การศึกษาผลกระทบต่าง ๆ ทางภาษาศาสตร์ต่อการรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทย

นายณัฐกร ทับทอง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2544

ISBN 974-03-1151-2
จุฬาลงกรณ์มหาวิทยาลัย

A STUDY OF VARIOUS LINGUISTIC EFFECTS ON TONE RECOGNITION IN THAI CONTINUOUS SPEECH

Mr.Nuttakorn Thubthong

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic year 2001

ISBN 974-03-1151-2

Thesis Title	A Study of Various Linguistic Effects on Tone Recognition in Thai Continuous Speech
By	Mr.Nuttakorn Thubthong
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor Boonserm Kijsirikul, Ph.D.
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 Punyakeow, D.Eng.)

THESIS COMMITTEE

..... Chairman

(Associate Professor Somchai Prasitjutrakul, Ph.D.)

..... Thesis Advisor

(Assistant Professor Boonserm Kijsirikul, Ph.D.)

..... Thesis Co-advisor

(Assistant Professor Sudaporn Luksaneeyanawin, Ph.D.)

..... Member

(Associate Professor Prabhas Chongstitvatana, Ph.D.)

..... Member

(Assistant Professor Thanaruk Theeramunkong, Ph.D.)

ณัฐกร ทับทอง : การศึกษาผลกระทบต่าง ๆ ทางภาษาศาสตร์ต่อการรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทย . (A STUDY OF VARIOUS LINGUISTIC EFFECTS ON TONE RECOGNITION IN THAI CONTINUOUS SPEECH) อ. ที่ปรึกษา : ผศ.ดร.บุญเสริม กิจศิริ-กุล, อ.ที่ปรึกษาร่วม : ผศ.ดร.สุดาพร ลักษณียนาวิน 125 หน้า. ISBN 974-03-1151-2.

การวิจัยนี้มุ่งศึกษาผลกระทบจากปัจจัยทางภาษาศาสตร์ อันได้แก่ โครงสร้างพยางค์ บริบท ทำนองเสียง และเสียงหนัก/เบา ต่อการรู้จำวรรณยุกต์ในคำพูดต่อเนื่องภาษาไทย และพัฒนาแบบจำลองวรรณยุกต์เพื่อแก้ปัญหาผลกระทบดังกล่าว การวิจัยเริ่มจากการศึกษาผลกระทบของหน่วยเสียงพยัญชนะต้น สระ และ พยัญชนะตัวสะกด ต่อการรู้จำเสียงวรรณยุกต์ในคำพูดเดี่ยว ผู้วิจัยเสนอลักษณะสำคัญของเสียงวรรณยุกต์ ชุดใหม่ซึ่งให้ผลการรู้จำที่ดีกว่าลักษณะสำคัญที่ใช้กันแต่เดิม นอกจากนี้ ผู้วิจัยได้ ศึกษาการผสมผสานตัวแยกแยะแทนการใช้ตัวแยกแยะเดี่ยว เพื่อเพิ่มอัตราการรู้จำ

ผู้วิจัยได้พัฒนากรอบงานการรู้จำเสียงวรรณยุกต์พื้นฐานสำหรับคำพูดต่อเนื่องภาษาไทย โดยกรอบงานประกอบด้วยแบบจำลองวรรณยุกต์ และตัวแยกแยะ โดยที่แบบจำลองวรรณยุกต์จะพิจารณาใช้องค์ประกอบที่สำคัญในการจำแนกเสียงวรรณยุกต์ คือ ลักษณะสำคัญของเสียงวรรณยุกต์ หน่วยความถี่มูลฐาน เทคนิคการปรับบรรทัดฐาน และส่วนประกอบของพยางค์ที่เป็นตัวเกาะของวรรณยุกต์ ขณะที่ตัวแยกแยะจะใช้ข่ายงานระบบประสาท

จากนั้น ผู้วิจัยได้ศึกษาผลกระทบของบริบทต่อเสียงวรรณยุกต์ และได้เสนอชุดลักษณะสำคัญ เรียกว่า ลักษณะสำคัญของเสียงวรรณยุกต์แบบพึ่งพาบริบท (contextual tone features) เพื่อแก้ผลกระทบจากบริบท พบว่าอัตราการลดลงของความผิดพลาดสูงสุดเท่ากับ 56.17 42.47 และ 42.42 เปอร์เซ็นต์ สำหรับฐานข้อมูล TPC PC-99 และ TASC ตามลำดับ นอกจากนั้น ผู้วิจัยได้ทดลองแบบจำลองวรรณยุกต์แบบขึ้นกับบริบท (context-dependent tone model) และเสนอแบบจำลองครึ่งวรรณยุกต์ (half-tone model) พบว่าแบบจำลองทั้งสองให้อัตราการรู้จำดีขึ้น แต่เวลาในการฝึกของแบบจำลองครึ่งวรรณยุกต์น้อยกว่าถึง 1 ใน 4 ของแบบจำลองวรรณยุกต์แบบขึ้นกับบริบท

จากนั้น ผู้วิจัยได้ศึกษาผลกระทบจากทำนองเสียง และเสนอวิธีการปรับทำนองเสียงเพื่อลดผลกระทบจากทำนองเสียง ซึ่งการปรับดังกล่าวทำให้อัตราการลดลงของความผิดพลาดสูงสุด คือ 22.20 และ 16.84 เปอร์เซ็นต์ สำหรับฐานข้อมูล TASC แบะ TPC ตามลำดับ

จากนั้น ผู้วิจัยได้ศึกษาผลกระทบจากเสียงหนัก/เบา โดยเริ่มจากการออกแบบวิธีการแยกเสียงหนักเบา โดยใช้ลักษณะทางสวนศาสตร์ต่าง ๆ คือ ระยะเวลา พลังงาน และความถี่มูลฐาน โดยศึกษาจากหน่วยเสียงขนาดต่าง ๆ คือ สระ หน่วยตาม (rhyme) และ พยางค์ จากการทดลองพบว่า การใช้หน่วยเสียงตามให้ผลการแยกแยะดีที่สุด ผู้วิจัยยังได้เสนอวิธีการแยกเสียงหนัก /เบา (separated stress method) และวิธีการรวมลักษณะสำคัญของเสียงหนัก /เบา (incorporated stress feature method) เพื่อลดผลกระทบของเสียงหนัก/เบา ที่มีต่อการรู้วรรณยุกต์ จากผลการทดลองพบว่า ทั้งสองวิธีช่วยเพิ่มอัตราการรู้จำ โดยมีอัตราการลดลงของความผิดพลาดสูงสุดที่ 32.43 และ 27.16 เปอร์เซ็นต์ สำหรับฐานข้อมูล TPC และ TASC ตามลำดับ

ท้ายสุด ผู้วิจัยได้นำแบบจำลองวรรณยุกต์ชนิดต่าง ๆ มาประยุกต์กับระบบการรู้จำเสียงพูดระดับพยางค์ ซึ่งจากการทดลองพบว่า อัตราการลดลงของความผิดพลาดสูงสุด คือ 85 .16 และ 75.06 เปอร์เซ็นต์ สำหรับฐานข้อมูล TPC และ TASC ตามลำดับ

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต.....ลายมือชื่ออาจารย์ที่ปรึกษา.....ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....
สาขาวิชา	วิศวกรรมคอมพิวเตอร์	
ปีการศึกษา	2544	

4171815021 :MAJOR COMPUTER ENGINEERING
KEY WORD: TONE RECOGNITION / SYLLABLE STRUCTURE / COARTICULATION / INTONATION / STRESS

NUTTAKORN THUBTHONG : A STUDY OF VARIOUS LINGUISTIC EFFECTS ON TONE RECOGNITION IN THAI CONTINUOUS SPEECH. THESIS ADVISOR : ASST. PROF. BOONSERM KIJSIRIKUL, Ph.D., THESIS COADVISOR : ASST. PROF. SUDAPORN LUKSANEEYANAWIN, Ph.D., 125 pp. ISBN 974-03-1151-2.

This thesis studies various linguistic effects, i.e., syllable structure, coarticulation, intonation and stress on tone recognition in Thai continuous speech. Tone models for compensating these effects are also developed. We first study the effect of initial consonants, vowels, and final consonants on tone recognition in isolation. Based on the observation on F_0 contours, we proposed a novel tone feature set. The new feature set achieved better recognition rates than the conventional tone feature sets. We also explored several combinations of classifier schemes and found that the combinations of classifiers were superior to a single classifier.

Next, we developed a basic tone recognition framework for Thai continuous speech. The framework consisted of tone models used to parameterize F_0 contours of tones and a classifier used to evaluate the performance of the tone models. We conducted experiments to construct the tone models by concentrating on tone features, frequency scales, normalization techniques, and tone critical segments. The classifier was developed using a feed-forward neural network.

Next, we focussed on tone coarticulation effect. We have proposed a feature set called "*contextual tone features*" that captured the F_0 realizations of the neighboring syllables. The features provided the best tone error reduction rates of 56.17%, 42.47%, and 42.42% for Thai Proverb Corpus (TPC), Potisuk-1999 Corpus (PC-99), and Thai Animal Story Corpus (TASC), respectively. Furthermore, we explored the context-dependent tone model (CD-T-175) and developed a novel model, *half-tone model* (H-T-30). Both models increased recognition rates, but the training time of H-T-30 was one-fourth of CD-T-175.

Next, we studied the effect of intonation on tone recognition. We obtained two methods, i.e., *beginning-point intonation normalization* and *center-point intonation normalization* methods to compensate the intonation effect. Both methods significantly increased recognition rates. The best error reduction rates of 22.20% and 16.84% were achieved for TASC and TPC, respectively.

Next, we concentrated on stress effect. We first performed two empirical experiments of stress detection on pairs of ambiguous words and poly-syllabic words. We explored acoustic features, i.e., duration, energy, and F_0 extracted from several linguistic units, i.e., vowel, syllable and rhyme units. The rhyme unit outperformed the other units for stress detection. We then performed an empirical study of tone recognition. We have proposed two methods, i.e., *separated stress method* (SSM) and *incorporated stress feature method* (ISFM). Both methods increased the tone recognition rates. We additionally incorporated ISFMs into the tone model and found that TSFM improved the recognition rates. The highest error reduction rates of 32.43% and 27.16% were reported for TPC and TASC, respectively.

Finally, we integrated several refined tone models into a syllable-based speech recognition system to enhance the recognition performance. We achieved the best error reduction rates of 85.16% and 75.06% for TPC and TASC, respectively.

Department	Computer Engineering	Student's signature .....
Field of study	Computer Engineering	Advisor's signature .....
Academic year	2001	Co-advisor's signature .....

ACKNOWLEDGEMENTS

Many people have contributed to the success of this thesis. I would like to especially thank my thesis advisor, Assistant Professor Boonserm Kijsirikul, for his insightful suggestions, comments, and feedback throughout this thesis. I would like to thank my thesis co-advisor, Assistant Professor Sudaporn Luksaneeyanawin, for many helpful comments and discussions on issues related to this thesis.

I would like to sincerely thank my other thesis committee members, Associate Professor Somchai Prasitjutrakul, Associate Professor Prabhas Chongstitvatana and Assistant Professor Thanaruk Theeramunkong, who provided valuable advice at committee meetings.

I thank all the volunteer speakers, especially Yanin Sawanakunanon, who participated in my experiments. Your patience in enduring the long recording sessions is highly appreciated. I also thank all of the member of the Machine Intelligence and Knowledge Discovery laboratory and the Centre for Research in Speech and Language Processing for engaging discussions. Very special thanks to Nuanwan Soonthornphisaj. Without you, my student life would have been much lonely.

I especially thank Kamonwadee Sirikarnjanawong, my best friend. Although we may not spend a lot of time together, I can always count on her whenever I need help.

I would link to thank Kittiwut Choochote, Sitara Nuanyai and Wisit Leelasiriwong for all their help. I also wish to thank Suthasinee Sithigasorn, Wityada Thongdaeng and Onwadee Rukkharangsarit for their friendship, their kindness, and their concern. I greatly appreciate Narumons Suwonjandee, a graduate student of University of Cincinnati, for the required papers that you downloaded and sent me.

Most of all, I especially thank Pantawan for good words and being a fun companion. Although the study was too stressful, I still survive because of your words. I would also like to extend my appreciation to my obstinate girl and my good girl for their patience, their trust, their humor, and their encouragement.

Of course all this would not have been possible without the constant encouragement and support provided by my parents. I would like to thank my parents for their support and encouragement throughout my graduate study.

In addition, I would like to thank office of the Civil Service Commission and Ministry of University Affairs for financial supporting.

CONTENTS

ABSTRACT (THAI) iv

ABSTRACT (ENGLISH) v

ACKNOWLEDGEMENT vi

CONTENTS vii

LIST OF TABLES xi

LIST OF FIGURES xv

1 INTRODUCTION 1

1.1 Motivation 1

1.1.1 Syllable structure 2

1.1.2 Coarticulatory effect 3

1.1.3 Intonation effect 3

1.1.4 Stress 4

1.2 Thesis Goals 5

1.3 Overview 5

2 EFFECT OF SYLLABLE STRUCTURE ON TONE RECOGNITION 7

2.1 Background on Thai Syllable Structure 7

2.2 Related Works 9

2.3 Methodology 10

2.3.1 Voiced portion detection 10

2.3.2 Tone feature extraction 11

2.3.3 Normalization 13

2.3.4 Neural network classifier 13

2.4 Experiments 14

2.4.1 Speech corpus 14

2.4.2 Training set sampling method 16

2.4.3 Experiment I : Dependent data experiment 16

2.4.4 Experiment II : Mixed data experiment 23

2.4.5 Experiment III : Cross data experiment 23
2.4.6 Discussion 24
2.4.7 Human perception test 26
2.5 Combination of Neural Networks 27
2.5.1 Method 27
2.5.2 Results and discussion 29
2.6 Summary 33

3 CONSTRUCTING TONE RECOGNITION FRAMEWORK FOR THAI CONTINUOUS SPEECH 34
3.1 Thai Speech Corpora 35
3.2 Classifiers 36
3.3 Experimental Setting 37
3.4 Tone Features 38
3.5 Frequency Scale 40
3.6 Normalization Technique 41
3.7 Tone-Critical Segment 43
3.8 Discussion 45
3.9 Summary 48

4 EFFECT OF COARTICULATION ON TONE RECOGNITION 49
4.1 Related Works 52
4.2 Methodology 53
4.2.1 Contextual tone features (CTF) 53
4.2.2 Tone models 54
4.3 Experiments 57
4.3.1 Experiments of different contextual tone features 57
4.3.2 Experiments of different tone models 59
4.4 Discussion 59
4.5 Summary 59

5 EFFECT OF INTONATION ON TONE RECOGNITION 62
5.1 Related Works 63
5.2 Methodology 66

5.3 Experiments 67
5.4 Discussion 69
5.5 Summary 71

6 EFFECT OF STRESS ON TONE RECOGNITION 73
6.1 Stress Detection 74
6.1.1 Accentual system of polysyllabic words in Thai 76
6.1.2 Acoustic features of stress 76
6.1.3 Speech corpora 78
6.1.4 Experiment I: Stress detection on pairs of ambiguous words . . . 79
6.1.5 Experiment II: Stress detection on polysyllabic words 83
6.1.6 Discussion 85
6.2 Tone Recognition 86
6.2.1 Stress feature methods 86
6.2.2 Experiments 87
6.2.3 Discussion 90
6.3 Experiments of Tone Recognition on Continuous Speech 91
6.4 Summary 92

7 COMPARISON OF SEVERAL REFINED TONE MODELS AND INCORPORATION OF TONE MODELS INTO SPEECH RECOGNITION 94
7.1 Comparison in Performance of Several Refined Tone Models 94
7.2 Incorporation of Tone Models into Speech Recognition 95
7.2.1 Syllable-based speech recognition framework 95
7.2.2 Integrating tone models into speech recognition 99
7.2.3 Experiments 100
7.3 Summary 100

8 SUMMARY AND FUTURE WORKS 102
8.1 Summary and Contributions 102
8.2 Further Works 107

REFERENCES 110

A THE INTERNATIONAL PHONETIC ALPHABET OF THAI PHONEMES 121

A.1 Consonants 121

A.2 Vowels 122

A.3 Tones 122

B PUBLICATIONS 123

B.1 National Conferences 123

B.2 International Conferences 123

B.3 International Journals 124

BIOGRAPHY 125

LIST OF TABLES

2.1 Thai consonants. 8

2.2 Thai vowels. 8

2.3 Recognition rates (%) and standard deviations (s.d.) of sets A, B, and C using different tone feature sets. 17

2.4 Recognition rates (%) and standard deviations (s.d.) of set A reported separately according to consonant and tone feature sets. 18

2.5 Recognition rates (%) and standard deviations (s.d.) of set B reported separately according to vowel and tone feature sets. 18

2.6 Recognition rates (%) and standard deviations (s.d.) of set C reported separately according to final consonant and tone feature sets. 18

2.7 Confusion matrices of tone recognition for (a) set A, (b) set B, and (c) set C using different tone feature sets. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively. 21

2.8 Recognition rates (%) and standard deviations (s.d.) of the mixed data experiment using different tone feature sets. 24

2.9 Recognition rates (%) and standard deviations (s.d.) of the cross data experiment using different tone feature sets. 24

2.10 Confusion matrices of human perception test. 27

3.1 Summary of all three corpora used in our experiments for each fold. . . 38

3.2 Recognition rates (%) and standard deviations (s.d.) of tone recognition with different tone features. The best results for each corpus are printed in **boldface**. 40

3.3 Measure of statistical difference of tone recognition with different tone features. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). . . 40

3.4 Recognition rates (%) and standard deviations (s.d.) of tone recognition with different frequency scales. The best results for each corpus are printed in **boldface**. 42

3.5 Measure of statistical difference of tone recognition with different frequency scales. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). 42

3.6 Recognition rates (%) and standard deviations (s.d.) of tone recognition with different normalization techniques. The best results for each corpus are printed in **boldface**. 43

3.7 Measure of statistical difference of tone recognition with different normalization techniques. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). 43

3.8 Recognition rates (%) and standard deviations (s.d.) of tone recognition with different tone critical segments. The experiment was performed on PC-99 (both inside and outside tests). The best results for each test are printed in **boldface**. 45

3.9 Measure of statistical difference of tone recognition with different tone critical segments. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). 45

3.10 Confusion matrices of tone recognition for (a) PC-99, (b) TPC, and (c) TASC using the best configurations. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively. 47

4.1 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition with different contextual tone features for PC-99 (outside test), TPC, and TASC. 58

4.2 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition with different tone models for PC-99 (outside test), TPC, and TASC. CTF34 was used. 60

4.3 Measure of statistical difference of tone recognition with different frequency scales. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).) 60

4.4 Training times of tone recognition with different tone models for PC-99 (outside test), TPC, and TASC. CTF34 was used. The experiments were run on a Pentium III 866MHz machine. 60

4.5 Confusion matrices of tone recognition for (a) PC-99, (b) TPC, and (c) TASC using CTF34 and CI-T-5. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively. 61

5.1 Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of tone recognition with different intonation normalization methods on Thai proverb corpus (TPC) and Thai animal corpus (TASC). The best recognition rates for each corpus are printed in **boldface**. . . . 69

5.2 Measure of statistical difference of tone recognition with different intonation normalization methods. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). 69

5.3 Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of tone recognition with the refinements of contextual tone features and different intonation normalization methods on Thai proverb corpus (TPC) and Thai animal corpus (TASC). 70

5.4 Measure of statistical difference of tone recognition with different intonation normalization methods. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05). 70

5.5 Confusion matrices of tone recognition for (a) TPC and (b) TASC using CIN. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively. 71

6.1 Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of stress detection for PC-96 with different linguistic units and stress feature sets. The best recognition rates for each unit are printed in **boldface**. 81

6.2 Confusion matrix of stress detection for PC-96 using SF5 and the rhyme unit. 81

6.3 Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of stress detection for TPNC using different linguistic units and stress feature sets. The best recognition rates for each unit are printed in **boldface**. 84
6.4 Confusion matrix of stress detection for TPC using SF5 and the rhyme unit. 84
6.5 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for PC-96 using different tone feature sets. The highest recognition rates for stressed, unstressed and total syllables are printed in **boldface**. 88
6.6 Confusion matrices of tone recognition of (a) stressed and (b) unstressed syllables for PC-96 using baseline+ISFM (SF5). 88
6.7 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for TPNC using different tone feature sets. The highest recognition rates for stressed, unstressed and total syllables are printed in **boldface**. 89
6.8 Confusion matrices of tone recognition of (a) stressed and (b) unstressed syllables for TPNC using ISFM (SF5). 89
6.9 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for TPC and TASC using different tone feature sets. The best recognition rate for each corpus is printed in **boldface**. 91

7.1 Tone recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of the simple tone model and more refined tone models on TPC and TASC. 95
7.2 Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of syllable-based speech recognition for TPC and TASC when incorporating stress feature (SF5) and several refined tone models. 101

LIST OF FIGURES

1.1 F_0 contours of the five Thai tones when syllables are spoken in isolation by a male speaker. 2

1.2 The waveform and F_0 contours of tones in an FHRL sequence when (a) each word is spoken in isolation and (b) when the whole utterance is naturally spoken. 2

2.1 The architecture of the isolated Thai tone recognition system. 10

2.2 The search algorithm proposed by (Jun et al. 1998). 11

2.3 Lists of hypothetical syllables in (a) set A, (b) set B, and (c) set C. . . . 15

2.4 Average *normalized* F_0 contours of the five Thai tones of (a) set A, (b) set B, and (c) set C. 19

2.5 Average *normalized* F_0 contours of the five Thai tones of (a) set A, (b) set B, and (c) set C. For each tone, the F_0 are plotted separately according to initial consonants, vowels, and final consonants, respectively. 20

2.6 Recognition rates of the cross data experiment with different training and test sets using different tone feature sets. 25

2.7 Average recognition rates of three main experiments using different tone feature sets. 25

2.8 The neural network combination. 28

2.9 Recognition rates (%) and standard deviations (s.d.) of ten experiments with different training and test sets using (a) probability combination rules and (b) voting techniques. 30

2.10 Recognition rates (%) and standard deviations (s.d.) of ten experiments with different training and test sets using different combination schemes. 31

3.1 Comparison performances of inside and outside tests on PC-99 with different configurations. 46

3.2 Comparison performances of all three corpora with different configurations. 46

4.1 Carry-over effect: effect of preceding tones on F_0 contour of following tone in /maː maː/ sequences in Thai. In each panel, the tones in the second syllable was held constant (the mid, the low, the fall, the high, and the rise in (a) to (e), respectively), and the tone of the first syllable was varied. Each curve was from a female speaker. 50

4.2 Anticitory effect: effect of following tones on F_0 contour of preceding tone in /maː maː/ sequences in Thai. In each panel, the tones in the first syllable was held constant (the mid, the low, the fall, the high, and the rise in (a) to (e), respectively), and the tone of the second syllable was varied. Each curve was from a female speaker. 51

4.3 The different time points of (a) the baseline tone features (simple tone models), (b) the contextual tone features (for CTF34), (c) the first-half tone features (for CTF34), and (d) the second-half tone features (for CTF34). The circle points represent five F_0 at different time points of a syllable and bar points represent the F_0 used for different tone features. The dash lines and dot lines represent syllable boundaries and onset-rhyme boundaries, respectively. 55

4.4 The process of the half-tone model. 57

4.5 Error rates of different contextual tone features. 58

5.1 The F_0 contours of three intonation types: (a) falling intonation, (b) rising intonation, and (c) convolution intonation, with all-points lines (A), baselines (B), and topline (T). 64

5.2 F_0 contours of all utterances. The '×' line represents the mean F_0 contour, with the upper and lower '+' lines for standard deviation. The dot line is the first-order polynomial regression line for the average F_0 contour. 67

5.3 F_0 contours: (a) before applying intonation normalization, (b) after applying beginning-point intonation normalization, and (c) after applying center-point intonation normalization. 68

5.4 Recognition rates of (a) 4-syllabic utterances, (b) 5-syllabic utterances, and (c) 6-syllabic utterances for TPC, plotted separately according to syllable positions. 70

6.1 Mean F_0 contours of all five tones in (a) stressed and (b) unstressed syllables. Data were normalized within speaker, and across tones and stress categories. 74
6.2 Accentual system of the polysyllabic words in Thai where ‘ˈ’ is an accented maker that is in front of an accented syllable (Luksaneeyanawin 1983). . 77
6.3 Histogram distribution of (a) duration, (b) total energy, (c) maximum energy, and (d) F_0, measured for the rhyme unit in PC-96. 82
6.4 Recognition rates of stress detection for TPNC along five stress feature sets with the rhyme unit. 85
6.5 The separated stress method (SSM). 87
6.6 Error rates (%) of tone recognition along different rhyme durations for TASC. 92

7.1 Tone recognition rates of all five tones using several refined tone models for (a) TPC and (b) TASC. 96
7.2 Architecture of the syllable-based speech recognition framework. 97
7.3 A syllable-based speech recognition framework based on a three-layer feedforward neural network. Each syllable is represented by 15 frame features of RASTA coefficients. 98
7.4 Normal distribution and the position of $minF_i$ and $maxF_i$. 99

A.1 The IPA of Thai consonants. 121
A.2 The IPA of Thai vowels. 122
A.3 The IPA of Thai tones. 122

CHAPTER 1

INTRODUCTION

During the past decade, speech recognition technology has undergone significant progress. Several applications of speech recognition to human-computer interface have been developed since speech is the most natural way of human communication and interaction. Most existing methods for speech recognition are developed mainly for spoken English, and some of them have been adapted to be applicable to Thai language. However, unlike English, Thai is a tone language. In such a language, the referential meaning of an utterance is dependent on the lexical tones (Jian 1998). Therefore, a tone classifier is an essential component of a speech recognition system of a tone language.

In this chapter, we first motivate the importance and the difficulties of tone recognition in continuous speech. We then introduce the general goal and approach of this thesis. Finally, we give a chapter by chapter overview.

1.1 Motivation

In Thai, there are five different lexical tones as follows: the mid /M/, the low /L/, the fall /F/, the high /H/, and the rise /R/. The following examples show the effect of tones on the meaning of an utterance (Luksaneeyanawin 1998): M /khāː/ ("a kind of grass"); L /khàː/ ("galangale"); F /khâː/ ("to kill"); H /kháː/ ("to trade"); and R /khǎː/ ("a leg"). The tone information is superimposed on the voiced portion[1] of each syllable. The identification of a Thai tone relies on the shape of the fundamental frequency[2] (F_0) contour. Figure 1.1 shows the average of F_0 contours of five different tones when syllables are spoken in isolation by a male speaker.

Although there are only five different tones, the tone behavior is very complicated in continuous speech. Figure 1.2 shows the comparison of F_0 realization of an FHRL sequence when each monosyllabic word is spoken in isolation (see top panel) and when all four tones are spoken naturally in running speech (see bottom panel). The tones

[1]Voiced portion is the portion of sound that is produced when the vocal cords are tensed together and they vibrate in a relaxation mode as the air pressure build up, forcing the glottis open, and then subsides as the air passes through (Owens 1993).

[2]Fundamental frequency is the acoustic correlate of pitch and is defined as the frequency of vibration of the vocal fold (Taylor 1992). Fundamental frequency descriptions are normally represented as F_0 *contours*, which are plots of F_0 against time.

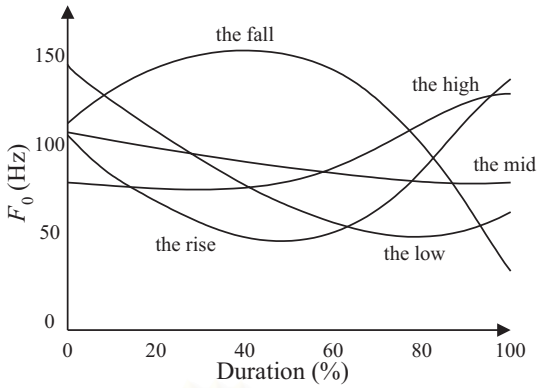


Figure 1.1: F_0 contours of the five Thai tones when syllables are spoken in isolation by a male speaker.

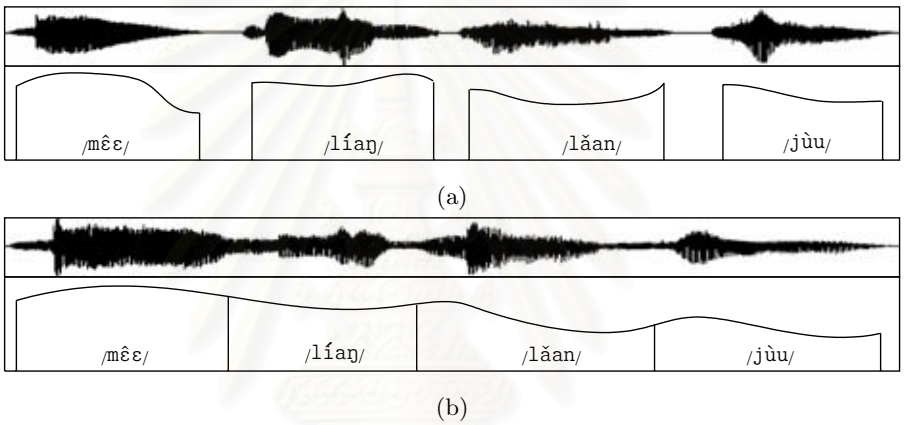


Figure 1.2: The waveform and F_0 contours of tones in an FHRL sequence when (a) each word is spoken in isolation and (b) when the whole utterance is naturally spoken.

produced on isolation words are very similar to those in Figure 1.1, while tones produced on words in continuous speech are much more difficult to identify. Several interacting factors affect F_0 realization of tones, e.g., syllable structure, coarticulation, intonation, stress, speaking rate, dialect, sex, age, and emotion (Botinis et al. 2001; Gandour et al. 1994; Potisuk et al. 1999; Potisuk, Gandour, and Harper 1996; Gandour et al. 1999; Chen and Chang 1992). However, in this thesis, we focus on the first four effects only.

1.1.1 Syllable structure

Syllable structure affects tone classification in term of phonology and acoustic phonetics. The former will be described in Section 2.1. The latter affects tone classification due to the *continuity effect* on the F_0 contour in terms of the voiced/unvoiced

property of contiguous phones (Chen and Chang 1992). A continuously voiced stretch of speech has a continuous F_0 contour; a stretch of speech with intervening voiceless obstruents has a discontinuous F_0 contour. There are also consonantally-induced perturbations on the F_0 contour of the following vowel (Potisuk et al. 1999). Furthermore, high vowels may have higher tonal realization than low vowels (Botinis et al. 2001). In the past few years, some studies of syllable structure on Thai tone recognition in isolation have been conducted. Thubthong (1995) proposed a set of tone features to recognize the five Thai tones. He concentrated on the effects of initial consonants, and vowels on tone recognition. Tungthangthum (1998) used raw F_0's and hidden Markov models to classify the tones in isolation. His study shows that tones and vowels are independent from each other but his result was done on single-speaker tone recognition only. However, the effect of syllable structure on Thai tone recognition is still not clear.

1.1.2 Coarticulatory effect

The tone of a neighboring syllable influences the shape and level of the F_0 contour (Wang and Chen 1994). The effects of the following syllable and the preceding syllable are called *anticipatory coarticulation* and *carry-over coarticulation*, respectively. The problem of coarticulatory effect has been studied by many researchers in linguistics and phonology (Gandour et al. 1994; Shen 1990; Xu 1994; Xu 1999). There are a number of studies of tone recognition on this effect in Mandarin. These studies were based on tone modelling (Wang and Seneff 2000; Zhang and Hirose 2000) and classification techniques (Chen and Wang 1995). In Thai, there is only one study of tone recognition in coarticulatory effect (Potisuk et al. 1999). The study used an extension to Fujisaki's model (Fujisaki 1983) for modelling three tone sequences. They conducted experiment on a linguistically designed corpus and classification result was given in term of a tone sequence (not a tone in each syllable). The effect of coarticulatory on tone recognition is an open problem for Thai. We will empirically study the effect on larger corpora and design the tone models for compensating this effect.

1.1.3 Intonation effect

Intonation is defined as the combination of tonal features into larger structural units associated with the acoustic parameter of *voice fundamental frequency* and its dis-

tinctive variations in the speech process (Botinis et al. 2001). The intonation effect plays an important role in tone recognition in term of F_0 height adjustment of tones (Potisuk et al. 1995). There is no un-criticized method available to quantitatively determine the slope and domain of intonation contour yet. Although intonational phonology has proposed a relatively simple framework for describing the intonation of an utterance, i.e., as a sequence of intonational phrases, each consisting of certain categorical constituents, a suitable set of descriptive units has been elusive (Wang 2001). Most studies of tone recognition on the intonation effect were proposed for Mandarin. These studies were based on acoustic features (Wang and Seneff 2000) and classification techniques (Chen and Wang 1995; Huang and Seide 2000).

There is a study of tone recognition on intonation effect in Thai (Potisuk et al. 1999). The limitation is described in the previous subsection. We think that intonation is an important issue for speech technology. The study on intonation is not only useful for improving tone recognition, but also useful for speech synthesis. This effect is an interesting issue that should be intensively studied.

1.1.4 Stress

In speech perception, stress refers to the relative perceptual prominence of a syllable or a word in a particular context (Ying 1998). The relative prominence is produced by a change in F_0, increased duration, increased intensity, and a change in vowel quality (or timbre).

The F_0 contours of stressed syllables are generally quite different from unstressed ones (Chen and Wang 1995). For standard Thai, despite systematic changes in F_0 contours, all five tonal contrasts are preserved in unstressed as well as stressed syllables. However, F_0 contours of stressed syllables more closely approximate the contours in citation forms than those of unstressed syllables (Potisuk, Gandour, and Harper 1996; Potisuk et al. 1999).

Potisuk, Gandour, and Harper (1996) investigated acoustic correlates of stress in Thai. Stimuli consisted of 25 pairs of sentences that the first member of each sentence pair contained a two-syllable noun-verb sequence exhibiting a strong-strong stress pattern, and the second member contained a two-syllable noun compound exhibiting a weak-strong stress pattern. Five prosodic features, i.e., duration, average F_0, F_0 stan-

dard deviation, average intensity, and intensity standard deviation, were used. Results indicated that duration is the predominant cue in signaling the distinction between stressed and unstressed syllables in Thai. However, we have not found any work on tone recognition under stress effect in the literature.

1.2 Thesis Goals

The general goal of this thesis is to study the effect of syllable structure, coarticulation, intonation and stress on tone recognition in Thai continuous speech and develop a tone recognition framework for compensating these effects.

Specifically, this thesis accomplishes the following tasks:

- The acoustic correlations between tones and the phonematic units constructing the syllables.
- Tone modelling for accounting coarticulatory effect.
- Tone modelling for accounting intonation effect.
- Stress modelling for Thai tone recognition.
- Using tone models to improve speech recognition performance.

1.3 Overview

The remainder of this thesis is organized into seven chapters. Chapter 2 describes empirical studies of the effect of initial consonants, vowels, and final consonants on tone recognition. We first provide some understanding of the phonology and phonetics of Thai syllable structure and some related works. After that, we present a methodology for tone recognition by considering the syllable structure effect. Then, several experiments are conducted to study the correlations between tones and the other phonemes. In addition, the study of various combination schemes is described to enhance the performance of the tone recognition.

Chapter 3 presents an empirical study for constructing the basic tone recognition framework for Thai continuous speech. The framework consists of the simple tone models used to parameterize F_0 contour and a classifier, a feedforward neural network, used to evaluate the performance of the tone models. The simple tone modelling is focussed

on the question of which configurations with respect to tone features, frequency scale, normalization technique, and tone critical segment should be used for tone recognition. We first introduce three Thai speech corpora used in our experiments: Potisuk-1999 corpus, Thai Proverb Corpus, and Thai Animal Stories Corpus. Then, we describe the experimental setting for evaluating all aspects. After that we demonstrate four experiments for answering the question of four configurations above.

Chapter 4 presents a tone recognition study focussing on the coarticulatory effect. We propose tone features to compensate this effect and the features are incorporated into the framework to refine the tone models. We also propose a novel model called *half-tone model* to improve the performance of tone recognition. We perform experiments on all three corpora and analyze the results in several directions.

Chapter 5 extends the framework described in Chapter 3 and 4 to model the intonation effect for improving tone recognition. We first present some related works on intonation and the effect of intonation on tone recognition. Then, we describe a method, called *intonation normalization*, to compensate this effect. After that, we evaluate the method by simulating a number of experiments on Thai proverb corpus and Thai animal story corpus. The experimental results are discussed in several aspects.

Chapter 6 describes the effect of stress on tone recognition. We first study the correlation of duration, energy, and F_0 measurements with stress on pairs of ambiguous words and polysyllabic words, and identify the most informative features of stress by experiments. We then describes tone recognition experiments. Based on the knowledge of stress, we have proposed two methods, i.e., *separated stress method* (SSM) and *incorporated stress features method* (ISFM) to alleviate the stress effect. Tone recognition experiments are conducted, and some analysis and interpretation of the tone recognition results are provided.

In Chapter 7, we first demonstrate the performance comparison of several refined tone models for accounting various interacting factors. We then describe the implementation of a mechanism in the recognition system for incorporating tone model constraints. A suite of speech recognition experiments are conducted to compare the contributions of using various tone models.

Chapter 8 summarizes the thesis, discusses contributions and suggests directions for future works.

CHAPTER 2

EFFECT OF SYLLABLE STRUCTURE ON TONE RECOGNITION

For a tone language such as Thai, fundamental frequency plays a critical role in charactering tones, which is an essential lexical feature (Wang 2001). There are many interacting factors affecting F_0 contours, e.g., syllable structure, coarticulation, intonation, stress, etc. In this chapter, we present empirical studies of the first factor, syllable structure (Thubthong et al. 2000a). The goals of this chapter are to study the effect of initial consonants, vowels, and final consonants on tone recognition and propose for isolated Thai tone recognition a method that extracts and makes use of the acoustic features of Thai tones. Moreover, recently, the classifier combination approach has been repeatedly proven to be more robust than the single classifier approach (Kirchhoff and Bilmes 1999). Therefore, we also apply combination approaches to improve the performance of isolated Thai tone recognition.

The following section provides some understanding of the phonology and phonetics of Thai syllable structure. We then describe some related works. Next, we present a methodology for tone recognition by considering syllable structure effect. Then, several experiments are conducted to study the correlations between tones and the other phonemes. In addition, the study of various combination schemes to enhance the performance of the recognition is described. Finally, we conclude this chapter with a brief summary.

2.1 Background on Thai Syllable Structure

The phonetic structure of Thai is based primarily upon the monosyllable. Thai syllable structure is /C(C)V(ː)(C)$^{\mathrm{T}}$/ where C, V, 'ː' and $^{\mathrm{T}}$ represent a consonant, vowel, vowel length, and lexical tone, respectively (Luksaneeyanawin 1998). There are 33 initial consonants (including clusters), 24 vowels, 8 final consonants, and 5 tones. The detail of each phoneme is shown below.

1. The consonant phonemes

   In Thai, there are 21 consonants as shown in Table 2.1. All of these 21 conso-

Table 2.1: Thai consonants.

		Labial	Alveolar	Palatal	Velar	Glottal
Stop	Voiceless Unaspirated	p*	t*	c	k*	ʔ*
	Voiceless Aspirated	ph	th	ch	kh	
	Voiced	b	d			
Non-stop	Nasal	m*	n*		ŋ*	
	Fricative	f	s			h
	Trill		r			
	Lateral		l			
	Approximant	w*		j*		

*Consonants with asterisks can occur in syllable initial and syllable final positions (Luksaneeyanawin 1993).

Table 2.2: Thai vowels.

	Front	Center	Back
High	i, iː	ɯ, ɯː	u, uː
Mid	e, eː	ɤ, ɤː	o, oː
Low	ɛ, ɛː	a, aː	ɔ, ɔː

nants can occur in the syllable initial position but there are only nine consonants that can also occur in the syllable final position (the ones with asterisks in Table 2.1). Except for borrowed words, there are 12 consonant clusters, i.e., /kr/, /kl/, /kw/, /khr/, /khl/, /khw/, /tr/, /thr/, /pr/, /pl/, /phr/ and /phl/.

2. The vowel phonemes

   Thai has 18 monophthongs, nine short and nine long. A pair of short and long monophthongs is quantitatively different but qualitatively quite similar. The vowel phonemes are shown in Table 2.2. There are six diphthongs in Thai, i.e., /ia/, /iːa/, /ɯa/, /ɯːa/, /ua/ and /uːa/.

3. The tone phonemes

   There are five different lexical tones in Thai: the mid /M/, the low /L/, the fall /F/, the high /H/, and the rise /R/. They can be divided into two groups: the static group consisting of three tones (the low, the mid, and the high) and the dynamic group consisting of two tones (the rise and the fall) (Luksaneeyanawin 1993).

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

syllables with long vowels ending with stops, we found only the low, the fall, and the high but the high in this type of syllable is also scarce.

The above detail shows the effect of syllable structure on tones in phonology. Moreover, the syllable structure also affects tones in acoustic phonetics. It affects tones due to the *continuity effect* on the F_0 contour in terms of the voiced/unvoiced property of contiguous phones (Chen and Chang 1992). A continuously voiced stretch of speech has a continuous F_0 contour; a stretch of speech with intervening voiceless obstruents has a discontinuous F_0 contour. There are also consonantally-induced perturbations on the F_0 contour of the following vowel (Potisuk et al. 1999). Furthermore, high vowels may have higher tonal realization than low vowels (Botinis et al. 2001).

2.2 Related Works

In the past few years, some methods for five-tone-recognition of isolated Thai syllables have been proposed. Thubthong (1995) proposed a method for Thai tone recognition by partitioning the F_0 contour of a syllable into four equal segments and identifying the five tones based on the normalized slopes of the segments and the F_0 level of the syllable. He built two speech data sets that consisted of hypothesis syllables with varying six vowels and ten consonants, respectively, from three male and three female speakers. The recognition rate of the first set was lower than that of the second set. The results imply that vowels have much effect on tones than consonants do. Tungthangthum (1998) used raw F_0 contours and hidden Markov models to classify the five tones. He built a speech data set consisting of hypothesis syllables with varying ten vowels and five tones from a male speaker. The speech data was separated into the training set, the first test set, and the second test set. The training and first test sets consisted of the first five vowels, and the second test set consists of the other five vowels. The recognition rates on two test sets were not significantly different. His study shows that tones were independent from the vowels but his result was done on single-speaker tone recognition only. However, the effect of syllable structure on Thai tone recognition is still not clear.

We have no complete knowledge to compensate this effect. However, we could get the idea from practice. The goals of the practice are to study the correlation of tones on other phonemes, and to propose a novel tone feature set to prevent these correlations.

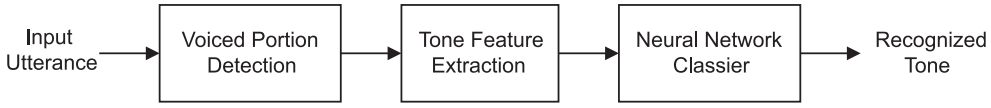


Figure 2.1: The architecture of the isolated Thai tone recognition system.

2.3 Methodology

The architecture of our isolated Thai tone recognition system is shown in Figure 2.1. The architecture starts through the voiced portion detection, and then proceeds by extracting tone feature parameters. After that, the tone feature parameters are normalized and fed into a feedforward neural network classifier to recognize the tone. The details are discussed in the following subsections.

2.3.1 Voiced portion detection

Since an F_0 does not exist in unvoiced and silent portions of speech (Chen and Wang 1995), an F_0 contour can be extracted in the voiced portion only. To detect the voiced portion, we use root mean square energy and zero-crossing rate defined in Equation (2.1) and (2.2), respectively (Lee et al. 1995). For these analyzes, we use 25 ms frame with 5 ms frame shift. The beginning and the end of voiced portions can be determined by backward and forward searching respectively, from the frame with the maximum frame energy (Lee et al. 1993). The beginning is located when the energy of a particular frame is smaller than an energy threshold, or when the zero-crossing rate is higher than a zero-crossing rate threshold. For locating the end point, only an energy threshold is used. The thresholds are determined by experiments.

$$E_n = \sqrt{\frac{1}{N}\sum_{i=1}^{N} s_n^2(i)} \tag{2.1}$$

$$Z_n = \frac{1}{N}\sum_{i=1}^{N} \frac{\mid sgn\{s_n(i)\} - sgn\{s_n(i-1)\} \mid}{2} \tag{2.2}$$

where

$$sgn\{s_n(i)\} = \begin{cases} +1 & \text{if } s_n(i) \geq 0 \\ -1 & \text{if } s_n(i) < 0 \end{cases} \tag{2.3}$$

Assume $F_1, F_2, \cdots, F_N$ is the F_0 profile of each utterance.

if $|F_n - F_{n-1}| > C_1$ and $|F_{n+1} - F_{n-1}| > C_2$
then $F'_n = 2 \times F_{n-1} - F_{n-2}$

if $|F_{n+1} - F_{n-1}| \leq C2$
then $F'_n = (F_{n-1} + F_{n+1})/2$

Where C_1 and C_2 are two thresholds, determined by experiments.

Figure 2.2: The search algorithm proposed by (Jun et al. 1998).

and E_n, Z_n, $s_n(i)$, and N is the energy of frame n, the zero-crossing rate of frame n, the i^{th} windowed speech sample in frame n, and the frame size, respectively.

2.3.2 Tone feature extraction

The Average Magnitude Different Function (AMDF) algorithm (Ross et al. 1974) is employed for F_0 extraction with 30 ms frame size and 5 ms frame shift. To correct some of the estimation errors and enforce continuity of the F_0 contour, a search algorithm (Jun et al. 1998) and the moving average smoothing are applied. The search algorithm is described in Figure 2.2. The moving average smoothing is determined by the following equation:

$$F'_n = \frac{1}{N} \sum_{i=n-N/2}^{n+N/2} F_i \tag{2.4}$$

where F_i is the i^{th} F_0, $F_n^{'}$ is the smoothed F_0 of frame n, and N is the frame size. Finally, if there are any detection errors, they are manually corrected.

Since not all syllables are of equal duration, F_0 contours are equalized for duration on a percentage scale (Gandour et al. 1994). We obtain F_0's at 11 different time points with the equal step size of 10% between 0% and 100% of the voiced portion. Each F_0 is then interpolated by Lagrange's interpolating polynomial (Gerald and Wheatley 1994) based on four points around its position. Thus the F_0 profile of each utterance has the same dimension of 11. The F_0 profile is denoted as $\{F_0(0), F_0(1), F_0(2), \cdots, F_0(10)\}$.

In this thesis, three sets of tone features are built to capture the characteristics of Thai tones.

1. **Tone Feature Set 1 (TF1)**

   As the shape of F_0 contour represents a tone, the derivatives of F_0 are extracted.

Thubthong (1995) used four slopes of F_0's, and the average of the F_0 level to classify tones. A slope of F_0's is denoted as delta F_0 (dF_0), which can be approximated by the following equation:

$$dF_0(n) = F_0(n+1) - F_0(n-1) \tag{2.5}$$

where n is the index of the F_0 profile. In this thesis, four dF_0's, computed at 20, 40, 60, and 80% of the voiced portion, are served as tone feature set 1.

2. Tone Feature Set 2 (TF2)

Tungthangthum (1998) used F_0's at every individual time instant with time step of 10 ms to recognize tones. However, tone identification does not rely on precise F_0 values at every individual time instant (Lee and Ching 1999). Therefore, we used only six F_0's as tone feature set 2. These are the initial F_0 (F_{0I}), the final F_0 (F_{0F}), and four F_0's at 20, 40, 60, and 80% of the voiced portion. The initial F_0 and the final F_0 are defined as follows:

$$F_{0I} = \frac{F_0(0) + F_0(1)}{2} \tag{2.6}$$

$$F_{0F} = \frac{F_0(9) + F_0(10)}{2} \tag{2.7}$$

3. Tone Feature Set 3 (TF3)

In the literature, the error of Thai tone recognition mainly resulted from the similarity of the F_0 contours of the low to the mid and the rise to the high (Abramson 1998; Luksaneeyanawin 1995). Therefore, we define another tone feature set to alleviate this problem. Our tone features are constructed based on the following characteristics of Thai tones observed from Figure 2.4.

1. An F_0 at the beginning of a syllable is a key factor in discriminating between the mid and the low.

2. The rise and the fall can be distinguished by the shape of the F_0 contour.

3. The F_0 level is helpful in discriminating between the high and the rise.

Based on these observations, we propose a novel tone feature set (tone feature set 3) that consists of the initial F_0, the final F_0, and four dF_0's at 20, 40, 60, and 80% of the voiced portion.

In summary, we have three tone feature sets:

(1) Tone feature set 1: $\{dF_0(2), dF_0(4), dF_0(6), dF_0(8)\}$.

(2) Tone feature set 2: $\{F_{0I}, F_0(2), F_0(4), F_0(6), F_0(8), F_{0F}\}$.

(3) Tone feature set 3: $\{F_{0I}, dF_0(2), dF_0(4), dF_0(6), dF_0(8), F_{0F}\}$.

2.3.3 Normalization

An F_0 is basically a physiologically determined characteristic and is regarded as being speaker dependent (Lee et al. 1995). For example, the dynamic F_0 range of a male voice is much narrower (90-180 Hz) than that of a female voice (150-240 Hz). Therefore, for independent-speaker tone recognition that uses the relative F_0 of each utterance as the main discriminative feature, a normalization procedure is needed to align the range of the F_0 level for different speakers. In this thesis, all parameters are normalized by the mean of the F_0 profile to reduce variation on speakers.

2.3.4 Neural network classifier

To evaluate each tone feature set, a feedforward neural network (multi-layer perceptron) is used. The network has an input layer of several units depending on the number of tone feature parameters, a hidden layer of 10 units, and an output layer of 5 units corresponding to 5 Thai tones. The tanh function is used as the activation function in the network. Since the network learns more efficiently if the inputs are normalized to be symmetrical around 0 (Tebelskis 1995), all feature parameters are normalized to lie between -1.0 and 1.0 using the following equation:

$$normF_i = 2.0 \times (\frac{F_i - minF_i}{maxF_i - minF_i}) - 1.0 \tag{2.8}$$

where F_i is the i^{th} feature under consideration, $minF_i$ and $maxF_i$ are the minimum and maximum values of F_i, and $normF_i$ is the normalized value of F_i. $MinF_i$ and $maxF_i$ are obtained from the 5^{th} and 95^{th} percentiles of a histogram of F_i generated on the training data (Muthusamy 1993).

The network is trained by the standard back-propagation algorithm for a maximum of 2000 epochs with 0.0001 learning rate and 0.9 momentum. Initial weights are set with random values between -1.0 and 1.0. The NICO (Neural Inference COmputation) toolkit (Ström 1997b) is used to build and train the network for the following experiments.

2.4 Experiments

The objectives of this section are to evaluate our tone recognition method, to study the effect of initial consonants, vowels, and final consonants on tone recognition, and to use the experiment results as baseline for comparison with other combination methods. Three data sets are built according to initial consonants, vowels, and final consonants. Three main experiments are evaluated, i.e., dependent data experiment, mixed data experiment, and cross data experiment. The details are given in the following subsections.

2.4.1 Speech corpus

To study the effect of these phonemes on tone recognition, three speech data sets were carefully designed:

Set A: the initial consonant set

To study the effect of initial consonants on tone recognition, ten consonants were selected to build a set of hypothetical syllables. These consonants are five stop consonants (i.e., /p/, /c/, /ph/, /ch/ and /d/) and five non-stop consonants (i.e., /n/, /f/, /s/, /r/ and /w/). These ten consonants together with the vowel /aː/ and five tones were used as variables to form 50 hypothetical syllables as shown in Figure 2.3 (a).

Set B: the vowel set

To study the effect of vowels on tone recognition, six vowels were selected. There are three monophthongs (i.e., /iː/, /aː/ and /uː/) and three diphthongs (i.e., /iːa/, /ɯːa/ and /uːa/). In this set, there are 30 hypothetical syllables with the initial consonant /p/, six vowels, and five tones used as variables as shown in Figure 2.3 (b).

pāː	pàː	pâː	páː	pǎː
cāː	càː	câː	cáː	cǎː
phāː	phàː	phâː	pháː	phǎː
chāː	chàː	châː	cháː	chǎː
dāː	dàː	dâː	dáː	dǎː
nāː	nàː	nâː	náː	nǎː
fāː	fàː	fâː	fáː	fǎː
sāː	sàː	sâː	sáː	sǎː
rāː	ràː	râː	ráː	rǎː
wāː	wàː	wâː	wáː	wǎː

(a)

pī̀ː	pìː	pîː	píː	pǐː
pāː	pàː	pâː	páː	pǎː
pūː	pùː	pûː	púː	pǔː
pī̀ːa	pìːa	pîːa	píːa	pǐːa
pūːa	pùːa	pûːa	púːa	pǔːa

(b)

pī̀ːm	pìːm	pîːm	píːm	pǐːm
pī̀ːn	pìːn	pîːn	píːn	pǐːn
pī̀ːŋ	pìːŋ	pîːŋ	píːŋ	pǐːŋ
pī̀ːw	pìːw	pîːw	píːw	pǐːw
	pìːp	pîːp	píːp	
	pìːt	pîːt	píːt	
	pìːk	pîːk	píːk	
pāːm	pàːm	pâːm	páːm	pǎːm
pāːn	pàːn	pâːn	páːn	pǎːn
pāːŋ	pàːŋ	pâːŋ	páːŋ	pǎːŋ
pāːj	pàːj	pâːj	páːj	pǎːj
pāːw	pàːw	pâːw	páːw	pǎːw
	pàːp	pâːp	páːp	
	pàːt	pâːt	páːt	
	pàːk	pâːk	páːk	
pūːm	pùːm	pûːm	púːm	pǔːm
pūːn	pùːn	pûːn	púːn	pǔːn
pūːŋ	pùːŋ	pûːŋ	púːŋ	pǔːŋ
pūːj	pùːj	pûːj	púːj	pǔːj
	pùːp	pûːp	púːp	
	pùːt	pûːt	púːt	
	pùːk	pûːk	púːk	

(c)

Figure 2.3: Lists of hypothetical syllables in (a) set A, (b) set B, and (c) set C.

Set C: the final consonant set

To study the effect of final consonants on tone recognition, eight consonants were selected. There are five sonorants (i.e., /m/, /n/, /ŋ/, /j/ and /w/) and three obstruents (i.e., /p/, /t/ and /k/). In this data set, there are 92 hypothetical syllables with the initial consonant /p/, three vowels (i.e., /iː/, /aː/ and /uː/), eight final consonants and five tones used as variables as shown in Figure 2.3 (c).

In Thai language, /j/ and /w/ cannot occur with the vowels /iː/ and /uː/, respectively in the syllable final position. For syllables ending with /p/, /t/ or /k/, only the low, the fall and the high tones can be found.

The data was collected from 20 native Thai speakers (10 male and 10 female speakers), aged from 18 to 29 years (mean=20.95 and s.d.=2.80). Each speaker read all sets for a trial. Therefore, the corpus comprises 1000, 600, and 1840 one-word utterances for sets A, B, and C, respectively. The speech signals were digitized by a 16-bit A/D converter of 11 kHz.

2.4.2 Training set sampling method

The k-fold cross-validation approach was considered. The advantage of this approach is that every sample is in a test set exactly once and the variance of results is reduced as k is increased (Schneider and Moore 1997). However, the training algorithm has to be rerun from scratch k times, which means that it takes k times as much computation to make an evaluation. In all experiments below (experiment I, II, and III), we performed five-fold cross-validation approach. The original utterances were partitioned into five disjoint sets of equal size. Each set contains utterances collected from two male and two female speakers. Five training sets were then constructed by overlapping the five disjoint sets and dropping out a different one systematically. The different sets, which were dropped, were used as test sets. The experimental results below are the averages of the five test sets.

2.4.3 Experiment I : Dependent data experiment

In this experiment, training and test sets were constructed from the same data set using the above five-fold cross-validation approach. The recognition rates (%) of all tone feature sets and their standard deviations (s.d.) are shown in Table 2.3. The best

Table 2.3: Recognition rates (%) and standard deviations (s.d.) of sets A, B, and C using different tone feature sets.

Training set	Test set	TF1		TF2		TF3		Average	
		%	s.d.	%	s.d.	%	s.d.	%	s.d.
A	A	95.20	2.73	**95.40**	2.86	**95.40**	3.05	95.33	2.88
B	B	96.33	3.36	**96.83**	3.70	96.67	3.33	96.61	3.46
C	C	97.23	1.53	96.74	1.22	**97.34**	1.09	97.10	1.28
Average		96.99	1.74	96.62	1.51	**97.12**	1.40		

results for each set are printed in boldface. Tone feature set 3 yields the best recognition rates for sets A and C (95.40% and 97.34%), while tone feature set 2 yields the best recognition rates for sets A and B (95.40% and 96.83%). Considering the average recognition rates and the average standard deviations across all data sets, we found that tone feature set 3 provides the highest recognition rate (97.12%) and the lowest standard deviation (1.40%). This shows that tone feature 3 is better than the other tone features. The average results demonstrate that set C gives the best recognition rate (97.10%) and the lowest standard deviation (1.28%), while set A gives the worst recognition rate (95.33%) with 2.88% standard deviation.

The results will be discussed in several aspects. Table 2.4 shows the recognition rates and standard deviations of set A reported separately according to initial consonants. The recognition rates vary dependently on initial consonants. This indicates that there are some correlations between initial consonants and tones, and the correlations are quite different for each consonant. As can be seen in Figure 2.4, the shapes of F_0 contours of all sets are quite similar. However, when we plotted the F_0 contours separately according to initial consonants, vowels, and final consonants (see Figure 2.5), we found some difference in the level and shape of the F_0 contours as described in the following. Figure 2.5 (a) shows the average *normalized* F_0 contours of the five Thai tones of set A. For each tone, F_0 contours are plotted separately according to initial consonants. The contours are different at the beginning points. This is the effect of initial consonants that makes the worst results for this set. Also there are some differences in the level of the ending part of F_0 in the fall, the high, and the rise. However, these are not the effect of initial consonants but these are the errors from the interpolation in tone feature extraction.

Table 2.5 shows the recognition rates and standard deviations of set B reported separately according to vowels. It is clear that, the recognition rates of monophthongs

Table 2.4: Recognition rates (%) and standard deviations (s.d.) of set A reported separately according to consonant and tone feature sets.

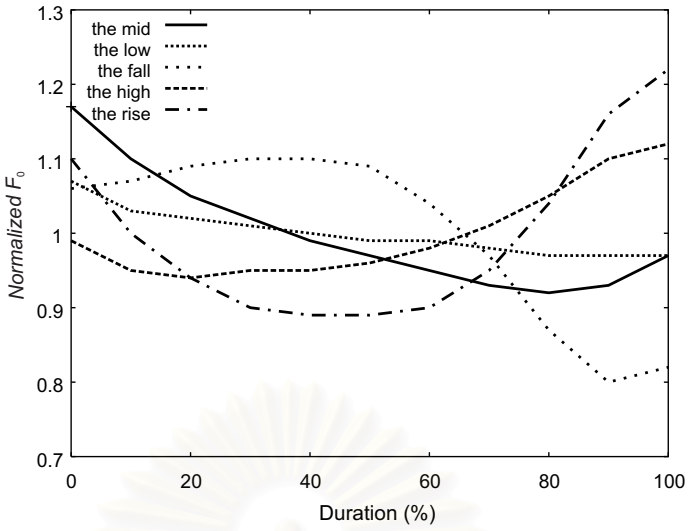
Consonant	TF1		TF2		TF3		Average	
	%	s.d.	%	s.d.	%	s.d.	%	s.d.
/p/	96.00	4.18	96.00	4.18	94.00	6.52	95.33	4.96
/c/	95.00	3.54	95.00	3.54	95.00	3.54	95.00	3.54
/ph/	95.00	8.66	96.00	6.52	96.00	6.52	95.67	7.23
/ch/	98.00	2.74	97.00	4.47	97.00	4.47	97.33	3.89
/d/	93.00	5.70	93.00	7.58	93.00	5.70	93.00	6.33
/n/	96.00	4.18	98.00	4.47	98.00	4.47	97.33	4.37
/f/	94.00	5.48	98.00	4.47	97.00	4.47	96.33	4.81
/s/	97.00	4.47	94.00	5.48	95.00	5.00	95.33	4.98
/r/	94.00	4.18	91.00	4.18	93.00	4.47	92.67	4.28
/w/	94.00	4.18	96.00	4.18	96.00	4.18	95.33	4.18

Table 2.5: Recognition rates (%) and standard deviations (s.d.) of set B reported separately according to vowel and tone feature sets.

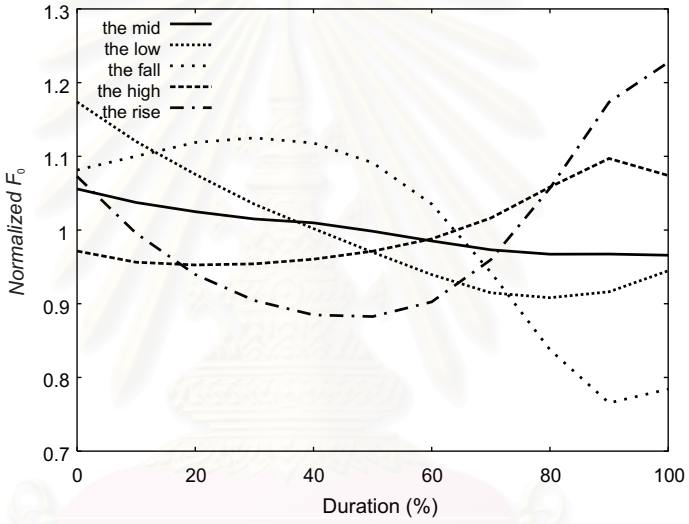
Vowel	TF1		TF2		TF3		Average	
	%	s.d.	%	s.d.	%	s.d.	%	s.d.
/iː/	97.00	2.74	97.00	2.74	97.00	4.47	97.00	3.32
/aː/	96.00	4.18	99.00	2.24	98.00	2.74	97.67	3.05
/uː/	99.00	2.24	98.00	2.74	98.00	2.74	98.33	2.57
/iːa/	95.00	5.00	95.00	7.07	94.00	8.22	94.67	6.76
/ɯːa/	96.00	8.94	97.00	4.47	96.00	6.52	96.33	6.64
/uːa/	95.00	6.12	95.00	6.12	97.00	4.47	95.67	5.57

Table 2.6: Recognition rates (%) and standard deviations (s.d.) of set C reported separately according to final consonant and tone feature sets.

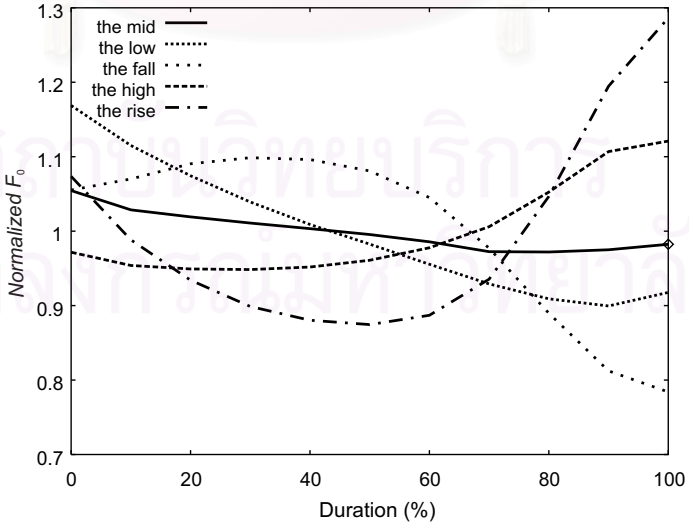
Final	TF1		TF2		TF3		Average	
consonant	%	s.d.	%	s.d.	%	s.d.	%	s.d.
/m/	95.33	2.17	97.00	2.98	97.00	1.83	96.44	2.33
/n/	98.00	1.39	98.33	2.36	98.33	1.67	98.22	1.81
/ŋ/	98.67	1.39	98.33	2.89	98.67	1.83	98.56	2.04
/j/	96.00	1.37	97.50	1.77	96.50	2.24	96.67	1.79
/w/	97.00	2.09	96.50	4.18	97.00	3.26	96.83	3.18
/p/	93.89	4.56	94.44	4.39	95.00	4.12	94.44	4.36
/t/	96.11	3.17	96.67	3.04	96.67	3.62	96.48	3.28
/k/	97.78	1.24	97.78	2.32	98.33	1.52	97.96	1.69



(a)



(b)



(c)

Figure 2.4: Average *normalized* F_0 contours of the five Thai tones of (a) set A, (b) set B, and (c) set C.

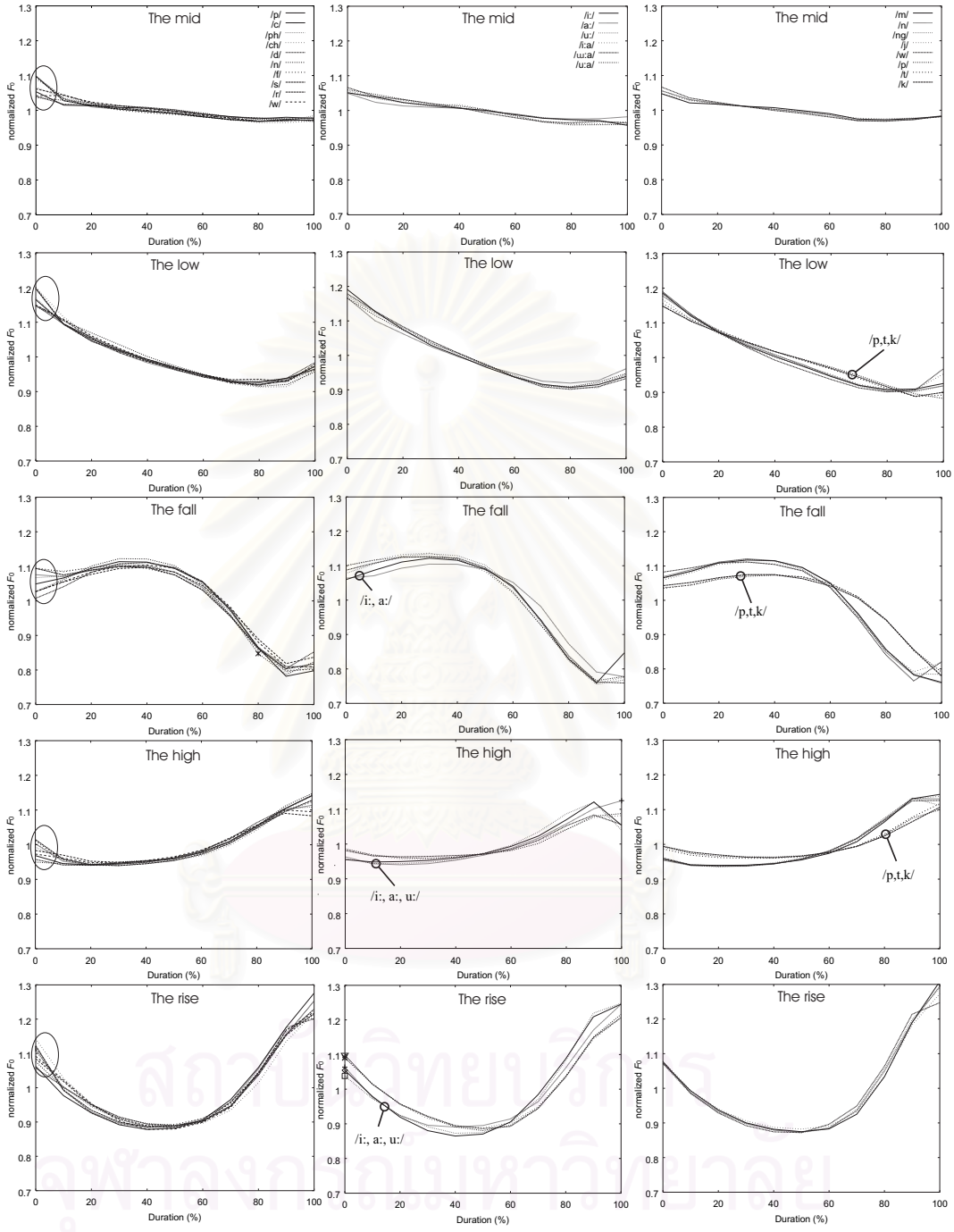


Figure 2.5: Average *normalized* F_0 contours of the five Thai tones of (a) set A, (b) set B, and (c) set C. For each tone, the F_0 are plotted separately according to initial consonants, vowels, and final consonants, respectively.

Table 2.7: Confusion matrices of tone recognition for (a) set A, (b) set B, and (c) set C using different tone feature sets. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively.

Reference	#tokens	Recognition rates (%)			Recognition results (#tokens)														
		TF1	TF2	TF3	M			L			F			H			R		
					TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3
M	200	94.00	94.50	95.00	**188**	**189**	**190**	11	11	9	1	0	1	0	0	0	0	0	0
L	200	89.50	90.50	90.00	19	17	16	**179**	**181**	**180**	0	0	0	0	0	0	2	2	4
F	200	99.50	99.50	99.50	1	1	1	0	0	0	**199**	**199**	**199**	0	0	0	0	0	0
H	200	99.00	96.50	98.00	0	5	1	0	0	0	0	0	0	**198**	**193**	**196**	2	2	3
R	200	94.00	96.00	94.50	0	0	0	2	0	0	0	0	0	10	8	11	**188**	**192**	**189**
Total	1000	95.20	95.40	95.40															

(a)

Reference	#tokens	Recognition rates (%)			Recognition results (#tokens)														
		TF1	TF2	TF3	M			L			F			H			R		
					TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3
M	120	91.67	94.17	96.67	**110**	**113**	**116**	8	6	3	0	0	0	2	1	1	0	0	0
L	120	96.67	96.67	92.50	3	2	7	**116**	**116**	**111**	0	0	0	0	0	0	1	2	2
F	120	99.17	98.33	98.33	0	0	1	1	2	1	**119**	**118**	**118**	0	0	0	0	0	0
H	120	95.00	95.00	95.83	2	2	2	3	3	1	0	0	0	**114**	**114**	**115**	1	1	2
R	120	99.17	100.00	100.00	0	0	0	0	0	0	0	0	0	1	0	0	**119**	**120**	**120**
Total	600	96.33	96.83	96.67															

(b)

Reference	#tokens	Recognition rates (%)			Recognition results (#tokens)														
		TF1	TF2	TF3	M			L			F			H			R		
					TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3	TF1	TF2	TF3
M	260	93.46	91.92	93.85	**243**	**239**	**244**	8	11	9	2	5	2	5	5	5	2	0	0
L	440	96.14	96.36	96.36	11	13	11	**423**	**424**	**424**	5	0	3	0	0	0	1	3	2
F	440	98.64	97.50	98.41	4	9	5	2	2	2	**434**	**429**	**433**	0	0	0	0	0	0
H	440	98.18	98.18	98.64	4	5	3	0	0	0	0	0	0	**432**	**432**	**434**	4	3	3
R	260	98.85	98.46	98.46	0	0	0	0	0	0	0	0	0	3	4	4	**257**	**256**	**256**
Total	1840	97.23	96.74	97.34															

(c)

(97.00%, 97.67%, and 98.33% in average for sets A, B, and C, respectively) are higher than those of diphthongs (94.67%, 96.33%, and 95.67% in average for sets A, B, and C, respectively). A diphthong is the sequence of two monophthongs with changing quality, and most parts of voiced portion represent the vowel area. This implies that vowels affect the F_0 contours and tone recognition. As shown in Figure 2.5 (b), the *normalized* F_0 contours between monophthongs and diphthongs are slightly different and separated from each other, especially for the fall, the high, and the rise. This also indicates the effect of vowels on the F_0 contour.

Table 2.6 shows the recognition rates and standard deviations of set C reported separately according to final consonants. Only three tones, i.e., the mid, the fall, and the high, are found for a syllable ending with an obstruent. In each experiment, utterances with final nasals (i.e., /n/, /m/ and /ŋ/), especially /ŋ/, get the highest recognition rates. The worst recognition rate is reported for final /p/. We can conclude that the recognition rates of most syllables with sonorant endings are better than those of obstruent endings, except final /k/, for all tone feature sets. Figure 2.5 (c) shows the average *normalized* F_0 contours of the five Thai tones of set C. For the low, the fall, and the high, the shapes of F_0 contours of syllables with sonorants differ remarkably from those with obstruents. The difference in the shape of F_0 contours at the end point can be taken into account in dividing the final consonants into the obstruents and the sonorants. The F_0 movements of sonorants change in direction at the end while those of obstruents do not. This is because the duration of a syllable ending with an obstruent is much shorter, and the energy contour sharply decreases. These two features may help to increase the recognition rate of Thai tone recognition. We plan to use them in the future. There are, at least, two more factors that impact the recognition results. Firstly, since the number of samples ending with obstruents is lower than those with sonorants, the classifier is probably biased. Secondly, in our experiments, a sample ending with a sononant or an obstruent was not separately trained and evaluated, and thus the classifier may provide errors by answering the mid or the rise for a sample ending with an obstruent.

In conclusion, the recognition rates of set A are lower than those of sets B and C (see Table 2.3), although the F_0 contours for each tone within set A are less different than those within sets B and C (see Figure 2.5). It is not claimed that initial consonants

have more effect on tones than vowels and final consonants do. But it is clear from our experiments that our tone feature sets are more sensitive to the effect of initial consonants than those of vowels and final consonants.

Table 2.7 shows confusion matrices of tone recognition for sets A, B, and C using different tone feature sets. The tone references and recognition results of each tone are represented in rows and columns, respectively. The correct results are printed in boldface. It can be seen that the fall provides the highest recognition rate for set A, and the rise gives the highest recognition rate for sets B and C. Through detailed error analysis, we found that the most errors come from the misclassification between the mid and the low, because all of them have very close F_0 levels as shown in Figure 2.5. Tone feature set 3 slightly reduces the errors. Moreover, some F_0 contours of the high and the rise share this similarity as shown by the misclassification between them.

2.4.4 Experiment II : Mixed data experiment

In this experiment, we used a mixed data set that combines data sets A, B, and C together. The results of this experiment are shown in Table 2.8. The best results are printed in boldface. It can be seen that tone feature set 3 provides the highest recognition rate (96.25%) and the lowest standard deviation (1.10%). Comparing these results with the average results across all data sets in Table 2.3, we found that the degradation in recognition performances are 1.47%, 0.98%, and 0.87% for tone feature sets 1, 2, and 3, respectively. Among these three feature sets, tone feature set 3 is the best as it provides the lowest degradation. The reason for the degradation can be given that initial consonants, vowels, and final consonants affect F_0 contours. The effect makes the contours of every tone slightly different, and thus decreases the recognition rates.

2.4.5 Experiment III : Cross data experiment

We applied the cross data method that uses one data set (set A, B or C) as a training set and the others as test sets. The results of this experiment are shown in Table 2.9. The best results are printed in boldface. Tone feature set 3 still yields the highest recognition rate (94.29% in average).

In Figure 2.6, let X/Y denote the experiment where set X is used as the training

Table 2.8: Recognition rates (%) and standard deviations (s.d.) of the mixed data experiment using different tone feature sets.

Training set	Test set	TF1		TF2		TF3		Average	
		%	s.d.	%	s.d.	%	s.d.	%	s.d.
A+B+C	A+B+C	95.52	1.38	95.64	1.53	**96.25**	1.10	95.80	1.34

Table 2.9: Recognition rates (%) and standard deviations (s.d.) of the cross data experiment using different tone feature sets.

Training set	Test set	TF1		TF2		TF3		Average	
		%	s.d.	%	s.d.	%	s.d.	%	s.d.
A	B	96.33	4.07	96.33	2.47	**97.17**	3.31	96.61	3.28
	C	94.89	1.71	94.84	1.75	**95.27**	1.69	95.00	1.72
B	A	93.80	2.66	93.50	1.84	**94.40**	1.64	93.90	2.05
	C	92.83	2.33	92.39	2.96	**93.04**	2.75	92.75	2.68
C	A	**94.80**	4.44	93.60	4.52	**94.80**	3.70	94.40	4.22
	B	94.33	3.30	**94.83**	2.73	94.50	3.31	94.55	3.11
Average		93.98	2.25	93.72	2.45	94.29	2.32		

set while set Y is used as the test set. The recognition rates of A/B and B/A are higher than those of A/C and B/C, respectively. This means that the effects of initial consonants (set A) and vowels (set B) on tones are more similar than those of initial consonants (set A) and final consonants (set C), and more similar than those of vowels (set B) and final consonants (set C). The performances of A/B and C/B are better than those of B/A and B/C, respectively. The reason is that the number of samples in set B is lower than those of the other sets, and thus the variations of tone feature parameters in set B are not enough to cover those in the other sets. The best recognition rate is reported when set A is used as the training set and the worst is reported when set B is used, respectively.

2.4.6 Discussion

Figure 2.7 shows the comparison of the average recognition rates of the three main experiments with different tone feature sets. The average recognition rate of the dependent data experiment is the best for every tone feature set. The average recognition rate of the mixed data experiment is better than that of the cross data experiment for each tone feature set. Tone feature set 3 provides the highest recognition rate for all experiments. In the dependent data and cross data experiments, the recognition rates for tone feature set 1 are better than those for tone feature set 2.

As shown in these experiments, although our feature sets are useful for the dependent data sets (the data sets of the dependent data experiment and the mixed data experiment), they are not robust for independent data set (the data set of the cross data

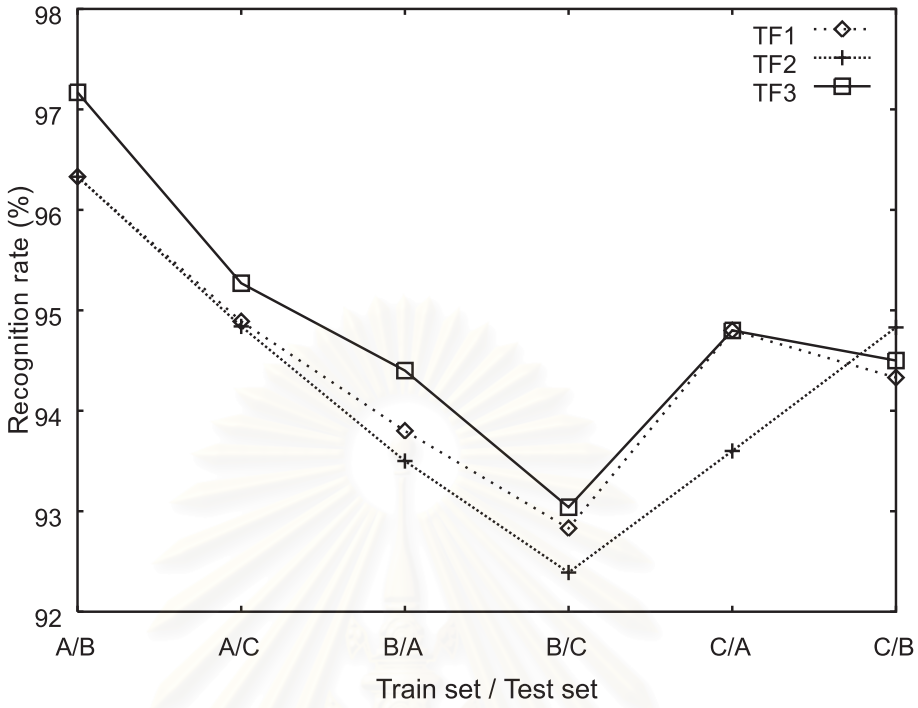


Figure 2.6: Recognition rates of the cross data experiment with different training and test sets using different tone feature sets.

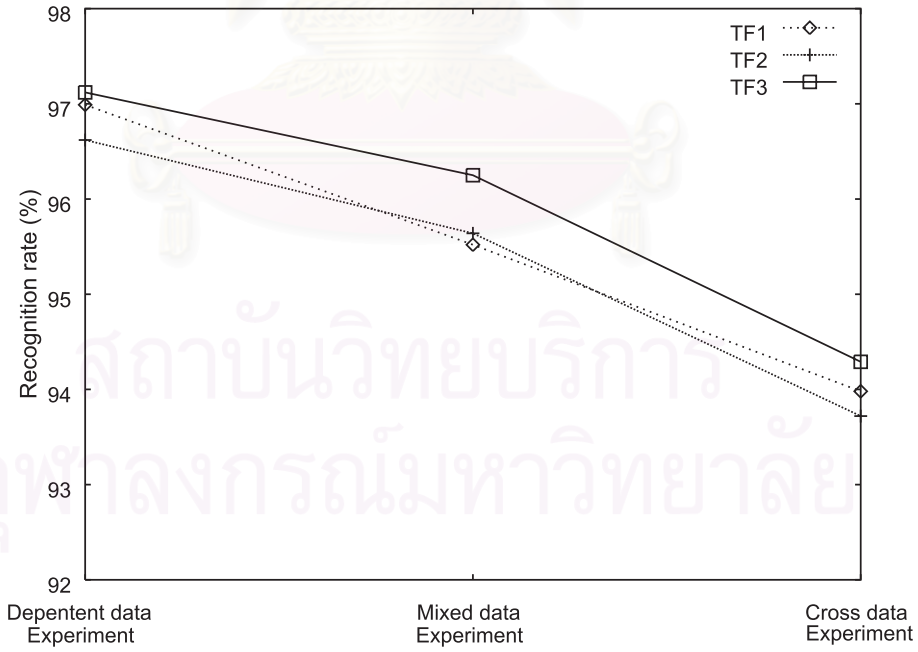


Figure 2.7: Average recognition rates of three main experiments using different tone feature sets.

experiment). Therefore, tone recognition for all available syllables needs the training examples that cover all combinations between tones and the phonematic units composing the syllables, and the size of the training set must be large enough.

2.4.7 Human perception test

In order to judge the above recognition rates on the basis of human perception, a listening test has been designed. Only 50 random utterances from all data sets were used so that the participants would not experience too much inattention. The test was taken by 20 participants: 8 male and 12 female listeners. Each utterance was played once, with a 5 second interval between the current and the next utterances. Each listener identified and entered one of the five choices (the mid, the low, the rise, the high, or the fall) before the computer proceeded to the next utterance.

Table 2.10 shows the total responses in a confusion matrix. The average perception rate and standard deviation of tone identification by individual listeners are 88.00% and 9.40%, respectively. The best perception rate is 95.00% for the rise, and the worst is 76.00% for the low. Like the machine, most listeners often confuse the low with the mid. The human's perception rate is much lower than that of the machine. This may suggest that humans are not good at recognizing meaningless words, or words without the context (words in isolation). Moreover, the basic unit that humans recognize well is a word (not phoneme), and therefore the human perception rate in our experiment is lower than the machine. On the other hand, a question is raised. If humans are not as good as machine in tone recognition but have much higher performance in speech recognition, is the improved performance on tone recognition going to be translated into improved performance in speech recognition? In fact, human uses many other linguistic cues, e.g., semantic, syntactic, pragmatic and morphology to recognize speech in human communication. The improvement of automatic speech recognition needs a good acoustic model for its bottom up speech recognition process and also a good linguistic model for its top down speech recognition process. Tone recognition is essential to word recognition because it recognizes lexical information that the tones carry. Tone recognition helps to extremely reduce the number of referential templates for recognizing speech. Thus, we believe that tone recognition is still very important and helpful to improve the performance of speech recognition in tone languages.

Table 2.10: Confusion matrices of human perception test.

Reference	#Tokens	Perception rate (%)	Perception results (#tokens)				
			M	L	F	H	R
M	180	90.00	**162**	13	1	2	2
L	200	76.00	34	**152**	10	1	3
F	220	90.00	11	2	**198**	9	0
H	220	89.56	2	4	10	**197**	7
R	180	95.00	0	1	2	6	**171**
Total	1000	88.00					

2.5 Combination of Neural Networks

Recently, the classifier combination approach has been repeatedly proven to be more robust than the single classifier approach (Kirchhoff and Bilmes 1999). The basic idea is to classify an input pattern by obtaining classification from several classifiers and then using a consensus scheme to decide the collective classification by vote (Hansen and Salamon 1990). There are basically two classifier combination scenarios (Kittler et al. 1998): (i) all classifiers use the same representation of the input pattern and (ii) each classifier uses its own representation of the input pattern. In this section, we study the usefulness of a combination approach of neural networks for isolated Thai tone recognition and focus on classifier combination in the second scenario only.

2.5.1 Method

Figure 2.8 shows the architecture of the neural network combination. Three neural networks (NNs) are employed to decide the final classification. Each NN uses a different tone feature set as the representation of the input pattern. The outputs of each network are fed to the combination module for deciding a final result. Several combination schemes have been proposed (Kittler et al. 1998). In this thesis, we apply three groups of combination schemes to tone recognition and consider the benefit of each combination scheme to this problem. These groups are described as follows.

Probability combination rules (PCRs)

Kittler et al. (1998) proposed various probability combination rules, i.e., product rule, sum rule, max rule, and min rule. The necessary notation and formulations of them are introduced in the following.

Consider a recognition problem where pattern Z is to be assigned to one of the M

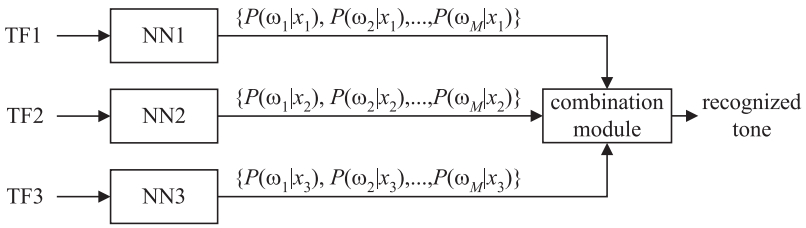


Figure 2.8: The neural network combination.

possible classes $(\omega_1, \ldots, \omega_M)$. Assume that we have N classifiers, and let x_i and $P(\omega_k|x_i)$ be the input representation used by the i^{th} classifier, and the *posteriori* probability for the k^{th} class given the input representation, respectively. Under the equal prior assumption, the pattern Z will be assigned to class ω_k^* using one of the following rules:

Product rule:

$$P(\omega_{k^*}|x_i) = \max_{k=1}^{M} \prod_{i=1}^{N} P(\omega_k|x_i) \tag{2.9}$$

Sum rule:

$$P(\omega_{k^*}|x_i) = \max_{k=1}^{M} \sum_{i=1}^{N} P(\omega_k|x_i) \tag{2.10}$$

Max rule:

$$P(\omega_{k^*}|x_i) = \max_{k=1}^{M} \max_{i=1}^{N} P(\omega_k|x_i) \tag{2.11}$$

Min rule:

$$P(\omega_{k^*}|x_i) = \max_{k=1}^{M} \min_{i=1}^{N} P(\omega_k|x_i) \tag{2.12}$$

Voting techniques (VTs)

The method based on voting techniques considers the result of each network as an expert judgement (Cho 1997). A variety of voting procedures can be adopted from *group decision marking theory*, e.g., unanimity, majority, plurality, borda count, and so on. In our experiments, we used two of them, i.e., the majority voting and the borda count.

Majority voting:

$$k^* = arg \max_{k=1}^{M} \sum_{i=1}^{N} \Delta_{ki} \tag{2.13}$$

where

$$\Delta_{ki} = \begin{cases} 1 & \text{if } P(\omega_k|x_i) = \max_{j=1}^{M} P(\omega_j|x_i) \\ 0 & \text{otherwise} \end{cases} \tag{2.14}$$

Borda count:

$$k^* = arg \max_{k=1}^{M} \sum_{i=1}^{N} B_{ki} \tag{2.15}$$

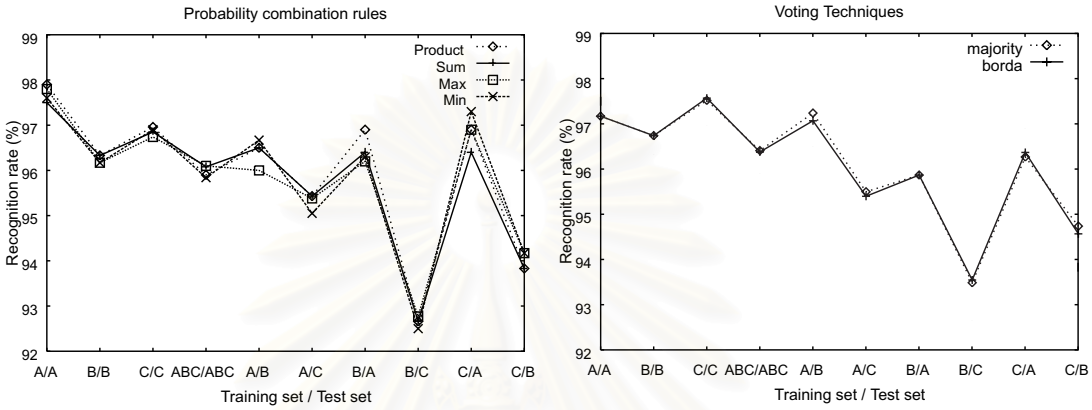
where B_{ki} is the number of classes ranked below the class k by the i^{th} classifier.

Non-linear combination method (NCM)

The outputs (*posteriori* probabilities) of three neural networks are used as inputs to a new neural network, which non-linearly combines the three networks. The new network is employed for evaluating the final classification. In this thesis, a feedforward neural network is used as the non-linear classifier. The network has an input layer of 15 units (corresponding to the outputs of all three input networks), a hidden layer of 30 units, and an output layer of 5 units. The network is trained by the standard back-propagation algorithm for a maximum of 300 epochs with 0.00001 learning rate and 0.9 momentum.

2.5.2 Results and discussion

Figure 2.9 shows the experimental results using probability combination rules (PCRs) and voting techniques (VTs). All best results for each experiment are printed in boldface. Considering the recognition rates of PCRs, we found that the recognition rates of all rules are not significantly different for each experiment (see Figure 2.9 (a)). However, it appears that the product rule frequently scores a best result. The average results of all experiments are 95.42%, 95.40%, 95.41%, and 95.27% for product, sum, max, and min rules, respectively. Although the product rule provides the highest recognition rates, this rule is sensitive to badly estimated *posteriori* probabilities (Kittler et al. 1998). If any of the classifiers reports the correct class a *posteriori* probability as zero, the output will be zero and the correct class cannot be identified. Instead a more robust mean, the sum rule is expected to work better. This rule is not very sensitive to very poor estimates (Duin and Tax 2000). For VTs (see Figure 2.9 (b)), these results are also not significantly different for each experiment. The majority voting often provides a best result. The average recognition performances are 95.76% and 95.75% for



(a) (b)

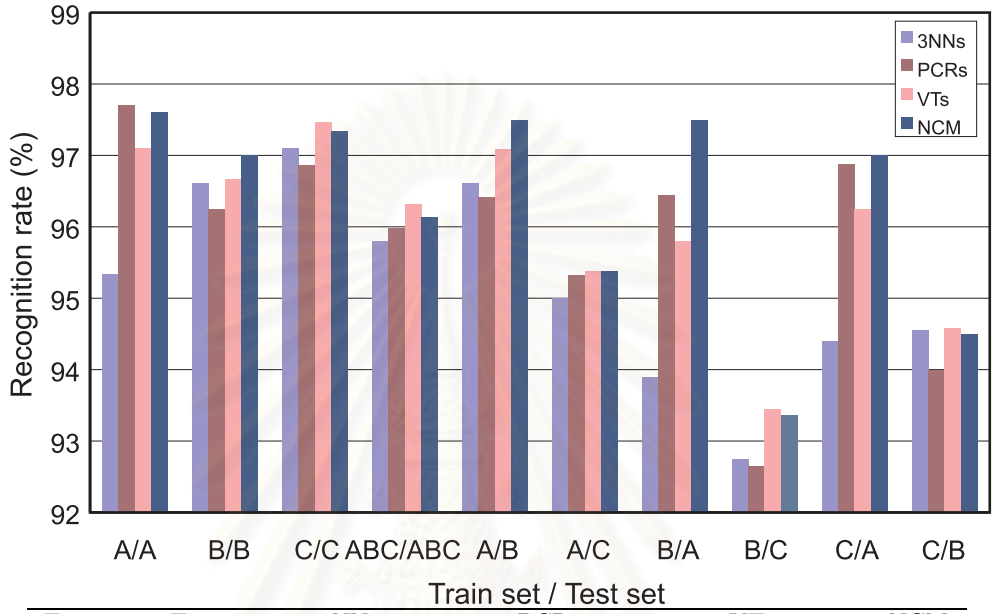
Tr./Te.	Product	Sum	Max	Min
A/A	**97.90**	97.50	97.80	97.60
B/B	**96.33**	**96.33**	96.17	96.17
C/C	**96.96**	96.85	96.74	96.90
Average	**97.01**	96.88	96.81	96.93
ABC/ABC	95.90	96.08	**96.10**	95.84
A/B	96.50	96.50	96.00	**96.67**
A/C	**95.43**	**95.43**	95.38	95.05
B/A	**96.90**	96.40	96.20	96.30
B/C	92.66	92.66	**92.77**	92.50
C/A	96.90	96.40	96.90	**97.30**
C/B	93.83	93.83	**94.17**	**94.17**
Average	**94.38**	94.33	94.36	94.15
Total avg.	**95.42**	95.40	95.41	95.27

(a)

Tr./Te.	Majority	Borda
A/A	**97.10**	**97.10**
B/B	**96.67**	**96.67**
C/C	97.45	**97.50**
Average	97.37	**97.42**
ABC/ABC	**96.34**	96.31
A/B	**97.17**	97.00
A/C	**95.43**	95.33
B/A	**95.80**	**95.80**
B/C	93.42	**93.48**
C/A	96.20	**96.30**
C/B	**94.67**	94.50
Average	**94.66**	94.63
Total avg.	**95.76**	95.75

(b)

Figure 2.9: Recognition rates (%) and standard deviations (s.d.) of ten experiments with different training and test sets using (a) probability combination rules and (b) voting techniques.



Training set	Test set	3NNs %	3NNs s.d.	PCRs %	PCRs s.d.	VTs %	VTs s.d.	NCM %
A	A	95.33	0.12	**97.70**	0.18	97.10	0.00	97.60
B	B	96.61	0.25	96.25	0.10	96.67	0.00	**97.00**
C	C	97.10	0.32	96.86	0.09	**97.47**	0.04	97.34
Average		96.91	0.26	96.91	0.09	**97.39**	0.03	97.34
ABC	ABC	95.80	0.39	95.98	0.13	**96.32**	0.02	96.13
A	B	96.61	0.48	96.42	0.29	97.08	0.12	**97.50**
A	C	95.00	0.24	95.33	0.18	**95.38**	0.08	**95.38**
B	A	93.90	0.46	96.45	0.31	95.80	0.00	**97.50**
B	C	92.75	0.33	92.65	0.11	**93.45**	0.04	93.37
C	A	94.40	0.69	96.88	0.37	96.25	0.07	**97.00**
C	B	94.56	0.25	94.00	0.19	**94.58**	0.12	94.50
Average		94.00	0.29	94.30	0.11	94.64	0.02	**94.73**
Total average		95.18	0.33	95.37	0.14	**95.75**	0.04	95.73

Figure 2.10: Recognition rates (%) and standard deviations (s.d.) of ten experiments with different training and test sets using different combination schemes.

majority voting and borda count, respectively.

Figure 2.10 shows the average results of three neural networks (3NNs), the average results of PCRs, the average results of VTs and the results of the non-linear combination method (NCM), with their standard deviations. All best results for each experiment are printed in boldface. The results of 3NNs are used as the baseline results. The recognition performances of most experiments are improved when the combination schemes are applied. This is not surprising because the combination schemes using different input representations almost always provide less correlation between the input vectors than the single classifier using an input representation. Although some feature set yields poor results, it still contains valuable information for the combination classifiers.

All three combination schemes can be divided into two groups according to the format of the individual classifiers used by the combiner (Kittler et al. 1998). Hard-level combination uses the output of the classifier after it is hard-thresholded (binarized). Soft-level combination uses the estimates of a *posteriori* probability of the class by each classifier. VT is a representative of the first category while the others are the soft-level combiners. Considering the average performances of all combination schemes, we can see that NCM and VT perform equally well and PCR provides the worst results. However, there is no overall winning combination scheme. VTs frequently give a best result for the dependent data and mixed data experiments whereas PCR often yields a best result for the cross data experiment. Compared to the baseline (3NNs), VT gives the best error reduction rates in average of 15.63% and 12.36% for the dependent data and mixed data experiments, respectively, while NCM provides the best error reduction rate in average of 12.24% for the cross data experiment. VT does not take into account the differences in the individual classifier capabilities. All classifiers are treated equally, which may not be preferable when we know that certain classifiers are more likely to be correct than others (Ho et al. 1994). In the opposite way, NCM takes into account the differences in the individual classifier capabilities (*posteriori* probabilities). However, in our experiments, the average recognition rates of all experiments for VT and NCM are not significantly different. One advantage of VT over NCM is that it needs no training whereas NCM requires training.

Considering the results of each experiment, we can see that when set A is used as test set, the recognition rates are much improved for all combination schemes. The

results of set A become the best although those of set A using the single classifier are the worst. When set B is used as test set, the recognition results of PCR are not improved, while those of VT are slightly increased. For NCM, the recognition performances are improved only for B/B and A/B. When set C is used as test set, VT and NCM provide the improvement in all experiments, while PCR improves recognition results in A/C only.

2.6 Summary

In this chapter, a method of isolated Thai tone recognition has been proposed. Three tone feature sets were used to capture the characteristics of Thai tones. The first two tone feature sets come from the previous works (Thubthong 1995; Tungthangthum 1998) and the last one is a novel tone feature set designed from our observation on the F_0 contour patterns. To evaluate the performance of our method and to study the effect of initial consonants, vowels, and final consonants on tone recognition, three data sets were built. Each data set was used to study the effect of each phoneme. Several experiments have been conducted using feedforward neural networks. The proposed tone feature set yielded the best performance for most experiments. The experimental results imply that there are some correlations between tones and the phonematic units constructing the syllables. Therefore, a tone recognition system for all available syllables needs the training examples that cover all combinations between tones and the other syllable types, and the size of the training set must be large enough. Human perception test was then employed to judge the recognition rate. The recognition rate of human perception test was much lower than that of the machine. This suggests that humans are not good at recognizing meaningless words, and words without context. The basic unit for human recognition is a word, not a phoneme. The combination of neural networks trained on different tone feature sets was studied. Several classifier combination schemes, i.e., PCR, VT, and NCM, were used to enhance the recognition rate. The experimental results demonstrated that the neural network combination was superior to a single network, and NCM and VT performed equally well whereas PCR did worst.

CHAPTER 3

CONSTRUCTING TONE RECOGNITION FRAMEWORK FOR THAI CONTINUOUS SPEECH

In the previous chapter, we proposed a tone feature set and demonstrated a series of experiments by considering the syllable structure effect. All speech data used were isolated syllables. Before studying the other effects, we will first concentrate on an issue of constructing a basic tone recognition framework. The framework consists of tone models used to parameterize tone F_0 contour and a classifier used to evaluate the performance of the tone models. The former is designed by an empirical study, while, for the latter, a three-layer feedforward neural network is applied. The framework will be used for the following chapters.

There are a number of studies of Thai tone recognition in isolation (Kongkachandra et al. 1998; Thubthong et al. 2000a; Thubthong and Kijsirikul 2000b; Thubthong and Kijsirikul 2001b; Tungthangthum 1998) and in continuous speech (Potisuk et al. 1999; Thubthong et al. 2000; Thubthong et al. 2000b; Thubthong and Kijsirikul 2001a; Thubthong and Kijsirikul 2001c). All studies concentrated on tone features and classifier techniques. However, there are other issues that should be considered. In this chapter, we are concerned with four questions: (i) which tone features are useful for tone recognition in continuous speech in term of performance, (ii) what kind of frequency scale should be used in order to provide the highest tone recognition rate, (iii) what is the effect of normalization and which normalization techniques should be used, and (iv) which part of the F_0 contour in a syllable should be used for tone recognition. We perform an empirical study of Thai tone recognition to answer these questions. The study contains four experiments each of which is performed for answering each question. The answers will be used as configurations for constructing the simple tone models (Thubthong and Kijsirikul 2002).

In the following sections, we first present three Thai speech corpora used in our experiments: Potisuk-1999 corpus, which consists of 11 linguistically designed sentences that are continuously voiced throughout; Thai Proverb Corpus, which contains 30 Thai proverb utterances (i.e., 10 four syllabic, 10 five syllabic, and 10 six syllabic utterances); and Thai Animal Stories Corpus, which is composed of 50 Thai sentences produced in

reading style. Then, we present the literature review of classification techniques for tone recognition and explain the reasons for using neural networks as the classifiers in our framework. We also describe the experimental setting for evaluating the above issues. After that we demonstrate four experiments for answering the four questions above. Finally, we summarize this chapter.

3.1 Thai Speech Corpora

Three Thai speech corpora of different complexities are used in our experiments: Potisuk-1999, Thai proverb, and Thai animal story corpora. The first one is a linguistically designed corpus, the second one is a short sentence corpus, and the last one is a read speech corpus. The corpora are used to evaluate the configurations for constructing Thai tone modelling. The detail of the corpora will be described in the following.

(1) Potisuk-1999 corpus (PC-99)

A corpus was built using the sentence list designed by (Potisuk et al. 1999). The list contained 11 sentences with varying tone sequences. Each sentence consisted of four monosyllabic words. To enhance coarticulatory effect, all four syllables began and ended with a sonorant, and the sentence was continuously voiced throughout. In order to eliminate the potential interaction between stress and tone, the stress pattern of the carrier sentence (strong strong strong strong) was invariant. The data was collected from 10 native Thai speakers (five male and five female speakers), ranging in age from 20 to 22 years (mean=20.8 and s.d.=0.78). Each speaker read all sentences for five trials at a conversational speaking rate. Therefore, the corpus contained 550 utterances (2,200 syllables). We named this corpus "Potisuk-1999 corpus".

(2) Thai proverb corpus (TPC)

The Thai proverb corpus contained 30 Thai proverbs: 10 four-syllabic, 10 five-syllabic and 10 six-syllabic proverbs. The data was collected from 40 native Thai speakers (20 male and 20 female speakers), ranging in age from 17 to 29 years (mean=20.78 and s.d.=2.35). Each speaker read all 30 proverbs in one trial at a conversational speaking rate. Therefore, the corpus consists of 1,200 utterances (6,000 syllables).

(3) Thai animal story corpus (TASC)

The Thai animal story corpus is a read speech corpus containing four animal stories (i.e., cat, monkey, elephant and buffalo). The corpus consisted of 50 different sentences. The data was collected from 20 native Thai speakers (10 male and 10 female speakers), ranging in age from 20 to 30 years (mean=22.20 and s.d.=2.01). Each speaker read all 50 sentences in one trial at a conversational speaking rate. Therefore, the corpus consisted of 1,000 sentence utterances (5,760 syllables).

The speech signals were digitized by a 16-bit A/D converter at 11 kHz. These were manually segmented and transcribed at syllable and onset-rhyme levels using audio-visual cues from a waveform display.

3.2 Classifiers

Many methods of tone recognition have been proposed for both isolated and continuous speech in Mandarin, Cantonese and Thai. They include the methods based on neural networks (NNs) for four-tone-recognition of isolated Mandarin syllables (Chang et al. 1990), for five-tone-recognition of continuous Mandarin speech (Wang and Chen 1994; Chen and Wang 1995) and for nine-tone-recognition of isolated Cantonese syllables (Lee et al. 1993; Lee et al. 1995).

There are also a number of works based on a hidden Markov model (HMM) for four-tone-recognition of isolated Mandarin syllables (Yang et al. 1988), for five-tone-recognition of continuous Mandarin speech (Cao et al. 2000; Huang and Seide 2000), and five-tone-recognition of isolated Thai syllables (Tungthangthum 1998). The fuzzy C-means-based method for four-tone-recognition of isolated Mandarin syllables (Li et al. 1999) and support vector machines (SVMs) for five-tone-recognition of isolated Thai syllables (Thubthong and Kijsirikul 2000b; Thubthong and Kijsirikul 2001b) have also been proposed. In addition, a few works based on other statistical or non-statistical classification methods have been proposed for four-tone-recognition of isolated Mandarin (Wang et al. 1990; Wu et al. 1991) and five-tone-recognition of isolated Thai (Charnvivit et al. 2001).

An HMM provides poor discrimination due to the fact that model parameters are estimated by maximum likelihood estimation instead of an estimation method that

attempts to explicitly minimize the classification error (Ström 1997a). It is not suitable for our framework because it needs a large number of features for training but the number of our tone features is quite small. In the case of the fuzzy C-means-based method, the computational model is quite complicate. An SVM is a new promising pattern classification technique. It aims to minimize the upper bound of the generalization error through maximizing the margin between the separating hyperplane and data (Vapnik 1998). The weakness of the SVM is that it is a binary classifier. Although the SVM can be adopted for solving multi-class problems and it often achieves good results, the training times are very long and the system is complicate and hard to implement. For these reasons, we decide to use neural networks as the classifiers for our tone recognition framework. An NN aims to minimize the empirical training error. The advantages of the NN are that (i) it does not require the underlying statistical distributions, (ii) it does not need a large number of features, (iii) it is a multi-class classifier, and (vi) it is easy to implement.

3.3 Experimental Setting

In the following experiments, we performed five-fold cross-validation approach (Dietterich 1997) for every corpus. For PC-99, we built two tests, i.e., inside test and outside test, to compare the recognition robustness against speaker variation. In the inside test, the data from the same speakers were used in both training and testing; while, in the outside test, the data for training and testing were from different speakers. The original utterances were partitioned into five disjoint parts of equal size. For the inside test, each part contained utterances collected from each trial of all speakers. For the outside test, each part contained utterances collected from one male and one female speakers. For TPC and TASC, we conducted only outside test. Each part of TPC consisted of utterances collected from four male and four female speakers; while each part of TASC contained utterances collected from two male and two female speakers. A summary of the corpora for each fold is shown in Table 3.1.

Every experiment was performed using a three-layer feedforward neural network. The network had three layers, i.e., input, hidden, and output layers. The number of input units depended on the number of tone features. The number of hidden and output units were 20 and 5, respectively. All feature parameters were normalized to lie

Table 3.1: Summary of all three corpora used in our experiments for each fold.

Data set	PC-99 (inside)		PC-99 (outside)		TPC		TASC	
	Training	Test	Training	Test	Training	Test	Training	Test
#Utterances	440	110	440	110	960	240	800	200
#Syllables	1760	440	1760	440	4800	1200	4608	1152
#Speakers	10	10	8	2	32	8	16	4

between -1.0 and 1.0. The network was trained using the standard back-propagation method. Initial weights were set to random values between -1.0 and 1.0. The NICO (Neural Inference COmputation) toolkit (Ström 1997b) was used to build and train the network. The experimental results were the average values of the five test sets.

3.4 Tone Features

Thubthong et al. (2000a) (see Chapter 2) have proposed a tone feature set for recognizing isolated syllables. The feature set is based on the effects of initial consonant, vowel and final consonant in a syllable on an F_0 contour. We refer to these effects as *internal effects*. In continuous speech, there are, however, many other interacting factors (e.g., co-articulation, intonation and stress) affecting an F_0 contour. We refer to them as *external effects*. The external effects affect the F_0 contour more than internal effects do. We therefore need a new accurate and convenient tone feature set. The feature set should be enhanced conveniently by other tone features in order to compensate for external effects. We built three basic tone feature sets. They were compared and the best one was selected as the basic tone feature set for the following experiments.

In order to extract tone features, we first applied the Average Magnitude Different Function (AMDF) algorithm (Ross et al. 1974) for F_0 extraction with 20 ms frame size and 5 ms frame shift. Since syllables were not of equal duration, the duration of F_0 contours of each syllable were equalized on a percentage scale (Gandour et al. 1994; Potisuk, Gandour, and Harper 1996). F_0-normalized data were then fitted with a third-order polynomial ($y = a_0 + a_1x + a_2x^2 + a_3x^3$) that has been proven to be successful for fitting F_0 contours of the five Thai tones (Gandour et al. 1999). Then three tone feature sets were built:

1. Tone feature set A

All four polynomial coefficients were used. The tone feature vector of each

syllable contained all four polynomial coefficients.

2. Tone feature set B

This feature set was obtained by extracting the F_0 heights at the beginning and end points of the *tone critical segment*[1]. Furthermore, by using the polynomial coefficients, we computed the slopes at five different time points between 0% to 100% throughout each tone critical segment with the equal step size of 25%. The two F_0 heights and five slopes were used together as tone feature set B. This feature set is similar to tone feature set 3 in Chapter 2.

3. Tone feature set C

Tone feature set C was created by enhancing tone feature set B. The F_0 height at 25%, 50% and 75% time points of the tone critical segment were also incorporated into tone feature set B.

To evaluate all three tone feature sets, PC-99 (only outside test), TPC, and TASC were used. Since we did not yet know, which the other configurations, i.e., frequency scale, normalization technique, and tone critical segment, should be used, we first explored the experiment by using the frequency scale in hertz. The tone features were extracted from the rhyme unit of a syllable without normalization.

The recognition rates (%) and standard deviations (s.d.) are shown in Table 3.2. Tone feature set C provides the best recognition rates for all corpora, while feature set A yields the worst. We performed McNemar's test (Gillick and Cox 1989) between each pair of classification outputs to examine whether the differences in tone performance are statistical significant. The McNemar significance level reflects the probability of the hypothesis that the differences between two classification results occur by chance. We set the threshold of the significance level to be 0.05 (95% confidence interval), which means that the differences are considered as statistically significant if the probability of the differences occurring due to chance is less than 0.05. As shown in Table 3.3, the differences in recognition rates between tone feature set C and the other tone feature sets are considered to be statistically significant for PC-99 and highly statistically significant for TPC and TASC. But the difference between tone feature sets A and B are not statistically significant for all corpora.

[1]Tone critical segment is the F_0 contour segment containing critical information for tone recognition (see Section 3.7).

Table 3.2: Recognition rates (%) and standard deviations (s.d.) of tone recognition with different tone features. The best results for each corpus are printed in **boldface**.

Feature set	PC-99		TPC		TASC	
	%	s.d.	%	s.d.	%	s.d.
A: 4 coef.	77.82	4.79	71.25	1.13	68.85	2.08
B: F_{0I}+5dF_0+F_{0F}	78.55	6.20	71.73	1.91	70.38	3.22
C: 5F_0+5dF_0	**79.86**	5.89	**74.67**	1.32	**74.48**	3.27

Table 3.3: Measure of statistical difference of tone recognition with different tone features. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Feature set	PC-99		TPC		TASC	
	B	C	B	C	B	C
A	0.3485	**0.0067**	0.3718	**< 0.0001**	0.0149	**< 0.0001**
B	-	**0.0385**	-	**< 0.0001**	-	**< 0.0001**

3.5 Frequency Scale

This section concerns the question of what kind of scale should be used in order to provide highest tone recognition rate. In physics, frequency is generally expressed in terms of the unit hertz (Hz). In various branches of hearing research, other units are used, e.g., semitone, ERB-rate, Mel scale, Bark scale, etc. Most researchers, in speech recognition community, have used Mel and Bark scales (Hermansky 1990; Hermansky and Morgan 1994; Hermansky 1998; Ström 1997b) for recognizing segmentals (consonants and vowels). Moreover, for suprasegmental especially intonation and tone, semitone and ERB-rate scales have been considered. In intonation studies, results of experiments have been accounted in terms of a hertz scale (Rietveld and Gussenhoven 1985), a semitone scale (Sagisaka and Kaiki 1992; Swerts et al. 1996; 't Hart 1981; Thorsen 1980) and an ERB-rate scale (Hermes and van Gestel 1991; Beaugendre et al. 2001). Most researchers have used hertz scale (Cao et al. 2000; Chen and Wang 1995; Lee et al. 1995; Thubthong et al. 2000a; Wang et al. 1997; Wang and Seneff 2000) in tone recognition for several languages, while some researchers have applied log scale (equivalent to simitone) (Huang and Seide 2000; Yang et al. 1988) and ERB-rate scale (Potisuk, Gandour, and Harper 1996; Potisuk et al. 1999). A few details of three frequency scales are described as follow:

1. **Hertz**

   Hertz (Hz) is a linear frequency scale. It is defined as the number of cycles per second (Ladefoged 1996).

2. Semitone

Semitone is a musical scale used to express the relative distance between two tones in a musical interval. It is a logarithmic frequency scale defined as follow:

$$\texttt{semitone} = 69 + 12\log_2\left|\frac{f}{440}\right| \tag{3.1}$$

where f is frequency in Hz.

3. Equivalent rectangular bandwidth rate

Equivalent rectangular bandwidth rate (ERB-rate) is a psychoacoustic scale. It has better representation of the perceived excursion size of prominence-lending pitch movements presented in different pitch register (Hermes and van Gestel 1991). As female and male voices differ in the excursion size of F_0 movements, a raw F_0 can be converted into an ERB-rate scale to normalize the excursion size between speakers. The ERB-rate scale is defined as follow (Moore and Glasberg 1983):

$$\texttt{ERB-rate} = 11.17\ln\left|\frac{f+312}{f+14675}\right| + 43.0 \tag{3.2}$$

where f is frequency in Hz.

Every frequency scale was applied to all corpora by using tone feature set C. As shown in Table 3.4, ERB-rate scale provides the best recognition rates for PC-99 and TCP, whereas semi-tone scale yields the best for TASC. However, as shown in Table 3.5, the statistically significant difference are found for the pair of hertz scale and semitone scale, and the pair of hertz scale and ERB-rate scale on TPC only. Since the average recognition rate of ERB-rate scale is slightly better those of the other scales and many researchers have successfully used this scale for prosodic studies, this scale is selected and used for the following sections.

3.6 Normalization Technique

An F_0 is basically a physiologically determined characteristic and is regarded as speaker dependent (Lee et al. 1995). Therefore, for speaker independent tone recognition that uses the relative F_0 of each syllable as the main discriminative feature, a normalization procedure is needed to account for the F_0 range of different speakers.

Table 3.4: Recognition rates (%) and standard deviations (s.d.) of tone recognition with different frequency scales. The best results for each corpus are printed in **boldface**.

Scale	PC-99		TPC		TASC	
	%	s.d.	%	s.d.	%	s.d.
Hertz	79.86	5.89	74.67	1.30	74.48	3.27
Semitone	79.86	5.69	74.83	1.26	**74.70**	2.91
ERB-rate	**80.18**	5.94	**74.93**	1.27	74.65	3.32

Table 3.5: Measure of statistical difference of tone recognition with different frequency scales. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Scale	PC-99		TPC		TASC	
	Semitone	ERB-rate	Semitone	ERB-rate	Semitone	ERB-rate
Hertz	0.9447	0.6538	**< 0.0001**	**< 0.0001**	0.4627	0.6285
Semitone	-	0.5655	-	0.8035	-	0.9132

Many researchers (Chen and Wang 1995; Lee et al. 1995; Thubthong et al. 2000a; Wang and Seneff 1998) have proposed to normalize all raw F_0's of each utterance by its own mean. Some studies (Gandour et al. 1999; Potisuk et al. 1999), however, suggest that z-score normalization is useful to account for the F_0 variation.

In this section, we conducted an experiment to compare different normalization techniques. The first one used raw F_0's without normalization. For the second one, raw F_0's of each utterance were normalized by its own mean. For the last one, raw F_0's were normalized by transforming the frequency values to a z-score values using the mean and standard deviation calculated from raw F_0's of all syllables within each speaker.

The results are shown in Table 3.6. Better recognition rates are reported when both normalization techniques are applied. The z-score technique gives the best recognition rates for all corpora. As shown in Table 3.7, all differences between each pair of recognition rates are highly significant. This means that the z-score technique is powerful to reduce F_0 variation across speakers. However, the drawback of this technique is that it needs the mean and the standard deviation of F_0's of all utterances for each speaker, which cannot be determined without having all utterances. During training process, these values can be calculated from the training examples of each speaker. However, in the testing process, we do not know in advance all raw F_0's from all utterances of each speaker, and thus we cannot determine these values. This means that z-score normalization is not convenient for a speaker-adaptation speech recognition system. However, although we cannot determine the exact values of the mean and

Table 3.6: Recognition rates (%) and standard deviations (s.d.) of tone recognition with different normalization techniques. The best results for each corpus are printed in **boldface**.

Normalization	PC-99		TPC		TASC	
	%	s.d.	%	s.d.	%	s.d.
None	80.18	5.94	74.93	1.27	74.65	3.32
Mean	84.14	5.23	82.33	1.59	80.00	2.08
Z-score	**90.05**	2.33	**84.07**	0.80	**82.48**	2.41

Table 3.7: Measure of statistical difference of tone recognition with different normalization techniques. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Normalization	PC-99		TPC		TASC	
	Mean	Z-score	Mean	Z-score	Mean	Z-score
None	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**	**< 0.0001**
Mean	-	**< 0.0001**	-	**< 0.0001**	-	**< 0.0001**

the standard deviation, we can approximate the values by calculating them from raw F_0's of an utterance. Therefore, we decide to use the z-score normalization for the last experiment.

3.7 Tone-Critical Segment

This section will concern the question of which part of the F_0 contour in the syllable should be used for tone recognition. The part containing critical information for tone recognition is referred to as the *tone-critical segment*.

Most studies in this area have been performed in Mandarin. Howie (1974) observed that tones in Mandarin are carried only by the syllable rhyme, while the syllable onset of the F_0 contour corresponding to an initial voiced consonant or glide is merely an adjustment for the voicing of initial consonants and then there is much F_0 perturbation. The studies of F_0 contour analysis and perception (Whalen and Xu 1992; Xu and Wang 1997) showed that segments of different locations in a syllabic F_0 contour may contribute differently to tone perception for the syllable. The F_0 contour segment of the rhyme portion of a syllable contains critical information for tone perception while the onset portion of the F_0 contour is subject to variation. Xu (1997) confirmed that the nasal part of a syllable carries tone information and the movement of F_0 continues all the way to the end of the syllable. Furthermore, Xu (1998) observed coarticulated tones and argued that the syllable was the appropriate domain for tone alignment, and the large perturbation seen at the syllable onset portion of the F_0 contour of a syllable

is the result of the carry-over effect[2] from the preceding syllable.

In tone recognition literature, most studies (Chen and Wang 1995; Lee et al. 1995; Cao et al. 2000) have used voiced segment within a syllable as a tone critical segment. Wang and Seneff (1998, 2000), however, proposed to use the F_0 contour from the syllable rhyme for tone modelling. Besides the above mentioned reasons, they argued that a speech recognition system in Mandarin generally uses syllable initials (onsets) and finals (rhymes) as acoustic modelling units and the presence of a syllable initial should reduce the carry-over effect. On the contrary, Zhang and Hirose (1998) argued that coda (final consonant) in the rhyme portion is also less important for tone perception of a syllable. They therefore used only the vowel nucleus as a tone-critical segment.

The definition of tone-critical segment is not yet well understood. We think that vowel nucleus is not enough to identify tones for Thai language. The vowel length is depend on the syllable structure. The vowel lengths in CVːN syllables are shorter than those in CVː syllables, where C, Vː, and N represent an initial consonant, a long vowel, and a final sonorant (Zhang 2001). Therefore, the vowel alone in CVːN cannot capture all tone information. Moreover, due to the Thai syllable structure (much similar to Mandarin), an onset-rhyme unit or a syllable unit is more suitable for speech recognition system than a phoneme unit. A number of studies in Thai have considered rhyme units for prosodic studies of stress (Potisuk, Gandour, and Harper 1996; Potisuk, Harper, and Gandour 1996; Thubthong and Kijsirikul 2001a) and tones (Potisuk et al. 1999; Thubthong and Kijsirikul 2001a), and syllable units for speech recognition (Demeechai and Mäkeläinen 2001; Thubthong and Kijsirikul 1999b; Thubthong and Kijsirikul 2000a). Thus, we will consider only rhyme and syllable units as tone-critical segments. We built an experiment for comparing the performance between syllable and rhyme units on tone recognition. In order to prevent the effect of voiced/unvoiced portions in a syllable, only Potisuk-1999 corpus was used. We applied tone feature set C in ERB-rate scale. It was extracted from both syllable and rhyme units. All feature parameters were normalized by the z-score technique. We performed both inside and outside tests.

The results are shown in Table 3.8. Rhymes yield better recognition rates than

[2]The carry-over effect is the effect of the preceding syllable as described in Chapter 4.

Table 3.8: Recognition rates (%) and standard deviations (s.d.) of tone recognition with different tone critical segments. The experiment was performed on PC-99 (both inside and outside tests). The best results for each test are printed in **boldface**.

Segment	Inside test		Outside test	
	%	s.d.	%	s.d.
Syllable	92.05	1.06	88.95	2.17
Rhyme	**92.86**	1.23	**90.05**	2.33

Table 3.9: Measure of statistical difference of tone recognition with different tone critical segments. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Domain	Inside test	Outside test
	Rhyme	Rhyme
Syllable	0.0700	**0.0298**

syllables for both inside and outside tests but only statistically significant difference is found for the outside test (see Table 3.9).

From (Potisuk et al. 1999), the best recognition rate is 89.10% for the inside test. Although we used the same list of 11 sentences for the test set, we cannot directly compare this result with our results (92.86% for the inside test). This is because the speech data were collected from the different speakers, the different number of speakers, and the different recording environment.

3.8 Discussion

This section describes three points: the usefulness of normalization, the comparison of the recognition rates between different corpora, and the confusion matrices of the classification results for each corpus. Figure 3.1 shows the comparison of the inside and outside tests on PC-99 with different configurations (i.e., tone features, frequency scales and normalization techniques). The performances of the inside test are higher than that of the outside test for all configurations. Normalization successfully reduces the difference of performances between the inside and outside tests. This confirms that normalization is needed to compensate for speaker variability.

Comparing the performances of three corpora, we found that the highest performances are achieved on PC-99 for all configurations, while the poorest performances are reported on TASC for all configurations (see Figure 3.2). These results can be explained by the complexity of the corpus. The speech utterances in PC-99 are very clean and all parts of utterances are voiced, while TPC contains both voiced and voiceless por-

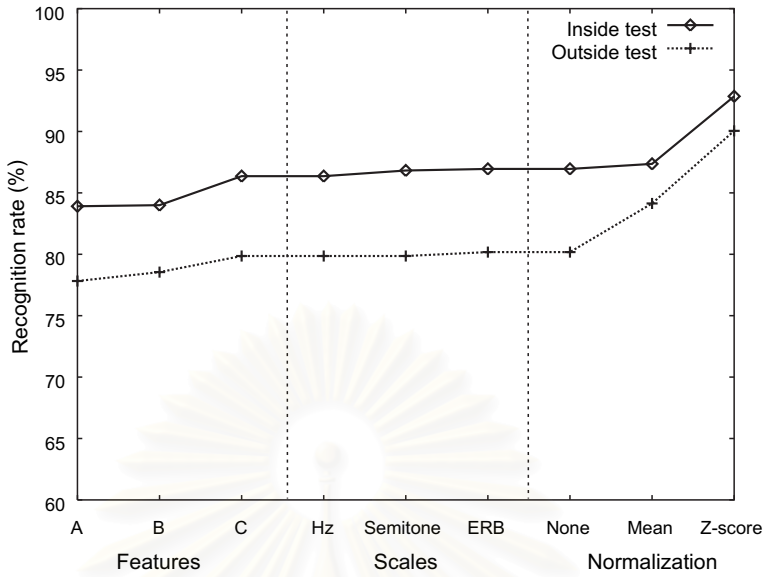


Figure 3.1: Comparison performances of inside and outside tests on PC-99 with different configurations.

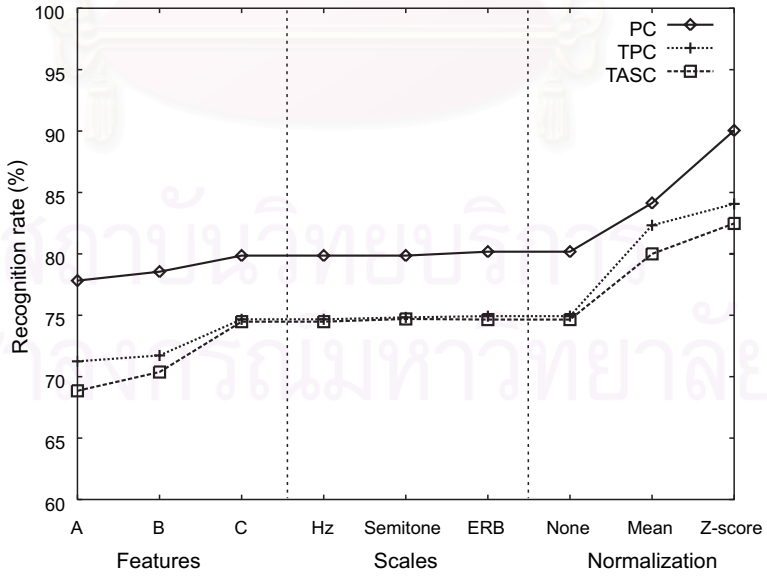


Figure 3.2: Comparison performances of all three corpora with different configurations.

Table 3.10: Confusion matrices of tone recognition for (a) PC-99, (b) TPC, and (c) TASC using the best configurations. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively.

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	200	72.50	**145**	43	0	11	1
L	600	89.17	33	**535**	1	6	25
F	450	**97.56**	0	4	**439**	7	0
H	600	93.17	7	3	11	**559**	20
R	350	86.57	4	16	0	27	**303**
Total	2200	90.05					

(a) PC-99

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	1920	85.16	**1635**	133	73	46	33
L	1200	79.08	183	**949**	6	2	60
F	1280	**92.03**	63	10	**1178**	29	0
H	640	70.00	83	3	59	**448**	47
R	960	86.88	36	56	2	32	**834**
Total	6000	84.07					

(b) TPC

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	2380	86.30	**2054**	114	110	72	30
L	1260	78.57	220	**990**	12	6	32
F	960	**84.06**	116	4	**807**	33	0
H	660	73.03	119	8	39	**482**	12
R	500	83.60	27	34	0	21	**418**
Total	5760	82.48					

(c) TASC

tions. TASC is a read speech corpus that includes some neutral syllables. Most neutral syllables are short vowel syllables produced in unaccented. Therefore, they have a very short syllable duration and a nonstable F_0 pattern. The neutral syllables are commonly excluded from experimental results on tone recognition in most studies. If we do not consider the neutral syllables, the performance will be improved from 82.48% to 85.27%.

Table 3.10 shows confusion matrices of tone recognition for PC-99, TPC, and TASC using the best configurations. The tone references and recognition results of each tone are represented in rows and columns, respectively. The correct results are printed in boldface. It can be seen that the fall provides the highest recognition rate for all corpora, while the mid yields the poorest for PC-99 and the high gives the worst for TPC and TASC. Through broad error analysis, we found that the most errors come from the misclassification between the mid and the low. The results are similar to those in Chapter 2. In TPC and TASC, as the number of syllables with the mid is very large (compared to the number of syllables with the other tones), the classifier may be biased.

Thus, the fall, the high, and the rise are commonly misclassified as the mid.

3.9 Summary

In this chapter, we have presented an empirical study for constructing a basic tone recognition framework. The framework consisted of tone models to parameterize tone F_0 contour and a classifier to evaluate the performance of the tone models. The former was designed by an empirical study, while, for the latter, a three-layer feedforward neural network was applied. To construct the tone models, we concentrated on the questions of which configurations with respect to tone features, frequency scale, normalization technique, and tone critical segment should be used for tone recognition. We built three corpora: Potisuk-1999, Thai Proverb, Thai Animal Story corpora. From the experimental results, we conclude that tone feature set C significantly outperforms the other tone feature sets, ERB-rate scale surpasses semi-tone and hertz scales, z-score normalization significantly exceeds mean normalization, and rhyme units are better than syllable units. Based on these results, we have therefore proposed a Thai tone modelling for tone recognition. The tone modelling is the set of five F_0's and their slopes at 0%, 25%, 50%, 75%, and 100% of the rhyme segment in ERB-rate scale using z-score normalization. It provides 92.86%, 90.05%, 84.07%, and 82.48% recognition rates for Potisuk-1999 (inside test), Potisuk-1999 (outside test), Thai Proverb and Thai Animal Story corpora, respectively.

In the following chapter, we will extend this study by enhancing the tone modelling for compensating external effects and use the basic tone recognition framework for evaluating the further tone models.

CHAPTER 4

EFFECT OF COARTICULATION ON TONE RECOGNITION

Tones in continuous speech can be influenced by many linguistic factors. Neighboring tones are important factors due to articulatory constraints. When produced in context, the tonal contours undergo certain variations depending on preceding and following tones (Xu 1997). Both the following syllable and preceding syllable influence the considering syllable. The effects of the following syllable and the preceding syllable are called *anticipatory coarticulation* and *carry-over coarticulation*, respectively.

Following the work of Xu (1997), we considered F_0 contours of Thai bi-syllables /maː maː/. Figure 4.1 shows the F_0 contour variations due to the influence of the preceding tones in the /maː maː/ sequences produced in isolation. Each panel in the figure plotted the same tone in the second syllable when preceded by five different tones. Each curve was obtained by averaging over two repetitions produced by a female speaker. The time scale was equalized for all the curves. A nasal segment was plotted with 10 points, while a vowel segment was plotted with 50 points. The F_0 contours plotted with duration of each segment proportional to the averaged actual duration of segment are also obtained. We can observe that, at the boundary between the two syllables, the starting F_0 of a given tone in the second syllable varies enormously depending on the tone of the first syllable. The differences due to the preceding syllable decrease gradually over time. There are rapid F_0 movements, which are larger when the adjacent values of two neighboring tones are far apart than when they are similar.

Figure 4.2 shows the variations in F_0 contour of the tones when followed by different tones in the /maː maː/ sequences. In contrast to the strong tendency for the F_0 of the second syllable to assimilate to the offset value of the first syllable, the contours of the first syllable are much less affected by the identity of the following tone.

In this chapter, we propose tone features to compensate coarticualtory effect and investigate a series of experiments on tone recognition using the basic tone recognition framework (as described in Chapter 3). We also propose a novel model called *half-tone model* to improve the performance of tone recognition. In the following sections, we first look at some related works of coarticulatory effect on tone recognition. Then, we present experiments concentrating on tone features and tone models for compensating the effect.

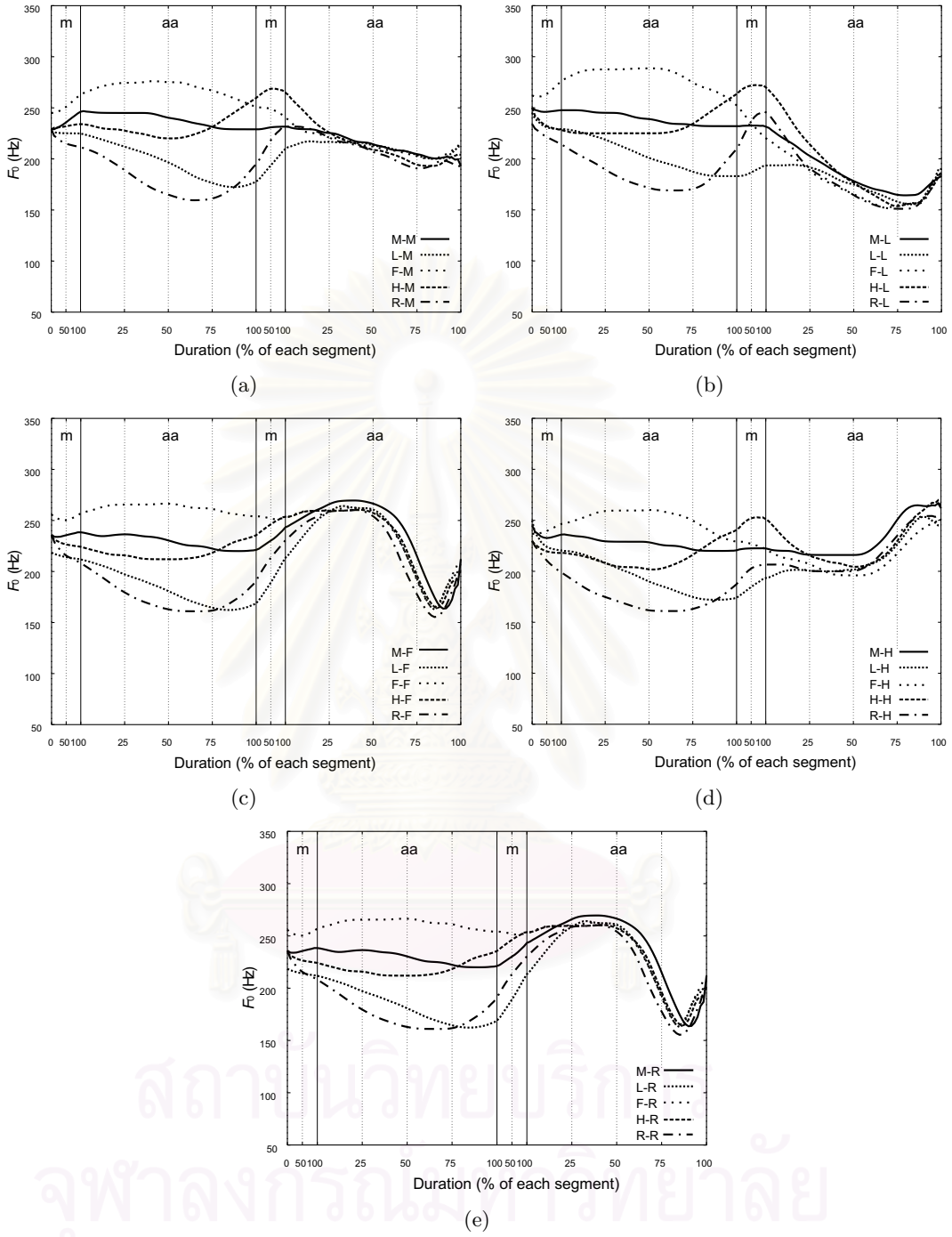


Figure 4.1: Carry-over effect: effect of preceding tones on F_0 contour of following tone in /maː maː/ sequences in Thai. In each panel, the tones in the second syllable was held constant (the mid, the low, the fall, the high, and the rise in (a) to (e), respectively), and the tone of the first syllable was varied. Each curve was from a female speaker.

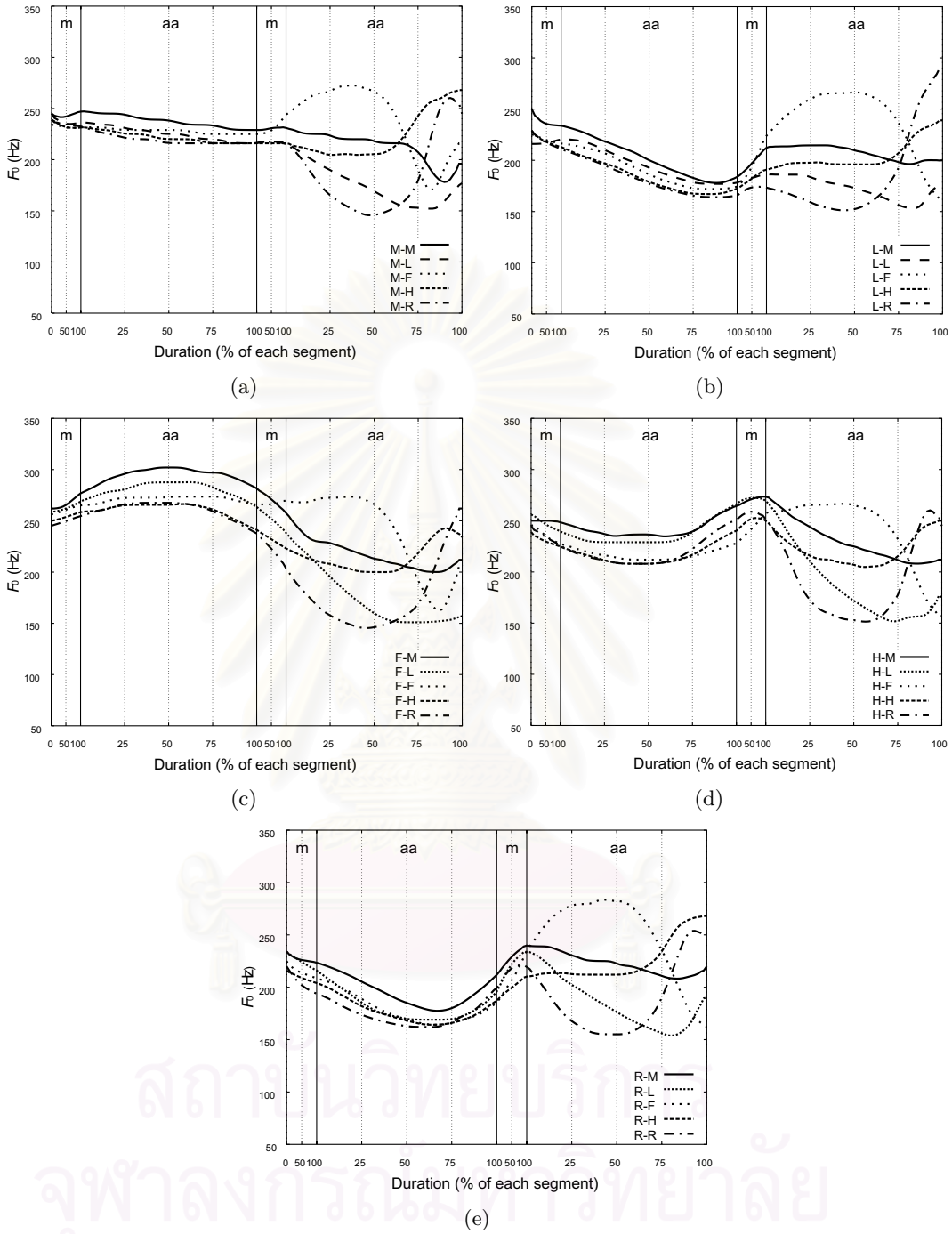


Figure 4.2: Anticitory effect: effect of following tones on F_0 contour of preceding tone in /maː maː/ sequences in Thai. In each panel, the tones in the first syllable was held constant (the mid, the low, the fall, the high, and the rise in (a) to (e), respectively), and the tone of the second syllable was varied. Each curve was from a female speaker.

After that, we perform experiments and analyze the results in several directions. Finally, we give the summary of this chapter.

4.1 Related Works

There are a number of studies on the problem of the coarticulatory effect in phonological, perception and engineering approaches. We first describe the studies on phonological and perception approaches. Most studies are for Mandarin. Shen (1990) studied all possible combinations of tones of Mandarin on /baː baː baː/ tri-syllables, and found that not only the onset and offset values but also the overall heights of a tone were affected, and the coarticulatory effects were bi-directional and symmetric. Xu (1994) conducted a perceptual study of coarticulated tones of Mandarin and found that human performance on tone identification was highly dependent on the availability of original tone context when the context was "conflicting" with the tone. Xu (1997, 1999) also studied F_0 contours of Mandarin bi-syllables /maː maː/ embedded in a number of carrier sentences. He found that anticipatory and carry-over effects differed both in magnitude and in nature: the carry-over effects were larger in magnitude and mostly assimilatory in nature, e.g., the onset F_0 value of a tone was assimilated to the offset value of a previous tone; the anticipatory effects were relatively small and mostly dissimilatory in nature, e.g., a low onset value of a tone raised the maximum F_0 value of a preceding tone.

Gandour et al. (1994) studied all possible combinations of Thai tones on di-syllables, and found that the coarticulatory effects were not symmetric. Thai tones were more influenced by carry-over than by anticipatory coarticulation. Carry-over coarticulation affected a greater number of Thai tones and extended further into adjacent tones. Carry-over effects extend forward to about 75% of the duration of the following syllable, while anticipatory effects extend backward to about 50% of the duration of the preceding syllable. Coarticulation affected primarily height but the slope was relatively unaffected. Magnitude of the effects was fairly uniform regardless of direction of coarticulation.

In engineering point of views, there are a number of studies of tone recognition on the coarticulatory effect. Chen and Wang (1995) focused on coarticulatory and intonation effects. For the coarticulatory effect, they concentrated on the effect of

neighboring syllables and sandhi rules (Lee et al. 1989). Some contextual features extracted from neighboring syllables and tones of neighboring syllables were added to compensate the effect. The intonation effect will be described in the next chapter. Zhang et al. (2000) claimed that carry-over played a more important role than anticipation for tone discrimination. They also proposed a new Chinese lexical tone recognition based on the *pitch anchoring* hypothesis. The hypothesis indicated that the tone offset of the preceding lexical tone and the tone onset of the succeeding lexical tone serve as anchor points for the pitch heights of the onset and offset of the sandwiched lexical tone. The experiments were carried out on the data of a female speaker in the data corpus HKU96. Continuous density HMMs with left-to-right were employed. The new approach provided a notable increase about 10% compared with the conventional one. Wang and Seneff (2000) proposed a method to improve tone recognition by normalizing coarticulatory, phrase boundary, and intonation effects. The tone classification errors on continuous digit strings were reduced by 26.1% from the baseline, when these effects were normalized.

In Thai, Potisuk et al. (1999) proposed an analysis-by-synthesis algorithm for recognizing Thai tones in continuous speech. This algorithm used an extension to Fujisaki's model (Fujisaki 1983) for a tone language. The study concentrated on the coarticulatory and intonation effects. They used 125 possible three tone sequences of five Thai tones and 11 sentences with varying tone sequences for training and testing, respectively. The classification result was given in term of a tone sequence (not a tone in each syllable). The recognition rate of 89.10% was achieved.

4.2 Methodology

This section presents *contextual tone features* and *half-tone model* proposed for alleviating coarticulatory effect.

4.2.1 Contextual tone features (CTF)

As mentioned previously, coarticulatory effect of neighboring syllables will make the F_0 contour of a syllable change in level and in shape to interfere with tone discrimination. The coarticulatory effect decreased when the distance between neighboring syllables increased. This means that the coarticulatory effect is inversely proportional

to the distance between neighboring syllables. Therefore, we used some features extracted from neighboring syllables and the distance between neighboring syllables to identify tones. They include: (i) a number of F_0 heights and slopes of neighboring syllables (both preceding and following syllables) and (ii) the two inversions of duration: from the ending point of the preceding syllable to the beginning point of the considering syllable (d_p) and from the ending point of the considering syllable to the beginning point of the following syllable (d_f) where the duration is measured in second. The former are the primary features for coping with the coarticulation while the latter are employed to implicitly represent the tightness of relationships between the considering syllable and the two neighboring syllables. These features are referred to as *contextual tone features*.

As described in (Gandour et al. 1994), carry-over effects extend forward to about 75% of the duration of the following syllable, while anticipatory effects extend backward to about 50% of the duration of the preceding syllable. However, they analyzed on a small set of hypothesis sentences. We therefore explored different pairs of F_0 points from preceding and following syllables. We refer to these features as CTFmn where m and n are the numbers of time points of the preceding and following syllables used to extract contextual tone features, respectively. For example, CTF34 is the features when the three F_0 heights and slopes at 50%, 75% and 100% of the preceding syllable and the four F_0 heights and slopes at 0%, 25%, 50%, and 75% of the following syllable are employed as shown in Figure 4.3. In the case of the beginning or ending syllable, F_0 heights and slopes are set to 0 and a duration feature is set to 1.

4.2.2 Tone models

This subsection describes two conventional tone models, i.e., context independent and context-dependent tone models, and a novel tone model called *half-tone model*. The half-tone model is proposed to enhance the performance of tone recognition in term of recognition rate and speed.

1. *Context-independent tone model* (CI-T)

   This model treats the considering syllable as being independent of its neighboring syllables. Therefore, there are only five different tones. This model will be used to evaluate our tone features in Subsection 4.3.1.

2. *Context-dependent tone model* (CD-T)

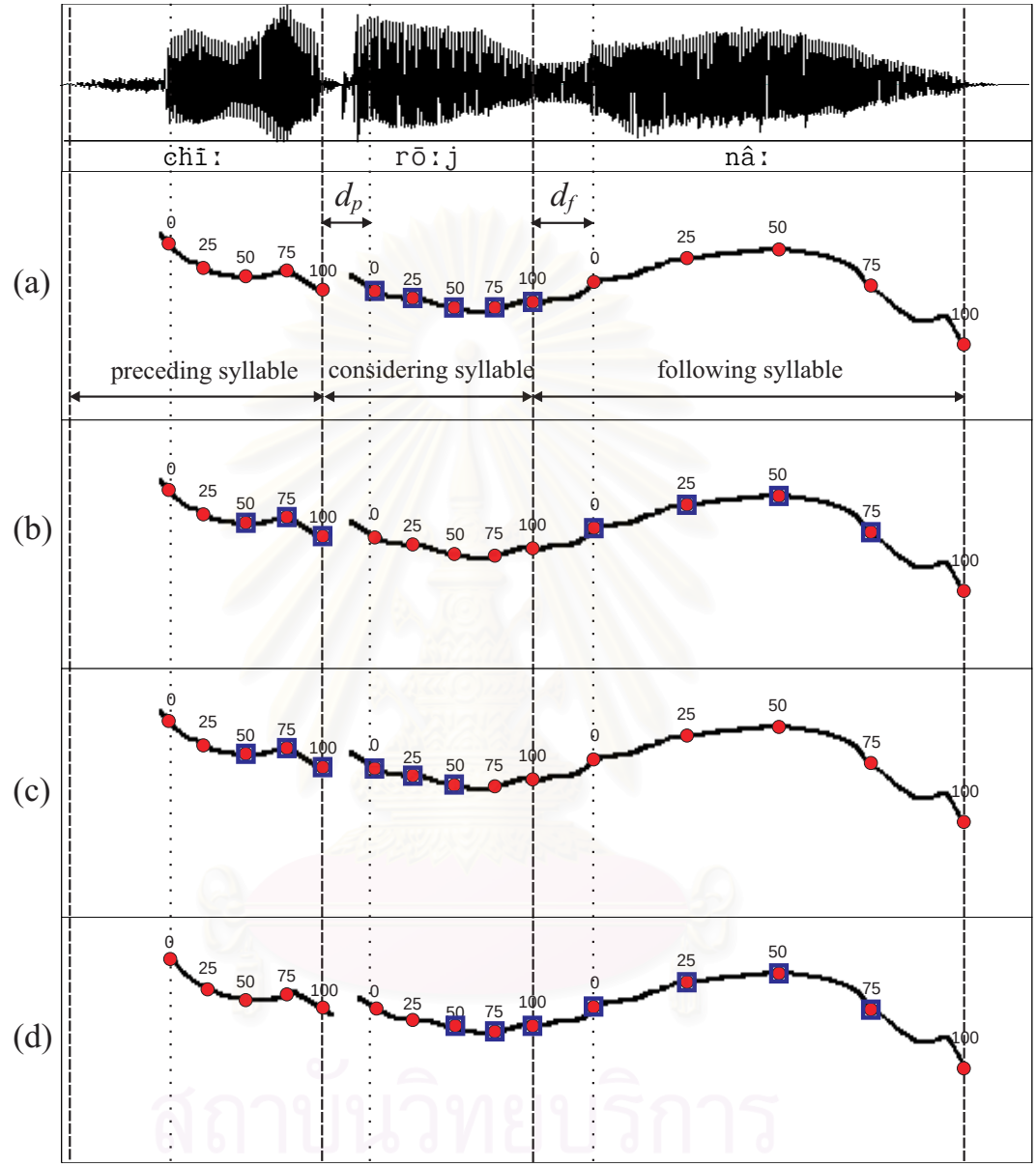


Figure 4.3: The different time points of (a) the baseline tone features (simple tone models), (b) the contextual tone features (for CTF34), (c) the first-half tone features (for CTF34), and (d) the second-half tone features (for CTF34). The circle points represent five F_0 at different time points of a syllable and bar points represent the F_0 used for different tone features. The dash lines and dot lines represent syllable boundaries and onset-rhyme boundaries, respectively.

In fact, the neighboring syllables affect the considering syllable. Potisuk et al. (1999) used three-tone sequences to measure this effect. There are 125 possible three-tone sequences. However, they concentrated only the syllable in the middle of a sentence. To enhance these sequences, we also consider the syllable at the beginning and the end of sentence. Therefore, a total of 175 sequences will be needed, i.e., 5^3 (in the middle of a sentence) + 5^2 (at the beginning of a sentence) + 5^2 (at the end of a sentence).

3. *Half-tone model* (H-T)

   The number of 175 possible sequences in CD-T is too large and its training time is very long. Thus, we propose a novel model called *half-tone model* (H-T). This model is based on the "divide-and-conquer" principle. A syllable is separated into two parts at the center. For the first half, a total of 30 sequences will be needed, i.e., 5 (at the beginning of a sentence)+ 5^2 (in the middle of a sentence). Also for the second half, a total of 30 sequences will be needed, i.e., 5 (at the end of a sentence)+ 5^2 (in the middle of a sentence). The first half is trained by using one classifier and the second half is trained by the other classifier. The outputs of two classifiers are then combined to determine the classification result (see Figure 4.4). The tone features of this model are the baseline tone features plus the contextual tone features. For example, when CTF34 is used, we obtain F_0 heights and slopes at three different time points at 0%, 25%, and 50% of the considering syllable and at three different time points at 50%, 75%, and 100% of the preceding syllable for the first half. In the same way, F_0 heights and slopes at three different time points at 50%, 75%, and 100% of the considering syllable and at four different time points at 0%, 25%, 50%, and 75% of the following syllable are obtained for the second half (see Figure 4.3 (c) and (d)).

For the context-independent tone model, the output of the model is the classification result, but for the context-dependent and half-tone models, the outputs are not. They need a final decision process. Thus, we employed two decision algorithms for:

1. *Context-dependent tone model*: All output sequences are grouped into five groups depending on the tone of the middle syllable in each sequence. The *posteriori* probabilities of each sequence in each group are summarized as the group score.

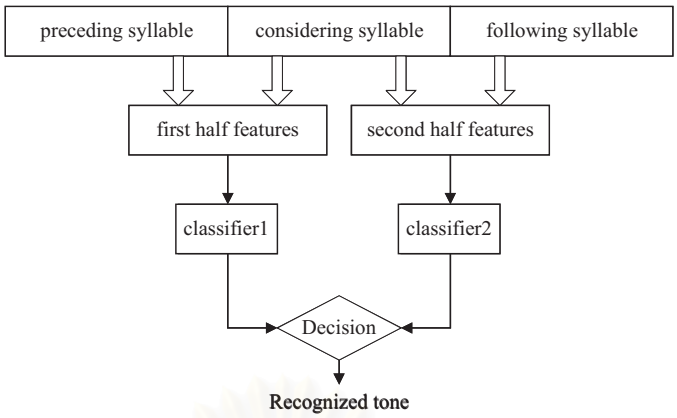


Figure 4.4: The process of the half-tone model.

The group providing the highest group score is chosen as a classification result.

2. *Half-tone model*: Each output sequence pair of two classifiers (one for first-half and the other for second-half) are grouped into five groups depending on the tone of the middle syllable in each sequence. The *posteriori* probabilities of each sequence in each group are summarized as the group score. Then, the classification result is the group that provides the highest group score.

4.3 Experiments

We conducted two main experiments. The first and second experiments are for evaluating the contextual tone features and tone models, respectively.

4.3.1 Experiments of different contextual tone features

To evaluate the contextual tone features, we performed experiments on PC-99 (only outside test), TPC, and TASC using the basic tone recognition framework as described in Chapter 3. We used different configurations of the contextual tone features, i.e., CTF12, CTF22, CTF23, CTF33, CTF34, and CTF44. The context-independent tone model was employed. The results for the simple tone models were used as the baseline results.

The recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of different refinements of contextual tone features are shown in Table 4.1. The

Table 4.1: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition with different contextual tone features for PC-99 (outside test), TPC, and TASC.

Tone model	PC-99 (outside test)			TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
Simple	90.05	2.33	-	84.07	0.80	-	82.48	2.41	-
+CTF12	92.77	3.22	27.40	92.23	0.81	51.26	89.06	2.24	37.56
+CTF22	93.55	2.66	35.16	92.37	0.97	52.09	88.89	1.07	36.57
+CTF23	93.95	2.80	39.27	92.95	0.63	55.75	89.34	2.18	39.15
+CTF33	94.23	2.84	42.01	**93.02**	0.84	56.17	89.41	1.89	39.54
+CTF34	**94.27**	2.83	42.47	92.93	0.69	55.65	**89.91**	2.02	42.42
+CTF44	94.23	2.95	42.01	92.67	1.15	53.97	89.74	2.13	41.43

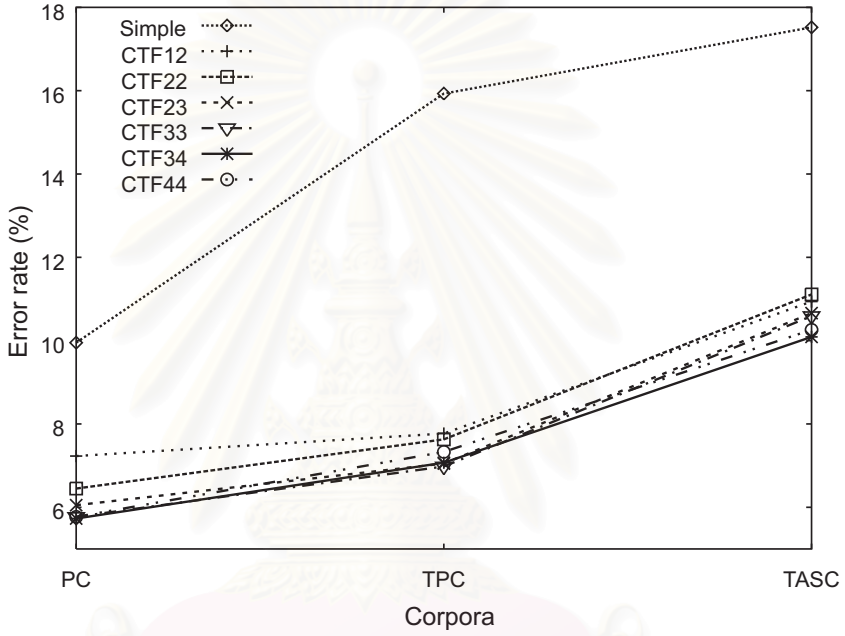


Figure 4.5: Error rates of different contextual tone features.

best recognition rates of 94.27%, 93.02% and 89.91% are reported for PC-99, TPC, and TASC, respectively. Compared to the baseline results, the highest error reduction rates are 42.47%, 56.17%, and 42.42% for PC-99, TPC, and TASC, respectively. Based on McNemar's test, the differences in performance are highly significant. Considering the different configurations of CTF, CTF34 provides the best recognition rates for PC-99 and TASC, while CTF33 yields the highest recognition rate for TPC. This supports the study of (Gandour et al. 1994) that carry-over effects extend forward to about 75% of the duration of the following syllable, while anticipatory effects extend backward to about 50% of the duration of the preceding syllable. However, the performances for all CTF configurations are close and not significantly different (see Figure 4.5).

4.3.2 Experiments of different tone models

For the second main experiment, the context-independent tone model (CI-T-5), the context-dependent tone model (CD-T-175), and the half-tone model (H-T-30) were conducted on PC-99 (only outside test), TPC, and TASC. Only CTF34 was applied. The results are shown in Table 4.2. The recognition rates of CD-T-175 and H-T-30 outperform those of CI-T-5 for all corpora. CD-T-175 provides the best recognition rates for TPC and TASC whereas H-T-30 yields the best for PC-99. The statistical significances in recognition rates are reported for the differences between CI-T-5 and CD-T-175 on TPC and TASC, as shown in Table 4.3. The best error reduction rates are 11.11%, 11.56%, and 11.53% for PC-99 (H-T-30), TPC (CD-T-175), and TASC (CD-T-175), respectively.

The recognition rates of H-T-30 are slightly higher than that of CD-T-175 for PC-99 but slightly lower than those of CD-T-175 for TPC and TASC. However, the training time for CD-T-175 is very long (see Table 4.4). It is about 19 time longer than those for CI-T-5, while the training time for H-T-30 is about 5 time longer than those for CI-T-5. In the other words, the training time of H-T-30 is about one-fourth of CD-T-175. This concludes that H-T-30 is a promising model. The model is the best choice when we want to optimize both recognition rates and training time.

4.4 Discussion

Table 4.5 shows confusion matrices of tone recognition for PC-99, TPC, and TASC using CTF34 and CI-T-5. The tone references and recognition results of each tone are represented in rows and columns, respectively. The correct results are printed in boldface. The results are similar to those in Table 3.10. The fall provides the highest recognition rate for all corpora, while the mid yields the poorest for PC-99 and the high gives the worst for TPC and TASC. Most errors come from the misclassification between the mid and the low. In addition, the fall, the high, and the rise are commonly misclassified as the mid.

4.5 Summary

In this chapter, we have considered coarticulatory effect on tone recognition. We have proposed the contextual tone features and used them to refine the simple tone

Table 4.2: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition with different tone models for PC-99 (outside test), TPC, and TASC. CTF34 was used.

Model	PC-99 (outside test)			TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
CI-T-5	94.27	2.83	-	92.93	0.69	-	89.91	2.02	-
CD-T-175	94.73	0.52	7.94	**93.75**	1.28	11.56	**91.08**	1.45	11.53
H-T-30	**94.91**	0.56	11.11	93.32	1.00	5.42	90.66	1.62	7.40

Table 4.3: Measure of statistical difference of tone recognition with different frequency scales. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).)

Model	PC-99 (outside test)		TPC		TASC	
	CD-T-175	H-T-30	CD-T-175	H-T-30	CD-T-175	H-T-30
CI-T-5	0.3428	0.1658	**0.0202**	0.2731	**0.0067**	0.062
CD-T-175	-	0.7404	-	0.2225	-	0.2711

Table 4.4: Training times of tone recognition with different tone models for PC-99 (outside test), TPC, and TASC. CTF34 was used. The experiments were run on a Pentium III 866MHz machine.

Model	PC-99 (outside test)		TPC		TASC	
	second	ratio	second	ratio	second	ratio
CI-T-5	1,640	1.0	4,280	1.0	4,120	1.0
CD-T-175	30,600	18.7	83,200	19.4	81,600	19.8
H-T-30	7,820	4.8	20,940	4.9	20,040	4.9

models for compensating this effect. We explored several configurations of contextual tone features, i.e., CTF12, CTF22, CTF23, CTF33, CTF34, and CTF44. To evaluate these refined tone models, we performed experiments on PC-99, TPC, and TASC using the basic tone recognition framework (described in Chapter 3). A context-independent tone model (CI-T-5) was employed. The experimental results showed that all refined tone models outperformed the simple tone models for all corpora. CTF34 provided the best recognition rates. These results confirmed the study of (Gandour et al. 1994) that carry-over effects extended forward to about 75% of the duration of the following syllable, while anticipatory effects extended backward to about 50% of the duration of the preceding syllable.

Furthermore, the context-dependent tone model (CD-T-175) was applied to enhance recognition rate. CD-T-175 provided better recognition rates than CI-T-5 for all corpora. However, the training times for CD-T175 was very long. Therefore, we have also proposed a novel model called *half-tone model* (H-T-30) to alleviate the drawback of CD-T-175. We found that H-T-30 also increased recognition rates over CI-T-5 for all

Table 4.5: Confusion matrices of tone recognition for (a) PC-99, (b) TPC, and (c) TASC using CTF34 and CI-T-5. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively.

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	200	86.50	**173**	24	2	1	0
L	600	93.33	30	**560**	0	3	7
F	450	**98.89**	1	1	**445**	3	0
H	600	96.83	2	0	6	**581**	11
R	350	90.00	3	4	0	28	**315**
Total	2200	94.27					

(a) PC-99

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	1920	93.65	**1798**	49	28	27	18
L	1200	90.58	73	**1087**	4	5	31
F	1280	**96.88**	31	1	**1240**	8	0
H	640	88.44	34	5	24	**566**	11
R	960	92.19	21	39	1	14	**885**
Total	6000	92.93					

(b) TPC

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	2380	92.35	**2198**	75	48	49	10
L	1260	85.71	132	**1080**	17	5	26
F	960	**92.92**	47	7	**892**	14	0
H	660	83.33	81	8	16	**550**	5
R	500	91.80	12	20	0	9	**459**
Total	5760	89.91					

(c) TASC

corpora. Considering the best recognition rates, we found that CD-T175 archived the highest recognition rates for TPC and TASC, while H-T-30 yielded the best for PC-99. However, H-T-30 was better than CD-T-175 in term of speed. The recognition rates of H-T-30 were slightly higher than those of CD-T-175 for PC-99 but slightly lower than those of CD-T-175 for TPC and TASC. This concludes that H-T-30 is a promising model. The model is the best choice when we want to optimize both recognition rates and training time.

CHAPTER 5

EFFECT OF INTONATION ON TONE RECOGNITION

Intonation is defined as the combination of tonal features into larger structural units associated with the acoustic parameter of *voice fundamental frequency* and its distinctive variations in the speech process (Botinis et al. 2001). A broad term "intonation" is the contour of F_0 throughout an utterance. At a phrasal level, the F_0 contours and inflections, which make up intonation, contribute information about semantics and syntax, as well as about emotion and meaning. In addition, the contrastive value and direction (e.g. the rise, the fall) of intonation may contribute to the perception of syllabic stress or even word meaning (Johnson 2000). On the other hand, intonation phrases may be associated with different sentence types (Botinis et al. 2001), which may define as statements, questions, commands, etc.

Luksaneeyanawin (1993) use the term "intonation" to refer to a distinctive pitch of an information unit, either a word or a set of words which are semantically and syntactically unified to other information units in the information whole. The function of intonation is to distinguish the grammatical meaning as well as the attitudinal meaning of the intonation unit.

There are three intonation contours in Thai: the Fall, the Rise, and the Convolution (Luksaneeyanawin 1993). The Fall conveys the semantic finality, closeness, and definiteness. The Rise conveys the semantic non-finality, openness, and non-definiteness. The Convolution conveys the semantic contrariety, conflicts, and emphasis. Luksaneeyanawin (1993) claimed that each tone has its own behavior when superimposed by different intonations. Figure 5.1 shows an example of three intonation contours. The first one is a statement sentence representing the falling intonation; the second one is a question sentence representing the rising intonation; and the last one is a statement emphatic sentence representing the convolution intonation. The falling and rising intonation contours can be represented by a straight line, e.g., topline (T), baseline (B) or all-points line (Lieberman et al. 1985). However, the convolution intonation needs more than one straight line or a nonlinear line to represent it.

In perception studies, the effect of declination is compensated for by the listener, thus two accents occurring at different times in a phrase can have equal prominence,

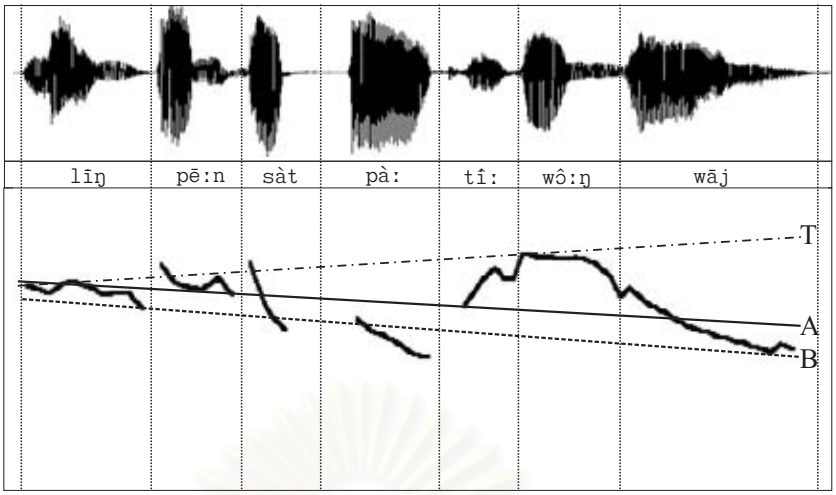
but widely differing F_0 values (Taylor 1992).

In acoustic point of view, the local F_0 will be adjusted to conform to the intonation pattern of the sentence (Wang and Chen 1994). For example, statements are generally characterized by a global rightward intonation lowering which is mainly related to junctures (alias boundary tones) and tonal range variations. The initial juncture, i.e., the tonal onset at the very beginning of the sentence, is usually higher than the final juncture, i.e., the tonal offset at the very end of the sentence. The F_0 contour of the entire sentence is in descending form from left to right. The general rightward lowering is referred to as *declination* (Botinis et al. 2001). Declination is defined as the tendency of F_0 to gradually decline over the course on an utterance (Shih 1997; Swerts et al. 1996; Thorsen 1980). F_0 values of tones are affected by the declination to varying degrees. The declination effect plays an important role in tone recognition in term of F_0 height adjustment of tones (Potisuk et al. 1995). For example, a falling tone at the beginning of an utterance is higher in pitch than at the end.

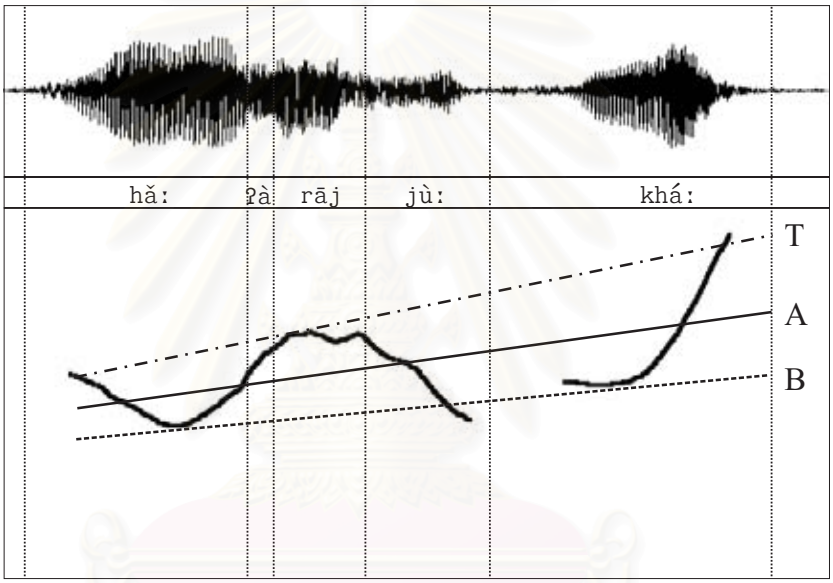
This chapter addresses the effect of intonation on tone recognition. The goal of this chapter is to compensate the effect and improve tone recognition performance. We will first model the intonation contour as a straight line and then subtract the intonation contour from the F_0 contour of each utterance. Due to the limitation of speech corpora, we focus on the effect of falling intonation (declination effect) only. In the following sections, we first present some related works on intonation and the effect of intonation on tone recognition. Then, we describe a method, called *intonation normalization*, to compensate this effect. After that, we will evaluate the method by simulating a number of experiments on Thai proverb corpus (TPC) and Thai animal story corpus (TASC) using the basic tone recognition framework as described in Chapter 3. The experimental results will be discussed in several aspects. Finally, we give the summary of this chapter.

5.1 Related Works

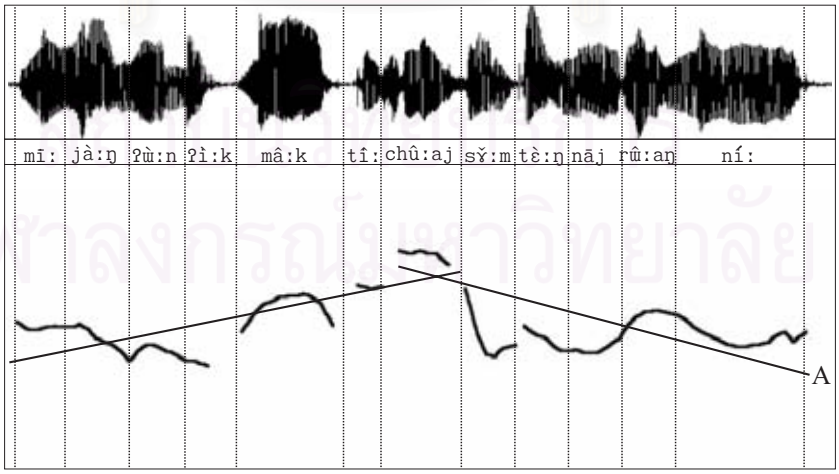
Several studies have been conducted on intonation for various languages (Jensen et al. 1994; Kochanski and Shih 2001; Madhukumar et al. 1993; Thorsen 1980). Some of them were the studies of tone recognition in the intonation effect. These studies were based on classification techniques (Chen and Wang 1995; Huang and Seide 2000) and acoustic features (Wang and Seneff 2000). For the first one, Chen and Wang (1995)



(a)



(b)



(c)

Figure 5.1: The F_0 contours of three intonation types: (a) falling intonation, (b) rising intonation, and (c) convolution intonation, with all-points lines (A), baselines (B), and topline (T).

studied the intonation pattern of sentence pronunciation. They used a hidden control neural net (HCNN) and a hidden state multilayer perceptron (HSMLP) to model the global intonation pattern of a sentential utterance as a hidden Markov chain and use a separated recognizer for tone discrimination. Error reduction rates of 6.45% and 12.63% were reported for the cases of using HCNN and HSMLP, respectively.

For the second one, Lieberman et al. (1985) determined three different declination measures, by fitting linear regression lines to local peaks (topline), local valleys (baseline) and all F_0 points. They argued that all-point line was a better descriptor of sentence F_0 contours in speech than either baseline- or topline-declination models. Swerts et al. (1996) proposed a comparison of F_0 declination in read-aloud and spontaneous speech in Swedish. They estimated the slope of the declination by fitting an all-points regression line to the F_0 points (with semitone scale). They found that both speaking styles revealed negative slopes, a steepness-duration dependency with declination being less steep in longer utterances than in short ones and resetting at utterance boundaries. However, there was a difference in degree of declination between the two speaking styles that read-aloud speech have steeper slopes, a more apparent time-dependency and stronger resetting than spontaneous speech. Shen (1989) conducted an experiment on a small set of read utterances to investigate if intonation can change the tone contours to beyond recognition. He found that both the shape and the scale of a given tone were perturbed by intonation. However, the basic tone shape is still preserved under different intonation types (e.g., statement or question intonations). Moreover, he also found that tones at sentence initial and final positions seem to behave differently from at other positions. Wang and Seneff (2000) also used all-points regression line to represent intonation contour and assumed that all utterances in a speech corpus have a similar underlying intonation contour. Therefore, an F_0 contour can be viewed as a "constant" intonation component with additive "random" perturbation. They smoothed out the "random" variations by averaging and the average could be obtained as the underlying intonation contour. They subtracted the intonation contour from each F_0 contour and re-trained tone models. The method provided 13.5% of error reduction rate.

In Thai, Potisuk et al. (1999) proposed an analysis-by-synthesis algorithm for recognizing Thai tones in continuous speech. This algorithm used an extension to Fujisaki's model (Fujisaki 1983) for a tone language. The model considers the F_0 contour

as the response of the phonatory system to a set of suprasegmental commands: the phrase and the accent commands. The phrase command produces the baseline component that captures the global variation (intonation effect); whereas the accent command produces the accent component of an F_0 contour that captures local variations (accent effect). In a tone language, the accent component can be replaced by the tone component. The tone component is able to capture tone types, tonal coarticulation, and stress effects. They concentrated on the coarticulatory and intonation effects. To compensate the intonation effect, they removed the effect by subtracting the exponential curve used as the response function of the phrase component. They used 125 possible three tone sequences of five Thai tones and 11 sentences with varying tone sequences for training and testing, respectively. The classification result was given in term of a tone sequence (not a tone in each syllable). The recognition rate of 89.10% was achieved. There was no the comparison of the results before and after removing the intonation effect.

5.2 Methodology

There is no un-criticized method available to quantitatively determine the slope and domain of intonation contour yet. According to the phonological approach to intonation (Ladd 1996; Thorsen 1978), the intonation contour was found to approach straight line of pitch accents and boundary tones whose slopes vary according to sentence types, and there is an overall downstep[1] trend of the F_0 height of the pitch accents.

Like (Swerts et al. 1996; Wang and Seneff 2000), we applied the all-point approach to determine the intonation line. We also assumed that all utterances in a speech corpus have a similar underlying intonation contour. Figure 5.2 shows the F_0 contours of all utterances. To deal with different utterances with different durations, the time scale of each utterance is normalized by the utterance duration. The plot shows that there is a steady downdrift[2] of the mean F_0 contour. The mean F_0 contour is then fitted with a first-order polynomial ($y = a_0 + a_1 x$). The straight line represents an F_0 downdrift.

To neutralize the intonation effect, each F_0 will be adjusted. The adjustment of the slope affects the F_0 contour of tones and then tone recognition rate may be dropped.

[1]Downstep refers to the phenomenon that a high (H) pitch target has lower F_0 height after a low (L) pitch target.

[2]Downdrift (often referred to as *downstep* and *declination*) is an important aspect of intonation. Declination refers to the tendency for F_0 to gradually decline over the course of an utterance. A broad term "downtrend" is used to describe the combined effects of the two.

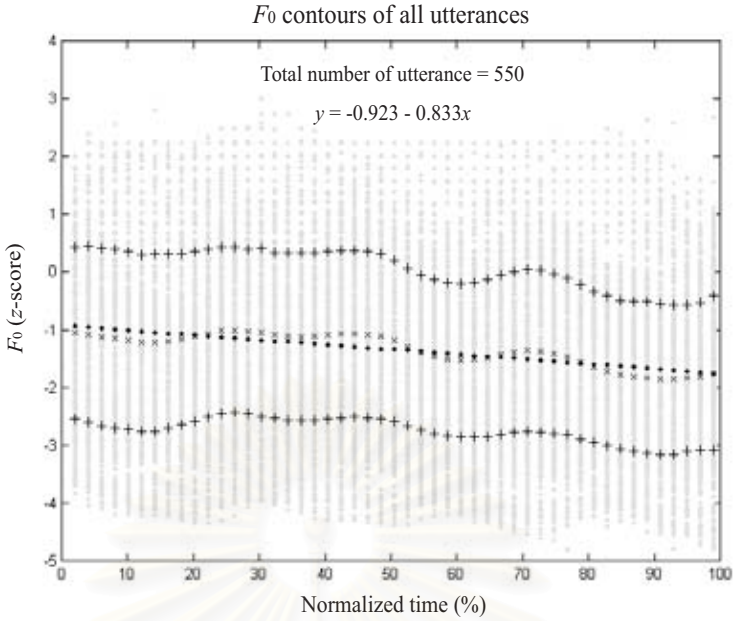


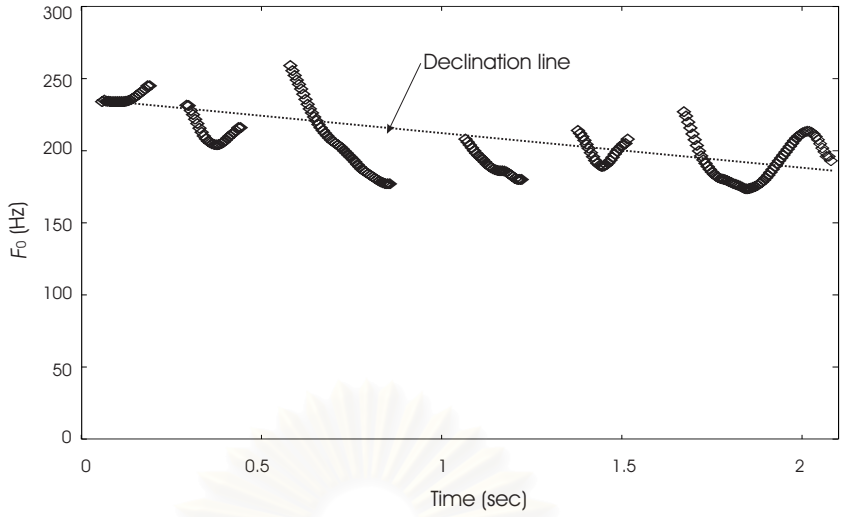
Figure 5.2: F_0 contours of all utterances. The '×' line represents the mean F_0 contour, with the upper and lower '+' lines for standard deviation. The dot line is the first-order polynomial regression line for the average F_0 contour.

Therefore, only F_0 heights will be considered to be adjusted. We have obtained two adjusting lines, i.e., *beginning-point adjusting line* and *center-point adjusting line*. The former uses the beginning point of the intonation line as a moment point, while the latter uses the center point of the intonation line as a moment point, as shown in Figure 5.3. Each F_0 will be adjusted to conform the adjusting lines. These adjustments are referred to as *beginning-point intonation normalization* (BIN) and *center-point intonation normalization* (CIN) for the former and latter adjusting lines, respectively.

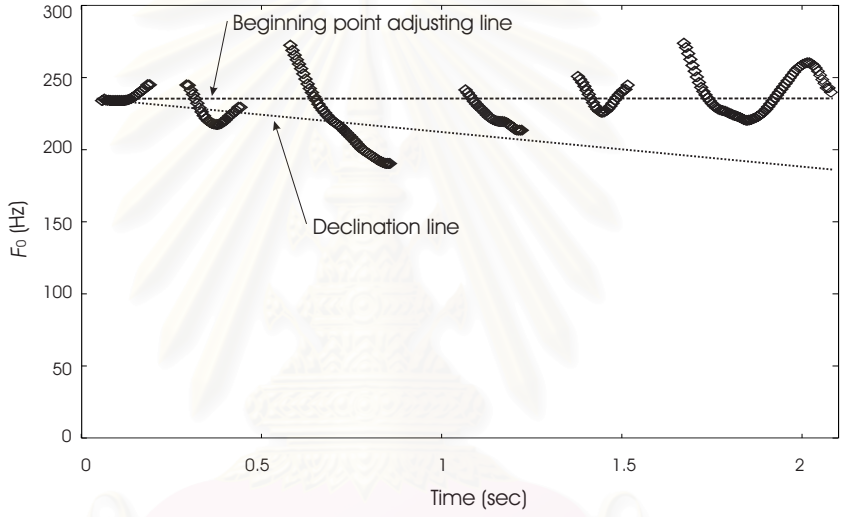
5.3 Experiments

In order to evaluate intonation normalization methods, we performed a number of experiments on TPC and TASC using the basic tone recognition framework. The results of the simple tone model were used as the baseline results.

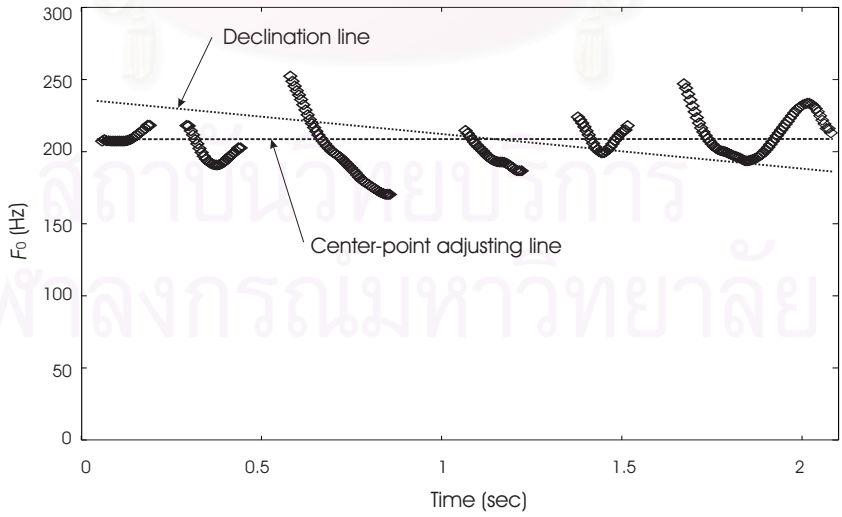
The recognition rates (%), standard deviations (s.d.) and error reduction rate (%ER) are shown in Table 5.1. Compared to the baseline results, both intonation normalization methods significantly increase recognition rates for both corpora (see Table 5.2). BIN provides the maximum error reduction rates of 22.20% for TASC, while



(a)



(b)



(c)

Figure 5.3: F_0 contours: (a) before applying intonation normalization, (b) after applying beginning-point intonation normalization, and (c) after applying center-point intonation normalization.

Table 5.1: Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of tone recognition with different intonation normalization methods on Thai proverb corpus (TPC) and Thai animal corpus (TASC). The best recognition rates for each corpus are printed in **boldface**.

Tone model	TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER
Simple	84.07	0.80	-	82.48	2.41	-
+BIN	86.65	1.10	16.21	**86.37**	2.45	22.20
+CIN	**86.77**	1.04	16.84	86.13	2.23	20.81

Table 5.2: Measure of statistical difference of tone recognition with different intonation normalization methods. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Tone model	TPC		TASC	
	+BIN	+CIN	+BIN	+CIN
Simple	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
+BIN	-	0.7182	-	0.3159

CIN yields the maximum error reduction rate of 16.84% for TPC. The differences in recognition rates between BIN and CIN are not statistically significant for both corpora (see Table 5.2).

We then conducted an experiment by incorporating the contextual tone features (CTF34) into the simple tone model. We also applied both intonation normalization methods for enhancement. The results are shown in Table 5.3. After applying CIN, better recognition rates for both corpora are reported, whereas better recognition rate for only TASC is achieved after employing BIN. The best error reduction rates of 1.42% and 6.71% are reported for TPC and TASC, respectively when using CIN. However, all better results are not statistically significantly different (see Table 5.4).

5.4 Discussion

The tones at sentence initial and final positions seem to behave differently from at other positions (Shen 1989). We further analyzed the effect of intonation on different syllable positions by examining the recognition rates of syllables located in different position of an utterance. Because the numbers of syllables in each utterance of TCP are fixed (4-syllabic, 5-syllabic, and 6-syllabic utterances), while those of TASC are varied, only TCP was used for discussion. Figure 5.4 shows the error rates of 4-syllabic, 5-syllabic and 6-syllabic utterances plotted separately according to syllable positions. It can be seen from the figure that the error rates at the beginning and ending of utterances

Table 5.3: Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of tone recognition with the refinements of contextual tone features and different intonation normalization methods on Thai proverb corpus (TPC) and Thai animal corpus (TASC).

Tone model	TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER
Simple	84.07	0.80	-	82.48	2.41	-
+CTF34	92.93	0.69	-	89.91	2.02	-
+CTF34+BIN	92.82	1.03	-1.65	90.38	1.35	4.65
+CTF34+CIN	**93.03**	0.94	1.42	**90.59**	1.31	6.71

Table 5.4: Measure of statistical difference of tone recognition with different intonation normalization methods. Significant differences are printed in **boldface**, while insignificant differences are shown in regular (based on a threshold of 0.05).

Tone model	TPC		TASC	
	+CTF34+BIN	+CTF34+CIN	+CTF34+BIN	+CTF34+CIN
+CTF34	0.7431	0.7792	0.2094	0.0628
+CTF34+BIN	-	0.4948	-	0.5508

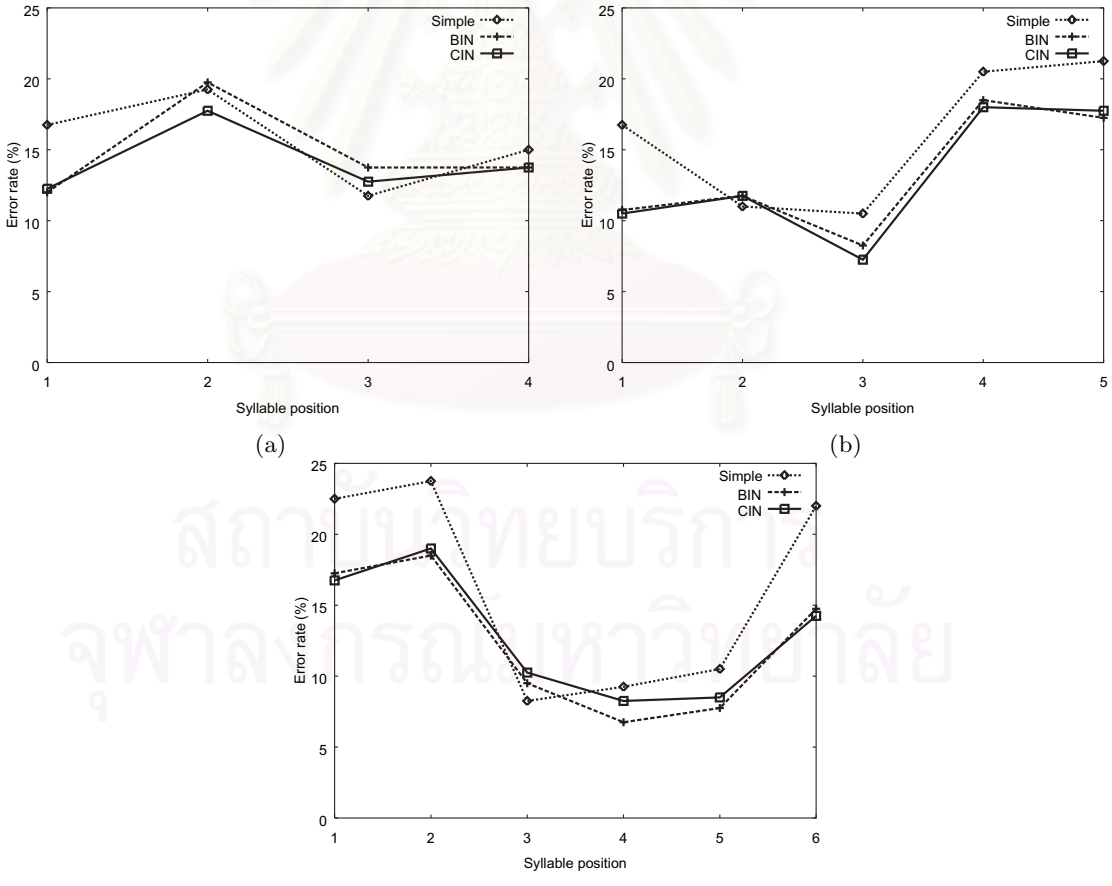


Figure 5.4: Recognition rates of (a) 4-syllabic utterances, (b) 5-syllabic utterances, and (c) 6-syllabic utterances for TPC, plotted separately according to syllable positions.

Table 5.5: Confusion matrices of tone recognition for (a) TPC and (b) TASC using CIN. M, L, F, H, and R denote the mid, the low, the fall, the high, and the rise, respectively.

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	1920	89.01	**1709**	84	56	55	16
L	1200	82.75	127	**993**	6	1	7
F	1280	**93.28**	52	5	**1194**	29	0
H	640	75.31	82	3	51	**482**	22
R	960	86.15	31	67	2	33	**827**
Total	6000	86.75					

(a) TPC

Reference	#Tokens	Recognition rate (%)	results (#tokens)				
			M	L	F	H	R
M	2380	90.46	**2153**	83	78	58	8
L	1260	80.08	189	**1009**	14	7	41
F	960	**91.04**	61	1	**874**	24	0
H	660	73.18	131	4	33	**483**	9
R	500	88.4	12	38	0	8	**442**
Total	5760	86.13					

(b) TASC

were dropped, especially for 6-syllabic utterances, when BIN or CIN was employed; while the error rates of the other syllable positions were consistent. This confirmed that the intonation normalization method can partially compensate the effect of the intonation contour.

We finally check the recognition rates for five tones. Table 5.5 shows the recognition rates of five tones for CIN. It is found that the recognition rates for the mid, the fall, and the rise are very good, but those for the high are still far below the average. The high is commonly misclassified as the mid. This mainly results from the number of syllables with the mid is very large (compared to the numbers of syllables with the other tones), and then the classifier may be biased.

5.5 Summary

In this chapter, we presented an empirical study to compensate intonation effect. We obtained two methods, i.e., beginning-point intonation normalization (BIN) and center-point intonation normalization (CIN) methods. These methods were evaluated on TPC and TASC using the basic tone recognition recognition framework. Both methods significantly increased recognition rates over the simple tone model. However, the recognition rates between these methods were not significantly different. The highest error reduction rates were 16.84% and 22.20% for TPC and TASC, respectively. We additionally applied the intonation normalization methods with the contextual tone

features. CIN slightly improved recognition rates for both corpora, while BIN slightly increased recognition rate for TASC only. The best error reduction rates were 1.42% and 6.71% for TPC and TASC, respectively. From error analysis, we found that the intonation normalization methods commonly increase recognition rates in the beginning and ending syllable of utterances.

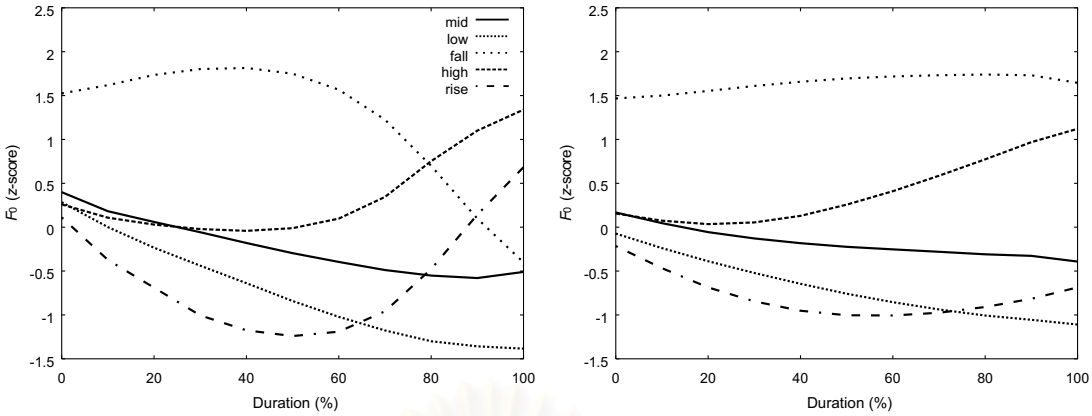
CHAPTER 6

EFFECT OF STRESS ON TONE RECOGNITION

Stress is an interacting effect on tone recognition. The F_0 contours of stressed syllables are generally quite different from unstressed ones (Chen and Wang 1995). For standard Thai, despite systematic changes in F_0 contours, all five tonal contrasts are preserved in unstressed as well as stressed syllables. However, F_0 contours of stressed syllables more closely approximate the contours in citation forms than those of unstressed syllables (Potisuk, Gandour, and Harper 1996; Potisuk et al. 1999). Figure 6.1 shows the mean of F_0 contours of all five tones in stressed and unstressed syllables. Data from Potisuk-1996 corpus (described in Subsection 6.1.3) were normalized within speaker, and across tones and stress categories. We observed that, in stressed syllables, the F_0 contours of each tone are quite different from each other. Thus, tones in stressed syllables are easy to recognize by the F_0 height and slope. In unstressed syllables, the shapes of F_0 contours among the five Thai tones are rather flat. The high and the rise with rising F_0 contours occur to be opposed to the other tones (the mid, the low, and the fall). The contours of unstressed syllables are preserved in the syntactic context and other factors such as speaking rate, coarticulation and declination. This makes the tone recognition in unstressed syllables to be a hard problem.

The goals of this chapter are to address: (i) which acoustic features should be used to discriminate stressed or unstressed syllable in running speech, and (ii) can such information improve tone recognition performance? To answer these questions, we first study the correlation of duration, energy, and F_0 measurements with lexical stress on pairs of ambiguous words and polysyllabic words, and identify the most informative features of stress by experiments. We then apply these features to compensate the stress effect on tone recognition. We explore two approaches: (a) building separate tone models for stressed and unstressed syllables, and (b) incorporating stress information into tone models.

In the following sections, we first present an empirical study of stress detection. The study consists of an accentual system of polysyllabic word, acoustic features of stress, speech corpora, two experiments and discussion. We then demonstrates an empirical study of tone recognition. We have proposed two methods, i.e., *separated stress*



(a) (b)

Figure 6.1: Mean F_0 contours of all five tones in (a) stressed and (b) unstressed syllables. Data were normalized within speaker, and across tones and stress categories.

method (SSM) and *incorporated stress features method* (ISFM) to alleviate the stress effect. After that, we describe an experiment by applying the methods to TPC and TASC. Finally, we summarize this chapter.

6.1 Stress Detection

In speech perception, stress refers to the relative perceptual prominence of a syllable or a word in a particular context (Ying 1998). Although stress can be reliably perceived by human listeners, its manifestations in the speech signal are very complex and depend on the language. Listeners perceive different levels of stress in an utterance through a complex combination of four features: *length*, *loudness*, *pitch*, and *quality of a segment*. These four perceptual features have four acoustic correlates: *duration*, *intensity*, *fundamental frequency*, and *formant structure*, respectively.

Many researchers have studied the acoustic correlates of stress in several languages (Potisuk, Gandour, and Harper 1996; Potisuk, Harper, and Gandour 1996; van Kuijk and Boves 1997; van Kuijk and Boves 1999; van Kuijk et al. 1996; Ying et al. 1996). The acoustic correlates of stress appear to vary from language to language (Potisuk, Gandour, and Harper 1996). Most researches have confirmed that duration is the most important acoustic correlate of stress. Lea (1980) found that, beside duration and energy, F_0 were also correlated with lexical stress in English. Some studies have also used spectral features, such as spectral change, measured as the average change of

spectral energy over the middle part of a syllable (Aull and Zue 1985; Waibel 1988); and spectral tilt, measured as spectral energy in various frequency sub-bands (Sluijter and van Heuven 1996; van Kuijk and Boves 1999).

The stress classification performance depends highly on the data, and also somewhat on the labelling of stress categories (Wang 2001). Early works on lexical stress modelling were for recognizing the complete stress patterns for isolated words (Aull and Zue 1985), and for distinguishing stress-minimal word pairs (Ying et al. 1996). These studies archived high accuracy. In (Waibel 1988), a variety of feature combinations were tested on four relatively small English speech databases, and the best average error rate of 12.44% was reported with energy integral, syllable duration, spectral change and F_0 maxima features. Jenkin and Scordilis (1996) performed an experiment on the English TIMIT database using various classifiers with six features from energy, duration and F_0. The stress labelling of the data was performed manually by two transcribers. The accuracy of about 80% was achieved. van Kuijk and Boves (1999) concentrated on a read telephone Dutch database. The stress labelling in this case was determined automatically from a dictionary. They achieved 72.6% of performance at best. Wang (2001) expected the classification accuracies to be higher if the stress labelling takes into consideration sentence-level effects, i.e., not all lexically stressed syllables are stressed (accented) in continuous speech (Shattuck-Hufnagel and Turk 1996).

In Thai, Potisuk, Gandour, and Harper (1996) investigated acoustic correlates of stress in Thai. Stimuli consisted of 25 pairs of sentences that the first member of each sentence pair contained a two-syllable noun-verb sequence exhibiting a strong-strong stress pattern, and the second member contained a two-syllable noun compound exhibiting a weak-strong stress pattern. The detail will be described in Subsection 6.1.3. They used five prosodic features, i.e., duration, average F_0, F_0 standard deviation, average intensity, and intensity standard deviation to discriminate stressed/unstressed syllables. Results indicated that duration is the predominant cue in signaling the distinction between stressed and unstressed syllables in Thai.

In the following subsections, we first present an accentual system of polysyllabic words in Thai. We then describe the acoustic features of stress used in this thesis. After that, two speech corpora, i.e., Potisuk-1996 corpus and Thai proper name corpus, are introduced. The corpora are used for evaluating these stress features. We then

demonstrate two experiments: (i) experiment of stress detection on pairs of ambiguous words and (ii) experiment of stress detection on polysyllabic words using PC-96 and TPNC, respectively. Finally, the discussion will be described in the last subsection.

6.1.1 Accentual system of polysyllabic words in Thai

Accent, which is a phonological term, is used to refer to potentiality of the syllable, or syllables in a word to be realized with stress either when the word occurs by itself as an utterance or with other words in an utterance (Luksaneeyanawin 1993). Stress is used to refer to the subjective complex of four objective phonetic features, i.e., pitch, loudness, length, and segmental qualities (Luksaneeyanawin 1993).

Thai is a fixed accent language. The accentual system of polysyllabic words is determined by the number of the component syllables in the word, and structure of its component syllable (Luksaneeyanawin 1983). The last syllable of the word is accented and always realized as a stressed syllable. Secondary accent is assigned according to the syllable structures of the component syllables. Secondary accents are realized as stressed syllables in formal speech, but as unstressed syllables in rapid conversational or casual speech.

Luksaneeyanawin (1983) proposed an accentual system of polysyllabic words in Thai. In the system, syllables are divided into two types, i.e., linker syllables (L) and non-linker syllables (O). Linker syllables are short syllables with an /a/ and end with a glottal stop, while non-linker syllables are other syllables besides linker syllables. The accentual system of the polysyllabic words in Thai is described in Figure 6.2.

6.1.2 Acoustic features of stress

Most studies have confirmed that duration is the most important acoustic correlate of stress. The duration can be calculated in a number of ways. For example, one could use the duration of a syllable, the duration of the vowel in the syllable or the duration of the rhyme in the syllable. Stress is assumed to be a feature of a syllable, but for practical purposes, stress can be attributed to the vowel (van Kuijk and Boves 1997). Many researches used vowel duration for representing stressed/unstressed syllables (van Kuijk and Boves 1997; van Kuijk and Boves 1999; van Kuijk et al. 1996; Ying et al. 1996) but some researchers proposed to use the rhyme duration for representing

1.Bisyllabic word			
L'L	L'O	'O'O	'O'L
2.Trisyllabic word			
('LL)'L	('LL)'O	('LO)'O	(L'O)'L
('OO)'O	('OO)'L	('OL)'L	('OL)'O
3.Tretrasyllabic word			
(L'O)O'O	(L'O)O'L	(L'O)L'O	(L'O)L'L
('OL)O'O	('OL)O'L	('OL)L'O	('OL)L'L
('LL)L'L	('LL)L'O	('LL)O'O	('LL)O'L
('OO)O'O	('OO)O'L	('OO)L'L	('OO)L'O

Figure 6.2: Accentual system of the polysyllabic words in Thai where ‘'’ is an accented maker that is in front of an accented syllable (Luksaneeyanawin 1983).

them (Sluijter and van Heuven 1995; Potisuk, Gandour, and Harper 1996; Potisuk, Harper, and Gandour 1996).

In this thesis, we use comprehensive sets of acoustic features (related to duration, energy and F_0) that have been reported to be related to stress. To find the best linguistic unit for stress detection in Thai, we employ all three units, i.e., syllable, vowel and rhyme units. The basic acoustic features are described in the following:

1. Duration. The duration is manually determined for each linguistic unit. Since the duration feature is highly context dependent (van Kuijk and Boves 1999), the duration of each unit in a syllable needs to be normalized. The duration are normalized by total duration of the unit in the same word. However, in the case of unknown word boundary, the duration are normalized by the z-score technique using the mean and standard deviation of all syllables of a speaker in a corpus (see Section 3.6). This feature is referred to as DURATION.

2. Energy. The energy of a particular speech frame is computed with a root mean square energy algorithm. Two energy features, i.e., the maximum energy (MAXENE) and the total energy (TOTENE) (van Kuijk and Boves 1999) are computed for each linguistic unit. MAXENE is defined as the log energy of the frame that has the highest energy value within each unit. The TOTENE of a unit is an integration of the energy over all frames within that unit. Therefore, TOTENE is implicitly sensitive to duration. Like duration, if we know the word boundary, the energy features are normalized by the sum of energy features of

all syllables in the same word. However, in the case of unknown word boundary, the z-score technique is used for normalization.

3. F_0. A raw F_0 is normalized by transforming the hertz values to a z-score within each speaker. The average of normalized F_0 of each linguistic unit is used as a stress feature. This feature is referred to as ZF.

We built five different stress feature sets. DURATION was used as the primary feature for stress detection. MAXENE, TOTENE and ZF were incorporated for generating these feature sets as follow:

1. SF1. DURATION
2. SF2. DURATION + TOTENE
3. SF3. DURATION + TOTENE + MAXENE
4. SF4. DURATION + TOTENE + ZF
5. SF5. DURATION + TOTENE + MAXENE + ZF

6.1.3 Speech corpora

We build two speech corpora, i.e., Potisuk-1996 corpus and Thai proper name corpus. The former is a corpus of pairs of ambiguous sentences, while the latter is a polysyllabic word corpus. The details are in the following.

Potisuk-1996 Corpus (PC-96)

The speech corpus is based on 25 pairs of ambiguous target sentences designed by Potisuk, Gandour, and Harper (1996). The two members of each pair contained six segmentally identical syllables including a two-syllable sequence that occurred at the beginning of the sentences. Vowel length was held constant within sentence pairs. Target syllables (i.e., the first syllable of the two-syllable sequence) in five sentence pairs contained short vowels; in the other 20 pairs, long vowels. Sentence pairs manifested one type of structural ambiguity. The first member contained a two syllable noun-verb sequence exhibiting a strong-strong stress pattern, the second member a two-syllable noun compound exhibiting a weak-strong stress pattern. To minimize tonal

coarticulation effects (Gandour et al. 1994), the two-syllable sequences were embedded at the beginning of the sentence, hence only anticipatory coarticulation on the first syllable was present while carry-over coarticulation was eliminated. The tones of the two-syllable sequences were also varied to represent all possible two-tone combinations of five Thai tones so that anticipatory coarticulation in all contexts was considered. Of the 25 two-tone combinations, only 4 were fully voiced throughout (MH, MR, LF, and FH); the other 21 two-tone combinations had intervening voiceless obstruents. We named this corpus "Potisuk-1996 corpus".

The data was collected from 9 native Thai speakers (three male and six female speakers), ranging in age from 20 to 27 years (mean=22.22 and s.d.=2.54). Each speaker read all pairs of sentences for five trials at a conversational speaking rate. Therefore, the corpus comprises 2,250 utterances. The speech signals were digitized by a 16-bit A/D converter of 11 kHz. These were manually segmented and transcribed as phonemes, rhymes and syllables using audio-visual cues from a waveform display.

Thai Proper Name Corpus (TPNC)

The Thai Proper Name Corpus (TPNC) is comprised of 100 Thai names: 3, 46, 45, and 6 of monosyllabic, bisyllabic, trisyllabic, and tetrasyllabic words, respectively. These names were separated into four sets with equal size of 25 names. The data collected from 40 native Thai speakers (20 male and 20 female speakers), ranging in age from 17 to 34 years (mean=21.60 and s.d.=3.58). Each speaker read two sets in one trial at a conversational speaking rate. The corpus therefore contains 2,000 word utterances. The speech signals were digitized by a 16-bit A/D converter of 11 kHz. They were manually segmented and transcribed as phonemes, rhymes and syllables using audio-visual cues from a waveform display. The stress categories were labelled using the mentioned accentual system (see Subsection 6.1.1).

6.1.4 Experiment I: Stress detection on pairs of ambiguous words

We performed several experiments on PC-96 with the different stress feature sets. All stress features were normalized by the z-score technique. To evaluate the stress feature sets on PC-96, the three-fold cross-validation approach (Dietterich 1997) was considered. Each fold contained the utterances collected from one male and two female

speakers. The experiments were run using a feedforward neural network. The network has three layers, i.e., input, hidden, and output layers. The number of input units depends on the number of stress features (1, 2, 3, 3, and 4 for SF1-SF5, respectively). The number of hidden and output units are 20 and 2, respectively. The tanh function was used as the activation in the network. Since the network learns more efficiently if the inputs are normalized to be symmetrical around 0 (Tebelskis 1995), all feature parameters were normalized to lie between -1.0 and 1.0. The network was trained by the error back-propagation method. Initial weights were set with random values between -1.0 and 1.0. The NICO (Neural Inference COmputation) toolkit [24] was used to build and train the network.

The recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of all stress feature sets are shown in Table 6.1. The vowel unit provides the best recognition rates (93.35% in average), while the syllable unit yields the worst (93.01% in average) for all stress feature sets. However, the results are not significantly different. TOTENE slightly increases recognition rates for all units. MAXENE also provides the slight improvement of recognition rates for the vowel and rhyme units, whereas ZF also slightly improves the recognition rates for the rhyme unit only.

Table 6.2 shows the confusion matrix of stress detection using SF5 with the rhyme unit. It is surprising that the recognition rate of unstressed syllables is higher than that of syllabled ones. The reason will be described in the following. However, both recognition rates are quite good (98.40% and 88.98% for unstressed and stressed syllables, respectively).

Figure 6.3 shows the histograms of four features for the two stress categories. Considering Figure 6.3 (a), we found that stressed syllables have longer duration than unstressed syllables. This is confirmed by TOTENE in Figure 6.3 (b) because this feature is implicitly sensitive to duration. Although there is good separation between the two categories, the small overlap is found due to intrinsic duration interferences. From observation, some stressed syllables have low durations and stay in the unstressed syllable area (see the small peaks of the stressed syllable lines in Figure 6.3 (a) and (b)). This seems to suggest that some stressed syllables are acoustically unstressed when judged by DURATION or TOTENE cues. This answers the recognition rate of unstressed syllables, which is higher than that of syllabled ones. In Figure 6.3 (c), MAXENE

Table 6.1: Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of stress detection for PC-96 with different linguistic units and stress feature sets. The best recognition rates for each unit are printed in **boldface**.

Tone feature	Syllable unit			Vowel unit			Rhyme unit		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
SF1. DURATION	92.58	2.87	-	**92.67**	2.04	-	**92.67**	2.04	-
SF2. DURATION + TOTENE	93.11	2.06	7.19	**93.42**	1.13	11.38	92.84	1.80	2.42
SF3. DURATION + TOTENE + MAXENE	93.07	1.65	6.59	**93.73**	0.82	15.57	93.29	1.56	8.48
SF4. DURATION + TOTENE + ZF	93.29	1.34	9.58	93.38	1.19	10.78	**93.47**	1.32	10.91
SF5. DURATION + TOTENE + MAXENE + ZF	93.02	2.08	5.99	93.64	1.10	14.37	**93.69**	1.09	13.94
Average	93.01	2.00	-	**93.35**	1.42	-	93.19	1.56	-

Table 6.2: Confusion matrix of stress detection for PC-96 using SF5 and the rhyme unit.

Reference	#Tokens	Recognition rate (%)	Recognition results	
			Stress	Unstress
Stress	1125	88.98	**1001**	124
Unstress	1125	98.40	18	**1107**
Total	2250	93.69		

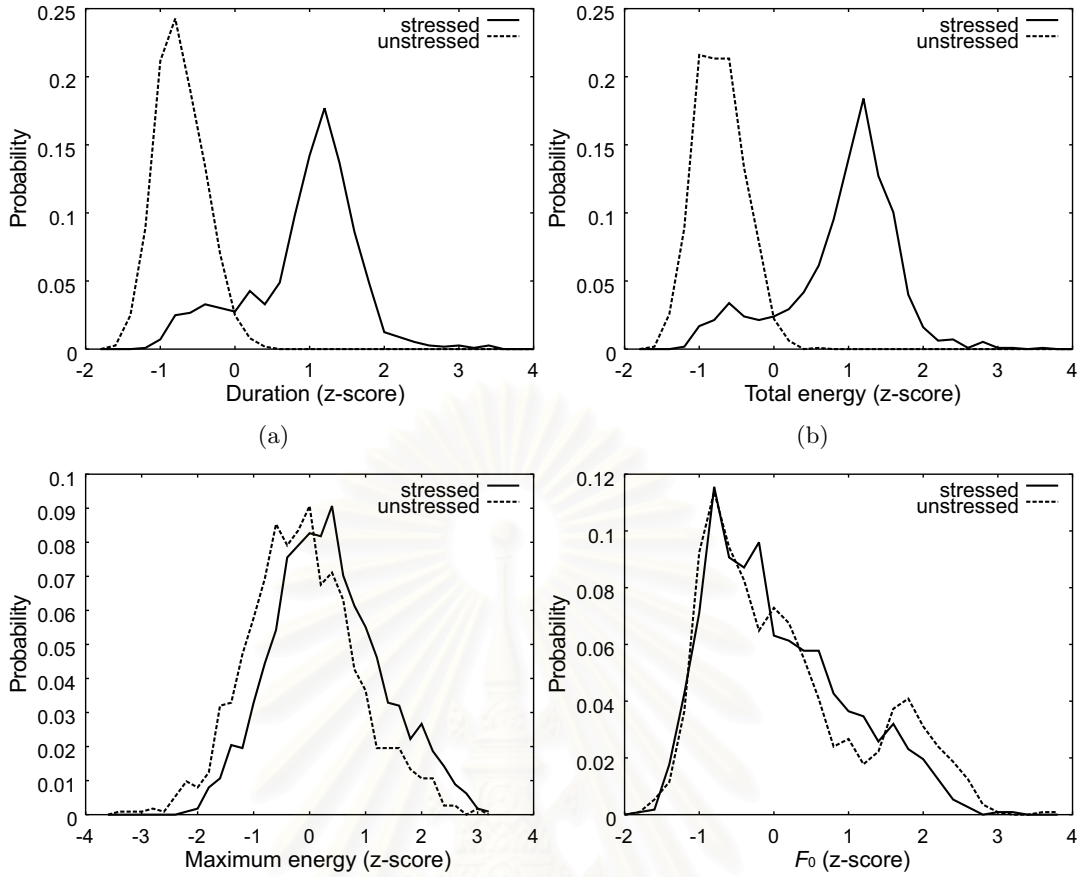


Figure 6.3: Histogram distribution of (a) duration, (b) total energy, (c) maximum energy, and (d) F_0, measured for the rhyme unit in PC-96.

shows very small differences among two stress categories. Contrary to some previous study (Potisuk, Gandour, and Harper 1996), ZF also shows very small differences among two stress categories (see Figure 6.3 (d)). The intrinsic F_0 contour for each tone still preserved its shape across stress categories. Potisuk, Gandour, and Harper (1996) successfully used the F_0 features (average F_0 and F_0 standard deviation) to discriminate stressed/unstressed syllables, because they separately measured and compared the F_0 features for each tone. They first separated all syllables into 5 groups depending on 5 tones. They then measured the F_0 features for each group and used the features for discriminating the stressed categories. However, we did not do that because we did not know the tone categories in advance. Therefore, the overlap of F_0 features is even more severe than that of duration and energy features.

6.1.5 Experiment II: Stress detection on polysyllabic words

Several experiments were performed on TPNC (monosyllabic words were excluded from the experiments) with the different stress feature sets using the five-fold cross-validation approach (Dietterich 1997). Each fold contained the utterances collected from four and four female speakers. As word boundaries were provided in TPNC, DURATION, TOTENE and MAXENE of a syllable were normalized by the sum of their feature values of all syllables in the same word, while ZF was normalized by the z-score tehcnique. The experiments were run using a feedforward neural network. The network configurations were similar to the previous experiment.

The recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of all stress feature sets are shown in Table 6.3. The rhyme unit provides the best recognition rates (83.65% in average) while the vowel unit yields the worst (74.74% in average) for all stress feature sets. TOTENE slightly increases recognition rates for the syllable and vowel units. MAXENE also provides the slight improvement of recognition rates for only the syllable unit. ZF also slightly improves the recognition rates for all three units.

Table 6.4 shows the confusion matrix of stress detection using SF5 and the rhyme unit. The recognition rate of stressed syllables is quite good (90.89%) while the rate of unstressed ones is not (67.22%). The reasons of this can be explained as follow. Firstly, in the recording process, the subjects spoke a word in their accentual styles. Thus, the accentual style of each speaker may differ from the standard accentual system as well as other speakers. Secondly, the number of unstressed syllables is small (compared to the number of stressed syllables) and so the classifier may be biased.

Figure 6.4 shows the recognition rates of stress detection along five stress feature sets with the rhyme unit, plotted separately by the number of syllables in a word. The bisyllabic words yield the highest results while the trisyllabic words do the worst. In the opposite way, the tetrasyllabic words achieve the highest improvement along the stress feature sets that reduced error from 18.96% to 17.08% whereas the bisyllabic words achieve the worst improvement that reduced error from 12.55% to 12.45%. This means that the energy and F_0 are not the prominent features for stress detection of bisyllabic words.

Table 6.3: Recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) of stress detection for TPNC using different linguistic units and stress feature sets. The best recognition rates for each unit are printed in **boldface**.

Stress feature set	Syllable unit			Vowel unit			Rhyme unit		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
SF1. DURATION	82.79	0.41	-	74.00	0.66	-	**83.49**	0.50	-
SF2. DURATION + TOTENE	82.89	0.16	0.58	74.30	0.50	1.15	**83.49**	0.41	0.00
SF3. DURATION + TOTENE + MAXENE	82.97	0.43	1.05	74.26	0.39	1.00	**83.49**	0.34	0.00
SF4. DURATION + TOTENE+ ZF	82.69	0.30	-0.58	75.70	0.31	6.54	**83.67**	0.42	1.09
SF5. DURATION + TOTENE + MAXENE + ZF	83.31	0.53	3.02	75.44	0.42	5.54	**84.10**	0.30	3.69
Average	82.93	0.37	-	74.74	0.46	-	**83.65**	0.39	-

Table 6.4: Confusion matrix of stress detection for TPC using SF5 and the rhyme unit.

Reference	#Tokens	Recognition rate (%)	Recognition results	
			Stress	Unstress
Stress	3580	90.89	**3254**	326
Unstress	1440	67.22	472	**968**
Total	5020	84.10		

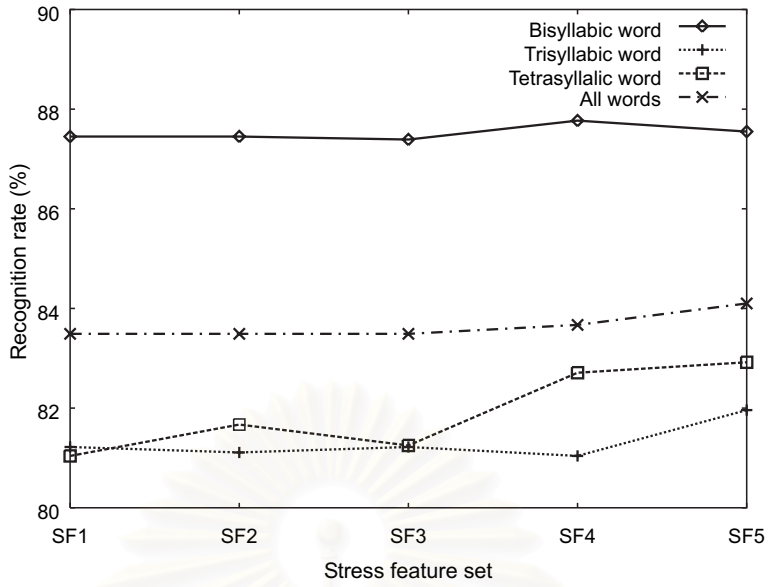


Figure 6.4: Recognition rates of stress detection for TPNC along five stress feature sets with the rhyme unit.

6.1.6 Discussion

This subsection describes some discussion about linguistic units (vowel, syllable, and rhyme) for stress detection. The syllable is a basic unit of human speech perception (Ohde and Sharf 1992) and the stress is an attribute of a syllable (Lehiste 1970). However, we have not found any work on stress detection using the syllable unit in the literature. The vowel unit was successfully used for stress detection in English (Wang 2001; Ying et al. 1996), Dutch (van Bergem 1993; van Kuijk et al. 1996; van Kuijk and Boves 1997; van Kuijk and Boves 1999); while the rhyme unit was used in Thai (Potisuk, Gandour, and Harper 1996; Potisuk, Harper, and Gandour 1996). The question of which linguistic unit should be used for general language is a hard problem because the unit selection depends on the syllable structure of each language.

From our experimental results for pairs of ambiguous words (see Table 6.1), the vowel unit provided the best average recognition rate but the results were not significantly different from those of the other two units. For polysyllabic words (see Table 6.3), the rhyme unit yielded better significantly recognition rate than the other units. These results support the works of (Potisuk, Gandour, and Harper 1996; Potisuk, Harper, and Gandour 1996), which demonstrate that the rhyme unit is a better correlate of stress in Thai than either the vowel or whole syllable unit. In Section 3.7, we described, in

phonology point of view, that the vowel unit is not suitable to recognize tones for Thai. Hence, we explored only the syllable and rhyme units for tone recognition experiments and found that the rhyme unit yielded better recognition rates than the syllable unit. In conclusion, we still believe that the rhyme unit is suitable for stress classification, tone recognition and speech recognition also.

6.2 Tone Recognition

This section presents experiments of Thai tone recognition. We first present two stress feature methods, i.e., separated stress method and incorporated stress feature method. Experiments with these methods are then demonstrated. The discussion finally will be described in the last subsection.

6.2.1 Stress feature methods

The F_0 contours of stressed syllables are generally quite different from unstressed ones. The most difference between the tone space for stressed and unstressed syllables is the reduction in excursion size of F_0 movements (Potisuk, Gandour, and Harper 1996) (see Figure 6.1). It makes the tone recognition of unstressed syllables to be a hard problem. In this thesis, we explore two approaches: (a) building separate tone models for stressed and unstressed syllables, and (b) incorporating stress information into tone models. Based on these approaches, we propose two methods to solve the stress effect as follow:

(1) Separated Stress Method (SSM)

All syllables are separated into stressed and unstressed groups using the stress detection algorithm as described in the previous section. Each group is trained and evaluated by its own classifier. The process is shown in Figure 6.5.

(2) Incorporated Stress Feature Method (ISFM)

The stress features (i.e., duration, energy, and F_0) are incorporated with tone features and are used together for training and evaluating. In the experiments, all stress feature sets (SF1 - SF5) determined within the rhyme unit were employed.

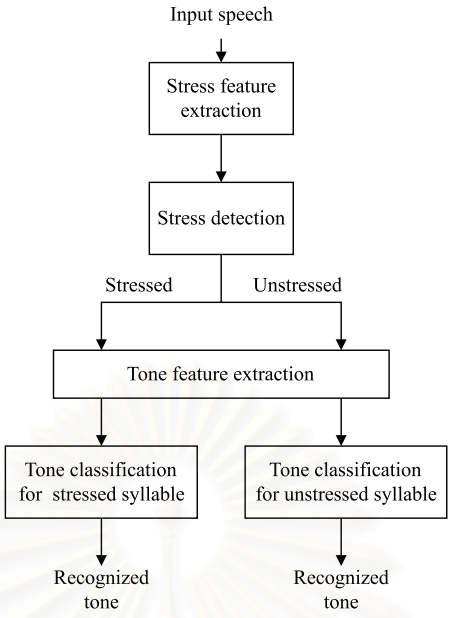


Figure 6.5: The separated stress method (SSM).

6.2.2 Experiments

A series of experiments with different tone feature sets were run on PC-96 and TPNC. Like the previous experiments, the three and five-fold cross-validation approach (Dietterich 1997) were used for PC-96 and TPNC, respectively. Each experiment was run using the basic tone recognition framework (described in Chapter 3).

For PC-96, the recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of all tone features are shown in Table 6.5. The results are reported separately by stressed, unstressed and total syllables. Comparing to the simple tone model, we found that the refined tone models with SSM and ISFM improve recognition rates for both stressed and unstressed syllables. The highest error reduction rates of 40.54%, 12.10%, and 18.61% are reported for stressed using ISFM (SF3), unstressed using SSM, and total syllables using ISFM (SF1), respectively.

In Table 6.6, the confusion matrices of tone recognition using ISFM (SF5) are shown. It can be seen that the fall provides the highest recognition rate for both stressed and unstressed syllables. This is similar to the previous experiments. However, it is surprising that the rise yields the poorest results for both syllabled and unstressed syllables. The rise are mostly misclassified as the high and the low for stressed and

Table 6.5: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for PC-96 using different tone feature sets. The highest recognition rates for stressed, unstressed and total syllables are printed in **boldface**.

Tone model	Stress			Unstress			Total		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
Simple	93.42	1.42	-	86.04	2.51	-	89.73	1.73	-
+ SSM	93.51	2.20	1.35	**87.73**	1.36	12.10	90.62	2.17	8.66
+ ISFM (SF1)	95.91	0.88	37.84	87.38	1.26	9.55	**91.64**	1.24	18.61
+ ISFM (SF2)	**96.09**	0.82	40.54	86.58	0.91	3.82	91.33	1.06	15.58
+ ISFM (SF3)	95.56	0.67	32.43	87.02	1.53	7.01	91.29	1.26	15.15
+ ISFM (SF4)	95.73	1.15	35.14	85.96	1.12	-0.64	90.84	1.39	10.82
+ ISFM (SF5)	95.02	1.20	24.32	86.13	1.89	0.64	90.58	1.61	8.23

Table 6.6: Confusion matrices of tone recognition of (a) stressed and (b) unstressed syllables for PC-96 using baseline+ISFM (SF5).

Reference	#Tokens	Recognition rate (%)	Recognition results (#tokens)				
			M	L	F	H	R
M	225	95.11	**214**	10	1	0	0
L	225	93.33	10	**210**	1	0	4
F	225	**99.56**	1	0	**224**	0	0
H	225	94.67	2	0	3	**213**	7
R	225	92.44	1	2	0	14	**208**
Total	1125	95.02					

(a)

Reference	#Tokens	Recognition rate (%)	Recognition results (#tokens)				
			M	L	F	H	R
M	225	81.78	**184**	16	3	11	11
L	225	85.78	15	**193**	0	0	17
F	225	**98.22**	0	0	**221**	4	0
H	225	91.56	9	0	4	**206**	6
R	225	73.33	18	34	0	8	**165**
Total	1125	86.13					

(b)

Table 6.7: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for TPNC using different tone feature sets. The highest recognition rates for stressed, unstressed and total syllables are printed in **boldface**.

Tone model	Stress			Unstress			Total		
	%	s.d.	%ER	%	s.d.	%ER	%	s.d.	%ER
Simple	86.45	2.29	-	64.03	2.92	-	80.02	2.47	-
+ SSM	87.29	1.37	6.19	74.58	3.34	29.34	83.65	1.83	18.15
+ ISFM (SF1)	88.60	1.62	15.88	74.38	3.21	28.76	84.52	1.95	22.53
+ ISFM (SF2)	88.85	1.61	17.73	74.38	2.71	28.76	84.70	1.90	23.43
+ ISFM (SF3)	88.94	1.58	18.35	75.49	2.82	31.85	85.08	1.85	25.32
+ ISFM (SF4)	**92.15**	0.95	42.06	77.29	2.17	36.87	**87.89**	1.01	39.38
+ ISFM (SF5)	91.62	0.89	38.14	**77.57**	2.12	37.64	87.59	0.97	37.89

Table 6.8: Confusion matrices of tone recognition of (a) stressed and (b) unstressed syllables for TPNC using ISFM (SF5).

Reference	#Tokens	Recognition rate (%)	Recognition results (#tokens)				
			M	L	F	H	R
M	2380	**94.87**	**2258**	58	5	49	10
L	420	86.90	40	**365**	2	3	10
F	80	82.50	13	0	**66**	1	0
H	400	80.25	58	19	0	**321**	2
R	300	90.00	3	20	0	7	**270**
Total	3580	91.62					

(a)

Reference	#Tokens	Recognition rate (%)	Recognition results (#tokens)				
			M	L	F	H	R
M	200	82.50	**165**	29	0	5	1
L	780	73.72	114	**575**	2	73	16
F	20	55.00	6	0	**11**	3	0
H	440	**83.18**	17	56	1	**366**	0
R	0	-	0	0	0	0	**0**
Total	1440	77.57					

(b)

unstressed syllables, respectively.

For TPNC, the recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of all tone features are shown in Table 6.7. The results are reported separately by stressed, unstressed and total syllables. Compared to the simple tone model, SSM and ISFM increase recognition rates for both stressed and unstressed syllables. The highest error reduction rates of 42.06%, 37.64%, and 39.38% are for stressed, unstressed, and total syllables, repectively. These results are reported when using ISFM (SF4), ISFM (SF5), and ISFM (SF4), respectively.

In Table 6.8, the confusion matrices of tone recognition using ISFM (SF5) are shown. It can be seen that the mid and the high provide the highest recognition rates for stressed and unstressed syllables, respectively. The results of stressed syllables are different from the previous studies (Luksaneeyanawin 1995; Thubthong et al. 2000a) that often reported low recognition rates for the mid. This is because of the similarity of the F_0 contours of the low to the mid, and so the most errors probably came from the misclassification between them. Only explicit reason making the best recognition rate for the mid is that the number of syllables with the mid is very large (compared to the number of syllables with the other tones) and thus the classifier may be biased. Considering the confusion matrix of unstressed syllables (see Table 6.8 (b)), the probable errors come from the misclassification between the low and the mid that are similar to the previous studies. Moreover, the probable errors also come from the misclassification between the low and the high. These errors frequently occur by linker syllables that speakers often speak confusedly between the low and the high. However, listeners can correctly recognize them. For example, even if a speaker wrongly says /nát tá kɔ̄ːn/ having a different tone (the high instead of the low) to the correct one /nát tà kɔ̄ːn/ ("source of sage"), the listener can correctly recognize it.

6.2.3 Discussion

This subsection will address the question of which method (SSM or ISFM) should be used for the further experiments. From the experimental results, we found that both SSM and ISFM outperform the simple tone model and ISFM provides better recognition rates than SSM. The advantage of SSM is that it breaks a data set into two small data sets. This reduces complexity of the search space during training process. The benefit

Table 6.9: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of tone recognition for TPC and TASC using different tone feature sets. The best recognition rate for each corpus is printed in **boldface**.

Tone model	TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER
Simple	84.07	0.80	-	82.48	2.41	-
+ ISFM (SF1)	88.17	1.24	25.73	86.27	2.59	21.61
+ ISFM (SF2)	88.15	1.20	25.63	86.77	2.34	24.48
+ ISFM (SF3)	88.47	0.93	27.62	83.92	2.77	8.23
+ ISFM (SF4)	88.42	0.84	27.30	86.65	2.07	23.79
+ ISFM (SF5)	**89.23**	0.61	32.43	**87.24**	1.99	27.16

is prominent when the data set is too big. However, SSM needs three networks and spends three training times. On the contrary, ISFM uses one network for training and testing and spend one training time. The prominent advantage of ISFM is that it does not need stressed/unstressed transcription. This is convenient for us to apply it to the other corpora (TPC and TASC), because these corpora have no stressed/unstressed labels. We therefore decide to use ISFM for the next section.

6.3 Experiments of Tone Recognition on Continuous Speech

In this section, we performed two additional experiments by applying a stress method on Thai proverb corpus (TPC) and Thai animal story corpus (TASC). We decided to use ISFM, because TPC and TASC were not yet labelled as stressed or unstressed. The basic tone recognition framework was used. The recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) are shown in Table 6.9. Compared to the results of the simple tone model, better results are achieved for all stress feature sets embedded in ISFM. ISFM using SF5 provides the highest error reduction rates of 32.43% and 27.16% for both TPC and TASC, respectively.

We then analyzed the tone error rates by plotting the error rates of the simple tone models and the refined tone models with ISMF (SF5) along different rhyme durations. As shown in Figure 6.6, we found that ISMF decreases error rates for all syllables along different rhyme durations. The extreme error reductions are reported for the syllables having rhyme duration below 100 ms. This confirmed the usefulness of our features on a very short syllable (such as neutral or unstressed syllable).

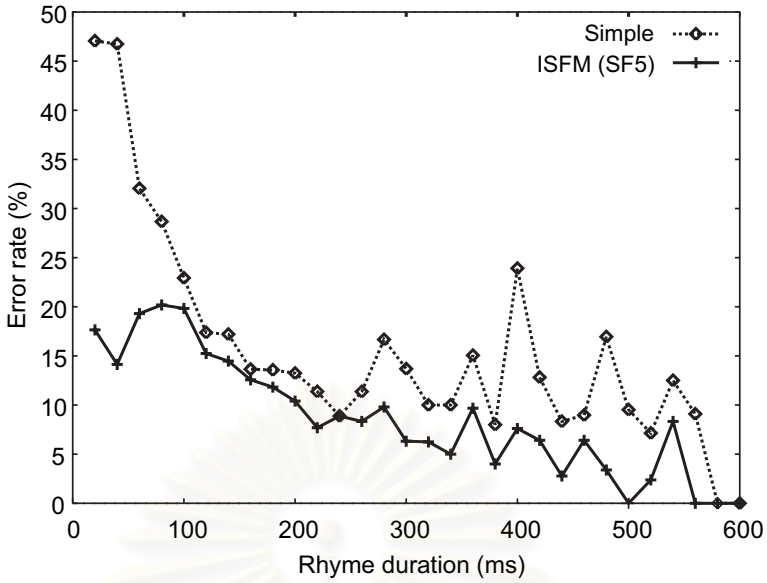


Figure 6.6: Error rates (%) of tone recognition along different rhyme durations for TASC.

6.4 Summary

In this chapter, we have demonstrated two empirical studies, i.e., stress detection and tone recognition. For stress detection, we conducted two experiments for pairs of ambiguous words and polysyllabic words. The former and the latter were performed on Potisuk-1996 corpus (PC-96) and Thai proper name corpus (TPNC), respectively. The acoustic features, i.e., duration, energy, and F_0, were considered. These features were extracted from several linguistic units, i.e., syllable, vowel and rhyme units, for comparison. A feedforward neural network was employed to evaluate the experiments. The experimental results show that the vowel unit provides the best average recognition rate for PC-96. However, the results of the vowel unit are not significantly different form those of the other units. For TPNC, the rhyme unit yields better recognition rate than the other units. We conclude that the rhyme unit outperforms the other units for stress detection.

For tone recognition, the stress effect have been considered. We have proposed two methods, i.e., *separated stress method* (SSM) and *incorporated stress feature method* (ISFM) to compensate this effect. To evaluate these methods, the basic tone recognition framework was employed. The experimental results show that both methods increase the tone recognition rates. The best recognition rate of 91.64% and 87.89% are achieved

for PC-96 and TPNC, respectively, when ISFM was applied.

We additionally employed ISFM for Thai proverb corpus (TPC) and Thai animal story corpus (TASC) using the basic tone recognition framework. We found that ISFM (SF5) yields higher recognition rates than the simple tone model. The maximum error reduction rates are 32.43% and 27.16% for TPC and TASC, respectively. We then analyzed the tone error rates by considering the rhyme duration of a syllable. We found that ISFM extremely reduces the error rates for the short syllables having the rhyme duration below 100 ms.

CHAPTER 7

COMPARISON OF SEVERAL REFINED TONE MODELS AND INCORPORATION OF TONE MODELS INTO SPEECH RECOGNITION

In Chapter 3, we first developed a tone recognition framework, which used the set of five normalized F_0's and their slopes in ERB-rate scale at 0%, 25%, 50%, 75%, and 100% of the rhyme segment as the simple tone model to parameterize F_0 contour and a three-layer feedforward neural network to evaluate the performance. We then tried to account for the tonal variation factors in tone modelling for improving tone recognition performance. Using this basic framework, we demonstrated that tone recognition performance of Thai continuous speech can be significantly improved by taking into account tone coarticulation, sentence declination, or stressed/unstressed syllables (Chapter 4, 5 and 6, respectively).

In this chapter, we first demonstrate the performance comparison of several refined tone models for accounting several interacting factors. We then present an empirical study for integrating the tone models into speech recognition to enhance speech recognition performance.

7.1 Comparison in Performance of Several Refined Tone Models

This section demonstrates the study of our tone recognition framework when several refined tone models were applied. The study was performed on Thai proper name corpus (TPC) and Thai animal story corpus (TASC).

The results are summarized in Table 7.1. The refinements to the tone models achieve significant recognition rate improvements for both corpora. Comparing the refinements with the contexture tone features (CTF34), the center intonation normalization (CIN), and the incorporated stress feature method (ISFM), we found that CTF34 yields the highest error reduction rates, while CIN provides the poorest error reduction rates. We then used the CTF34 refinement as the primary tone models. CIN and ISFM were then incorporated for enhancing the tone recognition rates. CIN slightly increases the recognition rates from 92.93% to 93.03%, and from 89.91% to 90.59% for TPC and TASC, respectively. The additional refined tone model with ISFM also improves recog-

Table 7.1: Tone recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of the simple tone model and more refined tone models on TPC and TASC.

Tone model	TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER
Simple	84.07	0.80	-	82.48	2.41	-
+ CTF34	92.93	0.69	55.65	89.91	2.02	42.42
+ CIN	86.75	1.04	16.84	86.13	2.23	20.81
+ ISFM (SF5)	89.23	0.61	32.43	87.24	1.99	27.16
+ CTF34 + CIN	93.03	0.94	56.28	90.59	1.31	46.28
+ CTF34 + CIN + ISFM (SF5)	**93.60**	1.05	59.83	**92.67**	1.16	58.18

nition rates. The over all highest recognition rates are 93.60% and 92.67% for TPC and TASC, respectively.

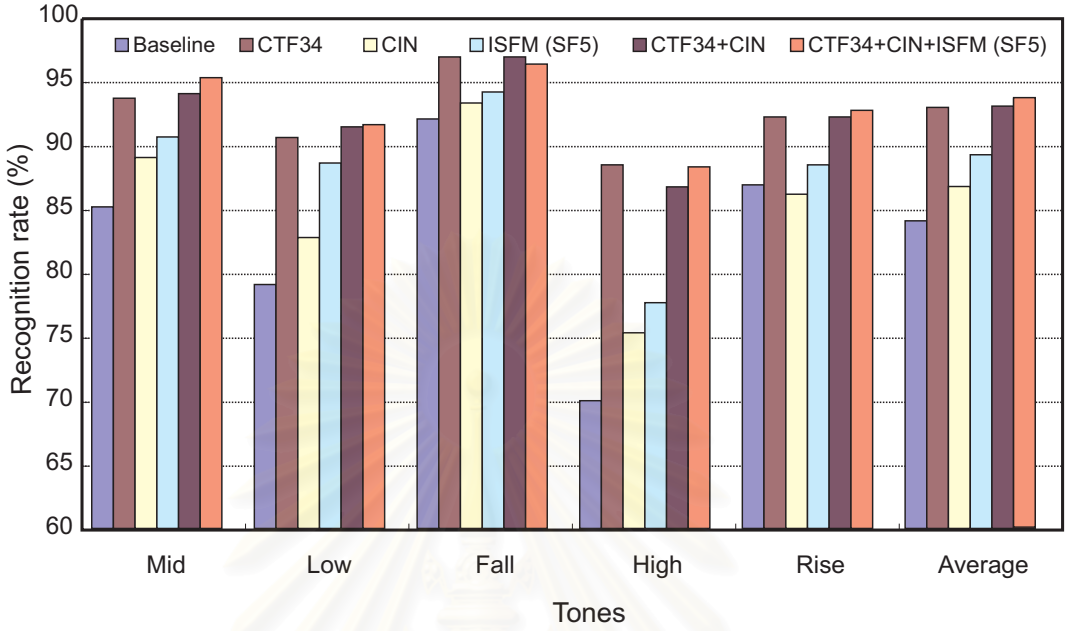
Figure 7.1 shows the tone recognition rates of all five tones using several refined tone models for TPC and TASC. The results for both corpora are in the same tendency. All refined tone models improve recognition rates for all tones, compared to the simple tone model. The refinement of CTF34+CIN+ISFM (SF5) provides the highest recognition rates for most tones.

7.2 Incorporation of Tone Models into Speech Recognition

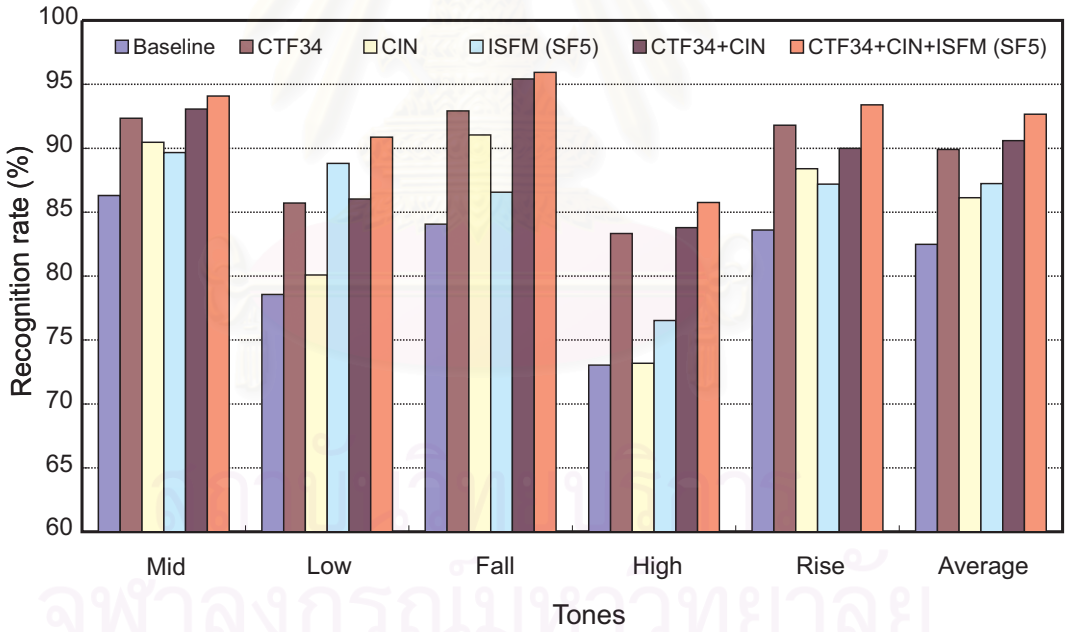
The goal of this section is to integrate the tone models into a syllable-based speech recognition for increasing the speech recognition performance. We first describe the basic syllable-based speech recognition framework. It includes feature selection, normalization, and a neural network classifier. Then, we present the method for integrating tone models into the framework. Finally, we discuss the experimental results.

7.2.1 Syllable-based speech recognition framework

For at least a decade, the phoneme-based methods have been the dominant methods for modelling acoustics in speech recognition. Since a phoneme unit spans an extremely short time interval, integration of spectral and temporal dependencies is not easy (Ganapathiraju et al. 2001). Therefore, we have shifted our focus to a larger acoustic context. The syllable is an attractive unit for recognition for several reasons: (i) the syllable lies in close connection to human speech perception and articulation, its integration of some coarticulation phenomena, and the potential for a relatively compact representation of conversational speech (Ganapathiraju et al. 2001), (ii) the relative duration of syllables is less dependent on variations in speaking rate than the



(a) TPC



(b) TASC

Figure 7.1: Tone recognition rates of all five tones using several refined tone models for (a) TPC and (b) TASC.

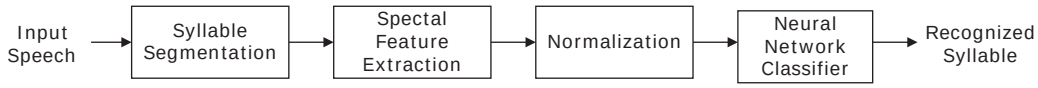


Figure 7.2: Architecture of the syllable-based speech recognition framework.

relative durations of phonemes (Greenberg 1996), (iii) syllables appear to offer a natural interface between speech acoustics and lexical access (Wu 1998; Wu et al. 1998), and (iv) syllables constitute a convenient framework for incorporating prosodic information into recognition (Wu 1998; Wu et al. 1998), particularly for tone language such as Thai. We believe that for Thai language a syllable-based speech recognition system is robust and more suitable than a phoneme-based one. There are a number of researches using syllable-based speech recognition approach in Thai speech (Demeechai and Mäkeläinen 2001; Thubthong and Kijsirikul 1999a; Thubthong and Kijsirikul 2000a).

In this thesis, we have used a syllable-based speech recognition as the basic speech recognition framework. Figure 7.2 shows the architecture of the framework. The architecture starts through the syllable segmentation, and then proceeds by extracting spectral feature vectors. The feature vectors are then normalized to increase learning efficiency. Finally, the normalized feature vectors are fed into a neural network classifier to recognize the syllable. The details are discussed in the following.

Feature extraction

Since not all syllables are of equal duration, we extracted spectral features from a number of frames at the time points between 5 to 95% of duration with the equal step size. In this thesis, we used 15 frames (see Figure 7.3). For each frame, the 12^{th} order of RASTA coefficients (Hermansky and Morgan 1994) computed within a Hamming-windowed 25 ms frame were used. Therefore, a syllable is represented by 180 feature parameters.

Normalization

Since a neural network learns more efficiently if the inputs are normalized to be symmetrical around 0 (Tebelskis 1995), all feature parameters are normalized to lie between -1.0 and 1.0 using the following equation:

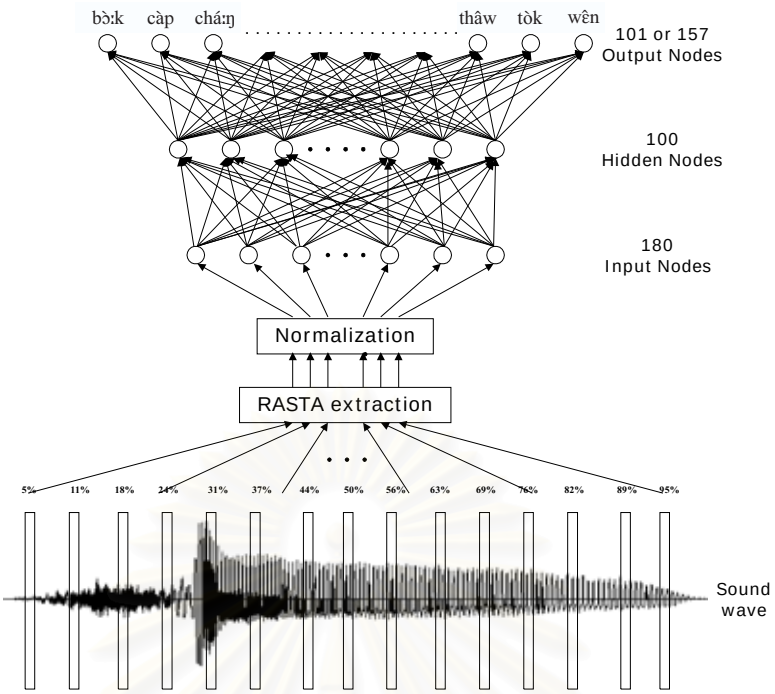


Figure 7.3: A syllable-based speech recognition framework based on a three-layer feed-forward neural network. Each syllable is represented by 15 frame features of RASTA coefficients.

$$normF_i = 2.0 \times (\frac{F_i - minF_i}{maxF_i - minF_i}) - 1.0 \tag{7.1}$$

where F_i is the i^{th} feature under consideration, $minF_i$ and $maxF_i$ are the minimum and maximum values of F_i, and $normF_i$ is the normalized value of F_i. $MinF_i$ and $maxF_i$ are obtained from the 5^{th} and 95^{th} percentiles of a histogram of F_i generated on the training data (see Figrue 7.4) (Muthusamy 1993).

Neural network classifier

A three-layer feedforward network was employed. The network had an input layer of several units depending on the number of the input features, a hidden layer of 100 units, and an output layer of several units corresponding to the number of different syllable classes. The tanh function was used as the activation function in the neural network. To train a classifier, we repetitively presented to the input units the feature vector of each syllable derived from the training utterances. The connection weights between neurons were then adjusted to reduce the error between the actual output

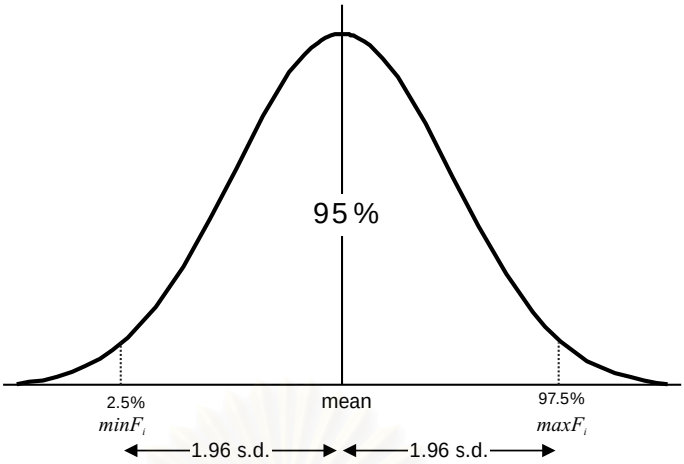


Figure 7.4: Normal distribution and the position of $minF_i$ and $maxF_i$.

pattern and the desired output pattern. We trained the network by the standard back-propagation algorithm for a maximum of 300 epochs with 0.0001 learning rate and 0.9 momentum. To test the performance of a properly trained classifier, the feature vector of a syllable was presented to the classifier and the output pattern was generated using the weights received from the training process. If the i^{th} output node exhibits the largest activation level (0.0-1.0), the syllable will be recognized to carry a syllable i. The NICO (Neural Inference Computation) toolkit (Ström 1997b) was applied to perform the experiment.

7.2.2 Integrating tone models into speech recognition

The goal of this study is to measure the performance improvement of tone modelling for speech recognition. In literature, there are two different approaches for recognition of tonal syllables. One is to build separate models for base syllables[1] and tones. This approach trains base syllable models and tone models separately. The output of the base syllable recognizer and the tone recognizer are combined to produce the final recognition result (Chen and Liao 1998; Lee and Ching 1999). The other approach is to augment the acoustic features with tone features, and build tone-dependent syllable models (Huang and Seide 2000; Thubthong and Kijsirikul 1999a; Thubthong and Kijsirikul 2000a; Wang 2001). This approach dose not need the final decision. The output

[1] A base syllable is a syllable disregarded tone.

of the syllable recognizer is the recognition result. The advantage of this approach is that we can easily incorporate the tone models into any existing recognizer without alternating its structure. However, this approach has the disadvantage of splitting the training data, i.e., the same tone from different rhymes as well as the same rhyme with different tones cannot be shared in training (Wang 2001). This approach is useful for a small vocabulary speech recognition system only.

In this thesis, we tried the second approach, because it is easy to implement and the numbers of output classes in our testing corpora are quite small (101 and 157 classes for TPC and TASC).

7.2.3 Experiments

We conducted speech recognition experiments on TPC and TASC. The five-fold cross-validation approach (Dietterich 1997) was used. The numbers of different syllables (classes) are 101 and 157 syllables for TPC and TASC. The recognition rates (%), standard deviations (s.d.) and error reduction rates (%ER) are shown in Table 7.2. The results for both corpora are in the same ways. The baseline features (RASTA oder 12) achieves 91.02% and 86.22% of recognition rates for TPC and TASC, respectively. These results were used as the baseline results. After incorporating the stress feature 5 (SF5), the error reduction rates of 17.63% and 16.75% are achieved for TPC and TASC, respectively. The use of the simple tone model greatly increases the syllable recognition rates from the baseline results. The error reduction rates are about 45% for both corpora. Compared to the simple tone model, the refined tone models further reduce the syllable error rates. The maximum error reduction rates of 85.16% (for TPC) and 75.06% (for TASC) are reported using the refined tone models of CTF34 + CIN + ISFM (SF5).

7.3 Summary

In this chapter, we first demonstrated the performance comparison of several refined tone models that accounted different interacting factors. Using the tone recognition framework, we conducted experiments on Thai proper name (TPC) and Thai animal story corpus (TASC). All refined tone models achieved significant recognition rate improvements for both corpora, compared to the simple tone model. The model

Table 7.2: Recognition rates (%), standard deviations (s.d.), and error reduction rates (%ER) of syllable-based speech recognition for TPC and TASC when incorporating stress feature (SF5) and several refined tone models.

Tone model	TPC			TASC		
	%	s.d.	%ER	%	s.d.	%ER
RASTA12	91.02	1.43	-	86.22	3.34	-
+ SF5	92.60	0.70	17.63	88.52	3.06	16.75
+ Simple	95.07	0.62	45.08	92.43	2.19	45.09
+ Simple + CTF34	98.38	0.14	82.00	96.20	1.36	72.42
+ Simple + CIN	95.15	0.54	46.01	92.45	2.16	45.21
+ Simple + ISFM (SF5)	96.15	0.69	57.14	93.32	2.06	51.51
+ Simple + CTF34 + CIN + ISFM (SF5)	**98.67**	0.11	85.16	**96.56**	1.20	75.06

using of all tone features (Simple+CTF34+CIN+ISF(SF5)) provided the highest recognition rates of 93.60% and 92.67% for TPC and TASC, respectively.

We then presented an empirical study for integrating several refined tone models into speech recognition to enhance speech recognition performance. The syllable-based speech recognition was used as the basic speech recognition framework. The 12^{th} order of RASTA coefficients extracted from 15 frames of a syllable were used as input features and fed into a three-layer feedforward neural network for training and testing. The refined tone models were also combined with spectral features together and fed into the network for enhancing speech recognition rates. The experiments were conducted on TPC and TASC. The experimental results showed that the integrating of the refined tone models highly improved speech recognition rates for both corpora. The best error reduction rates of 85.16% and 75.06% were achieved for TPC and TASC, respectively.

CHAPTER 8

SUMMARY AND FUTURE WORKS

8.1 Summary and Contributions

Tone information is very important to speech recognition in a tone language such as Thai. The tones produced on isolation words are very easy to identify. However, tones produced on words in continuous speech are much more difficult to identify. Several interacting factors affect F_0 realization of tones, e.g., syllable structure, coarticulation, intonation, stress, speaking rate, dialect, sex, age, and emotion.

In this thesis, we have studied several linguistic effects on tone recognition in Thai continuous speech. We focus on the effects of syllable structure, coarticulation, intonation, and stress. In the following, we briefly recapitulate the methodologies and main results of our explorations, followed by a summary of the main contributions of this thesis.

Effect of Syllable Structure

We performed empirical studies of the effect of initial consonants, vowels, and final consonants on tone recognition (Chapter 2). We explored three tone feature sets used to capture the characteristics of Thai tones. The first two tone feature sets come from the previous works (Thubthong 1995; Tungthangthum 1998) and the last one is a novel tone feature set designed from our observation on the F_0 contour patterns. We built three data sets of isolated syllables. Each data set was used to study the effect of each phoneme. Several experiments have been conducted using a three-layer feedforward neural network. The experimental results showed that the proposed tone feature set yielded the best performance for most experiments. Although the proposed feature set is useful for the dependent data experiments, it is not robust for independent data experiments. Therefore, a tone recognition system for all available syllables needs the training examples that cover all combinations between tones and the other syllable types, and the size of the training set must be large enough. The analysis of results implied that there were some correlations between tones and the phonematic units constructing the syllables. Human perception test was then employed to judge the recognition rate. The recognition rate of human perception test was much lower than

that of the machine. This suggests that humans are not good at recognizing meaningless words, and words without context. The basic unit for human recognition is a word, not a phoneme. Furthermore, the combination of neural networks trained on different tone feature sets was studied. We explored several classifier combination schemes to enhance the recognition rates. The experimental results demonstrated that the neural network combination was superior to a single network.

Constructing Tone Recognition Framework for Thai Continuous Speech

Before studying the other effects, we first developed a basic tone recognition framework for Thai continuous speech. The framework consisted of tone models used to parameterize tone F_0 contour and a classifier used to evaluate the performance of the tone models. The former was designed by an empirical study, while, for the latter, a three-layer feedforward neural network was applied. To construct the tone models, we concentrated on the question of which configurations with respect to tone features, frequency scale, normalization technique, and tone critical segment should be used for tone recognition (Chapter 3). To address the questions, we conducted four experiments according to the four configurations in the question. Three corpora, i.e., Potisuk-1999 Corpus (PC-99), Thai Proverb Corpus (TPC), Thai Animal Story Corpus (TASC), were built and used in the studies. In the first experiment, we explored three tone feature sets and found that the tone feature set containing five F_0's and their slopes at 0%, 25%, 50%, 75%, and 100% of the rhyme provided the highest recognition rates. We then conducted the second experiment, respect to the frequency scales. We tried three scales, i.e., hertz, semitone, and ERB-rate. The experimental results showed that ERB-rate scale outperformed semi-tone and hertz scales. In the third experiment, we concentrated on different normalization techniques. We found that the z-score normalization provided the highest recognition rates. Finally, we focused on the tone critical segment. We explored two segments, i.e., rhyme and syllable. The experiments using the rhyme segment archived better recognition rates than those using the syllable one. These configurations providing highest performance were used to construct the tone model referred to as the *simple tone model*.

Effect of Coarticulation

We considered the coarticulatory effects on tone recognition (Chapter 4). These effects were from the neighboring syllables. The effects of the following syllable and the preceding syllable are called *anticipatory coarticulation* and *carry-over coarticulation*, respectively. Using the basic tone recognition framework (described in Chapter 3), we focussed on tone modelling and took into account tone coarticulation aspects for improving the simple tone model. We have proposed a feature set called "*contextual tone features*" that captured the F_0 realizations of the neighboring syllables. We explored several configurations of contextual tone features (CTFs) depending to the numbers and positions of the extracted features (F_0's and their slopes) of a syllable. We performed experiments on PC-99, TPC, and TASC using the context-independent tone model (CI-T-5). The experimental results showed that all CTF's improved recognition rates for all corpora. CTF34 provided the best recognition rates of 94.27%, 92.93% and 89.91% for PC-99, TPC, and TASC, respectively. These results confirmed the study of (Gandour et al. 1994) that carry-over effects extend forward to about 75% of the duration of the following syllable, while anticipatory effects extend backward to about 50% of the duration of the preceding syllable.

Furthermore, we applied the context-dependent tone model (CD-T-175) to enhance tone recognition rate. CD-T-175 provided better recognition rates than CI-T-5 for all corpora. However, the training times for CD-T175 was very long. Therefore, we have also proposed a novel model called *half-tone model* (H-T-30) to alleviate the drawback of CD-T-175. We found that H-T-30 also increased recognition rates over CI-T-5 for all corpora. Considering the best recognition rates, we found that CD-T175 archived the highest recognition rates for TPC and TASC, while H-T-30 yielded the best for PC-99. However, H-T-30 was better than CD-T-175 in term of speed. The recognition rates of H-T-30 were slightly higher than those of CD-T-175 for PC-99 but slightly lower than those of CD-T-175 for TPC and TASC. This concluded that H-T-30 is a promising model. The model is the best choice when we want to optimize both recognition rates and training time.

Effect of Intonation

Intonation is an important effect on tone recognition. There are three intonation contours in Thai (Luksaneeyanawin 1993): the Fall, the Rise, and the Convolution. Due

to the lack of speech database, which has sufficient and accurate annotation of several intonation contours, we focused on the Fall only (Chapter 5). The Fall can be referred to as "declination" defined as the tendency of F_0 to gradually decline over the course on an utterance. We performed an empirical study to compensate the declination effect. The motivation is that the local F_0 will be adjusted to conform to the intonation pattern of the sentence; thus, the normalization by subtracting the intonation pattern from the local F_0 can improve tone recognition rate. We obtained two methods, i.e., *beginning-point intonation normalization* (BIN) and *center-point intonation normalization* (CIN) methods. Using the basic framework, we evaluated these methods on TPC and TASC. The experimental results showed that both methods significantly increased recognition rates. The recognition rates between these methods were not significantly different. As CIN tended to archive higher recognition rates, we decide to consider only CIN. CIN provided the recognition rates of 86.77% and 86.13% for TPC and TASC, respectively. As pointed out by (Shen 1989) that tones at sentence initial and final positions seem to behave differently from at other positions, we analyzed the error rates according to syllable positions of sentences. We found that the intonation normalization methods decreased the error rates of syllables in the beginning and ending syllable of sentences.

Effect of Stress

The F_0 contours of stressed syllables are generally quite different from unstressed ones. To compensate stress effect, we explored two approaches: (a) building separate tone models for stressed and unstressed syllables, and (b) incorporating stress information to tone models (Chapter 6). For the first one, we needed a stress detection. Therefore, we first performed two empirical experiments of stress detection on pairs of ambiguous words and polysyllabic words using a three-layer feedforward neural network. The former and the latter were performed on Potisuk-1996 corpus (PC-96) and Thai proper name corpus (TPNC), respectively. The acoustic features, duration, energy, and F_0, were considered. These features were extracted from several linguistic units, i.e., syllable, vowel and rhyme units, for comparison. The experimental results showed that the vowel unit provided the best average recognition rate for PC-96. However, the results were not significantly different from those of the other two units. For TPNC, the rhyme unit yielded better significantly recognition rate than the other units. We

concluded that the rhyme unit outperformed the other units for stress detection.

Then, we performed an empirical study of tone recognition. Based on the above approaches, we have proposed two methods, i.e., *separated stress mehtod* (SSM) and *incorporated stress feature method* (ISFM). Using the tone recognition framework, we tested these methods on PC-96 and TPNC. The experimental results showed that both methods increased the tone recognition rates. ISFM provided better recognition rates than SSM. The best recognition rate of 91.64% and 87.89% were achieved for PC-96 and TPNC, respectively, when ISFMs were applied. An additional advantage of ISFM was that it did not need stressed/unstressed transcription. Thus, we decided to use ISFM for the following experiments.

We additionally incorporated ISFMs into our simple tone model and performed experiments on TPC and TASC. We found that TSFM improved the recognition rates. The highest recognition rates were 89.23% and 87.24% for TPC and TASC, respectively. We then analyzed the tone error rates by considering the rhyme duration of a syllable. This showed that ISFM extremely reduced the error rates for the short syllables having the rhyme duration below 100 ms.

Incorporation of Tone Modelling into Speech Recognition

We presented an empirical study for integrating several refined tone models into a speech recognition system to enhance the recognition performance. The syllable-based speech recognition was used as the basic speech recognition framework. The 12^{th} order of RASTA coefficients extracted from 15 frames of a syllable were used as input features and fed into a three-layer feedforward neural network for training and testing. The refined tone models were also combined with spectral features together and fed into the network for enhancing speech recognition rates. The experiments were run on TPC and TASC. The experimental results showed that the integration of the refined tone models highly improved speech recognition rates for both corpora. The best error reduction rates of 85.16% and 75.06% were achieved for TPC and TASC, respectively.

Thesis Contributions

This thesis contributes to the advancement of speech recognition technology by presenting methods of tone modelling for improving tone and speech recognition by

machines. The main contribution of this thesis involves the empirical studies of various effects on tone recognition. The following contributions were made in this thesis:

- The empirical studies of Thai tones and tonal variations, which analyzes the effects of syllable structure, tone coarticulation, intonation, and stress, on the acoustic realizations of tones.
- The development and analysis of a tone recognition framework based on tone features, frequency scale, normalization technique and critical tone segment.
- The tone models that can be conveniently enhanced and easily integrated into a speech recognition system.
- An example of a mechanism, which is able to combine multiple classifiers.
- The stress detection that is an important task for speech recognition in continuous speech.

8.2 Further Works

In this thesis, we have empirically studied various linguistic effects on tone recognition in Thai continuous speech. Many aspects of the work presented in this thesis can be improved or extended. Some methodologies and empirical results are also potentially useful for other applications. In this section, we mention some of these directions for future works.

Effect of Syllable Structure

In this thesis, we studied effect of syllable structure on three sets of hypothetical syllables that each set was for concentrating each phonematic unit constructing the syllables. We have proposed a tone recognition method for compensating this effect based on these data. In the future, we plan to extend the method for all potential Thai syllables.

Effect of Coarticulation

The proposed features, *Contextual Tone Features* (CTF), are context dependent features. The use of CTF highly improves recognition rates because it was evaluated

on two closed sentence corpora. The sentence utterances for training and test set were uttered using the same scribes. We plan to apply CTF on an opened sentence corpus in which the training and test sets will be uttered using different scribe utterances.

Effect of Intonation

There are three intonation contours in Thai: the Fall, the Rise, and the Convolution. However, due to the lack of speech data, we have studied the Fall only. We plan to build a large corpus covering all three intonation contours and apply our intonation normalization methods on the corpus. We will additionally try a nonlinear line instead of a straight line for modelling the intonation contour. Moreover, we have found that the change of intonation contour depends on the length of the sentence and the number of syllables in the sentence. This information has not yet been applied in our modelling. We plan to intensively consider this issue for improving our tone models in the near future.

The empirical study on Thai tones and intonation is not only useful for improving tone recognition, but also useful for Thai speech synthesis. The intonation modelling can be improved for synthesis applications. We will try this knowledge for enhancing the quality of Thai speech synthesis.

Effect of Stress

We have used the acoustic features, i.e., duration, energy and F_0 for identifying stressed or unstressed syllables. The recognition rates were quite well. However, there are some spectral features, i.e., spectral change and spectral tilt, used in the literature. We are interested in trying them.

The study of stress effect relies on the availability of a corpus with stress labels. SSM could not be applied on TPC and TASC for tone recognition experiments because the lack of stress labels in these corpora. Hence, if we have a useful corpus with stress labels, we can employ both SSM and ISFM for comparison.

Incorporation of Tone Modelling into Speech Recognition

In this thesis, we applied tone modelling for improving the performance of speech recognition using a simple approach that augments the acoustic features with tone fea-

tures and builds tone-dependent syllable model. The advantage of this approach is that we can easily incorporate the tone models into any existing recognizer without alternating its structure. However, this approach requires a large amount of training data, because a syllable with different tones cannot be shared in training (Huang and Seide 2000; Wang 2001). This approach is useful for a small vocabulary speech recognition system only. We plan to intensively study the issue of how to incorporate the tone information into a large vocabulary speech recognition system. We will shift our extension to the other approach that builds separate models for base syllables and tones. In the new approach, the models will be trained separately to take advantage of data sharing. However, this approach needs a good search algorithm (Cao et al. 2000) to integrate the acoustic and the tone scores to find an optimal solution in the entre search space. Furthermore, the post-processing approach (Wang 2001) that applies the tone models to resort the recognizer an N-best list is also an interesting approach to be incorporated. In this approach, the tone score is added to the tonal path score for each syllable in an A^* path. The N-best hypotheses are then resorted according to the *adjusted total scores* to give a new best sentence hypothesis.

REFERENCES

Abramson, A. S. 1998. Thai tones on a reference system. In T. G. et al. (Ed.), Thai Linguistics in Honour of Fang Kuei Li, pp. 1–12. Chulalongkorn University Press.

Aull, A. M. and Zue, V. W. 1985. Lexical stress determination and its application to large vocabulary speech recongnition. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 1549–1552.

Beaugendre, F.; House, D.; and Hermes, D. J. 2001. Accentuation boundaries in Dutch, French and Swedish. Speech Communication 33(4): 305–318.

Botinis, A.; Granström, B.; and Möbious, B. 2001. Developments and paradigms in intonation research. Speech Communication 33(4): 263–296.

Cao, Y.; Deng, Y.; Zhang, H.; Huang, T.; and Xu, B. 2000. Decision tree based Mandarin tone model and its application to speech recognition. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1759–1752.

Chang, P. C.; Sue, S. W.; and Chen, S. H. 1990. Mandarin tone recognition by multi-layer perceptron. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 1, pp. 517–520.

Charnvivit, P.; Jitapunkul, S.; Ahkuputra, V.; and Maneenoi, E. 2001. F_0 feature extraction by polynomial regression function for monosyllabic Thai tone recognition. In Proc. Int. Conf. Eurospeech.

Chen, S.-H. and Chang, S. 1992. A statistical model based fundamental frequency synthesizer of Mandarin speech. Journal of the Acoustical Society of America 92(1): 114–120.

Chen, S.-H. and Liao, Y.-U. 1998. Modular recurrent neural networks for Mandarin syllable recognition. IEEE Transactions on Neural Networks 9(6): 1430–1441.

Chen, S.-H. and Wang, Y.-R. 1995. Tone recognition of continuous Mandarin speech based on neural networks. IEEE Transactions on Speech Audio Processing 3(2): 146–150.

Cho, S. 1997. Combining modular neural networks developed by evolutionary algorithm. In Proc. IEEE Int. Conf. Evolutionary Computation, pp. 647–650.

Demeechai, T. and Mäkeläinen, K. 2001. Recognition of syllables in a tone language. Speech Communication 33(3): 2410–254.

Dietterich, T. G. 1997. Machine learning research: four current directions. AI Magazine 4(18): 353–362.

Duin, R. P. W. and Tax, D. M. J. 2000. Experiments with classifier combining rules. In Proc. Int. Workshop on Multiple Classifier Systems (MCS), pp. 16–29.

Fujisaki, H. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (Ed.), The Production of Speech, pp. 39–55. Berlin, Germany: Springer-Verlag.

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

Gandour, J.; Potisuk, S.; and Dechnongkit, S. 1994. Tonal coarticulation in Thai. Journal of Phonetics 22: 477–492.

Gandour, J.; Tumtavitikul, A.; and Satthamnuwong, N. 1999. Effects of speaking rate on Thai tones. Phonetica 56: 123–134.

Gerald, C. F. and Wheatley, P. O. 1994. Applied Numerical Analysis (5 ed.). Addison-Wesley Publishing Company.

Gillick, L. and Cox, S. 1989. Some statistical issues in the comparison fo speech recognition algorithms. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 532–535.

Greenberg, S. 1996. Understanding speech understanding: Towards a unified theory of speech perception. In ESCA Workshop on The Auditory Basis of Speech Perception.

Hansen, L. K. and Salamon, P. 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(10): 993–1001.

Hermansky, H. 1990. Perceptual Linear Predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87(4): 1738–1752.

Hermansky, H. 1998. Should recognizers have ears? Speech Communication 25(2-3): 3–27.

Hermansky, H. and Morgan, N. 1994. RASTA processing of speech. IEEE Transactions on Speech Audio Processing 2(4): 578–589.

Hermes, D. J. and van Gestel, J. C. 1991. The frequency scale of speech intonation. Journal of the Acoustical Society of America 90(1): 97–102.

Ho, T. K.; Hull, J. J.; and Srihari, S. N. 1994. Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(1): 66–75.

Howie, J. M. 1974. On the domain of tone in Mandarin. Phonetica 30: 129–148.

Huang, H. C. H. and Seide, F. 2000. Pitch tracking and tone features for Mandarin speech recognition. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1523–1526.

Jenkin, K. L. and Scordilis, M. S. 1996. Development and comparison of three syllable stress classifiers. In Proc. Int. Conf. Spoken Language Processing, pp. 733–736.

Jensen, U.; Moore, R. K.; Dalsgaard, P.; and Lindberg, B. 1994. Modelling intonation contours at the phrase level using continuous density hidden Markov models. Computer Speech and Language 8: 247–260.

Jian, F. H. L. 1998. Classification of Taiwanese tones based on pitch and energy movement. In Proc. Int. Conf. Spoken Language Processing, Vol. 2, pp. 329–332.

Johnson, M. 2000. Incorporating prosodic information and language structure into speech recognition systems. Ph. D. Thesis, Purdue University.

Jun, L.; Zhu, X.; and Luo, Y. 1998. An approach to smooth fundamental frequencies in tone recognition. In Proc. Int. Conf. Communication Technology, Vol. 1, pp. S16–10–1–S16–10–5.

Kirchhoff, K. and Bilmes, J. A. 1999. Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 2, pp. 693–696.

Kittler, J.; Hatef, M.; Duin, R. P. W.; and Matas, J. 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3): 226–239.

Kochanski, G. and Shih, C. 2001. Automated modelling of Chinese intonation in continuous speech. In Proc. Int. Conf. Eurospeech, pp. 911–914.

Kongkachandra, R.; Pansang, S.; Sripra, T.; and Kimpan, C. 1998. Thai intonation analysis in harmonic-frequency domain. In Proc. IEEE Asia-Pacific Conf. Circuits and System, pp. 165–168.

Ladd, R. D. 1996. Intonational Phonology, Vol. 2. Cambridge University Press.

Ladefoged, P. 1996. Elements of Acoustic Phonetics (2 ed.). The University of Chicago Press.

Lea, W. A. 1980. Prosodic aids to speech recognition. In W. A. Lea (Ed.), Trends in Speech Recognition, pp. 166–205. Englewood Cliffs, New Jersey: Prentice-hall, Inc.

Lee, L. S.; Tseng, C. Y.; and Ouh-Young, M. 1989. The synthesis rules in a Chinese text-to-speech system. IEEE Transactions on Acoustics, Speech, Signal Processing 37(9): 1309–1320.

Lee, T. and Ching, P. C. 1999. Cantonese syllable recognition using neural networks. IEEE Transactions on Speech Audio Processing 7(3): 466–472.

Lee, T.; Ching, P. C.; Chan, L. W.; Cheng, Y. H.; and Mark, B. 1993. An NN based tone classifier for Cantonese. In Proc. Int. Joint Conf. Neural Networks, Vol. 1, pp. 287–290.

Lee, T.; Ching, P. C.; Chan, L. W.; Cheng, Y. H.; and Mark, B. 1995. Tone recognition of isolated Cantonese syllables. IEEE Transactions on Speech Audio Processing 3(3): 204–209.

Lehiste, I. 1970. Suprasegmentals. The M.I.T. Press.

Li, J.; Xia, X.; and Gu, S. 1999. Mandarin four-tone recognition with the fuzzy C-means algorithm. In Proc. IEEE Int. Conf. Fuzzy Systems, Vol. 2, pp. 1059–1062.

Lieberman, P.; Katz, W.; Jongman, A.; Zimmerman, R.; and Miller, M. 1985. Measures of the sentence intonation of read and spontaneous speech in American English. Journal of the Acoustical Society of America 77(2): 649–657.

Luksaneeyanawin, S. 1983. Intonation in Thai. Ph. D. Thesis, University of Edinburgh.

Luksaneeyanawin, S. 1993. Speech computing and speech technology in Thailand. In Proc. Symposium on Natural Language Processing, pp. 276–321.

Luksaneeyanawin, S. 1995. Tone transformation. In Proc. the 2^{nd} Symposium on Natural Language Processing, pp. 345–353.

Luksaneeyanawin, S. 1998. Intonation in Thai. In D. Hirst and A. D. Cristo (Eds.), Intonation Systems A Survey of Twenty Language, pp. 376–394. Cambridge University Press.

Madhukumar, A. S.; Rajendran, S.; and Yegnanarayana, B. 1993. Intonation component of a text-to-speech system for Hindi. Computer Speech and Language 7: 283–301.

Moore, B. C. J. and Glasberg, B. R. 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America 74(3): 750–753.

Muthusamy, Y. K. 1993. A Segmental Approach to Automatic Language Identification. Ph. D. Thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology.

Ohde, R. N. and Sharf, D. J. 1992. Phonetic Analysis of Normal and Abnormal Speech. Macmillan Publishing Company.

Owens, F. J. 1993. Signal Processing of Speech. Macmillan Publishing Company.

Potisuk, S.; Gandour, J.; and Harper, M. P. 1996. Acoustic correlates of stress in Thai. Phonetica 53: 200–220.

Potisuk, S.; Harper, M. P.; and Gandour, J. 1995. Speaker-independent automatic classification of Thai tones in connected speech by analysis-by-synthesis method. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 632–635.

Potisuk, S.; Harper, M. P.; and Gandour, J. 1996. Using stress to disambiguate spoken Thai sentences containing syntactic ambiguity. In Proc. Int. Conf. Spoken Language Processing, pp. 805–808.

Potisuk, S.; Harper, M. P.; and Gandour, J. 1999. Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method. IEEE Transactions on Speech Audio Processing 7(1): 95–102.

Rietveld, A. C. M. and Gussenhoven, C. 1985. On the relation between pitch excursion size and prominence. Journal of Phonetics 13: 299–308.

Ross, M. J.; Shaffer, H. L.; Cohen, A.; Freudberg, R.; and Manley, H. J. 1974. Average magnitude difference function pitch extractor. IEEE Transactions on Acoustics, Speech, Signal Processing ASSP22: 353–362.

Sagisaka, Y. and Kaiki, N. 1992. Optimization of intonation control using statistical F_0 resetting characteristics. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 49–52.

Schneider, J. and Moore, A. W. 1997. A Locally Weighted Learning Tutorial Using Vizier 1.0. http://www.cs.cmu.edu/∼schneide/tut5/tut5.html.

Shattuck-Hufnagel, S. and Turk, A. E. 1996. A prosody tutorial for investigators of auditory sentence processing. Journal of Psycholinguistic Research 25(2): 193–247.

Shen, X.-N. 1989. Interplay of the four citation tones and intonation in Mandarin Chinese. Journal of Chinese Linguistics 17(1): 61–74.

Shen, X.-N. 1990. Tonal coarticulation in Mandarin. Journal of Phonetics 18: 281–295.

Shih, C. 1997. Declination in Mandarin. In An European Speech Communication Association Workshop, pp. 293–296.

Sluijter, A. and van Heuven, V. 1995. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. Phonetica 52: 71–89.

Sluijter, A. and van Heuven, V. 1996. Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100(4): 2471–2485.

Ström, N. 1997a. Automatic Continuous Speech Recognition with Rapid Speaker Adaptation for Human/Machine Interaction. Ph. D. Thesis, Deparment of Speech, Music, and Hearing, KTH (Royal Institute of Technology).

Ström, N. 1997b. Phoneme probability estimation with dynamic sparsely connected artificial neural networks. The Free Speech Journal 1(5).

Swerts, M.; Stranger, E.; and Heldner, M. 1996. F_0 declination in read-aloud and

spontaneous speech. In Proc. Int. Conf. Spoken Language Processing, pp. 1501–1504.

't Hart, J. 1981. Differential sensitivity to pitch distance, particularly in speech. Journal of the Acoustical Society of America 69: 811–821.

Taylor, P. A. 1992. A Phonetic Model of English Intonation. Ph. D. Thesis, University of Edinburgh.

Tebelskis, J. 1995. Speech Recognition Using Neural Networks. Ph. D. Thesis, School of Computer Science, Carnegie Mellon University.

Thorsen, N. 1978. An acoustical analysis of Danish intonation. Journal of Phonetics 6: 151–175.

Thorsen, N. 1980. A study of the perception of sentence intonation-evidence from Danish. Journal of the Acoustical Society of America 67: 1014–1030.

Thubthong, N. 1995. A Thai tone recognition system based on phonemic distinctive features. In Proc. the 2^{nd} Symposium on Natural Language Processing, pp. 379–386.

Thubthong, N. and Kijsirikul, B. 1999a. Improving isolated Thai digit speech recognition using tone modelling. In Proc. the 22^{nd} Electrical Engineering Conference, pp. 163–166.

Thubthong, N. and Kijsirikul, B. 1999b. A syllable-based connected Thai digit speech recognition using neural network and duration modeling. In Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems, pp. 785–788.

Thubthong, N. and Kijsirikul, B. 2000a. Improving connected Thai digit speech recognition using prosodic information. In Proc. the 4^{th} National Computer Science and Engineering Conference, pp. 63–68.

Thubthong, N. and Kijsirikul, B. 2000b. Support vector machines for Thai phoneme recognition. In Proc. Int. Conf. Intelligent Technologies, pp. 206–213.

Thubthong, N. and Kijsirikul, B. 2001a. Stress and tone recognition of polysyllabic words in Thai speech. In Proc. Int. Conf. Intelligent Technologies, pp. 356–364.

Thubthong, N. and Kijsirikul, B. 2001b. Support vector machines for Thai phoneme recognition. International Journal of Uncertainty, Fuzziness and Knowledge-Based

Systems 9(6): 803–813.

Thubthong, N. and Kijsirikul, B. 2001c. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(6): 815–825.

Thubthong, N. and Kijsirikul, B. 2002. An empirical study for constructing Thai tone models. In Proc. the 5^{th} Symposium on Natural Language Processing and Oriental COCOSDA.

Thubthong, N.; Pusittrakul, A.; and Kijsirikul, B. 2000a. An efficient method for isolated Thai tone recognition using combination of neural networks. In Proc. the 4^{th} Symposium on Natural Language Processing, pp. 224–242.

Thubthong, N.; Pusittrakul, A.; and Kijsirikul, B. 2000b. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. In Proc. Int. Conf. Intelligent Technologies, pp. 229–234.

Thubthong, N.; Pusittrakul, A.; Sookawatand, T.; and Kijsirikul, B. 2000. Tone recognition of continuous Thai speech using half-Tone model. In Proc. the 4^{th} National Computer Science and Engineering Conference, pp. 69–74.

Tungthangthum, A. 1998. Tone recognition for Thai. In Proc. IEEE Asia-Pacific Conf. Circuits and System, pp. 157–160.

van Bergem, D. 1993. Acoustic vowel reduction as a function of sentence accent, work stress, and word class. Speech Communication 12: 1–23.

van Kuijk, D. and Boves, L. 1997. Acoustic characteristics of lexical stress in continuous speech. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 1655–1658.

van Kuijk, D. and Boves, L. 1999. Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication 27: 95–111.

van Kuijk, D.; van den Heuvel, H.; and Boves, L. 1996. Using lexical stress in continuous speech recognition for Dutch. In Proc. Int. Conf. Spoken Language Processing, pp. 1736–1739.

Vapnik, V. 1998. Statistical Learning Theory. Wiley.

Waibel, A. 1988. Prosody and Speech Recognition. London: Pitman.

Wang, C. 2001. Prosodic modeling for improved speech recognition and understanding. Ph. D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Instutute of Technology.

Wang, C. and Seneff, S. 1998. A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition. In Proc. Int. Conf. Spoken Language Processing, pp. 635–638.

Wang, C. and Seneff, S. 2000. Improved tone recognition by normalizing for coarticulation and intonation effects. In Proc. Int. Conf. Spoken Language Processing.

Wang, C.-F.; Fujisaki, H.; and Hirose, K. 1990. Chinese four tone recognition based on the model for process of generatin F_0 contours of sentences. In Proc. Int. Conf. Spoken Language Processing, pp. 221–224.

Wang, H.-M.; Ho, T.-H.; Yang, R.-C.; Shen, J.-L.; Bai, B.-R.; Hong, J.-C.; Chen, W.-P.; Yu, T.-L.; and Lee, L.-S. 1997. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. IEEE Transactions on Speech and Audio Processing 5(2): 195–200.

Wang, Y. R. and Chen, S. H. 1994. Tone recognition of continuous Mandarin speech assisted with prosodic information. Journal of the Acoustical Society of America 96(5): 1738–1752.

Whalen, D. H. and Xu, Y. 1992. Information for Mandarin tones in amplitude contour and in brief segments. Phonetica 49: 25–47.

Wu, S. L. 1998. Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition. Ph. D. Thesis, Department of Electrical Engineering and Computer Science, University of California, Berkeley.

Wu, S. L.; Kingsbury, B. E. D.; Morgan, N.; and Greenberg, S. 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 721–724.

Wu, Y.; Hemmi, K.; and Inoue, K. 1991. A tone recognition of polysyllabic Chinese words using an approximation model of four tone pitch patterns. In Proc. Int. Conf. Industrial Electronics, Control and Instrumentation, pp. 2115–2119.

Xu, Y. 1994. Production and perception of coarticulated tones. Journal of the Acoustical Society of America 95(4): 2240–2253.

Xu, Y. 1997. Contextual tonal variations in Mandarin. Journal of Phonetics 25: 61–83.

Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structure and speaking rate. Phonetica 55: 179–203.

Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F_0 contours. Journal of Phonetics 27: 55–105.

Xu, Y. and Wang, Q. E. 1997. What can tone studies tell us about intonation? In ESCA Workshop on Intonation: Theory, Models and Applications, pp. 337–340.

Yang, W. J.; Lee, J. C.; Chang, Y. C.; and Wang, H. C. 1988. Hidden Markov model for Mandarin lexical tone recognition. IEEE Transactions on Speech Audio Processing 36(7): 988–992.

Ying, G. S. 1998. Automatic measurement and representation of prosodic features. Ph. D. Thesis, Purdue University.

Ying, G. S.; Jamieson, L. H.; Chen, R.; and Michell, C. D. 1996. Lexicl stress detection on stress-minimal word pairs. In Proc. Int. Conf. Spoken Language Processing, pp. 1612–1615.

Zhang, J. 2001. The Effects of Duration and Sonority on Contour Tone Distribution Typological Survey and Formal Analysis. Ph. D. Thesis, University of California, Los Angeles.

Zhang, J.-S. and Hirose, K. 1998. A roboust tone recognitoin method of Chinese based on sub-syllabic F_0 contour. In Proc. Int. Conf. Spoken Language Processing, Vol. 3, pp. 703–706.

Zhang, J.-S. and Hirose, K. 2000. Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1419–1422.

Zhang, J.-S.; Nakamura, S.; and Hirose, K. 2000. Discriminating Chinese lexical tones by anchoring F_0 feature. In Proc. Int. Conf. Spoken Language Processing.

APPENDICES

APPENDIX A

THE INTERNATIONAL PHONETIC ALPHABET OF THAI PHONEMES

A.1 Consonants

p	ป
t	ต ฏ
c	จ
k	ก
ʔ	อ
ph	พ ภ ผ
th	ท ธ ฒ ฑ ถ ฐ
ch	ช ฌ ฉ
kh	ข ค ฆ
b	บ
d	ด ฎ
m	ม หม- (หมี)
n	น ณ หน- (หนู)
ŋ	ง หง- (เหงา)
f	ฝ ฟ
s	ส ศ ษ ซ
h	ห ฮ
r	ร หร- (เหรียญ)
l	ล ฬ หล- (หลาน)
w	ว หว- (หวี)
j	ย ญ หย- หญ- (หย่า หญิง)

Figure A.1: The IPA of Thai consonants.

A.2 Vowels

i	อิ
i:	อี
e	เอะ
e:	เอ
ɛ	แอะ
ɛ:	แอ
ɯ	อึ
ɯ:	อื
ɤ	เออะ
ɤ:	เออ
a	อะ
a:	อา
u	อุ
u:	อู
o	โอะ
o:	โอ
ɔ	เอาะ
ɔ:	ออ
ia	เอียะ
i:a	เอีย
ɯa	เอือะ
ɯ:a	เอือ
ua	อัวะ
u:a	อัว

Figure A.2: The IPA of Thai vowels.

A.3 Tones

ˉ	the mid	สามัญ
ˋ	the low	เอก
ˆ	the fall	โท
ˊ	the high	ตรี
ˇ	the rise	จัตวา

Figure A.3: The IPA of Thai tones.

APPENDIX B

PUBLICATIONS

The following published papers by the author are directly connected with the research described in this thesis and are appended here.

B.1 National Conferences

1. Thubthong, N., and Kijsirikul B. 1999. Improving isolated Thai digit speech recognition using tone modelling. In Proc. the 22^{nd} Electrical Engineering Conference, pp. 163-166.

2. Thubthong, N., and Kijsirikul B. 2000. Improving connected Thai digit speech recognition using prosodic information. In Proc. the 4^{th} National Computer Science and Engineering Conference, pp. 63-68.

3. Thubthong, N.; Pusittrakul, A.; Sookawat, T.; and Kijsirikul B. 2000. Tone recognition of continuous Thai speech using half-tone model. In Proc. the 4^{th} National Computer Science and Engineering Conference, pp. 69-74.

B.2 International Conferences

1. Thubthong, N., and Kijsirikul, B. 1999. A syllable-based connected Thai digit speech recognition using neural network and duration Modeling. In Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems, pp. 785-788.

2. Thubthong, N.; Pusittrakul, A.; and Kijsirikul, B. 2000. An efficient method for isolated Thai tone recognition using combination of neural networks. In Proc. The 4^{th} Symposium on Natural Language Processing, pp. 224-242.

3. Thubthong, N.; Pusittrakul, A.; and Kijsirikul, B. 2000. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. In Proc. Int. Conf. Intelligent Technologies, pp. 229-234.

4. Thubthong, N.; Kijsirikul, B.; and Luksaneeyawin, S. 2001. Stress and Tone Recognition of Polysyllabic Words in Thai Speech. In Proc. Int. Conf. Intelligent

Technologies, pp. 356-364.

5. Thubthong, N.; Kijsirikul, B.; and Luksaneeyawin, S. 2002. An empirical study for constructing Thai tone models. In Proc. the 5^{th} Symposium on Natural Language Processing and Oriental COCOSDA.

B.3 International Journals

1. Thubthong, N., and Kijsirikul, B. 2001. Tone recognition of continuous Thai speech under tonal assimilation and declination effects using half-tone model. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(6): 815-825.

2. Thubthong, N.; Kijsirikul, B.; and Pusittrakul, A. A method for isolated Thai tone recognition using combination of neural networks. Computational Intelligence, to appear.

BIOGRAPHY

Name	Nuttakorn Thubthong
Sex	Male
Date of Birth	June 23, 1969
Marital Status	Single
Work	Lecturer in Department of Physics, Chulalonkorn University
Work address	Department of Physics, Faculty of Science, Chulalongkorn University, Phayathai Road, Bangkok, 10330 Thailand.

Education

2002	Ph.D. in Computer Engineering, Chulalongkorn University
1995	M.Sc. in Computer Science, Chulalongkorn University
1991	B.Sc. in Physics, Chulalongkorn University