The sharing of onset units with similar context helps reduce the number of onset units. This model has 279 onset units, which are grammatically occurred in Thai language.

- Every onset unit incorporates transitional period between a releasing consonant and its adjacent vowel nucleus of the same syllable.

- The onset-rhyme models are overlapped segment models in which the onset units overlap into the rhyme units in modelling to include the transitional period into the onset units.

- The onset-rhyme models are context-dependent speech units in which every releasing and arresting consonants are modelled in all possible vowel contexts.

- The onset-rhyme models always occur in pairs of an onset unit and a rhyme unit, which make up syllables, words, and sentence respectively. The phonotactic or phonological rules are embedded in each pair of onset and rhyme units, where language modelling is automatically integrated.

- The onset and rhyme units in both models cover all potential speech units of the language, which are grammatically existed.

C. Fixed and Variable Duration Overlap Schemes for the Onset Units

The onset units overlap into the rhyme units in both of the proposed contextual and phonotactic onset-rhyme models. The overlap duration must be sufficient to cover the whole transitional period between a releasing consonant and its adjacent vowel. Therefore, two techniques are introduced to determine duration of overlap into the rhyme unit—the fixed and variable duration overlap. Details of the fixed and variable duration overlap are summarized as follows.

- The fixed duration overlap provide constant length of overlap over the vowel segment of a rhyme unit at 10 ms, 20 ms, or 30 ms. These figures are based on the minimum length of a vowel segment in a syllable. The shortest vowel segment is about 30 ms long in the speech corpus. Using longer duration beyond 30 ms might cover the whole vowel segment including a coda or even the whole syllable.

- The variable duration overlap provide varied length depending on the length of vowel. The duration is at 5%, 10%, 15%, 20%, or 25% of the vowel length. The concept of this technique is based on results of acoustic analysis on Thai syllables. The length of transitional period is varied according to the length of a vowel nucleus.

D. Acoustic Analysis on the Thai Continuous Speech

Acoustic-phonetic analysis is the study of acoustical properties in relation to phonetic characteristics of sounds. Acoustic-phonetic analysis on Thai continuous speech had been conducted prior to creating acoustic models. A set of Thai continuous speech was extensively analysed and study on their acoustical properties. Many acoustic knowledge was obtained from the analysis, which provide solid background for acoustic modelling. The analyses were focused on the Thai syllables including the syllable nucleus and its marginal sounds.

Results of the analysis provide understanding of Thai continuous speech. Some experiments were also conducted including classification of the Thai monophthongs using acoustic-phonetic features. The result of this analysis and experiment was writtern in a technical article as located in the Appendix B of this dissertation.

E. 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 had been created and analysed to contain sufficient samples of each onset unit and rhyme unit available for training. The text corpus was then used for recording of Thai continuous speech. Procedures in preparing the Thai text corpus is described as follows.

- The text corpus contains text from many sources including some Aesop's fables. Most of the text are created by the author and his colleage.

- The Thai text are transferred into a computer by typing and segmented into sentences. Every sentence of Thai text is transcribed into phonetic transcriptions.

- The transcribed phonetic transcriptions of each sentence are then analysed to compute statistical distribution of onset and rhyme units.

- The whole process is repeated until there are sufficient samples of each unit.

After text analysis, the completed text corpora are used in recording of Thai continuous speech on the sentence-by-sentence basis. The speech were recorded in reading or dictation style. The sets of recorded speech are manually labelled by their phonetic transcriptions. The Thai speech corpora contain only utterances spoken by the author. Details of the Thai speech corpora are shown as follows.

- Labelled Corpus — contains 553 sentences with manually labelled transcriptions.
- Unlabelled Corpus — contains 400 sentences with sentence transcriptions.
- Test Corpus — contains 32 sentences with sentence transcriptions.

F. Speech Analysis Tools

Many tools have been created for the research in Thai continuous speech recognition. The tools include speech analysis tool, speech labelling tool, Thai text parser, speech unit analysis tool, for instance. This dissertation contributes these tools for research in Thai continuous speech recognition in the future.

- Speech analysis tool — for analysing on acoustic properties of speech.
- Speech labelling tool — for labelling of continuous speech by their phonetic transcriptions
- Thai text parser — for conversion of Thai text into phonetic transcriptions
- Speech unit analysis tool — for analysing the amount of onset and rhyme units with their statistical distribution

5.3 Future Research in Acoustic Modelling

Eventhough high recognition accuracy was achieved using the onset-rhyme models, there are many issues that need some improvements.

- This dissertation focuses only at the onset unit. Extensive analysis and modelling of the rhyme unit is needed to complete the whole onset-rhyme models.
- More text and speech corpora are needed to cover the whole onset and rhyme units since this dissertation covers only a partial set of both units. This partial set of units are resulted from using only one speaker for both training and testing. However, each onset and rhyme units has sufficient samples for creating and training a stable model.
- Due to limited resources, only a single male speaker was used in training and testing. There are many issues on the use of only one speaker. Firstly, verification of the proposed acoustic models could be conducted on a small-scale corpus and system. Then, speech data of a single speaker should be sufficient with certain amount of speech units. Secondly, recording and labelling of utterances take a long time to complete and very labor-intensive. Both text and speech corpora took over six months to complete in this dissertation. Therefore, more speakers of both male and female are needed to sufficiently model and test the acoustic models in speaker-independent environment.
- Compares recognition performance of the onset-rhyme models to every available speech unit, i.e., diphones, triphones, demisyllable, initial-final, etc. These experiments should be conducted on a large-scale basis.

REFERENCES

Abercrombie, D. 1967. Elements of General Phonetics. Edinburgh University Press.

Abramson, A. S. 1962. The Vowels and Tones of Standard Thai : Acoustical Measurements and Experiments. International Journal of American Linguistics. 28: 30-38.

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., Luksaneeyanawin, S., and Maneenoi, E. 2000. Direct Classification of Thai Monophthongs on Two-dimensional Acoustic-Phonetic Feature Spaces in Linear, Mel, Bark, and Bark-difference Frequency Scales. Proceedings of Joint Meeting: 140th Meeting of the Acoustical Society of America and NOISE-CON 2000.

Ahkuputra V., Jitapunkul S., Maneenoi E., and Luksaneeyanawin S. Acoustic Modelling of Vowel Articulation on the Nine Thai Spreading Vowels – to be published in the International Journal of Computer Processing of Oriental Languages (IJCPOL)

Ahkuputra V., Jitapunkul S., Jittiwarangkul N., Maneenoi E., and Sawit K. A Comparison of Thai Speech Recognition Systems using Hidden Markov Model, Neural Network, and Fuzzy-Neural Network Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia, December 1998.

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 (APCCAS'98), Chiengmai, Thailand, November 1998.

Ahkuputra V., Jitapunkul S., Wutiwiwatchai C., Maneenoi E., and Kasuriya, S. "Comparison of Thai Speech Recognition Systems using Different Techniques" Proceedings of the 1998 IASTED International Conferences on Signal and Image Processing (SIP'98), Las Vegas, Nevada, U.S.A., October 1998.

Ahkuputra V., Jitapunkul S., Pornsukchandra W., and Luksaneeyanawin S. "A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model". Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'97), Victoria, Canada, August 1997, pages 593-599.

Ahkuputra V., Jitapunkul S., Pornsukchandra W. and Luksaneeyanawin S. "A Speaker Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model". Proceedings of 1997 Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand, December 1997, pages 281-286.

Ahkuputra V., Jitapunkul S., and Luksaneeyanawin S. "A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model". Proceedings of 19th Electrical Engineering Conference (EECON 19), Khonkaen, Thailand, November 1996, pages DS-56 - DS-60.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. 1982. Vowel Identification: Orthographic, Perceptual, and Acoustic Aspects. Journal of Acoustical Society of America. 71: 975-989.

Bernstein, J. 1981. Formant-based representation of auditory similarity among vowel-like sounds. Journal of Acoustical Society of America. 69: 1132-1144.

Deller, J. R., Proakis, J. G., Hansen, J. H. L. 1993. Discrete-Time Processing of Speech Signals. Maxwell Macmillan International.

Demeechai, T., Makelainen, K., 2001. Recognition of syllables in a tone language. Speech Communication 33: 241-254.

Di Benedetto, M.-G. 1989a. Vowel representation: Some observations on temporal and spectral properties of the first formant frequency. Journal of Acoustical Society of America. 86: 55-66.

Di Benedetto, M.-G. 1989b. Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. Journal of Acoustical Society of America. 86: 67-77.

Digalakis, V. V., Ostendorf, M., Rohlicek, J. R., 1992. Fast algorithm for phone classification and recognition using segment-based models. IEEE Transaction on Signal Processing 40 (12): 2885-2896.

Fant, G. 1960. Acoustic Theory of Speech Production. Mouton & Co., The Hague.

Fant, G. 1968. Analysis and Synthesis of Speech Processes. in Malmberg, B. (eds.), Manual of Phonetics. 2nd ed. North-Holland Publishing Co.

Flanagan, J. L. 1972. Speech Analysis, Synthesis and Perception. 2nd ed., Springer-Verlag.

Fujimura, O., and Lovins, J., 1978. Syllables as concatenative phonetic elements. In: Bell, A. and Hooper, J.B. (Eds), Syllables and segments. North-Holland.

Fujimura, O., Macchi, M. J., and Lovins, J. B., 1977. Demisyllables and affixes for speech synthesis. Proceedings of the 9th ICA 1: 513.

Furui, S. 2001. Digital Speech Processing, Synthesis, and Recognition. 2nd Eds. Marcel Dekker Inc.

Ganapathiraja, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G. R., 2001. Syllable-based large vocabulary continuous speech recognition. IEEE Transaction on Speech and Audio Processing 9 (4), 358-366.

Gao, S., Lee, T., Wong, Y. W., Xu, B., Ching, P. C., and, Huang, T., 2000. Acoustic modeling for Chinese speech recognition: a comparative study of Mandarin and Cantonese. Proceedings of 2000 International Conference on Acoustic, Speech, and Signal Processing (ICASSP'2000): 1261-1264.

Gao, Y., Hon, H.-W., Lin, Z., Loudon, G., Yogananthan, S., and, Yuan, B., 1995. Tangerine: a large vocabulary Mandarin dictation system. Proceedings of 1995 International Conference on Acoustic, Speech, and Signal Processing (ICASSP'1995). 77-80.

Hamaker, J., Ganapathiraju, A., and Picone J., 1997. Syllable-based speech recognition. ISIP Technical Report. Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University.

Hanpanich, S. 1993. The Statistical Distribution of Consonants, Vowels and Tones within Thai Syllables Existing as Word. Master's thesis, Department of Linguistics, Chulalongkorn University.

Huang, X. D., 1992. Phoneme classification using semicontinuous hidden markov models. IEEE Transaction on Signal Processing. 40 (5): 1062-1067.

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Rosenfeld, R., 1992. The SPHINX-II speech recognition system: an overview. CMU Technical Report Number CMU-CS-92-112. School of Computer Science, Carnegie Mellon University.

Huang, X., Acero, A., and Hon, H.-W., 2001. Spoken language processing, a guide to theory, algorithm, and system development Prentice Hall PTR.

Hwang, M.-Y., 1993. Subphonetic acoustic modeling for speaker-independent continuous speech recognition Ph.D. Dissertation, Department of Electrical Engineering, Carnegie Mellon University.

Jennings, D.T., Westaway, L., and Curtis, K.M. 1997. Automatic Demi-Syllable Extraction for Speech Synthesis Utilizing Artificial Neural Networks. Proceedings of the 1997 13th International Conference on Digital Signal Processing. 2: 579-581.

Jitapunkul S., Luksaneeyanawin S., Ahkuputra V., Maneenoi E., Kasuriya S., Amornkul P. "Recent Advance of Thai Speech Recognition in Thailand" Proceedings of the 1998 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'98), Chiengmai, Thailand, November 1998.

Jitapunkul S., Ahkuputra V., Wutiwiwatchai W., Maneenoi E., Kasuriya S., Amornkul P., and Luksaneeyanawin S. “Thai Automatic Speech Recognition Model”. Interdisciplinary Approaches to Language Processing, Burnham D., Luksaneeyanawin S., Davis C., and Lafourcade M. eds., 2000, ISBN: 1-86341-859-8, pages 195-213.

Juang, B.-H. and Furui, S. 2000. Automatic recognition and understanding of spoken language—a first step toward natural human-machine communication. Proceedings of the IEEE 88 (8): 1142-1165.

Ladefoged, P. 1967. Three Areas of Experimental Phonetics. Oxford University Press.

Ladefoged, P. 1971. Preliminaries to Linguistic Phonetics. The University of Chicago Press.

Lee, C.H., Rabiner, L. R., Pieraccini, R., and Wilpon, J. G. 1990. Acoustic Modeling for Large Vocabulary Speech Recognition. Computer, Speech, and Language 4: 127-165.

Lee, K. F. and Hon, H. W., 1988. Large-vocabulary speaker-independent continuous speech recognition using HMM. Proceedings of 1988 International Conference on Acoustic, Speech, and Signal Processing (ICASSP’1988) 123-126.

Lee, K. F., 1989. Automatic speech recognition : the development of the sphinx recognition system Kluwer Academic Publishers.

Lee, K. F., 1990. Context-dependent phonetic hidden markov models for continuous speech recognition. IEEE Transaction on Acoustics, Speech, and Signal Processing 38 (4): 599-609.

Lee, K.-F., Hon, H.-W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Transaction on Acoustics, Speech, and Signal Processing 37 (11): 1641-1648.

Lee, K.-F., Hon, H.-W., Reddy, R. 1990. An overview of the SPHINX speech recognition system. IEEE Transaction on Acoustics, Speech, and Signal Processing 38 (1): 35-45.

Leelasiriwong, W. 1991. A Study of Acoustic Characteristics of the Vowels /i,a,u/ in Thai and its Use in Speaker Identification. Master’s thesis, Department of Linguistics, Chulalongkorn University.

Liu, F.-H., Picheny, M., Srinivasa, P., Monkowski, M., and Chen, J., 1996. Speech recognition on Mandarin call home: a large-vocabulary, conversational, and telephone speech corpus. Proceedings of 1996 International Conference on Acoustic, Speech, and Signal Processing (ICASSP’1996). 157-160.

Luksaneeyanawin, S. 1989. A Thai Text-to-Speech System. Proceedings of the Regional Workshop on Computer Processing of Asian Languages (CPAL). Asian Institute of Technology. 305-315.

Luksaneeyanawin, S. 1992a. A Thai Text-to-Speech System. Proceedings of the 4th Conference on Research and Development in National Electronic and Computer Technology. National Electronic and Computer Technology Center (NECTEC), Ministry of Science, Technology and Environment: 65-75.

Luksaneeyanawin, S. 1992b. Three Dimensional Phonology : A Historical Implication. In Pan-Asiatic Linguistics : Proceedings of the Third International Symposium on Language and Linguistics. Chulalongkorn University.

Luksaneeyanawin, S. 1993. Linguistics research and thai speech technology. In: Proceedings of the 5th International Conference on Thai Studies. School of Oriental and African Studies, University of London.

Maneenoi, E., Jitapunkul, S., Luksaneeyanawin, S., Ahkuputra, V., Thathong, U., and Thampanichawong, B. 2000. Thai Monophthongs Recognition Using Continuous Density Hidden Markov Model and LPC Cepstral Coefficients. Proceedings of the 2000 International Conference on Spoken Language Processing (ICSLP'2000).

Maneenoi, E., Jitapunkul, S., Luksaneeyanawin, S., and Ahkuputra, V. 2000. Thai Monophthongs Classification Using CDHMM. Proceedings of Joint Meeting: 140th meeting of the Acoustical Society of America and NOISE-CON 2000.

Nearey, T. M. & Assmann P. F. 1986. Modeling the role of inherent spectral change in vowel identification. Journal of Acoustical Society of America. 80: 1297-1308.

O'Shaughnessy, D. 1987. Speech Communication: Human and Machine. Addison-Wesley Publishing Company.

Ostendorf, M. and Roukos, S. 1989. A stochastic segment model for phoneme-based continuous speech recognition. IEEE Transaction on Signal Processing 37 (12): 1857-1869.

Pensiri, R. 1995. Speaker-Independent Thai Numeral Voice Recognition by using Dynamic Time Warping. Master's thesis, Department of Electrical Engineering, Chulalongkorn University.

Pensiri, R. and Jitapunkul, S. 1995. Speaker-Independent Thai Numerical Voice Recognition by using Dynamic Time Warping. Proceedings of the 18th Electrical Engineering Conference. Chonburi, Thailand. 977-981.

Phatrapornnant, T. 1995. 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.

Phatrapornnant, T. and Jitapunkul, S. 1995. Speaker-Independent Isolated Thai Spoken Vowel Recognition by using Spectrum Distance Measurement and Dynamic Time Warping. Proceedings of the 18th Electrical Engineering Conference. Chonburi, Thailand. 988-993.

Philipp Schmid. 1996. Explicit Formant Features for Segment--Based Speech Recognition. Ph.D. Dissertation, Center for Spoken Language Processing, Department of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology.

Picone, J. 1993. Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE. 81.

Picone, J. 1996. Fundamentals of Speech Recognition: A Short Course. Institute for Signal and Information Processing, Mississippi State University.

Pornsukchandra, W. and Jitapunkul, S. 1996. Speaker-Independent Thai Numeral Speech Recognition using LPC and the Back Propagation Neural Network. Proceedings of the 19th Electrical Engineering Conference. Khonkaen, Thailand, 977-981.

Potisuk, S. and Harper, M. P. 1995. Speaker-Independent Automatic Classification of Thai Tones in Connected Speech by Analysis-Synthesis Method. Proceedings of the 1995 IEEE International Conferences on Acoustics, Speech, and Signal Processing (ICASSP'95).

Prathumthan T. 1986. Thai Speech Recognition using Syllable Units. Master's thesis, Department of Computer Engineering, Chulalongkorn University.

Rabiner, L. R. and Juang, B. H. 1993. Fundamentals of Speech Recognition. Prentice Hall,.

Rabiner, L. R. and Schafer, R. W. 1978. Digital Processing of Speech Signals. Prentice-Hall.

Reddy, D. R. 1996. An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave. Technical Report No. C549, Computer Science Department, Stanford University.

Saravari, C. and Satoshi, I., 1983. A demisyllable approach to speech synthesis of Thai a tone language. Journal of Acoustical Society of Japan 4 (2): 97-106.

Saravari, C. and Satoshi, I., 1984. An automatic procedure for extracting demisyllables from isolated monosyllabic source words for use in speech synthesis-by-rule of Thai. Journal of the Acoustical Society of Japan 5 (2): 71-83.

Schaikoff, R. J. 1992. Pattern Recognition: Statistical, Structural, and Neural Approaches. John Wiley & Sons.

Sriraksa, U. 1995. Acoustic Characteristics Signaling Syllable Boundary In Thai Connected Speech. Master's thesis, Department of Linguistics, Chulalongkorn University.

Stevens, K. N. 1980. Acoustic correlates of some phonetic categories. Journal of Acoustical Society of America. 68 (3), 836-842.

Stevens, K. N. 1998. Acoustic Phonetic. MIT Press.

Stevens, K. N. and Blumstein, S. E. 1978. Invariant cues for place of articulation in stop consonants. Journal of Acoustical Society of America. 64 (5): 1358-1368.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. 1992. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. Journal of Acoustical Society of America. 91 (5): 2979-3000.

Sussman, H. M. 1990. Acoustic correlates of the front/back vowel distinction: A comparison of transition onset versus "steady state". Journal of Acoustical Society of America. 88: 87-96.

Syrdal, A. K. and Gopal, H. S. 1986. A perceptual model of vowel recognitoin based on the auditory representation of American English vowels. Journal of Acoustical Society of America. 79: 1086-1100.

Tarnsakun, W. 1988. An Acoustic Analysis of Stop Consonants in Thai. Master's thesis, Department of Linguistics, Chulalongkorn University.

Thamphothong, P. 1990. Multispeaker Speech Recognition System. Master's thesis, Department of Computer Engineering, Chulalongkorn University.

Thubthong, N. 1995. A Thai Speech Recognition System based on Phonemic Distinctive Features. Master's thesis, Department of Computer Engineering, Chulalongkorn University.

Torkkola, K. 1991. Short-Time Feature Vector Based Phonemic Speech Recognition with the Aid of Local Context. Ph.D. Dissertation, Helsinki University of Technology.

Trongdee, T. 1987. An Acoustic Analysis Of Non-Stop Consonants In Thai. Master's thesis, Department of Linguistics, Chulalongkorn University.

Trubetzkoy, N. A. 1969. Principles of Phonology. Berkeley and Los Angeles University of California Press, 1969.

Wang, H.M., Ho, T.H., Yang, R.C., Shen, J.L., Bai, B.R., Hong, J.C., Chen, W.P., Yu, T.L., Lee, L.S. 1997. Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data. IEEE Transaction on Speech and Audio Processing. 5 (2): 195-200.

Wutiwiwatchai, C. 1997. Speaker Independent Thai Numeral Speech Recognition Using Neural Network and Fuzzy Technique. Master's thesis, Department of Electrical Engineering, Chulalongkorn University.

Yoshida, K., Watanabe, T., and Koga, S. 1989. Large Vocabulary Word Recognition based on Demi-Syllable Hidden Markov Model using Small Amount of Training Data. Proceedings of the 1989 IEEE International Conference on Acoustic, Speech, and Signal Processing. 1: 1-4.

Young, S. 1996. Large Vocabulary Continuous Speech Recognition: a Review. Speech Vision and Robotics Group Publications, Department of Engineering, Cambridge University.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.

Zahorian, S. A. and Jagharghi, A. J. 1991. Speaker-normalization of static and dynamic vowel spectral features. Journal of Acoustical Society of America. 90 (1):67-75.

Zahorian, S. A. and Jagharghi, A. J. 1993. Spectral-shape features versus formants as acoustic correlates vowels. Journal of Acoustical Society of America. 94 (4): 1966-1982.

Zue, V. 1994. Towards Systems that Understand Spoken Language. IEEE Expert Magazine. 9: 51-59.

Zue, V. et al. 1990. The VOYAGER Speech Understanding Systerm : Preliminary Development and Evaluation. Proceedings of the 1990 IEEE International Conference on Acoustic, Speech, and Signal Processing. 1: 51-59.

APPENDICES

APPENDIX A

The Thai Text Corpus

The list of test sentences in Thai is collected in this chapter. List of test sentences is shown in Table A1.1. Other collection of the whole text corpus is too large to keep in this dissertation. Please contact the author if any one would like to see the whole corpus.

Table A1.1 List of Thai test sentences

1.	เขาปักธงลายผ้าขาวม้าสีขาว ไว้ที่ปากทางเข้าบ้าน
2.	กว่าที่ชายหนุ่มจะเอ่ยปากพูดจากับหญิงสาว ฝนก็ตกโปรยปรายลงมาพอดี
3.	คนร้ายบุกรุกเข้าครอบครองสารเคมี ที่เป็นส่วนประกอบของสิ่งเสพติด ในจังหวัดเชียงราย
4.	เครื่องยนต์ทั้งสองเครื่องของเครื่องบิน หยุดทำงานพร้อมกันเมื่อน้ำมันหมด
5.	เป้าหมายของผู้ก่อการร้าย คือจับผู้เข้าประกวดยกน้ำหนัก ตอนเกิดเพลิงไหม้ได้อย่างแยบยล
6.	โรงเรียนตั้งอยู่ท่ามกลางแมกไม้ ริมแม่น้ำกลางเมืองเชียงใหม่
7.	ภายหลังการสอบสวน ก็ยังไม่สามารถหาหน่วยงานผู้รับผิดชอบ ต่อเหตุการณ์ที่เกิดขึ้นได้
8.	ร้านค้าขายสินค้าในราคาย่อมเยา ไม่เอาเปรียบผู้ซื้อ และไม่เก็งกำไรเกินสมควร
9.	เขาโดนมีดโกนบาดที่ข้างแก้มก้นจนเลือดสาด ตอนนั่งฟังเพลงอยู่ข้างกองฟาง
10.	หญิงสาววางแจกันถัดไปจากถาดและกระเช้าดอกไม้ข้างกระโถน
11.	ชายผู้นั้นเมื่อยล้า จึงละเลยป้ายห้ามล่าสัตว์ป่า เลยยางแตกต้องปะกันในป่า
12.	เจ้าหน้าที่ตำรวจ ร่วมกันจัดแผนการณ์ป้องกันเหตุร้ายในช่วงวันหยุด
13.	หญิงสาวเลาะชายกระโปรงที่ขาดออกทิ้งไป เพราะไปเกี่ยวโดนล้อรถจนขาด
14.	ต้นเงาะ ต้นถั่ว ต้นโพธิ์ และต้นหลิว ต่างก็เป็นต้นไม้ทั้งสิ้น
15.	ทั้งเด็กชายและเด็กหญิง ได้เป็นตัวแทนของประเทศ เดินทางไปแข่งขันในต่างประเทศ
16.	ยามเดินมาซื้อยำกุ้งแห้ง ให้ยามอีกคนไปใส่พริกบีบมะนาว กินเป็นอาหารกลางวัน
17.	หนิงเป็นคนเก่งละเอียดรอบคอบ แต่กลัวแมลงสาบ จึงทำไฟไหม้เสื้อผ้าฝ้ายหมดทั้งตัว
18.	เขาเดาว่า เสือดาวเข้ามาทำร้ายคนในหมู่บ้านเมื่อคืนนี้
19.	ในตอนกลางวัน กระแสลมพัดข้าวของกระจุยกระจายไปทั่วทั้งหมู่บ้าน
20.	จังหวัดเชียงใหม่และจังหวัดเชียงราย ร่วมกันจัดงานดอกไม้เมืองหนาวในช่วงปลายปี
21.	ชัยเป็นยามเจ้าหน้าที่ ใช้ถาดใส่ยำกั้นไม่ให้มีการบุกรุกป่าชายแดน
22.	นายพรานเลี้ยงนกอินทรีย์ นกยูง และนกเอี้ยงไว้ในป่า
23.	กระดาน กระชอน เครื่องหมาย ถั่วเหลือง ดอกไม้ ทบทวน ดินเหนียว
24.	ทางหลวง น้ำตาล หมูหยอง ประมง ขวาง ประเพณี ปัญหา โรงเรียน ลำดับ
25.	ลูกสาว หน้าต่าง ลูกชาย ไหวพริบ ลูกปัด หมู่บ้าน ลูกเหล็ก ขวด กำนัน
26.	หนึ่ง กล่าวหา สอง กล้าหาญ สาม ข้าวโพด สี่ ขิง ห้า คนเก่ง หก เงิน เจ็ด
27.	ฝากฝัง ฝักข้าวโพด ทาสชาย ข้างแรม ภูเขาสีขาว เดาดวงดาว น้ำหอมนำเข้าจากต่างประเทศ
28.	ฝ่ายชายเป็นฝ่ายได้ชัย ฝ่ายหญิงมีไฝใส่ผ้าฝ้ายติดไฟง่าย นั่งฟังอยู่ข้างกองฟาง
29.	แม่บ้าน ระเบียบ แม่น้ำ ระเบียง ภาคอีสาน ลวดลาย ภาษาไทย สายลม
30.	เคลื่อนไหว ช่องว่าง ชาวบ้าน ต้นเงาะ บิดงอ เช่าบ้าน ตอนเช้า
31.	เขาซื้อพานในราคาสามพันบาท
32.	เล็กเขียนตัวเลขได้อย่างสวยงาม

Table A1.2 Statistics of the Thai onset units on their frequency.

Units	Amount
n_a	333
m_a	247
th_ii	235
n_aa	229
m_aa	225
kh_a	213
th_a	208
p_a	199
k_a	181
c_a	172
j_aa	165
s_a	143
l_x	125
kh_o	123
l_a	121
h_a	118
kh_aa	118
kh_@@	114
w_a	114
w_aa	113
m_ii	110
d_aa	109
ch_aa	100
k_aa	99
j_uu	96
p_e	92
r_aa	92
h_aa	91
r_a	90
k_@@	89
c_o	88
kr_a	87
c_aa	86
ch_a	86
s_aa	85
t_aa	83
l_aa	80
t_xx	76
l_o	75
c_v	74
n_ii	73
th_aa	73
t_uua	70
ph_aa	69
l_xx	64
z_aa	63
khw_aa	61
t_a	61
pr_a	58
b_aa	55
z_a	55
s_@@	54
t_@@	54
ph_a	53
ph_uu	53
ph_vva	53
kh_v	52
r_@@	51
l_uu	49
t_o	48
th_uu	48
j_a	47
m_vva	47
b_o	46
d_a	46
z_@@	46
r_vva	45
b_a	44
n_v	43
p_aa	43
t_@	43
n_@@	42
ng_aa	42
s_o	42
d_uu	41
d_uua	40
th_uua	40
d_qq	39
k_i	39
s_iia	37
l_e	36
l_qq	35
d_e	34
d_ii	34
phr_@	34
n_o	33
th_u	33
w_i	33
th_v	32
n_i	31
f_a	30
d_oo	29
kl_aa	29
m_@@	29
m_o	29
ph_e	29
ph_i	29
s_ii	29
s_uua	29
ch_iia	28
r_o	28
s_vv	28
kh_u	27
pl_@@	27
r_uua	27
th_iia	27
z_iia	27
b_iia	26
ph_@@	26
pl_@	26
k_xx	25
kh_ii	25
kw_aa	25
ph_xx	25
r_iia	25
s_u	25
z_ii	25
b_@@	24
ch_vva	24
d_i	24
l_@@	24
l_vva	24
m_xx	24
s_i	24
j_i	23
j_u	23
ph_o	23
r_uu	23
s_vva	23
th_@@	23
th_i	23
h_e	22
c_e	21
j_@	21
khr_a	21
m_vv	21
ph_qq	21
z_uua	21
c_i	20
f_aa	20
j_oo	20
l_u	20
r_xx	20
s_uu	20
b_i	19
b_u	19
c_@@	19
j_o	19
kl_a	19
l_ee	19
pr_iia	19
tr_ii	19
tr_o	19
h_xx	18
j_xx	18
khl_a	18
ph_iia	18
z_ee	18
h_@	17

k_e	17
kh_x	17
n_u	17
n_vva	17
n_xx	17
phr_@@	17
r_ii	17
r_u	17
th_x	17
f_o	16
h_o	16
l_oo	16
pr_aa	16
r_e	16
t_u	16
th_ee	16
th_xx	16
ch_o	15
j_e	15
j_vv	15
k_o	15
k_oo	15
k_qq	15
k_uua	15
khl_@@	15
kl_uua	15
s_xx	15
z_vv	15
ch_@@	14
ch_uua	14
h_uua	14
kh_iia	14
kh_uua	14
kh_vv	14
khr_aa	14
l_i	14
l_uua	14
ph_uua	14
phl_qq	14
r_i	14
s_ee	14
h_@@	13
j_vva	13
k_@	13
kh_qq	13
khl_u	13
khr_uua	13
m_e	13
ph_x	13
phr_i	13
pl_aa	13
pl_o	13
pl_xx	13
r_qq	13
t_i	13
th_qq	13
w_ee	13
b_xx	12

c_ii	12
ch_vv	12
d_xx	12
j_@@	12
k_u	12
khl_vva	12
ph_vv	12
phl_aa	12
s_e	12
t_ii	12
th_vv	12
b_uu	11
c_iia	11
c_uua	11
ch_qq	11
d_@@	11
d_iia	11
d_o	11
k_vva	11
kh_i	11
kh_uu	11
khw_a	11
kr_uua	11
l_vv	11
n_iia	11
ng_oo	11
ng_q	11
ng_u	11
p_uua	11
ph_ee	11
t_oo	11
z_@	11
z_i	11
z_oo	11
z_u	11
b_oo	10
c_qq	10
ch_@	10
ch_ii	10
d_vva	10
h_i	10
h_x	10
k_ee	10
kh_e	10
khl_aa	10
khl_oo	10
khl_vv	10
m_uu	10
ng_a	10
p_@	10
p_o	10
ph_oo	10
ph_u	10
ph_v	10
phl_u	10
pr_@@	10
t_uu	10
t_x	10

z_qq	10
c_oo	9
ch_e	9
h_oo	9
k_iia	9
khr_vva	9
m_u	9
m_uua	9
n_@	9
p_ii	9
phl_xx	9
phr_aa	9
pl_iia	9
pr_@	9
r_@	9
s_oo	9
s_v	9
t_e	9
tr_uua	9
b.@	8
b_vva	8
b_x	8
ch_i	8
d_vv	8
f_v	8
f_xx	8
h_ee	8
h_vva	8
k_x	8
kh_vva	8
khl_ee	8
khr_uu	8
kl_i	8
kr_o	8
l_iia	8
l_q	8
l_v	8
n_uua	8
p_i	8
ph_ii	8
phl_ee	8
phl_oo	8
r_oo	8
t_vv	8
th_o	8
th_q	8
tr_iia	8
w_@	8
w_o	8
w_xx	8
z_o	8
c_@	7
c_u	7
d_u	7
f_uu	7
h_ii	7
h_uu	7
k_ii	7

kh_ee	7
kh_xx	7
khr_@@	7
kl_@@	7
kl_u	7
kr_x	7
l_@	7
m_iia	7
m_oo	7
n_uu	7
n_x	7
ng_@	7
ng_qq	7
p_iia	7
p_vva	7
phl_a	7
phl_i	7
phr_a	7
r_vv	7
s_@	7
t_iia	7
th_oo	7
tr_a	7
tr_aa	7
tr_uu	7
b_ii	6
c_q	6
c_uu	6
ch_oo	6
j_ee	6
khl_xx	6
kr_@	6
m_i	6
m_qq	6
n_e	6
n_qq	6
ng_@@	6
ng_ee	6
ng_o	6
p_qq	6
p_uu	6
ph_@	6
pl_i	6
pr_oo	6
t_v	6
w_e	6
khl_ii	5
phr_oo	5
pl_ee	5
r_v	5
tr_xx	5

Table A1.3 Statistics of the Thai rhyme units on their frequency.

Units	Amount
a_j	911
aa	668
ii	543
a	502
a_n	376
aa_j	332
o_n	317
aa_ng	315
aa_n	285
uu	252
aa_m	244
a_w	236
a_ng	233
a_m	225
@@.ng	196
e_n	186
v_ng	151
aa_k	140
vva	139
@@	136
uua	135
a_p	131
aa_w	129
x	123
xx	120
@@.n	116
a_k	115
o_ng	107
i_n	100
i_ng	100
uu_k	99
aa_t	94
@@.k	92
o_t	91
@.ng	87
iia_ng	87
uua_j	85
o_k	84
a_t	77
@	75
vv	73
vv_n	71
xx_ng	71
i_t	67
qq_j	67
u_k	64
x_ng	64
vva_ng	61
iia_n	59
o_m	58
uua_n	58
xx_w	57
uua_t	56
u_n	55
@@.p	54
iia_w	54
v_n	52
@_j	51
@@.m	50
xx_k	50
qq_n	47
e_k	46
@@.j	45
@@_t	45
o_p	45
oo_n	44
u	44
ee_t	43
xx_n	42
e_t	39
oo_ng	38
i	37
u_t	36
ii_p	35
u_m	35
iia_p	34
oo	34
u_ng	34
qq	33
oo_t	32
vva_n	32
e_p	31
aa_p	29
ee	29
i_m	29
x_n	29
@.n	28
ee_ng	28
@.m	27
i_w	27
v_k	27
uu_ng	24
xx_p	24
iia	23
oo_j	23
q	23
uu_t	23
vva_m	23
iia_m	22
oo_k	22
uua_ng	22
xx_m	22
qq_t	21
vva_p	21
iia_t	20
i_p	19
ii_k	19
ii_n	19
qq_m	19
qq_ng	18
vv_t	18
uua_m	17
vva_k	17
uua_k	16
e_m	15
e_ng	15
i_k	15
u_j	15
vva_t	15
ii_m	14
qq_k	14
ii_t	13
ee_k	12
oo_m	12
oo_p	12
uu_n	12
v_p	12
xx_t	12
u_p	11
vv_m	11
vva_j	11
e_w	10
ee_m	10
ee_w	10
iia_k	10
q_n	10
uu_p	10
uua_p	10
ee_n	9
ee_p	8
x_w	6
e	5
o	5
v	5

Table A1.4 Statistics of the Thai tones in the speech corpus.

/0/	1051	33.6428%
/1/	684	21.8950%
/2/	697	22.3111%
/3/	419	13.4123%
/4/	273	8.7388%

Table A1.5 Statistics of the Thai monophthongs in the speech corpus.

/a/	954	32.3609%
/aa/	628	21.3026%
/@@/	218	7.3948%
/ii/	186	6.3094%
/i/	156	5.2917%
/v/	127	4.3080%
/o/	125	4.2402%
/xx/	96	3.2564%
/uu/	92	3.1208%
/u/	63	2.1370%
/e/	63	2.1370%
/oo/	53	1.7978%
/qq/	46	1.5604%
/@/	41	1.3908%
/x/	30	1.0176%
/vv/	29	0.9837%
/ee/	28	0.9498%
/q/	13	0.4410%

Table A1.6 Statistics of the Thai diphthongs in the speech corpus.

/uua/	80	45.4545%
/vva/	73	41.4773%
/iia/	23	13.0682%
/ia/	0	0.0000%
/va/	0	0.0000%
/ua/	0	0.0000%

Table A1.7 Statistics of the Thai releasing consonants in the speech corpus.

/m-/	311	10.4503%
/n-/	310	10.4167%
/c-/	218	7.3253%
/k-/	216	7.2581%
/th-/	196	6.5860%
/t-/	179	6.0148%
/kh-/	151	5.0739%
/l-/	148	4.9731%
/s-/	141	4.7379%
/r-/	135	4.5363%
/d-/	128	4.3011%
/p-/	127	4.2675%
/h-/	125	4.2003%
/w-/	120	4.0323%
/j-/	120	4.0323%
/ph-/	102	3.4274%
/z-/	92	3.0914%
/ch-/	64	2.1505%
/b-/	46	1.5457%
/ng-/	32	1.0753%
/f-/	15	0.5040%

Table A1.8 Statistics of the Thai arresting consonants in the speech corpus.

/-n/	561	25.5000%
/-j/	446	20.2727%
/-ng/	427	19.4091%
/-k/	191	8.6818%
/-m/	181	8.2273%
/-w/	161	7.3182%
/-t/	128	5.8182%
/-p/	105	4.7727%

Table A1.9 Statistics of the Thai releasing consonant clusters in the speech corpus.

/kr-/	36	24.3243%
/khw-/	21	14.1892%
/kl-/	17	11.4865%
/tr-/	15	10.1351%
/khr-/	13	8.7838%
/phr-/	11	7.4324%
/kw-/	10	6.7568%
/phl-/	8	5.4054%
/pl-/	8	5.4054%
/pr-/	7	4.7297%
/khl-/	2	1.3514%
/thr-/	0	0.0000%
/fr-/	0	0.0000%
/br-/	0	0.0000%
/dr-/	0	0.0000%

A2 Unit Statistics on Test Sentences

The 32 test sentences are composed of 180 phonotactic onset units and 99 rhyme units. Statistics of the phones in the test sentences are shown as follows.

Table A2.1 Number of releasing consonants in the test sentences

Phones	Qty	Phones	Qty	Phones	Qty
/kh-/	46	/p-/	27	/c-/	19
/m-/	43	/t-/	26	/f-/	16
/th-/	37	/j-/	26	/b-/	15
/n-/	37	/r-/	25	/h-/	14
/k-/	35	/ch-/	24	/w-/	13
/s-/	35	/d-/	23	/z-/	9
/l-/	33	/ph-/	21	/ng-/	8

Table A2.2 Number of arresting consonants in the test sentences

Phones	Qty	Phones	Qty
/-p/	18	/-ng/	99
/-t/	44	/-n/	93
/-k/	34	/-w/	38
/-m/	34	/-j/	88

Table A2.3 Number of consonant clusters in the test sentences

Phones	Qty	Phones	Qty
/pr/	11	/khw/	1
/kr/	8	/kw/	1
/kl/	7	/thr/	0
/khr/	6	/tr/	0
/phr/	5	/fr/	0
/phl/	2	/br/	0
/pl/	1	/dr/	0
/khl/	1		

Table A2.4 Number of monophthongs in the test sentences

Vowels	Qty	Vowels	Qty	Vowels	Qty
/aa/	164	/uu/	19	/qq/	9
/a/	136	/xx/	16	/@/	9
/@@/	35	/e/	14	/x/	7
/o/	33	/oo/	11	/vv/	4
/ii/	23	/u/	9	/v/	4
/i/	23	/ee/	9	/q/	1

Table A2.5 Number of diphthongs in the test sentences

Diphtongs	Qty	Diphtongs	Qty
/uua/	21	/ia/	0
/iia/	14	/va/	0
/vva/	14	/ua/	0

Table A2.6 Number of tones in the test sentences

Tones	Qty
/0/	201
/2/	124
/1/	121
/3/	65
/4/	64

VITAE

Visarut Ahkuputra was born in Bangkok, Thailand in 1972. He received B.Eng. and M.Eng. in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1993 and 1996 respectively. From 1994 to 1996, he was awarded a fellowship by the Honours Program of Graduated Electrical Engineering Students Scholarship for his research on Thai speech recognition system. From 1997 to 2002, he was a Ph.D. student and Research Assistant at Digital Signal Processing Research Laboratory (DSPRL), Department of Electrical Engineering, Chulalongkorn University. The Ph.D. fellowship awarded to him by the Telecommunication Consortium Scholarship of the National Science and Technology Development Agency (NSTDA). This research is partially supported by the DSPRL, Department of Electrical Engineering and the Graduate School of Chulalongkorn University. His research interest is on speech processing and speech recognition of Thai language. Some of his publications are listed as follows.

Ahkuputra V., Jitapunkul S., Maneenoi E., and Luksaneeyanawin S. "Acoustic Modelling of Vowel Articulation on the Nine Thai Spreading Vowels" – **to be published in the International Journal of Computer Processing of Oriental Languages (IJCPOL)**

Ahkuputra V., Jitapunkul S., Maneenoi E., and Luksaneeyanawin S. "Modelling of Acoustic Speech Unit for Tone Language : the Onset and Rhyme model" (unpublished)

Jitapunkul S., Ahkuputra V., Wutiwiwatchai W., Maneenoi E., Kasuriya S., Amornkul P., and Luksaneeyanawin S. "Thai Automatic Speech Recognition Model". Interdisciplinary Approaches to Language Processing, Burnham D., Luksaneeyanawin S., Davis C., and Lafourcade M. eds., 2000, ISBN: 1-86341-859-8, pages 195-213.

Ahkuputra V., Jitapunkul S., Luksaneeyanawin S., Maneenoi E., Thathong U., and Thampanitchawong, B. "Direct Classification of Thai Monophthongs on Two Dimensional Acoustic-Phonetic Feature Spaces in Linear, Mel, Bark, and Bark-Difference Frequency Scales" Proceedings of the 140th Meeting of the Acoustical Society of America, Newport, California, U.S.A., December 2000.

Ahkuputra V., Jitapunkul S., Jittiwarangkul N., Maneenoi E., and Sawit Kasuriya "A Comparison of Thai Speech Recognition Systems using Hidden Markov Model, Neural Network, and Fuzzy-Neural Network" Proceedings of the International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia, December 1998.

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 (APCCAS'98), Chiengmai, Thailand, November 1998.

Ahkuputra V., Jitapunkul S., Pornsukchandra W., and Luksaneeyanawin S. "A Speaker-Independent Thai Polysyllabic Word Recognition Using Hidden Markov Model". Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'97), Victoria, Canada, August 1997, pages 593-599.