ระบบรู้จำคำเรียกพยัญชนะไทย
: การศึกษาการวัดลักษณะสำคัญแบบต่าง ๆ

นางสาวอุมาวสี ทาทอง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2544
ISBN 974-03-0413-3

SPEECH RECOGNITION SYSTEM FOR THAI CONSONANT NAMING
: A STUDY OF VERIOUS FEATURE MEASUREMENTS

Miss UMAVASEE THATHONG

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2001

ISBN 974-03-0413-3

หัวข้อวิทยานิพนธ์	ระบบรู้จำคำเรียกพยัญชนะไทย : การศึกษาการวัดลักษณะสำคัญแบบต่าง ๆ
โดย	นางสาวอุมาวสี ทาทอง
สาขาวิชา	วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล
อาจารย์ที่ปรึกษาร่วม	ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยาพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร. สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. วาทิต เบญจพลกุล)

.....อาจารย์ที่ปรึกษา
(รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล)

.....อาจารย์ที่ปรึกษาร่วม
(ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณียนาวิน)

.....กรรมการ
(ดร. จุฬารัตน์ ตันประเสริฐ)

อุมาวสี ทาทอง : ระบบรู้จำคำเรียกพยัญชนะไทย : การศึกษาการวัดลักษณะสำคัญแบบต่างๆ. (SPEECH RECOGNITION SYSTEM FOR THAI CONSONANT NAMING: A STUDY OF VARIOUS FEATURE MEASUREMENTS) อาจารย์ที่ปรึกษา : รศ.ดร.สมชาย จิตะพันธ์กุล, อาจารย์ที่ปรึกษาร่วม : ผศ.ดร.สุดาพร ลักษณียนาวิน, 91 หน้า. ISBN 974-03-0413-3

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อพัฒนาระบบรู้จำคำเรียกพยัญชนะไทยโดยใช้ขั้นตอนวิธีการฐานความรู้ (Knowledge-based Algorithm) คำเรียกพยัญชนะไทยประกอบด้วยเสียงทั้งหมด 28 เสียง แบ่งออกเป็นเสียงสามัญ 21 เสียงและเสียงจัตวา 7 เสียง ขั้นตอนวิธีการฐานความรู้ ใช้การจำแนกเสียงออกเป็นกลุ่มตามลักษณะการเปล่งเสียง และตามฐานเสียง โดยรวมเอาขั้นตอนวิธีการจำแนกเสียงวรรณยุกต์เข้าไว้ด้วย ค่าลักษณะสำคัญ 5 ลักษณะที่ใช้ในงานวิจัยนี้ ได้แก่ สัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น สัมประสิทธิ์การประมาณพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัมที่คำนวณจากสัมประสิทธิ์การประมาณพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัมที่คำนวณจากการแปลงฟูริเยร์ และสัมประสิทธิ์เซปสตรัมบนความถี่เมล โดยปรับค่าอันดับ และจำนวนเกาส์เซียนมิกซ์เจอร์ในแบบจำลองฮิดเดนมาร์คอฟชนิดต่อเนื่อง เพื่อหาค่าที่เหมาะสมที่สุดกับระบบ ผลการรู้จำของระบบสามารถรู้จำเสียงวรรณยุกต์ได้ร้อยละ 100 และลักษณะสำคัญที่ให้อัตราการรู้จำสูงที่สุดคือสัมประสิทธิ์เซปสตรัมบนความถี่เมล ขั้นตอนวิธีการฐานความรู้ให้อัตราการรู้จำเพิ่มขึ้นร้อยละ 18.8 เมื่อทดสอบกับข้อมูลเสียง 1680 ตัวอย่าง ประกอบด้วยเสียงของผู้พูดจำนวน 60 คนเป็นชาย 33 คน และหญิง 27 คน แบ่งเป็นข้อมูลชุดทดสอบซึ่งประกอบด้วยชาย 11 คน และหญิง 9 คน ผลของอัตราการรู้จำรวมทั้งระบบเท่ากับร้อยละ 83.75

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

4070571621 : MAJOR ELECTRICAL ENGINEERING

KEYWORD: CONSONANT NAMING RECOGNITION / FEATURE MEASUREMENTS / SPEECH RECOGNITION /

UMAVASEE THATHONG : SPEECH RECOGNITION SYSTEM FOR THAI CONSONANT NAMING: A STUDY OF VARIOUS FEATURE MEASUREMENTS. THESIS ADVISOR : ASSOC. PROF. SOMCHAI JITAPANKUL, Dr.Ing., THESIS COADVISOR : ASSIST. PROF. SUDAPORN LUKSANEEYANAWIN, Ph.D. 91 pp. ISBN 974-03-0413-3

The objective of this thesis is to develop a Thai consonant naming recognition system - CNRS using knowledge-based algorithms. Thai CNRS is composed of 28 consonant sounds, 21 of middle tones and 7 of rising tones. In this research, the purely knowledge-based algorithm was implemented using manner of articulation and place of articulation incorporated the tone classification algorithm. Five features used for the analysis of the optimal feature of consonant naming are linear frequency cepstrum coefficients, linear prediction coefficients, cepstrum coefficients derived from linear prediction coefficients, cepstrum coefficients derived from fourier transform and mel frequency cepstrum coefficients. The orders of features and the number of gaussian mixture in Continuous Hidden Markov Model are varied to obtain the optimal system parameters. The results of this system can be concluded as follows, recognition rate of tone classification algorithm is 100 percent, the optimal feature for consonant naming is mel frequency cepstrum coefficients, knowledge-based algorithms can improve 18.8% recognition rate. The data used for training contained 1680 consonant naming spoken by 60 speakers (33 males and 27 females). The system was tested on 20 speakers (11 males and 9 females). The total Recognition rate is 83.75 percent.

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

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

หน้า

สารบัญคำศัพท์

ขั้นตอนวิธีการ	algorithm
อัตสหสัมพันธ์	autocorrelation
สัญญาณรบกวนพื้นหลัง	background noise
อัตราการตัดผ่านระดับกำหนด	band crossing rate
สัมประสิทธิ์**เซปสตรัม**	cepstrum coefficients
เสียงพูดแบบต่อเนื่อง	continuous speech
การเปรียบเทียบทางเวลาแบบพลวัต	dynamic time warping
ลักษณะสำคัญ	feature
การสกัดลักษณะสำคัญ	feature extraction
ชุดวงจรกรอง	filter bank
วงจรกรองดิจิตอลอันดับที่หนึ่ง	first-order digital filter
ส่วนย่อย, กรอบ	frame
ความถี่มูลฐาน	fundamental frequency
แบบจำลองฮิดเดนมาร์คอฟ	hidden markov model
สัมประสิทธิ์การประมาณพันธะเชิงเส้น	linear prediction coefficients
สัมประสิทธิ์เซปสตรัมบนความถี่เมล	mel frequency cepstrum coefficients
ลำดับ	order
แบบจำลองที่ใช้พารามิเตอร์ในการจำลอง	parametric model
การจำแนกรูปแบบ	pattern classification
หน่วยเสียง	phoneme
การเน้นล่วงหน้า	preemphasis
ฟังก์ชันความหนาแน่นของความน่าจะเป็น	probability density function
อัตราการชักตัวอย่าง	sampling rate
การประมวลผลสัญญาณเบื้องต้น	signal preprocessing
การวางกรอบขนาดสัญญาณ	smoothing window
ปริภูมิ	space
โครงร่างสเปกตรัม	spectral envelope
โครงสร้างแบบละเอียดของสเปกตรัม	spectral fine structure
ข้อมูลเสียงพูด	speech data

สารบัญคำศัพท์ (ต่อ)

สัญญาณเสียงพูด	speech signal
ไม่เปลี่ยนแปลงตามเวลา	stationary
ข้อมูลชุดทดสอบ	testing data
จุดเริ่มเปลี่ยน	threshold
ข้อมูลชุดฝึกฝน	training data
เสียงอโฆษะ, เสียงไม่ก้อง	unvoiced
เส้นเสียง	vocal cord
ช่องทางเดินเสียง	vocal tract
เสียงโฆษะ, เสียงก้อง	voiced
หน่วยเสียงสระ	vowel phoneme
ฟังก์ชันกรอบ	window function
อัตราการตัดผ่านศูนย์	zero crossing

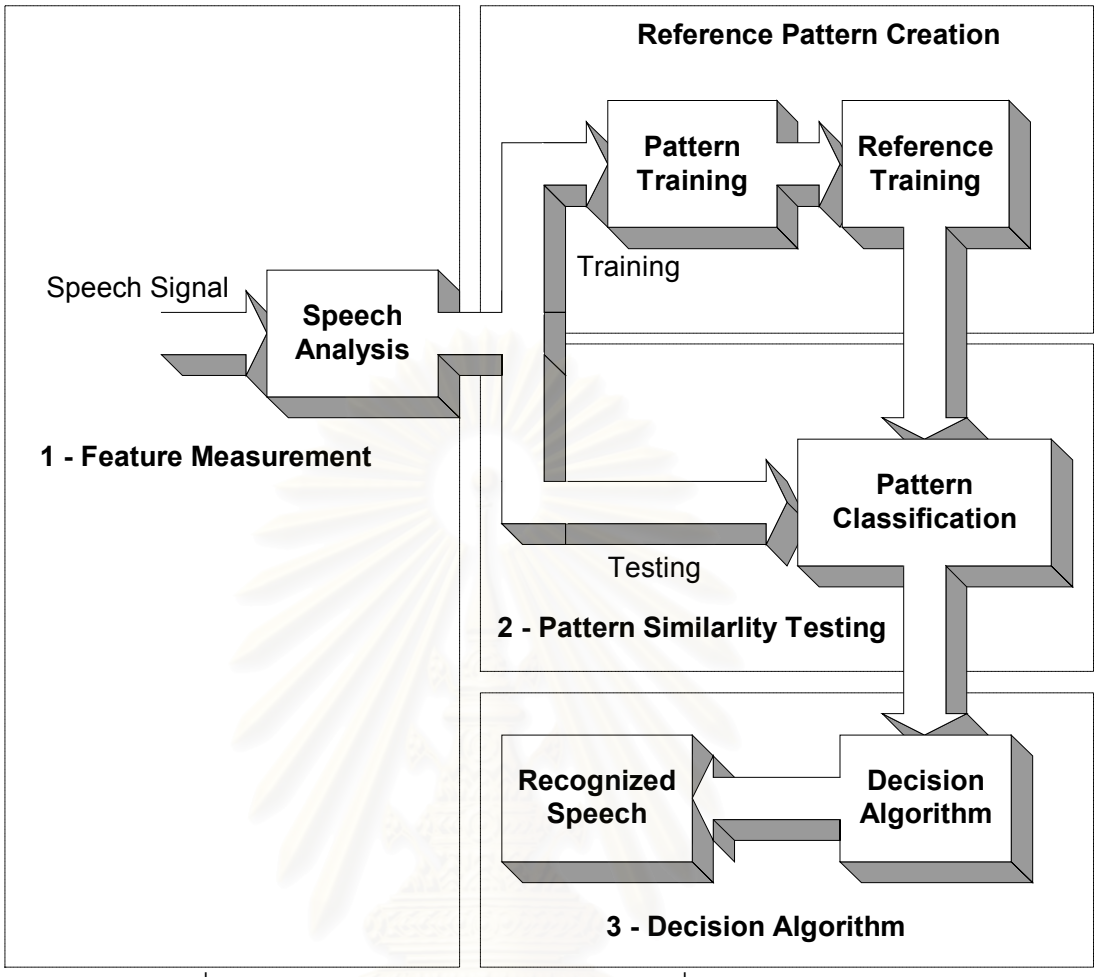
บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การสื่อสารโดยพื้นฐานของมนุษย์ประกอบด้วย การมองเห็น การได้ยิน และ การสัมผัส หากการสื่อสารทางการมองเห็นและการสัมผัสถูกจำกัด การสื่อสารด้วยเสียงยังคงทำให้มนุษย์เรียนรู้และเข้าใจกันและกันได้ จะเห็นได้จากการเติบโตอย่างรวดเร็วของโทรศัพท์มีสายและไร้สาย ซึ่งเข้ามามีบทบาทสำคัญต่อการดำรงชีวิตประจำวันอย่างมาก เทคโนโลยีที่นำมาประยุกต์ใช้กับการสั่งงานด้วยเสียงจึงถูกพัฒนาขึ้นเช่น การบันทึกข้อมูลเสียงลงเป็นข้อความ (Speech to Text) การค้นหาเส้นทางสำหรับรถยนต์โดยอัตโนมัติ (Car Navigator) ซึ่งใช้กับระบบการรู้จำคำพูดเป็นคำ (Speech Recognition) แต่ในระบบที่ต้องการรายละเอียดเฉพาะ เช่น การเก็บชื่อเฉพาะ ชื่อที่อยู่ ต้องอาศัยการรู้จำเสียงด้วยการสะกดคำ (Alphabet Recognition) เช่น การจองบัตรโดยสารเครื่องบินผ่านทางโทรศัพท์ ผู้โดยสารต้องสะกดชื่อเพื่อทำให้ได้ชื่อที่ถูกต้อง การต่อสายโทรศัพท์ไปยังแผนกหรือบุคคล (SmarTspelL พานาโซนิคเทคโนโลยี) (Rigazio, et al., 1998) การแก้ไขคำผิดด้วยการสะกดคำในระบบรู้จำคำพูดในกรณีที่เกิดการรู้จำคำผิดพลาดหรือคำพ้องเสียงเป็นต้น

ระบบเสียงในภาษาไทยประกอบด้วย พยัญชนะ 44 ตัว (21 รูป 28 เสียง) สระ 24 ตัว (สระเดี่ยว 18 ตัวและสระคู่ 6 ตัว) งานวิจัยที่มีอยู่แล้วคือการรู้จำเสียงสระภาษาไทย (Vowel Recognition) (ธีระ ภัทราพรนันท์, 2538; เอกฤทธิ์ มณีน้อย, 2541; Tungthangthum, 1998) ดังนั้นในงานวิจัยนี้จึงวิเคราะห์เฉพาะเสียงพูดพยัญชนะเท่านั้นเพื่อพัฒนาใช้ร่วมกับระบบรู้จำเสียงด้วยการสะกดคำต่อไป

ขั้นตอนในการรู้จำเสียงพูดโดยทั่วไปแบ่งได้เป็น 3 ขั้นตอนหลักคือ การวิเคราะห์และวัดค่าลักษณะสำคัญ (Speech Analysis and Feature Measurement) การจำแนกแบบรูป (Pattern Classification) หรือ การทดสอบความคล้ายคลึงกันของแบบรูป (Pattern Similarity Testing) และขั้นตอนวิธีการตัดสินใจ (Decision Algorithm) (วิศรุต อาขุบุตร, 2539) ดังรูปที่1.1 กระบวนการของระบบ เริ่มต้นจากการสร้างและเก็บลักษณะสำคัญ (Feature) แล้วนำไปฝึกฝน (Training) เพื่อจดจำแบบรูปอ้างอิง โดยแบบรูปอ้างอิงจะนำไปใช้ในการเปรียบเทียบกับเสียงพูดใหม่ที่ยังไม่ทราบแบบรูป เมื่อต้องการรู้จำเสียงพูดใหม่ระบบจะนำเสียงพูดที่ผ่านการวัดค่าลักษณะสำคัญแล้ว ไปผ่านขั้นตอนของการทดสอบ (Testing) ขั้นตอนสุดท้ายระบบจะตัดสินใจเลือกแบบรูปอ้างอิงที่มีความใกล้เคียงกับกับแบบรูปที่จะรู้จำมากที่สุด



รูปที่ 1.1 แบบจำลองแบบรูปการรู้จำทางสถิติที่ใช้ในการรู้จำเสียงพูด

การรู้จำเสียงพูดคำพยางค์เดียว สองพยางค์และ สามพยางค์ (วิศรุต อาขุบุตร, 2539) ให้อัตราการรู้จำร้อยละ 89.91 แต่เมื่อนำมารู้จำเสียงคำเรียกพยัญชนะได้อัตราการรู้จำเพียงร้อยละ 13 เท่านั้นเนื่องจากการเปล่งเสียงพยัญชนะมีช่วงระยะเวลาสั้นและการสะกดเสียงคำเรียกพยัญชนะไทยนั้นประกอบด้วยเสียงสระออ /@@/ ในทุกคำที่เปล่งเสียงเช่น เสียง กอ (/k@@0/) เสียง ขอ (/kh@@4/) เสียง คอ (/kh@@0/) เป็นต้น จากการเปล่งเสียงดังกล่าวทำให้มีความคล้ายคลึงกันมากไม่สามารถพิจารณาได้จากลักษณะสำคัญอย่างใดอย่างหนึ่งเท่านั้น จึงต้องหาลักษณะสำคัญหลายๆ ลักษณะที่สามารถแยกส่วนของความต่างของแต่ละเสียงพยัญชนะออกจากกัน โดยมีแนวความคิดพื้นฐานที่ว่าเสียงแต่ละเสียงมีคุณลักษณะที่แตกต่างกัน ถ้าสามารถแยกความแตกต่างนั้นได้จะสามารถพัฒนาวิธีการหาคุณลักษณะสำคัญที่รวมเอาความต่างทั้งหมดมารวมกันเพื่ออัตราการรู้จำที่สูงขึ้น

สำหรับคำเรียกที่มีลักษณะพ้องเสียงกันเช่น เสียง ญอ (/j@@0/) และ เสียง ยอ (/j@@0/) ซึ่งผู้พูดแต่ละบุคคลมีคำเรียกที่ต่างกันไปเช่น ยอ-ยักษ์ ญอ-หญิง หรือ ญอ-ผู้-หญิง คำเรียกในส่วน

ที่ใช้ระบุเช่น ยักษ์ หญิง หรือ ผู้หญิง ซึ่งเป็นเสียงคำพยางค์เดียวและสองพยางค์จะถูกนำไปใช้ในระบบรู้จำเสียงพูดพยางค์เดียว สองพยางค์และสามพยางค์ ต่อไป

ลักษณะสำคัญของเสียงสามารถวิเคราะห์หาได้ทั้งทางเวลา (Time Domain) และทางความถี่ (Frequency Domain) ดังนี้

1. ในทางเวลาจากรูปคลื่นเสียง (Speech Waveform) จะได้คุณลักษณะสำคัญต่างๆ ได้แก่ พลังงาน (Energy) อัตราการตัดผ่านศูนย์ (Zero Crossing) โดยค่าพลังงานบอกถึงจุดเริ่มต้นและสิ้นสุดคำ ค่าอัตราการตัดผ่านศูนย์บอกถึงเสียงก้องหรือเสียงไม่ก้องเสียดแทรก (ณัฏฐชา จิตติวรางกูล, 2541)

2. ในทางความถี่จากสเปกตรัมของสัญญาณเสียงช่วงเวลาสั้นๆ (Short-Time Spectrum) จะได้คุณลักษณะสำคัญเชิงสเปกตรัม (Spectral Feature) ที่สามารถคำนวณหาได้จากกรรมวิธีต่างๆ เช่น ค่าสัมประสิทธิ์การทำนายพันธะเชิงเส้น (Linear Prediction Coefficient) ชุดวงจรกรองแบบดิจิตอล (Digital Filter Bank) สัมประสิทธิ์เซปสตรัม (Cepstrum Coefficient) สัมประสิทธิ์เซปสตรัมบนความถี่เมล (Mel Frequency Cepstrum Coefficient) เป็นต้น (Tuzun et al., 1994; Rabiner and Juang, 1993) ซึ่งจะกล่าวโดยละเอียดในบทที่ 2

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาวิธีการจำแนกความแตกต่าง การวัดค่าลักษณะสำคัญสำหรับเสียงพูดพยัญชนะไทย

2. เพื่อเป็นแนวทางนำไปประยุกต์ใช้กับระบบรู้จำเสียงพูดแบบอัตโนมัติที่ต้องอาศัยการสะกดเพื่อความถูกต้อง

ขอบเขตของการวิจัย

สามารถรู้จำเสียงคำเรียกพยัญชนะไทย 28 เสียงได้ไม่น้อยกว่าร้อยละ 85

คำจำกัดความที่ใช้ในการวิจัย

คำเรียกพยัญชนะไทย ในที่นี้หมายถึงเสียงพยัญชนะไทยที่เรียกด้วยคำเรียกที่สั้นที่สุด เช่น ก ข ค คำเรียกคือ กอ (/k@@0/) ขอ (/kh@@4/) คอ (/kh@@0/)

/@@/ คือเสียงสระออ ตัวเลขที่อยู่ข้างท้ายคือเสียงวรรณยุกต์ โดย 0 คือ เสียงสามัญ 1 คือ เสียงเอก 2 คือ เสียงโท 3 คือ เสียงตรี และ 4 คือ เสียงจัตวา

ประโยชน์ที่คาดว่าจะได้รับ

1. รายละเอียดและคุณสมบัติของลักษณะสำคัญสำหรับเสียงพูดพยัญชนะไทย 28 เสียง

2. กรรมวิธีที่เหมาะสมในการรู้จำเสียงพูดพยัญชนะไทย 28 เสียง
3. สามารถนำไปประยุกต์ใช้กับระบบรู้จำเสียงพูดแบบอัตโนมัติที่ต้องอาศัยการสะกดเพื่อความถูกต้อง เช่น การแก้คำผิดในการพิมพ์งานจากเสียง การต่อโทรศัพท์ การจองบัตรโดยสารเครื่องบินหรือรถโดยสาร เป็นต้น

วิธีดำเนินการวิจัย

1. ศึกษา ค้นคว้าข้อมูลเกี่ยวกับ
    - คุณลักษณะสำคัญของเสียงพยัญชนะไทยทั้ง 28 เสียง
    - ทบทวนวรรณกรรมเกี่ยวกับวิธีการวิเคราะห์การวัดค่าลักษณะสำคัญและการจำแนกความแตกต่างของเสียงพูดที่มีความคล้ายคลึงกัน
2. การพัฒนาระบบการรู้จำ
    - เก็บข้อมูลเสียงคำเรียกพยัญชนะไทยของกลุ่มตัวอย่างจำนวน 60 คน แบ่งเป็นชาย 33 คน หญิง 27 คน
    - วิเคราะห์และพัฒนาโปรแกรมการเลือกค่าลักษณะสำคัญ และวิธีการจำแนกความแตกต่างของเสียงพูดที่เหมาะสม
3. ทดสอบ แก้ไขและปรับปรุงโปรแกรม
4. สรุปผลการวิจัยและจัดทำรายงานการวิจัยของวิทยานิพนธ์

บทที่ 2
แนวคิดทฤษฎีและเอกสารงานวิจัยที่เกี่ยวข้อง

แนวคิดและทฤษฎี

ในบทนี้จะกล่าวถึงรายละเอียดและทฤษฎีที่นำมาใช้ในการวิเคราะห์สัญญาณเสียงเป็นสำคัญ โดยเรียงหัวข้อการนำเสนอตามลำดับดังนี้ ทฤษฎีทางด้านภาษาศาสตร์ ทฤษฎีการวิเคราะห์สัญญาณเสียง และเอกสารงานวิจัยที่เกี่ยวข้อง

ทฤษฎีทางด้านภาษาศาสตร์

โครงสร้างพยางค์ในภาษาไทยประกอบด้วยหน่วยเสียง 3 ประเภทคือ พยัญชนะ สระ และวรรณยุกต์ ดังสมการที่ (2.1) โดยแบ่งออกเป็น เสียงพยัญชนะ 21 เสียง เสียงสระ 24 เสียง และเสียงวรรณยุกต์ 5 เสียง

$$S = C_i(C_i)V(V)(C_f)^{T} \quad (2.1)$$

C_i คือพยัญชนะต้น, C_f คือตัวสะกด, V คือสระ, T คือวรรณยุกต์ (Luksaneeyanawin, 1993)

ตารางที่ 2.1 ตารางการจำแนกพยัญชนะไทยตามลักษณะการเปล่งเสียง
(สุดาพร ลักษณียนาวิน, 2529)

ลักษณะการเปล่งเสียง (Manner of Articulation)	ฐาน (Place of Articulation)		ปาก (Bilabial)	ปุ่มเหงือก (Aveolar)	เพดานแข็ง (Palatal)	เพดานอ่อน (Velar)	ช่องเส้นเสียง (Glottal)
หยุก (Stop)	อโฆษะ (ไม่ก้อง, Voiceless)	ไม่พ่นลม (Unaspirated)	p (ป)	t (ต,ฏ)	c (จ)	k (ก)	@ (อ)
		พ่นลม (Aspirated)	ph (พ,ภ,ผ)	th (ท,ธ,ถ,ฑ,ฒ,ฐ)	ch (ช,ฌ,ฉ)	kh (ค,ฆ,ข)	
	โฆษะ (ก้อง, Voiced)		b (บ)	d (ด,ฎ)			
ไม่หยุก (Non-Stop)	นาสิก (Nasal)		m (ม)	n (น,ณ)		ng (ง)	
	เสียดแทรก (Fricative)		f (ฟ,ฝ)	s (ศ,ษ,ส,ซ)			h (ห,ฮ)
	เสียงลิ้นรัว (Trill)			r (ร)			
	เสียงข้างลิ้น (Lateral)			l (ล,ฬ)			
	ต่อเนื่อง (Approximant)		w (ว)		j (ย,ญ)		

ตารางที่ 2.2 คำเรียกพยัญชนะไทย 28 เสียง

เสียงวรรณยุกต์	เสียงพยัญชนะ			
สามัญ	ก /k@@0/	ค, ฆ /kh@@0/	ง /ng@@0/	จ /c@@0/
	ช, ฌ /ch@@0/	ซ /s@@0/	ย, ญ /j@@0/	ด, ฎ /d@@0/
	ต, ฏ /t@@0/	ท, ธ, ฒ, ฑ /th@@0/	น, ณ /n@@0/	บ /b@@0/
	ป /p@@0/	พ, ภ /ph@@0/	ฟ /f@@0/	ม /m@@0/
	ร /r@@0/	ล, ฬ /l@@0/	ว /w@@0/	อ /@@0/
	ฮ /h@@0/			
จัตวา	ข /kh@@4/	ฉ /ch@@4/	ถ, ฐ /th@@4/	ผ /ph@@4/
	ฝ /f@@4/	ส, ษ, ศ /s@@4/	ห /h@@4/	

เสียงพยัญชนะและ สระต่างก็มีลักษณะการกำเกิดของเสียงที่แตกต่างกัน โดยหัวข้อต่อไปนี้จะอธิบายถึงกลไกการกำเนิดเสียงและความแตกต่างของแต่ละเสียงพยัญชนะ

1. การจำแนกพยัญชนะไทยตามลักษณะการเกิดเสียง

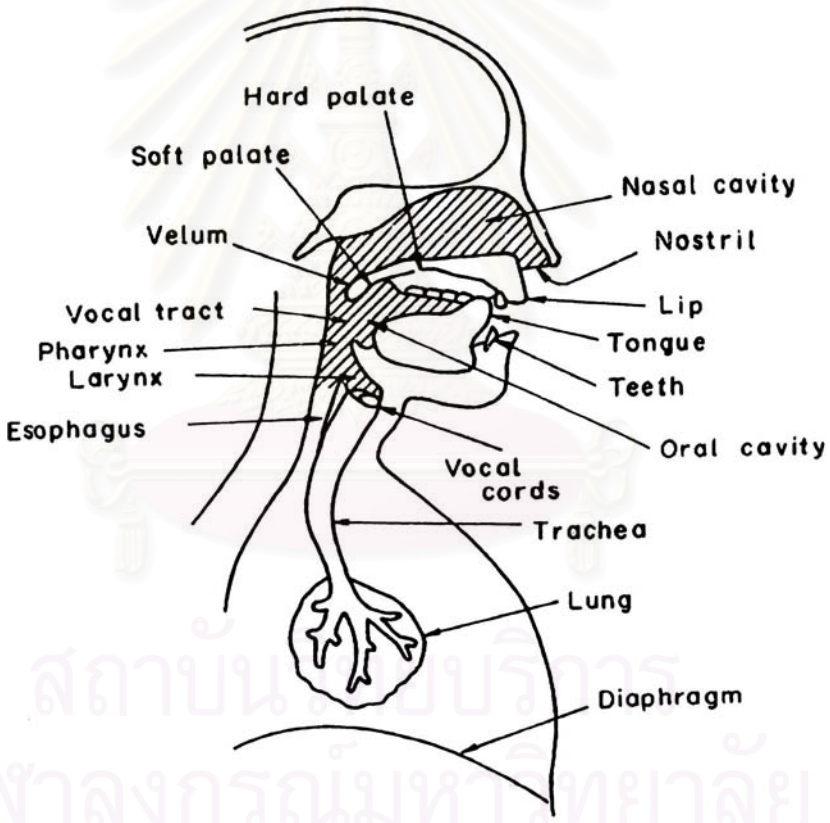
พยัญชนะในภาษาไทยมี 21 หน่วยเสียงแบ่งเป็นพยัญชนะกัก 11 หน่วยเสียง และพยัญชนะไม่กัก 10 หน่วยเสียง แสดงในตารางที่ 2.1 แต่เมื่อพิจารณาถึงเสียงวรรณยุกต์ ซึ่งเป็นส่วนหนึ่งของโครงสร้างภาษาไทยพบว่าเสียงกัก-ไม่ก้อง-พ่นลม และเสียงไม่กัก-เสียดแทรก มีวรรณยุกต์เสียงจัตวาเพิ่มขึ้นอีก 7 เสียง คือ เสียงผอ /ph@@4/, เสียงถอ /th@@4/, เสียงฉอ /ch@@4/, เสียงขอ /kh@@4/, เสียงฝอ /f@@4/, เสียงสอ /s@@4/, เสียงหอ /h@@4/ ดังนั้นระบบรู้จำคำเรียกพยัญชนะจึงประกอบด้วยเสียงคำเรียกพยัญชนะ 28 เสียง แบ่งเป็นเสียงสามัญ 21 เสียงและเสียงจัตวาอีก 7 เสียง ดังตารางที่ 2.2

2. กลไกการกำเนิดเสียง (Speech Production Mechanism)

อวัยวะในการเปล่งเสียงของมนุษย์ประกอบไปด้วย ปอด (Lungs) หลอดลม (Trachea) กล่องเสียง (Larynx) คอหอยหรือช่องคอ (Pharynx) จมูก (Nasal) ช่องปาก (Oral cavities) ซึ่งเรียงตัวกันเป็นช่องทางเดินอากาศดังรูปที่2.1 ส่วนที่อยู่เหนือกล่องเสียงขึ้นไปเรียกช่องทางเดินเสียง (Vocal Tract) สามารถเปลี่ยนรูปร่างโดยการเคลื่อนตัวของ กราม (Jaw) ลิ้น (Tongue) ริมฝีปาก (Lips) และส่วนอื่นๆ ช่องจมูกแยกจากช่องคอและช่องปากด้วยการยกตัวขึ้นและลงของ (Velum) หรือเพดานอ่อน (Soft Palate)

เมื่อกล้ามเนื้อช่องท้อง (Abdominal Muscles) ดันตัวยกกระบังลม (Diaphragm) ขึ้น แรงกดอากาศจะถูกปล่อยออกจากปอด ทำให้ลมผ่านจากหลอดลมและช่องว่างระหว่างเส้นเสียง

(Glottis) เข้าสู่กล่องเสียง โดยปกติช่องว่างระหว่างเส้นเสียงจะเปิดในขณะหายใจและจะแคบลงเมื่อพยายามเปล่งเสียง ลมที่ผ่านช่องว่างระหว่างเส้นเสียงจะถูกขัดจังหวะโดยการเปิดและปิดของเส้นเสียงเกิดเป็นเสียงซึ่งจำลองแบบได้ด้วยคลื่นสามเหลี่ยมอสมมาตร (Asymmetrical Triangular Waves) (Furui, 1989) เมื่อเส้นเสียงตึงและลมที่ออกจากปอดทำให้ความกดอากาศจากปอดถึงเส้นเสียง (Subglottal Air Pressure) สูง เส้นเสียงจะเปิดและปิดอย่างรวดเร็ว ความเป็นรายคาบของเสียงก็สูงตามไปด้วยและเมื่อความกดอากาศจากปอดถึงเส้นเสียงต่ำ เส้นเสียงเปิดและปิดอย่างช้าๆ ความเป็นรายคาบก็น้อยลง (คาบกว้างขึ้น) เช่นกัน จะเห็นได้ว่าความถี่ของช่วงเวลาการเปิดและปิดเส้นเสียงทำให้เกิดความเป็นรายคาบของเสียงเรียกว่าความถี่มูลฐาน (Fundamental Frequency)



รูปที่ 2.1 แผนภาพแผนผังแสดงอวัยวะในการเปล่งเสียง (Furui, 1989)

3. เสียงพยัญชนะ

เสียงก้อง (Voiced) คือเสียงที่เกิดขึ้นพร้อมกับการสั่นของเส้นเสียง ส่วนเสียงที่เส้นเสียงไม่สั่นจะเรียกว่าเสียงไม่ก้อง (Voiceless) ตัวอย่างเสียงที่ไม่เกิดการสั่นของเส้นเสียงเช่น เสียงจากการพ่นลมหรือการผิวปาก โดยเส้นเสียงจะมีลักษณะที่ค่อยๆ เปิดออกจากกันโดยการไหลอย่างปั่นป่วนของอากาศ (Turbulent Flow) ที่ช่องว่างระหว่างเส้นเสียง

เสียงเสียดแทรก (Fricative) เป็นเสียงที่มีความคล้ายสัญญาณรบกวน (Noise-like) เกิดโดยลิ้นหรือริมฝีปากสร้าง Turbulent Flow ของอากาศผ่านเส้นเสียงที่หดตัว เช่น เสียงสอ (/s@@4/) และเสียงฝอ (/f@@4/)

เสียงนาสิก (Nasal) เกิดจากการลดต่ำลงของ Velum ทำให้ช่องจมูกถูกเปิดออกต่อกับช่องปากเพื่อให้ลมผ่านออกมาพร้อมกับการกั้นลมไว้ที่ส่วนหนึ่งของช่องปาก

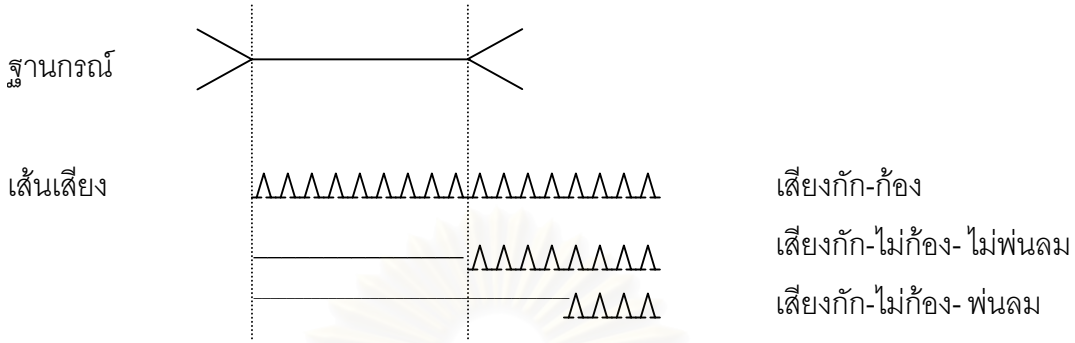
เสียงพยัญชนะกัก (Stop Consonant) เกิดจากการที่ลมจากปอด (Lung) ถูกกักไว้ภายในช่องปาก ณ ฐานกรณ์ที่กำเนิดเสียงซึ่งปิดสนิท ในขณะที่เพดานอ่อนเลื่อนขึ้นไปปิดผนังคอ (Pharynx) ทำให้ลมไม่สามารถผ่านไปยังช่องจมูกได้ เมื่ออวัยวะในช่องปากกักลมไว้ระยะหนึ่งจะทำให้ความกดอากาศในช่องปากเหนือเส้นเสียง (Supraglottal Air Pressure) สูงกว่าความกดอากาศภายนอกช่องปาก เมื่ออวัยวะที่ปิดกั้นลมไว้เปิดออกอย่างรวดเร็ว ลมที่ถูกกักไว้จะพุ่งออกทางช่องปาก (Impulsive Sound) เสียงพยัญชนะกักที่เปล่งด้วยกระแสลมจากปอดเดินทางออกนอกช่องปากนี้ เรียกว่า เสียงระเบิด (Plosive) ดังแสดงในรูปที่ 2.2 โดยแบ่งเป็น 3 ขั้นตอน

รูปที่ 2.2 รูปจำลองการทำงานของอวัยวะในช่องปากในขณะเปล่งเสียงพยัญชนะกัก
(จิตราวดี สิงหนิยม, 2542)

ขั้นที่ 1 ฐานกรณ์เคลื่อนเข้าหากันเพื่อปิดกั้นลม
ขั้นที่ 2 ฐานกรณ์ปิดกั้นลมชั่วระยะหนึ่ง
ขั้นที่ 3 ฐานกรณ์เปิดออกจากกันเพื่อให้ลมที่ถูกกักไว้เคลื่อนที่ออกมาอย่างรวดเร็ว

การทำงานของเส้นเสียงในขณะเปล่งเสียงพยัญชนะกัก สามารถแยกเสียงพยัญชนะกักได้คือ เสียงกัก-ก้อง เส้นเสียงมีการสั่นอย่างเป็นจังหวะสม่ำเสมอก่อนฐานกรณ์เปิดออกจากกัน แต่ถ้าเส้นเสียงเปิดออกจากกัน (ไม่สั่น) ในขณะปิดกั้นลม เราเรียกเสียงกักแบบนี้ว่า เสียงกัก-ไม่ก้อง เสียงกัก-ไม่ก้อง-ไม่พ่นลม เส้นเสียงเริ่มสั่นในทันทีที่ฐานกรณ์เปิดออกจากกัน ส่วนเสียงกัก-ไม่ก้อง-

พ่นลม เส้นเสียงเริ่มสั่นหลังจากฐานกรณ์เปิดออกจากกันแล้ว (จิตราวดี สิงหนิยม, 2542) แสดงในรูปที่ 2.3



รูปที่ 2.3 แสดงการทำงานของเส้นเสียงในขณะเปล่งเสียงพยัญชนะกักแบบต่างๆ (จิตราวดี สิงหนิยม, 2542)

เสียงลิ้นรัว เสียงข้างลิ้น และ เสียงต่อเนื่องไม่ได้เกิดจาก Turbulent Flow เหมือนเสียงเสียดแทรกและไม่ได้เกิดจากพัลส์เช่นเสียงกัก แต่เกิดจากคลายตัวออกของเส้นเสียงที่หดตัวและการเคลื่อนตัวอย่างช้าๆ ของอวัยวะในการเปล่งเสียง มีลักษณะคล้ายเสียงสระที่ใช้การเปลี่ยนแปลงรูปร่างของอวัยวะในการออกเสียงจึงเรียกเสียงกลุ่มนี้ว่าเสียงกึ่งสระ (Semivowel)

4. เสียงสระ

เสียงสระเป็นเสียงที่ไม่มีฐานกรณ์กักกั้นทางเดินของอากาศ แต่มีรูปร่างของช่องคอและช่องปากที่แตกต่างกันในขณะที่อากาศกำลังผ่านออกจากปากไป จึงทำให้เสียงสระต่างกันไป โดยคลื่นเสียงของสระทุกตัวมีลักษณะเป็นรายคาบ (เอกฤทธิ์ มณีน้อย, 2541)

5. Voice Onset Time (VOT)

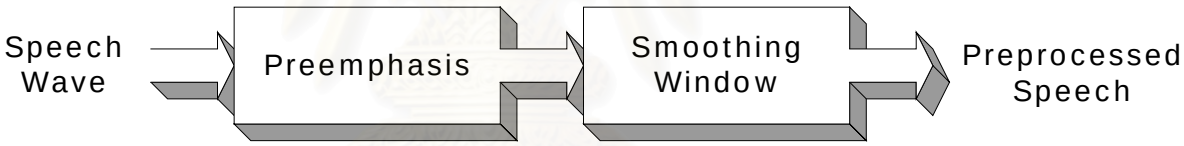
VOT คือค่าระยะเวลาระหว่างจุดเริ่มต้นของการเปิดฐานกรณ์กับจุดที่เส้นเสียงเริ่มสั่น ลิสเกอร์และ เอบรัมสัน (Lisker and Abramson, 1964) พบว่าค่า VOT ของหน่วยเสียงพยัญชนะกักในภาษาไทยสามารถแยกเสียงพยัญชนะกักได้เป็น 3 กลุ่มคือ เสียงกัก-ก้อง มีค่า VOT ติดลบ (Voicing lead) (ดังรูปที่ 2.3 เส้นเสียงเริ่มสั่นก่อนที่ฐานกรณ์เปิด เมื่อนำเวลาขณะที่เส้นเสียงสั่นลบด้วยเวลาขณะที่ฐานกรณ์เปิดจะได้ค่าติดลบ) เสียงกัก-ไม่ก้อง-ไม่พ่นลม มีค่า VOT เป็นบวก (Voicing lag) แต่ใกล้เคียงกับศูนย์ (เส้นเสียงเริ่มสั่นพร้อมๆ กับการเปิดของฐานกรณ์) และเสียงกัก-ไม่ก้อง-พ่นลมมีค่า VOT เป็นบวกและมีค่ามากกว่าศูนย์ (เส้นเสียงเริ่มสั่นหลังจากการเปิดของฐานกรณ์ชั่วระยะเวลาหนึ่ง)

ทฤษฎีทางด้านการวิเคราะห์สัญญาณเสียง (Spectral Analysis)

2.1 การประมวลผลสัญญาณเบื้องต้น (Signal Preprocessing)

การประมวลผลสัญญาณเบื้องต้นเป็นการแปลงสัญญาณเสียงพูด (Speech Wave) ที่ได้จากการบันทึกเสียงมาเป็นสัญญาณดิจิตอลที่เป็นข้อมูลเสียง (Speech Data) เพื่อนำไปใช้กับการประมวลผลสัญญาณขั้นต่อไป แต่เนื่องจากสัญญาณเสียงพูดเป็นสัญญาณที่ไม่เสถียร และเปลี่ยนแปลงตามเวลา (Nonstationary) ทำให้ไม่สามารถจำลองสัญญาณเสียงพูดด้วยค่าทางสถิติได้ จึงแบ่งสัญญาณเสียงพูดออกเป็นช่วงเวลาสั้นๆ (Frame) โดยมีกรอบเสียงพูด (Speech Frame) ประมาณช่วงละ 10 – 40 มิลลิวินาที เพื่อให้คุณสมบัติในแต่ละกรอบเสียงพูดนั้นเปลี่ยนแปลงตามเวลาน้อยมาก จนสามารถนิยามให้ส่วนย่อยนี้เป็นสัญญาณที่ไม่แปรเปลี่ยนตามเวลา (Stationary) พร้อมที่จะคำนวณหาค่าทางสถิติแทนสัญญาณเสียงพูดต่อไป

ขั้นตอนในการประมวลผลสัญญาณเบื้องต้นสามารถแบ่งได้เป็น 2 ขั้นตอน คือ ขั้นตอนกรรมวิธีการเน้นล่วงหน้า (Preemphasis) และขั้นตอนกรรมวิธีการวางกรอบขนาดสัญญาณ (Smoothing Window) ดังรูปที่ 2.4



รูปที่ 2.4 การประมวลผลสัญญาณเบื้องต้น

2.1.1 ขั้นตอนกรรมวิธีการเน้นล่วงหน้า (Preemphasis)

เป็นการบีบอัดสัญญาณเสียงพูดในช่วงพิสัยพลวัต (Dynamic Range) มีผลทำให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) มีค่าสูงขึ้น โดยนำสัญญาณเสียงพูดผ่านวงจรกรองแบบดิจิตอลอันดับที่หนึ่ง (First-Order Digital Filter) ที่มีฟังก์ชันถ่ายโอน $H(z)$ ดังแสดงในสมการที่ (2.2) และข้อมูลขาออกที่ผ่านการเน้นล่วงหน้าดังสมการที่ (2.3) (Furui, 1989)

$$H(z) = 1 - az^{-1} \tag{2.2}$$

$$\tilde{s}(n) = s(n) - as(n-1) \tag{2.3}$$

โดยที่ a คือค่าสัมประสิทธิ์ของวงจรกรอง

$\tilde{s}(n)$ เป็นค่าของสัญญาณเสียงพูดขาออกที่ผ่านกรรมวิธีล่วงหน้าที่ n

$s(n)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าที่ n

$s(n-1)$ เป็นค่าของสัญญาณเสียงพูดขาเข้าที่ $n-1$

โดยทั่วไปนิยมกำหนดให้ค่าสัมประสิทธิ์ของวงจรกรองเท่ากับ 0.95 (Rabiner and Juang, 1993)

2.1.2 ขั้นตอนกรรมวิธีการวางกรอบขนาดสัญญาณ (Smoothing Window)

เป็นการแบ่งสัญญาณเสียงพูดเป็นส่วนย่อยๆ โดยการคูณแต่ละค่าของสัญญาณในกรอบข้อมูลเสียงพูดด้วยค่าฟังก์ชันกรอบ (Window Function) ในงานวิจัยได้เลือกใช้ฟังก์ชันกรอบชนิด Hamming Window ดังแสดงในรูปที่ 2.5 มีผลทำให้เกิดการลดทอนแอมปลิจูดอย่างช้าๆ ที่บริเวณปลายแต่ละข้างของกรอบสัญญาณเสียงพูด เพื่อป้องกันการเปลี่ยนแปลงที่ไม่ต่อเนื่องบริเวณจุดปลาย โดยสมการที่ (2.4) แสดงค่าฟังก์ชันกรอบ $w(n)$ และสมการที่ (2.5) แสดงค่าสัญญาณเสียงพูดที่ผ่านกรรมวิธีการวางกรอบ

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

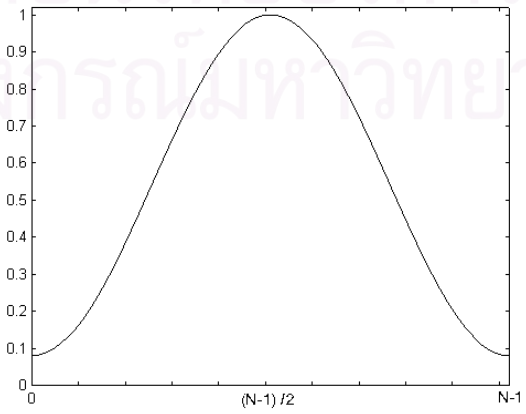
เมื่อ $n = 0,1,\ldots,N-1$

โดยที่ n คือลำดับข้อมูลในกรอบสัญญาณเสียงพูด, N คือจำนวนของข้อมูลในแต่ละกรอบสัญญาณเสียงพูด

$$\tilde{x}_l(n) = x_l(n)w(n) \tag{2.5}$$

เมื่อ $l = 0,1,\ldots,L-1$

โดยที่ l คือลำดับของกรอบสัญญาณเสียงพูด, L คือจำนวนของกรอบสัญญาณเสียงพูด, $x_l(n)$ เป็นค่าสัญญาณเสียงพูดของข้อมูลที่ n, $\tilde{x}_l(n)$ เป็นค่าสัญญาณเสียงพูดที่ผ่านกรรมวิธีการวางกรอบ โดยสัญญาณเสียงที่ได้จะมีความยาวเท่ากับจำนวนของกรอบสัญญาณเสียงพูดและมีหน่วยเป็น L เฟรม



รูปที่ 2.5 ฟังก์ชันกรอบชนิด Hamming Window

2.2 การวิเคราะห์ค่าเชิงเวลา (Time-Domain Analysis)

การวิเคราะห์ค่าเชิงเวลาเป็นการวิเคราะห์ลักษณะสำคัญของเสียงจากลักษณะสำคัญเชิงเวลา (Time-Domain Features) ได้แก่ พลังงาน, อัตราการตัดผ่านศูนย์ (Zero Crossing)

2.2.1 พลังงาน

ค่าพลังงานนิยมใช้ในการหาจุดเริ่มต้นและจุดสิ้นสุดของเสียงพูด เพราะสามารถแสดงความแตกต่างของสัญญาณเสียงพูดและสัญญาณรบกวนพื้นหลัง (Background Noise) ได้อย่างสะดวกและรวดเร็ว เนื่องจากการคำนวณที่ไม่ซับซ้อนดังสมการที่ (2.6)

$$E(n) = \sum_{i=0}^{N-1} s_n{}^2(i) \tag{2.6}$$

โดย $E(n)$ เป็นค่าระดับพลังงานของข้อมูลเสียงพูดเฟรมที่ n

$S_n(i)$ คือค่าสัญญาณเสียงพูดที่ i ในเฟรมที่ n

N คือจำนวนตัวอย่างเสียงพูดใน 1 เฟรม (Deller, Proakis, and Hansen, 1993)

2.2.2 อัตราการตัดผ่านศูนย์ (Zero Crossing)

อัตราการตัดผ่านศูนย์เป็นการวัดจำนวนครั้งการตัดผ่านแกนเวลาที่ระดับศูนย์คือมีการเปลี่ยนเครื่องหมายจากบวกเป็นลบหรือจากลบเป็นบวกของสัญญาณเสียงที่อยู่ติดกันดังสมการที่ (2.7) (Lee et. al., 1995) คุณสมบัติดังกล่าวใช้ตรวจสอบความถี่ของสเปกตรัมและคาบของสัญญาณเสียง แต่อัตราการตัดผ่านศูนย์ง่ายต่อการถูกกระทบได้จากสัญญาณเสียงรบกวน (Furui, 1989) จึงมีการปรับปรุงอัตราการตัดผ่านศูนย์ให้เป็น **อัตราการตัดผ่านระดับกำหนด (Band Crossing Rate)** ดังสมการที่ (2.8) (ณัฏฐชา จิตติวรางกูล, 2541)

$$Z(n) = \frac{1}{N}\sum_{i=0}^{N} \frac{|\operatorname{sgn}\{S_n(i)\} - \operatorname{sgn}\{S_n(i-1)\}|}{2} \tag{2.7}$$

ขณะที่ $$\operatorname{sgn}\{S_n(i)\} = \begin{cases} +1, & S_n(i) \geq 0 \\ -1, & S_n(i) < 0 \end{cases}$$

โดย Z_n คืออัตราการตัดผ่านศูนย์ของเฟรมที่ n

$S_n(i)$ คือสัญญาณเสียงพูดที่ i ในเฟรมที่ n

และ N คือจำนวนตัวอย่างเสียงพูดใน 1 เฟรม

$$B(n) = \sum_{i=0}^{N} |\operatorname{sgn}\{S_n(i)\} - \operatorname{sgn}\{S_n(i-1)\}| \tag{2.8}$$

ขณะที่ $$\text{sgn}\{S_n(i)\} = \begin{cases} +1, & S_n(i) \geq L \\ \text{sgn}\{S_n(i-1), & -L \leq S_n(i) \geq L \\ -1, & S_n(i) < L \end{cases}$$

โดย L คือความสูงของระดับกำหนด

2.3 การวิเคราะห์ค่าเชิงสเปกตรัม (Spectral Analysis)

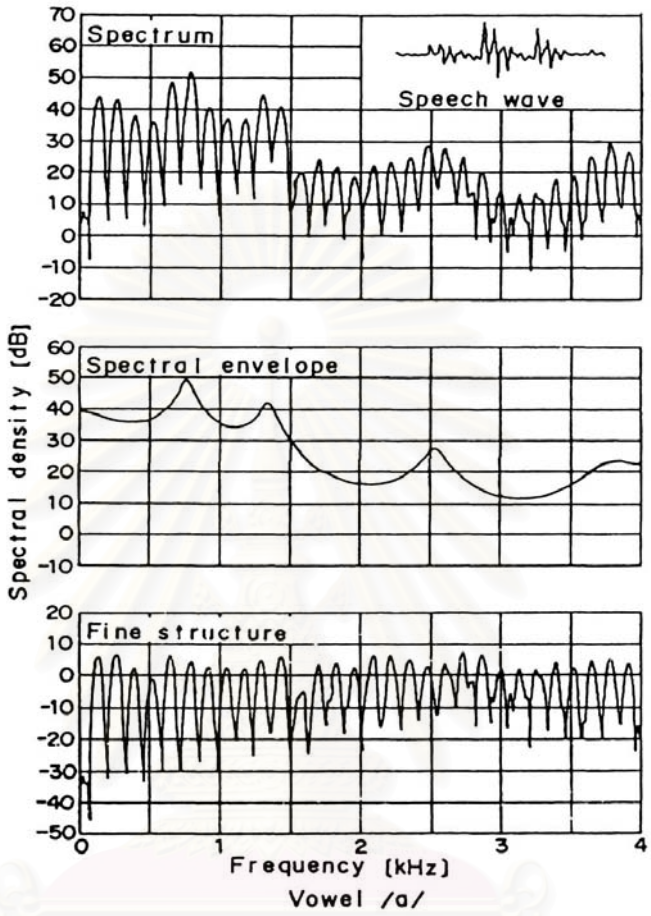
การวิเคราะห์ค่าเชิงสเปกตรัมเป็นการวิเคราะห์ลักษณะสำคัญของเสียงจากลักษณะสำคัญเชิงสเปกตรัม (Spectral Feature) เช่น สเปกตรัมความถี่ (Frequency Spectrum) หรือฟังก์ชันอัตสหสัมพันธ์ (Autocorrelation Function) แทนการวิเคราะห์โดยตรงจากสัญญาณเสียง เนื่องจากสัญญาณเสียงประกอบไปด้วย การรวมสัญญาณไซน์ (Sinusoidal Waves) โดยเปลี่ยนแอมพลิจูดและเฟสไปอย่างช้าๆ แต่การเปลี่ยนแปลงค่าเฟสไม่สามารถบอกถึงลักษณะสำคัญที่ใช้ในการรับรู้เสียงได้ดีเท่าลักษณะสำคัญเชิงสเปกตรัมที่สามารถบอกถึง การก้องของเสียงซึ่งอธิบายได้จากความเป็นรายคาบ (Periodic Pattern) ใช้บ่งความต่างของเสียงก้องและไม่ก้อง การเกิดหรือไม่เกิดการเรโซแนนท์ (Resonance) ของอวัยวะในการเปล่งเสียง (Articulatory Organs) และรูปร่างของเส้นเสียง ริมฝีปากและโพรงจมูก ที่มีผลต่อสเปกตรัมที่ออกมา

ลักษณะสำคัญเชิงสเปกตรัมประกอบด้วย โครงร่างสเปกตรัม (Spectral Envelope) ซึ่งเปลี่ยนแปลงอย่างช้าๆ และโครงสร้างแบบละเอียดของสเปกตรัม (Spectral Fine Structure) ที่เปลี่ยนแปลงอย่างรวดเร็วดังรูปที่ 2.6 โดยค่าโครงร่างสเปกตรัมถูกคำนวณในขั้นตอนของการสกัดคุณลักษณะสำคัญซึ่งจะกล่าวถึงในหัวข้อถัดไป ส่วนค่าโครงสร้างแบบละเอียดของสเปกตรัม คำนวณจากสัญญาณเสียงเข้าโดยตรง (Furui, 1989)

การสกัดคุณลักษณะสำคัญ (Feature Extraction) แบ่งออกเป็น 2 วิธีคือ วิธีวิเคราะห์พาราเมตริก (Parametric Analysis) และวิธีวิเคราะห์นอนพาราเมตริก (Nonparametric Analysis) โดยวิธีวิเคราะห์พาราเมตริกเป็นการสร้างแบบจำลองของช่องทางเดินเสียง (Vocal Tract) เพื่อหาค่าที่ใกล้เคียงกับสัญญาณเสียงขาเข้า ซึ่งตรงข้ามกับวิธีวิเคราะห์นอนพาราเมตริกที่แปรค่าไปตามการรับรู้เสียงของมนุษย์ (Tuzun, Demirekler and Nakiboglu, 1994)

ในงานวิจัยนี้นำเสนอลักษณะสำคัญทางพาราเมตริกที่คำนวณจาก ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear Prediction Coefficient - LPC) และลักษณะสำคัญนอนพาราเมตริกที่คำนวณจากค่าสัมประสิทธิ์เซปสตรัม (Cepstrum Coefficient - CEP) ความถี่มูลฐาน ความถี่ฟอร์แมนท์ ค่าสัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น (Linear Frequency Cepstrum Coefficients - LFCC) ค่าสัมประสิทธิ์เซปสตรัมบนความถี่เมล (Mel Frequency

Cepstrum Coeffieceint - MFCC) และวิธีปริภูมิย่อย เนื่องจากเป็นลักษณะสำคัญที่นิยมใช้กับระบบการรู้จำเสียง



รูปที่ 2.6 ลักษณะสำคัญเชิงสเปกตรัม (Furui, 1989)

2.3.1 สัมประสิทธิ์ของการประมาณพันธะเชิงเส้น (Linear Prediction Coefficient, LPC)

การประมาณพันธะเชิงเส้น (Linear Prediction) เป็นเทคนิคที่นิยมเพราะสามารถแสดงลักษณะสำคัญของช่องทางเดินเสียงได้อย่างดี ประกอบกับการคำนวณที่แม่นยำและง่ายต่อการนำไปประยุกต์ใช้งานจริง โดยใช้การคำนวณหาค่าพารามิเตอร์จากแบบจำลองช่องทางเดินเสียง และใช้การประมาณข้อมูลแบบ Linear Least Square หรือเรียกว่า Prediction (Sorenson, 1970)

หลักการพื้นฐานคือคำนวณหาค่าอนาคตของสัญญาณจากการประมาณผลรวมเชิงเส้น (Linear Combination) ของค่าของสัญญาณก่อนหน้า โดยแบบจำลองการเกิดสัญญาณ $s(n)$ จะประกอบไปด้วยแหล่งกำเนิดที่กระตุ้น (Excitation Source) $U(z)$ และวงจรกรองเชิงความถี่ $H(z)$ ใช้การแปลงแบบลาปลาซจะได้ $S(z)$

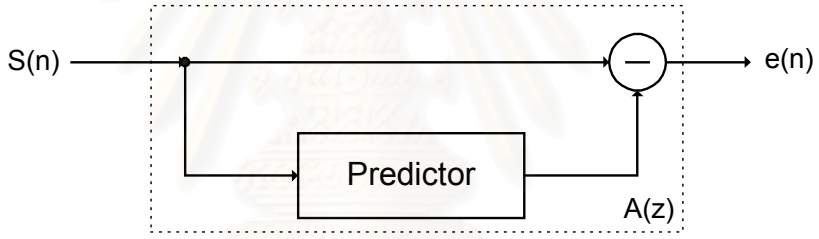
$$S(z) = U(z)H(z) \tag{2.9}$$

กำหนดให้วงจรกรอง $H(z)$ มี Pole เท่ากับ p และมี Zero เท่ากับ q ซึ่งหมายความว่าการจำลองสัญญาณ $\hat{s}(n)$ เป็นผลรวมเชิงเส้นระหว่างค่า p และค่า $q+1$ ของตัวอย่างขาออกกับตัวอย่างขาเข้าก่อนหน้า ดังสมการที่ (2.10)

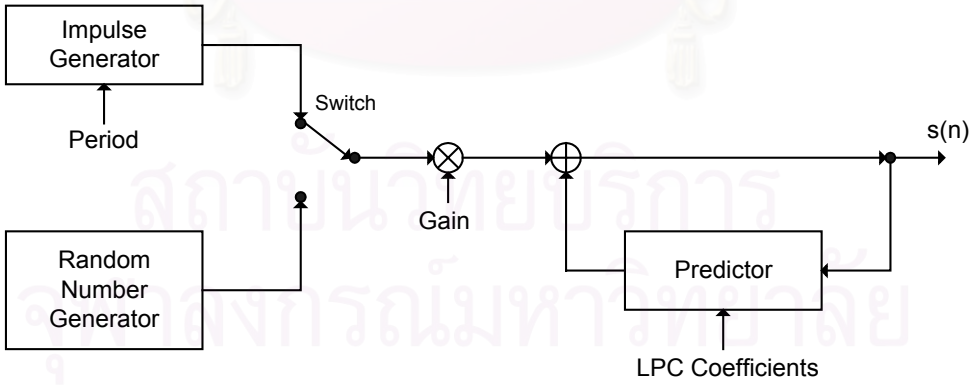
$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) + G\sum_{l=0}^{q} b_l \hat{u}(n-l) \tag{2.10}$$

โดย G คืออัตราขยาย (Gain Factor) ของการกระตุ้น สมมติให้ $b_0 = 1$ และ a_k คือค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น จากสมการที่ (2.10) จะได้ฟังก์ชันการถ่ายโอนดังสมการที่ (2.11)

$$\hat{H}(z) = \frac{\hat{S}(z)}{\hat{U}(z)} = G\frac{1+\sum_{l=1}^{q} b_l z^{-l}}{1-\sum_{k=l}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{2.11}$$



(ก)



(ข)

รูปที่ 2.7 แบบจำลอง all-pole สำหรับวิเคราะห์หาค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (O'Shaughnessy, 1988)

() แบบจำลองการวิเคราะห์การประมาณพันธะเชิงเส้นสำหรับเสียงพูด

() แบบจำลองการสังเคราะห์เสียงพูดจากแบบจำลองการประมาณพันธะเชิงเส้น

วงจรกรองที่ใช้ในการหาค่าสัมประสิทธิ์เป็นแบบจำลอง all-pole โดยกำหนดให้ $q = 0$ (Markel and Gray, 1980; O'Shaughnessy, 1988) ดังนั้นวงจร Predictor ในรูปที่ 2.7 (ก) แสดงได้ด้วยสมการที่ (2.12) และค่าผิดพลาด $e(n)$ ในสมการที่ (2.13) ส่วนค่าสัมประสิทธิ์ a_k คำนวณจากกรรมวิธี Least-squares Error โดยการพยายามทำให้ค่าผิดพลาดกำลังสองเฉลี่ยของ $e(n)$ ในแต่ละเฟรมมีค่าน้อยที่สุด

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{2.12}$$

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2.13}$$

กรรมวิธี Least-squares Error

ใช้ในการเลือกค่าสัมประสิทธิ์ a_k ให้มีค่าพลังงานของความผิดพลาดน้อยที่สุด โดยกำหนดให้ E คือพลังงานของค่าความผิดพลาด เมื่อ $e(n)$ คือค่าผิดพลาดของสัญญาณอินพุต $x(n)$ ดังสมการที่ (2.14)

$$E = \sum_{n=-\infty}^{\infty} e^2(n) \tag{2.14}$$

$$= \sum_{n=-\infty}^{\infty} \left[x(n) - \sum_{k=l}^{p} a_k x(n-k) \right]^2 \tag{2.15}$$

$$= \sum_{n=-\infty}^{\infty} x^2(n) - \sum_{n=-\infty}^{\infty} \left[2x(n) \sum_{k=1}^{p} a_k x(n-k) \right] + \sum_{n=-\infty}^{\infty} \left[\sum_{k=1}^{p} a_k x(n-k) \right]^2 \tag{2.16}$$

$$= \sum_{n=-\infty}^{\infty} x^2(n) - 2\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} x(n)x(n-k) + \sum_{n=-\infty}^{\infty} \left[\sum_{k=1}^{p} a_k x(n-k) \right]^2 \tag{2.17}$$

กำหนดค่า a_k ที่ทำให้ E มีค่าต่ำที่สุด โดยให้ $\partial E / \partial a_k = 0$ เมื่อ $k = 1,2,3,\ldots,p$ จะได้สมการเชิงเส้น p สมการและค่า a_k ที่ไม่ทราบค่า p ตัวดังนี้

$$\sum_{-\infty}^{\infty} x(n-i)x(n) = \sum_{k=l}^{p} a_k \sum_{n=-\infty}^{\infty} x(n-i) \cdot x(n-k) \tag{2.18}$$

เนื่องจากพจน์แรกเป็นอัตสหสมพันธ์ $R(i)$ ของ $x(n)$ มีความยาวจำกัดจะได้

$$\sum_{k=l}^{p} a_k R(i-k) = R(i), \qquad 1 \le i \le p \tag{2.19}$$

เมื่อ $$R(i) = \sum_{n=i}^{N-1} x(n)x(n-i) \tag{2.20}$$

สมการที่ (2.19) ประกอบไปด้วยสมการเชิงเส้น p สมการเขียนให้อยู่ในรูปเมตริกซ์ $Ra = r$ ได้ในสมการที่ (2.21)

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & & \vdots \\ \vdots & & \ddots & r_1 \\ r_{p-1} & \cdots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} \tag{2.21}$$

เมตริกซ์ R เป็นเมตริกซ์สมมาตรที่มีค่าในแนวทแยงมุมเท่ากันทั้งหมด เรียกเมตริกซ์ชนิดนี้ว่า Toepliz ซึ่งมีวิธีการแก้สมการที่อยู่ในรูปเมตริกซ์ Toepliz วิธีการหนึ่งคือขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin (Picone, 1996; O'Shaughnessy, 1988)

ขั้นตอนวิธีการวนซ้ำของ Levinson-Durbin

เป็นเทคนิคที่ใช้ในการคำนวณหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น a_i ในเมตริกซ์ a เมื่อ $i = 1,2,3,\ldots,p$ โดย p เป็นอันดับของการวิเคราะห์หาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นจากวิธีอัตสหสมพันธ์ แบ่งออกเป็น 4 ขั้นตอนดังนี้

ขั้นที่ 1 กำหนดค่าเริ่มต้น : $E_0 = R(0)$ และ $\alpha_0 = 0$

ขั้นที่ 2 คำนวณหาค่าสัมประสิทธิ์การสะท้อน (Reflection coefficient)

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} \alpha_{i-1}(j) R(i-j)}{E_{i-1}}$$

เมื่อ $R(i)$ และ $R(i-j)$ คำนวณได้จากสมการที่ (2.20)

ขั้นที่ 3 คำนวณค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น

ให้ $\alpha_i(i) = k_i$

และ $\alpha_i(j) = \alpha_{i-1}(j) - k_i \alpha_{i-1}(i-j)$, $j = 1,2,3,\ldots,i-1$

ขั้นที่ 4 คำนวณค่าผิดพลาดใหม่

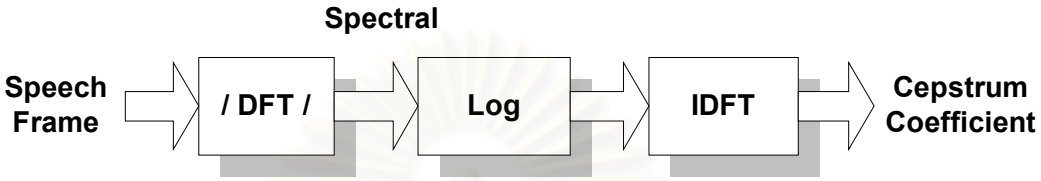
$E_i = (1 - k_i^2) E_{i-1}$

$i = i + 1$ วนซ้ำขั้นที่ 2 ถึง 4 เมื่อ $i < p$

เมื่อ $i = p$ แล้ว $a_k = \alpha_p(k)$ โดยที่ k คืออันดับของค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น

2.3.2 **สัมประสิทธิ์เซปสตรัม** (Cepstrum Coefficients - CEP)

สัมประสิทธิ์เซปสตรัมคำนวณได้จากการแปลงดีสครีตฟูริเยร์ (Discrete Fourier Transform - DFT) ดังรูปที่2.8 และจากค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้นดังสมการที่ (2.22) และ (2.23) (Rabiner and Juang, 1993)



รูปที่ 2.8 การคำนวณหาค่าสัมประสิทธิ์เซปสตรัมจากการแปลงดีสครีตฟูริเยร์

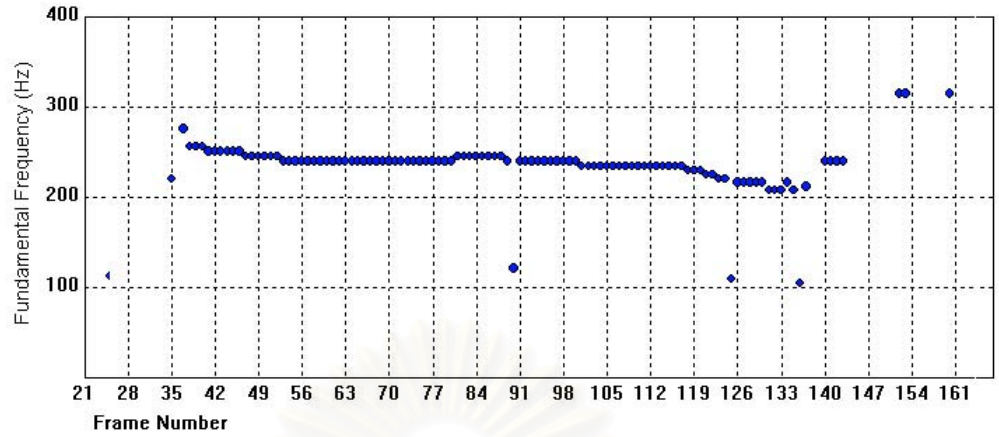
$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \qquad 1 \le m \le p \tag{2.22}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \qquad m > p \tag{2.23}$$

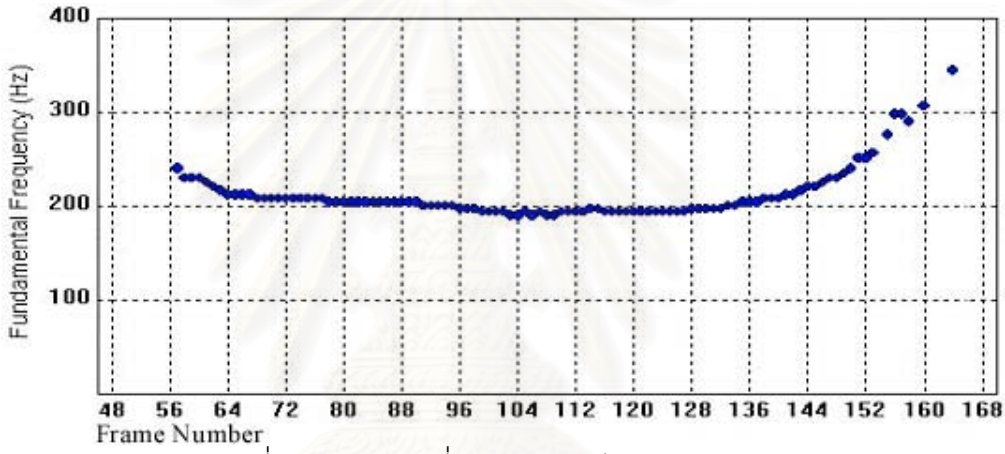
เมื่อ c_m คือค่าสัมประสิทธิ์เซปสตรัมอันดับที่ m, a_m คือค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (เอกฤทธิ์ มณีน้อย, 2541) และ p คืออันดับของค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้น

2.3.3 **ความถี่มูลฐาน** (Fundamental Frequency)

ความถี่มูลฐานเกิดจากการสั่นของเส้นเสียงทำให้เกิดคลื่นเสียงที่มีค่าความถี่มูลฐานต่างๆ คือเมื่อเส้นเสียงตึงอัตราการสั่นของเส้นเสียงจะสูง ความถี่มูลฐานก็จะสูง เสียงที่ได้ยินก็จะสูงเป็นปฏิภาคกัน ในการศึกษาเสียงสูงต่ำในภาษานั้นเราศึกษาจากค่าของความถี่มูลฐาน (สุดาพร ลักษณียนาวิน, 2529) ดังนั้นเสียงวรรณยุกต์จึงตรวจสอบได้ด้วยความถี่มูลฐาน โดยลักษณะคงที่ของความถี่มูลฐานบอกถึงเสียงสามัญดังรูปที่ 2.9 (ก) แตกต่างกับเสียงจัตวาที่มีความถี่มูลฐานเพิ่มขึ้นดังรูปที่ 2.9 (ข)

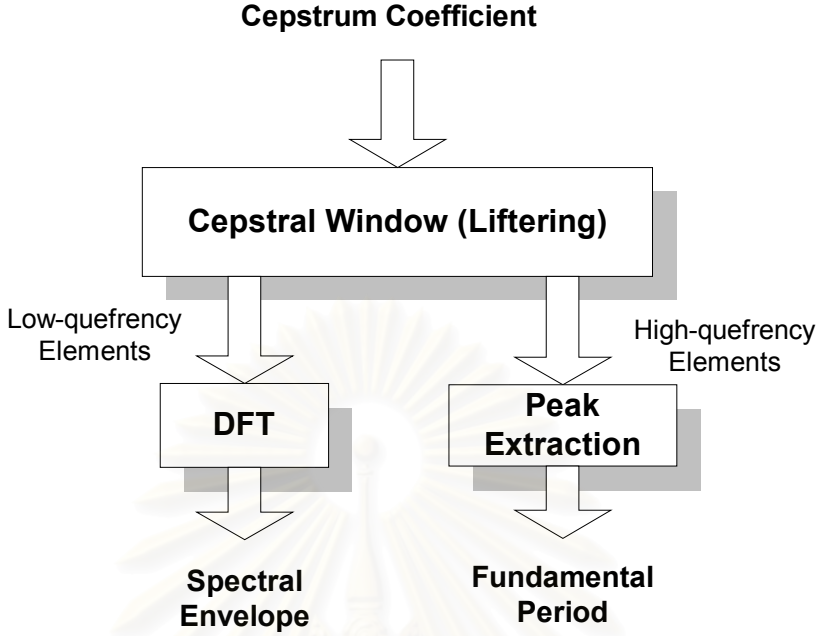


รูปที่ 2.9 (ก) ความถี่มูลฐานของเสียงกอ /k@@0/



รูปที่ 2.9 (ข) ความถี่มูลฐานของเสียงขอ /kh@@4/

ความถี่มูลฐานมีกรรมวิธีการหา 2 วิธีคือ วิธีฟังก์ชันผลต่างเฉลี่ย (Average Magnitude Difference Function, AMDF) และ Real Cepstrum ในงานวิจัยนี้เลือกใช้กรรมวิธี Real Cepstrum ที่ใช้เวลาในการคำนวณน้อยและให้ความแม่นยำมาก (เอกฤทธิ์ มณีน้อย, 2541) โดยนำค่าสัมประสิทธิ์เซปสตรัมที่คำนวณจาก การแปลงดีสครีตฟูริเยร์ (แปลงจาก Time Domain ให้อยู่ในรูปของ Frequency Domain) คำนวณลอการิทึม และแปลงดีสครีตฟูริเยร์แบบผกผัน (Inverse Discrete Fourier Transform - IDFT) สัญญาณที่ได้จะอยู่ในรูปของ Quefrency Domain แล้วนำมาผ่านการแยกองค์ประกอบออกเป็น 2 องค์ประกอบ คือLow-quefrency Elements และ High-quefrency Elements โดยค่า Low-quefrency จะคำนวณจากข้อมูลที่อยู่ในช่วง Low-quefrency Region ในช่วงตั้งแต่ 0 ถึง 2 หรือ 4 มิลลิวินาที ดังรูปที่ 2.10 (Furui, 1989)



รูปที่ 2.10 กรรมวิธีการหาความถี่มูลฐาน (Furui, 1989)

2.3.4 **ความถี่ฟอร์แมนท์** (Formant Frequency)

เกิดจากการเรโซแนนซ์ของช่องทางเดินเสียง (Vocal Tract) โดยช่องทางเดินเสียงจะทำหน้าที่เหมือนกับท่ออากาศและมีความถี่ธรรมชาติค่าหนึ่ง ซึ่งจะตอบสนองต่อคลื่นเสียงที่มีความถี่ตรงกับความถี่ธรรมชาติของช่องทางเดินเสียง และกำเนิดสเปกตรัมที่มีจุดยอดอยู่ที่ความถี่ธรรมชาติของช่องทางเดินเสียง เมื่อท่ออากาศใหญ่ขึ้นหรือเล็กลง การตอบสนองต่อความถี่ก็จะเปลี่ยนแปลงตามไปด้วย สเปกตรัมที่ได้ก็จะมีลักษณะแตกต่างกันไปตามการเปลี่ยนแปลงรูปร่างของช่องทางเดินเสียง โดยความถี่ที่มีค่าต่ำที่สุดเรียกว่าความถี่ฟอร์แมนท์ที่หนึ่งและเรียกความถี่ที่สูงขึ้นไปว่าความถี่ฟอร์แมนท์ที่สองและสามตามลำดับ

2.3.5 **สัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น** (Linear Frequency Cepstrum Coefficients - LFCC)

สัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้นคำนวณดังสมการที่ (2.24) (Davis and Mermelstein, 1980)

$$LFCC_{(i)} = \frac{1}{K}\sum_{m=1}^{K/2-1} \log X_m \cos\left(\frac{2\pi i m}{K}\right) \qquad i = 1,2,...,N \tag{2.24}$$

เมื่อ X_{m} คือค่าความถี่ที่ m ของสัญญาณเสียงที่ผ่านการแปลงดีสครีตฟูริเยร์

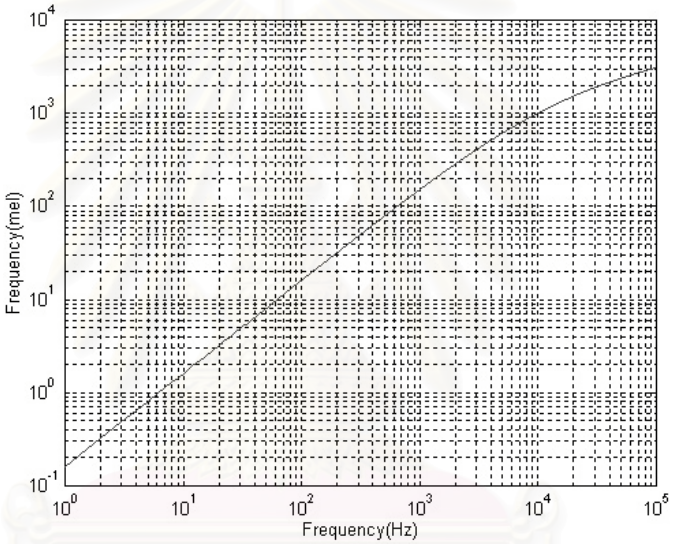
K คือขนาดของการแปลงฟูริเยร์

N คืออันดับของสัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น

2.3.6 สัมประสิทธิ์เซปสตรัมบนความถี่เมล

(Mel Frequency Cepstrum Coefficients - MFCC)

สัมประสิทธิ์เซปสตรัมบนความถี่เมลคำนวณจากการวิเคราะห์นอนพาราเมตริก โดยนำสเปกตรัมสัญญาณผ่านชุดวงจรกรองแบบดิจิตอล (Digital Filter Bank) โดยเลือกใช้การวิเคราะห์แกนความถี่แบบความถี่เมล (Mel Scale) เนื่องจากเป็นความถี่ที่จำลองแบบตามการได้ยินเสียงของหูมนุษย์ซึ่งได้ยินเสียงเป็นเชิงเส้นตั้งแต่ 0-1,000 Hz แต่หลังจากนั้นการได้ยินจะเปลี่ยนเป็นแบบลอการิทึมดังรูปที่ 2.11 แทนด้วยสมการที่ (2.25) โดยนำวงจรกรองแถบผ่านแบบสามเหลี่ยม (Triangular Band Pass Filter) มาใช้ในชุดวงจรกรองแบบดิจิตอลดังรูปที่ 2.12 (Claes and et al., 1998)



รูปที่ 2.11 ความถี่แบบเมล (Tolba and O'Shaughnessy, 1998)

$$mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{2.25}$$

เมื่อ f คือความถี่เชิงเส้น, $mel(f)$ คือความถี่เมล แบนด์วิดท์ของวงจรกรองคำนวณจากสมการดังนี้ (Tolba and O'Shaughnessy, 1998)

$$BW_{critical} = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69} \tag{2.26}$$

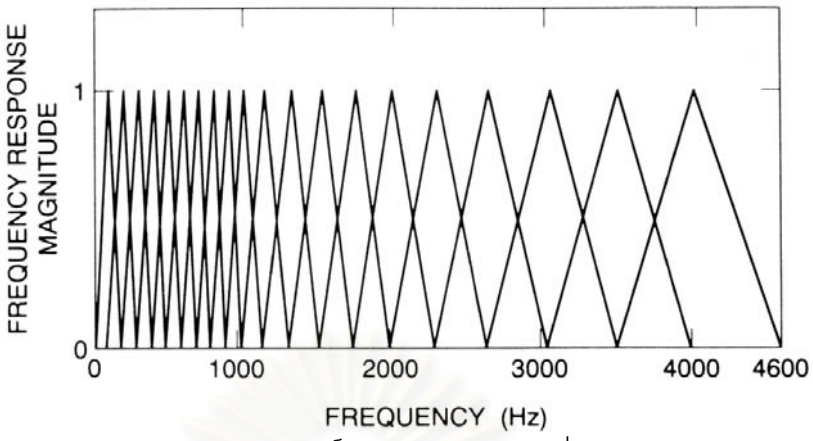
ค่าสัมประสิทธิ์เซปสตรัมบนความถี่เมลคำนวณจากสมการที่ (2.27) (Loizou, 1995)

$$m_i = \sum_{k=1}^{P} Y_k \cos\left[k(i-\frac{1}{2})\frac{\pi}{P}\right] \qquad i = 1,2,\ldots,N \tag{2.27}$$

เมื่อ N คืออันดับของสัมประสิทธิ์เซปสตรัมบนความถี่เมล

P คือจำนวนของวงจรกรองที่ใช้ในชุดวงจรกรอง

Y_k คือค่าลอการิทึมของพลังงานที่ผ่านชุดวงจรกรอง โดยที่ $k = 1,2,\ldots,P$



รูปที่ 2.12 ชุดวงจรกรองสำหรับสัมประสิทธิ์เซปสตรัมบนความถี่เมล (Rabiner and Juang, 1993)

2.3.7 สัมประสิทธิ์เซปสตรัมบนความถี่เมลแบบความต่าง (MFCC Delta) และสัมประสิทธิ์เซปสตรัมบนความถี่เมลแบบบวกความต่าง (MFCC Delta Different)

เป็นการหาค่าความแตกต่างระหว่างกรอบสัญญาณขนาด 5 เฟรมโดยนำค่าสัมประสิทธิ์ของ 2 กรอบสัญญาณทางด้านซ้ายลบด้วยค่าสัมประสิทธิ์ของ 2 กรอบสัญญาณทางด้านขวา ดังสมการที่ (2.28) และ (2.29)

$$w_i(m) = \sum_{k=-2}^{2} kC_{i-k}(m) \tag{2.28}$$

$$x_i(m) = \sum_{k=-2}^{2} kw_{i-k}(m) \tag{2.29}$$

เมื่อ $m = 1,...,M$ โดยที่ M คือค่าอันดับของสัมประสิทธิ์ MFCC

$i = 2,...,N-2$ โดยที่ N คือจำนวนเฟรมของข้อมูลเสียง

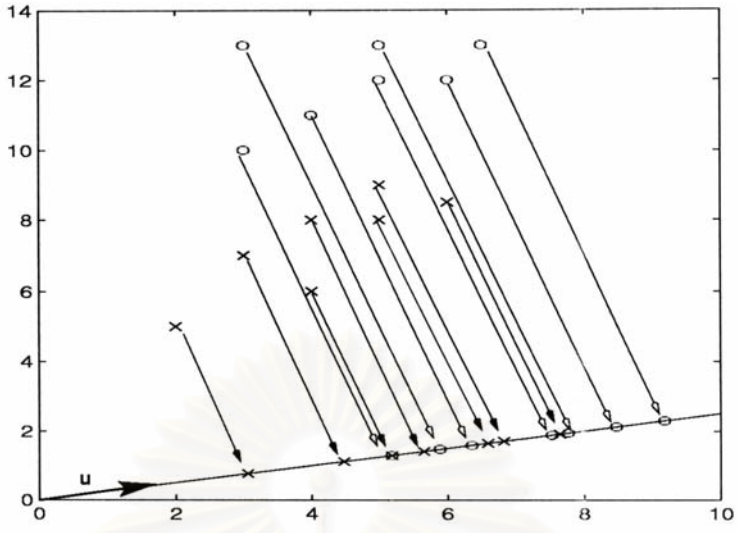
$C_i(m)$ คือสัมประสิทธิ์ MFCC

$w_i(m)$ คือสัมประสิทธิ์ MFCC Delta

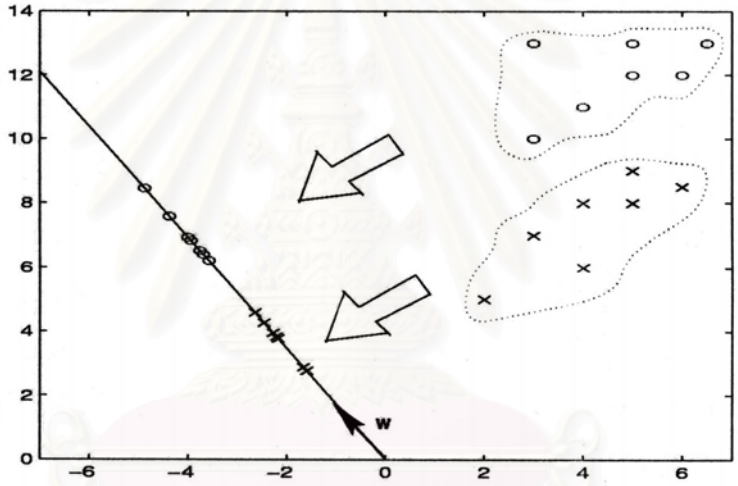
และ $x_i(m)$ คือสัมประสิทธิ์ MFCC Delta Different

2.3.8 วิธีปริภูมิย่อย (Subspace Approach)

วิธีปริภูมิย่อยเป็นวิธีการหาเมตริกซ์การแปลง (Transformation Matrix) โดยฉาย (Project) เวกเตอร์ข้อมูลลงไปบนปริภูมิย่อยเพื่อทำให้ข้อมูลที่อยู่ในกลุ่มเดียวกันเข้ามาอยู่ใกล้กันและข้อมูลที่อยู่นอกกลุ่มอยู่ห่างออกไปดังรูปที่ 2.13 วิธีการดังกล่าวสามารถแยกข้อมูลที่ไม่ต้องการออกจากกันได้อย่างชัดเจน (Schalkoff, 1992) ดังนั้นในงานวิจัยนี้จึงนำวิธีการหาขอบเขตของข้อมูลที่ต้องการแยกมาใช้ 2 วิธีคือ วิธี Fisher's Linear Discriminant และวิธี Divergence Measures เพื่อหาเมตริกซ์การแปลงที่บ่งความต่างได้มากที่สุด



รูปที่ 2.13 (ก) การฉายข้อมูลลงบนเวกเตอร์ $\underline{u}$



รูปที่ 2.13 (ข) การฉายข้อมูลลงบนเวกเตอร์ $\underline{w}$

2.3.8.1 Fisher's Linear Discriminant

การฉายข้อมูลลงบนปริภูมิย่อยใดๆ ในรูปที่ 2.13 อาจทำให้เกิดความสับสนได้ถ้าเลือกทิศทางของเวกเตอร์ $\underline{u}$ ได้ไม่ดีพอดังรูปที่ 2.13 (ก) ตรงกันข้ามกับเวกเตอร์ $\underline{w}$ ที่ทำให้ข้อมูลแยกจากกันได้อย่างชัดเจนดังรูปที่ 2.13 (ข) ดังนั้นฟิชเชอร์จึงนำเสนอ Fisher Linear Discriminant Function $\underline{w}^T \underline{x}$ เพื่อฉายข้อมูล $\underline{x}$ ลงไปบนเวกเตอร์ $\underline{w}$

สมมุติให้กลุ่มข้อมูล C ประกอบด้วยข้อมูล 2 กลุ่มคือ C_1 และ C_2

$$C = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_n\} = \{C_1, C_2\} \tag{2.30}$$

ข้อมูลถูกฉายลงบนปริภูมิย่อย

$$y_i = \underline{w}^T \underline{x}_i = \langle \underline{w}, \underline{x}_i \rangle \qquad i = 1,2,...,n \tag{2.31}$$

คำนวณหาเวกเตอร์ $\underline{w}$ ด้วยวิธี Fisher's Criterion คือหาค่าที่มากที่สุดของ $J(\underline{w})$ ดังสมการที่ (2.32) โดยที่ $J(\underline{w})$ คืออัตราส่วนระหว่าง ค่าความแตกต่างของค่าเฉลี่ยในแต่ละกลุ่มข้อมูล กับ ความแปรปรวนรวมของข้อมูล

$$J(\underline{w}) = \frac{\underline{w}^T S_B \underline{w}}{\underline{w}^T S_W \underline{w}} \tag{2.32}$$

$$S_B = (\underline{m}_1 - \underline{m}_2)(\underline{m}_1 - \underline{m}_2)^T \tag{2.33}$$

$$S_w = \sum_{\underline{x} \in C_1} (\underline{x} - \underline{m}_1)(\underline{x} - \underline{m}_1)^T + \sum_{\underline{x} \in C_2} (\underline{x} - \underline{m}_2)(\underline{x} - \underline{m}_2)^T \tag{2.34}$$

เมื่อ S_B และ S_W เป็น *Between-class* และ *Within-class Scatter Matrices* ตามลำดับ, C_1 คือกลุ่มข้อมูลกลุ่มที่1, C_2 คือกลุ่มข้อมูลกลุ่มที่2, $\underline{m}_1$ คือค่าเฉลี่ย (Mean) ของข้อมูลกลุ่มที่ 1, $\underline{m}_2$ คือค่าเฉลี่ยของข้อมูลกลุ่มที่ 2

2.3.8.2 Divergence Measures

พิจารณาข้อมูล 2 กลุ่ม C_1 และ C_2 มีความหนาแน่นของความน่าจะเป็น (Probability Densities) $p_1(x) = p(x|w_1)$ และ $p_2(x) = p(x|w_2)$ ตามลำดับ ค่าระยะทางระหว่างข้อมูล 2 กลุ่มวัดด้วยไดเวอร์เจนท์ดังสมการที่ (2.35)

$$J_x(1,2) = \int_x [p_1(x) - p_2(x)] \ln \left[\frac{p_1(x)}{p_2(x)} \right] dx \tag{2.35}$$

จากสมการที่ (2.35) ให้ข้อมูล n ใน C_1 และ C_2 กระจายตัวแบบปกติ (Normal Distributions) ด้วยค่าเฉลี่ย μ_1 และ μ_2 และมีค่าเมตริกซ์ความแปรปรวนร่วม (Covariance Matrices) Σ_1 และ Σ_2 จะได้

$$J_x(1,2) = \frac{1}{2} tr\left[\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1\right] - n + \frac{1}{2} tr\left[\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)\Delta\right] \tag{2.36}$$

$$\Delta = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{2.37}$$

ให้เมตริกซ์ A เป็นเมตริกซ์การแปลงโดยที่ $y = A^T x$ ขณะที่ x เป็นเวกเตอร์ขนาด $n \times 1$, เวกเตอร์ y มีขนาด $m \times 1$ โดยที่ $m < n$

เมื่อ tr คือ ผลรวมในแนวทแยงมุม ดังนั้นไดเวอร์เจนท์ของข้อมูลที่ถูกฉายลงบนปริภูมิย่อยคือ

$$J_y(1,2) = \frac{1}{2} tr\left[D_1^{-1} D_2 + D_2^{-1} D_1\right] - m + \frac{1}{2} tr\left[\left(D_1^{-1} + D_2^{-1}\right)\Delta_1\right] \tag{2.38}$$

โดยที่ $D_1 = A^T \Sigma_1 A$, $D_2 = A^T \Sigma_2 A$ และ $\Delta_1 = A^T \Delta A$

หาค่าเมตริกซ์ A ที่ทำให้ค่า $J_y(1,2)$ มีค่ามากที่สุดโดยให้ $\frac{\partial J_y(1,2)}{\partial A} = 0$

$$\frac{\partial J_y(1,2)}{\partial A} = \Sigma_1 A\left(D_2^{-1} - D_1^{-1}(D_2+\Delta_1)D_1^{-1}\right) + \Sigma_2 A\left(D_1^{-1} - D_2^{-1}(D_1+\Delta_1)D_2^{-1}\right) + \Delta A\left(D_1^{-1} + D_2^{-1}\right) \quad (2.39)$$

จากวิธี Steepest Ascent Method หาคำตอบของเมตริกซ์ A โดยให้ค่าเมตริกซ์เริ่มต้นเป็น $A^{(0)}$ และวนซ้ำปรับเปลี่ยนค่า $A^{(k)}$ ดังนี้

$$A^{(k+1)} = A^{(k)} + \gamma \frac{\partial J_y(1,2)}{\partial A^{(k)}} \qquad k = 0,1,2,... \quad (2.40)$$

โดยค่า k คือจำนวนรอบของการวนซ้ำ, γ คือค่าคงที่ และค่า $A^{(0)}$ คำนวณจากค่าเวกเตอร์เจาะจง (Eigen Vector) m ค่าของ $\Sigma_2^{-1}\Sigma_1$ ภายใต้เงื่อนไข

$$\lambda_i + \frac{1}{\lambda_i} \geq \lambda_j + \frac{1}{\lambda_j} \qquad i = 1,2,...,m \quad \text{และ} \quad j = 1,2,...,n \ : i \neq j \quad (2.41)$$

เมื่อได้ค่า A แล้วนำไปแทนในสมการที่ (2.39) (Loizou, 1995) จะได้

$$\frac{\partial J_y(1,2)}{\partial A} \approx \Sigma_1 A\left(E_2 - E_3\Delta_1 E_3\right) + \Sigma_2 A\left(E_1 - E_4\Delta_1 E_4\right) + \Delta A\left(E_3 + E_4\right) \quad (2.42)$$

$$(E_1)_{ii} = \frac{1}{S_1(i)} - \frac{S_1(i)}{S_2^2(i)} \quad (2.43)$$

$$(E_2)_{ii} = \frac{1}{S_2(i)} - \frac{S_2(i)}{S_1^2(i)} \quad (2.44)$$

$$(E_3)_{ii} = \frac{1}{S_1(i)} \quad (2.45)$$

$$(E_4)_{ii} = \frac{1}{S_2(i)} \quad (2.46)$$

โดยที่ $S_1(i) = \left(A^T\Sigma_1 A\right)_{ii}$, $S_2(i) = \left(A^T\Sigma_2 A\right)_{ii}$, $i = 1,2,...,m$ โดยที่ E_{ii} คือค่าในแนวทแยงมุมของเมตริกซ์ E แถวที่ i

เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยการรู้จำเสียงภาษาไทย

ในระยะแรกการรู้จำเสียงใช้กรรมวิธีการเปรียบเทียบทางเวลาแบบพลวัต (Dynamic Time Warping, DTW)

-ระพีพัฒน์ เพ็ญศิริ ทำการวิจัยรู้จำเสียงตัวเลข (0-9) ใช้กรรมวิธีการเปรียบเทียบทางเวลาแบบพลวัต ได้ผลการรู้จำร้อยละ 79.25 (ระพีพัฒน์ เพ็ญศิริ, 2538)

-ธีระ ภัทราพรนันท์ วิจัยการรู้จำเสียงสระโดยมีการแปลงดีสครีตฟูริเยร์ และ ค่าลอการิทึมของพลังงาน เป็นค่าลักษณะสำคัญ ใช้วิธีวัดสเปกตรัมดิสแตนซ์และกรรมวิธีการเปรียบเทียบทาง

เวลาแบบพลวัต ได้อัตราการรู้จำร้อยละ 86.17 สำหรับเสียงสระจำนวน 24 เสียงและร้อยละ 81 สำหรับเสียงวรรณยุกต์ 15 เสียง (ธีระ ภัทราพรนันท์, 2538)

ต่อมาจึงมีการนำวิธีฮิดเดน มาร์คอฟมาใช้

-เสาวลักษณ์ อารีพงศา ทำการวิจัยรู้จำเสียงตัวเลข (0-9) โดยมีสัมประสิทธิ์การทำนายพันธะเชิงเส้น 10 อันดับกับเวกเตอร์ชุดรหัสขนาด 64 ชุดรหัสคำ เป็นค่าลักษณะสำคัญ ได้อัตราการรู้จำร้อยละ 82 (เสาวลักษณ์ อารีพงศา, 2538)

-วิศรุต อาขุบุตร ทำงานวิจัยรู้จำคำไทยหลายพยางค์ซึ่งเป็นคำพยางค์เดียว สองพยางค์และสามพยางค์ 70 คำ มีสัมประสิทธิ์การทำนายพันธะเชิงเส้น 10 อันดับ กับเวกเตอร์ชุดรหัสขนาด 128, 256 เป็นค่าลักษณะสำคัญ ได้อัตรารู้จำร้อยละ 89.91 (วิศรุต อาขุบุตร, 2539)

-Tungthangthum ทำงานวิจัยรู้จำเสียงสระภาษาไทย ใช้ค่าความถี่มูลฐานร่วมกับความถี่ฟอร์แมนท์เป็นลักษณะสำคัญ ได้อัตรารู้จำร้อยละ 91 (Tungthangthum, 1998)

จากงานข้างต้นพบว่าอัตราการรู้จำเพิ่มสูงขึ้นเมื่อนำวิธีฮิดเดน มาร์คอฟมาใช้ ต่อมาเริ่มมีการนำวิธีโครงข่ายประสาทเทียมมาใช้เช่น

-ในงานวิจัยของเสรี ปานซาง ทำการรู้จำเสียงพูดตัวเลขไทยเฉพาะบุคคล ใช้การวิเคราะห์สเปกโตรแกรม ความถี่ฮาร์โมนิค-เวลา-พลังงาน เป็นลักษณะสำคัญ ได้อัตรารู้จำร้อยละ 87.5 (เสรี ปานซาง, 2540)

-งานวิจัยของไชยันต์ สุวรรณชีวะศิริ ทำการรู้จำเสียงพูดตัวเลขภาษาไทยแบบหลายผู้พูด ใช้ความถี่ฟอร์แมนท์ สัมประสิทธิ์พหุนามของคาบเวลาพิตช์ และความถี่สเปกตรัม ได้อัตรารู้จำร้อยละ 91.6 (ไชยันต์ สุวรรณชีวะศิริ, 2541)

-งานวิจัยของวุฒิพงษ์ พรสุขจันทราทำการรู้จำเสียงตัวเลขไทย โดยใช้สัมประสิทธิ์การทำนายพันธะเชิงเส้น 10 อันดับ เป็นค่าลักษณะสำคัญและใช้วิธีโครงข่ายประสาทเทียมแบบแบ็กพรอพาเกชัน ได้อัตรารู้จำร้อยละ 89.40 (วุฒิพงษ์ พรสุขจันทรา, 2539)

จะเห็นได้ว่าอัตราการรู้จำเพิ่มสูงขึ้นเล็กน้อยเมื่อเทียบกับงานรู้จำเสียงตัวเลขโดยวิธีฮิดเดน มาร์คอฟ ส่วนเทคนิคแบบฟัซซีได้ถูกนำมาใช้ด้วยเช่นกันในงานวิจัยของชัย วุฒิวิวัฒน์ชัย โดยใช้ร่วมกับโครงข่ายประสาทเทียม

-ชัย วุฒิวิวัฒน์ชัย ทำการวิจัยรู้จำคำพยางค์เดียว สองพยางค์และสามพยางค์ซึ่งเป็นชุดคำเดียวกับงานของวิศรุต อาขุบุตร (2539) ใช้สัมประสิทธิ์การทำนายพันธะเชิงเส้น 10 อันดับเป็นค่าลักษณะสำคัญ ได้อัตราการรู้จำร้อยละ 91.90 (ชัย วุฒิวิวัฒน์ชัย, 2540)

จากงานที่ผ่านมาพบว่าการรู้จำคำเป็นพยางค์ต้องอาศัยการเก็บข้อมูลเป็นจำนวนมาก จึงมีแนวความคิดในการนำหน่วยเสียงมาใช้ในงานรู้จำเพื่อลดการจัดเก็บข้อมูลจำนวนมากและ

สามารถเพิ่มเติมจำนวนคำศัพท์จากหน่วยเสียงที่มีอยู่ได้โดยไม่ต้องทำการฝึกฝนแบบจำลองใหม่ทั้งหมด เพียงแต่เพิ่มแบบรูปการประกอบคำของหน่วยเสียงเท่านั้น

-เอกฤทธิ์ มณีน้อย ทำการวิจัยรู้จำหน่วยเสียงสระภาษาไทยจำนวน 9 เสียง โดยมีค่าสัมประสิทธิ์การทำนายพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัม ความถี่ฟอร์แมนท์และ ความเข้มของสเปกตรัมเป็นค่าลักษณะสำคัญ ใช้โครงข่ายประสาทเทียมในการรู้จำ ได้อัตราการรู้จำร้อยละ 90.34 (เอกฤทธิ์ มณีน้อย, 2541)

งานรู้จำหน่วยเสียงสระของเอกฤทธิ์ มณีน้อย สามารถรู้จำเสียงสระในคำพูด สามารถประยุกต์ใช้ในงานรู้จำคำพูดต่อเนื่อง (Continuous Speech) ได้โดยอาศัยแบบรูปการประกอบคำตามหลักภาษาศาสตร์ แต่เมื่อมีชื่อเฉพาะที่ไม่เป็นไปตามหลักภาษาหรือคำพ้องเสียงจะไม่สามารถระบุได้ว่าเป็นคำใด ดังนั้นในงานวิจัยนี้จึงวิเคราะห์เสียงคำเรียกตัวอักษรไทยโดยพิจารณาเฉพาะเสียงพยัญชนะไทยเพื่อสามารถพัฒนาใช้ร่วมกับระบบการรู้จำคำต่อไป

ตัวอย่างงานรู้จำเสียงคำเรียกตัวอักษรในภาษาต่างประเทศ

งานวิจัยการรู้จำเสียงคำเรียกตัวอักษรในภาษาต่างประเทศ เช่น ภาษาญี่ปุ่น อังกฤษ และจีน

-Itakura (Itakura, 1975) นำ DTW มาใช้ในงานวิจัยรู้จำเสียงคำเรียกตัวอักษร (Alphabet Recognition) ในการรู้จำชื่อเมือง 200 เมือง แบบขึ้นกับผู้พูดในภาษาญี่ปุ่นและได้ความถูกต้องในการจำแนก (Classification) ร้อยละ 98.3 แต่เมื่อนำมาใช้กับ Alphadigits ซึ่งประกอบด้วย Alphabet, Digits (0-9) และ 3 Control words (Yes, No, O.K.) ความสามารถจะลดลงเหลือเพียงร้อยละ 88 เท่านั้น ซึ่งแสดงให้เห็นถึงผลกระทบของเสียงคำเรียกตัวอักษรที่มีความคล้ายคลึงกัน

-Fanty, et al. ทำงานวิจัยการรู้จำเสียงคำเรียกตัวอักษรภาษาอังกฤษด้วยวิธีโครงข่ายประสาทเทียม ได้อัตราการรู้จำร้อยละ 89 (Fanty et al., 1993)

-Loizou และ Spanias วิจัยการรู้จำเสียงคำเรียกตัวอักษรภาษาอังกฤษด้วยวิธีฮิดเดน มาร์คอฟ ได้อัตราการรู้จำร้อยละ 85.0 (Loizou and Spanias, 1996)

-Junqua, et al. วิจัยการรู้จำเสียงคำเรียกตัวอักษรภาษาอังกฤษด้วยวิธีฮิดเดน มาร์คอฟ ได้อัตรารู้จำร้อยละ 90 (Claude, et al., 1997)

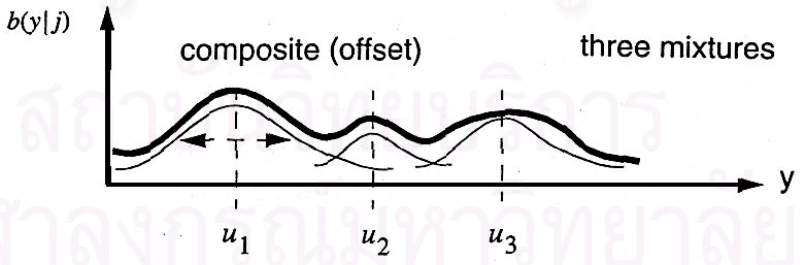
-Jun, Zuoying และ Yansong ทำงานวิจัยการรู้จำเสียงคำเรียกตัวอักษรภาษาจีนด้วยวิธีฮิดเดน มาร์คอฟ ได้อัตรารู้จำร้อยละ 86.3 (Jun, 1994)

จากงานวิจัยข้างต้นพบว่าวิธีโครงข่ายประสาทเทียมมีอัตราการรู้จำที่น้อยกว่าวิธีฮิดเดน มาร์คอฟเพราะวิธีโครงข่ายประสาทเทียมไม่สามารถมองเห็นความแตกต่างของหน่วยเสียง

ในตัวอักษรได้ เนื่องจากมีการกำหนดโนดอินพุตคงที่และมีการปรับความยาวเสียงที่มีความสั้นยาวต่างกันให้มีความยาวเท่ากันโดยมองเสียงทั้งเสียงเป็นหนึ่งคำ ตรงกันข้ามกับวิธีฮิดเดน มาร์คอฟที่สามารถรู้จำเสียงที่มีความยาวแตกต่างกันได้และแยกความแตกต่างของหน่วยเสียงได้โดยวิธีการกำหนดจำนวนสถานะของระบบให้ปรับเปลี่ยนตามความยาวของหน่วยเสียง ยกตัวอย่างหน่วยเสียง เช่น เสียง B แยกหน่วยเสียงได้เป็น /b iy/, เสียง H แยกได้ /ey tcl ch/, เสียง Q แยกได้ /k y uw/ เป็นต้น

ในงานวิจัยนี้ได้ทำการทดสอบเบื้องต้นโดยใช้วิธีโครงข่ายประสาทเทียมและวิธีฮิดเดน มาร์คอฟเพื่อทดสอบผลของการรู้จำเสียงพยัญชนะไทย 28 เสียงของผู้พูดคนเดียว พบว่าวิธีฮิดเดน มาร์คอฟให้ผลการรู้จำร้อยละ 30 ส่วนวิธีโครงข่ายประสาทเทียมให้ผลการรู้จำร้อยละ 13 จึงเลือกใช้แบบจำลองฮิดเดน มาร์คอฟ โดยแบบจำลองฮิดเดน มาร์คอฟมีทั้งแบบต่อเนื่อง (Continuous Hidden Markov Model – CHMM) และแบบไม่ต่อเนื่อง (Discrete Hidden Markov Model - DHMM)

ในงานวิจัยนี้เลือกใช้แบบจำลองฮิดเดน มาร์คอฟแบบต่อเนื่อง เนื่องจากคุณสมบัติของลักษณะสำคัญที่มีความต่อเนื่อง ถ้านำมาใช้กับแบบจำลองฮิดเดน มาร์คอฟแบบไม่ต่อเนื่อง ต้องทำการปรับขนาดของข้อมูลด้วยการทำเวกเตอร์ควอนไทซ์ (Vector Quantize - VQ) ต่างกับแบบจำลองฮิดเดน มาร์คอฟแบบต่อเนื่อง ซึ่งใช้การประสมเชิงเส้นแบบให้น้ำหนัก (Weighted Linear Combination) ของการแจกแจงแบบเกาส์ (Gaussians Distribution) หรือเรียกว่า Gaussians Mixture ในการเพิ่มจำนวนของ Gaussians Mixture เป็นการเพิ่มความซับซ้อนให้กับการแทนข้อมูลเสียง เพื่อทำให้ได้ข้อมูลที่ยังคงลักษณะสำคัญของข้อมูลเสียงได้ครบถ้วนมากขึ้น



รูปที่ 2.14 การประสมเชิงเส้นแบบให้น้ำหนักของการแจกแจงแบบเกาส์

บทที่ 3
วิธีดำเนินการวิจัย

วิธีดำเนินการวิจัยในบทนี้กล่าวถึงรายละเอียดของการเลือกเสียงตัวอย่าง การเก็บข้อมูลเสียง ขั้นตอนการรู้จำเสียงคำเรียกพยัญชนะ วิธีการวัดลักษณะสำคัญแบบต่างๆ วิธีการฝึกฝนและทดสอบระบบ

3.1 การกำหนดเสียงตัวอย่าง

เนื่องจากคำเรียกพยัญชนะไทยสามารถเรียกได้หลายแบบ ดังนั้นในงานวิจัยนี้จึงสำรวจวิธีการเรียกพยัญชนะไทยของคน 40 คนช่วงอายุระหว่าง 15-60 ปีด้วยแบบสอบถาม โดยให้ผู้ตอบแบบสอบถามกรอกคำอ่านหรือวิธีที่ใช้ในการเรียกพยัญชนะไทยทั้ง 44 ตัว พบเสียงคำเรียกแตกต่างกันยกตัวอย่างดังตารางที่ 3.1 โดยเสียงสระออจะเป็นเสียงที่ทุกคนใช้ และจากวัตถุประสงค์ที่ต้องการนำไปใช้ร่วมกับระบบรู้จำคำพูด จึงเลือกเสียงที่สั้นที่สุดและลงท้ายด้วยสระออเป็นเสียงตัวอย่าง (ตารางที่ 2.2)

ตารางที่ 3.1 ตัวอย่างคำเรียกพยัญชนะไทย

พยัญชนะ	คำเรียกแบบที่ 1	คำเรียกแบบที่ 2
ญ	ยอ-หญิง	ยอ-ผู้-หญิง
ฐ	ถอ-ถาน	ถอ-สัน-ถาน
ฑ	ทอ-มน-โท	ทอ-นาง-มน-โท
ศ	สอ-คอ	สอ-สา-ลา

3.1.1 เครื่องมือที่ใช้ในการวิจัย

1. การ์ดเสียง Sound Blaster 16 ของบริษัท Creative Technology
2. ไมโครโฟน Sony Uni-directional Dynamic Microphone Impedance 600 โอห์ม
3. โปรแกรม Goldwave Version 4.23
4. ระบบปฏิบัติการ Microsoft Window 2000
5. ระบบปฏิบัติการ MS-DOS Version 5.00.2195

3.1.2 การเก็บรวบรวมข้อมูล

ข้อมูลเสียงจำนวน 1680 เสียง จากผู้พูดจำนวน 60 คนแบ่งเป็นชาย 33 คน หญิง 27 คน อายุระหว่าง 15-60 ปี โดยบันทึกเสียงคำเรียกพยัญชนะไทย 28 เสียงต่อผู้พูด 1 คน ซึ่งผู้พูดเป็นคนละชุดกับผู้กรอกแบบสอบถามดังตารางที่ 3.1

หลักเกณฑ์ในการบันทึกเสียง

1. บันทึกเสียงพูดด้วยไมโครโฟนผ่านการ์ดเสียง โดยมีอัตราการชักตัวอย่างที่ 11025 เฮิรตซ์
2. ผู้พูดเป็นผู้ที่มีการออกเสียงเป็นปกติและใช้ภาษาไทยสำเนียงกรุงเทพฯ เป็นภาษาพูด
3. บันทึกเสียงภายใต้สภาพแวดล้อมของที่ทำงานทั่วไป

ข้อมูลเสียงถูกแบ่งเป็น 2 ชุดคือ ชุดฝึกฝน (Training Set) และชุดทดสอบ (Testing Set) ชุดฝึกฝนประกอบด้วยเสียงของผู้พูดจำนวน 40 คนแบ่งเป็นชาย 22 คน หญิง 18 คนอายุระหว่าง 16-60 ปี นำไปสร้างและฝึกฝนระบบ ส่วนชุดทดสอบอีก 20 คนประกอบด้วยชาย 11 คน และหญิง 9 คน อายุระหว่าง 19-60 ปี นำไปทดสอบระบบรู้จำเสียงคำเรียกพยัญชนะไทย

3.2 ระบบรู้จำเสียงคำเรียกพยัญชนะ

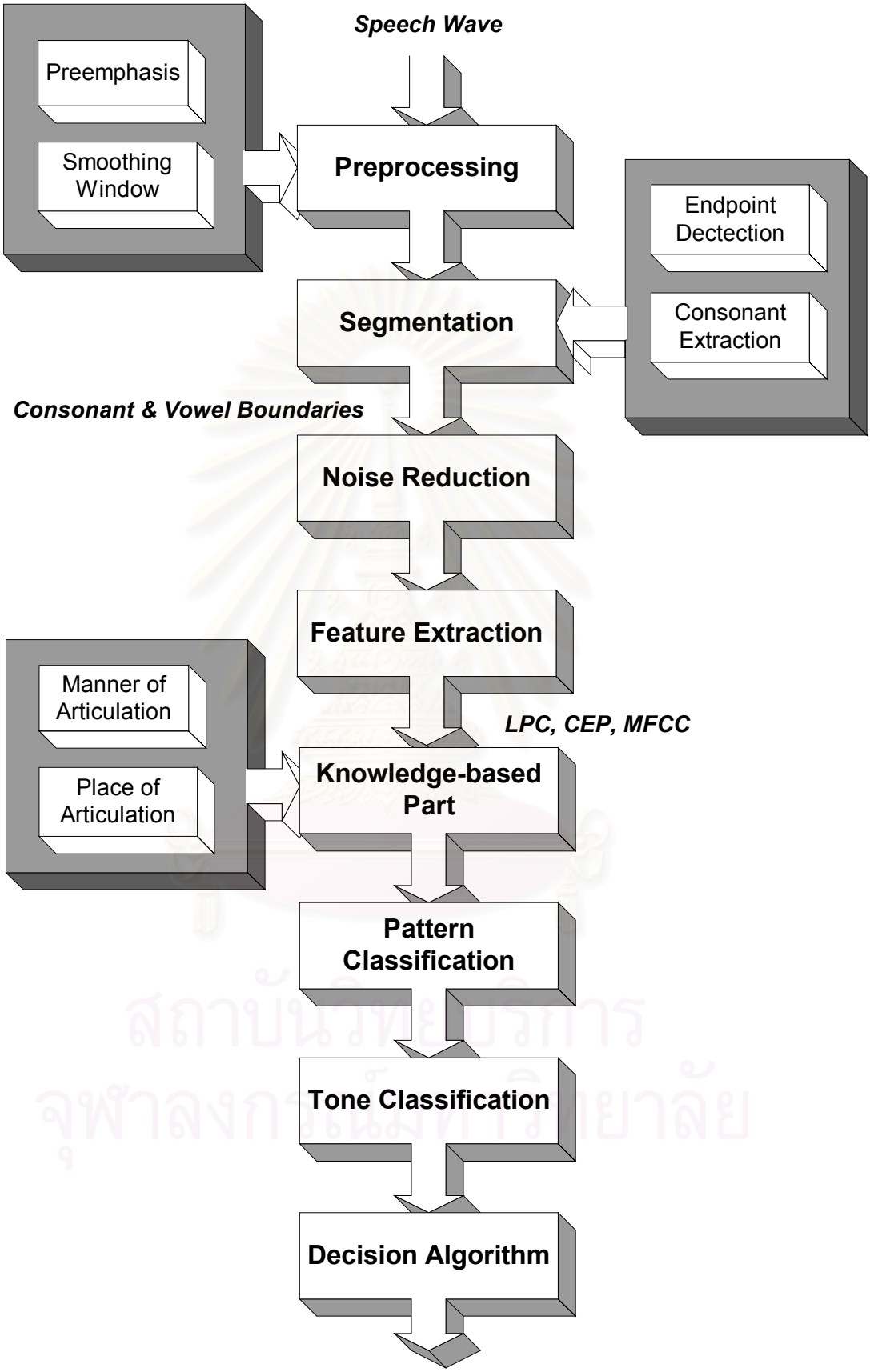
การรู้จำเสียงคำเรียกพยัญชนะแบ่งออกเป็น ส่วนการประมวลผลสัญญาณเบื้องต้น ส่วนการกำหนดขอบเขตข้อมูล ส่วนการขจัดสัญญาณรบกวน ส่วนการวิเคราะห์และคำนวณค่าลักษณะสำคัญ ส่วนขั้นตอนวิธีการฐานความรู้ ส่วนการทดสอบระบบ ส่วนการแยกเสียงวรรณยุกต์ และส่วนของขั้นตอนวิธีการตัดสินใจ ดังรูปที่ 3.1

3.2.1 การประมวลผลสัญญาณเบื้องต้น (Preprocessing)

การประมวลผลสัญญาณเบื้องต้นแบ่งออกเป็น 2 ขั้นตอนคือ

1. ขั้นตอนกรรมวิธีการเน้นล่วงหน้า นำสัญญาณเสียงพูดที่บันทึกผ่านไมโครโฟนมาผ่านกรรมวิธีการเน้นล่วงหน้า เพื่อปรับสเปกตรัมให้มีความเรียบและเพิ่มอัตราส่วนของสัญญาณเสียงพูดต่อสัญญาณรบกวน

2. ขั้นตอนกรรมวิธีการวางกรอบขนาดสัญญาณ โดยเลือกใช้ฟังก์ชันกรอบชนิด Hamming ที่มีความกว้างของกรอบสัญญาณ 25 มิลลิวินาทีต่อ 1 เฟรมและมีการเลื่อนของฟังก์ชันกรอบครั้งละ 10 มิลลิวินาที เพื่อให้การเปลี่ยนแปลงของสัญญาณเป็นไปอย่างช้าๆ หลีกเลี่ยงความไม่ต่อเนื่องของสัญญาณ



รูปที่ 3.1 แผนภาพแสดงขั้นตอนการรู้จำคำเรียกพยัญชนะไทย

3.2.2 การกำหนดขอบเขตข้อมูล (Segmentation)

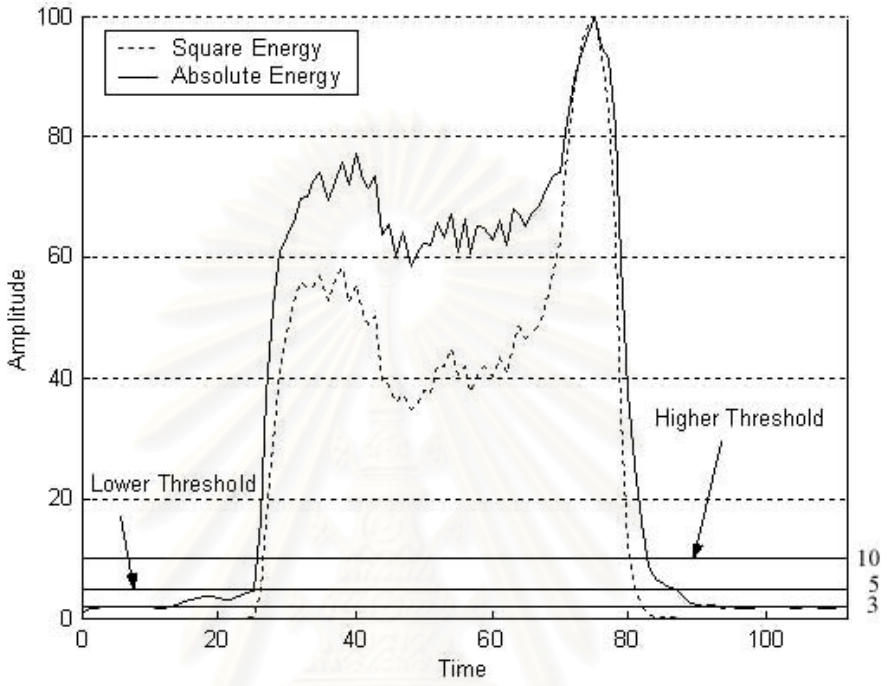
แบ่งเป็น 2 ขั้นตอนคือ ขั้นตอนการหาจุดเริ่มต้นและสิ้นสุดพยางค์ และขั้นตอนการกำหนดขอบเขตหน่วยเสียง

1. ขั้นตอนการหาจุดเริ่มต้นและสิ้นสุดพยางค์ (Endpoint Detection) จากงานวิจัยการรู้จำเสียงคำพยางค์เดียว สองพยางค์และสามพยางค์ (วิศรุต อาชุบุตร, 2539; ชัย วุฒิวิวัฒน์ชัย, 2540) ที่อาศัยแผนภูมิเส้นระดับพลังงาน (Energy Level Contour) ของเสียงพูดโดยการกำหนดจุดเริ่มเปลี่ยนระดับพลังงาน (Lower Threshold) ถ้าพลังงานสูงขึ้นจนถึงระดับพลังงานขั้นสูง (Higher Threshold) ให้กำหนดจุด Lower Threshold เป็นจุดเริ่มต้นพยางค์ โดยเลือกค่าที่ร้อยละ 5 และร้อยละ 10 ของค่าระดับพลังงานสูงสุด ดังรูปที่ 3.2 แต่สำหรับเสียงคำเรียกพยัญชนะไทยในงานวิจัยนี้พบว่าเกิดความสับสนระหว่างสัญญาณรบกวนและเสียงที่มีลักษณะคล้ายสัญญาณรบกวน (Noise-like) เช่นเสียงเสียดแทรก จากรูปที่ 3.2 ข้อมูลเสียงเริ่มตั้งแต่จุดเริ่มเปลี่ยนที่ระดับพลังงานร้อยละ 3 และร้อยละ 5 แต่เมื่อนำไปใช้กับข้อมูลเสียงดังรูปที่ 3.3 กำหนด Lower Threshold และ Higher Threshold คือค่าที่ร้อยละ 3 และร้อยละ 5 ของค่าระดับพลังงานสูงสุด จะได้จุดเริ่มต้นพยางค์ที่ Y แต่จุดเริ่มต้นพยางค์ที่ถูกต้องควรอยู่ที่ X ดังนั้นจึงไม่ควรกำหนดจุดเริ่มเปลี่ยนด้วยค่าคงที่ แต่ควรคำนึงถึงการเปลี่ยนแปลงระดับพลังงาน โดยในงานวิจัยนี้ใช้การวางกรอบสัญญาณขนาด 100 มิลลิวินาที (10 เฟรม) และคำนวณการเปลี่ยนแปลงที่เฟรมแรก และเฟรมสุดท้ายของกรอบสัญญาณ ถ้ามีขนาดเพิ่มขึ้นมากกว่า A เท่า จึงกำหนดเป็นจุดเริ่มเปลี่ยนระดับพลังงาน (กำหนดให้ $A = \frac{\text{ขนาดของแอมพลิจูดที่จุด } X_1 \text{ (ร้อยละ)}}{\text{ขนาดของแอมพลิจูดที่จุด } X_2 \text{ (ร้อยละ)}} = 350$)

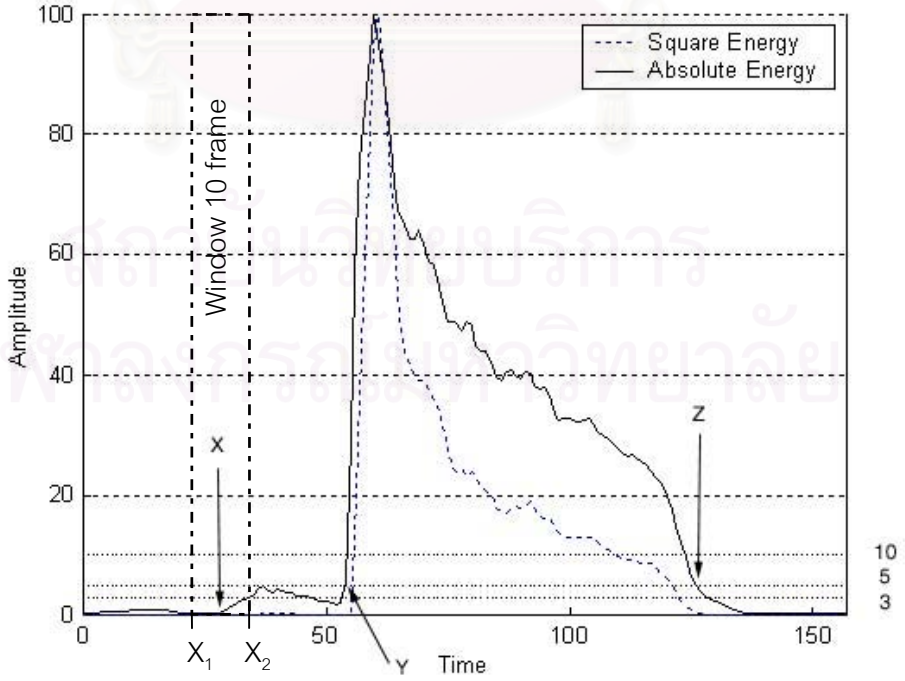
ส่วนการหาจุดสิ้นสุดพยางค์ยังคงใช้การกำหนดจุดเริ่มเปลี่ยนพลังงานที่ร้อยละ 5 และ 10 เช่นเดิม ดังรูปที่ 3.3 เมื่อกำหนดให้ Z คือจุดสิ้นสุดพยางค์เมื่อใช้วิธีเปลี่ยนระดับพลังงาน

2. ขั้นตอนการกำหนดขอบเขตหน่วยเสียง (Consonant Extraction) เนื่องจากแต่ละข้อมูลเสียงประกอบด้วยหน่วยเสียงพยัญชนะ และหน่วยเสียงสระออซึ่งมีความคล้ายคลึงกันในทุกๆ เสียง ดังนั้นในงานวิจัยนี้จึงมีแนวความคิดที่แยกเสียงพยัญชนะออกมาเพื่อหลีกเลี่ยงผลกระทบของหน่วยเสียงสระที่เหมือนกันในทุกๆ ข้อมูล และยังเป็นการลดขนาดของข้อมูลลง รวมทั้งเวลาในการคำนวณอีกด้วย การวิเคราะห์เสียงสระอาศัยคุณสมบัติความเป็นรายคาบของสระร่วมกับความถี่มูลฐาน โดยพิจารณาส่วนของสระจากจุดกึ่งกลางของพยางค์ (เอกฤทธิ์ มณีน้อย, 2541) ถ้าความเป็นรายคาบของสระมีลักษณะเปลี่ยนแปลงเข้าสู่ส่วนของพยัญชนะให้ตัดสินใจเป็นจุดเริ่มต้นเสียงสระดังรูปที่ 3.4 (ค) การตัดสินจุดเปลี่ยนแปลงรายคาบในงานวิจัยนี้ใช้วิธีกำหนดคาบเวลาของสระที่กึ่งกลางพยางค์เป็นค่าระดับกำหนด (H) ถ้าคาบที่อยู่ถัดไปทางซ้าย

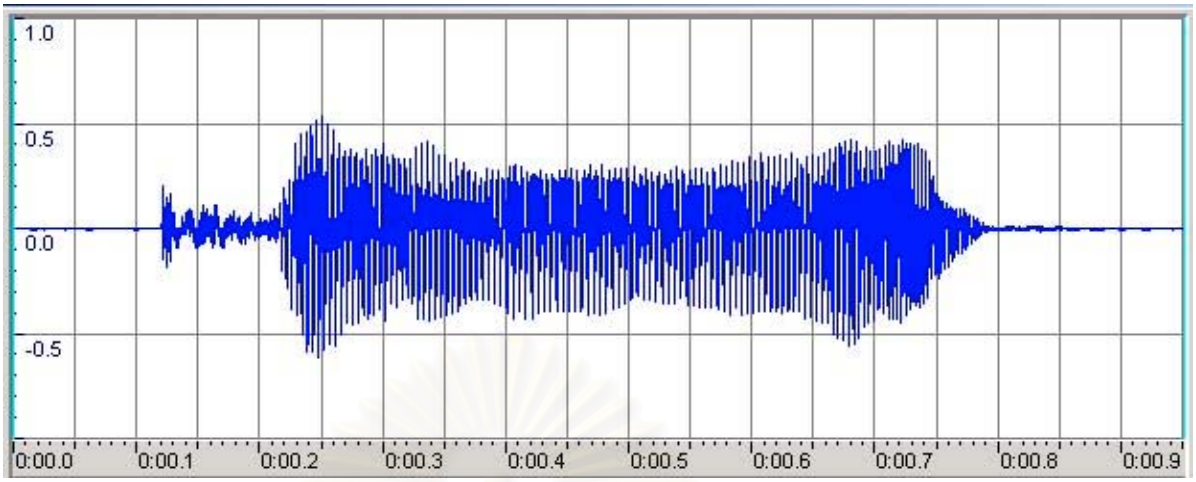
มือมีคาบเวลาไม่อยู่ในช่วง $H \pm V$ และมีความแตกต่างของแอมพลิจูดมากกว่า 3,000 ให้กำหนดเป็นจุดเริ่มต้นเสียงสระ โดยกำหนดให้ V เท่ากับ 3 มิลลิวินาที



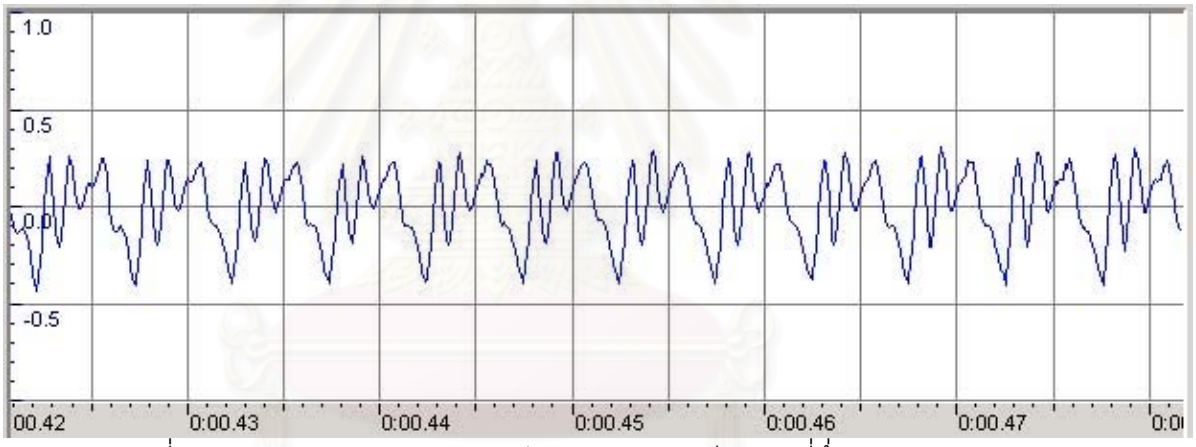
รูปที่ 3.2 แผนภูมิเส้นระดับพลังงาน



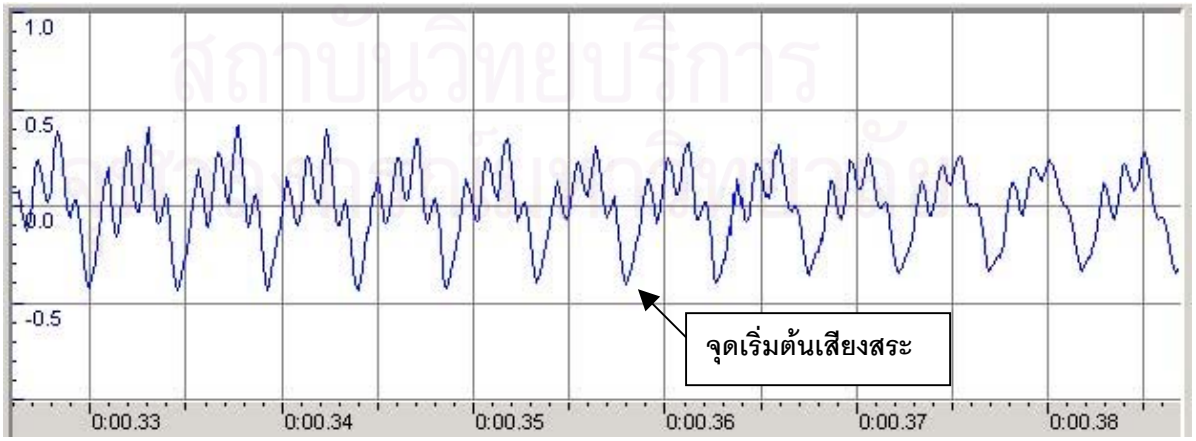
รูปที่ 3.3 แผนภาพแสดงการหาจุดเริ่มต้นและสิ้นสุดพยางค์ของเสียงเสียดแทรก



รูปที่ 3.4 (ก) แผนภาพเสียงที่บันทึกผ่านไมโครโฟน



รูปที่ 3.4 (ข) แผนภาพแสดงความเป็นรายคาบของเสียงสระที่กึ่งกลางพยางค์



รูปที่ 3.4 (ค) แผนภาพแสดงจุดเริ่มต้นเสียงสระ

3.2.3 **การขจัดสัญญาณรบกวน** (Noise Reduction)

ข้อมูลเสียงตัวอย่างถูกบันทึกในสภาพแวดล้อมของที่ทำงาน ซึ่งมีเสียงรบกวนทั้งแบบต่อเนื่องและไม่ต่อเนื่อง แบบต่อเนื่อง เช่น เสียงการทำงานของเครื่องใช้ไฟฟ้า ไมโครโฟนที่ไม่ได้ต่อสายดิน โดยสัญญาณรบกวนจะเกิดต่อเนื่องไปตลอดทั้งข้อมูล แบบไม่ต่อเนื่อง เช่น เสียงการพูดคุย เสียงกริ่งโทรศัพท์ ที่เกิดขึ้นบางช่วงของข้อมูลเป็นต้น จากข้อมูลเสียงตัวอย่างพบว่าสัญญาณรบกวนจากไมโครโฟนที่ไม่ได้ต่อสายดินไว้จะให้สัญญาณที่ความถี่ต่างๆ กันไปตามสถานที่ของแหล่งจำหน่ายไฟ ในงานวิจัยนี้พบสัญญาณรบกวนแบบต่อเนื่องที่ความถี่ 5, 1700, และ 4,000 เฮิรตซ์ ตามลำดับ ดังรูปที่ 3.5 ส่วนเสียงแบบไม่ต่อเนื่องใช้จุดเริ่มต้นของพยางค์ที่คำนวณได้จากขั้นตอนการกำหนดขอบเขตข้อมูลข้างต้นแล้วทำการกำจัดสัญญาณรบกวนที่มีลักษณะเหมือนสัญญาณที่เกิดขึ้นก่อนหน้าจุดเริ่มต้นพยางค์ ดังรูปที่ 3.6 โดยเลือกใช้ฟังก์ชัน Noise Reduction ในโปรแกรม Goldwave Version 4.23 ทำการกำจัดสัญญาณรบกวน

3.2.4 **การสกัดคุณลักษณะสำคัญ** (Feature Extraction)

คุณลักษณะสำคัญที่ใช้ประกอบด้วย ค่าพลังงาน อัตราการตัดผ่านศูนย์ อัตราการตัดผ่านระดับกำหนด ความถี่มูลฐาน สัมประสิทธิ์ LPC, CEPL, CEPF, LFCC, MFCC

ค่าพลังงานถูกนำมาใช้ในขั้นตอนการกำหนดขอบเขตข้อมูลและขั้นตอนการรู้จำเสียงในกระบวนการของการตรวจสอบระดับพลังงาน

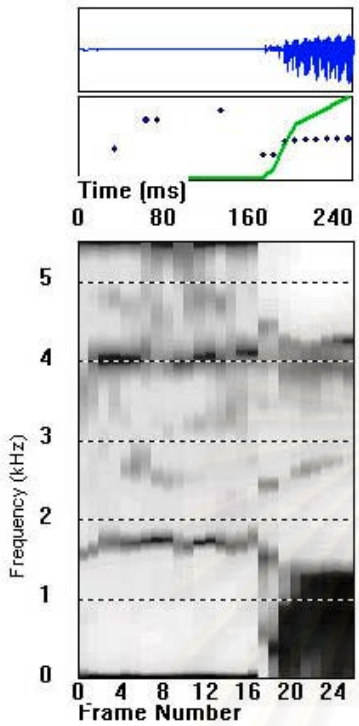
อัตราการตัดผ่านศูนย์บอกถึงความถี่ของข้อมูลเสียงในการตัดผ่านแกนนอน โดยเสียง Noise-like ที่มีลักษณะคล้ายสัญญาณรบกวนจะมีการตัดผ่านแกนศูนย์อย่างรวดเร็ว ทำให้ค่าอัตราการตัดผ่านศูนย์มีค่าสูง

อัตราการตัดผ่านระดับกำหนด (Band Crossing Rate) ถูกนำมาใช้เนื่องจากปัญหาของสัญญาณรบกวนเช่นเดียวกับขั้นตอนการหาจุดเริ่มต้นและสิ้นสุดพยางค์ ดังนั้นระดับกำหนดจึงคำนวณจากระดับของสัญญาณรบกวนซึ่งมีค่าไม่เท่ากันทุกเสียง เนื่องจากสภาพแวดล้อมที่บันทึกต่างกัน โดยคำนวณหาค่าเฉลี่ยของพลังงานจากเฟรมที่ศูนย์ไปจนถึงจุดเริ่มต้นพยางค์เป็นระดับกำหนด

ความถี่มูลฐานใช้บอกถึงการมีคุณลักษณะรายคาบซึ่งเกิดจากการเปิด-ปิดของเส้นเสียงและใช้แยกเสียงจัตวาออกจากเสียงสามัญในส่วนของการแยกเสียงวรรณยุกต์

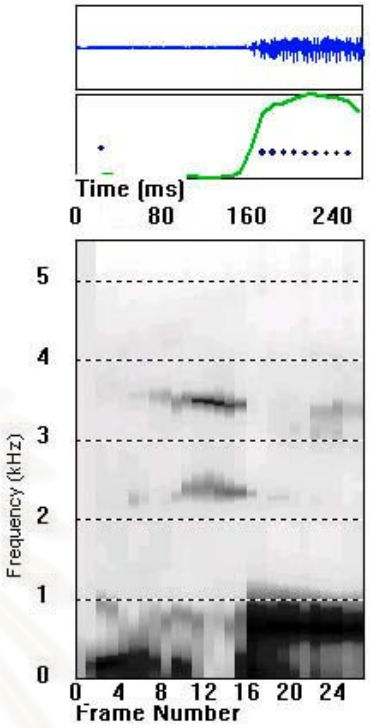
สัมประสิทธิ์ LPC, LFCC, CEPL, CEPF, MFCC นำมาใช้ในส่วนของการจำแนกแบบรูปด้วยแบบจำลองฮิดเดนมาร์คอฟโดยทำการทดสอบปรับค่าอันดับของสัมประสิทธิ์ และจำนวนGaussian Mixture ที่ใช้ในแบบจำลอง

เสียงกอ (/k@@0/) กัก-ไม่ก้อง-ไม่พ่นลม

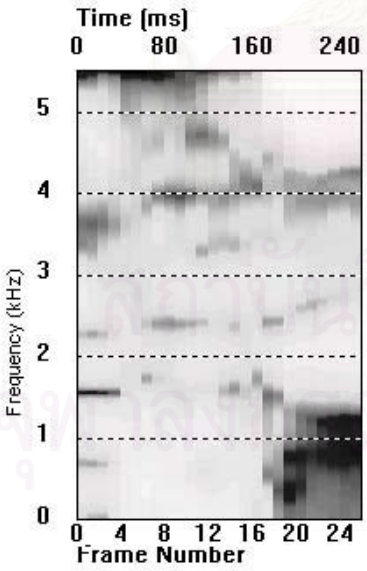


รูปที่ 3.5 (ก) สเปกโตรแกรมแสดงสัญญาณรบกวนที่ 5, 1700 และ 4000 เฮิรตซ์

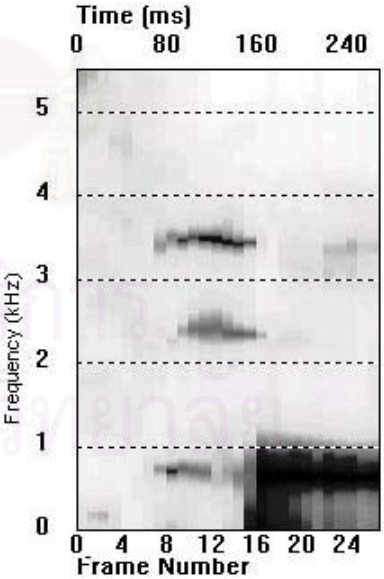
เสียงหอ (/h@@4/) ไม่กัก-ไม่ก้อง-เสียดแทรก



รูปที่ 3.6 (ก) สเปกโตรแกรมแสดงสัญญาณรบกวนที่ 600 เฮิรตซ์



รูปที่ 3.5 (ข) สเปกโตรแกรมของสัญญาณเสียงหลังจากกำจัดสัญญาณรบกวน



รูปที่ 3.6 (ข) สเปกโตรแกรมของสัญญาณเสียงหลังจากกำจัดสัญญาณรบกวน

3.2.5 **ส่วนของฐานความรู้** (Knowledge-based Part)

ขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง

เสียงกัก-ไม่ก้อง-ไม่พ่นลม เกิดจากลมที่ถูกกักไว้ระเบิดออกมา ทำให้แอมพลิจูดของค่าพลังงานมีลักษณะเพิ่มขึ้นอย่างรวดเร็ว ในขณะที่อัตราการตัดผ่านศูนย์มีค่าลดลง และอัตราการตัดผ่านระดับกำหนดมีค่าสูงขึ้น ดังรูปที่ 3.7

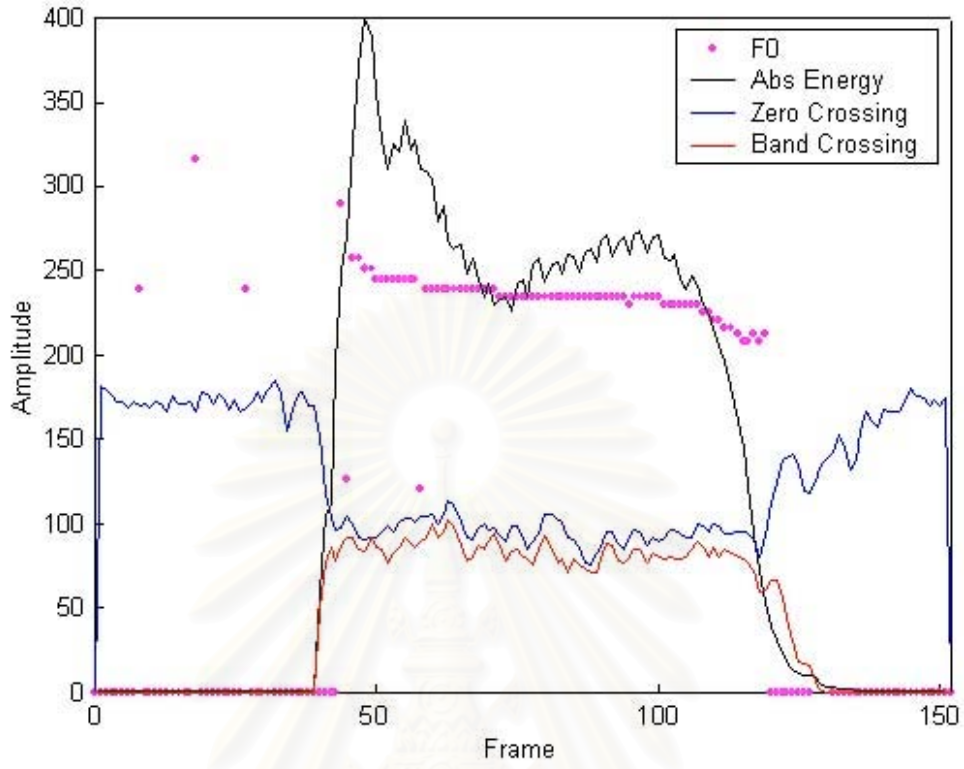
เสียงกัก-ไม่ก้อง-พ่นลม เกิดจากการระเบิดของลมออกมาเช่นเดียวกัน แต่มีการพ่นของลมออกมาด้วย ดังนั้นสัญญาณของเสียงกัก-ไม่ก้อง-พ่นลมจึงมีลักษณะของพัลส์ โดยมีการเพิ่มขึ้นและลดลงของแอมพลิจูดอย่างรวดเร็ว ประกอบกับลมที่พ่นออกมาจะมีลักษณะคล้ายกับสัญญาณรบกวน ดังนั้นในช่วงของค่าพลังงานเพิ่มสูงขึ้น อัตราการตัดผ่านศูนย์และอัตราการตัดผ่านระดับกำหนดจะเพิ่มสูงขึ้นด้วย และเมื่อค่าพลังงานลดต่ำลง อัตราการตัดผ่านศูนย์และอัตราการตัดผ่านระดับกำหนดก็ลดลงด้วย ดังรูปที่ 3.8

เสียงกัก-ก้อง เกิดจากการระเบิดของลมเช่นกัน แต่มีการสั่นของเส้นเสียงก่อนการระเบิดของลม ทำให้ช่วงที่เส้นเสียงสั่นแต่ยังไม่ระเบิดลมออกมา มีสัญญาณเสียงแล้ว ดังนั้นอัตราการตัดผ่านศูนย์จึงลดลง ขณะที่ค่าพลังงานสูงขึ้นไม่มากนัก และเมื่อมีการระเบิดของลม อัตราการตัดผ่านศูนย์ และอัตราการตัดผ่านระดับกำหนดจะเพิ่มสูงขึ้น โดยที่ค่าพลังงานก็เพิ่มขึ้นอย่างรวดเร็ว ดังรูปที่ 3.9

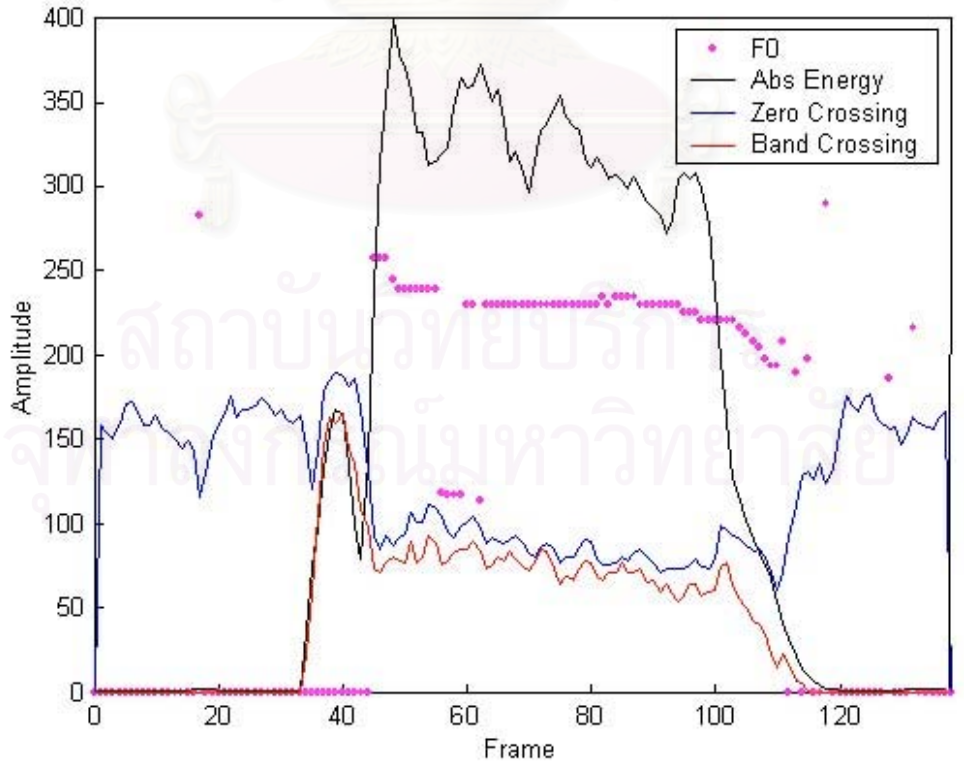
เสียงเสียดแทรก (ไม่กัก-ไม่ก้อง) ไม่ได้เกิดจากการระเบิดของลม (ไม่กัก) ไม่มีการสั่นของเส้นเสียง (ไม่ก้อง) แต่เกิดจาก Turbulent Flow ของลมทำให้เสียงที่ได้มีลักษณะเสียดแทรกคือมีความถี่สูงแต่มีค่าพลังงานต่ำ ซึ่งมีลักษณะคล้ายสัญญาณรบกวน ทำให้อัตราการตัดผ่านศูนย์และอัตราการตัดผ่านระดับกำหนดเพิ่มขึ้น ในขณะที่ค่าพลังงานสูงขึ้นน้อยมาก ดังรูปที่ 3.10

เสียงนาสิก (ไม่กัก-ก้อง) เกิดจากการสั่นของเส้นเสียงพร้อมกับลมที่ออกมาจากช่องจมูกและช่องปาก การรวมพื้นที่ของช่องจมูกและช่องปากทำให้ปริมาตรของลมเพิ่มมากขึ้นกว่าปกติ ดังนั้นเมื่อเปล่งเสียงนาสิก ลมที่ออกมาจึงมีค่าพลังงานสูง ส่วนช่วงที่เส้นเสียงสั่นจะมีอัตราการตัดผ่านศูนย์ลดลง และค่าพลังงานที่เพิ่มสูงขึ้น ดังรูปที่ 3.11

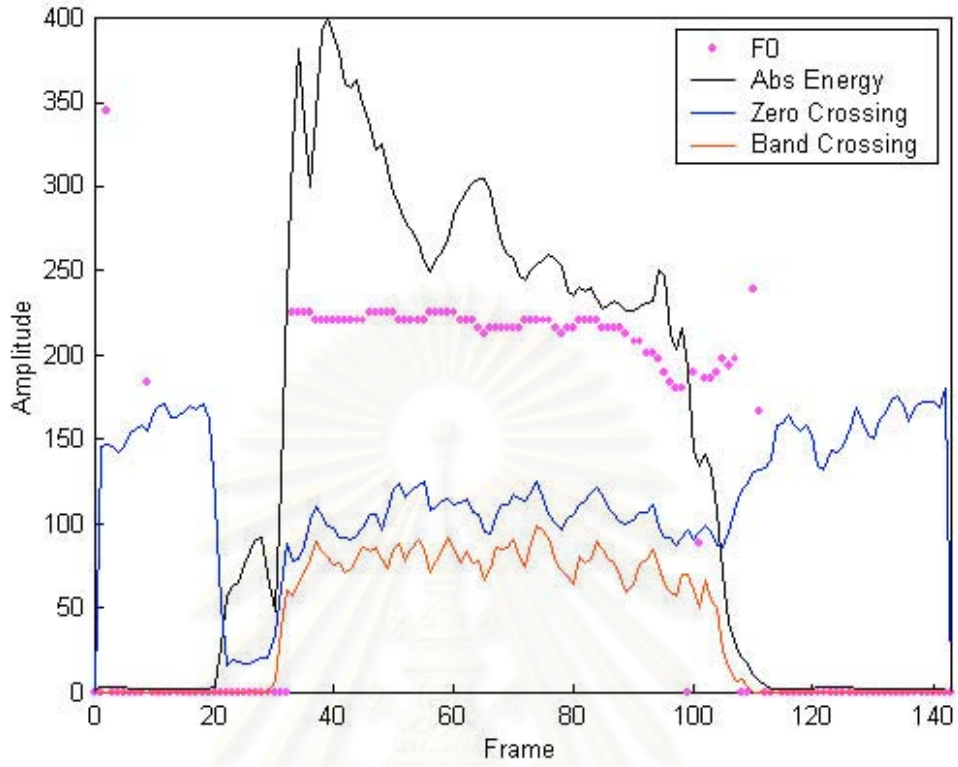
เสียงกึ่งสระ ประกอบด้วย เสียงลิ้นรัว เสียงข้างลิ้น เสียงต่อเนื่อง (ไม่กัก-ก้อง) เกิดจากการสั่นของเส้นเสียง และมีลมที่ออกมาพร้อมๆ กับการเปลี่ยนแปลงรูปร่างของช่องปากอย่างช้าๆ ทำให้ช่วงที่เส้นเสียงสั่นมีอัตราการตัดผ่านศูนย์ลดลง ส่วนค่าพลังงานค่อยๆ สูงขึ้นอย่างช้าๆ ตามการเปลี่ยนแปลงรูปร่างของช่องปาก ดังรูปที่ 3.12 และ 3.13 ส่วนเสียงลิ้นรัวจะมีการเพิ่มขึ้นและลดลงของค่าพลังงาน ตามลักษณะการรัวของลิ้น ดังรูปที่ 3.14



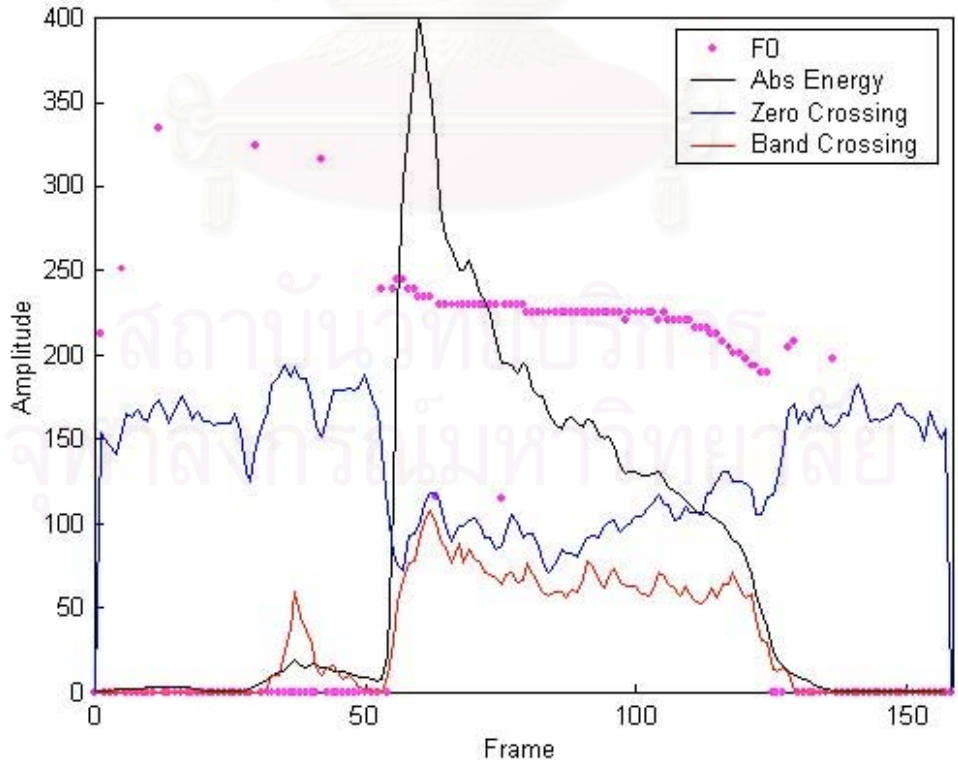
รูปที่ 3.7 ตัวอย่างเสียง กัก-ไม่ก้อง-ไม่พ่นลม /k@@0/ (เสียงกอ)



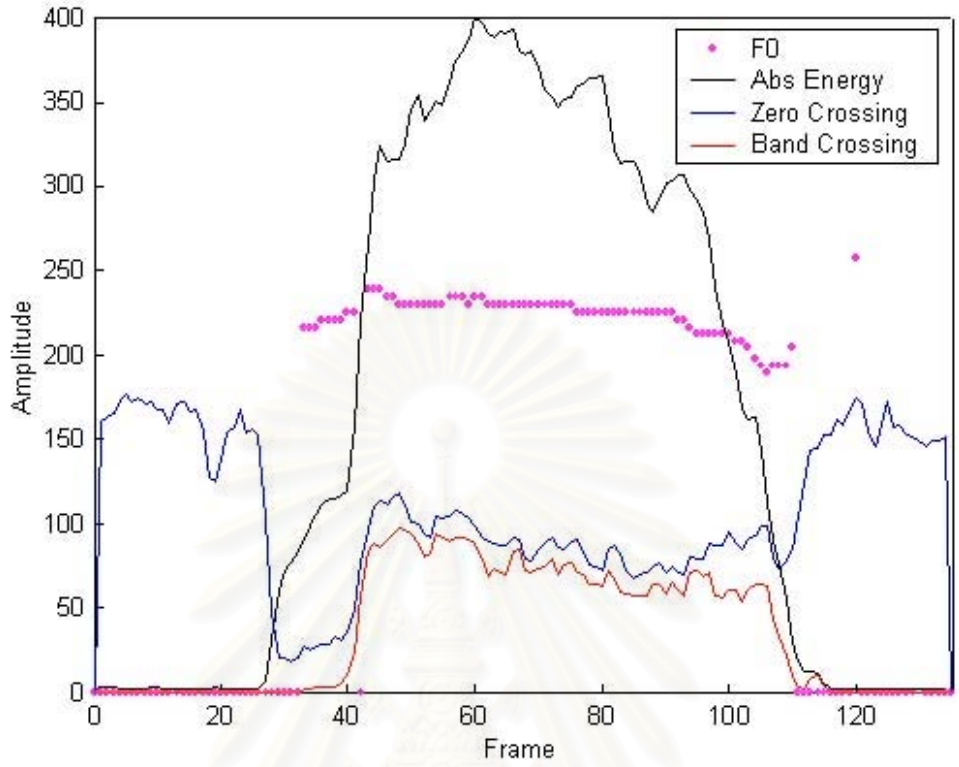
รูปที่ 3.8 ตัวอย่างเสียง กัก-ไม่ก้อง-พ่นลม /th@@0/ (เสียงทอ)



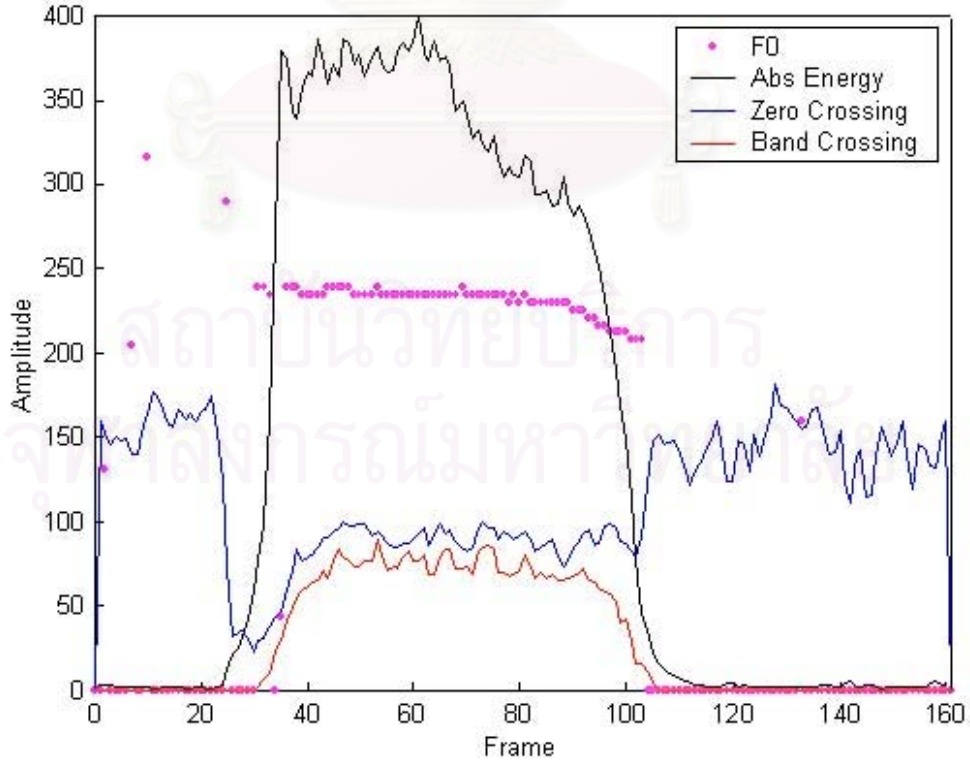
รูปที่ 3.9 ตัวอย่างเสียงกัก-ก้อง /d@@0/ (เสียงดอ)



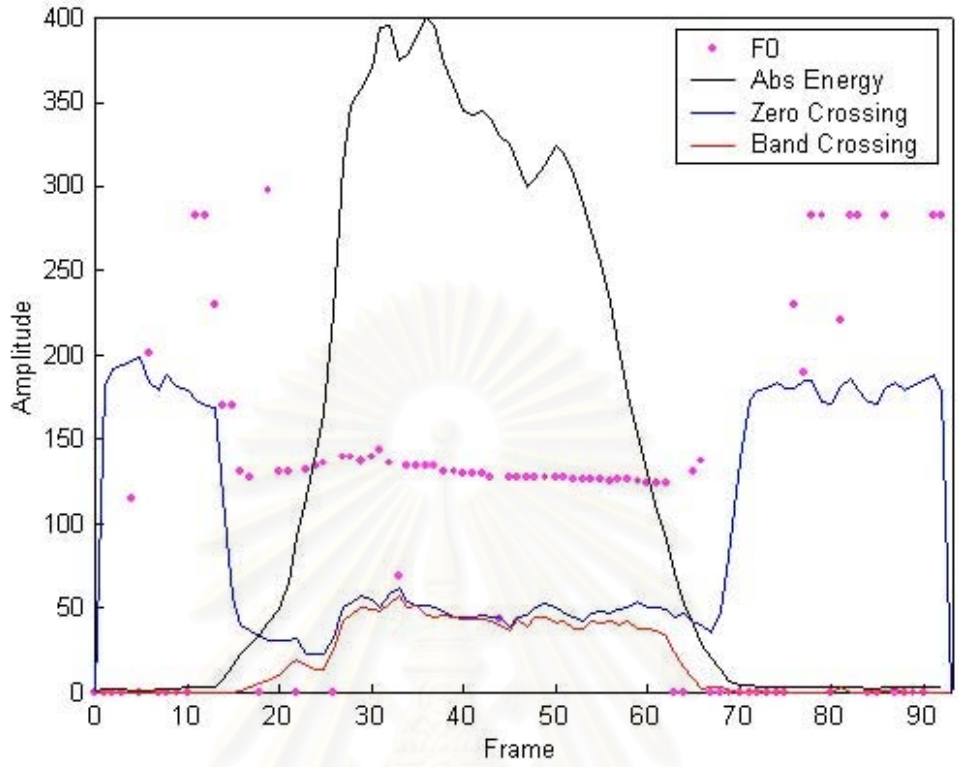
รูปที่ 3.10 ตัวอย่างเสียง ไม่กัก-ไม่ก้อง-เสียดแทรก /f@@4/ (เสียงฝอ)



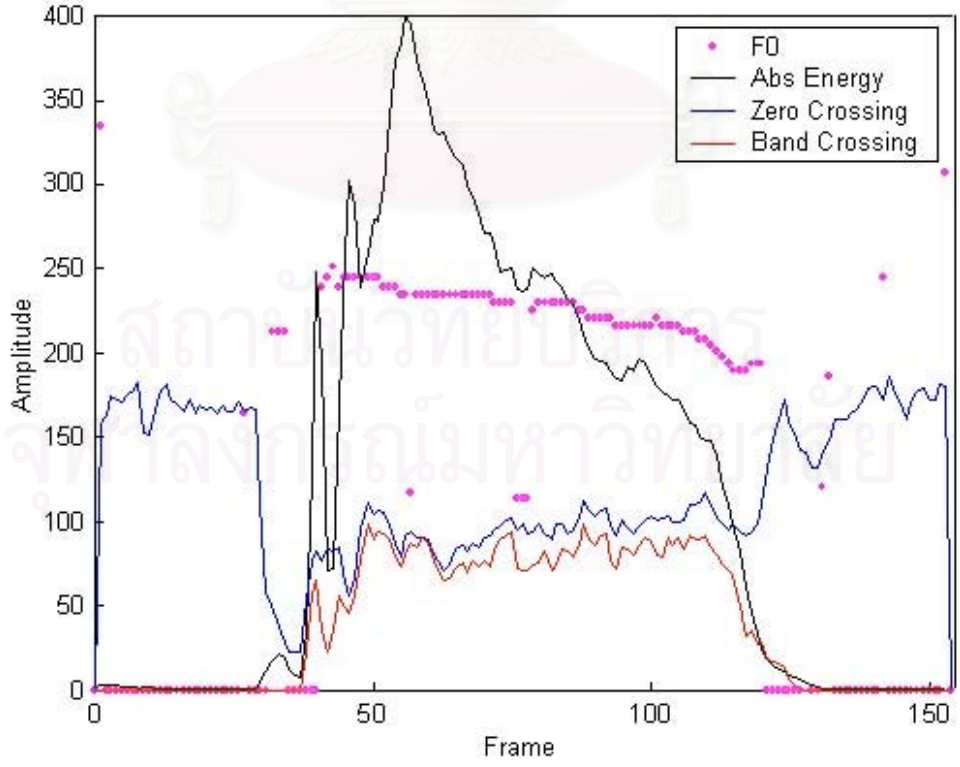
รูปที่ 3.11 ตัวอย่างเสียง ไม่กัก-ก้อง-นาสิก /n@@0/ (เสียงนอ)



รูปที่ 3.12 ตัวอย่างเสียงไม่กัก-ก้อง-เสียงต่อเนื่อง /w@@0/ (เสียงวอ)



รูปที่ 3.13 ตัวอย่างเสียงไม่กัก-ก้อง-เสียงข้างลิ้น /l@@0/ (เสียงลอ)



รูปที่ 3.14 ตัวอย่างเสียงไม่กัก-ก้อง-เสียงลิ้นรัว /r@@0/ (เสียงรอ)

จากคุณสมบัติข้างต้นของเสียงจะเห็นได้ว่าเมื่อเริ่มพิจารณาตั้งแต่ส่วนเริ่มต้นของเสียงพยัญชนะจะพบความก้องหรือไม่ก้องของเสียงก่อน และเมื่อพิจารณาต่อไปจะพบลักษณะ Noise-like อย่างชัดเจนในกลุ่มเสียงกัก-ไม่ก้อง-พ่นลมและเสียงเสียดแทรกเท่านั้น ซึ่งเสียง Noise-like ทั้งสองต่างก็เป็นเสียงไม่ก้อง ส่วนเสียงไม่ก้องที่เหลืออยู่ก็คือเสียงกัก-ไม่ก้อง-ไม่พ่นลม ส่วนเสียงก้องซึ่งพบในกลุ่มเสียงกัก เสียงนาสิกและเสียงกึ่งสระ เมื่อพิจารณาจากคุณสมบัติของเสียงกักที่มีการระเบิดของลมจะพบการเพิ่มขึ้นของแอมพลิจูดอย่างรวดเร็วต่างกับเสียงนาสิกที่มีแอมพลิจูดสูงระดับหนึ่ง ขณะที่เสียงกึ่งสระจะมีแอมพลิจูดค่อยๆ เพิ่มขึ้น ดังนั้นในงานวิจัยนี้จึงเลือกใช้คุณสมบัติของความก้อง ความคล้ายสัญญาณรบกวน และค่าพลังงาน มาใช้ในการแยกเสียงออกจากกัน โดยแบ่งออกเป็น 4 กรรมวิธีคือ

กรรมวิธี Voiced Detection ใช้แยกเสียงก้อง (Voiced) ออกจากเสียงไม่ก้อง (Voiceless) เสียงก้องประกอบด้วย เสียงกัก-ก้อง เสียงนาสิกและเสียงกึ่งสระ ส่วนเสียงไม่ก้องประกอบด้วย เสียงกัก-ไม่ก้อง-ไม่พ่นลม เสียงกัก-ไม่ก้อง-พ่นลมและเสียงเสียดแทรก โดยเสียงก้องจะมีการเปิด-ปิดของเส้นเสียง ดังรูปที่3.9 ที่เวลา 20 มิลลิวินาที เริ่มมีค่าพลังงานขณะที่อัตราการตัดผ่านศูนย์ลดต่ำลง และอัตราการผ่านระดับกำหนดยังไม่มีค่า ตรงข้ามกับเสียงไม่ก้องที่มีอัตราการตัดผ่านศูนย์ลดต่ำลง พร้อมกับการเพิ่มสูงขึ้นของอัตราการตัดผ่านระดับกำหนด (รูปที่ 3.7)

กรรมวิธี Noise Detection แยกเสียงไม่ก้องออกเป็น 2 กลุ่มตามลักษณะของความคล้ายและไม่คล้ายสัญญาณรบกวน (Noise และ Non-noise) โดยเสียงที่มีลักษณะคล้ายสัญญาณรบกวนจะมีการตัดแกนศูนย์มากกว่าดังรูปที่ 3.8 และ 3.10 ทำให้ค่าอัตราการตัดผ่านศูนย์ และค่าอัตราการตัดผ่านระดับกำหนดมีค่าสูง ดังนั้นจึงเลือกใช้ค่าอัตราการตัดผ่านศูนย์ร่วมกับอัตราการตัดผ่านระดับกำหนดมาพิจารณา

กรรมวิธี Amplitude Detection ใช้แยกเสียงกัก-ก้อง เสียงนาสิก และเสียงกึ่งสระออกจากกัน จากคุณสมบัติของเสียงกัก ซึ่งเกิดจากการระเบิดของลมที่กักไว้ ทำให้ลักษณะของค่าพลังงานเพิ่มสูงขึ้นอย่างรวดเร็ว ดังรูปที่ 3.9 ในขณะที่เสียงนาสิกมีลมออกมาจากช่องจมูกและช่องปากทำให้แอมพลิจูดค่อนข้างสูงดังรูปที่ 3.11 ส่วนเสียงกึ่งสระที่เกิดจากการเคลื่อนตัวอย่างช้าๆ ของอวัยวะภายในช่องปากทำให้แอมพลิจูดมีลักษณะค่อยๆ เพิ่มขึ้นและค่าพลังงานจะต่ำกว่าเมื่อเทียบกับเสียงกักและเสียงนาสิก

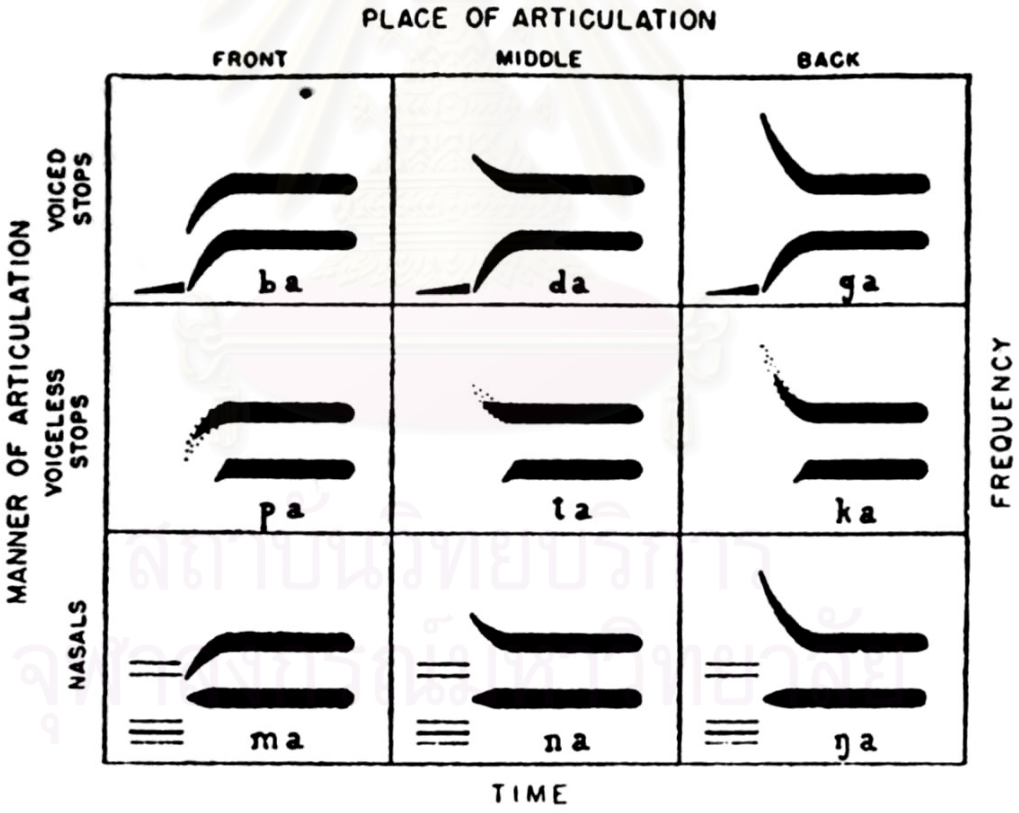
กรรมวิธี Amplitude Detection แยกกลุ่มเสียง Noise ออกเป็นเสียงกัก-ไม่ก้อง-พ่นลม และ เสียงเสียดแทรก เนื่องจากการระเบิดของลมในเสียงกัก-ไม่ก้อง-พ่นลมทำให้แอมพลิจูดสูงมากกว่าเสียงเสียดแทรกที่มีลมแบบ Turbulent Flow ดังรูปที่ 3.8 และ 3.10 จะพบว่าช่วงที่อัตรา

การตัดผ่านระดับกำหนดมีค่าสูง (Noise Period) ค่าพลังงานของเสียงกัก-ไม่ก้อง-พ่นลมจะมีลักษณะเป็นพัลส์สูง ตรงข้ามกับเสียงเสียดแทรกที่มีค่าพลังงานต่ำ

เมื่อได้กรรมวิธีการแยกเสียงออกตามลักษณะการเกิดเสียงแล้ว จึงพิจารณาการรู้จำเสียงตามฐานของการเกิดเสียงต่อไป เนื่องจากลักษณะของพลังงาน อัตราการตัดผ่านศูนย์ และอัตราการตัดผ่านระดับกำหนดไม่เพียงพอในการแยกความแตกต่างตามฐานของเสียง

ขั้นตอนวิธีการรู้จำเสียงตามฐานเสียง

เสียงกัก-ก้อง ประกอบด้วยเสียงบอ /b@@/ และเสียงดอ /d@@/ โดยเสียงบอ /b@@/ เกิดที่ฐานปาก ส่วนเสียงดอ /d@@/ เกิดที่ฐานปุ่มเหงือก (รูปที่ 3.16 (ก) และ รูปที่ 3.16 (ข)) ทำให้มีแบบรูปการเรียงตัวของความถี่ฟอร์แมนท์ระหว่างช่วงต่อของเสียงสระกับเสียงพยัญชนะ ยกตัวอย่างเสียงบา /b@/ ที่มีลักษณะของความถี่ฟอร์แมนท์ที่สองลดลง และเสียงดา /d@/ ที่มีลักษณะเพิ่มขึ้นของฟอร์แมนท์ที่สอง ดังรูปที่ 3.15

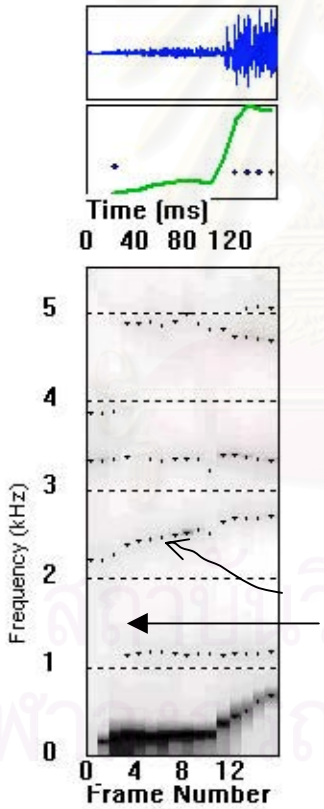


รูปที่ 3.15 แบบรูปการเรียงตัวของความถี่ฟอร์แมนท์ในเสียงกักและเสียงนาสิก (Liberman, 1995)

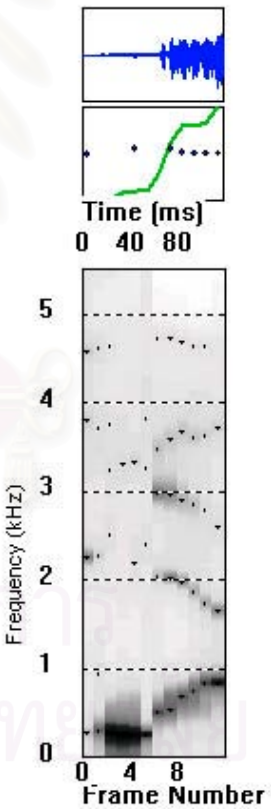
เสียงนาสิกประกอบด้วยเสียงมอ (/m@@0/) เสียงนอ (/n@@0/) และเสียงงอ (/ng@@0/) โดยเสียงมอ /m@@0/ เกิดที่ฐานปาก ส่วนเสียงนอ /n@@0/ เกิดที่ฐานปุ่มเหงือก และเสียงงอ

(/ng@@0/) เกิดที่ฐานเพดานอ่อน พบความแตกต่างของเสียงได้จากความถี่ฟอร์แมนท์ระหว่างเสียงสระกับเสียงนาสิก (Vowel-to-Nasal) ดังรูปที่ 3.15 โดยเสียงมอมีลักษณะของความถี่ฟอร์แมนท์ตกลง ขณะที่เสียงนอจะค่อนข้างเพิ่มขึ้น และเสียงงอจะมีฟอร์แมนท์เพิ่มขึ้นอย่างชัดเจน ยกตัวอย่างดังรูปที่ 3.17

เสียงกึ่งสระประกอบด้วยเสียงลิ้นรัวหรือเสียงรอ (/r@@0/) เสียงข้างลิ้นหรือเสียงลอ (/l@@0/) และเสียงต่อเนื่องประกอบด้วยเสียงวอ (/w@@0/) และเสียงยอ (/j@@0/) โดยเสียงรอ (/r@@0/) จะเกิดจากปุ่มเหงือกเช่นเดียวกับเสียงลอ (/l@@0/) แต่การรัวของลิ้นทำให้เสียงรอมีการเปลี่ยนแปลงของแอมพลิจูดอย่างชัดเจนดังรูปที่ 3.18 (ข) ส่วนเสียงวอและเสียงยอต่างกันที่เสียงวอเกิดที่ฐานปาก ส่วนเสียงยอเกิดที่ฐานเพดานแข็ง ซึ่งคุณลักษณะของเสียงที่เกิดที่ฐานเพดานแข็งจะให้พลังงานที่ความถี่สูงอย่างชัดเจนดังรูปที่ 3.18 (ง)



รูปที่ 3.16 (ก) ตัวอย่างเสียงบอ /b@@0/



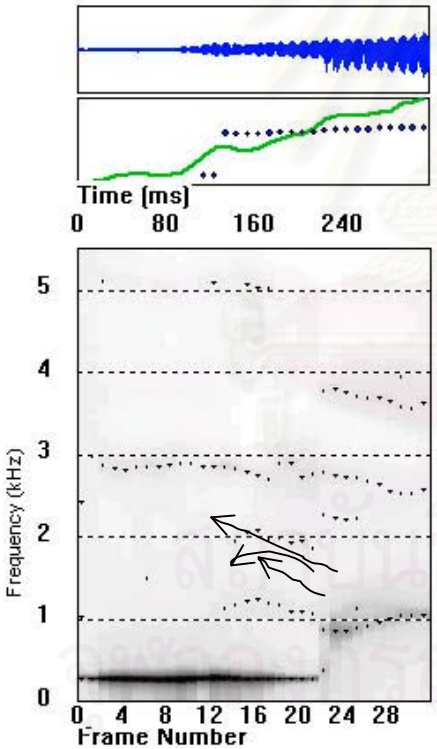
รูปที่ 3.16 (ข) ตัวอย่างเสียงดอ /d@@0/

เสียงกัก-ไม่ก้อง-ไม่พ่นลม ประกอบด้วยเสียงปอ (/p@@0/) เสียงตอ (/t@@0/) เสียงจอ (c@@0/) เสียงกอ (/k@@0/) เสียงออ (/@@0/) โดยเสียงปอเกิดที่ฐานปาก เสียงตอเกิดที่ฐานปุ่มเหงือก เสียงจอเกิดที่ฐานเพดานแข็ง และเสียงกอเกิดที่เพดานอ่อน เสียงปา /p@/ จะมี

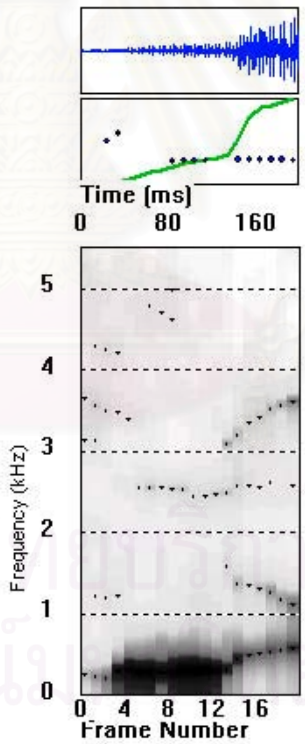
ฟอร์แมนท์ต่ำลง เสียงตา /t@/ จะมีฟอร์แมนท์ค่อนข้างสูงขึ้นและเสียงกา /k@/ จะมีฟอร์แมนท์สูงที่สุด เสียงจอที่เกิดที่เพดานแข็งก็จะมีลักษณะของความถี่สูง ส่วนเสียงออก็มีลักษณะของเสียงสระที่มีความคงตัวของความถี่ฟอร์แมนท์ ยกตัวอย่างดังรูปที่ 3.19

เสียงกัก-ไม่ก้อง-พ่นลม ประกอบด้วยเสียงพอ (/ph@@0/) เสียงทอ (/th@@/) เสียงชอ (/ch@@0/) และเสียงคอ (/kh@@0/) โดยมีลักษณะการเกิดเช่นเดียวกับเสียงกัก-ไม่พ่นลม ยกตัวอย่างดังรูปที่ 3.20

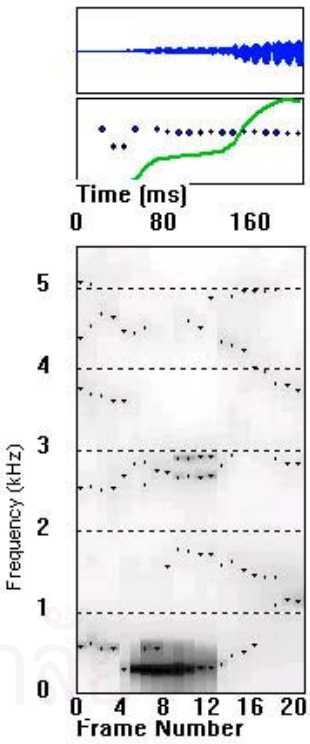
เสียงเสียดแทรกประกอบด้วยเสียงฟอ (/f@@0/) เสียงซอ (/s@@0/) เสียงฮอ (/h@@0/) เสียงเสียดแทรกจะมีข้อมูลเสียงส่วนใหญ่อยู่ที่ความถี่สูง โดยเสียงฟอเกิดที่ฐานปากจะมีพลังงานค่อนข้างต่ำเมื่อเทียบกับเสียงซอและเสียงฮอ โดยเสียงซอเกิดที่ฐานปุ่มเหงือกมีพลังงานส่วนใหญ่ที่ความถี่สูง และเสียงฮอเกิดที่ช่องเส้นเสียงมีพลังงานส่วนใหญ่ที่ความถี่ที่ต่ำกว่าเสียงซอ ยกตัวอย่างดังรูปที่ 3.21



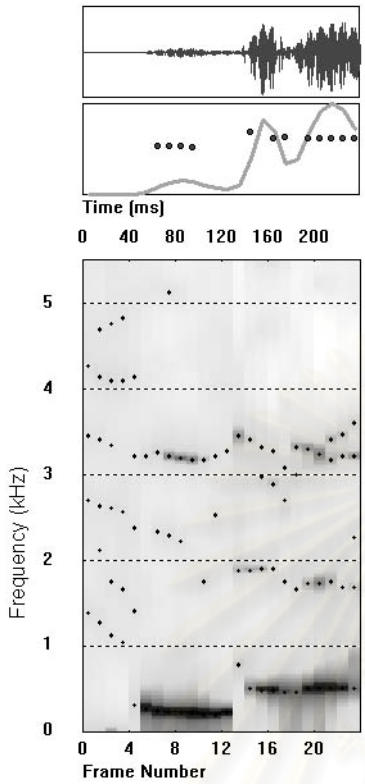
รูปที่ 3.17 (ก) ตัวอย่างเสียงมอ /m@@0/



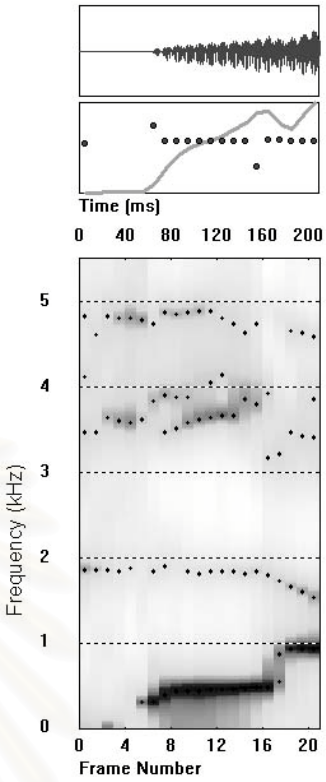
รูปที่ 3.17 (ข) ตัวอย่างเสียงนอ /n@@0/



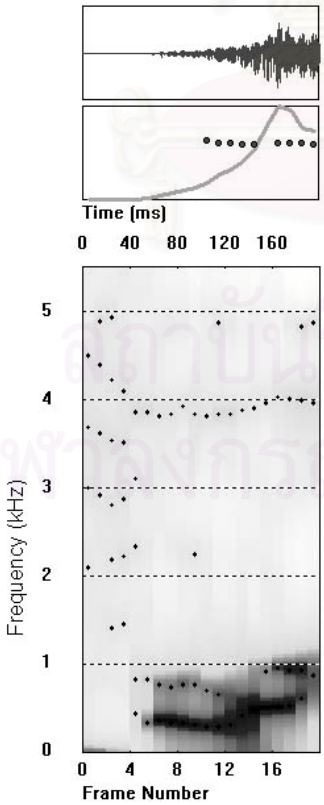
รูปที่ 3.17 (ค) ตัวอย่างเสียงงอ /ng@@0/



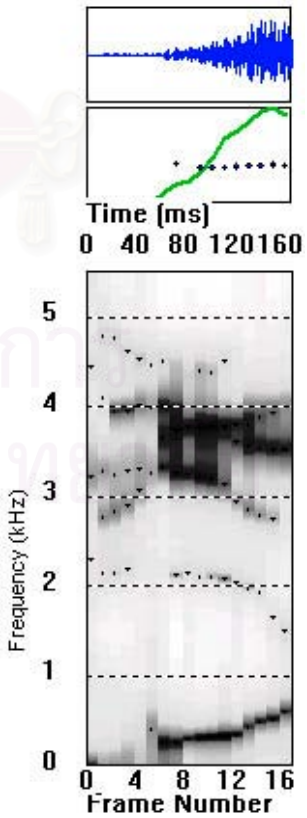
รูปที่ 3.18 (ก) ตัวอย่างเสียงรอ /r@@0/



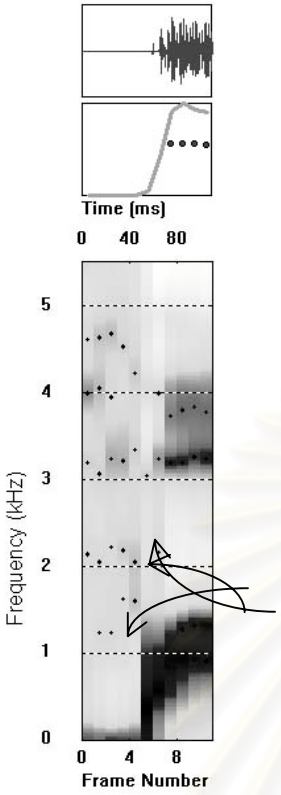
รูปที่ 3.18 (ข) ตัวอย่างเสียงลอ /l@@0/



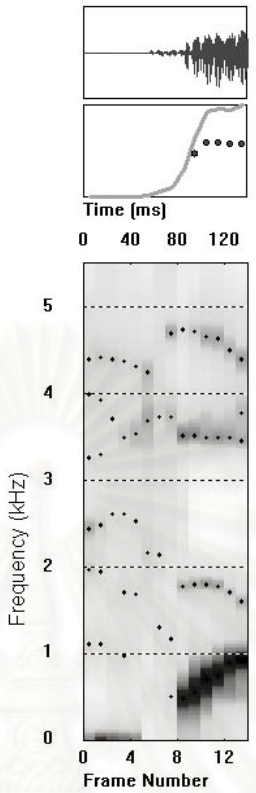
รูปที่ 3.18 (ค) ตัวอย่างเสียงวอ /w@@0/



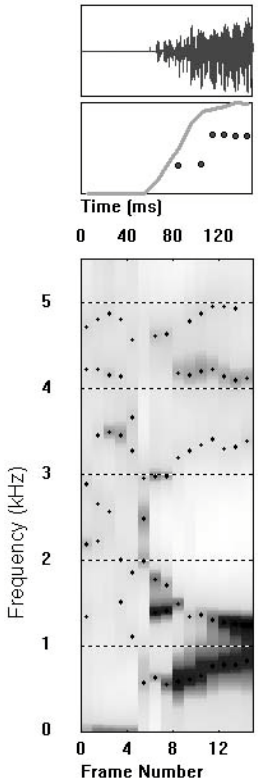
รูปที่ 3.18 (ง) ตัวอย่างเสียงยอ /j@@0/



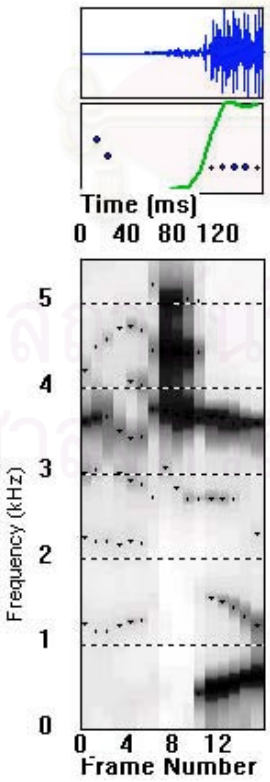
รูปที่ 3.19 (ก) ตัวอย่างเสียงปอ /p@@0/



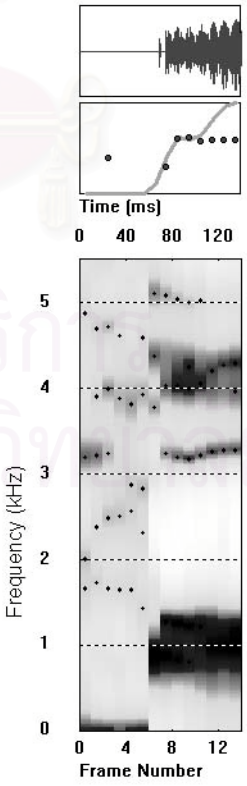
รูปที่ 3.19 (ข) ตัวอย่างเสียงตอ /t@@0/



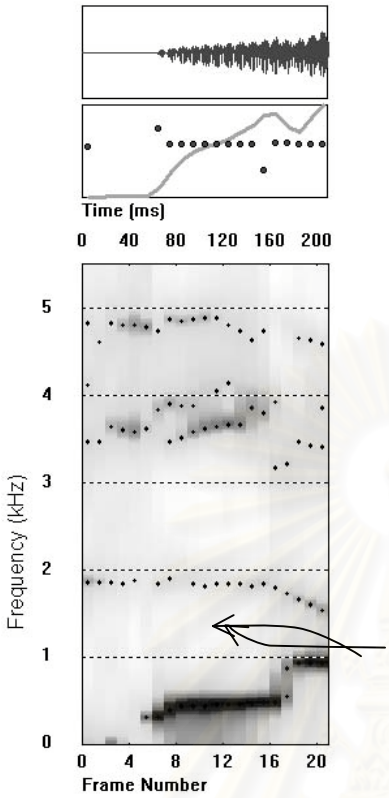
รูปที่ 3.19 (ค) ตัวอย่างเสียงกอ /k@@0/



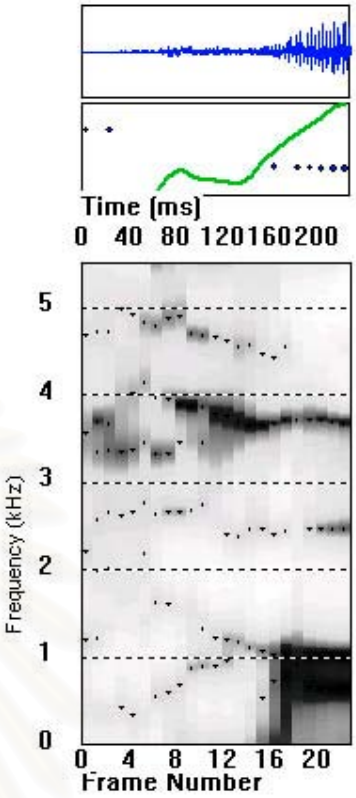
รูปที่ 3.19 (ง) ตัวอย่างเสียงจอ /c@@0/



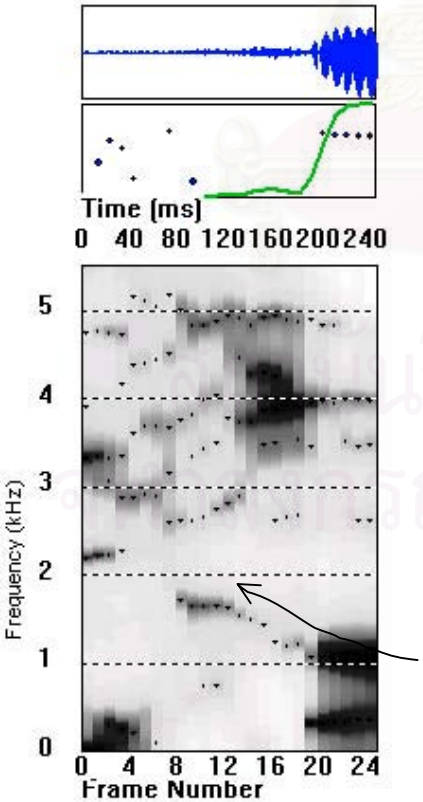
รูปที่ 3.19 (จ) ตัวอย่างเสียงออ /@@0/



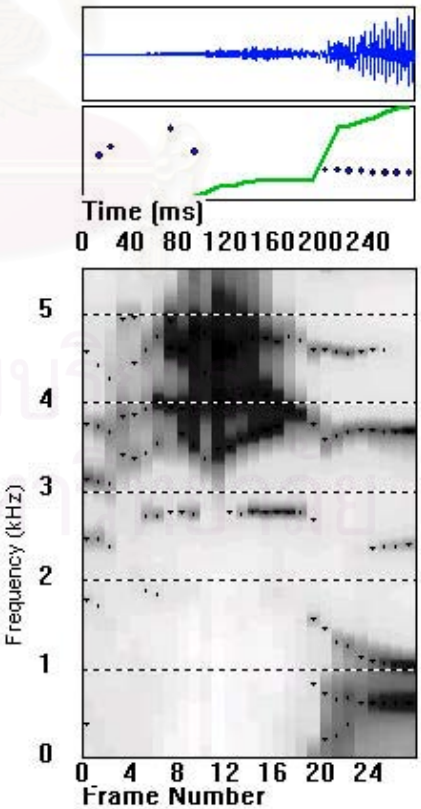
รูปที่ 3.20 (ก) ตัวอย่างเสียงพอ /ph@@0/



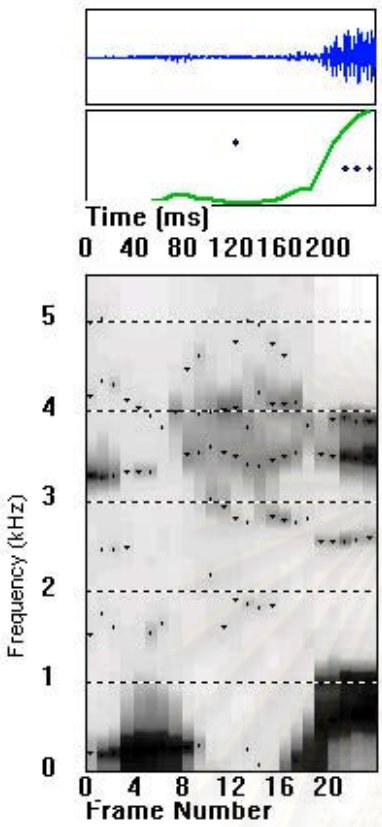
รูปที่ 3.20 (ข) ตัวอย่างเสียงทอ /th@@0/



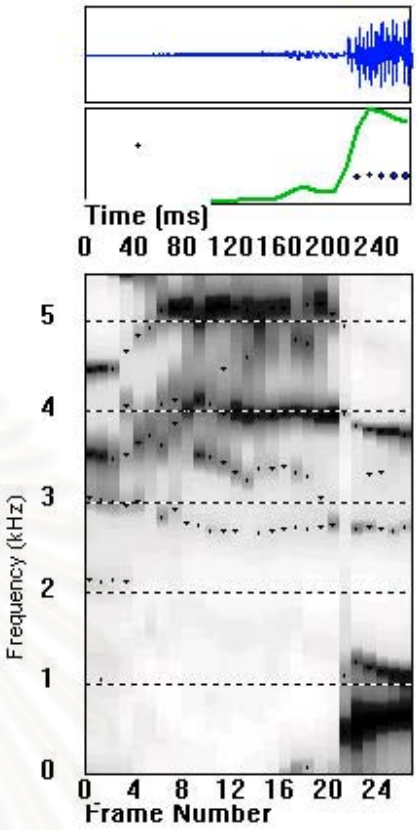
รูปที่ 3.20 (ค) ตัวอย่างเสียงคอ /kh@@0/



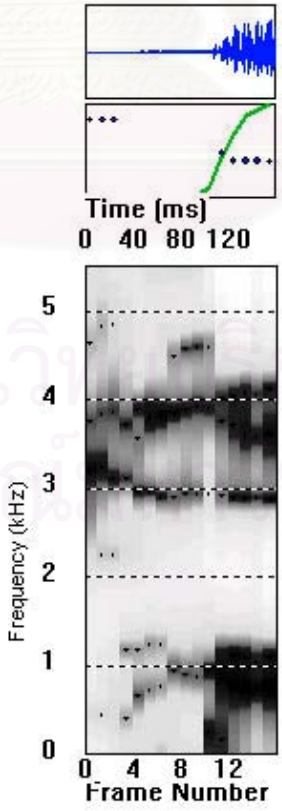
รูปที่ 3.20 (ง) ตัวอย่างเสียงชอ /ch@@0/



รูปที่ 3.21 (ก) ตัวอย่างเสียงฟอ /f@@0/



รูปที่ 3.21 (ข) ตัวอย่างเสียงซอ /s@@0/



รูปที่ 3.21 (ค) ตัวอย่างเสียงฮอ /h@@0/

จากฐานความรู้ดังกล่าวสามารถสรุปเป็นขั้นตอนวิธีการฐานความรู้ 10 ขั้นตอนดังนี้คือ

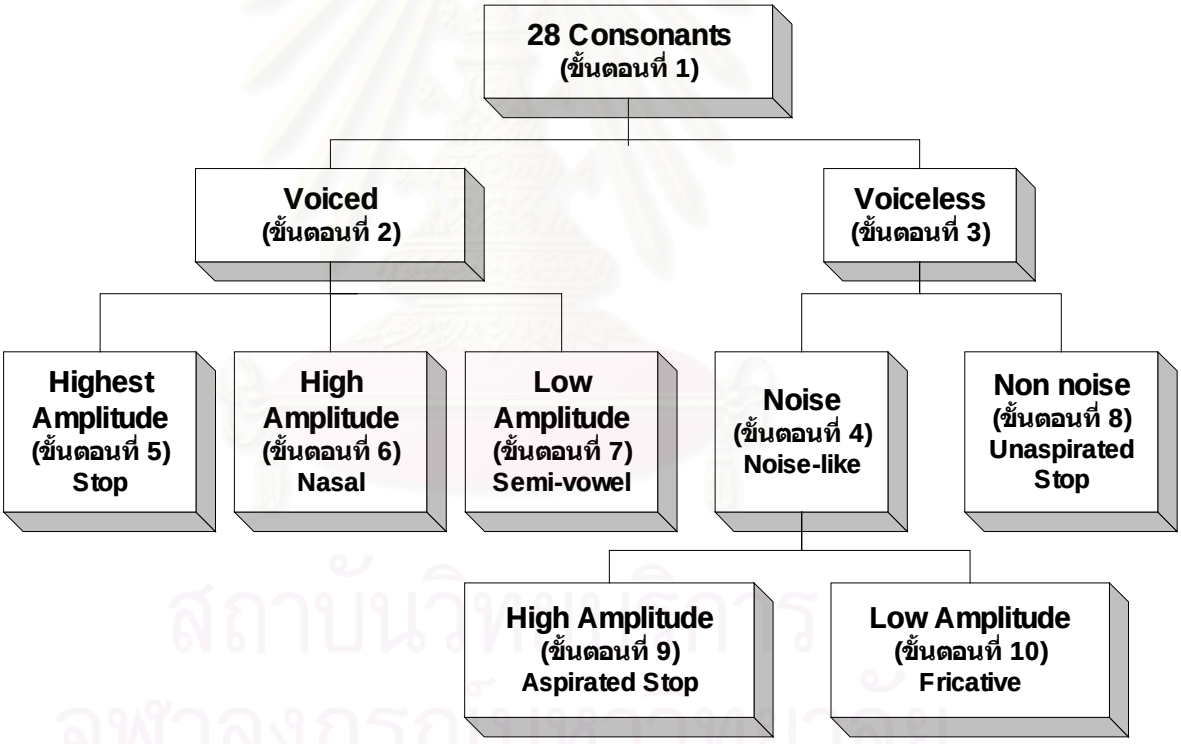
ขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง (Preclassified Algorithm) ประกอบด้วยขั้นตอนที่ 1-4 ดังรูปที่ 3.22

ขั้นตอนที่ 1 การแยกเสียงก้อง (Voiced) ออกจากเสียงไม่ก้อง (Voiceless)

ขั้นตอนที่ 2 การแยกเสียงก้อง แบ่งเสียงออกเป็น 3 กลุ่มคือ กลุ่มเสียงกัก-ก้อง กลุ่มเสียงนาสิก และกลุ่มของเสียงกึ่งสระ (เสียงลิ้นรัว เสียงข้างลิ้น และเสียงต่อเนื่อง)

ขั้นตอนที่ 3 การแยกเสียงไม่ก้อง แบ่งเสียงออกเป็น 2 กลุ่มตามลักษณะของความคล้ายและไม่คล้ายสัญญาณรบกวน (Noise และ Non-noise)

ขั้นตอนที่ 4 การแยกกลุ่มเสียง Noise แบ่งเสียงออกเป็นเสียงกัก-ไม่ก้อง-พ่นลม และเสียงเสียดแทรก



รูปที่ 3.22 แผนภาพแสดงขั้นตอนการรู้จำเสียงพยัญชนะ 21 เสียง

ขั้นตอนวิธีการรู้จำเสียงตามฐานเสียง

ขั้นตอนที่ 5 การแยกกลุ่มเสียงกัก-ก้อง (Stop Voiced Classifier)

ขั้นตอนที่ 6 การแยกเสียงกลุ่มเสียงนาสิก (Nasal Classifier)

ขั้นตอนที่ 7 การแยกเสียงกึ่งสระ (Trill, Lateral and Approximant Classifier)

ขั้นตอนที่ 8 การแยกกลุ่มเสียงไม่ก้องและ Non-noise หรือเสียงกัก-ไม่ก้อง-ไม่พ่นลม (Stop Voiceless Unaspirated)

ขั้นตอนที่ 9 การแยกกลุ่มเสียงกัก-ไม่ก้อง-พ่นลม (Stop Voiceless Aspirated)

ขั้นตอนที่ 10 การแยกกลุ่มเสียงเสียดแทรก (Fricative Classifier)

3.2.6 **ส่วนของการจำแนกแบบรูป** (Pattern Classification)

ใช้แบบจำลองฮิดเดน มาร์คอฟ ชนิดต่อเนื่อง (Continuous Hidden Markov Model) โดยทำการปรับจำนวน Gaussian Mixture ที่เหมาะสมกับระบบดังรายละเอียดในบทที่ 2

3.2.7 **ขั้นตอนการแยกเสียงวรรณยุกต์** (Tone Classification)

ในงานวิจัยนี้นำเสนอการแยกเสียงวรรณยุกต์ด้วยความถี่มูลฐาน โดยพิจารณาการเพิ่มขึ้นของความถี่มูลฐานในเสียงจัตวาและการลดลงของความถี่มูลฐานในเสียงสามัญที่บริเวณท้ายพยางค์ พบปัญหาการขาดหายของความถี่มูลฐานดังรูปที่ 3.23 ซึ่งเกิดจากการที่ค่าพลังงานลดต่ำลง ทำให้ไม่สามารถตรวจหาจุดยอดของสัญญาณเสียงได้ในขั้นตอนของ Peak Extract (รูปที่ 2.10) และยังพบปัญหาจากการเน้นเสียงของผู้พูดที่ท้ายพยางค์ทำให้ความถี่มูลฐานบริเวณท้ายพยางค์มีลักษณะเพิ่มขึ้น ยกตัวอย่างดังรูปที่ 3.24 ซึ่งเป็นเสียงสามัญที่ ดังนั้นในงานวิจัยนี้จึงนำเสนอการแยกเสียงจัตวาออกจากเสียงสามัญด้วยอัลกอริทึมแยกเสียงวรรณยุกต์ 4 ขั้นตอนดังนี้

ขั้นตอนที่1 หาจุดเริ่มต้น และจุดสิ้นสุดของความถี่มูลฐาน

1.1) วิธีการหาจุดเริ่มต้น ให้จุดเริ่มต้นสระเป็นจุดเริ่มต้นของความถี่มูลฐานหรือเรียกว่า *เฟรมเริ่ม*

1.2) วิธีการหาจุดสิ้นสุดของความถี่มูลฐาน กำหนดให้จุดสิ้นสุดของความถี่มูลฐานให้เป็น *เฟรมท้าย* เริ่มหาจากจุดสิ้นสุดพยางค์ย้อนมา จนถึงครึ่งหนึ่งของจำนวนเฟรมทั้งหมด โดยเลือกเฟรมที่อยู่ทางซ้ายของเฟรมที่ค่าความถี่มูลฐานเท่ากับศูนย์ แต่ความยาวของ *ช่วงของเฟรมที่เลือก* คือตั้งแต่ *เฟรมเริ่ม* ไปจนถึง *เฟรมท้าย* ให้มีค่ามากกว่า 2 ใน 3 ของขอบเขตของสระที่คำนวณได้จากขั้นตอนการกำหนดขอบเขตหน่วยเสียง

เมื่อได้จุดเริ่มต้นและจุดสิ้นสุดของความถี่มูลฐานแล้วจึงแบ่งความถี่มูลฐานออกเป็น 3 ส่วนคือ *ส่วนต้นเฟรม ส่วนกลางเฟรม* และ *ส่วนท้ายเฟรม*

การแบ่งความถี่มูลฐานออกเป็น 3 ส่วนนี้ใช้เพื่อเปรียบเทียบระดับของความถี่มูลฐานว่า *ส่วนท้ายเฟรม*มีลักษณะเพิ่มขึ้น หรือลดลงหรือไม่ เมื่อเทียบกับ *ส่วนกลางเฟรม* และ *ส่วนต้นเฟรม* โดยนำไปใช้แก้ปัญหาของการเน้นเสียงที่ท้ายพยางค์สำหรับเสียงสามัญที่ *ส่วนท้ายเฟรม*

มีลักษณะเพิ่มขึ้นเช่นรูปที่ 3.24 แต่เมื่อเทียบกับ *ส่วนต้นเฟรม* จะต้องพบว่าค่าความถี่มูลฐานของ *ส่วนท้ายเฟรม* ไม่สูงไปกว่า *ส่วนต้นเฟรม*

ขั้นตอนที่2 การปรับเรียบความถี่ของข้อมูล

ข้อมูลที่ได้จากขั้นตอนที่1 มีความไม่ต่อเนื่องทางด้านความถี่ จึงแบ่งกลุ่มข้อมูลตามความถี่โดยทำการฉายข้อมูลลงบนแกนความถี่แล้วทำการนับจำนวนความหนาแน่นที่แต่ละช่วงความถี่ [0,1), [1,2),[2,3)…[399,400) จากนั้นจึงตรวจหาช่วงความถี่ที่ต้องการ โดยในช่วงดังกล่าวจะต้องเป็นช่วงที่มีความหนาแน่นมากที่สุด และยอมให้มีความหนาแน่นที่เท่ากับศูนย์ในช่วงนั้นติดกันได้ไม่เกิน 25 เฮิรตซ์ เมื่อได้ช่วงความถี่ที่ต้องการแล้วจึงกำจัดความถี่ที่ไม่อยู่ในช่วงดังกล่าวทิ้งไปพร้อมทั้งเลื่อนเฟรมถัดไปเข้ามาและลดขนาด *เฟรมท้าย* เพื่อแก้ปัญหาการขาดหายของความถี่ ดังรูปที่ 3.23 จากนั้นจึงคำนวณหาค่าที่ใช้อ้างอิงดังสมการ

$$\textit{ช่วงของเฟรมที่เลือก} = \textit{เฟรมท้าย} - \textit{เฟรมเริ่ม} \quad (3.1)$$

$$\textit{เฟรมอ้างอิง} = 2/3 \text{ ของ } \textit{ช่วงของเฟรมที่เลือก} \quad (3.2)$$

$$\textit{ค่าอ้างอิง} = \frac{\text{ผลรวมของค่าความถี่มูลฐาน 10 เฟรมที่อยู่รอบๆ } \textit{เฟรมอ้างอิง}}{10} \quad (3.3)$$

ค่าอ้างอิง ใช้จัดระดับค่าความถี่ออกเป็น *ระดับสูง ระดับกลาง* และ *ระดับต่ำ*(ต่ำกว่า *ค่าอ้างอิง*) ซึ่งนำไปช่วยในการตัดสินใจว่าเป็นเสียงจัตวาหรือไม่

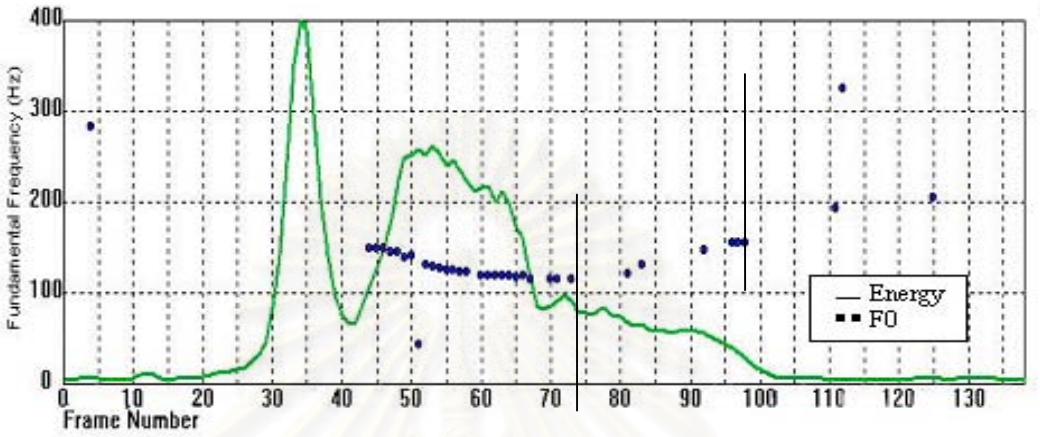
ขั้นตอนที่3 ตรวจสอบแนวโน้มการเปลี่ยนแปลงระดับของความถี่มูลฐานโดยแบ่งออกได้ 2 กรณีคือ

กรณี 1) ความถี่มูลฐานเพิ่มขึ้นอย่างชัดเจน พิจารณาค่าของเฟรมที่อยู่ติดกันใน *ช่วงของเฟรมที่เลือก* ถ้าเฟรมที่อยู่ถัดไปทางขวามีค่ามากกว่าให้เรียกเฟรมนั้นว่า *เฟรมเพิ่มขึ้น* ถ้ามีค่าเท่ากับเฟรมที่อยู่ติดกันเรียกเฟรมทางขวานั้นว่า *เฟรมเท่ากัน* แต่ถ้ามีค่าน้อยกว่าให้เรียกว่า *เฟรมลดลง* จากนั้นจึงตรวจสอบเงื่อนไข 3 ข้อดังนี้

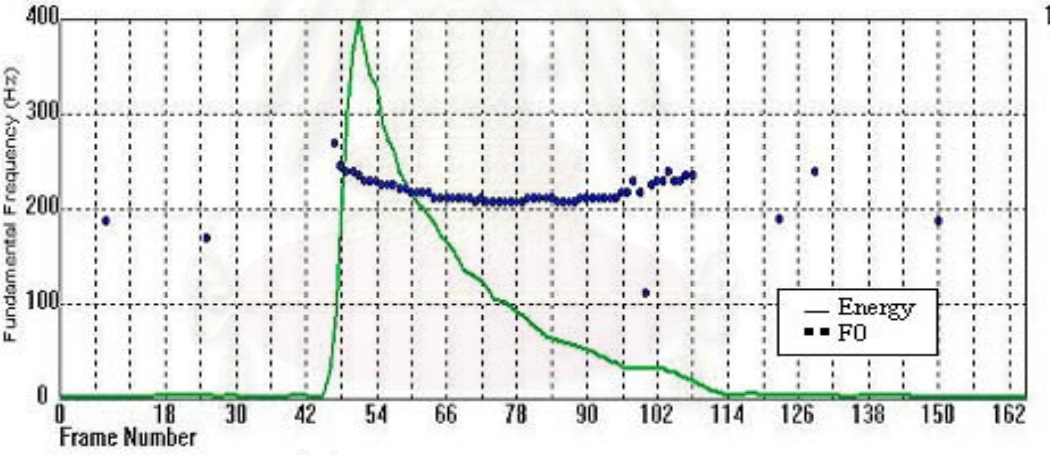
- จำนวนของ *เฟรมเพิ่มขึ้น* มีค่ามากกว่าร้อยละ 20 ของ *ช่วงของเฟรมที่เลือก* ซึ่งคำนวณจากการทดสอบของกลุ่มฝึกฝน โดยนำข้อมูลเสียงสามัญทั้งหมดมาคำนวณหาสัดส่วน *เฟรมเพิ่มขึ้น* กับ *ช่วงของเฟรมที่เลือก* แล้วเลือกค่าที่มากที่สุด
- ค่าความถี่มูลฐานของทุกๆ *เฟรมเพิ่มขึ้น* มีค่าสูงกว่า *ค่าอ้างอิง*
- ค่าความถี่มูลฐานที่มากที่สุดลบด้วย *ค่าอ้างอิง* มากกว่า 30 เฮิรตซ์ โดยคำนวณจากการทดสอบของกลุ่มฝึกฝนในชุดของกลุ่มเสียงสามัญ

ถ้าเป็นไปตาม 3 เงื่อนไขข้างต้นให้ตัดสินใจว่าเป็นเสียงจัตวา ยกตัวอย่างดังรูปที่ 3.25

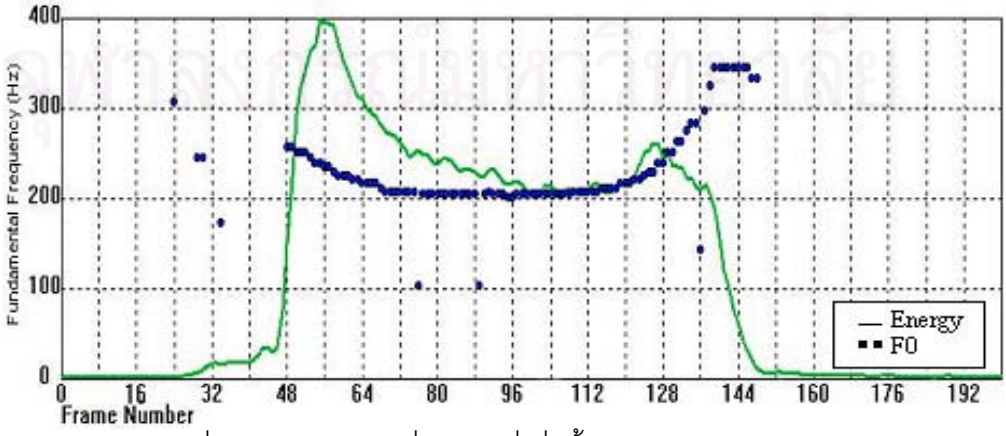
กรณี 2) ความถี่มูลฐานค่อนข้างเพิ่มขึ้น พิจารณาจากจำนวนของ *เฟรมเพิ่มขึ้น* มากกว่าร้อยละ 40 ของ *เฟรมอ้างอิง* และค่าที่มากที่สุดของ *เฟรมเพิ่มขึ้น* มากกว่า 15 เฮิรตซ์ (ได้จากการทดสอบของกลุ่มฝึกฝน) ดังรูปที่ 3.26



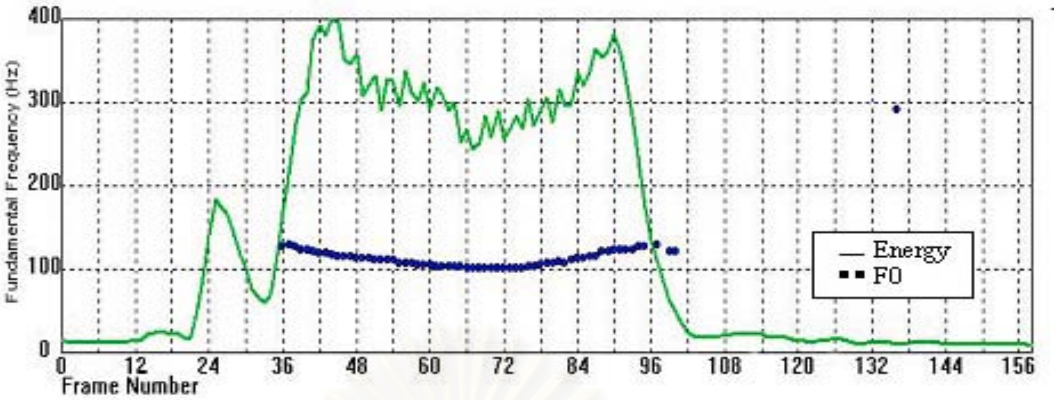
รูปที่ 3.23 กรณีของการขาดหายของความถี่มูลฐาน /ch@@4/



รูปที่ 3.24 การเน้นเสียงของผู้พูดที่ท้ายพยางค์ของผู้พูดเพศหญิง /t@@0/



รูปที่ 3.25 เสียงจัตวาที่มีความถี่เพิ่มขึ้นอย่างชัดเจน /s@@4/



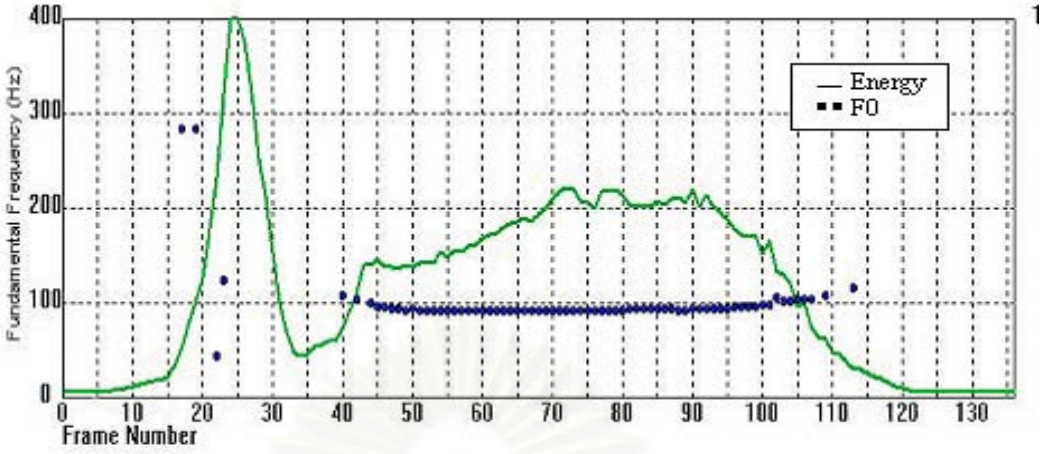
รูปที่ 3.26 เสียงจัตวาที่มีความถี่ค่อนข้างเพิ่ม /kh@@4/

ขั้นตอนที่4 สำหรับเสียงที่มีความกำกวมมาก มีการแกว่งของข้อมูลขึ้นๆ ลงๆ ดังรูปที่ 3.24 และรูปที่ 3.27 จากการทดสอบพบว่าเสียงตัวอย่างไม่ได้มีเฉพาะเสียงสามัญและจัตวาเท่านั้น แต่ในผู้พูดบางคนยังใช้การเน้นเสียงที่ท้ายพยางค์ ซึ่งเสียงที่ได้จะมีลักษณะคล้ายเสียงตรี (รูปที่ 3.24) อีกด้วย ซึ่งส่วนใหญ่จะเป็นผู้พูดเพศหญิงที่นิยมใช้ความถี่สูง ค่าความถี่มูลฐานที่ได้จะมีการแกว่งขึ้นลง ตรงข้ามกับเสียงจัตวาที่มีแนวโน้มขึ้นอย่างชัดเจน (รูปที่ 3.25) ส่วนผู้พูดเพศชายมักนิยมใช้ความถี่ต่ำ (รูปที่ 3.23, 3.26, 3.27 และ 3.28) ไม่มีการแกว่งขึ้นลงของข้อมูลแต่มีการตกลงอย่างชัดเจนของความถี่มูลฐานสำหรับเสียงสามัญ (รูปที่ 3.28) ดังนั้นจำนวนของ *เฟรมเพิ่มขึ้น* *เฟรมเท่ากัน* และ *เฟรมลดลง* จึงถูกนำมาใช้ในการตัดสินใจ โดยทำการปรับเรียบข้อมูลด้วยการกรองความถี่แบบกำหนดค่าตั้งแต่ *เฟรมอ้างอิง* ไปจนถึง *เฟรมท้าย* (ถ้าเฟรมถัดไปมีความต่างมากกว่า 20 เฮิรตซ์จะถูกกำจัดเพื่อรักษาข้อมูลที่มีความต่อเนื่องเท่านั้น) และตรวจสอบเงื่อนไข 2 ข้อต่อไปนี้

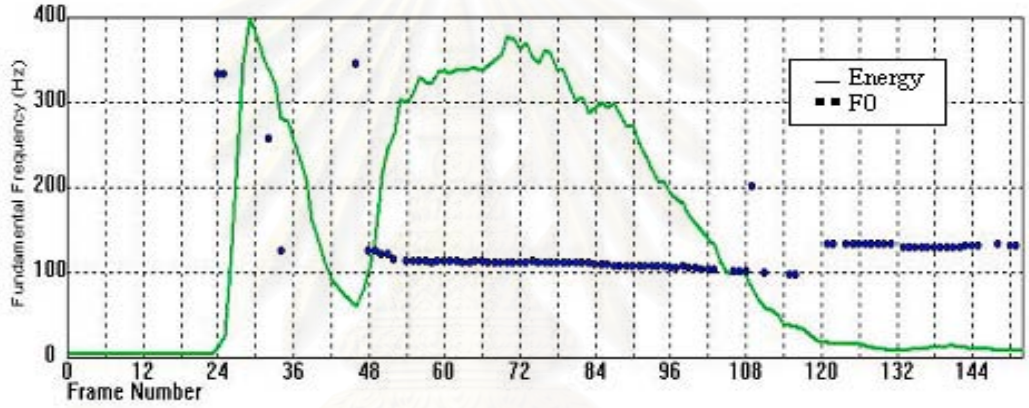
- สำหรับทุกๆ เฟรมตั้งแต่ *เฟรมอ้างอิง* ไปจนถึง *เฟรมท้าย* ถ้า *เฟรมเพิ่มขึ้น* หรือ *เฟรมเท่ากัน* มีจำนวนมากกว่า *เฟรมลดลง* จะคำนวณหาความชันของเสียงตัวอย่างนั้น ดังสมการที่ (3.4) ถ้าความชันมีค่ามากกว่า 2 เฮิรตซ์ จึงตัดสินให้เสียงตัวอย่างนั้นเป็นเสียงจัตวา
- สำหรับเสียงที่มีความกำกวมมาก ให้คำนวณตามสมการที่ (3.5) ถ้าเป็นจริงให้ตัดสินใจเป็นเสียงจัตวา

ความชัน = (ค่าความถี่ *เฟรมท้าย – ค่าอ้างอิง*) / (*เฟรมสุดท้าย - เฟรมอ้างอิง*) (3.4)

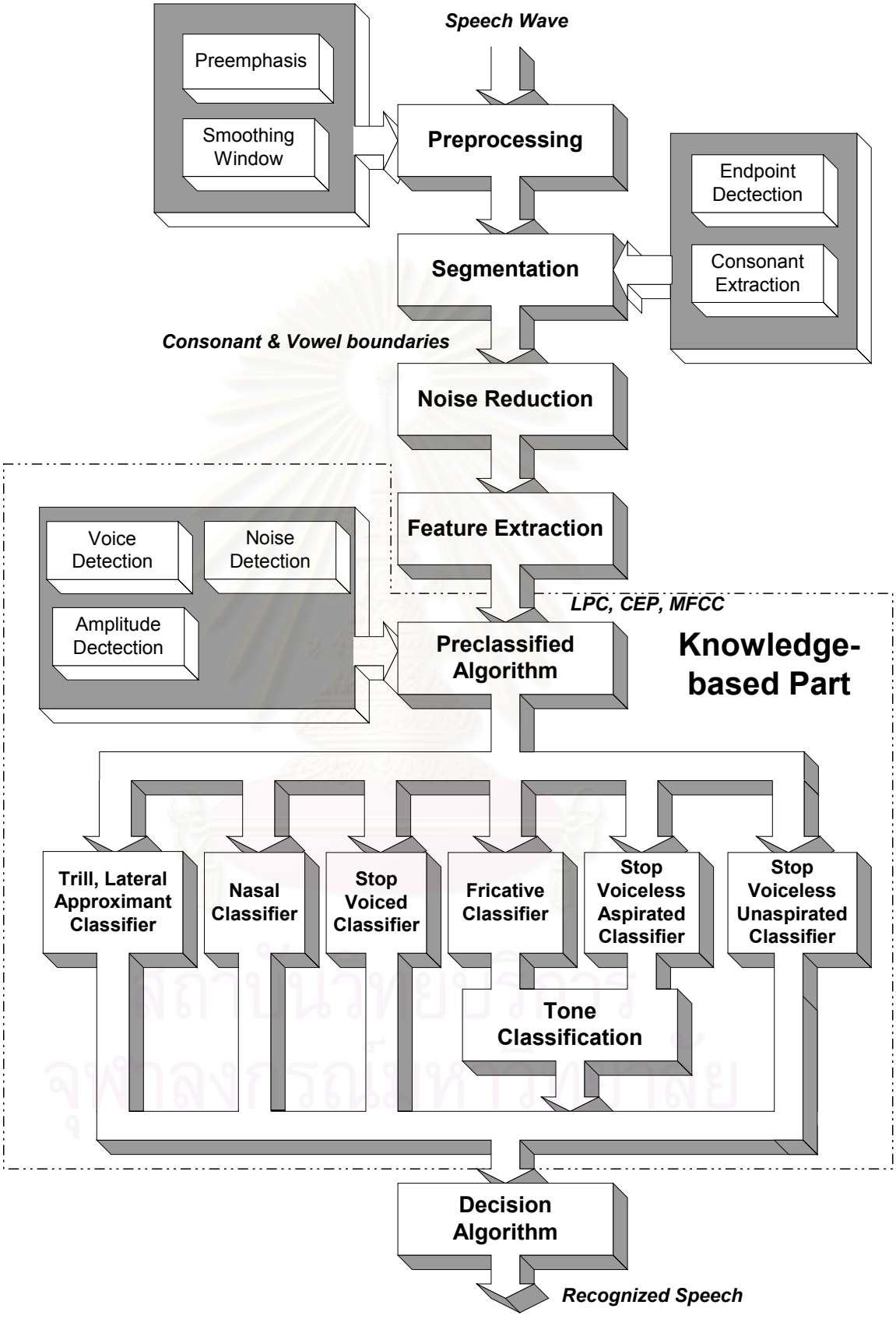
เฟรมเพิ่มขึ้น + *เฟรมลดลง* - *เฟรมเท่ากัน* > 0.125 เท่าของ *ช่วงของเฟรมที่เลือก* (3.5)



รูปที่ 3.27 เสียงที่มีความกำกวมมาก /ch@@4/



รูปที่ 3.28 เสียงสามัญของผู้พูดเพศชาย /ph@@0/



รูปที่ 3.29 แผนภาพแสดงระบบการรู้จำเสียงพยัญชนะ 28 เสียง

บทที่ 4
ผลการวิเคราะห์ข้อมูล

ในบทนี้กล่าวถึงรายละเอียดขั้นตอนการทดสอบการปรับเปลี่ยนพารามิเตอร์ต่างๆ ผลการทดสอบ และการวิเคราะห์ผลที่ได้จากการทดสอบ

กรรมวิธีการรู้จำเสียงคำเรียกพยัญชนะไทย

การรู้จำเสียงคำเรียกพยัญชนะไทยมีการเปรียบเทียบขั้นตอนการรู้จำ 2 แบบคือ 1) แบบของขั้นตอนวิธีการฐานความรู้ (Knowledge-based Algorithm) เป็นการแยกเสียงออกเป็น 10 ขั้นตอนตามลักษณะการเปล่งเสียงและตามฐานเสียง 2) แบบที่ไม่มีฐานความรู้ โดยทำการรู้จำข้อมูลเสียงทั้งหมดเป็น 28 ชุดต่างๆ กัน และทำการปรับค่าตัวแปร 3 ส่วนคือส่วนของข้อมูลเสียง ส่วนของลักษณะสำคัญ และส่วนของระบบจำลองฮิดเดน มาร์คอฟ

ส่วนของข้อมูลเสียง ทดสอบการกำจัดสัญญาณรบกวนและการตัดหน่วยเสียงสระ

ส่วนของลักษณะสำคัญ เปรียบเทียบค่าอันดับของลักษณะสำคัญ 5 ชนิดประกอบด้วยสัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น (LFCC) สัมประสิทธิ์การประมาณพันธะเชิงเส้น (LPC) ค่าสัมประสิทธิ์เซปสตรัมที่คำนวณจากการประมาณพันธะเชิงเส้น (CEPL) สัมประสิทธิ์เซปสตรัมที่คำนวณจากการแปลงดีสครีตฟูริเยร์ (CEPF) และสัมประสิทธิ์เซปสตรัมบนความถี่เมล (MFCC)

ส่วนของระบบจำลองฮิดเดน มาร์คอฟ ทำการปรับจำนวนของ Gaussian Mixture

ข้อมูลเสียงพูดประกอบด้วยเสียงของผู้พูดจำนวน 60 คนเป็นชาย 33 คน หญิง 27 คน ชุดฝึกฝนประกอบด้วยเสียงของผู้พูดจำนวน 40 คนแบ่งเป็นชาย 22 คน หญิง 18 คนนำไปสร้างและฝึกฝนระบบ ส่วนชุดทดสอบอีก 20 คนประกอบด้วยชาย 11 คนและหญิง 9 คน (รายละเอียดดังภาคผนวก ก)

ผลการเปรียบเทียบ

การทดสอบการกำจัดสัญญาณรบกวนนี้ ใช้การรู้จำของระบบแบบที่ไม่มีฐานความรู้ โดยรู้จำข้อมูลเสียงเป็น 28 ชุดต่างๆ กัน

4.1 ผลการทดสอบการกำจัดสัญญาณรบกวน

เสียงคำเรียกพยัญชนะในกลุ่มเสียงกัก-ไม่ก้อง-พ่นลม และเสียงเสียดแทรกมีลักษณะคล้ายสัญญาณรบกวน ดังนั้นในงานวิจัยนี้จึงทำการรู้จำสัญญาณเสียงทั้งที่ไม่ผ่านการกำจัดสัญญาณรบกวนและผ่านการกำจัดสัญญาณรบกวนโดยแบ่งสัญญาณรบกวนออกเป็น แบบที่มีความต่อเนื่อง คือสัญญาณรบกวนที่เกิดขึ้นอย่างต่อเนื่องที่ความถี่หนึ่งตลอดทั้งข้อมูลเสียงเช่น เสียงของสัญญาณที่ผ่านเข้าสู่ไมโครโฟนในกรณีที่ไม่ได้ทำการต่อสายดินไว้ และแบบไม่ต่อเนื่องคือสัญญาณเสียงที่เกิดขึ้นบ้างไม่เกิดขึ้นบ้างเช่น เสียงกริ่งโทรศัพท์ เสียงตะโกน เป็นต้น

ตารางที่ 4.1 ผลการรู้จำด้วยสัมประสิทธิ์ MFCC ที่อันดับ 10

ชนิดของสัญญาณรบกวน	ไม่ได้กำจัดสัญญาณรบกวน	แบบต่อเนื่อง	แบบไม่ต่อเนื่อง
อัตราการรู้จำ(ร้อยละ)	13.31	26.67	30.00

การกำจัดสัญญาณรบกวนแบบต่อเนื่องให้อัตราการรู้จำสูงขึ้นร้อยละ 13.36 ดังตารางที่ 4.1 และเมื่อกำจัดสัญญาณรบกวนที่มีลักษณะไม่ต่อเนื่องได้อัตราการรู้จำเพิ่มขึ้นอีกร้อยละ 3.33

4.2 ผลการทดสอบการตัดหน่วยเสียงสระ

1) ผลการทดสอบอัตราการรู้จำด้วยการนำข้อมูลที่ผ่านการหาจุดเริ่มต้นและสิ้นสุดของพยางค์ โดยปรับความยาวของข้อมูล 4 ลักษณะคือ ข้อมูลความยาวของเสียงทั้งหมด (หน่วยเสียงพยัญชนะ+หน่วยเสียงสระ) ข้อมูลเสียงตั้งแต่เฟรมศูนย์ไปจนถึงครึ่งหนึ่งของพยางค์ ข้อมูลเสียงตั้งแต่เฟรมศูนย์ไปจนถึงหนึ่งในสามของพยางค์ และข้อมูลหน่วยเสียงพยัญชนะดังตารางที่ 4.2 โดยเสียงคำเรียกพยัญชนะไทยมีข้อมูลเสียงสระออเหมือนกันทั้งหมด เสียงสระออไม่ได้ถูกใช้ในการแยกความแตกต่าง ดังนั้นการแยกข้อมูลของหน่วยเสียงพยัญชนะออกมาทำให้ได้ข้อมูลเฉพาะเสียงพยัญชนะที่ใช้แยกความแตกต่างได้ดีขึ้น

ตารางที่ 4.2 ผลการรู้จำด้วยสัมประสิทธิ์ MFCC ที่อันดับ 10

	ความยาวของเสียง	อัตราการรู้จำ(ร้อยละ)
แบบที่ไม่ให้ความรู้กับระบบ	ข้อมูลความยาวของเสียงทั้งหมด	30.00
	เสียงตั้งแต่เฟรมศูนย์ไปจนถึงครึ่งหนึ่งของพยางค์	34.10
	เสียงตั้งแต่เฟรมศูนย์ไปจนถึงหนึ่งในสามของพยางค์	37.30
	เสียงของหน่วยเสียงพยัญชนะ	41.00
แบบที่ให้ความรู้กับระบบ	เสียงของหน่วยเสียงพยัญชนะ	59.90

จากตารางที่ 4.2 พบว่าขั้นตอนวิธีการฐานความรู้สามารถเพิ่มอัตราการรู้จำให้กับระบบสูงขึ้นร้อยละ 18.8 เนื่องจากช่วยจำแนกความแตกต่างให้กับระบบและลดความกำกวมของเสียงที่เกิดขึ้นข้ามกลุ่ม ดังตารางที่ 4.3 แสดงตัวอย่างการรู้จำผิดพลาดข้ามกลุ่มของขั้นตอนการรู้จำที่ไม่มีฐานความรู้ เช่นเสียงปอ รู้จำผิดพลาดเป็นเสียงฟอและเสียงวอ โดยเสียงปอเป็นเสียงกลุ่มกัก-ไม่ก้อง-ไม่พ่นลม ที่มีลักษณะของแอมพลิจูดสูง ส่วนเสียงฟอที่มีลักษณะแอมพลิจูดต่ำแต่เสียงทั้งสองต่างก็มีลักษณะของ Noise-like ที่เหมือนกัน ส่วนเสียงวอเป็นเสียงกึ่งสระที่มีแอมพลิจูดต่ำกว่าและยังเป็นเสียงก้องอีกด้วย ดังนั้นในงานวิจัยนี้จึงทำการทดสอบหาค่าลักษณะสำคัญและค่าพารามิเตอร์ต่างๆ ที่เหมาะสมกับระบบฐานความรู้ต่อไป

	ป	ต	จ	ก	อ	พ	ผ	ท	ถ	ช	ฉ	ค	ข	บ	ด	ม	น	ง	ฟ	ฝ	ซ	ส	ฮ	ห	ร	ล	ว	ย		%	
ป	3	1	1	1	1		2			3				1			1		3	1					1		2		0.15	15	stop
ต	1	1	2	1	1			2		1			1			1			4	2	1	1	1				1		0.05	5	voiceless
จ	1	1	3	1	1			1				1					1	1	2		5	1			2		1	1	0.15	15	unaspirated
ก		1	1	3	2					2		1						1	1			3			3		2		0.15	15	
อ	2	1	2	1	5		1	1		1	1								1	1			1		1		3		0.25	25	
พ		1				3	3	2	1	2	1	1	1	1					2	3				1			1		0.3	30	stop
ผ		2			1	3	4	2	1	2	1	1	1			1			1	1							1		0.35	35	voiceless
ท		2			1	1	1	6	1	1	1	1	1						2	1				2	1				0.3	30	aspirated
ถ	1	1			1	1	2	3	1	1	1	1	1						2	3				2			2		0.15	15	
ช		1				1	1	1	1	5	7	1	1						1	1	2						1		0.6	60	
ฉ		1	1			1	1	1	1	6	5	1	1						1	1	2	2					1		0.5	50	
ค		1		1		1	1	1	1	1	1	5	6						3						1				0.65	65	
ข		3		1	1	1	1	1	1	1	1	6	2				2		1				1	1					0.4	40	
บ														13	1	2	2								2		1		0.65	65	stop
ด														1	13	1	2	1							1			2	0.65	65	voiced
ม				1	1									6	1	8	2	1									1		0.4	40	nasal
น			1	2										1	2	1	12	1							1				0.6	60	
ง				1			1									1	1	8		1			1		1		2	1	0.4	40	
ฟ		1	1	1		1	1	1		1	1							1	6	3	1	1	1	1	2				0.45	45	fricative
ฝ		1				1				1								1	4	6	1	1	1	1	2		3		0.5	50	
ซ								1		2					1		1		1	1	5	6	1	1				2	0.55	55	
ส								1			1	1					1		1	2	4	7	1	1	2				0.55	55	
ฮ	1	3			2		1	2					1						1	1	1	1	7	2					0.45	45	
ห	1	2		1	3		2	1				1							1	1	1	1	7	1					0.4	40	
ร		1		1													1	1					1		12		1	2	0.6	60	trill
ล		1	4											3				4							6	1	2		0	0	lateral
ว	1							1							1	2			2						1		12	1	0.6	60	approximant
ย			1												1			1							3		1	14	0.7	70	
Total																													0.41	41	

ตารางที่ 4.3 ผลการรู้จำหน่วยเสียงพยัญชนะทั้ง 28 เสียงด้วยสัมประสิทธิ์ MFCC อันดับ 10

2) ผลของการเปรียบเวลาที่ใช้ในการประมวลผลและขนาดของข้อมูลเสียงก่อนและหลังการปรับความยาวข้อมูล สำหรับข้อมูลเสียง 100 ตัวอย่าง ดังตารางที่ 4.4 (ก) และ 4.4 (ข) จะเห็นได้ว่าการปรับลดความยาวของข้อมูลทำให้เวลาที่ใช้ในระบบรู้จำลดลงจาก 10 นาทีเป็น 2 นาทีและขนาดของข้อมูลก็ลดลงประมาณ 4 เท่า

ตารางที่ 4.4 (ก) เวลาที่ใช้ในการประมวลผลก่อนและหลังการปรับความยาวข้อมูล (ข้อมูล 100 ตัวอย่าง)

เวลาที่ใช้	ความยาวข้อมูลเสียง	
	เสียงทั้งหมด (พยัญชนะ+สระ)	หน่วยเสียงพยัญชนะ
การปรับความยาวเสียง (วินาที)	-	30
การคำนวณค่าลักษณะสำคัญ (วินาที)	15	3
การฝึกฝนและทดสอบระบบ (นาที)	10	2
เวลารวมที่ใช้ทั้งระบบ	10 นาที 15 วินาที	2 นาที 33 วินาที

ตารางที่ 4.4 (ข) ขนาดของข้อมูลเสียงก่อนและหลังการปรับความยาวข้อมูล (ข้อมูล 100 ตัวอย่าง)

	ความยาวข้อมูลเสียง	
	เสียงทั้งหมด (พยัญชนะ+สระ)	หน่วยเสียงพยัญชนะ
ขนาดของข้อมูล (กิโลบิต)	2160	489
ขนาดของข้อมูลหลังจากการคำนวณค่าลักษณะสำคัญ (กิโลบิต)	1280	287

4.3 ผลการทดสอบลักษณะสำคัญ

เปรียบเทียบค่าลักษณะสำคัญ 5 ลักษณะพร้อมทั้งปรับอันดับสัมประสิทธิ์ที่อันดับ 10, 15, 20, 25 และ 30 โดยใช้ค่าพารามิเตอร์ของระบบจำลองฮิดเดน มาร์คอฟที่ 5 States, 1 Mixture

จากตารางที่ 4.5 พบว่าสัมประสิทธิ์ LFCC, LPC, และ CEPL ให้อัตราการรู้จำต่ำกว่า CEPF และ MFCC เพราะต่างก็มาจากคุณสมบัติของความเป็นเชิงเส้นซึ่งไม่เหมาะสมต่อการแทนคุณสมบัติของเสียงที่มีลักษณะของสัญญาณรบกวน เช่น เสียงกัก-ไม่ก้อง-พ่นลม หรือเสียงเสียดแทรก ซึ่งเป็นเสียงครึ่งหนึ่งในเสียงคำเรียกพยัญชนะไทย โดย LFCC เป็นการแปลงสัญญาณเสียงให้อยู่ในรูปความถี่เชิงเส้นแล้วผ่านลอการิทึมเน้นสัญญาณในช่วงความถี่ต่ำให้อัตราการรู้จำร้อยละ 14.30 ส่วนสัมประสิทธิ์ LPC มีวิธีการเลือกค่าสัมประสิทธิ์ให้มีค่าเฉลี่ยของค่าผิดพลาด (ระหว่างข้อมูลเสียงต้นแบบกับค่าลักษณะสำคัญที่คำนวณได้ในแต่ละเฟรม) มีค่าน้อยที่สุดจึงให้อัตราการรู้จำที่สูงขึ้น ในขณะที่ CEPL นำค่าสัมประสิทธิ์ LPC มาผ่านขั้นตอน

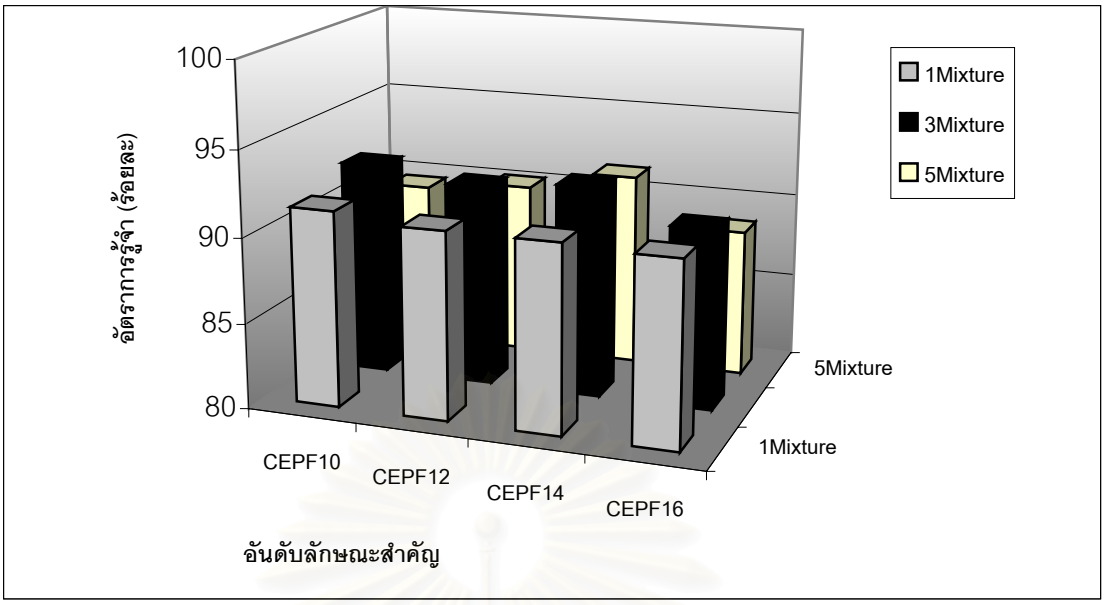
Cepstrum Liftering ที่แยกลักษณะโครงร่างสเปกตรัม (Global Pattern) ออกจากความถี่มูลฐาน ทำให้ได้อัตรารู้จำที่สูงสุดของคุณสมบัติความเป็นเชิงเส้น ส่วน CEPF ผ่านขั้นตอนของ Cepstrum Liftering เช่นเดียวกันแต่คำนวณจากการแปลงฟูริเยร์โดยตรง ในขณะที่สัมประสิทธิ์ MFCC ผ่านการแปลงสัญญาณเสียงด้วยความถี่เมล (วงจรกรองสามเหลี่ยม) ซึ่งไม่เป็นเชิงเส้น ทำให้ได้อัตราการรู้จำสูงที่สุด

ตารางที่ 4.5 ผลการรู้จำด้วยลักษณะสำคัญ LFCC, LPC, CEPL, CEPF และ MFCC ที่อันดับ 10 15 20 25 และ 30

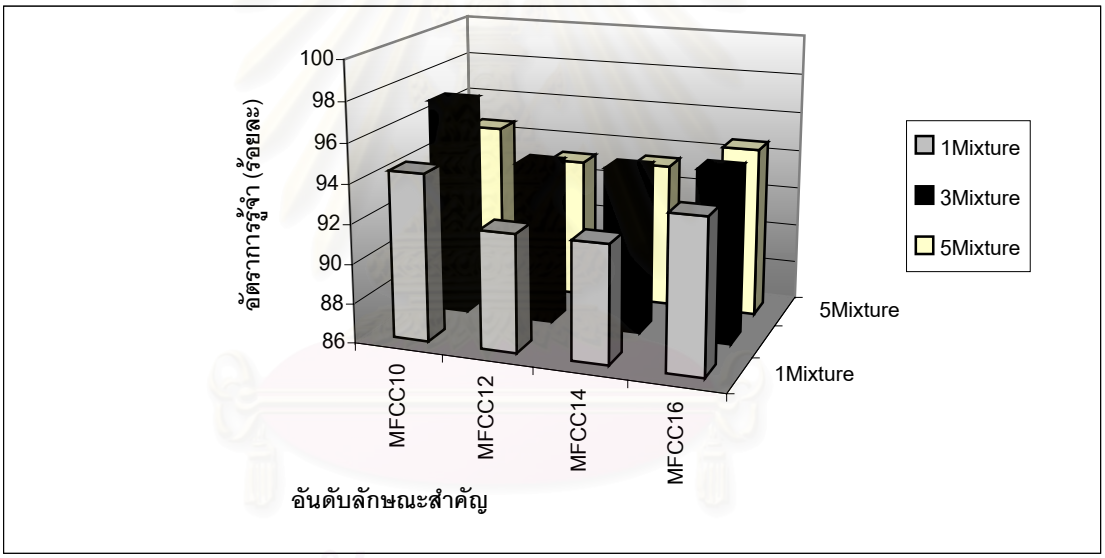
อัตราการรู้จำ(ร้อยละ)					
ลักษณะสำคัญ	**สัมประสิทธิ์อันดับที่**				
	10	15	20	25	30
LFCC	14.30	12.37	12.52	11.67	11.29
LPC	21.30	25.37	24.58	24.86	23.20
CEPL	36.60	43.26	11.01	8.05	7.03
CEPF	51.72	50.37	50.12	47.00	45.45
MFCC	59.90	54.29	50.76	48.53	46.44

เนื่องจากสัมประสิทธิ์ไม่เชิงเส้นเหมาะสมต่อระบบการรู้จำเสียงคำเรียกพยัญชนะมากกว่าแบบเชิงเส้นดังนั้นจึงเลือกใช้สัมประสิทธิ์ MFCC และสัมประสิทธิ์ CEPF ที่ให้อัตราการรู้จำสูงที่สุด โดยสัมประสิทธิ์ MFCC และ CEPF ให้อัตราการรู้จำสูงที่สุดที่อันดับ10 และ 15 ดังนั้นจึงทดสอบปรับค่าอันดับที่ 10, 12, 14, 16 และปรับค่าอันดับของสัมประสิทธิ์ CEPF ที่ 10, 12, 14, 16 โดยปรับค่าพารามิเตอร์ระบบจำลองฮิดเดน มาร์คอฟที่ 1 Mixture, 3 Mixture และ 5 Mixture โดยคงค่าสถานะที่ 5 States

จากผลการทดสอบดังรูปที่ 4.1 (ก) และ 4.1 (ข) พบว่าสัมประสิทธิ์ MFCC และ CEPF อันดับ 10 ที่ 3 Gaussian Mixture ให้อัตราการรู้จำดีที่สุดถึงแม้ว่าการเพิ่มจำนวนของ Gaussian Mixture จากหนึ่งเป็นสามทำให้อัตราการรู้จำสูงขึ้นแต่เวลาที่ใช้ก็สูงขึ้นด้วยเนื่องจากการคำนวณที่ซับซ้อนขึ้น โดยการคำนวณที่ 1 Mixture ใช้เวลาเฉลี่ยประมาณ 1 ถึง 2 นาที ส่วน 3 Mixture ใช้เวลาเฉลี่ยประมาณ 13 ถึง 14 นาที ในขณะที่ 5 Mixture ใช้เวลาเฉลี่ยประมาณ 20-22 ชั่วโมง ดังนั้นจึงเลือกใช้สัมประสิทธิ์ MFCCและ CEPF อันดับ 10 ที่ 3 Gaussian Mixture ในการปรับปรุงขั้นต่อไป



รูปที่ 4.1 (ก) ผลการรู้จำด้วยสัมประสิทธิ์ CEPF อันดับ 10, 12, 14 และ 16



รูปที่ 4.1 (ข) ผลการรู้จำด้วยสัมประสิทธิ์ MFCC อันดับ 10, 12, 14 และ 16

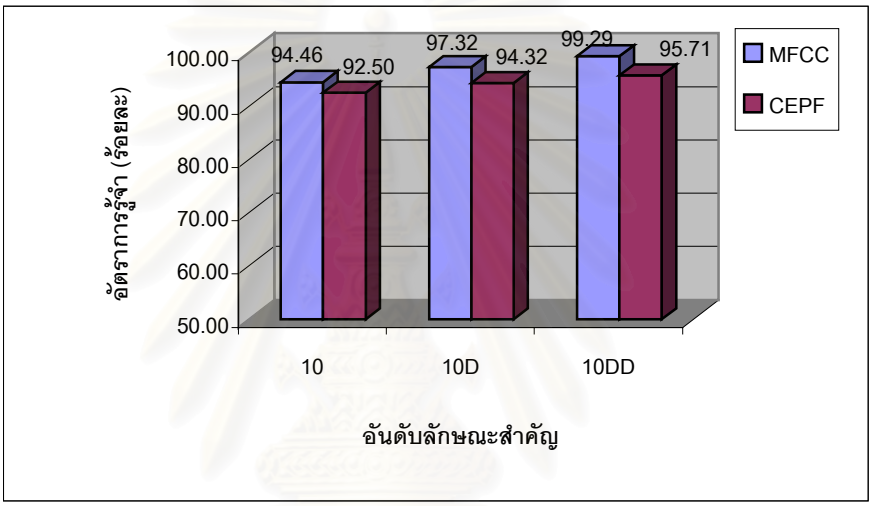
4.4 ขั้นตอนวิธีการฐานความรู้

ส่วนของขั้นตอนวิธีการฐานความรู้ ประกอบด้วย 2 ส่วนคือ ส่วนของขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง และส่วนของขั้นตอนการรู้จำเสียงตามฐานเสียง ซึ่งแบ่งออกเป็นขั้นตอนในการรู้จำทั้งหมด 10 ขั้นตอน อยู่ในส่วนของขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง 4 ขั้นตอน และอยู่ในส่วนของขั้นตอนการรู้จำเสียงตามฐานเสียงอีก 6 ขั้นตอน ดังต่อไปนี้

ขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง (Preclassified Algorithm)

ในส่วนนี้เป็นรายละเอียดและผลการรู้จำของขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียง ดังรายละเอียดในบทที่ 3 ซึ่งประกอบไปด้วยการเปรียบเทียบผลการรู้จำของสัมประสิทธิ์ MFCC, MFCC Delta (MFCCD), MFCC Delta Different (MFCCDD) สัมประสิทธิ์ CEPF, CEPF Delta (CEPFD) และ CEPF Delta Different (CEPFDD) อันดับ10 ที่ 3 Gaussian Mixture

1) ขั้นตอนที่ 1 การแยกเสียงก้อง (เสียงก้อง) ออกจากเสียงไม่ก้อง (Voiceless) จากรูปที่ 4.2 พบว่าสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ให้อัตราการรู้จำสูงสุดที่ร้อยละ 99.29 โดยแสดงผลการรู้จำดังตารางที่ 4.6



รูปที่ 4.2 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 1

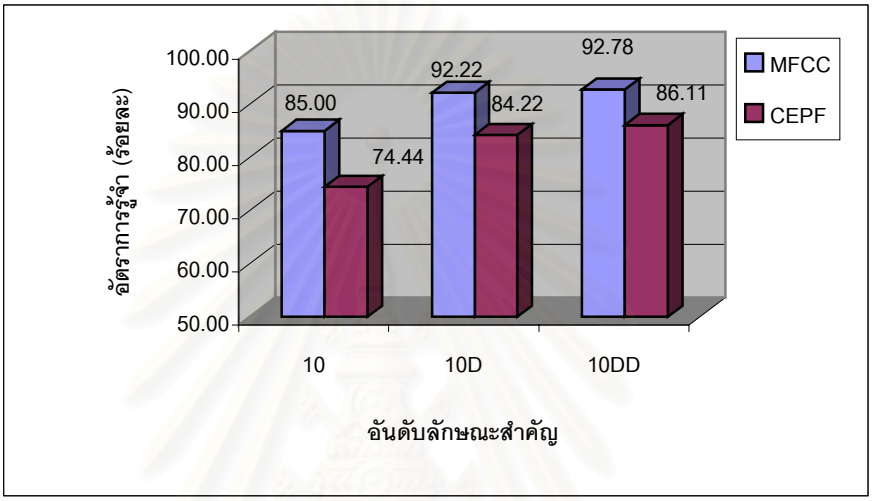
ตารางที่ 4.6 ผลการรู้จำในขั้นตอนที่ 1

กลุ่มเสียง	จำนวนข้อมูล (560)	รู้จำเป็นเสียงก้อง	รู้จำเป็นเสียงไม่ก้อง
เสียงก้อง	180	178 (98.88%)	2 (1.12%)
เสียงไม่ก้อง	380	2 (0.53.%)	378 (99.47%)
ขั้นตอนที่ 1 อัตราการรู้จำเฉลี่ยร้อยละ 99.29			

ลักษณะการเกิดของเสียงก้องจะมีการสั่นของเส้นเสียงก่อนการเปิดของฐานกรณ์ ซึ่งสามารถแยกความแตกต่างระหว่างเสียงก้องและไม่ก้องได้จากลักษณะสำคัญหลายๆ ลักษณะ เช่น การลดต่ำลงของอัตราการตัดผ่านศูนย์ อัตราการตัดผ่านระดับกำหนดมีค่าต่ำในขณะที่ค่าพลังงานก็มีค่าต่ำ และการมีค่าพลังงานที่ความถี่ต่ำในช่วงที่เส้นเสียงสั่น ซึ่งสามารถตรวจสอบ

ได้ง่ายและเมื่อมีลักษณะสำคัญใดไม่สามารถบอกความแตกต่างได้ชัดเจน ก็ยังคงมีลักษณะสำคัญอื่นๆ ให้ตรวจสอบอีก ดังนั้นอัตราการรู้จำจึงมีค่าสูง

2) ขั้นตอนที่ 2 การแยกเสียงก้องออกเป็น 3 กลุ่มคือกลุ่มเสียงกัก-ก้อง กลุ่มเสียงนาสิก และกลุ่มเสียงกึ่งสระ พบว่าสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ให้อัตราการรู้จำสูงสุดที่ร้อยละ 92.78 ดังรูปที่ 4.3



รูปที่ 4.3 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 2

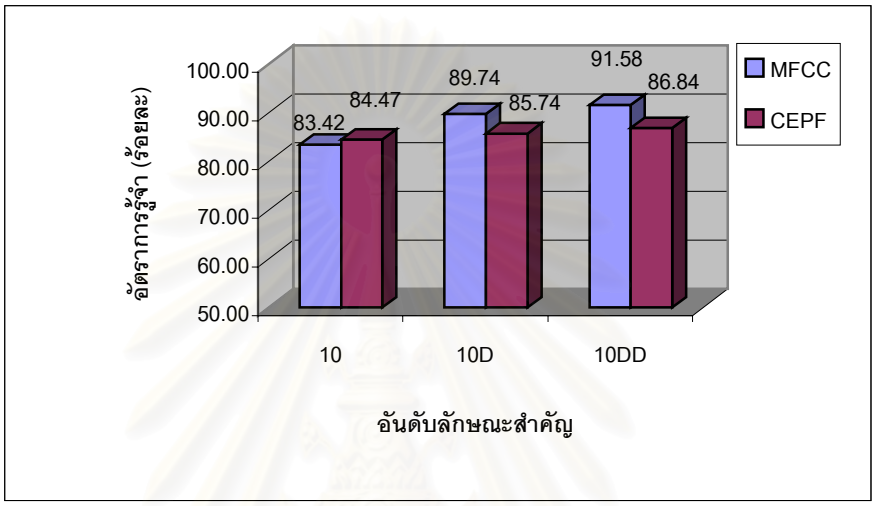
ตารางที่ 4.7 ผลการรู้จำในขั้นตอนที่ 2

กลุ่มเสียง	จำนวนข้อมูล (180)	รู้จำเป็นเสียงกัก-ก้อง	รู้จำเป็นเสียงนาสิก	รู้จำเป็นเสียงกึ่งสระ
เสียงกัก-ก้อง	40	36 (90.00%)	4 (10.00%)	0
เสียงนาสิก	60	2 (3.33%)	56 (93.34%)	2 (3.33%)
เสียงกึ่งสระ	80	1 (1.25%)	4 (5.00%)	75 (93.75%)
ขั้นตอนที่ 2 อัตราการรู้จำเฉลี่ยร้อยละ 92.78				

เนื่องจากในขั้นตอนนี้ใช้วิธี Amplitude Detection ในการแยกกลุ่มของข้อมูลโดยเสียงกัก-ก้องจะเป็นเสียงที่มีแอมพลิจูดสูงที่สุด เสียงนาสิกมีขนาดของแอมพลิจูดขนาดกลางคือต่ำกว่าเสียงกัก-ก้องแต่สูงกว่าเสียงกึ่งสระ และเสียงกึ่งสระจะมีแอมพลิจูดของพลังงานต่ำที่สุด ดังนั้นจากตารางที่ 4.7 จะพบว่าเสียงกัก-ก้องซึ่งมีแอมพลิจูดสูงที่สุดรู้จำผิดพลาดไปเป็นเสียงนาสิกซึ่งมีแอมพลิจูดขนาดกลางแต่ไม่รู้จำผิดพลาดไปเป็นเสียงกึ่งสระซึ่งมีแอมพลิจูดต่ำที่สุด ส่วนเสียงนาสิกซึ่งมีแอมพลิจูดขนาดกลางจึงมีการรู้จำผิดพลาดไปเป็นเสียงที่มีแอมพลิจูดขนาดสูงและ

ขนาดต่ำอย่างละ 2 เสียงเท่าๆ กัน ส่วนเสียงกึ่งสระซึ่งมีแอมพลิจูดต่ำที่สุดมีการรู้จำผิดพลาดไปเป็นเสียงที่มีแอมพลิจูดขนาดกลาง 4 เสียงและรู้จำเป็นเสียงที่มีแอมพลิจูดสูง 1 เสียงเท่านั้น

3) ขั้นตอนที่ 3 การแยกเสียงไม่ก้องออกเป็น 2 กลุ่มตามลักษณะของความคล้ายและไม่คล้ายสัญญาณรบกวน (Noise และ Non-noise) โดยสัมประสิทธิ์ MFCCDD อันดับ 10 ให้อัตราการรู้จำสูงที่สุดคือร้อยละ 91.58 ดังรูปที่ 4.4



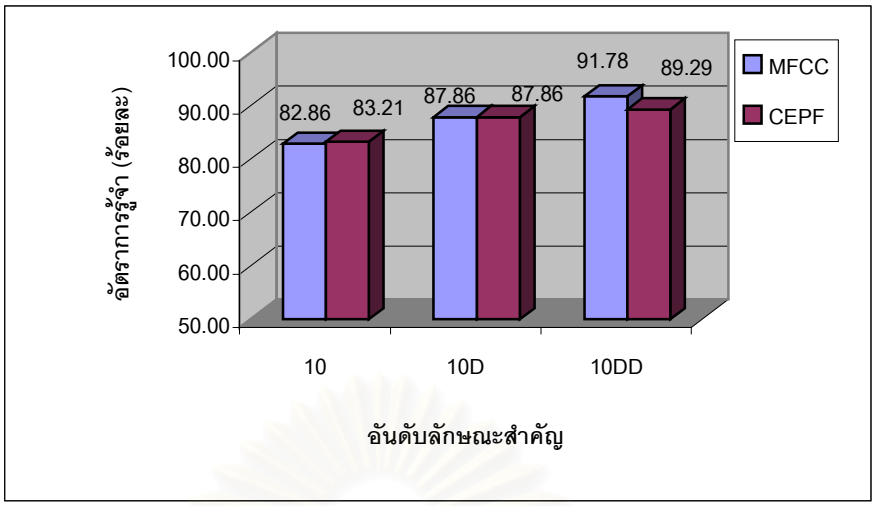
รูปที่ 4.4 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 3

ตารางที่ 4.8 ผลการรู้จำในขั้นตอนที่ 3

กลุ่มเสียง	จำนวนข้อมูล (380)	รู้จำเป็นเสียง Noise	รู้จำเป็นเสียง Non-Noise
Noise	280	256 (91.43%)	24 (8.57%)
Non-Noise	100	8 (8.00%)	92 (92.00%)
ขั้นตอนที่ 3 อัตราการรู้จำเฉลี่ยร้อยละ 91.58			

ขั้นตอนนี้ใช้วิธี Noise Detection ในการแยกเสียงกัก-ไม่ก้อง-ไม่พ่นลม (Non-Noise) ออกจากเสียงกัก-ไม่ก้อง-พ่นลมและเสียงเสียดแทรก ซึ่งมีลักษณะคล้ายสัญญาณรบกวน ดังนั้นการมีสัญญาณรบกวนจึงส่งผลกระทบต่ออัตราการรู้จำได้ โดยแสดงผลการรู้จำดังตารางที่ 4.8

4) ขั้นตอนที่ 4 การแยกกลุ่มเสียง Noise ออกเป็นเสียงกัก-ไม่ก้อง-พ่นลม และ เสียงเสียดแทรก โดยสัมประสิทธิ์ MFCCDD อันดับ 10 ให้อัตราการรู้จำสูงสุดร้อยละ 91.78 ดังรูปที่ 4.5



รูปที่ 4.5 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 4

ตารางที่ 4.9 ผลการรู้จำในขั้นตอนที่ 4

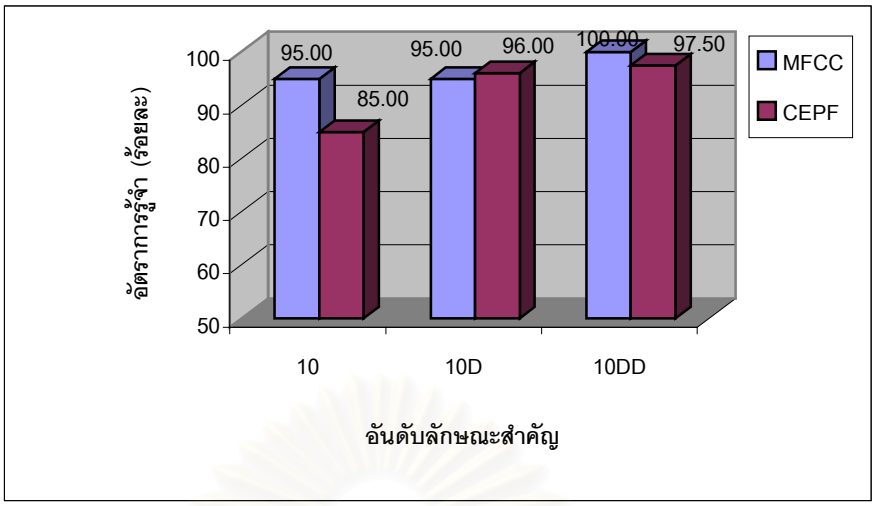
กลุ่มเสียง	จำนวนข้อมูล (280)	รู้จำเป็นเสียงกัก-ไม่ก้อง-พ่นลม	รู้จำเป็นเสียงเสียดแทรก
เสียงกัก-ไม่ก้อง-พ่นลม	160	148 (92.50%)	12 (7.50%)
เสียงเสียดแทรก	120	11 (9.16%)	109 (90.84%)
ขั้นตอนที่ 4 อัตราการรู้จำเฉลี่ยร้อยละ 91.78			

ขั้นตอนนี้ใช้วิธี Amplitude Detection แยกความแตกต่างระหว่างเสียงที่มีแอมพลิจูดสูง (เสียงกัก-ไม่ก้อง-พ่นลม) กับเสียงที่มีแอมพลิจูดต่ำ (เสียงเสียดแทรก) ให้ผลการรู้จำดังตารางที่ 4.9

ขั้นตอนการรู้จำเสียงตามฐานเสียง

จากส่วนของขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียงในหัวข้อที่แล้ว จะแบ่งกลุ่มของเสียงออกเป็นกลุ่มใหญ่ๆ ได้ 4 กลุ่ม ดังนั้นในส่วนนี้จึงทำการรู้จำเสียงตามฐานเสียงเพื่อแยกเสียงทั้งหมดให้ออกเป็น 21 เสียง ซึ่งประกอบไปด้วย ขั้นตอนที่ 5 ถึง ขั้นตอนที่ 10

5) ขั้นตอนที่ 5 การแยกกลุ่มเสียงกัก-ก้อง (Stop Voiced Classifier) พบว่าสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ให้อัตราการรู้จำสูงสุดที่ร้อยละ 100 ดังรูปที่ 4.6



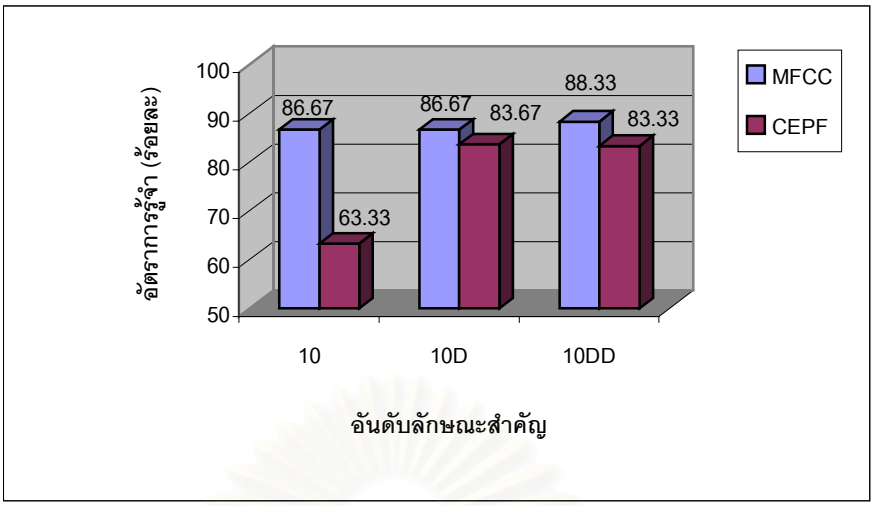
รูปที่ 4.6 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 5

ตารางที่ 4.10 ผลการรู้จำในขั้นตอนที่ 5

กลุ่มเสียง	จำนวนข้อมูล (40)	รู้จำเป็นเสียงบอ/b@@/	รู้จำเป็นเสียงดอ /d@@/
เสียงบอ /b@@/	20	20 (100%)	0
เสียงดอ /d@@/	20	0	20 (100%)
ขั้นตอนที่ 5 อัตราการรู้จำเฉลี่ยร้อยละ 100.00			

เสียงบอจะมีข้อมูลเสียงส่วนใหญ่อยู่ที่ความถี่ต่ำ ตรงกันข้ามกับเสียงดอที่มีข้อมูลเสียงส่วนใหญ่อยู่ที่ความถี่สูง และยังมีคุณลักษณะของความถี่ฟอร์แมนท์ที่สองต่ำลงในเสียงบอแต่สูงขึ้นสำหรับเสียงดอ (รายละเอียดดังบทที่ 3) ดังนั้นการรู้จำเสียงระหว่างเสียงบอและเสียงดอจึงให้อัตราการรู้จำสูงที่ร้อยละ 100 ดังตารางที่ 4.10

6) ขั้นตอนที่ 6 การแยกเสียงกลุ่มเสียงนาสิก (Nasal Classifier) โดยสัมประสิทธิ์ MFCCDD อันดับ 10 ให้อัตราการรู้จำสูงที่สุดที่ร้อยละ 88.33 ดังรูปที่ 4.7



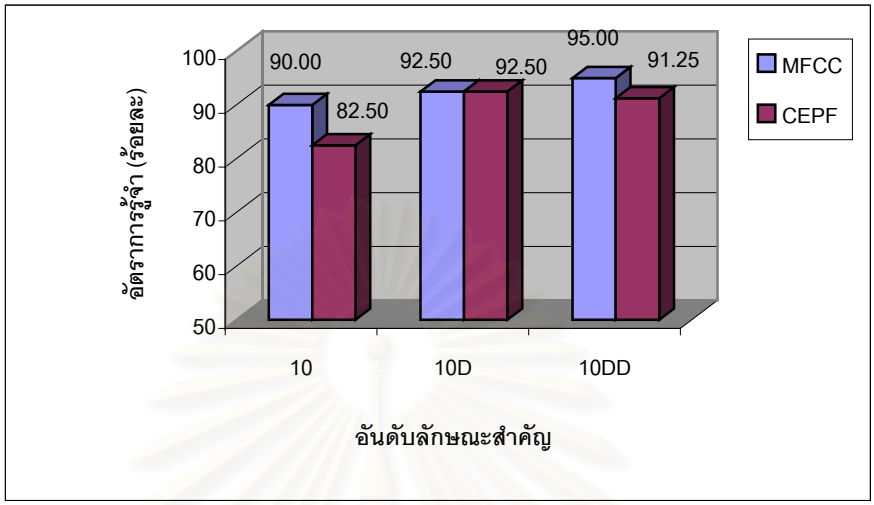
รูปที่ 4.7 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 6

ตารางที่ 4.11 ผลการรู้จำในขั้นตอนที่ 6

เสียง	จำนวนข้อมูล (60)	รู้จำเป็นเสียงมอ /m@@/	รู้จำเป็นเสียงนอ /n@@/	รู้จำเป็นเสียงงอ /ng@@/
เสียงมอ /m@@/	20	17 (85.00%)	2 (10.00%)	1 (5.00%)
เสียงนอ /n@@/	20	0	20 (100.00%)	0
เสียงงอ /ng@@/	20	0	4 (20.00%)	16 (80.00%)
ขั้นตอนที่ 6 อัตราการรู้จำเฉลี่ยร้อยละ 88.33				

เสียงนาสิกมีลักษณะของสเปกตรัมที่คล้ายคลึงกันมากโดยจะมีคุณลักษณะของความถี่ฟอร์แมนท์ที่ใช้แยกความแตกต่างดังนี้คือ เสียงมอที่เกิดขึ้นที่ฐานปากจะมีการตกลงของฟอร์แมนท์ที่สอง เสียงนอเกิดที่ฐานปุ่มเหงือกจะมีฟอร์แมนท์ที่สองสูงขึ้นและเสียงงอที่เกิดที่ฐานเพดานอ่อนจะมีฟอร์แมนท์ที่สองเพิ่มสูงขึ้นมากที่สุด จะเห็นได้ว่าการเปรียบเทียบแบบสัมพัทธ์ (Relative) ของฟอร์แมนท์ที่สองจะเกิดความกำกวมขึ้นระหว่างการสูงขึ้นของฟอร์แมนท์ที่สองในเสียงนอกับการสูงขึ้นในเสียงงอ โดยเสียงงอรู้จำผิดไปเป็นเสียงนอ 4 เสียงดังตารางที่ 4.11 และความกำกวมของฟอร์แมนท์ที่สองในเสียงมอซึ่งมีการรู้จำผิดพลาดไปเป็นเสียงนอและเสียงงอเนื่องจากระดับของความถี่ฟอร์แมนท์ มีระดับของ การตกลง การคงที่ ค่อนข้างคงที่ หรือ ค่อนข้างเพิ่ม โดยถ้าตั้งระดับกำหนด (Threshold) ของฟอร์แมนท์ที่สองสำหรับเสียงงอให้สูงขึ้นจะทำให้เสียงมอไม่รู้จำผิดเป็นเสียงงอ แต่ก็ทำให้เสียงงอรู้จำผิดพลาดเป็นเสียงนอมากขึ้นเช่นกัน

7) ขั้นตอนที่ 7 การแยกเสียงกลุ่มของเสียงลิ้นรัว เสียงข้างลิ้นและเสียงต่อเนื่อง (Trill, Lateral และ Approximant Classifier) พบว่าสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ให้อัตราการรู้จำสูงสุดที่ร้อยละ 95.00 ดังรูปที่ 4.8



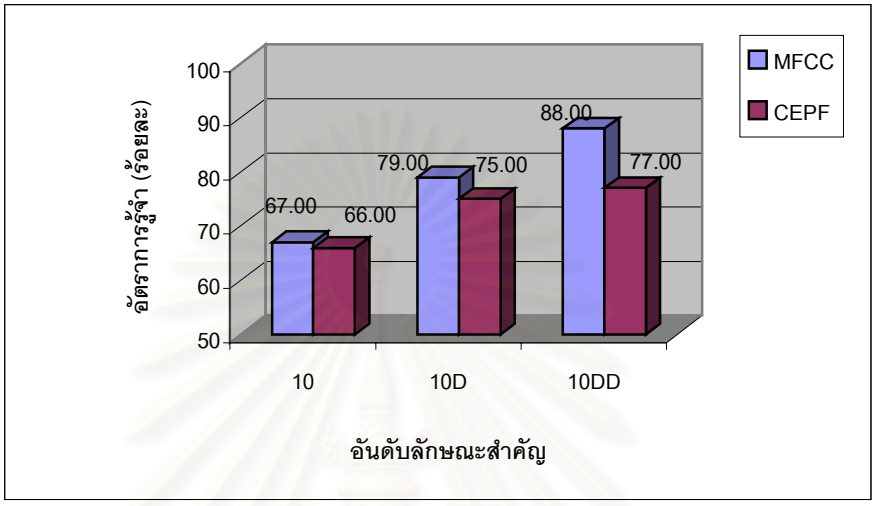
รูปที่ 4.8 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 7

ตารางที่ 4.12 ผลการรู้จำเฉลี่ยในขั้นตอนที่ 7

เสียง	จำนวนข้อมูล (80)	รู้จำเป็นเสียงรอ /r@@/	รู้จำเป็นเสียงลอ /l@@/	รู้จำเป็นเสียงวอ /w@@/	รู้จำเป็นเสียงยอ /j@@/
เสียงรอ /r@@/	20	19 (95.00%)	1 (5.00%)	0	0
เสียงลอ /l@@/	20	1 (5.00%)	19 (95.00%)	0	0
เสียงวอ /w@@/	20	0	2 (10.00%)	18 (90.00%)	0
เสียงยอ /j@@/	20	0	0	0	20 (100.00%)
ขั้นตอนที่ 7 อัตราการรู้จำเฉลี่ยร้อยละ 95.00					

เสียงรอกับเสียงลอเป็นเสียงที่มีลักษณะการเปล่งเสียงที่ต่างกันคือเสียงรอเป็นเสียงลิ้นรัว ส่วนเสียงลอเป็นเสียงข้างลิ้น แต่เกิดขึ้นที่ฐานปุ่มเหงือกเหมือนกันซึ่งสำหรับข้อมูลเสียงจริงผู้พูดบางคนไม่สามารถออกเสียงลิ้นรัวได้ (จะเห็นได้ชัดจากการออกเสียงควบกล้ำ ร-เรือ ล-ลิง) ดังนั้นจึงมีการรู้จำผิดพลาดที่เกิดขึ้นระหว่างสองเสียงนี้ ส่วนเสียงวอและเสียงยอเป็นเสียงต่อเนื่องซึ่งเสียงวอเกิดที่ฐานปาก ส่วนเสียงยอเกิดขึ้นที่ฐานเพดานแข็ง โดยฐานเพดานแข็งนี้เป็นฐานที่มีคุณลักษณะของพลังงานที่ความถี่สูงอย่างชัดเจน ทำให้เสียงยอมีอัตราการรู้จำสูงที่สุดในกลุ่มเสียงกึ่งสระและไม่มีเสียงอื่นๆ ที่รู้จำผิดพลาดมาเป็นเสียงยอ รวมถึงเสียงวอที่มีลักษณะการเปล่งเสียงเหมือนกันแต่กลับรู้จำผิดพลาดไปเป็นเสียงลอดังตารางที่ 4.12

8) ขั้นตอนที่ 8 การแยกกลุ่มเสียงไม่ก้องและ Non-noise ซึ่งก็คือเสียงกัก-ไม่ก้อง-ไม่พ่นลม (Stop Voiceless Unaspirated) โดยสัมประสิทธิ์ MFCCDD อันดับ 10 ให้อัตราการรู้จำสูงที่สุดที่ร้อยละ 88.00 ดังรูปที่ 4.9



รูปที่ 4.9 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 8

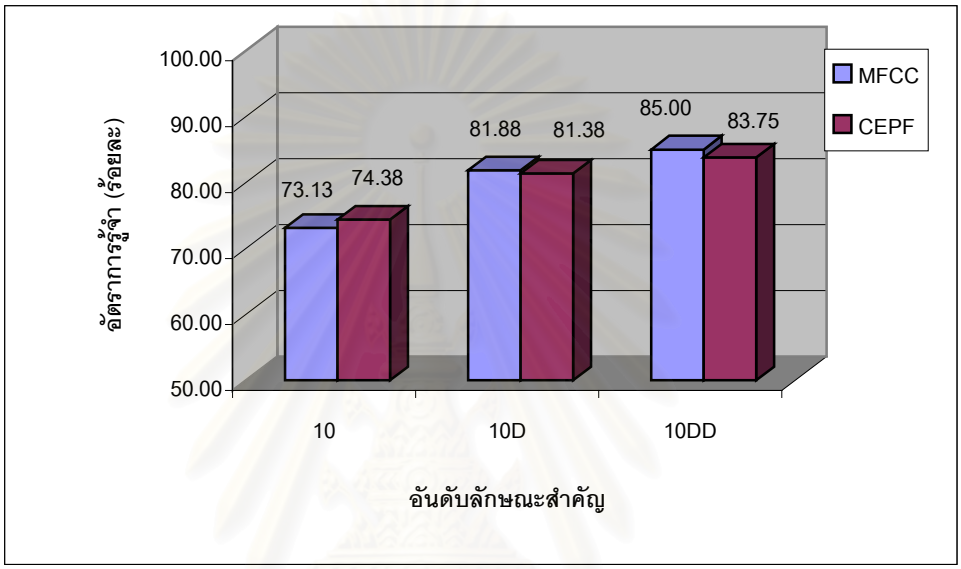
ตารางที่ 4.13 ผลการรู้จำในขั้นตอนที่ 8

เสียง	จำนวนข้อมูล (100)	รู้จำเป็นเสียง ปอ /p@@/	รู้จำเป็นเสียง ตอ /t@@/	รู้จำเป็นเสียง จอ /c@@/	รู้จำเป็นเสียง กอ /k@@/	รู้จำเป็นเสียง ออ /@@/
เสียงปอ /p@@/	20	16 (80.00%)	3 (15.00%)	0	1 (5.00%)	0
เสียงตอ /t@@/	20	0	18 (90.00%)	0	2 (10.00%)	0
เสียงจอ /c@@/	20	0	0	20 (100.00%)	0	0
เสียงกอ /k@@/	20	0	0	0	20 (100.00%)	0
เสียงออ /@@/	20	1 (5.00%)	5 (25.00%)	0	0	14 (70.00%)
ขั้นตอนที่ 8 อัตราการรู้จำเฉลี่ยร้อยละ 88.00						

เสียงปอ ตอและกอคล้ายกับเสียงมอ นอและงอคือต่างก็มีลักษณะของสเปกตรัมที่คล้ายคลึงกันและใช้ระดับสัมพัทธ์ของฟอร์แมนท์ที่สองแยกความแตกต่าง โดยเสียงปอมีฟอร์แมนท์ที่สองต่ำลง เสียงตอมีฟอร์แมนท์ที่สองเพิ่มขึ้น ส่วนเสียงกอมีฟอร์แมนท์ที่สองสูงที่สุด จากตารางที่ 4.13 ถึงแม้ว่าเสียงกอจะให้อัตราการรู้จำที่ร้อยละ 100 แต่พบว่าเสียงปอและเสียงตอต่างก็รู้จำผิดมาเป็นเสียงกอ ตรงกันข้ามกับเสียงจอที่มีฐานการเกิดที่เพดานแข็งและมีคุณลักษณะ

ของพลังงานที่ความถี่สูงอย่างชัดเจนจึงไม่มีการรู้จำผิดไปเป็นเสียงอื่น และเสียงอื่นในกลุ่มก็ไม่รู้จำผิดมาเป็นเสียงจอด้วยเช่นกัน ส่วนเสียงออที่เกิดจากฐานของช่องเส้นเสียงมีการรู้จำผิดพลาดเป็นเสียงปอและเสียงตอที่เกิดจากฐานปากและฐานปุ่มเหงือก

9) ขั้นตอนที่ 9 การแยกกลุ่มเสียงกัก-ไม่ก้อง-พ่นลม (Stop Voiceless Aspirated) พบว่าสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ให้อัตราการรู้จำสูงสุดที่ร้อยละ 85.00 ดังรูปที่ 4.10



รูปที่ 4.10 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 9

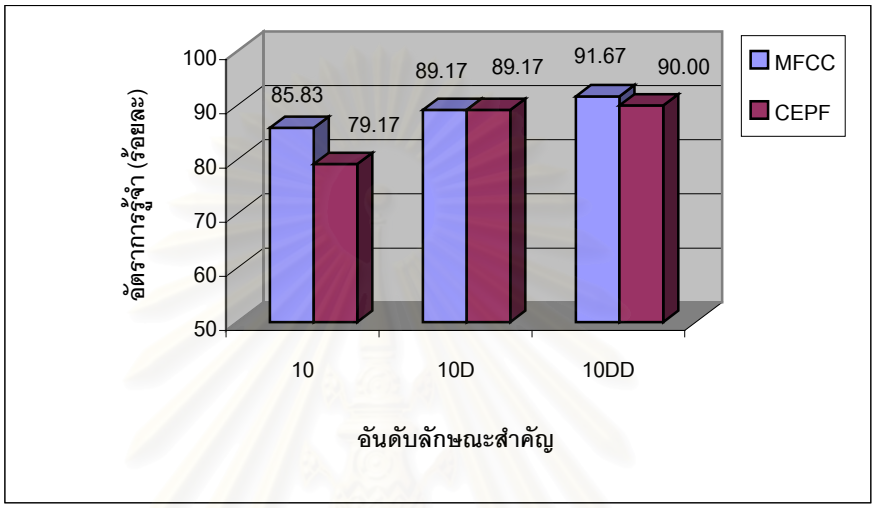
ตารางที่ 4.14 ผลการรู้จำในขั้นตอนที่ 9

	จำนวนข้อมูล (160)	รู้จำเป็นเสียงพอ /ph@@/	รู้จำเป็นเสียงทอ /th@@/	รู้จำเป็นเสียงชอ /ch@@/	รู้จำเป็นเสียงคอ /kh@@/
เสียงพอ /ph@@/	40	28 (70.00%)	12 (30.00%)	0	0
เสียงทอ /th@@/	40	6 (15.00%)	30 (75.00%)	1 (2.50%)	3 (7.50%)
เสียงชอ /ch@@/	40	0	0	40 (100.00%)	0
เสียงคอ /kh@@/	40	2 (5.00%)	0	0	38 (95.00%)
ขั้นตอนที่ 9 อัตราการรู้จำเฉลี่ยร้อยละ 85.00					

เสียงกัก-ไม่ก้อง-พ่นลมมีฐานการเกิดเสียงเช่นเดียวกับเสียงกัก-ไม่ก้อง-ไม่พ่นลม (ในขั้นตอนที่ 8) คือใช้ระดับสัมพัทธ์ของฟอร์แมนท์ที่สอง ทำให้เสียงพอที่มีฟอร์แมนท์ที่สองต่ำลง เสียงทอที่มีฟอร์แมนท์ที่สองเพิ่มขึ้น และเสียงคอซึ่งมีฟอร์แมนท์ที่สองสูงที่สุดนั้นต่างก็รู้จำผิด

พลาดในกลุ่มของมันเอง ส่วนเสียงซอที่เกิดที่เพดานแข็งมีคุณลักษณะของพลังงานที่ความถี่สูงเช่นเดียวกับเสียงจอทำให้มีอัตราการรู้จำสูงที่สุดคือร้อยละ 100 ดังตารางที่ 4.14

10) ขั้นตอนที่ 10 การแยกกลุ่มเสียงเสียดแทรก (Fricative Classifier) โดยสัมประสิทธิ์ MFCCDD อันดับ 10 ให้อัตราการรู้จำสูงที่สุดคือร้อยละ 91.67



รูปที่ 4.11 ผลการรู้จำด้วยลักษณะสำคัญ MFCC, MFCCD, MFCCDD, CEPF, CEPFD และ CEPFDD อันดับ 10 ในขั้นตอนที่ 10

ตารางที่ 4.15 ผลการรู้จำในขั้นตอนที่ 10

เสียง	จำนวนข้อมูล (120)	รู้จำเป็นเสียงฟอ /f@@/	รู้จำเป็นเสียงซอ /s@@/	รู้จำเป็นเสียงฮอ /h@@/
เสียงฟอ /f@@/	40	34 (85.00%)	6 (15.00%)	0
เสียงซอ /s@@/	40	3 (7.50%)	37 (92.50%)	0
เสียงฮอ /h@@/	40	1 (2.50%)	0	39 (97.50%)
ขั้นตอนที่ 10 อัตราการรู้จำเฉลี่ยร้อยละ 91.67				

เสียงฟอเกิดที่ฐานปากจะมีพลังงานค่อนข้างต่ำ (Weak Initial Frication) เมื่อเทียบกับเสียงซอและเสียงฮอ เสียงซอเกิดที่ฐานปุ่มเหงือกจะมีพลังงานส่วนใหญ่อยู่ที่ความถี่สูง และเสียงฮอเกิดที่เกิดขึ้นที่ช่องเส้นเสียงจะมีพลังงานสูงและพลังงานส่วนใหญ่อยู่ที่ความถี่ที่ต่ำกว่าเสียงซอ

จากผลการรู้จำด้วยขั้นตอนวิธีการฐานความรู้ทั้ง 10 ขั้นตอนนี้ เมื่อนำมาคำนวณหาอัตราการรู้จำรวมทั้งระบบด้วยแผนภูมิต้นไม้ (Decision Tree) จะได้อัตราการรู้รวมร้อยละ 79.64

ดังรูปที่ 4.12 ซึ่งแสดงผลของอัตราการรู้จำในแต่ละขั้นตอนของแผนภูมิต้นไม้ โดยเริ่มจากข้อมูลเสียงกลุ่มทดสอบทั้งหมด 560 เสียง ซึ่งประกอบด้วยกลุ่มเสียงก้อง 180 เสียง และกลุ่มเสียงไม่ก้อง 380 เสียง แต่สามารถรู้จำได้ถูกต้องเป็นกลุ่มเสียงก้อง 178 เสียง และกลุ่มเสียงไม่ก้อง 378 เสียง ดังนั้นในขั้นตอนถัดมา เสียงที่ทำการรู้จำจึงมีจำนวนเหลืออยู่ 556 เสียง แบ่งออกเป็น เสียงกัก-ก้อง 40 เสียง เสียงนาสิก 59 เสียง เสียงกึ่งสระ 79 เสียง กลุ่มเสียง Noise 278 เสียง และกลุ่มเสียง Non-noise อีก 100 เสียง

4.5 ขั้นตอนการแยกเสียงวรรณยุกต์

กรรมวิธีการแยกเสียงวรรณยุกต์ที่นำเสนอในงานวิจัยนี้สามารถแยกเสียงวรรณยุกต์ของคำเรียกพยัญชนะไทยในชุดทดสอบได้ถูกต้องร้อยละ 100 ดังตารางที่ 4.16

ตารางที่ 4.16 ผลการรู้จำเสียงวรรณยุกต์

กรรมวิธีในการรู้จำ	กลุ่มเสียง	จำนวนข้อมูล (560)	รู้จำเป็นเสียงสามัญ	รู้จำเป็นเสียงจัตวา
กรรมวิธีการตรวจสอบความถี่มูลฐาน	เสียงสามัญ	420	420 (100%)	0
	เสียงจัตวา	140	0	140 (100%)

วิธีการปรับปรุงระบบ

อัตราการรู้จำรวมทั้งระบบของสัมประสิทธิ์ MFCC10DD คือร้อยละ 79.64 จึงปรับปรุงระบบด้วย 2 วิธีคือ

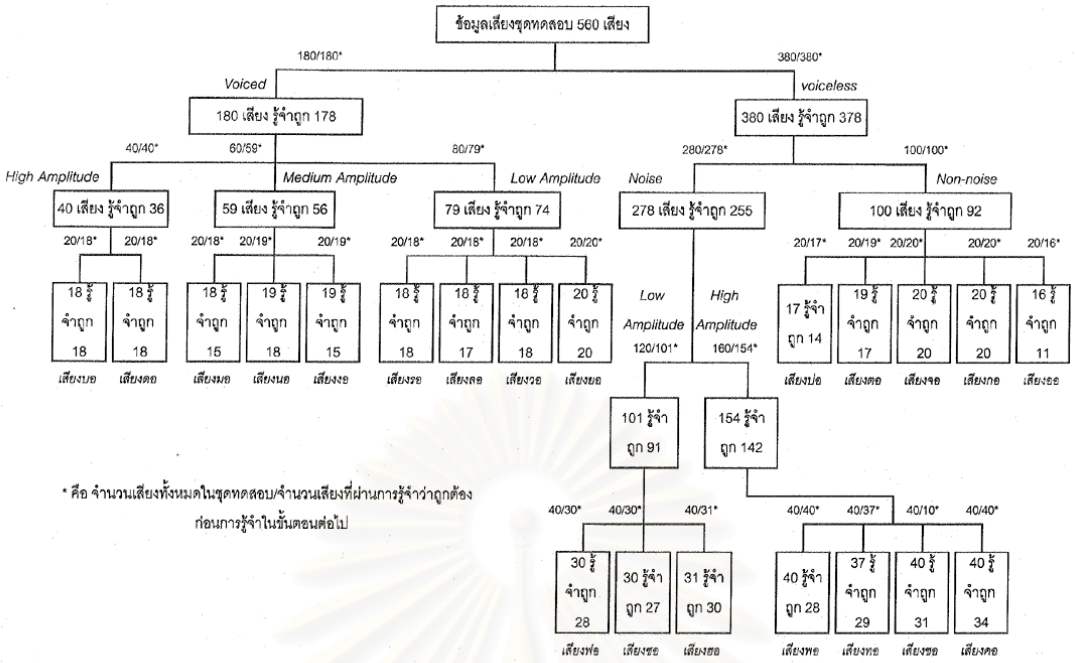
1) นำสัมประสิทธิ์ CEPFDD ที่อันดับ 10 มาใช้ร่วมกับสัมประสิทธิ์ MFCCDD ที่อันดับ 10 ทำให้อัตราการรู้จำเพิ่มขึ้นเป็นร้อยละ 81.43 โดยนำค่าความผิดพลาด (คำนวณได้จากขั้นตอนการตัดสินใจในขั้นตอนสุดท้าย) ที่น้อยที่สุด 2 อันดับมาหาค่าความผิดพลาดรวมแล้วจึงตัดสินใจเลือกคำตอบที่มีให้ค่าความผิดพลาดน้อยที่สุด

2) ทำการแยกเสียงวรรณยุกต์ก่อนขั้นตอนการแยกเสียงตามลักษณะการเกิดเสียงและขั้นตอนการแยกเสียงตามฐานเสียงดังรูปที่ 4.13 ได้อัตราการรู้จำเพิ่มจากร้อยละ 81.43 เป็นร้อยละ 83.75 เนื่องจากช่วยลดการรู้จำเสียงพยัญชนะจัตวาผิดพลาดในขั้นตอนที่ 1 และ 3

จากวิธีการปรับปรุงระบบทั้ง 2 วิธีข้างต้นจะแสดงผลได้ดังตารางที่ 4.17 และขั้นตอนวิธีการโดยรวมของระบบแสดงดังรูปที่ 4.13 ซึ่งประกอบด้วย การประมวลผลสัญญาณเบื้องต้น การกำหนดขอบเขตข้อมูล การกำจัดสัญญาณรบกวน การวัดค่าลักษณะสำคัญ ขั้นตอนวิธีการแยกเสียงวรรณยุกต์ ขั้นตอนวิธีการฐานความรู้ และขั้นตอนวิธีการตัดสินใจ

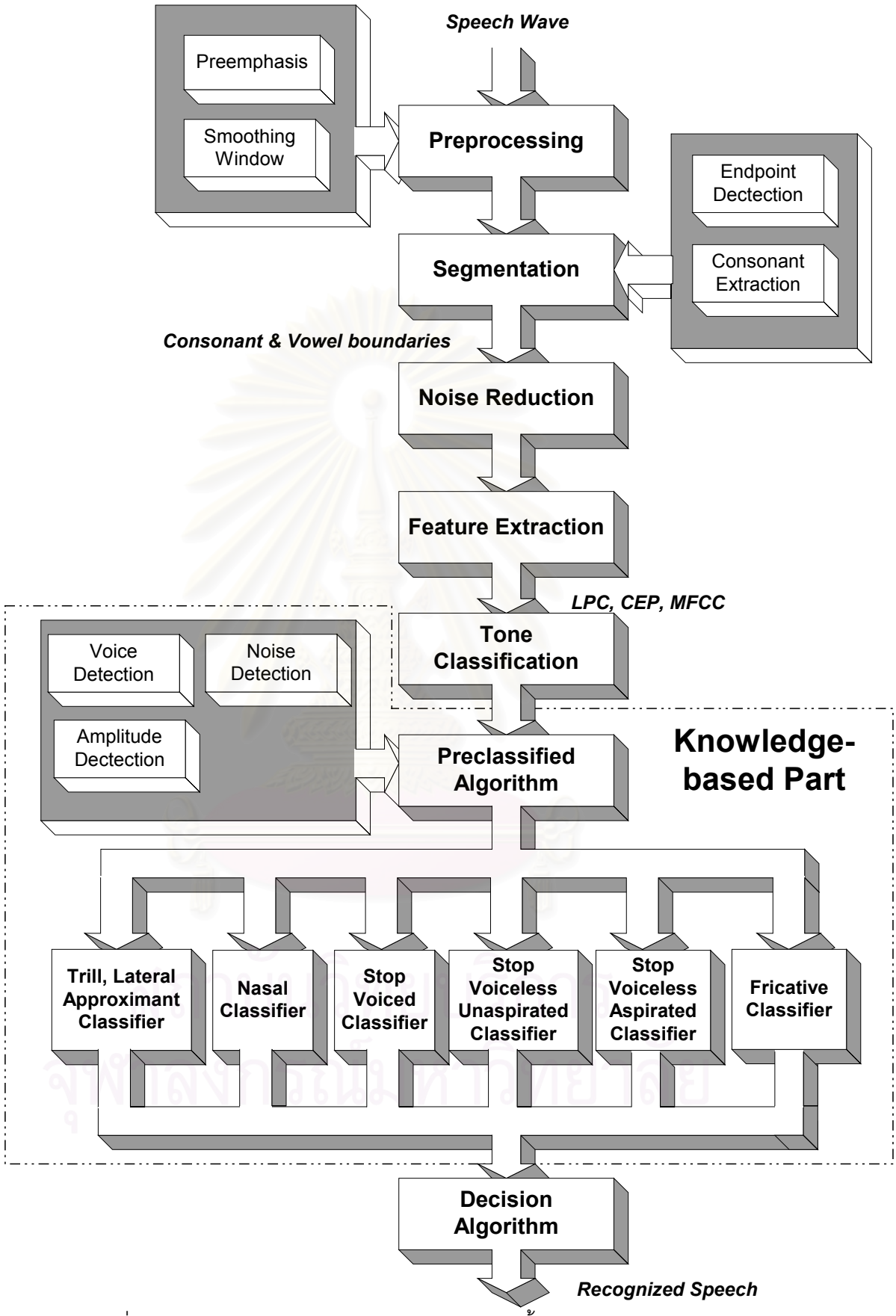
ตารางที่ 4.17 ผลการปรับปรุงการรู้จำด้วยสัมประสิทธิ์ MFCC และ CEPF แบบบวกความต่างที่อันดับ 10, 3 Gaussian Mixture

อัตราการรู้จำ (ร้อยละ)			
ขั้นตอนที่	**ชนิดของสัมประสิทธิ์**		
	สัมประสิทธิ์ MFCCDD อันดับ10	**แบบปรับปรุง**	
		ใช้ร่วมกับสัมประสิทธิ์ CEPFDD ที่อันดับ 10	**เสียงวรรณยุกต์**
ขั้นตอนที่ 1	99.29	99.29	99.64
ขั้นตอนที่ 2	92.78	93.89	93.89
ขั้นตอนที่ 3	91.58	91.58	94.74
ขั้นตอนที่ 4	91.78	91.78	91.78
ขั้นตอนที่ 5	100.00	100.00	100.00
ขั้นตอนที่ 6	88.33	93.33	93.33
ขั้นตอนที่ 7	95.00	95.00	95.00
ขั้นตอนที่ 8	88.00	88.00	88.00
ขั้นตอนที่ 9	85.00	88.12	88.12
ขั้นตอนที่ 10	91.67	91.67	91.67
อัตราการรู้จำรวม	79.64	81.43	83.75



* คือ จำนวนเสียงทั้งหมดในชุดทดสอบ/จำนวนเสียงที่ผ่านการรู้จำว่าถูกต้องก่อนการรู้จำในขั้นตอนต่อไป

รูปที่ 4.12 แผนภาพแสดงอัตราการรู้จำรวมของทั้งระบบ



รูปที่ 4.13 ระบบการรู้จำคำเรียกพยัญชนะไทยแบบขั้นตอนวิธีการฐานความรู้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

1. ขั้นตอนวิธีการฐานความรู้ที่นำเสนอในงานวิจัยนี้ ใช้การแยกเสียงออกเป็น 10 ขั้นตอน ตามลักษณะการเปล่งเสียงและตามฐานเสียง ให้อัตราการรู้จำสูงขึ้นร้อยละ 18.8 เนื่องจากช่วยลดปัญหาความกำกวมของเสียงที่เกิดขึ้นข้ามกลุ่มและช่วยทำให้ระบบแยกความแตกต่างได้มากขึ้น

2. การเปรียบเทียบลักษณะสำคัญ 5 ชนิดซึ่งประกอบด้วย สัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น สัมประสิทธิ์การประมาณพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัมที่คำนวณจากสัมประสิทธิ์การประมาณพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัมที่คำนวณจากการแปลงดีสครีตฟูริเยร์ และสัมประสิทธิ์เซปสตรัมบนความถี่เมล พบว่าสัมประสิทธิ์ที่คำนวณจากความถี่เชิงเส้นหรือใช้การประมาณพันธะเชิงเส้นไม่เหมาะสมกับระบบรู้จำคำเรียกพยัญชนะไทย เนื่องจากคุณสมบัติของเสียงพยัญชนะที่มีลักษณะคล้ายสัญญาณรบกวน เช่นเสียงเสียดแทรก และเสียงกัก-ไม่ก้อง-พ่นลม ไม่สามารถแทนได้ด้วยคุณสมบัติเชิงเส้น โดยสัมประสิทธิ์เซปสตรัมบนความถี่เชิงเส้น สัมประสิทธิ์การประมาณพันธะเชิงเส้น สัมประสิทธิ์เซปสตรัมที่คำนวณจากสัมประสิทธิ์การประมาณพันธะเชิงเส้น ให้อัตราการรู้จำน้อยกว่าสัมประสิทธิ์เซปสตรัมที่คำนวณจากการแปลงดีสครีตฟูริเยร์ และสัมประสิทธิ์เซปสตรัมบนความถี่เมลให้อัตราการรู้จำสูงที่สุดที่ร้อยละ 79.64 และเมื่อนำสัมประสิทธิ์ทั้งสองมาใช้ร่วมกันทำให้อัตราการรู้จำเพิ่มขึ้นเป็นร้อยละ 81.43

3. เนื่องจากเสียงพยัญชนะในกลุ่มของเสียงกัก-ไม่ก้อง-พ่นลมและเสียงเสียดแทรกที่มีลักษณะคล้ายสัญญาณรบกวน ดังนั้นการรบกวนของสัญญาณรบกวนในกลุ่มเสียงที่ไม่มีลักษณะคล้ายสัญญาณรบกวน จะทำให้เพิ่มความกำกวมและลดอัตราการรู้จำของเสียง จากการทดสอบพบว่าอัตราการรู้จำจะเพิ่มขึ้นร้อยละ 16 เมื่อกำจัดสัญญาณรบกวนออกไป

4. กรรมวิธีการหาขอบเขตหน่วยเสียงพยัญชนะทำให้สามารถกำจัดเสียงสระ ซึ่งไม่จำเป็นในการแยกความแตกต่างของเสียงพยัญชนะ ทำให้อัตราการรู้จำเพิ่มสูงขึ้นร้อยละ 11

5. การคำนวณค่าความแตกต่างระหว่างค่าสัมประสิทธิ์ด้วยวิธี Delta และวิธี Delta Different ทำให้อัตราการรู้จำเพิ่มขึ้น โดยสัมประสิทธิ์เซปสตรอลบนความถี่เมลแบบบวกความต่างเป็นลักษณะสำคัญที่ให้อัตราการรู้จำสูงที่สุด เนื่องจากเป็นการเพิ่มปริภูมิของความแตกต่างมากที่สุด

6. จำนวนของ Gaussian Mixture ที่เพิ่มขึ้นมีผลต่ออัตราการรู้จำของระบบ ทำให้อัตราการรู้จำเพิ่มสูงขึ้น แต่เวลาที่ใช้กับความซับซ้อนในการคำนวณก็สูงขึ้นด้วย ดังนั้นเมื่อเพิ่มจำนวน Gaussian Mixture จาก 1 เป็น 3 ทำให้อัตราการรู้จำเพิ่มขึ้นประมาณร้อยละ 3 เวลาที่ใช้เพิ่มสูงขึ้นประมาณ 12 นาที ในขณะที่เมื่อเพิ่มจำนวน Gaussian Mixture เป็น 5 เวลาที่ใช้เพิ่มสูงขึ้นประมาณ 20 ชั่วโมง และอัตราการรู้จำที่ได้ยังมีค่าลดลงเท่ากับจำนวน Gaussian Mixture ที่ 1 เนื่องจากข้อมูลเสียงมีการกระจายตัวไม่มากนัก จำนวน Gaussian Mixture เท่ากับ 3 ก็สามารถครอบคลุมข้อมูลทั้งหมดได้อย่างพอดี

5.2 ข้อเสนอแนะ

1. คุณภาพของเสียงมีผลต่ออัตราการรู้จำ ควรปรับปรุงวิธีการกำจัดสัญญาณรบกวนหรือใช้ลักษณะสำคัญที่ทนทานต่อสัญญาณรบกวน

2. การขาดหายไปของความถี่มูลฐานที่เกิดขึ้นเมื่อค่าพลังงานมีค่าต่ำที่บริเวณต้นพยางค์และท้ายพยางค์ ถ้าสามารถปรับปรุงขั้นตอนวิธีการหาความถี่มูลฐานจากค่าพลังงานที่มีแอมพลิจูดต่ำได้ จะทำให้ได้ค่าความถี่มูลฐานที่ครบถ้วนและช่วยลดความกำกวมระหว่างเสียงวรรณยุกต์และ เสียงก้องกับไม่ก้องได้ถูกต้องเพิ่มขึ้น

3. ความกำกวมที่เกิดขึ้นระหว่างเสียงสามัญของผู้พูดเพศหญิงกับเสียงจัตวาของเสียงผู้พูดเพศผู้ชาย ที่เกิดจากการเน้นเสียงที่ท้ายพยางค์ของผู้พูดเพศหญิงจะทำให้ความถี่มูลฐานมีลักษณะเพิ่มสูงขึ้น คล้ายกับการเพิ่มสูงขึ้นของความถี่มูลฐานในเสียงสามัญสำหรับผู้พูดเพศชาย ดังนั้นการแยกเสียงออกเป็นเพศชายและหญิง ด้วยค่าความถี่ (Charnvivit et al, 2000) ก่อนการรู้จำเสียงวรรณยุกต์จะทำให้ช่วยลดความกำกวมที่เกิดขึ้นได้ เนื่องจากเสียงของผู้พูดเพศหญิงจะมีความถี่ที่สูงกว่าเสียงของผู้พูดเพศชาย

4. ควรปรับปรุงขั้นตอนวิธีการตัดสินใจของระบบ เนื่องจากขั้นตอนวิธีการฐานความรู้ที่นำเสนอในงานวิจัยนี้มีลักษณะการตัดสินใจแบบแผนภูมิต้นไม้ (Decision Tree) โดยแบ่งออกได้เป็น 10 ขั้นตอน ดังนั้นในแต่ละขั้นตอนจะมีการสะสมความผิดพลาดที่เกิดจากการรู้จำผิดในขั้นตอนก่อนหน้า ทำให้ผลของอัตราการรู้จำรวมทั้งระบบมีค่าน้อยกว่าอัตราการรู้จำที่รู้จำได้ในขั้นตอนแรกๆ ถ้าเปลี่ยนวิธีการตัดสินใจของระบบโดยเลือกใช้กรรมวิธีที่มีการตัดสินใจจากแต่ละขั้นตอนแบบขนานกันไปเช่น กรรมวิธี Voting Classification และกรรมวิธี Bagging Prediction ที่คำนวณหาค่าความผิดพลาดน้อยที่สุดให้กับลักษณะสำคัญหลายๆ แบบอย่างขนานกันไป (Bauer and Kohavi, 1999)

รายการอ้างอิง

ภาษาไทย

ชัย วุฒิวิวัฒน์ชัย, การรู้จำเสียงพูดคำไทยหลายพยางค์แบบไม่ขึ้นต่อผู้พูดโดยใช้เทคนิคแบบฟัซซีและนิวรอลเน็ตเวอร์ค. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2540.

ไชยันต์ สุวรรณชีวะศิริ, การรู้จำเสียงพูดตัวเลขภาษาไทยแบบหลายผู้พูดด้วยนิวรอลเน็ตเวอร์ค. เอกสารรวมเล่มการประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 21, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 2541.

ณัฏฐชา จิตติวรางกูล, กรรมวิธีการหาขอบเขตพยางค์สำหรับคำพูดต่อเนื่องภาษาไทย. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541.

ธีระ ภัทราพรนันท์, การรู้จำเสียงพูดสระภาษาไทยโดดๆ ไม่ขึ้นกับผู้พูด โดยการวัดสเปกตรัมดิสแตนซ์และใช้ไดนามิกไทม์วาร์ปปิง. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2538.

ระพีพัฒน์ เพ็ญศิริ, การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นกับผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิง. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2538.

วิศรุต อาขุบุตร, ระบบรู้จำคำไทยหลายพยางค์แบบไม่ขึ้นกับผู้พูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2539.

วุฒิพงษ์ พรสุขจันทรา, การรู้จำเสียงตัวเลขภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยใช้แอลพีซีและโครงข่ายประสาทเทียมแบบแบ็กพรอพาเกชัน. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2539.

สุดาพร ลักษณียนาวิน, สัทศาสตร์และภาษาศาสตร์. กรุงเทพมหานคร: ห้างหุ้นส่วนจำกัด เทคเพรสเซอร์วิส จำกัด, 2529.

เสรี ปานซาง, การรู้จำเสียงพูดคำไทยเฉพาะบุคคลด้วยนิวรอลเน็ตเวอร์ค. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, 2540.

รายการอ้างอิง (ต่อ)

เสาวลักษณ์ อารีพงศา, การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดน มาร์คอฟโมเดลและเวกเตอร์ควอนไตซ์เซชั่น. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2538.

เอกฤทธิ์ มณีน้อย, การรู้จำหน่วยเสียงภาษาไทยโดยใช้โครงข่ายประสาทเทียม. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, 2541.

ภาษาอังกฤษ

Davis, S.B. and Mermelstein P., “Comparison of Parametric Representations for MonoSyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Transaction on Acoustics, Speech and Signal Processing 28, 4 (August 1980): 357-366.

Claes, T., et al., “A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition”, IEEE Transactions on Speech and Audio Processing 6, 6 (1998): 549-557.

Charnvivit, P., et al., “F0 Feature Extraction by Polynomial Regression Function for Monosyllabic Thai Tone Recognition”, Proceedings of the 7th European Conference on Speech Communication and Technology 4 (2001): 2753-2756.

Deller, J.R., et al., Discrete-Time Proceessing of Speech Signals. Arizona: Mcmillan Publishhing Company, 1993.

Fanty, M, et al., “City Name Recognition over the Telephone”, IEEE Transaction on Acoustics, Speech, and Signal Processing 1 (1993): 549-552.

Furui, S., Digital Speech Processing, Synthesis, and Recognition. New York: Marcel Dekker, 1989.

Itakura, F. “Minimum Prediction Residual Applied to Speech Recognition”, IEEE Transactions on Acoutics, Speech and Signal Processing 23, 1 (1975): 67-72.

Jun, W., et al., “Stochastic Language Models for Chinese Speech Recognition Based on Chinese Spelling”, IEEE Proceedings on Speech, Image Processing and Neural Network (1994): 674-677.

รายการอ้างอิง (ต่อ)

Junqua, J.C., "SmarTspelL™: A Multipass Recognition System for Name Retrieval Over the Telephone", IEEE Transactions on Speech and Audio Processing 5, 2 (March 1997): 173-182.

Lee, T., et al., "Tone Recognition of isolated Cantonese Syllables", IEEE Transactions on Speech and Audio Processing 3, 3 (1995): 204-209.

Liberman, M. A., Speech: A Special Code. Massachusetts: MIT Press, 1995.

Loizou, P.C., "Robust Speaker-Independent Recognition of a Confusable Vocabulary", Doctoral Dissertation, Arizona State University, 1995.

Loizou, P.C. and Spanias, A., "High-Performance Alphabet Recognition", IEEE Transactions on speech and Audio Processing 4, 6 (November 1996): 430-445.

O'Shaughnessy, D., "Linear Predictive Coding", IEEE Potentials (February 1988): 29-32.

Picone, J., Fundamentals of Speech Recognition: A Short Course. Institute for Signal and Information Processing Department of Electrical and Computer Engineering, Mississippi State University (1996).

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

Rigazio, L., et al., "Multilevel Discriminative Training for Spelled Word Recognition", IEEE Proceedings on Acoustics, Speech and Signal Processing 1 (1998): 489-492.

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

Tuzun, O.B., et al., "Comparison of Parametric and Non-Parametric Representations of Speech for Recognition", Proceedings of the Electrotechnical Conference 7th Mediterranean 1 (1994): 65-68.

Tolba, H. and O'Shaughnessy, D., "Automatic Speech Recognition Based on Cepstral Coefficients and a Mel-Based Discrete Energy Operator", IEEE Proceedings on Acoustic, Speech and Signal Processing 2 (1998): 973-976.

Tungthangthum, A., "Tone Recognition for Thai", Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Chiang Mai Thailand (November 1998): 157-160.

ภาคผนวก

ภาคผนวก ก

ข้อมูลผู้บอกภาษา

ข้อมูลเสียงพูดกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ ส่วนข้อมูลเสียงพูดกลุ่มที่ 2 และ กลุ่มที่ 3 เป็นข้อมูลชุดฝึกฝน

กลุ่มที่ 1

ชาย (11)	อายุ (ปี)	หญิง (9)	อายุ (ปี)
Adisak (ADS)	28	Akkrapat (AKT)	22
Akkrapol (AKL)	22	Dayennapa (AMP)	26
Boonchoung (BCG)	25	Chulaluk (CLL)	20
Boonchai (BOC)	21	Chantima (CTM)	21
Chanan (CHN)	19	Nattaree (EUY)	23
Udom (DAD)	60	Hatairat (HTR)	21
Panachit (GNG)	22	Umavasee (JAN)	22
Jamorn (JAM)	21	Jirakanya (JRK)	23
Ekkarit (JOK)	24	Metavee (KAY)	20
Suppachet (JOO)	26		
Jadsada (JSD)	23		

กลุ่มที่ 2

ชาย (11)	อายุ (ปี)	หญิง (9)	อายุ (ปี)
Suppakiat (KIA)	24	Kritika (KTK)	27
Kumphon (KPN)	24	Kantip (KTP)	22
Kittiphong (KTI)	23	Kittiya (KTY)	22
Visarut (MMM)	28	Latchana (LCN)	21
Nattapon (NPO)	19	Souvanee (MOM)	59
Navapat (NVP)	21	Nattaporn (NTP)	22
Denpong (ORM)	24	Patcharee (PCR)	22

ชาย (11)	อายุ (ปี)	หญิง (9)	อายุ (ปี)
Pongsatorn (PST)	23	Pimjana (PJN)	22
Sawit (SAW)	24	Painporn (PPR)	22
Supattarachai (SPC)	21		
Supakiat (SPK)	22		

กลุ่มที่3

ชาย (11)	อายุ (ปี)	หญิง (9)	อายุ (ปี)
Sirichai (SRC)	21	Proesaya (PSY)	16
Suchai (SUC)	23	Piyawan (PYW)	31
Sukree (SUK)	21	Rassamee (RSM)	19
Suwit (SUW)	23	Sudarat (SDR)	20
Theeraphong (TOO)	38	Suthasenee (STN)	18
Urat (URT)	34	Siriluk (TIP)	22
Vorawut (VRW)	20	Tilya (TLY)	19
Rathaphon (WAT)	32	Thanyaporn (TYP)	22
Wllard (WIL)	23	Weena (WNA)	23
Vorawit (WIT)	22		
Warathorn (WRT)	22		

ภาคผนวก ข

ตารางที่ ข.1 รายละเอียดผลการรู้จำทั้ง 10 ขั้นตอนของสัมประสิทธิ์ MFCC แบบบวกความต่างที่อันดับ 10, 3 Gaussian Mixture

ขั้นตอนที่ 1

กลุ่มเสียง	จำนวนข้อมูล	เสียงก้อง	เสียงไม่ก้อง	%
เสียงก้อง	180	178	2	98.88
เสียงไม่ก้อง	380	2	378	99.47
Total	560			99.29

ขั้นตอนที่ 2

กลุ่มเสียง	จำนวนข้อมูล	เสียงกัก	เสียงนาสิก	เสียงกึ่งสระ	%
เสียงกัก	40	36	4	0	90.00
เสียงนาสิก	60	2	56	2	93.34
เสียงกึ่งสระ	80	1	4	75	93.75
Total	180				92.78

ขั้นตอนที่ 3

กลุ่มเสียง	จำนวนข้อมูล	Noise	Non-Noise	%
Noise	280	256	24	91.43
Non-Noise	100	8	92	92.00
Total	380			91.58

ขั้นตอนที่ 4

กลุ่มเสียง	จำนวนข้อมูล	เสียงกัก	เสียงเสียดแทรก	%
เสียงกัก	160	148	12	92.50
เสียงเสียดแทรก	120	11	109	90.84
Total	280			91.78

ขั้นตอนที่ 5

เสียง	จำนวนข้อมูล	/b@@0/	/d@@0/	%
/b@@0/	20	20	0	100.00
/d@@0/	20	0	20	100.00
Total	40			100.00

ขั้นตอนที่ 6

เสียง	จำนวนข้อมูล	/m@@0/	/n@@0/	/ng@@0/	%
/m@@0	20	17	2	1	85.00
/n@@0/	20	0	20	0	100.00
/ng@@0/	20	0	4	16	80.00
Total	60				88.33

ขั้นตอนที่ 7

เสียง	จำนวนข้อมูล	/w@@0/	/r@@0/	/l@@0/	/j@@0/	%
/w@@0/	20	19	1	0	0	95.00
/r@@0/	20	1	19	0	0	95.00
/l@@0	20	0	2	18	0	90.00
/j@@0/	20	0	0	0	20	100.00
Total	80					95.00

ขั้นตอนที่ 8

เสียง	จำนวนข้อมูล	/p@@0/	/t@@0/	/c@@0/	/k@@0/	/@@0/	%
/p@@0/	20	16	3	0	1	0	80.00
/t@@0/	20	0	18	0	2	0	90.00
/c@@0/	20	0	0	20	0	0	100.00
/k@@0/	20	0	0	0	20	0	100.00
/@@0/	20	1	5	0	0	14	70.00
Total	100						88.00

ขั้นตอนที่ 9

เสียง	จำนวนข้อมูล	/ph@@/	/th@@/	/ch@@/	/kh@@/	%
/ph@@/	40	28	12	0	0	70.00
/th@@/	40	6	30	1	3	75.00
/ch@@/	40	0	0	40	0	100.00
/kh@@/	40	2	0	0	38	95.00
Total	160					85.00

ขั้นตอนที่ 10

เสียง	จำนวนข้อมูล	/f@@/	/s@@/	/h@@/	%
/f@@/	40	34	6	0	85.00
/s@@/	40	3	37	0	92.50
/h@@/	40	1	0	39	97.50
Total	120				91.67

ตารางที่ ข.2 รายละเอียดผลการรู้จำทั้ง 10 ขั้นตอนของสัมประสิทธิ์ MFCC แบบบวกความต่างที่อันดับ 10, 3 Gaussian Mixture แบบปรับปรุง

ขั้นตอนที่1

กลุ่มเสียง	จำนวนข้อมูล	เสียงก้อง	เสียงไม่ก้อง	%
เสียงก้อง	180	178	2	98.89
เสียงไม่ก้อง	380	0	380	100.00
Total	560			99.64

ขั้นตอนที่ 2

กลุ่มเสียง	จำนวนข้อมูล	เสียงกัก	เสียงนาสิก	เสียงกึ่งสระ	%
เสียงกัก	40	36	4	0	90.00
เสียงนาสิก	60	1	58	1	96.67
เสียงกึ่งสระ	80	1	4	75	93.75
Total	180				93.89

ขั้นตอนที่ 3

กลุ่มเสียง	จำนวนข้อมูล	Noise	Non-Noise	%
Noise	280	268	12	95.71
Non-Noise	100	8	92	92.00
Total	380			94.74

ขั้นตอนที่ 4

กลุ่มเสียง	จำนวนข้อมูล	เสียงกัก	เสียงเสียดแทรก	%
เสียงกัก	160	148	12	92.50
เสียงเสียดแทรก	120	11	109	90.83
Total	280			91.78

ขั้นตอนที่ 5

เสียง	จำนวนข้อมูล	/b@@0/	/d@@0/	%
/b@@0/	20	20	0	100
/d@@0/	20	0	20	100
Total	40			100

ขั้นตอนที่ 6

เสียง	จำนวนข้อมูล	/m@@0/	/n@@0/	/ng@@0/	%
/m@@0/	20	19	1	0	95.00
/n@@0/	20	0	20	0	100.00
/ng@@0/	20	0	3	17	85.00
Total	60				93.33

ขั้นตอนที่ 7

เสียง	จำนวนข้อมูล	/w@@0/	/r@@0/	/l@@0/	/j@@0/	%
/w@@0/	20	18	0	2	0	90.00
/r@@0/	20	0	19	1	0	95.00
/l@@0/	20	0	1	19	0	95.00
/j@@0/	20	0	0	0	20	100.00
Total	80					95.00

ขั้นตอนที่ 8

เสียง	จำนวนข้อมูล	/p@@0/	/t@@0/	/c@@0/	/k@@0/	/@@0/	%
/p@@0/	20	16	3	0	1	0	80.00
/t@@0/	20	0	18	0	2	1	90.00
/c@@0/	20	0	0	20	0	0	100.00
/k@@0/	20	0	0	0	20	0	100.00
/@@0/	20	1	5	0	0	14	70.00
Total	100						88.00

ขั้นตอนที่ 9

เสียง	จำนวนข้อมูล	/ph@@/	/th@@/	/ch@@/	/kh@@/	%
/ph@@/	40	33	7	0	0	82.50
/th@@/	40	5	30	1	4	75.00
/ch@@/	40	0	0	40	0	100.00
/kh@@/	40	2	0	0	38	95.00
Total	160					88.12

ขั้นตอนที่ 10

เสียง	จำนวนข้อมูล	/f@@/	/s@@/	/h@@/	%
/f@@/	40	34	6	0	85.00
/s@@/	40	3	37	0	92.50
/h@@/	40	1	0	39	97.50
Total	120				91.67

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอุมาวสี ทาทอง เกิดเมื่อวันที่ 1 มกราคม พ.ศ.2519 ที่จังหวัด อุบลราชธานี สำเร็จการศึกษาปริญญาตรี หลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมไฟฟ้า มหาวิทยาลัยเกษตรศาสตร์ ในปีการศึกษา 2539 และเข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อ พ.ศ. 2540