การเปรียบเทียบวิธีการหาลักษณะเด่นของเสียงสำหรับการรู้จำคำไทย

A COMPARISON OF FEATURE EXTRACTION METHODS FOR THAI WORD RECOGNITION

เกรียงศักดิ์ เตมีย์
KREANGSAK TAMEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2544

ISBN 974-648-357-9

การเปรียบเทียบวิธีการหาลักษณะเด่นของเสียงสำหรับการรู้จำคำไทย

A COMPARISON OF FEATURE EXTRACTION METHODS FOR THAI WORD RECOGNITION

เกรียงศักดิ์ เตมีย์

KREANGSAK TAMEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2544

ISBN 974-648-357-9

A COMPARISON OF FEATURE EXTRACTION METHODS FOR THAI WORD RECOGNITION

KREANGSAK TAMEE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2001

ISBN 974-648-357-9

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การเปรียบเทียบวิธีการหาลักษณะเด่นของเสียงสำหรับการรู้จำคำไทย
A COMPARISON OF FEATURE EXTRACTION METHODS FOR THAI WORD RECOGNITION

ชื่อนักศึกษา นายเกรียงศักดิ์ เตมีย์

รหัสประจำตัว 41061179

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชา วิศวกรรมไฟฟ้า

อาจารย์ผู้ควบคุมวิทยานิพนธ์ รศ.ดร.ชม กิ้มปาน

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ดร.วิศิษฏ์	หิรัญกิตติ	
รศ.ดร.บุญธีร์	เครือตราชู	
รศ.ประทีป	บัญญัตินพรัตน์	
รศ.ดร.ชม	กิ้มปาน	

วัน/เดือนปี ที่สอบ 23 พฤษภาคม 2544 เวลา 12.00 – 13.00 น.

สถานที่สอบ ณ อาคาร 12 ชั้น ชั้น 4 (ห้อง E12-403)

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.บุญวัฒน์ อัตชู)

คณบดีบัณฑิตวิทยาลัย

วันที่ 23 **เดือน** ..... **พ.ศ.** 2544

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการหาลักษณะเด่นของเสียงสำหรับการรู้จำคำไทย
นักศึกษา	นายเกรียงศักดิ์ เตมีย์
รหัสประจำตัว	41061179
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
พ.ศ.	2544
อาจารย์ผู้ควบคุม	รศ.ดร.ชม กิ้มปาน

บทคัดย่อ

วิทยานิพนธ์นี้มีวัตถุประสงค์ในการศึกษา และเปรียบเทียบประสิทธิภาพ วิธีดึงลักษณะที่สำคัญของเสียงพูด หรือเรียกว่าวิธีการประมวลผลเบื้องต้น ระหว่างวิธี LPCC MFCC และ PLP โดยดูจากผลการทดสอบกับระบบการรูจำเสียงภาษาไทยแบบคำโดดไม่ขึ้นอยู่กับผู้พูด ใช้แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Models) ทำการฝึกสอน และรู้จำ ระบบรู้จำได้ทั้งหมด 20 คำ คือ คำพูดที่เป็นเลขศูนย์ถึงเก้า และคำที่ใช้ในการควบคุมอุปกรณ์ เสียงพูดที่ใช้ในการทดลองได้จากผู้พูดเพศชาย อายุระหว่าง 20-30 ปี แบ่งเป็น 2 กลุ่ม คือ กลุ่มที่ใช้ฝึกสอน และกลุ่มที่ใช้ทดสอบ กลุ่มละ 15 คน และ 5 คน ตามลำดับ มีทั้งเสียงที่ปราศจากสัญญาณรบกวน และมีสัญญาณรบกวนแบบเกาส์เซียน แบบเสียงพูดแทรก และแบบเสียงรถ ที่ระดับ –45dB -40dB -35dB และ -30dB จากการทดสอบพบว่าระบบมีประสิทธิภาพในการรู้จำสูงพอๆ กัน ทั้ง 3 วิธี เมื่อทดสอบกับเสียงที่ปราศจากสัญญาณรบกวน แต่เมื่อทดสอบกับเสียงที่มีสัญญาณรบกวนแบบต่างๆ ประสิทธิภาพในการรู้จำจะลดลง ขึ้นอยู่กับระดับของสัญญาณรบกวน อย่างไรก็ตามการประมวลผลเบื้องต้นแบบ PLP มีประสิทธิภาพสูงกว่าการประมวลผลแบบอื่นๆ โดยเฉพาะอย่างยิ่งเมื่อทดสอบกับเสียงที่มีสัญญาณรบกวน

Thesis Title	A Comparison of Signal Processing Front Ends for Thai Word Recognition
Student	Mr.Kreangsak Tamee
Student ID.	41061179
Degree	Master of Engineering
Programme	Electrical Engineering
Year	2001
Thesis Advisor	Assoc. Prof. Dr.Chom Kimpan

ABSTRACT

The main purpose of this thesis is to study and comparing the performance of feature extraction as signal processing front ends between LPCC, MFCC and PLP .Using these methods with independent speaker Thai isolated word speech recognition trained by Hidden Markov Models. The speech words used in this experiment are the words for number one to nine in Thai and keywords used for controling electrical equipment, for the total of 20 kinds of speech words. Those speech words are obtained from male speakers, where ages are between 20-30 years old, divided into 2 groups. The first group has 15 members, used for training the system. The other group has 5 member, used for testing the system. We recorded clean speech, Gaussian noise speech, babble noise speech and car noise speech at –45dB, –40dB, -35dB and –30dB.

From the experiment. We see that the system performs well for all three methods of processing front ends. But when applying to noised speech, PLP method give the best result.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยคำแนะนำและคำปรึกษาเกี่ยวกับระบบการรู้จำเสียงพูดภาษาไทย รวมทั้งการประเมินผลงานจาก รศ.ดร.ชม กิ้มปาน ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่านและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ อ.รัชฎา คงคะจันทร์ จากมหาวิทยาลัยธรรมศาสตร์ ที่ช่วยเหลือแก้ไขและให้คำแนะนำในบางจุดที่ผู้วิจัยติดปัญหาบางอย่าง ซึ่งมีส่วนช่วยทำให้ผู้วิจัยเข้าใจในปัญหานั้น

ขอขอบพระคุณอาจารย์ผู้ควบคุมการสอบวิทยานิพนธ์ ที่ให้คำแนะนำทำให้วิทยานิพนธ์นี้ความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ คุณพ่อ คุณแม่ และญาติๆ ที่ได้ให้ทุนสนับสนุนการศึกษา และให้กำลังใจมาโดยตลอด

ขอขอบคุณเพื่อนๆ นักศึกษาทุกคน ที่ช่วยเหลือด้านข้อมูลตัวอย่างเสียง และให้คำแนะนำต่างๆ อย่างใกล้ชิด

คุณค่าและประโยชน์อันพึงจะมีจากวิทยานิพนธ์นี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

เกรียงศักดิ์ เตมีย์

สารบัญ

สารบัญ(ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

จากงานวิจัยที่ผ่านมาหลายปี แสดงให้เห็นว่าระบบการรู้จำเสียง เป็นเทคโนโลยีที่มีความจำเป็นในชีวิตประจำวันของเราในอนาคตอันใกล้นี้ ในปัจจุบันนี้การรู้จำเสียงพูดมีความก้าวหน้าไปมาก และเป็นรูปเป็นร่างมากขึ้นในการประยุกต์ใช้งานเพื่ออำนวยความสะดวกสบาย เช่น การกดหมายเลขโทรศัพท์ด้วยเสียง การสั่งงานคอมพิวเตอร์ด้วยเสียง และนอกจากนั้นยังเป็นประโยชน์อย่างยิ่งในระบบรักษาความปลอดภัย

ระบบการรู้จำส่วนใหญ่จะมีโมเดลตามรูปที่ 1.1 คือ Y จะเป็นพารามิเตอร์ที่แสดงคุณลักษณะเด่นของข้อมูลเสียงพูด X โดยผ่านการดึงลักษณะเด่นของข้อมูล (Feature Extraction) หรือเรียกว่าการประมวลผลเบื้องต้น (Front Ends) แล้วนำเข้าไปทำการจัดแบ่งกลุ่ม (Classifier) ของคำแต่ละคำ W

รูปที่ 1.1 โมเดลของระบบการรู้จำเสียง

การรู้จำเสียงพูดนั้นมีการศึกษากันอย่างแพร่หลาย แนวทางในการศึกษาวิจัยก็มีหลากหลายว่าจะเน้นในด้านใดเนื่องจากกระบวนการรู้จำเสียงพูด ต้องการขั้นตอนและวิธีที่แตกต่างกันไป ทั้งนี้อาจแบ่งออกเป็นแนวทางหลักๆ ได้เป็นการศึกษาวิจัยที่เน้น กลุ่มผู้พูด ลักษณะการพูด วิธีการฝึกสอนและรู้จำ วิธีและเทคนิคเสริมอื่นๆ [4] เป็นต้น

<u>การวิจัยที่เน้นตามกลุ่มผู้พูด</u> แบ่งย่อยได้เป็น

1. แบบขึ้นกับผู้พูด (Speaker-Dependent)
2. แบบไม่ขึ้นกับผู้พูด (Speaker-Independent)
3. แบบตรวจรู้ผู้พูด (Speaker Identification)

งานวิจัยส่วนใหญ่เน้นที่แบบไม่ขึ้นกับผู้พูด เพราะสามารถนำไปประยุกต์ใช้งานได้กว้างขวางกว่าแบบอื่น อีกทั้งลักษณะสำคัญที่ต้องใช้จะไม่มีข้อกำหนดที่เข้มงวด ต่างจาก 2 แบบหลัง อย่างไรก็ตาม การศึกษาวิจัยในระยะแรก เริ่มจากแบบขึ้นกับผู้พูด เพราะสามารถควบคุมความแปรปรวนของกลุ่มตัวอย่างได้ดีกว่า อีกทั้งผลิตภัณฑ์ทางด้านการรู้จำเสียงพูดที่เผยแพร่ในระยะ

แรกๆ จนถึงปัจจุบัน ก็เป็นแบบขึ้นกับผู้พูดหรืออยู่ในแนวทางดังกล่าว เพราะต้องมีการฝึกฝนผู้ที่จะนำผลิตภัณฑ์เหล่านั้นไปใช้งานเป็นระยะเวลาหนึ่ง ในขณะที่แทบไม่ปรากฏงานวิจัยแบบตรวจรู้ผู้พูด เพราะเรายังไม่สามารถสร้างแบบจำลองกลไกการพูดของมนุษย์ที่สมบูรณ์ ที่สำคัญการหาลักษณะสำคัญเฉพาะในเสียงพูดของแต่ละบุคคล เป็นเรื่องที่ยังทำไม่ได้ด้วยความรู้ที่มีในปัจจุบัน

<u>การศึกษาวิจัยที่เน้นตามลักษณะการพูด</u> สามารถจำแนกตามเสียงพูดได้เป็น

1. แบบคำโดด (Isolated Word) เป็นการศึกษาวิจัยที่เริ่มก่อนแบบอื่น เพราะรูปแบบของสัญญาณง่ายต่อการวิเคราะห์และวัดหาลักษณะสำคัญโดยรวมที่สุด มีงานวิจัยจำนวนมากที่เดินไปในแนวทางนี้แม้จนถึงปัจจุบัน เป็นต้นว่า งานวิจัยของ วุฒิพงษ์ พรสุขจันทรา [6] และเสาวลักษณ์ อารีย์พงศา [5]

2. แบบคำติดกัน (Connected Word) หรือคำหลายพยางค์ (Polysyllabic Word) เป็นแบบที่มีการศึกษาวิจัยในลำดับต่อจากแบบแรก เพื่อพยายามก้าวไปสู่แบบที่สาม ดังตัวอย่างงานวิจัยของ Rebiner L.R. and Schmidt C.E. [12]

3. แบบเสียงพูดต่อเนื่อง (Continuous Speech) เป็นแบบที่ให้ความสนใจทำการศึกษาวิจัยกันมากในปัจจุบัน ดังเช่นงานวิจัยของ Chen S.H. and Wang Y.R. [8]

<u>การศึกษาวิจัยที่เน้นตามลักษณะของเสียงพูด</u> เป็นการวิจัยที่มีการกำหนดขอบเขตหรือเป้าหมายของงานวิจัยหรือประโยชน์ที่จะนำไปประยุกต์ แบ่งเป็น

1. แบบจำกัดเฉพาะเสียงตัวเลข
2. แบบเสียงคำสั่งเฉพาะงาน ที่เป็นคำๆ หรือคำหลายพยางค์ มีจำนวนคำไม่มาก
3. แบบเสียงคำทั่วไปที่มีคำศัพท์จำนวนมากพอประมาณ
4. แบบเสียงคำทั่วไปที่มีคำศัพท์จำนวนมาก

<u>การศึกษาวิจัยที่เน้นตามการฝึกสอนและรู้จำ</u> ที่แบ่งได้เป็น 3 วิธี คือ

1. การเข้าคู่ต้นแบบ (Template matching)

เทคนิคที่อยู่ในแนวทางกรรมวิธีนี้และเป็นที่นิยมกันมากคือ Dynamic time Wraping (DTW) ซึ่งมีกรรมวิธีรุ่นแรกๆ ที่นำมาวิจัยกัน มีความง่ายต่อการพัฒนา การฝึกฝนใช้เวลาน้อย แต่เวลาในการรู้จำขึ้นกับจำนวนแบบอ้างอิงและสมรรถนะของการรู้จำจะต่ำกว่ากรรมวิธีอื่นๆ และลดลง ถ้าเพิ่มจำนวนคำศัพท์มากขึ้น ตัวอย่างงานวิจัยของ ระพีพัฒน์ อารีย์พงศา [3]

2. แบบจำลองฮิดเดนมาร์คอฟ (Hidden Morkov Model : HMM)

วิธีแบบจำลองฮิดเดนมาร์คอฟ เป็นที่นิยมกันมากที่สุดในปัจจุบัน เพราะข้อดีหลายประการ เป็นต้นว่าความยืดหยุ่นในการปรับให้ใช้กับการรู้จำเสียงพูดที่มีลักษณะการพูดต่างๆ กัน สมรรถนะในการรู้จำที่สูง ความสามารถรองรับคำศัพท์จำนวนมาก อย่างไรก็ดีข้อเสียของมันคือ ต้องการเวลาในการฝึกฝนและรู้จำ โดยเฉพาะเมื่อมีคำศัพท์ใหม่ๆ เพิ่มขึ้นจะต้องเริ่มกระบวนการฝึกฝนใหม่ทุกครั้งไป ตัวอย่างงานวิจัยทางด้านนี้ ได้แก่ เสาวลักษณ์ อารีย์พงศา [5] และจิตรลดา จารุมิศร์ [1]

3. เครือข่ายนิวรอล (Neural Network)

เครือข่ายนิวรอล หรือนิวรอลเน็ตเวิร์ก เป็นกรรมวิธีที่เริ่มมีผู้สนใจทำวิจัยกัน โดยการเลียนแบบเครือข่ายประสาทของมนุษย์ ที่มีความเร็วในการรู้จำเหนือกว่ากรรมวิธีอื่น ข้อเสียของกรรมวิธีนี้จะเหมือนกับของฮิดเดนมาร์คอฟ ที่ต้องการเวลาในการฝึกฝน และต้องเริ่มการฝึกฝนใหม่ทุกครั้งที่มีการเพิ่มคำศัพท์ ตัวอย่างของงานวิจัยในแนวทางนี้ ได้แก่ วุฒิพงษ์ พรสุขจันทรา [6]

การค้นคว้าวิจัยในด้านการรู้จำเสียงพูดภาษาไทยที่มีในประเทศไทยนั้น เท่าที่สามารถสืบค้นเอกสารได้ ปรากฏว่ามีไม่มากนัก และเริ่มมีการศึกษาค้นคว้าวิจัยอย่างจริงจังตั้งแต่ประมาณปี พ.ศ. 2525 ดังจะเห็นได้จากผลงานวิจัยต่างๆ ได้แก่ การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดนมาร์คอฟโมเดล และเวกเตอร์ควอนไตซ์เซชั่น [5] การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นกับผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิ่ง [3] การออกแบบ แบบจำลองในการรู้จำเสียงวรรณยุกต์สำหรับภาษาไทย โดยใช้เทคนิคการควอนไตซ์พิทช์ และ Hidden Markov Modeling [1] และการรู้จำคำพูดภาษาไทยโดยใช้ลักษณะบ่งความต่างของหน่วยเสียง [2] โดยงานวิจัยเหล่านี้ล้วนเป็นการวิจัยเกี่ยวกับคำโดดภาษาไทย แต่ไม่เคยมีการศึกษาเชิงเปรียบเทียบถึงวิธีต่างๆ ของการดึงลักษณะเด่นของข้อมูล (Feature Extraction) หรือการประมวลผลเบื้องต้น (Front end) เพื่อพิจารณาวิธีที่ควรจะปรับใช้กับเสียงพูดภาษาไทย ได้อย่างมีประสิทธิภาพ

จากผลงานทางวิชาการของผู้เขียนวิทยานิพนธ์ [15], [16] ได้นำการแปลงข้อมูลแบบคาฮูเน็นเลิฟ (Karhunen-Loeve Tranfrom,KLT) กระทำกับสัญญาณเสียงทั้งที่อยู่ในโดเมนความถี่ และโดเมนของเวลา ผลลัพธ์ที่ได้จากการแปลงข้อมูลจะได้ เวกเตอร์ไอเก็น (Eigen Vector) และสัมประสิทธ์ (Coefficients) จาก [15] นำค่าสัมประสิทธ์ไปทำการวิเคราะห์แยกเสียงวรรณยุกต์ภาษาไทยสามัญ เอก โท ตรี จักวา และจาก [16] นำค่าสัมประสิทธ์เป็นตัวแทนเสียงพูดแต่ละเสียง เป็นข้อมูลให้กับระบบการเรียนรู้ และรู้จำ โดยใช้โครงข่ายประสาทเทียม จากการทดลองกับเสียงพูดคำไทยที่เป็นตัวเลข 0-9 ได้อัตราการรู้จำ 80% (LVQ1) และ 93% (LVQ2)

จากผลงานวิจัยวิธีการประมวลผลเบื้องต้นที่ใช้หาพารามิเตอร์เพื่อเป็นตัวแทนของเสียง มีอยู่ด้วยกันหลายวิธี เช่น วิธีการประมวลผลเบื้องต้นแบบ LPCC [7] และวิธีการประมวลผลเบื้องต้นแบบ MFCC [7] ซึ่งวิธีการทั้งหมดนี้เป็นที่ยอมรับและนิยมใช้ ในการพัฒนาระบบการรู้จำเสียงพูด

สำหรับระบบการรู้จำเสียงพูดที่มีลักษณะเป็นคำโดด (Isolate ,Word Recognition) จากงานวิจัยทั้งต่างประเทศ และภายในประเทศ ได้ถูกพัฒนาจนมีประสิทธิภาพสูง มีความถูกต้องสูง แต่ระบบเหล่านี้ใช้ได้ดีเฉพาะเสียงที่ถูกควบคุมคุณภาพเสียง (เสียงที่ใช้ในการฝึกสอน และเสียงที่ใช้ในการทดสอบไม่มีความแปรปรวนมากนัก) แต่สำหรับข้อมูลเสียงที่มีสัญญาณรบกวน ซึ่งอาจจะเกิดจากอุปกรณ์ภายใน หรือเสียงสอดแทรกอย่างอื่นเข้ามา เมื่อนำมาใช้กับระบบเหล่านี้ประสิทธิภาพจะลดลงอย่างมาก ดังนั้นจึงได้มีการวิจัยอีกแนวทางหนึ่งเพื่อพยายามให้ระบบการรู้จำเสียงสามารถนำไปใช้ได้ในทุกกรณี เรียกว่า Robust Speech Recognition โดยกลุ่มหนึ่งจะพยายามทำการ

ลดสัญญาณรบกวน หรือปรับคุณภาพเสียง เช่น [13] ใช้วิธี Spectral Subtraction วิธีแบบนี้ทำให้ระบบต้องมีขั้นตอนการประมวลผลเพิ่ม ซึ่งทำให้เสียเวลาไม่เหมาะในการนำไปประยุกต์ใช้งาน และอีกกลุ่มหนึ่งได้พยายามปรับวิธีการประมวลผลเบื้องต้น เช่น งานวิจัย [9] ได้นำวิธีการวิเคราะห์เสียงทั้งในโดเมนเวลา และโดเมนความถี่นำมารวมกัน เรียกว่า PLP

จากที่กล่าวมาข้างต้นทำให้วิทยานิพนธ์นี้จึงมุ่งเน้นในการวิจัย 2 ลักษณะ

1. วิจัยเชิงเปรียบเทียบวิธีการประมวลผลเบื้องต้นในการหาพารามิเตอร์ที่แสดงลักษณะเด่นของข้อมูลเสียงแบบต่างๆ เพื่อไปเป็นตัวแทนให้กับการเรียนรู้และรู้จำด้วยแบบจำลองฮิดเดนมาร์คอฟ โดยเราจะวัดที่ประสิทธิภาพในการรู้จำเสียง
2. เป็นการศึกษาเบื้องต้นถึงผลกระทบของเสียงที่มีสัญญาณรบกวน กับระบบการรู้จำเสียง ที่เรียกว่า Robust Speech Recognition

1.2 วัตถุประสงค์ในการทำวิทยานิพนธ์

ในวิทยานิพนธ์นี้ได้ศึกษาระบบการรู้จำเสียงพูดภาษาไทยแบบคำโดด ทั้งหมด 20 คำ โดยมีวัตถุประสงค์ดังนี้

1. ศึกษาวิธีการประมวลผลเบื้องต้นที่ใช้หาพารามิเตอร์แสดงลักษณะเด่นของข้อมูลเสียง ทั้ง LPCC และ MFCC
2. สร้างระบบการรู้จำเสียงพูดภาษาไทย โดยใช้แบบจำลองฮิดเดนมาร์คอฟ ในกระบวนการฝึกสอน และรู้จำ
3. ทำการทดลองระบบการรู้จำ และเปรียบเทียบประสิทธิภาพของระบบ
4. นำข้อมูลเสียงที่มีสัญญาณรบกวน ทดสอบกับระบบที่ทำขึ้น ศึกษาผลกระทบจากผลการทดลอง
5. แก้ไขปรับปรุงระบบการรู้จำ โดยใช้วิธีการประมวลผลเบื้องต้นแบบ PLP เข้ามาประยุกต์ใช้ และทำการทดสอบระบบ
6. จากการศึกษาทั้งหมดนี้นำเข้าไปสู่การพัฒนาระบบการรู้จำเสียงพูดแบบคำโดด เพื่อในอนาคตอันใกล้นี้ เราจะได้นำไปประยุกต์ใช้งานได้จริงในชีวิตประจำวัน

1.3 เป้าหมายและขอบเขตงานวิจัย

สามารถหาวิธีการประมวลผลเบื้องต้นที่เหมาะสมที่สุด กับระบบการรู้จำเสียงภาษาไทยแบบคำโดด

1.4 ขั้นตอนและวิธีการดำเนินงาน

1. ค้นคว้าและเก็บข้อมูลรายละเอียดที่เกี่ยวข้องดังนี้

- ทฤษฎีเกี่ยวกับการวิเคราะห์เสียง

- การประมวลผลเบื้องต้นแบบต่างๆ ดังนี้
  - LPCC
  - MFCC
  - PLP
- ระบบการรู้จำเสียงโดยใช้ แบบจำลองฮิดเดนมาร์คอฟ

2. เก็บข้อมูลเสียงพูดภาษาไทย
   - เก็บข้อมูลเสียงในห้องทดลองที่ควบคุมสัญญาณรบกวน (clean speech)
   - สร้างสัญญาณรบกวนแบบ Gaussian
   - เก็บสัญญาณรบกวนที่เกิดขึ้นจริงตามสถานที่ต่างๆ
3. วิเคราะห์และพัฒนาโปรแกรมในแต่ละส่วน
4. ทดสอบระบบการรู้จำเสียงที่มีการประมวลผลเบื้องต้นแบบต่างๆ และทดสอบทั้ง เสียงที่ปราศจากสัญญาณรบกวน (clean speech) และเสียงที่มีสัญญาณรบกวน (noise speech)
5. สรุปรวบรวมผลการทดลองทั้งหมด

1.5 ข้อกำหนดในการทำวิทยานิพนธ์

1. งานวิจัยนี้จะทำการศึกษาบนเครื่องคอมพิวเตอร์ส่วนบุคคล โดยใช้โปรแกรม MATLAB 5.0 โดยมีอุปกรณ์เพิ่มเติมได้แก่ การ์ดเสียง โปรแกรมอัดเสียง (COOL96) ไมโครโฟน และลำโพง
2. เสียงพูดที่ใช้ในการทดสอบ โดยมีผู้พูดเป็นเพศชาย 20 คน อายุระหว่าง 20-30 ปี ระบบนี้จะทำการรู้จำเสียงแบบคำโดด 20 คำ โดยเสียงที่ใช้ในการฝึกสอนทั้งหมด 900 เสียง และสำหรับทดสอบการรู้จำอีก 300 เสียง
3. สำหรับสัญญาณรบกวนที่เราใช้จะมีระดับของสัญญาณรบกวนต่างๆ กัน

1.6 โครงประกอบของวิทยานิพนธ์

แบ่งออกเป็น 6 บทดังนี้

บทที่ 1 ดังได้กล่าวแล้วข้างต้น

บทที่ 2 กล่าวถึงทฤษฎีการวิเคราะห์เสียง และระบบการรู้จำเสียงด้วยแบบจำลองฮิดเดนมาร์คอฟ

บทที่ 3 กล่าวถึงการประมวลผลเบื้องต้น เป็นวิธีดึงคุณลักษณะเด่นของข้อมูลเสียงพูด เพื่อเป็นตัวแทนให้กับระบบการรู้จำ ในวิทยานิพนธ์นี้เราอธิบายการประมวลผลเบื้องต้น 3 วิธี LPCC, MFCC และ PLP

บทที่ 4 การทดลอง และผลการทดลอง

บทที่ 5 สรุปผลการทดลอง

บทที่ 2
ทฤษฎีการวิเคราะห์เสียง และระบบการรู้จำเสียง

การวิเคราะห์เสียงพูด คือการหาค่าพารามิเตอร์ต่างๆ ที่เป็นลักษณะเฉพาะเพื่อเป็นตัวแทนของเสียง ซึ่งทำให้ข้อมูลมีจำนวนน้อยลง แต่ยังคงแสดงคุณสมบัติของสัญญาณสียงได้อย่างถูกต้อง โดยทั่วไปแล้วเราจะทำการวิเคราะห์เสียงโดยใช้ลักษณะเด่นเชิงสเปกตรัม (spectral feature) การวิเคราะห์เสียงพูดนั้นมีอยู่หลายวิธี มีทั้งในด้านโดเมนเวลา (time domain) คือ วิเคราะห์จากรูปคลื่นของสัญญาณเสียงตามแกนเวลาโดยตรง และในด้านโดเมนความถี่ (frequency domain) คือวิเคราะห์จากสเปกตรัมของสัญญาณเสียง

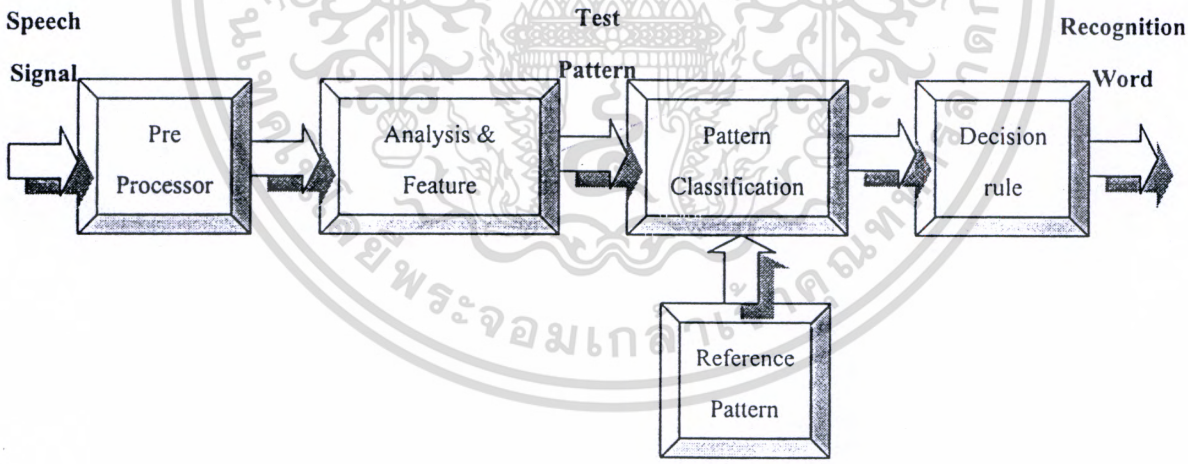
ระบบการรู้จำเสียงที่ใช้ในวิทยานิพนธ์นี้เป็นโครงสร้างตามรูปที่ 2.1 จะประกอบไปด้วยขั้นตอนการดำเนินงาน 4 ขั้นตอนหลัก [11]

1.การเตรียมข้อมูลเบื้องต้นก่อนการวิเคราะห์เสียง (Preprocessing)

2. การวิเคราะห์เสียงพูด (Speech Analysis)

3. การจำแนกรูปแบบ (Pattern Classification)

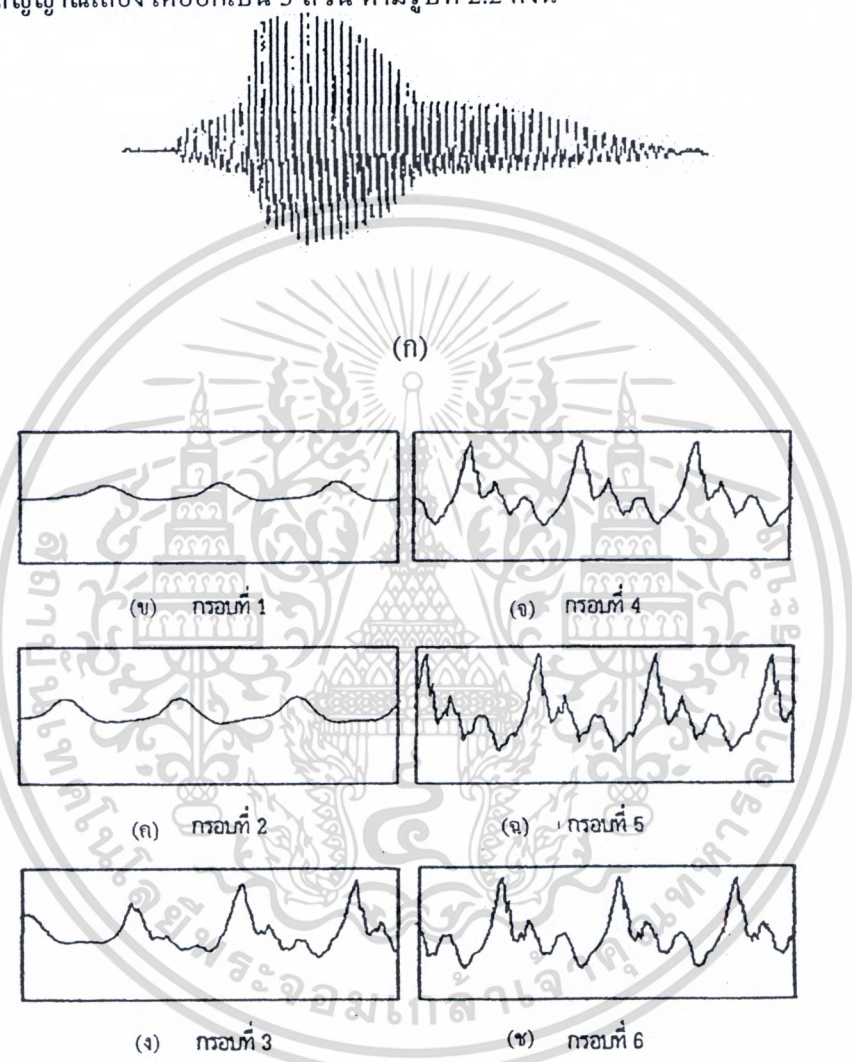
4. ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)



รูปที่ 2.1 โครงสร้างของระบบการรู้จำเสียงพูด

2.1 การเตรียมข้อมูลเบื้องต้นก่อนการวิเคราะห์เสียง (Preprocessing)

การเตรียมข้อมูลเบื้องต้นเป็นขั้นตอนในการจัดเตรียมข้อมูล จากข้อมูลดิบของเสียงพูดที่ได้จากการบันทึกเสียง เพื่อใช้ในการประมวลผลในขั้นตอนต่อไป เนื่องจากสัญญาณเสียงพูดโดยรวมจะมีสัญญาณรบกวน มีการแปรเปลี่ยนตามเวลา (Nonstationary) ดังรูปที่ 2.2(ก) สามารถแบ่งลักษณะของสัญญาณเสียงได้ออกเป็น 3 ส่วน ตามรูปที่ 2.2 ดังนี้



รูปที่ 2.2 สัญญาณเสียงพูด " หนึ่ง" ในแต่ละกรอบเสียงพูด ขนาดกรอบ 25 มิลลิวินาที

1. ช่วงที่ยังไม่มีการเปล่งเสียงหรือช่วงเงียบ (silence) เสียงในช่วงนี้ค่อนข้างเรียบถ้าไม่มีสัญญาณรบกวนจากภายนอก ตามรูปที่ 2.3ข และ 2.3ค
2. ช่วงก่อนที่จะเปล่งเสียงออกมาหรือเรียกว่า เสียงอโฆษะ (unvoice speech) ในช่วงนี้แอมพลิจูดเสียงพูดจะต่ำและไม่มีความเป็นคาบ ตามรูปที่ 2.3ง
3. ช่วงที่เป็นคำพูดหรือเรียกว่า เสียงโฆษะ (voice speech) ในช่วงนี้เสียงพูดจะมีลักษณะเป็นคาบจะมีแอมพลิจูดสูง ตามรูปที่ 2.3จ และ 2.3ช

ดังนั้นก่อนที่จะทำการวิเคราะห์เสียงพูดเราต้องผ่านขั้นตอนการเตรียมข้อมูลเบื้องต้นก่อน เช่น กำจัดสัญญาณรบกวน (pre-emphasis) การแบ่งช่วงสัญญาณ (frame blocking) โดยแต่ละช่วงของสัญญาณเสียงมีความยาวประมาณ 10-40 มิลลิวินาที ซึ่งถือได้ว่าเป็นช่วงที่มีข้อมูลเสียงจริงๆ จะมีความเสถียรและไม่แปรเปลี่ยนตามเวลา (stationary) ตามรูปที่ 2.3ง ถึง 2.3ฉ และการวินโดว์ (windowing) เป็นการเน้นข้อมูลส่วนกลางเฟรมของข้อมูลที่จะนำไปวิเคราะห์ จากนั้นจึงสามารถนำไปทำการวิเคราะห์สัญญาณเสียงพูดในแต่ละเฟรม

2.1.1 พรีเอมฟาซิส (pre-emphasis)

องค์ประกอบส่วนใหญ่ของสัญญาณเสียงพูดจะอยู่ที่บริเวณช่วงความถี่ต่ำ เมื่อเทียบกับแถบความถี่ (bandwidth) ไม่เกิน 4 kHz จึงทำให้ช่วงบริเวณความถี่สูงจะมีอัตราส่วนสัญญาณเสียงต่อสัญญาณรบกวน (signal to noise ratio : SNR) สูง เพื่อให้อัตราส่วนสัญญาณเสียงต่อสัญญาณรบกวนมีค่าค่อนข้างคงที่ตลอดช่วงแถบความถี่ทั้งหมด จึงต้องมีการพรีเอมฟาซิส (pre-emphasis) ก็คือการกรองสัญญาณด้วยวงจรกรองความถี่สูงผ่านซึ่งมักนิยมใช้วงจรกรองอันดับหนึ่ง ตามสมการที่ 2.1 ค่า a อยู่ในช่วงระหว่าง 0.9 ถึง 1 ค่าที่นิยมใช้คือ 0.95 เป็นการเน้นให้สัญญาณช่วงความถี่สูงมีขนาดสูงขึ้นนั้นเอง

$$s'(n) = s(n) - a * s(n-1) \qquad (2.1)$$

2.1.2 การแบ่งช่วงสัญญาณ (frame blocking)

สัญญาณเสียงพูดเป็นสัญญาณที่มีลักษณะไม่คงที่ (nonstationary signal) ดังนั้นเทคนิคในการวิเคราะห์เสียงพูดส่วนใหญ่แล้ว จะสมมุติให้สัญญาณเสียงมีคุณสมบัติที่เปลี่ยนแปลงสัมพันธ์กับเวลาอย่างคงที่ นั่นก็คือเราจะต้องแบ่งทำการวิเคราะห์หาพารามิเตอร์ของสัญญาณเสียงพูดในช่วงเวลาสั้นๆ ที่เรียกว่าการแบ่งช่วงสัญญาณ (frame blocking) เพื่อทำการวิเคราะห์หาพารามิเตอร์ การกำหนดขนาดที่จะใช้ขึ้นอยู่กับ

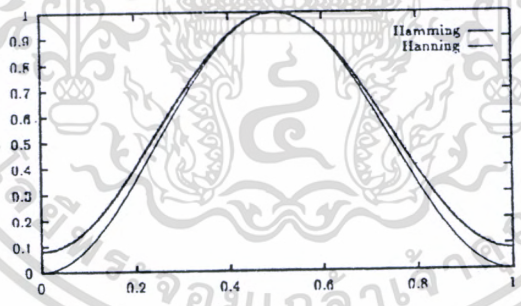
1) จะต้องสั้นพอที่จะทำให้คุณสมบัติของเสียงที่กำลังพิจารณาไม่มีการเปลี่ยนแปลงอย่างมีนัยสำคัญ
2) จะต้องยาวพอที่จะทำให้การจัดเตรียมตัวอย่างของเสียงเพื่อจะนำไปคำนวณหาพารามิเตอร์ให้ได้ตามต้องการอย่างเช่น ในกรณีที่มีสัญญาณรบกวนเข้ามาแทรกอยู่บางช่วงในสัญญาณเสียงด้วย ถ้าเราเลือกใช้ช่องแคบที่มีขนาดใหญ่กว่า เมื่อทำการหาค่าพารามิเตอร์โดยเฉลี่ยแล้ว ก็จะทำให้ส่วนประกอบของสัญญาณรบกวนถูกตัดทิ้งหรือมองข้ามไป
3) ขนาดที่เหมาะสม ไม่ควรสั้นเกินกว่าช่วงหนึ่งคาบของสัญญาณเสียงในช่วงที่กำลัง

วิเคราะห์ เงื่อนไขนี้จะมีผลต่อค่าเฟรมเรท (frame rate) ซึ่งก็คือ จำนวนครั้งต่อวินาทีที่ทำการวิเคราะห์สัญญาณเสียง โดยการขยับกรอบการวิเคราะห์ไปเป็นคาบๆ ตามแกนเวลา) ตามปกติเฟรมเรทจะมีค่าประมาณ 2 เท่าของส่วนกลับของขนาดกรอบการวิเคราะห์ นั้นก็คือซ้อนทับกัน 50 เปอร์เซ็นต์

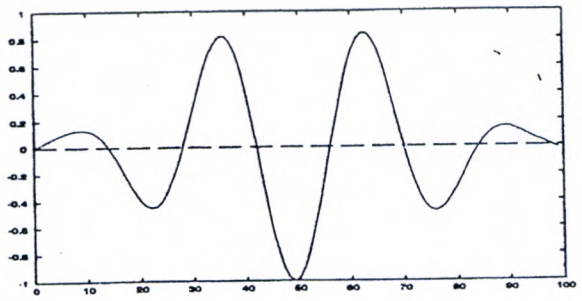
2.1.3 การวินโดว์ (Windowing)

โดยปกติการประยุกต์ส่วนมากจะใช้การแบ่งช่วงสัญญาณ ที่มีช่วงกว้างกว่าช่วงของข้อมูลที่คงที่ และเน้นเฉพาะข้อมูลตรงกลางของกรอบการวิเคราะห์ให้เป็นส่วนของข้อมูลที่จะทำการวิเคราะห์ ด้วยวินโดว์ฟังก์ชัน (windows funtion) เช่น ข้อมูลสัญญาณเสียงที่มีลักษณะคงที่ในช่วงเวลา 10 มิลลิวินาที ก็อาจจะแบ่งช่วงสัญญาณ ขนาด 20 มิลลิวินาที โดยช่วงกึ่งกลางขนาด 10 มิลลิวินาที จะมีการเน้นเพิ่มน้ำหนักให้มากกว่าช่อง 5 วินาที ที่ริมทั้งสองข้างของกรอบการวิเคราะห์ วินโดว์ฟังก์ชัน (windows funtion) ที่นิยมใช้กันมากในการวิเคราะห์สัญญาณเสียงก็คือ ฟังก์ชัน Hamming ตามสมการที่ 2.2 รูปที่ 2.3 แสดงวินโดว์ฟังก์ชั่น แบบ Hamming และ แบบ Hanming รูปที่ 2.4 เป็นลักษณะสัญญาณหลังการนำวินโดว์ฟังก์ชั่น Hamming กระทำกับสัญญาณซายส์

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n)/(N-1) \ldots\ldots n = 0,1,2\ldots\ldots,N-1 \\ 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots n = otherwise \end{cases} \quad (2.2)$$



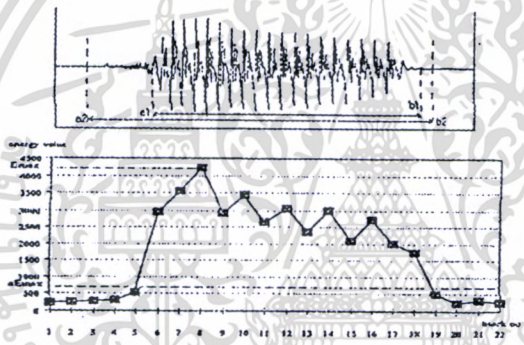
รูปที่ 2.3 แสดง Hamming windows และ Hanming windows



รูปที่ 2.4 แสดงสัญญาณซายส์กระทำกับ Hamming windows

2.1.4 การหาจุดสิ้นสุดของเสียงพูด (Endpoint Detection)

การหาจุดสิ้นสุดของเสียงพูดทั้งทางส่วนหน้า และส่วนท้ายเป็นการหาช่วงของข้อมูลเสียงพูดจริงๆ ที่ได้จากการบันทึกเสียง มีหลายวิธีสำหรับการประมวลผลในขั้นตอนนี้ เช่น การหาจุดสิ้นสุดโดยใช้แอมพลิจูด ใช้ค่าพลังงาน และอัตราการตัดศูนย์ แต่ในวิทยานิพนธ์นี้เราเลือกใช้วิธีใช้ค่าพลังงาน โดยหาค่าพลังงานในแต่เฟรม เพื่อหาจุดที่มีพลังงานมากกว่าระดับที่กำหนดไว้ ติดต่อกันนานกว่าช่วงเวลาที่กำหนด จุดเริ่มต้นของเสียงพูดจะอยู่ก่อนจุดที่ตรวจพบ และการหาส่วนท้ายของเสียงก็ใช้วิธีเดียวกัน หาจุดที่มีพลังงานน้อยกว่าระดับที่กำหนดไว้ ติดต่อกันนานกว่าช่วงเวลาที่กำหนด แต่จุดสิ้นสุดจะอยู่หลังจุดที่ตรวจพบ ถึงแม้ว่าวิธีนี้ยังไม่ใช้วิธีที่ดีสุดเพราะมีโอกาสกำหนดช่วงของเสียงพูดผิดพลาดได้ แต่ก็เป็นวิธีที่นิยมใช้ และเข้าใจง่าย ตามรูปที่ 2.5 a1 เป็นจุดที่มีค่าพลังงานสูงการระดับที่ตั้งไว้ และ b1 เป็นจุดที่พลังงานต่ำกว่าระดับ ดังนั้นช่วงของข้อมูลเสียงพูดก็คือช่วงจาก a2 ถึง b2



รูปที่ 2.5 การหาจุดสิ้นสุดของเสียงพูดโดยใช้วิธีหาค่าพลังงาน

2.2 การวิเคราะห์เสียงพูด (Speech Analysis)

เป็นเทคนิคการลดจำนวนข้อมูล โดยที่ข้อมูลจำนวนมากจะถูกแปลงเป็นชุดของข้อมูลที่มีจำนวนน้อยลง และยังคงแสดงคุณสมบัติสำคัญของรูปคลื่นสัญญาณเสียงได้อย่างถูกต้อง โดยทั่วไปสัญญาณเสียงถูกวิเคราะห์โดยใช้ลักษณะเด่นเชิงสเปกตรัม (Spectral feature) เพราะลักษณะเด่นส่วนใหญ่สำหรับการรับรู้เสียงพูดโดยหูของมนุษย์รวมอยู่ในข้อมูลเชิงสเปกตรัม วิธีการสกัดเอนเวโลปเชิงสเปกตรัม (Spectral envelope) แบ่งออกเป็นการวิเคราะห์โดยใช้พารามิเตอร์ (Parametric analysis) และการวิเคราะห์โดยไม่ใช้พารามิเตอร์ (nonparametric analysis) การวิเคราะห์โดยใช้พารามิเตอร์จะเลือกแบบจำลองที่เหมาะสมกับสัญญาณ และปรับแต่งพารามิเตอร์ลักษณะเด่นที่ใช้แทนแบบจำลองนั้น ในขณะที่การวิเคราะห์โดยไม่ใช้พารามิเตอร์สามารถประยุกต์ใช้กับสัญญาณหลายชนิดได้ เพราะการวิเคราะห์วิธีนี้ไม่ได้สร้างแบบจำลองสัญญาณ ถ้าแบบจำลองที่ใช้มีความ

เหมาะสมกับสัญญาณ การวิเคราะห์โดยใช้พารามิเตอร์จะสามารถแสดงลักษณะเด่นของสัญญาณได้ดีกว่า

การวิเคราะห์โดยไม่ใช้พารามิเตอร์ มีวิธีการหลักๆ ดังนี้

1. ชุดวงจรกรองความถี่ (band-pass filter bank) วิธีนี้นำสัญญาณเสียงมาผ่านวงจรกรองความถี่หลายวงจรที่มีช่วงความถี่ผ่านแตกต่างกัน วงจรกรองความถี่แต่ละวงจรจะให้สัญญาณเอาต์พุตที่สัมพันธ์กับพลังงานของสัญญาณในช่วงความถี่ของวงจรกรองนั้น วิธีนี้มีข้อดีคือสร้างเป็นฮาร์ดแวร์ได้ง่ายและเหมาะสมสำหรับการประมวลผลเวลาจริง (real-time processing)
2. การวิเคราะห์การตัดค่าศูนย์ (zero-crossing analysis) จะนับจำนวนการเปลี่ยนเครื่องหมายของสัญญาณ ซึ่งเป็นการประมาณค่าความถี่ฟอร์เมนท์ (Formant frequency) คือ ความถี่ที่มีพลังงานสูงสุด การวิเคราะห์วิธีนี้มักใช้ร่วมกับวิธีชุดวงจรกรองผ่านแถบ
3. การวิเคราะห์โดยใช้เซ็ปสตรัม (cepstrum) การวิเคราะห์วิธีนี้มีข้อดีคือ สามารถแยกเอนเวโลปเชิงสเปกตรัมและโครงสร้างย่อยเชิงสเปกตรัม (spectral fine structure) ออกจากกันได้ในโดเมนคิวเฟรินซี่ (quefrency domain) ซึ่งเป็นพารามิเตอร์ในโดเมนเวลา แต่มีข้อเสียคือต้องคำนวณผลการเปลี่ยนแปลงฟูริเยร์แบบเร็ว (fast fourier transform) 2 ครั้ง และคำนวณค่าลอการิทึม ซึ่งต้องใช้เวลาในการคำนวณมาก

การวิเคราะห์โดยใช้พารามิเตอร์ มีวิธีการหลักๆ ดังนี้

1. การวิเคราะห์โดยการสังเคราะห์ (analysis-by-synthesis) วิธีนี้สามารถสร้างแบบจำลองที่ถูกต้องแม่นยำได้ โดยการใช้พารามิเตอร์หลายค่าเช่น ค่าความถี่ฟอร์แมนท์, ความกว้างแถบ (bandwidth), เอนเวโลปเชิงสเปกตัมและอื่นๆ แต่มีข้อเสียคือต้องใช้เวลาคำนวณในการวนซ้ำมากเพราะค่าพารามิเตอร์หลายค่ามีผลกระทบต่อกัน
2. การประมาณพันธะเชิงเส้น (linear predictive coding) ซึ่งเป็นวิธีที่ใช้กันแพร่หลายในการหาลักษณะที่สำคัญของเสียง โดยการสร้างแบบจำลองของสเปกตรัมอย่างง่ายโดยใช้โพล (all-pole spectrum modeling) พารามิเตอร์สามารถประมาณค่าได้จากค่าความแปรปรวนร่วมหรือค่าอัตสหสัมพันธ์ โดยไม่ใช้การวนซ้ำ วิธีนี้มีข้อดีคือ สามารถแทนสัญญาณเสียงได้อย่างมีประสิทธิภาพโดยใช้พารามิเตอร์จำนวนน้อยและใช้การคำนวณที่ค่อนข้างง่าย

การวิเคราะห์และวัดค่าลักษณะสำคัญเป็นการวิเคราะห์สัญญาณเสียงพูด เพื่อเก็บรวบรวมลักษณะสำคัญของเสียงพูดแต่ละเสียง สำหรับการฝึกสอนระบบให้รับรู้ถึงความแตกต่างของเสียง

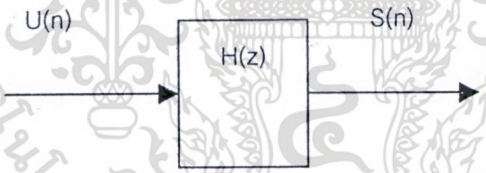
พูดแต่ละเสียงและเพื่อใช้ในการเปรียบเทียบแบ่งแยกความแตกต่างของเสียงพูดแต่ละเสียงออกจากกัน ด้วยแบบจำลองฮิดเดนมาร์คอฟ

2.2.1 การประมาณพันธะเชิงเส้น (Linear Prediction Coefficients : LPC)

การประมาณพันธะเชิงเส้นเป็นขบวนการทางคณิตศาสตร์ที่ใช้ในการหาเอกลักษณ์ของเสียง โดยพิจารณาว่าเสียงเกิดจากผลรวมเชิงเส้น (linear combination) ของสัญญาณที่ทราบค่าแล้ว โดยใช้วิธีกำลังสองน้อยที่สุด (least-squares method) ในการเลือกค่าพารามิเตอร์ของระบบ

หลักการประมาณพันธะเชิงเส้นมีวิธีการหลัก 2 วิธี คือ วิธีการหาค่าความแปรปรวนร่วมและวิธีอัตสหสัมพันธ์ เนื่องจากวิธีอัตสหสัมพันธ์ใช้การคำนวณน้อยกว่าวิธีความแปรปรวนร่วมและมีความแน่นอนด้านเสถียรภาพ ดังนั้นงานวิจัยนี้จึงเลือกใช้การประมาณพันธะเชิงเส้นโดยวิธีอัตสหสัมพันธ์ การประมาณพันธะเชิงเส้นสามารถแสดงคุณสมบัติได้ใกล้เคียงกับพื้นฐานรูปแบบการกำเนิดเสียงของมนุษย์ ข้อมูลสัญญาณเสียงพูด s(n) เมื่อใช้การประมาณด้วยผลรวมเชิงเส้น (linear combination) ของจำนวน p ข้อมูลที่ผ่านมาตามสมการที่ 2.3

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p) \tag{2.3}$$



รูปที่ 2.6 โมเดลการกำเนิดเสียงพูด

จากโมเดลของการกำเนิดเสียงพูดตามรูปที่ 2.6 สัญญาณการกระตุ้น U(n) ผ่าน H(z) เป็นฟังชั่นถ่ายโอน (transfer funtion) แสดงถึงคุณสมบัติของทางเดินเสียง ได้สัญญาณเสียง S(n) ตามต่อไปนี้

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k 2^{-k}} \tag{2.4}$$

ดังนั้น

$$(1-\sum_{k=1}^{p}a_k Z^{-k})S(n)=AU(n) \tag{2.5}$$

หรือ

$$s(n)=\sum_{k=1}^{p}a_k s(n-k)+Au(n) \tag{2.6}$$

เมื่อ A = อัตราการขยาย(Gain)

u(n) = สัญญาณการกระตุ้น (normalized excitation)

a_k = สัมประสิทธ์ของ digital filter

จากสมการที่ 2.3 จะได้

$$s(n)\approx s'(n)=\sum_{k=1}^{p}a_k s(n-k) \tag{2.7}$$

จากสมการที่ 2.6 และ 2.7 จะได้ค่าความผิดพลาดดังนี้

$$e(n)=s(n)-s'(n)=s(n)-\sum_{k=1}^{p}a_k s(n-k)=Au(n) \tag{2.8}$$

การวิเคราะห์ด้วยวิธีนี้ จะคำนึงถึงแต่ส่วนสัญญาณที่ผ่านช่องทางเดินเสียง ไม่คำนึงถึงสัญญาณการกระตุ้น Au(n) = 0 ดังนั้นเราจึงใช้วิธีการของความผิดพลาดกำลังสองน้อยที่สุด เพื่อหาค่าสัมประสิทธิ์ a_k ของฟังชั่นถ่ายโอน ให้ E_n แทนค่าผลรวมของกำลังสองของความคลาดเคลื่อนซึ่งมีค่าดังนี้

$$E_n=\sum_m e_n^2(m)=\sum_m[s_n(m)-\sum_{k=1}^{p}a_k s_n(m-k)]^2 \tag{2.9}$$

โดยที่ n คือช่วงที่ n ของสัญญาณที่ใช้คำนวณ และ m คือข้อมูลในช่วงของ n ค่าสัมประสิทธิ์ a_k ที่ทำให้ E มีค่าน้อยที่สุดสามารถหาได้โดยการแก้สมการ $\frac{\partial E}{\partial a_k}=0$

$$\frac{\partial E}{\partial a_k} = -2s_n(m-i)\sum_m [s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k)] = 0 \qquad (2.10)$$

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} a_k \sum_m s_n(m-i)s_n(m-k) \qquad (2.11)$$

เนื่องจากพจน์ด้านซ้ายมือของสมการคือค่าอัตสหสัมพันธ์ R(i) ของ S(n) จะได้

$$\sum_{k=1}^{p} R_n(|i-k|)a_k = R_n(i) \qquad (2.12)$$

จากสมการที่ 2.12 เขียนให้อยู่ในรูปเมตริกซ์ได้เป็น

$$\begin{bmatrix} R_n(0)..... R_n(1)..... ..... R_n(p-1) \\ R_n(1)..... R_n(0)..... ..... R_n(p-2) \\ . \\ . \\ R_n(p-1).R_n(p-2)..... ...R_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ . \\ . \\ R_n(p) \end{bmatrix} \qquad (2.13)$$

การแก้สมการเชิงเส้น p สมการของเมตริกซ์ pxp นี้จะได้ค่าสมประสิทธิ์ a_k เนื่องจากเมตริกซ์ของค่าอัตสหสัมพันธ์ (Autocorrelation) อยู่ในรูปของเมตริกซ์ Toepliz ทำให้แก้สมการเพื่อหาค่าสัมประสิทธิ์ โดยใช้ Durbin Method โดยทำการคำนวณลำดับของสมการสำหรับ m=1,2,.....,p

$$\begin{aligned} k_m &= \frac{R(m) - \sum_{i=1}^{m-1} a_{m-1}(i)R(m-i)}{E_{m-1}} \\ a_m(m) &= k_m \\ a_m(i) &= a_{m-1}(i) - k_m a_{m-1}(m-i)..... \ 1 \le i \le m \\ E_m &= (1-k_m^2)E_{m-1} \end{aligned} \qquad (2.14)$$

โดยที่ $E_0=R(0)$ และ $a_0=0$ ในแต่ละรอบของ m ค่าสัมประสิทธิ์ $a_m(i)$ เมื่อ i=1,2,.....,m แทนการประมาณพันธะเชิงเส้นอันดับ (order) m ที่เหมาะที่สุด

2.2.2 การวิเคราะห์ในโดเมนความถี่ช่วงเวลาสั้นๆ (Short-Term Fourier Analysis)

การวิเคราะห์ตามแกนความถี่ ก็คือการวิเคราะห์ทางสเปกตรัมนั่นเอง โดยจะวิเคราะห์หรือพิจารณาสัญญาณเสียงในด้านของโดเมนความถี่ เหตุที่ต้องมีการวิเคราะห์ทางด้านสเปกตรัมนั้นก็เนื่องจากพารามิเตอร์สัญญาณเสียงส่วนใหญ่ จะสามารถหาได้จากโดเมนความถี่ และสัญญาณที่สร้างออกมาจากการเปลี่ยนแปลงรูปร่างของทางเดินเสียงของคน ก็จะสามารถพิจารณาเป็นรูปแบบในโดเมนของความถี่ได้ง่าย และคงที่กว่าวิเคราะห์ในโดเมนเวลา

การแปลงฟูริเยร์ จะสามารถแทนค่าสัญญาณเสียงออกมาในเทอมของพลังงานและความถี่ ถ้าเรามองทางเดินเสียงของคนเป็นระบบเชิงเส้นอันหนึ่ง การแปลงฟูริเยร์ของสัญญาณเสียงก็คือ ผลลัพธ์ที่เกิดจากการกระตุ้นของเส้นเสียงกับการตอบสนองของทางเดินเสียงนั่นเอง นิยามของการแปลงฟูริเยร์ของสัญญาณ s(n) ใดๆ ดังนี้

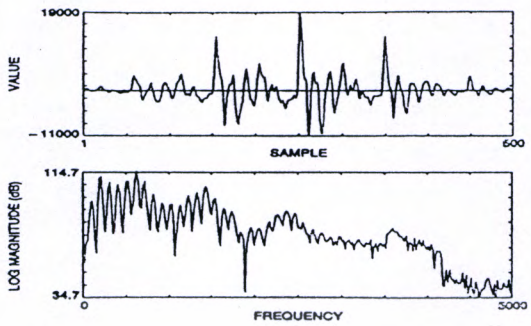
$$S(f_i) = \sum_{n=0}^{N-1} s(n)e^{-j(2\pi f_i/f_s)n}$$
$$f_i = \frac{f_s}{N} i \quad \ldots\ldots\ldots\ldots\ldots 1 \le i \le N/2 \qquad (2.15)$$

การวิเคราะห์ทางสเปกตรัมในช่วงเวลาสั้นๆ เป็นวิธีการดั้งเดิมอันหนึ่งของการวิเคราะห์สัญญาณเสียงที่สำคัญ ด้วยข้อสมมติฐานที่ว่าสัญญาณเสียงในช่วงเวลายาวๆ จะมีการเปลี่ยนแปลงอยู่ตลอดเวลาไม่มีสภาพของการคงที่ แต่ถ้าวิเคราะห์ในช่วงเวลาสั้นๆ ช่วงหนึ่งจะสามารถแทนเป็นค่าสเปกตรัมของสัญญาณเสียงในช่วงเวลาดังกล่าวเป็นอย่างดี การวิเคราะห์สเปกตรัมของสัญญาณเสียงในช่วงเวลาอันสั้นนั้น จะเป็นพื้นฐานของเทคนิคการวิเคราะห์เสียงต่างๆ

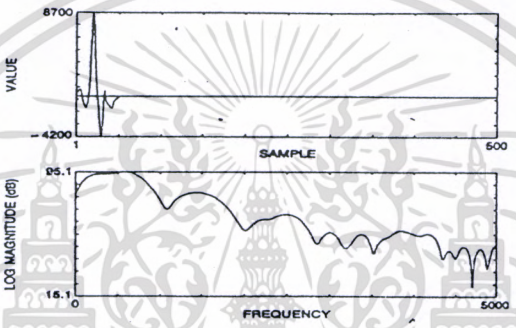
การแสดงสเปกตรัมของสัญญาณเสียง โดยการใช้การแปลงฟูริเยร์ในช่วงเวลาสั้นๆ ของสัญญาณเสียงมาพล็อตรูปกราฟโดยมีแกนนอนเป็นแกนของความถี่ และแกนตั้งแสดงค่าของแมกนิจูด ดังรูปที่ 2.7 ซึ่งแสดงสเปกตรัมแบบช่วงกรองกว้าง และช่วงกรองแคบ

สเปกตรัมแบบช่วงกรองแคบ (narrowband spectrogram) จะใช้วิเคราะห์สัญญาณเมื่อต้องการความละเอียดสูง ตามรูปที่ 2.7 (ก)

สเปกตรัมแบบช่วงกรอบกว้าง (wide-band spectrogram) จะใช้วิเคราะห์สัญญาณที่ไม่ต้องการมีความละเอียดสูงตามรูปที่ 2.7 (ข)



(ก)



(ข)

รูปที่ 2.7 (ก) ลักษณะสเปกตรัมแบบช่วงกรองกว้าง และ (ข) ช่วงกรองแคบ

2.2.3 การวิเคราะห์เซ็ปสตรัม (Cepstral analysis)

เนื่องจากกระบวนการในการผลิตสัญญาณเสียงพูด ได้จากการที่มีสัญญาณการกระตุ้นคอนโวลูชัน (convolution) กับการตอบสนองของทางเดินเสียงของคน เมื่อเขียนให้อยู่ในรูปผลคูณในโดเมนความถี่ได้ตามนี้

$$S(e^{i\theta}) = H(e^{i\theta})E(e^{i\theta}) \tag{2.16}$$

รูปที่ 2.8 จะเห็นได้ว่ามี 2 ส่วน คือ $H(e^{i\theta})$ แสดงถึง เอนเวโลปเชิงสเปกตัม (spectrum envelope) ของสเปกตรัมสัญญาณเสียงพูด และ ส่วนที่มียอดเล็กๆ จำนวนมาก $E(e^{i\theta})$ แสดงรายละเอียดของสัญญาณกระตุ้น หรือรายละเอียดของสเปกตรัม จากสมการข้างบนเมื่อนำ log มากระทำจะเขียนได้ใหม่ดังนี้

$$\log(S(e^{i\theta})) = \log(H(e^{i\theta})) + \log(E(e^{i\theta})) \tag{2.17}$$

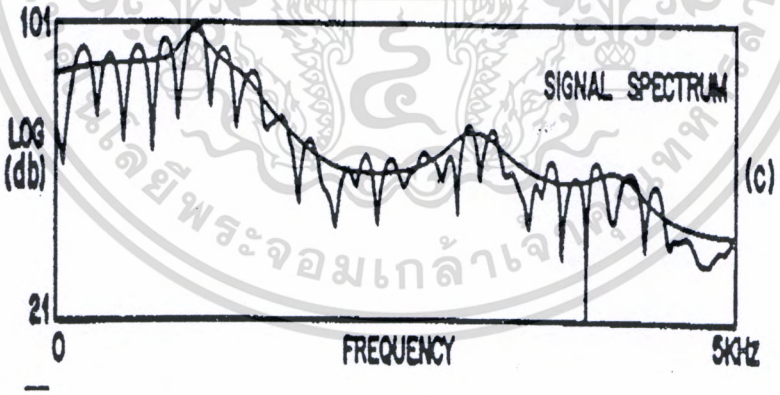
ในการวิเคราะห์หาพารามิเตอร์เพื่อเป็นตัวแทนของสัญญาณเสียงพูดส่วนใหญ่จะใช้แต่ส่วนที่เป็นแอมพลิจูดของสัญญาณ

$$\log(|S(e^{i\theta})|) = \log(|H(e^{i\theta})|) + \log(|E(e^{i\theta})|) \tag{2.18}$$

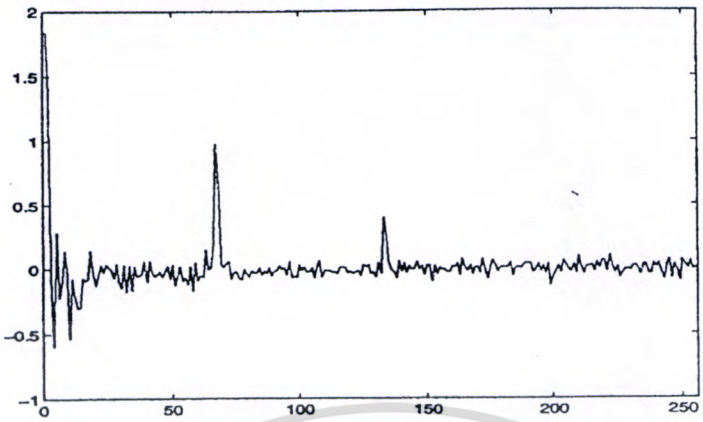
ในการหาสัมประสิทธิ์เซ็ปสตรัม ก็คือการแยกส่วนของเอนเวโลปเชิงสเปกตัมสัญญาณเสียงพูด ออกจากรายละเอียดย่อยของมันนั้นเอง กระทำได้โดยใช้การแปลงดิสครีดฟูริเยร์กลับ (discrete fourier transform : IDFT) กระทำกลับ log สเปกตรัมของสัญญาณเสียงพูด จะได้สัมประสิทธิ์เซ็ปสตรัม ตามนี้

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|S(k)|e^{(2\pi/N)kn} \tag{2.19}$$

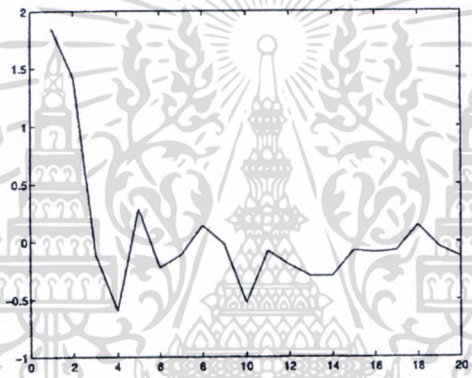
จากรูปที่ 2.9 ที่บริเวณ n เข้าใกล้ 0 เป็นข้อมูลที่มีค่ามากจึงมีความสำคัญสูง นั้นก็คือส่วนของการตอบสนองของทางเดินเสียงคนนั้นเอง ส่วนยอดเล็กๆ ซึ่งจะแสดงบริเวณเวลาเท่ากับจำนวนเท่าของคาบ คือข้อมูลรายละเอียดของสัญญาณกระตุ้น ดังนั้นข้อมูลที่เราจะนำไปใช้เป็นพารามิเตอร์ของเซ็ปสตรัม เราจะนำข้อมูลส่วนที่อยู่เฉพาะในคาบแรกเท่านั้น ตามรูปที่ 2.10



รูปที่ 2.8 สเปกตรัมของสัญญาณเสียงพูด



รูปที่ 2.9 สัมประสิทธิ์เซ็ปสตรัมของสัญญาณเสียงพูด



รูปที่ 2.10 สัมประสิทธิ์เซ็ปสตรัมของสัญญาณเสียงพูดที่นำไปเป็นพารามิเตอร์

2.3 การจำแนกรูปแบบ (Pattern Classification)

การจำแนกรูปแบบที่ถูกใช้ในระบบการรู้จำเสียงมี 3 วิธีการ ได้แก่ ไดนาร์มิคไทม์วาร์ปิง นิวร์อลเน็ตเวิร์ก และแบบจำลองฮิดเดนมาร์คอฟ จากงานวิจัยที่ผ่านมาแสดงให้เห็นว่าวิธีการแบบจำลองฮิดเดนมาร์คอฟ เป็นวิธีการที่มีความยืดหยุ่นสูงกว่าวิธีการทั้งสอง [4] ดังนั้นในวิทยานิพนธ์นี้เราจะใช้แบบจำลอง ฮิดเดนมาร์คอฟ เพื่อการจำแนกรูปแบบ มีรายละเอียดดังนี้

2.3.1 ทฤษฎีแบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model : HMM)

แบบจำลองฮิดเดนมาร์คอฟเป็นแบบจำลองทางสถิติซึ่งพัฒนามาเพื่อแบ่งกลุ่มของอนุกรมทางเวลา หรือสัญญาณที่ไม่คงที่ นั่นคือใช้สำหรับจัดกลุ่มของสัญญาณที่ไม่รู้จัก (Unknown signal) ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่งของสัญญาณ นำมาประยุกต์ใช้ในการรู้จำเสียงพูด

แบบจำลองฮิดเดนมาร์คอฟแบ่งออกเป็น 2 ประเภท คือ แบบต่อเนื่อง (Continuous) และแบบไม่ต่อเนื่อง (Discrete – time) ในการทำการวิจัยนี้ได้มีการใช้แบบจำลองแบบต่อเนื่อง แต่เนื้อ

หาตอนต้นจะอธิบายแบบไม่ต่อเนื่องเพราะง่ายต่อการเข้าใจ ในตอนท้ายของบทจะกล่าวถึงระบบการรู้จำแบบคำโดด (Isolated Word Recognition) ที่ใช้ในการทดลองในวิทยานิพนธ์ ซึ่งการสร้างแบบจำลองขึ้นมาเป็นตัวแทนข้อมูลเสียงนั้นจะใช้การสร้างแบบจำลองหนึ่งแบบแทนคำพูด 1 คำ

2.3.1.1 ส่วนประกอบของแบบจำลอง Markov

พารามิเตอร์สำคัญที่เกี่ยวข้องในการสร้างแบบจำลองอ้างอิง ที่ต้องรู้จักได้แก่

1. T คือ ความยาวของลำดับข้อมูล ซึ่งมีขนาดความยาวของลำดับเท่ากับจำนวนเฟรมทั้งหมดในเสียงแต่ละเสียง ซึ่งจะใช้เป็นข้อมูลอินพุทในส่วนของ HMM โดยต่อไปจะเรียกแทนว่า "ลำดับของค่าปรากฏ" (Observation sequence)
2. N คือ จำนวน state ในแบบจำลอง ถ้ากำหนดให้เซตของ state เป็น $\{ 1 , 2 , \dots , N \}$ จะสามารถแทน state ที่เปลี่ยนไปตามเวลา t ด้วย เซตของ $Q = \{ q_1 , q_2 , \dots , q_N \}$
3. M คือ จำนวนของค่าปรากฏที่สามารถเป็นไปได้ต่อหนึ่ง state แทนสัญลักษณ์ด้วย $V = \{ v_1 , v_2 , \dots , v_M \}$
4. ค่าความน่าจะเป็นในการย้าย state : $A = \{ a_{ij} \}$

   โดย a_{ij} แทนการย้าย state จาก i ไป j เมื่อ

   $$a_{ij} = P [q_t = j | q_{t-1} = i] \quad ; 1 \le i , j \le N \qquad (2.20)$$

5. การกระจายความน่าจะเป็น ของค่าปรากฏที่สามารถเป็นไปได้ภายใน state : $B = \{ b_j (k) \}$

   $$b_j (k) = P [v_k \text{ ที่เวลา } t] \; q_j \text{ ที่เวลา } t] \quad ; 1 \le k \le M \qquad (2.21)$$

   เป็นนิยามการกระจายใน state j เมื่อ $j = 1 , 2 , \dots , N$
6. ค่าความน่าจะเป็นของการเป็น state เริ่มต้น : $\pi = \{ \pi_i \}$

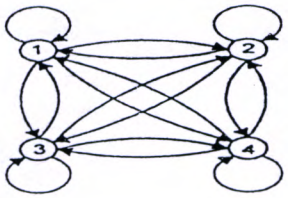
เมื่อ $\pi_i = P[q_i \text{ ที่เวลา } t = 1] \quad ; 1 \le i \le N \qquad (2.22)$

จะเห็นว่าแบบจำลองฮิดเดนมาร์คอฟ ต้องการพารามิเตอร์ของแบบจำลองคือ กลุ่มของความน่าจะเป็น A,B,π ดังนั้นในการแสดงเซตของพารามิเตอร์ที่สมบูรณ์ของแบบจำลองอ้างอิง จะแทนด้วยสัญลักษณ์

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

2.3.1.2 ชนิดของ HMM

แบ่งชนิดตามการย้าย state ของเมตริกซ์ A



(ก)

(ข)

(ค)

รูปที่ 2.11 แบบจำลองชนิดต่าง ๆ ของ HMM

HMM แบบ Egordic Model หรือ Fully Connected Model

การย้าย state สามารถย้ายไปยังทุก ๆ state ของแบบจำลอง ดังรูปที่ 2.11 (ก) เป็นตัวอย่างของ
แบบจำลองที่มี N = 4 ซึ่งจากรูปนี้มีค่าของเมตริกซ์ A เป็น

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

HMM แบบ Left – Right Model หรือ Bakis Model

การย้าย state จะย้ายจากซ้ายไปขวา ซึ่งจะมีคุณสมบัติของสัมประสิทธิ์ในการย้าย state ดังนี้

$$a_{ij} = 0 \quad , j < i$$

คือจะไม่มีการย้าย state ไปยัง state ที่ต่ำกว่า state ปัจจุบัน และนอกจากนี้ก็ยังมีความน่าจะเป็น

ของ state เริ่มต้นดังนี้

$$\pi_i = \begin{cases} 0 & ,i \neq 1 \\ 1 & ,i = 1 \end{cases}$$

คือลำดับของ state จะต้องเริ่มที่ state ที่ 1 เสมอ และ Left – Right Model นี้มักมีกฎบังคับการย้าย state เพื่อไม่ให้มีการเปลี่ยนแปลงดัชนีของ state มากนัก กล่าวคือ

$$a_{ij} = 0 \quad ,j > i + \Delta i$$

ดังรูปที่ 2.11(ข) ค่าของ $\Delta i = 2$ คือจะไม่มีการย้ายข้าม state ไปเกิน 2 state และมีเมตริกซ์ในการย้าย state เป็น

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

จะเห็นว่า state สุดท้าย สัมประสิทธิ์การย้าย state จะเป็น

$$a_{NN} = 1$$

$$a_{Ni} = 0 \quad ,i < N$$

แบบจำลองแบบนี้จะเหมาะกับสัญญาณที่มีลักษณะเปลี่ยนแปลงตามเวลาอย่างต่อเนื่อง เช่น เสียงพูด

HMM แบบ Parallel Left – Right Model

เป็นแบบจำลองที่มีความยืดหยุ่นมากกว่าแบบที่ 2 แสดงได้ดังรูปที่ 2.11(ค)

2.3.1.3 ปัญหาพื้นฐานของแบบจำลองฮิดเดนมาร์คอฟ

ปัญหาของ HMM มี 3 ข้อ ซึ่งต้องใช้ Algorithm วิธีต่างๆ ในการคำนวณเพื่อแก้ปัญหานี้

ปัญหาที่ 1 เมื่อมีลำดับของค่าปรากฏ $O = \{ O_1\ O_2\ O_3 \ldots O_T \}$ และมีแบบจำลอง $\lambda = (A,B,\pi)$ จะคำนวณหาค่าความน่าจะเป็น $P(O|\lambda)$ ของลำดับค่าปรากฏนั้นได้อย่างไร

ปัญหาที่ 2 เมื่อมีลำดับของค่าปรากฏ $O = \{ O_1\ O_2\ O_3 \ldots O_T \}$ และแบบจำลอง $\lambda = (A,B,\pi)$ จะคำนวณหาลำดับ state $q = \{ q_1\ q_2\ q_3 \ldots q_T \}$ ที่เหมาะสมกับลำดับค่าปรากฏนั้นได้อย่างไร

ปัญหาที่ 3 จะปรับพารามิเตอร์ของแบบจำลอง $\lambda = (A,B,\pi)$ เพื่อให้ได้ค่า $P(O|\lambda)$ สูงสุดได้อย่างไร

2.3.1.4 การคำนวณเพื่อแก้ปัญหาของ HMM

การแก้ปัญหาที่ 1

การแก้ปัญหาที่ 1 เป็นการคำนวณหาว่าแบบจำลอง λ ใด ๆ มีโอกาสจะให้ค่าลำดับเป็นไปตามลำดับของค่าปรากฏนั้น ด้วยค่าของความน่าจะเป็นมาก – น้อยเท่าใด

การแก้ปัญหาสามารถทำได้โดยระบุ state ให้กับลำดับของค่าปรากฏซึ่งยาว T (โดยที่ค่าปรากฏหนึ่งตัวมีความเป็นไปได้ที่จะอยู่ใน state ได้ N state) ซึ่งสามารถเป็นไปได้ถึง N^T แบบ ให้ state ต่าง ๆ แทนด้วย

$$q = q_1\, q_2\, q_3 \cdots q_T \tag{2.23}$$

เมื่อ q_1 เป็น state เริ่มต้นที่เวลา $t = 1$ ความน่าจะเป็นของลำดับของค่าปรากฏ O ที่กำหนดคือ

$$P(O|q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda) \tag{2.24a}$$

ความน่าจะเป็นในการเกิดค่าปรากฏคือ

$$P(O|q, \lambda) = b_{q_1} O_1 \cdot b_{q_2} O_2 \cdot \ldots \cdot b_{q_T} O_T \tag{2.24b}$$

และความน่าจะเป็นในการย้ายข้าม state q จะเป็น

$$P(q|\lambda) = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdot \ldots \cdot a_{q_{T-1} q_T} \tag{2.24c}$$

ดังนั้นเมื่อนำความน่าจะเป็นของการเกิดค่าปรากฏ O และค่าความน่าจะเป็นในการย้าย state q มารวมกัน ซึ่งนั่นก็คือความน่าจะเป็นที่ O และ q จะเกิดขึ้นพร้อมกัน จะได้

$$\begin{aligned} P(O, q|\lambda) &= P(O|q, \lambda)\, P(q|\lambda) \\ &= (b_{q_1} O_1 \cdot b_{q_2} O_2 \cdot \ldots \cdot b_{q_T} O_T)(\pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdot \ldots \cdot a_{q_{T-1} q_T}) \end{aligned} \tag{2.25}$$

จากสมการด้านบนนั้นจะเห็นได้ว่าในความเป็นจริงนั้นมีเพียงลำดับของ O เท่านั้นที่รู้แต่ลำดับของ state นั้นถูกซ่อนอยู่ ทำให้เป็นเหตุผลว่าทำไมถึงเรียกว่า Hidden Markov

โดยที่ความน่าจะเป็นของ O ได้มาจากผลรวมของความน่าจะเป็นที่ O และ q เกิดขึ้นพร้อมกัน โดยคิดจากทุก state q ที่จะเป็นไปได้ ดังนี้

$$P(O|\lambda) = \sum_{allq} P(O|q,\lambda)\,P(q|\lambda) \tag{2.26}$$

$$= \sum_{q_1 q_2 \ldots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_T O_T \tag{2.27}$$

ที่เวลาเริ่มต้น (t = 1) จะอยู่ที่ state q_1 ด้วยค่าความน่าจะเป็น π_{q1} และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏ O_1 ที่ state นี้ด้วย $b_{q_1} O_1$

ที่เวลา เพิ่มขึ้นจาก $t \longrightarrow t+1$ (t = 2) แทนการย้าย state จาก state q_1 ไปยัง q_2 ด้วยค่าความน่าจะเป็น $a_{q_1 q_2}$ และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏเป็น O_2 ด้วยค่าความน่าจะเป็น $b_{q_2} O_2$ จนกระทั่ง ที่เวลา T แทนการย้าย state จาก state q_{T-1} ไปยัง q_T ด้วยค่าความน่าจะเป็น $a_{q_{T-1} q_T}$ และแทนค่าความน่าจะเป็นในการเกิดค่าปรากฏเป็น OT ด้วยค่าความน่าจะเป็นb_{q_T} (O_T)

จะเห็นว่าสมการนี้มีการคำนวณที่ยุ่งยากเนื่องจากมีการคูณกันเป็นจำนวนมากในรูปของลำดับ $2T*N^T$ ดังนั้นจึงมีการคิดหาวิธีมาช่วย ซึ่งแบ่งออกเป็น

กระบวนการไปข้างหน้า (Forward Procedure) ; $\alpha_t(i)$ = Forward variable

นิยาม

$$\alpha_t(i) = P(O_1 O_2 \ldots O_T, q_t = i|\lambda) \tag{2.28}$$

คือ ความน่าจะเป็นของการเกิดลำดับของค่าปรากฏ $O_1\ O_2 \ldots O_T$ และอยู่ที่ state q_i ณ เวลา t โดยมีแบบจำลองเป็น λ แล้วสามารถหา $\alpha_t(i)$ ได้ดังนี้

1. การเริ่มต้น (Initialization)

$$\alpha_t(i) = \pi_i b_i(O_1) \ ; \ 1 \le i \le N \tag{2.29}$$

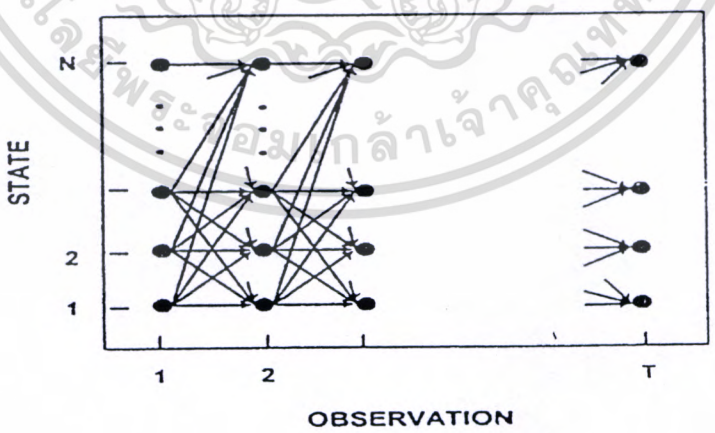
เริ่มด้วยการกำหนดความน่าจะเป็นไปข้างหน้าซึ่งเป็นความน่าจะเป็นร่วมของ state i และมีเหตุการณ์เริ่มต้นเป็น O_1

2. การเหนี่ยวนำ (Induction)

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad ; \quad \begin{matrix} 1 \le t \le T-1 \\ 1 \le j \le N \end{matrix} \tag{2.30}$$

หมายความว่า state j ที่เวลา t + 1 สามารถมาได้จาก state ก่อนหน้านี้ซึ่งเป็นไปได้ถึง N state (state i ณ เวลา t โดยที่ $1 \le i \le N$) ดังรูป 2.12(ก)

(ก)



(ข)

รูปที่ 2.12 กระบวนการไปข้างหน้า

จากรูป 2.12(ข) แสดงให้เห็นว่าการคำนวณค่าความน่าจะเป็นแบบไปข้างหน้า (Forward probability) มีโครงสร้างการคำนวณคล้าย ๆ ลักษณะของโครงผลึก และเนื่องจากมีจำนวน state เพียง N state (แทนด้วยจำนวน nodeในแต่ละช่วงเวลา t ใด ๆ ในโครงผลึก) จำนวนลำดับ state จะถูกจัดเรียงลงใน node เหล่านี้ โดยในเวลา t = 1 จะทำการคำนวณค่าของ α_t (i) ในทุก ๆ state $1 \le i \le N$ และที่เวลา t = 2 , 3 , ... , T จะทำการคำนวณค่าของ α_t (j) ในทุก ๆ state $1 \le j \le N$ โดยในแต่ละค่าจะทำการคำนวณมาจาก α_{t-1}(i) จำนวน N ค่าก่อนหน้านี้

3. การสิ้นสุด (Termination)

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad ; 1 \le i \le N \tag{2.31}$$

สามารถหา $P(O|\lambda)$ ได้จากผลรวมของ α_T (j) จากทุก ๆ state

กระบวนการย้อนกลับ (Backward Procedure) ; $\beta_t(i)$ = Backward variable

นิยาม

$$\beta_t(i) = P(O_{t+1}O_{t+2}...O_T | \; i_t = q_i, \lambda) \tag{2.32}$$

คือ ความน่าจะเป็นของลำดับค่าปรากฏส่วนหลังจากเวลา t+1 ไปจนจบโดยกำหนดว่าต้องอยู่ที่ state i ที่เวลา t และมีแบบจำลองเป็น λ จะคำนวณหา β_t (i) ได้ดังนี้

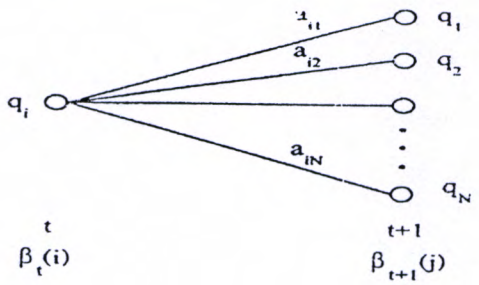
1. การเริ่มต้น (Initialization)

$$\beta_t(i) = 1 \qquad ; 1 \le i \le N \tag{2.33}$$

2. การเหนี่ยวนำ (Induction)

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{2.34}$$

เมื่อ t= T-1, T-2, ..., 1 ,$1 \le i \le N$



รูปที่ 2.13 กระบวนการย้อนกลับ

จากรูป 2.13 เพื่อที่จะให้ค่าปรากฏอยู่ที่ state i ณ เวลา t โดยคาดคะเนจากลำดับค่าปรากฏจากเวลา t+1 ซึ่งจะต้องพิจารณาจาก state j ที่เป็นไปได้ทั้งหมด โดยจะขึ้นอยู่กับค่า a_{ij} และ b_j (O_{t+1})

การแก้ปัญหาที่ 2 ใช้ Viterbi Algorithm

เพื่อที่จะหาลำดับ state ที่ดีที่สุด $q = (q_1\ q_2\ q_3 \dots q_T)$ ให้กับลำดับของค่าปรากฏ $O = \{O_1\ O_2\ O_3 \dots O_T\}$ ที่มีอยู่

โดยนิยามให้

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P\left[q_1 q_2 \dots q_{t-1}, q_t = i, O_1 O_2 \dots O_t \middle| \lambda\right] \qquad (2.35)$$

เมื่อ $\delta_t(i)$ คือ ความน่าจะเป็นสูงสุด (highest probability ของเส้นทาง (path) ซึ่งจะหาได้จากค่าความน่าจะเป็นสูงสุด เมื่อเทียบกับ state ทุก state ในการให้ค่าปรากฏเป็นไปตามค่าปรากฏที่กำหนดให้ ที่ขณะเวลา t ใด ๆ และจากการอาศัยคุณสมบัติของการเหนี่ยวนำจะได้

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \cdot b_j(O_{t+1}) \qquad (2.36)$$

โดยกำหนดให้ $\psi_t(j)$ เป็นอาเรย์ที่เก็บตำแหน่งของ state ที่ให้ค่าความน่าจะเป็นสูงที่สุดที่คำนวณได้ในแต่ละเวลา t และแต่ละลำดับ j ซึ่งจะสามารถหาลำดับ state ที่ดีที่สุดได้โดยใช้กระบวนการต่อไปนี้

1. การเริ่มต้น (Initialization)

$$\delta_1(i) = \pi_i b_i(O_1) \qquad ; 1 \le i \le N \tag{2.37a}$$

$$\psi_1(i) = 0 \tag{2.37b}$$

2. การย้อนกลับ (Recursion)

$$\delta_t(j) = [\max_{1 \le i \le N} \delta_{t-1}(i) a_{ij}] \cdot b_j(O_t) \quad ; \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{2.38a}$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] \quad ; \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{2.38b}$$

3. การสิ้นสุด (Termination)

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \tag{2.39a}$$

$$q_T = \arg \max_{1 \le i \le N} [\delta_T(i)] \tag{2.40b}$$

4. เส้นทางเดินย้อนกลับ (Backtracking)

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \qquad ; t = T-1, T-2, \ldots, 1 \tag{2.41}$$

การแก้ปัญหาที่ 3

จากที่กล่าวมาแล้วข้างต้นว่าแบบจำลองของเสียงจะแทนด้วยคุ่พารามิเตอร์ λ = (A,B,π) ดังนั้นเมื่อมีลำดับของค่าปรากฏจำนวนหนึ่ง เพื่อที่จะนำมาสร้างแบบจำลองอ้างอิงจะต้องทำการคำนวณหาค่าพารามิเตอร์ A, B, π ของแบบจำลองซึ่งจะอยู่ในรูปของค่าความน่าจะเป็น โดยวิธีที่เลือกใช้ก็คือ วิธีของ Baum-Welch method หรือเรียกอีกชื่อหนึ่งว่า EM (Expectation-Maximization method) โดยมี

นิยาม 1. คือ

$$\gamma_t(i) = P(q_t = i|O,\lambda) \tag{2.42}$$

เมื่อ $\gamma_t(i)$ คือ ค่าความน่าจะเป็นที่จะอยู่ที่ state i ที่ขณะเวลา t โดยให้ลำดับของค่าปรากฏด้วยโมเดล λ โดยที่กำหนดลำดับของค่าปรากฏให้ สามารถแสดงค่า $\gamma_t(i)$ ได้ดังนี้

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

$$= \frac{P(O,q_t = i|\lambda)}{P(O|\lambda)}$$

$$= \frac{P(O,q_t = i|\lambda)}{\sum_{i=1}^{N} P(O,q_t = i|\lambda)} \tag{2.43}$$

เนื่องจาก $P(O,q_t = i|\lambda)$ มีค่าเท่ากับ $\alpha_t(i)\beta_t(i)$ ดังนั้นสามารถเขียน $\gamma_t(i)$ ได้เป็น

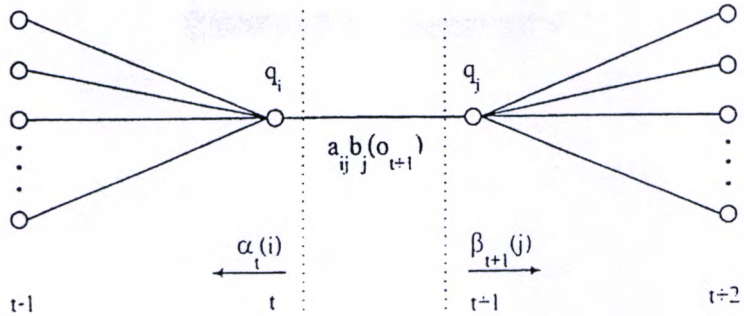
$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{2.44}$$

โดย $\alpha_t(i)$ เริ่มจาก $O_1, O_2, \ldots, O_t$ จนถึง state i ที่เวลา t

โดย $\beta_t(i)$ เริ่มจาก $O_{t+1}, O_{t+2}, \ldots, O_T$ จนถึง state $q_t = i$ ที่เวลา t

นิยาม 2. $$\varepsilon_t(i,j) = P(q_t = i, q_{t+1} = j|O,\lambda) \tag{2.45}$$

เมื่อ $\varepsilon_t(i,j)$ คือความน่าจะเป็นที่จะอยู่ที่ state i ที่เวลา t และ state j ที่เวลา t+1 เมื่อกำหนดแบบจำลองและลำดับค่าปรากฏให้



รูปที่ 2.14 ลำดับการคำนวณการเกิดค่าปรากฏร่วมซึ่งจะอยู่ที่ state i ที่เวลา t และอยู่ที่ state j ที่เวลา t-1

จากรูปแสดง ลำดับการคำนวณการเกิดค่าปรากฏร่วม ซึ่งระบบจะอยู่ใน state i ที่เวลา t และอยู่ที่ state j ที่เวลา t+1 โดย $\alpha_t(i)$ เริ่มจากเวลา t = 1 ที่ค่าปรากฏแรก จนถึง state q_i ที่เวลา t และ $a_{ij}b_jO_{t+1}$ เป็นการเปลี่ยน state ที่เวลา t ไปเป็น q_j ที่เวลา t+1 และให้ค่าปรากฏเป็น O_{t+1}

ซึ่งจากนิยามของตัวแปรไปข้างหน้า $\alpha_t(i)$ และตัวแปรย้อนกลับ $\beta_t(i)$ สามารถนำมาสัมพันธ์กับ $\varepsilon_t(i,j)$ ได้เป็น

$$\varepsilon_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \qquad (2.46)$$

จากที่ได้นิยาม $\gamma_t(i)$ แล้ว นำมาสัมพันธ์กับ $\varepsilon_t(i,j)$ ได้เป็น

$$\gamma_t(i) = \sum_{j=1}^{N}\varepsilon_t(i,j) \qquad (2.47)$$

เมื่อ

$\sum_{t=1}^{T-1}\gamma_t(i)$ = จำนวนของการย้าย state จาก state i ในลำดับค่าปรากฏ O (2.48a)

$\sum_{t=1}^{T-1}\varepsilon_t(i,j)$ = จำนวนของการย้าย state จาก state i ไป j ในลำดับค่าปรากฏ O (2.48b)

ดังนั้น สามารถคำนวณหาค่าของพารามิเตอร์ได้ดังนี้

π_i' = จำนวนครั้งในการอยู่ที่ state i ที่เวลา t = 1

$$\pi_i' = \gamma_1(i) \quad ;1 \le i \le N \tag{2.49a}$$

$$a_{ij}' = \frac{\text{จำนวนครั้งที่คาดไว้ของการย้ายสเตทจาก i ไป j}}{\text{จำนวนครั้งที่คาดว่าจะย้ายจากสเตท i}}$$

$$a_{ij}' = \frac{\sum_{t=1}^{T-1}\varepsilon_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{2.49b}$$

$$b_j'(k) = \frac{\text{จำนวนครั้งที่คาดว่าจะอยู่ในสเตท j และเกิดค่าปรากฏเป็น } V_K}{\text{จำนวนครั้งที่คาดว่าจะอยู่ที่สเตท j}}$$

$$b_j'(k) = \frac{\sum_{t=1,O_t=V_K}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \tag{2.49c}$$

จากกระบวนการข้างต้นถ้าให้ λ = (A,B,π) เป็นแบบจำลองปัจจุบัน และใช้ λ นี้คำนวณในด้านขวาของสมการที่ (2.49a-c) และให้แบบจำลองที่ได้จากการคำนวณซ้ำเป็น $\lambda' = (A', B', \pi')$ เป็นแบบจำลองที่ได้จากด้านซ้ายของสมการที่ (2.49a-c) ซึ่งจะได้จุดวิกฤตของฟังก์ชันความน่าจะเป็นในกรณีที่ $\lambda' = \lambda$ หรือถ้า λ' มีความน่าจะเป็นมากกว่าแบบจำลอง λ $\left[P(O|\lambda') > P(O|\lambda)\right]$ นั่นคือจะได้แบบจำลอง λ' ใหม่ ที่น่าจะทำให้เกิดลำดับของค่าปรากฏ O ที่ดีกว่า

2.3.1.5 ลำดับของค่าปรากฏหลายลำดับ (Multiple Observation Sequences)

ในการสร้างแบบจำลองด้วย Left-Right Model จำเป็นจะต้องใช้จำนวนลำดับของเหตุการณ์หลาย ๆ ลำดับเข้ามาเทรนเพื่อให้การประมาณค่าพารามิเตอร์ของแบบจำลองที่ได้มีความน่าเชื่อถือที่สุด ถ้ากำหนดให้ k แทน เซตของลำดับค่าปรากฏ ดังนี้

$$O = \left[O^{(1)}, O^{(2)}, ..., O^{(k)}\right] \tag{2.50}$$

เมื่อ $O^{(k)} = \left(O_1^{(k)} O_2^{(k)} ... O_T^{(k)}\right)$ คือ ลำดับค่าปรากฏอันดับที่ k โดยสมมติให้แต่ละอันดับของค่าปรากฏเป็นอิสระต่อกัน โดยมีจุดประสงค์ เพื่อที่จะปรับค่าพารามิเตอร์ของแบบจำลอง λ ให้มีค่าเหมาะสมมากที่สุด

2.3.1.6 แบบจำลองฮิดเดนมาร์คอฟแบบต่อเนื่อง (Continuous Density HMM)

มีความแตกต่างกับแบบจำลองแบบไม่ต่อเนื่องที่อธิบายไว้แล้วข้างต้น ตรงที่การหาการกระจายความน่าจะเป็น ของค่าปรากฏ $b_j(k)$ แบบจำลองแบบต่อเนื่องจะแสดงโดยใช้ ค่า mean (μ_j) และ covariance (Σ_j) ของแต่ละ state ดังนั้นพารามิเตอร์ของแต่ละแบบจำลองจะแสดง λ = (A,μ_j,Σ_j,π) แทน ดังนั้นการคำนวณซ้ำเพื่อให้พารามิเตอร์ที่เหมาะสมที่สุดของ μ_j และ Σ_j เป็นไปตามสมการดังนี้

$$\mu_j' = \frac{\sum_{t=1,}^{T} \gamma_t(j) * O_t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.51}$$

$$\Sigma_j' = \frac{\sum_{t=1,}^{T} \gamma_t(j) * (O_t - \mu_j)(O_t - \mu_j)'}{\sum_{t=1}^{T} \gamma_t(j)} \tag{2.52}$$

และการกระจายความน่าจะเป็น ของค่าปรากฏ $b_j(o_t)$ คือ

$$b_j(o_t) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_j\right|^{1/2}} \exp\left[-1/2(o_t - \mu_j)\Sigma_j(o_t - \mu_j)'\right] \tag{2.53}$$

D = ขนาดสัมประสิทธิในแต่ละเวกเตอร์ (สัมประสิทธิที่ได้จากการประมวลผลเบื้องต้น)

2.4 ขั้นตอนวิธีการตัดสินใจ (Decision Algorithm)

วิธีการตัดสินใจมีด้วยกันหลายวิธี แต่ในที่นี้เราจะกล่าวถึงวิธีของ Viterbi ซึ่งเป็นการแก้ปัญหาพื้นฐานข้อที่ 2 จัดเป็นขั้นตอนวิธีการตัดสินใจเลือกรูปแบบที่เหมาะสมที่สุดในการรู้จำ โดยมีขั้นตอนวิธีการดังแสดงไว้ในการแก้ปัญหาพื้นฐานข้อที่ 2 ที่กล่าวไว้แล้ว

บทที่ 3
การประมวลผลเบื้องต้น (Front Ends)

การประมวลผลเบื้องต้น (front ends) ก็คือการหาพารามิเตอร์ที่เป็นลักษณะเฉพาะของเสียงแต่ละคำพูด ไม่ขึ้นอยู่กับผู้พูดหรือการเปลี่ยนแปลงอุปกรณ์ใดๆ โดยมีเป้าหมายเพื่อลดจำนวนข้อมูลเสียง และลดการซ้ำซ้อนของข้อมูลเสียงพูดลง แล้วจึงนำข้อมูลเหล่านี้ไปเป็นตัวแทนเสียงพูดแต่ละคำ เข้าไปสู่ระบบการฝึกสอน และทดสอบการรู้จำด้วยแบบฮิดเดนมาร์คอฟ

3.1 การเตรียมข้อมูลเบื้องต้น (Preprocessing)

ในการประมวลผลเบื้องต้น (front ends) จะผ่านขั้นตอนนี้ก่อนทุกครั้ง เนื่องจากเราต้องทำการวิเคราะห์เสียงเป็นช่วงเวลาสั้นๆ เพื่อให้ข้อมูลเสียงพูดที่จะทำการวิเคราะห์ช่วงนั้นมีความเสถียรและไม่แปรเปลี่ยนตามเวลา

3.1.1 พรีเอมฟาซีส (pre-emphasis) เป็นขั้นตอนทำให้อัตราส่วนของสัญญาณต่อสัญญาณรบกวนมีค่าคงที่ ตลอดทุกช่วงความถี่ จากสมการที่ 2.1 เราแทนค่า a=0.95 [10]

3.1.2 การแบ่งช่วงสัญญาณ (frame blocking) สัญญาณที่ผ่านการพรีเอมฟาแล้ว s'(n) จะถูกตัดออกเป็นช่วงๆ หรือ เฟรม ช่วงละ N ข้อมูล (เราใช้ 256 ข้อมูล หรือ ประมาณ 32 ms) และแต่ละเฟรมจะมีส่วนที่เหลื่อมกัน 128 ข้อมูล หรือประมาณ 16 ms เพื่อให้ข้อมูลในการวิเคราะห์มีความต่อเนื่อง ตามสมการดังนี้

$$x_l(n) = s'(Ml + n) \tag{3.1}$$

โดยที่ n=1,2,3,.....,N เป็นจำนวนข้อมูลใน 1 เฟรม

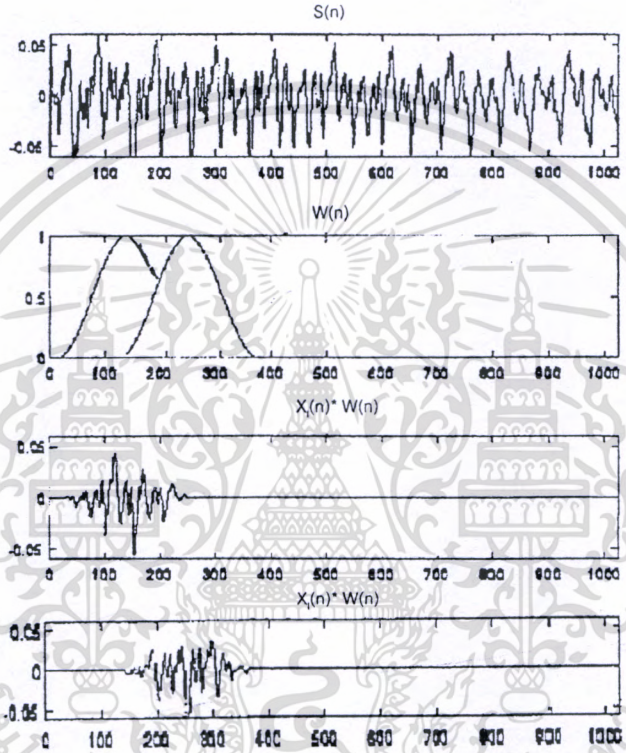
l =0,1,2,.....,L-1 เป็นจำนวนเฟรมทั้งหมดของสัญญาณเสียง

M=เป็นจำนวนข้อมูลที่ ใน 1 ส่วยย่อยไม่มีส่วนเหลื่อมของข้อมูล

3.1.3 การวินโดว์ (windowing)

เป็นการนำสัญญาณเสียงพูดแต่ละเฟรมมาผ่านฟังชั่น hamming window ซึ่งจะค่อยๆ ลดทอนแอมพลิจูดที่ปลายทั้งสองด้านของเฟรม ตามสมการที่ 2.2

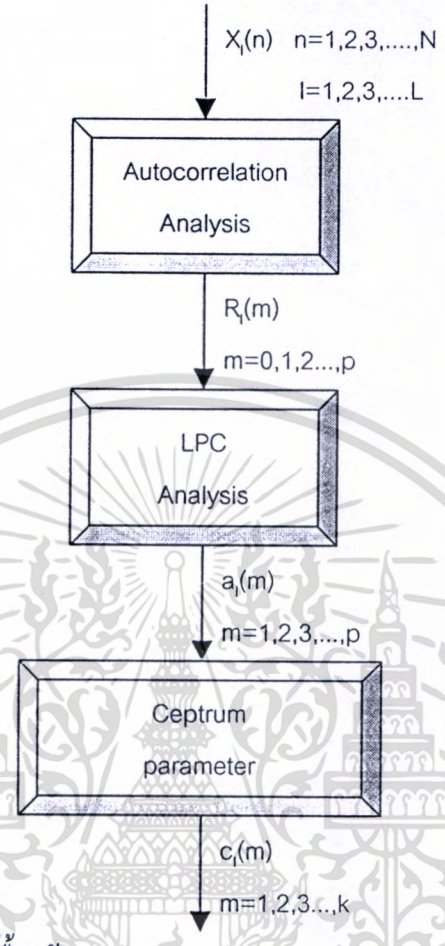
จากขั้นตอนทั้งหมดเราจะได้ข้อมูลที่อยู่ในเวกเตอร์ $X_l(n)$ โดยที่ n=1,2,.....,256 ตามจำนวนเฟรมที่ได้ l=1, 2, 3,.....,L เฟรม มีลักษณะดังรูปที่ 3.1



รูปที่ 3.1 การเตรียมข้อมูลเบื้องต้น (preprocessing)

3.2 การประมวลผลเบื้องต้นแบบ Linear predictive cepstrum coefficience (LPCC)

การประมาณพันธะเชิงเส้น (linear predictive coding : LPC) เป็นการสร้างแบบจำลองของสเปกตรัมอย่างง่ายโดยใช้โพล (all-pole spectrum modeling) เมื่อได้สเปกตรัมของสัญญาณเสียงจากสัมประสิทธิ์การประมาณเชิงเส้นแล้ว นำไปหาสัมประสิทธิ์เซ็ปสตรัม (cepstrum coefficience) เพื่อนำแต่ ส่วนที่บ่งบอกลักษณะการตอบสนองของทางเดินเสียง ขั้นตอนการประมวลผลเบื้องต้นแบบ LPCC มีลักษณะตามรูปที่ 3.2 มีรายละเอียดดังนี้



รูปที่ 3.2 ขั้นตอนการประมวลผลเบื้องต้นแบบ LPCC

3.2.1 การวิเคราะห์ค่าอัตสหสัมพันธ์ (autocorrelation analysis)

$$R_l(m) = \sum_{n=0}^{N-1-m} x_l(n)x_l(n+m) \quad ; \text{m=0,1,2,.....,p} \tag{3.2}$$

x(n) เป็นข้อมูลเสียงในแต่ละเฟรม

N เป็นจำนวนข้อมูลในแต่ละเฟรม (N=256)

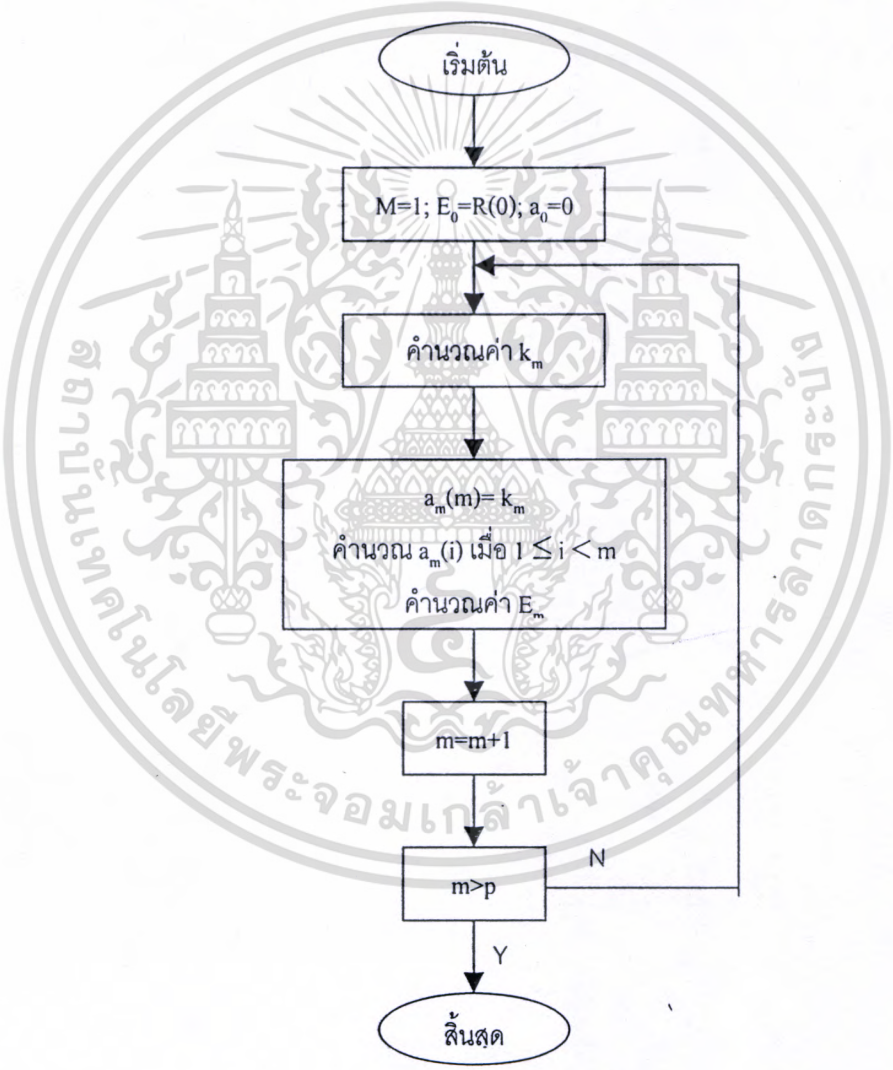
P เป็นลำดับการวิเคราะห์

3.2.2 การวิเคราะห์ค่าสัมประสิทธิ์ (LPC)

จากการหาค่าอัตสหสัมพันธ์ p เป็นลำดับ (order) ของการวิเคราะห์ระบบ โดยทั่วไปค่าลำดับที่ใช้ในการวิเคราะห์มีค่าอยู่ระหว่าง 8-18 เงื่อนไขในการเลือกค่าลำดับสำหรับการประมาณพันธะเชิงเส้นเป็นการปรับแต่งระหว่าง ความถูกต้องของสเปกตรัม เวลาในการคำนวณ และหน่วย

ความจำ ในระบบการรู้จำในวิทยานิพนธ์นี้เราจะใช้ ลำดับที่ 10 เป็นค่าที่เหมาะสำหรับการรู้จำเสียงพูดแบบคำโดด [4]

การหาค่าสัมประสิทธิ์ LPC เราใช้วิธีของ Levinson-Durbin ที่กล่าวไว้ในหัวข้อที่ 2.2.1 ตอนท้าย ตามสมการที่ 2.14 เป็นสูตรในการคำนวณ และรูปที่ 3.3 แสดงขั้นตอนของการคำนวณค่าสัมประสิทธิ์ LPC จะเริ่มคำนวณจากนำค่าอัตสหสัมพันธ์ (autocorrelation) ของแต่ละเฟรม ทำการคำนวณตามขั้นตอนรูปที่ 3.3 แล้วจะได้ค่าสัมประสิทธิ์ $a_m(1), a_m(2), a_m(3),\ldots\ldots, a_m(p)$ ซึ่งเป็นการประมาณพันธะเชิงเส้นที่เหมาะที่สุดสำหรับสัญญาณเสียงพูดแต่ละเฟรม ดังนั้นสัญญาณเสียงพูดแต่ละคำจะถูกแทนด้วย ชุดของสัมประสิทธิ์ LPC ขนาด p (p=10) จำนวน L ชุด



รูปที่ 3.3 ขั้นตอนการหาค่าสัมประสิทธิ์ LPC

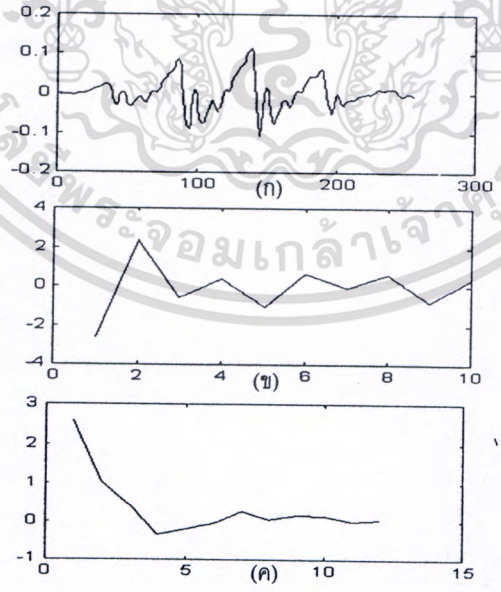
3.2.3 เปลี่ยนพารามิเตอร์ LPC เป็น สัมประสิทธิ์เซ็ปสตรัม (Cepstrum)

ในการรู้จำเสียงพูดนั้น สัมประสิทธิ์เซ็ปสตรัมนี้เป็นพารามิเตอร์ที่มีลักษณะน่าเชื่อถือได้ดีกว่าสัมประสิทธิ์ LPC ทั้งยังมีความสัมพันธ์ใกล้ชิดกับการรู้จำเสียง ตามความรู้สึกการได้ยินของมนุษย์ สัมประสิทธิ์เซ็ปสตรัมสามารถหาได้โดยตรงจากสัมประสิทธิ์ LPC ดังนี้

$$C_m = a_m + \sum_{k=1}^{m-1}(\frac{k}{m})C_k a_{\text{m-k}} \ldots\ldots, 1 \le m \le p \tag{3.3}$$

$$C_m = \sum_{k=1}^{m-1}(\frac{k}{m})C_k a_{\text{m-k}} \ldots\ldots\ldots, m > p \tag{3.4}$$

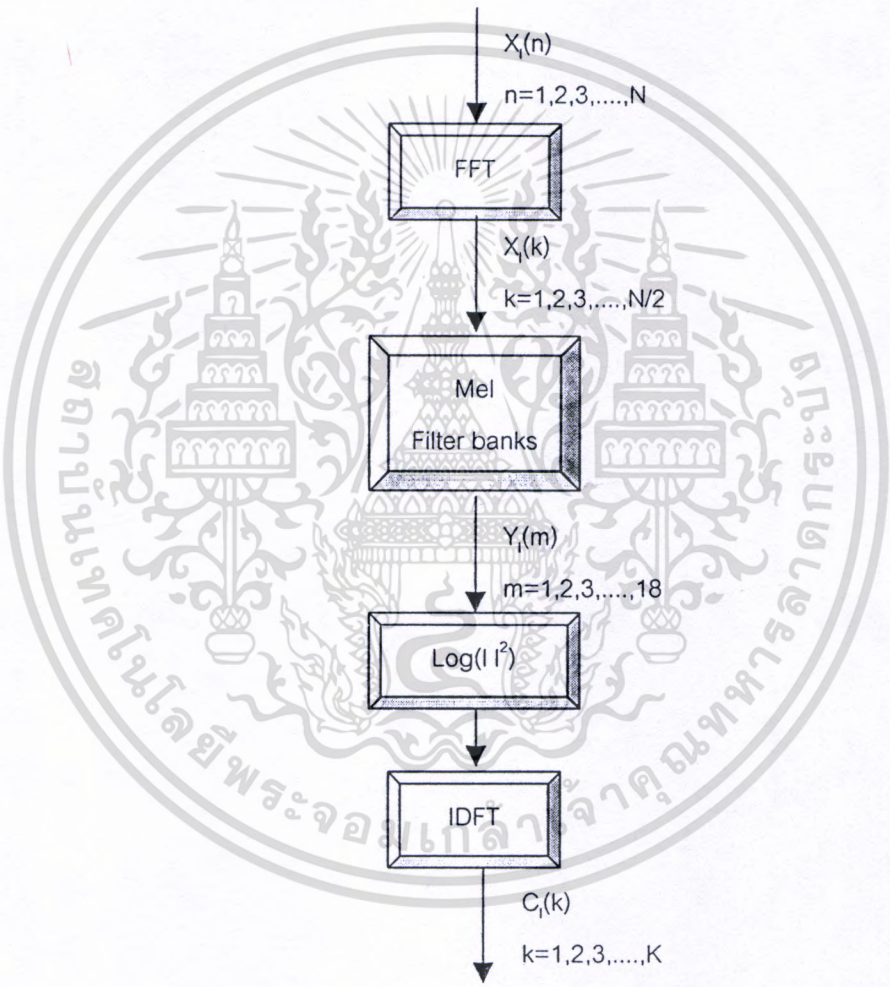
จากขั้นตอนทั้งหมดของการประมวลผลเบื้องต้นแบบ Linear predictive cepstrum coefficience (LPCC) เราจะได้ข้อมูลที่แสดงลักษณะเด่นของเสียงแต่ละเสียง และนำข้อมูลที่ได้เป็นอินพุทให้กับแบบจำลองฮิดเด็นมาร์คอฟ (HMM) โดยขนาดของข้อมูลทั้งหมดจะขึ้นอยู่กับจำนานสัมประสิทธิ์ในแต่ละเฟรม m และจำนวนเฟรมทั้งหมดของเสียง L รูปที่ 3.4 เป็นตัวอย่างของข้อมูลตามขั้นตอนที่ได้จากการประมวลผลเบื้องต้นแบบ LPCC ของเสียง “สอง” เฟรมที่ 20 รูปที่ 3.4(ก) แสดงข้อมูลเสียงที่ผ่านการเตรียมข้อมูลเบื้องต้น (ข) แสดงสัมประสิทธิ์ LPC (p=10) และ (ค) แสดงสัมประสิทธิ์เซ็ปสตรัม (m=12)



รูปที่ 3.4 เป็นตัวอย่างของข้อมูลตามขั้นตอนที่ได้จากการประมวลผลเบื้องต้นแบบ LPCC ของเสียง “ศูนย์” เฟรมที่ 20

3.3 การประมวลผลเบื้องต้นแบบ Mel Frequency Cepstral Coefficients (MFCC)

การประมวลผลเบื้องต้นแบบ MFCC เป็นการแทนสัญญาณเสียงพูดด้วยสเปกตรัม โดยใช้ Fast Fourier Transform (FFT) แล้วทำการปรับสเกลของความถี่แบบไม่เป็นเชิงเส้นโดยใช้วิธี Mel scale นำไปหาสัมประสิทธิ์เซ็ปสตรัม (cepstrum coefficience) เพื่อนำแต่ส่วนที่บ่งบอกลักษณะการตอบสนองของทางเดินเสียง ขั้นตอนการประมวลผลเบื้องต้นแบบ MFCC มีลักษณะตามรูปที่ 3.5 มีรายละเอียดดังนี้



รูปที่ 3.5 ขั้นตอนการประมวลผลเบื้องต้นแบบ MFCC

3.3.1 แปลงสัญญาณให้อยู่ในโดเมนความถี่

การเปลี่ยนสัญญาณจากโดเมนของเวลาให้อยู่ในรูปของโดเมนความถี่ ก็คือสเปกตรัมของสัญญาณเสียงนั้นเอง $x_l(n) \rightarrow x_l(k)$ ทำให้เราสามารถวิเคราะห์สัญญาณได้ดีขึ้น เพราะคุณสมบัติส่วนใหญ่จะแสดงอยู่ในรูปของสเปกตรัม ดังนั้นการแปลงข้อมูลโดยใช้ Discrete Fourier Transform (DFT) ของทุกเฟรม เป็นไปตามสมการที่ 2.15

เนื่องจากขั้นตอนการเตรียมข้อมูลเบื้องต้นหัวข้อที่ 3.1 เราแบ่งให้แต่ละเฟรมมีข้อมูลเท่ากับ N=256 ข้อมูล N มีค่าเท่ากับยกกำลังของ 2 ($N=2^p$ p เป็นเลขจำนวนเต็ม) ดังนั้นเพื่อให้การคำนวณเร็วขึ้นเราจะใช้การแปลงข้อมูลแบบ Fast Fourier Transform (FFT) แทน Discrete Fourier Transform (DFT) เมื่อผ่านการแปลงข้อมูลแล้วเราจะได้สัญญาณอยู่ในรูปสเปกตรัม นอกจากนั้นจำนวนข้อมูลลงเหลือ 128 ข้อมูล การแปลงฟูริเยร์ของสัญญาณเสียงก็คือ ผลลัพธ์ที่เกิดจากการกระตุ้นของเส้นเสียงกับการตอบสนองของทางเดินเสียงนั่นเอง

3.3.2 Mel Filter Banks

เป็นวิธีปรับสเกลของความถี่ (f) ให้มีความสัมพันธ์ใกล้ชิดกับระบบการรับรู้ของหูมนุษย์ เรียกว่า Mel scales โดยได้จากแบบจำลองการรับรู้ของหูมนุษย์ที่วิศวกรสร้างขึ้น

การปรับสเกล เป็นการปรับข้อมูลสเปกตรัมที่ได้จากหัวข้อที่ผ่านมา โดยพ่านตัวกรองสัญญาณ จากการทดลองของ Holmes filter bank [10] ซึ่งได้รับการยอบรับและใช้กันมากในระบบการรู้จำเสียงได้กำหนดให้มีตัวกรองสัญญาณ 18 ช่อง วิเคราะห์ความถี่ช่วง 0-4000 Hz สำหรับสัญญาณที่มีอัตราการชักตัวอย่างเท่ากับ 8 KHz ลักษณะของตัวกรองสัญญาณมีคุณสมบัติตามตารางที่ 3.1

เรากำหนดให้การกรองสัญญาณเป็นแบบ triangular window $U_{\Delta_m}(k)$ ตามสมการที่ 3.6 เมื่อ k คือตำแหน่งของความถี่ และ $2\Delta_m$ เป็นขนาดช่วงความถี่ผ่าน

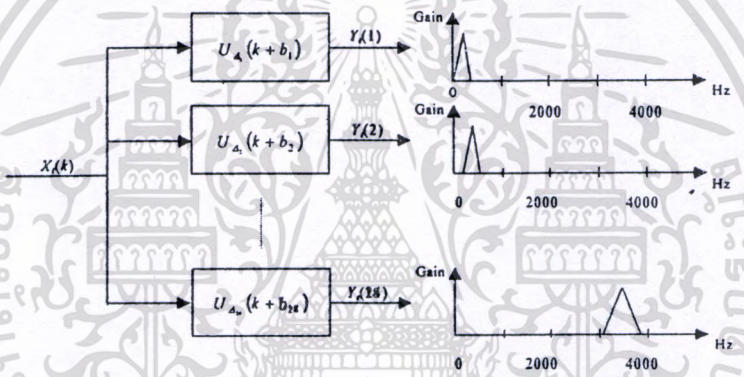
ตารางที่ 3.1 Mel และ Bark scales โดย Holmes

Channel number	Mel		Bark	
	Centre Frequency/Hz	Bandwidth	Centre Frequency/Hz	Bandwidth
1	240	120	50	100
2	360	120	150	100
3	480	120	250	100
4	600	120	350	100
5	720	120	450	110
6	840	120	570	120
7	1000	150	700	140
8	1150	150	840	150
9	1300	150	1000	160
10	1450	150	1170	190
11	1600	150	1370	210
12	1800	200	1600	240
13	2000	200	1850	280
14	2200	200	2150	320
15	2400	200	2500	380
16	2700	300	2900	450
17	3000	300	3400	550
18	3300	300	4000	700
19	3800	500	4800	900

$$U_{\Delta_m}(k) = \begin{cases} 1 - |k| / \Delta_m \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots |k| < \Delta_m \\ 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots |k| \geq \Delta_m \end{cases} \quad (3.6)$$

สมการที่ 3.7 เป็นการกระทำระหว่างสเปกตรัมของสัญญาณเสียงที่ได้จาก FFT คือ $x_l(k)$ นำไปผ่านตัวกรองสัญญาณตามสมการ 3.6 โดยมีคุณสมบัติของตัวกรองตามตารางที่ 3.1 กำหนดให้ b_m คือ centre frequency เราจะได้ข้อมูลของสเปกตรัมที่เปลี่ยนแปลงตามสเกลของ Mel คือ $Y_l(m)$ โดย m=1,2,3…,18 รูปที่ 3.6 แสดงการนำ $x_l(k)$ ผ่านตัวกรองสัญญาณ

$$Y_l(m) = \sum_{k=b_m-\Delta_m}^{b_m+\Delta_m} x_l(k) U_{\Delta_m}(k + b_m) \quad (3.7)$$



รูปที่ 3.6 การทำ Mel Filter banks

3.3.3 การคำนวณ log energy

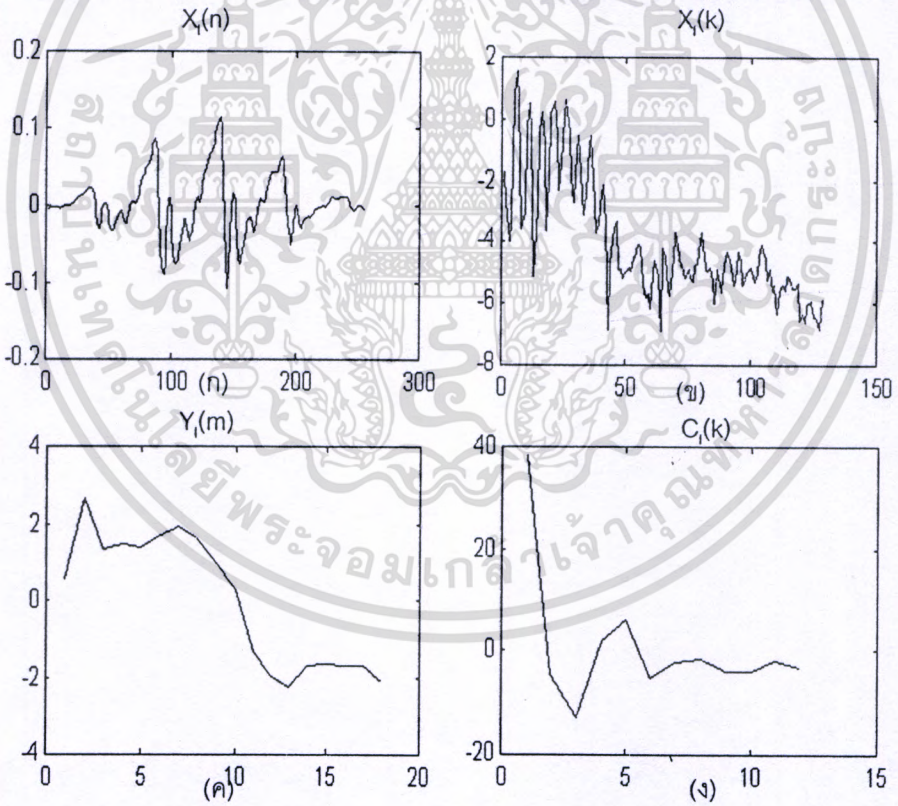
จากขั้นตอนก่อนหน้านี้เป็นการปรับสเกล หรือ เป็นการทำ smoothing สเปกตรัมของสัญญาณเสียงนั้นเอง โดยทำให้มีความสัมพันธ์ใกล้ชิดมากขึ้นกับระบบรับรู้สัญญาณเสียงของมนุษย์ ในขั้นตอนนี้จะเป็นการนำ logarithm ไปกระทำกับ ค่าพลังงานของสเปกตรัม ($Y_l(m)$) ที่ได้จากสมการที่ 3.7 ดังนี้ $\log(|Y_l(m)|^2)$ การคำนวณขนาดของสัญญาณก็เป็นการตัดส่วนที่เป็นเฟส (phase) เพราะในระบบการรู้จำเสียงเราจะทำการวิเคราะห์สัญญาณโดยไม่คำนึงถึงเฟส และยังเป็นการปรับระดับความดังของเสียงให้อยู่ในหน่วยที่มนุษย์เข้าใจ หลังการคูณด้วยค่า log แล้วนำข้อมูลที่ได้ไปหาค่าเซ็ปสตรัม (ceptrum) จะอธิบายในหัวข้อต่อไป

3.3.4 การคำนวณเซ็ปสตรัม (ceptrum)

ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการประมวลผลเบื้องต้นแบบ MFCC โดยนำ $\log(|Y_i(m)|^2)$ ไปผ่าน Inverse discread fourier transform (IDFT) ตามสมการที่ 3.8

$$C_i(k) = \sum_{m=1}^{M} \left(\log(|Y_i(m)|^2)\right) * \cos\left(k(m-1/2)\frac{\pi}{M}\right), \quad k = 0,1,2,\ldots,K \qquad (3.8)$$

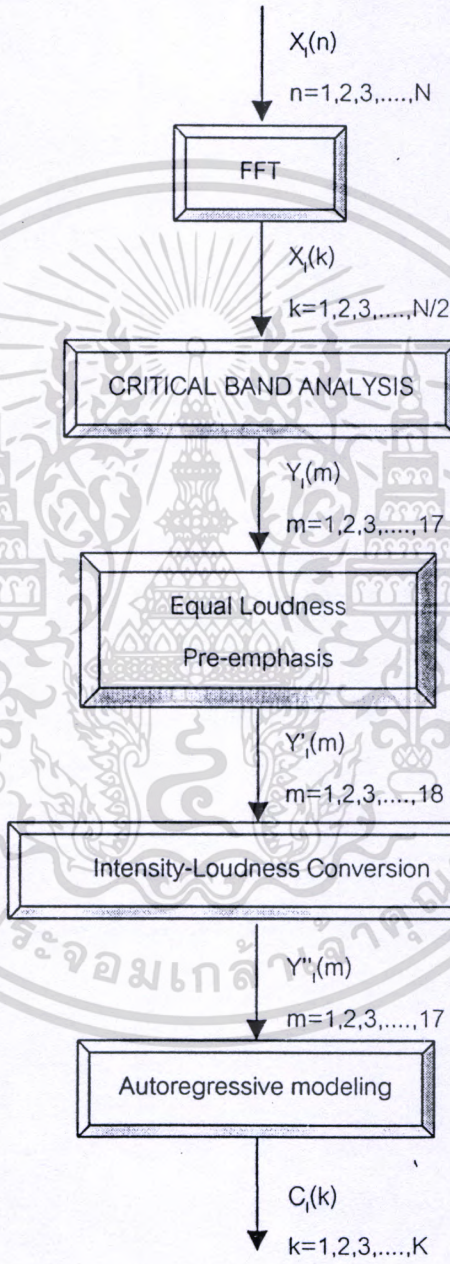
ผลลัพธ์ที่ได้จากสมการที่ 3.8 จะเป็นสัมประสิทธิ์เซ็ปสตรัม (ceptrum coefficience) ใช้สำหรับเป็นพารามิเตอร์เป็นตัวแทนของสัญญาณเสียง โดยที่ข้อมูลแต่ละเฟรมจะขึ้นอยู่กับขนาดสัมประสิทธิ์ (k) ที่จะใช้เป็นอินพุตเข้าสู่ระบบการของแบบจำลองฮิดเดนมาร์คอฟ รูปที่ 3.7 แสดงตัวอย่างข้อมูลของเสียง "สอง" เฟรมที่ 20 รูปที่ 3.7(ก) เป็นข้อมูลที่ได้จากการเตรียมข้อมูลเบื้องต้น 3.7(ข) แสดงค่า log(สเปกตรัมของสัญญาณเสียงพูด) ของข้อมูลที่ได้หลังจาก FFT 3.7(ค) ค่าสเปกตรัมที่ผ่านการ smoothing โดย Mel Filter banks 3.7(ง) สัมประสิทธิ์เซ็ปสตรัม (k=12)



รูปที่ 3.7 ตัวอย่างตามขั้นตอนการประมวลผลเบื้องต้นแบบ MFCC

3.4 การประมวลผลเบื้องต้นแบบ Perceptual Linear Predictive (PLP)

การประมวลผลเบื้องต้นแบบนี้ เป็นวิธีของ Hermansky [9] ที่ใช้หาคุณลักษณะของเสียง โดยทำให้เสียงมีความใกล้ชิดกับระบบการรับรู้ของหูมนุษย์มากกว่าวิธีการที่กล่าวมาข้างต้น จากผลการทดลองตามเอกสารอ้างอิงที่ [9] แสดงให้เห็นว่าการประมวลผลเบื้องต้นแบบ PLP ใช้ได้ดีกับระบบการรู้จำเสียงที่เรียกว่า Robuse Speech Recognition โดยมีโครงสร้างการคำนวณตามรูปที่ 3.8



รูปที่ 3.8 ขั้นตอนการประมวลผลเบื้องต้นแบบ PLP

3.3.1 แปลงสัญญาณให้อยู่ในโดเมนความถี่

เป็นขั้นตอนการแปลงสัญญาณเสียงพูดจากโดเมนเวลา ให้อยู่ในรูปโดเมนความถี่ โดยใช้ FFT จะได้ $X_t(k)$

3.3.2 Critical band analysis

เมื่อได้ค่าพลังงานของสเปกตรัม $P_t(k) = |X_t(k)|^2$ แล้วทำการปรับข้อมูลให้มีความราบเรียบ (smoothing) ขึ้นโดยใช้หลักของ Bark Filter Bank คล้ายกับการทำ Mel Filter Bank ตามที่กล่าวมาแล้ว ตามตารางที่ 3.1 แสดงคุณสมบัติของตัวกรองสัญญาณซึ่งได้จากการทดสอบกับระบบการรับรู้ของหูมนุษย์ของ Holmes [9] เราจะใช้ตัวกรองทั้งหมด 17 ช่อง สำหรับวิเคราะห์ช่วงความถี่ 0-4000 Hz

3.3.3 Equal-loudness Pre-emphasis

$$E(m) = \frac{(m^2 + 56.8 \times 10^6) m^4}{(m^2 + 6.3 \times 10^6)(m^2 + 0.38 \times 10^9)} \tag{3.9}$$

สมการที่ 3.9 เป็นการทำให้มีระดับความดัง 12 dB/oct ที่ช่วงความถี่ 0-300 Hz, 0 dB/oct ที่ช่วงความถี่ 300-1200 Hz, 6 dB/oct ที่ช่วงความถี่ 1200-3100 Hz และ 0 dB/oct ที่ช่วงความถี่ 3100-3000 Hz ขั้นตอนนี้มีจุดประสงค์เพื่อปรับระดับความดังของเสียงให้มีขนาด 40 dB ตลอดทุกช่วงความถี่ (เป็นระดับความดังที่หูมนุษย์ได้ยินชัดเจน [9]) การปรับมีสมการดังนี้

$$Y_t'(m) = Y_t(m) E(m) \tag{3.10}$$

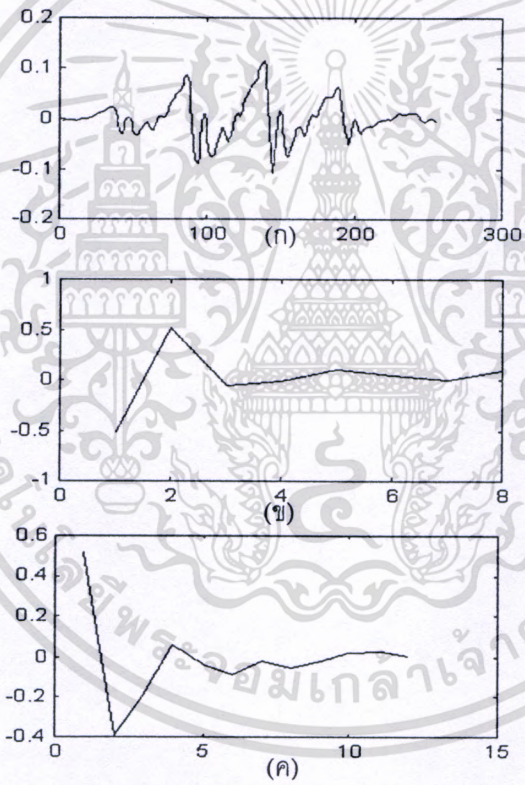
3.3.4 Intensity-loudness power-law

เป็นวิธีการปรับให้ความสัมพันธ์ระหว่าง ความเข้มของเสียงกับเสียงที่ได้รับโดยคุณลักษณะการรับรู้ของหู มีลักษณะไม่เป็นเชิงเส้น ซึ่งทำให้เสียงที่ได้รับมีความใกล้ชิดกับความรู้สึกกับหูมนุษย์มากขึ้น ตามสมการนี้

$$Y_t''(m) = (Y_t'(m))^{0.33} \tag{3.11}$$

3.3.5 Autoregressive modeling

นำข้อมูลที่ได้ไปผ่าน Inverse discread fourier transform (IDFT) กำหนดให้มีขนาด 34 ข้อมูล ก็คือเป็นการปรับข้อมูลให้มาอยู่ในโดเมนเวลา หลังจากนั้นผ่านขั้นตอนการประมาณเชิงเส้น คือ การวิเคราะห์ค่าอัตสหสัมพันธ์ (autocorrelation analysis) และหาค่าสัมประสิทธิ์ LPC โดยใช้การวิเคราห์ลำดับที่ p เท่ากับ 8 [9] สุดท้ายปรับข้อมูลให้อยู่ในรูปของเซ็ปสตรัม เพื่อใช้เป็นตัวแทนของเสียงแต่ละเฟรม เหมือนกับหัวข้อที่ 3.2.3 ที่กล่าวไว้แล้ว รูปที่ 3.9 จะแสดงข้อมูลที่ได้ตามขั้นตอนของการประมวลผลเบื้องต้นแบบ PLP ของเสียง "สอง" เฟรมที่ 20 รูปที่ 3.9(ก) เป็นข้อมูลที่ได้จากการเตรียมข้อมูลเบื้องต้น 3.9(ข) แสดงค่าของข้อมูลที่ได้หลังจากขั้นตอนการประมาณเชิงเส้น (p=8) 3.9(ค) สัมประสิทธิ์เซ็ปสตรัม (k=12)



รูปที่ 3.9 เป็นตัวอย่างของข้อมูลตามขั้นตอนที่ได้จากการประมวลผลเบื้องต้นแบบ PLP ของเสียง "ศูนย์"

บทที่ 4
การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการทดลองและแสดงผลการทดลองใช้แบบจำลองฮิดเดนมาร์คอฟ (HMM) ทำการแบ่งแยกคำพูดแต่ละคำออกจากกัน โดยมุ่งเน้นการเปรียบเทียบประสิทธิภาพของการประมวลผลเบื้องต้นแบบต่างๆ

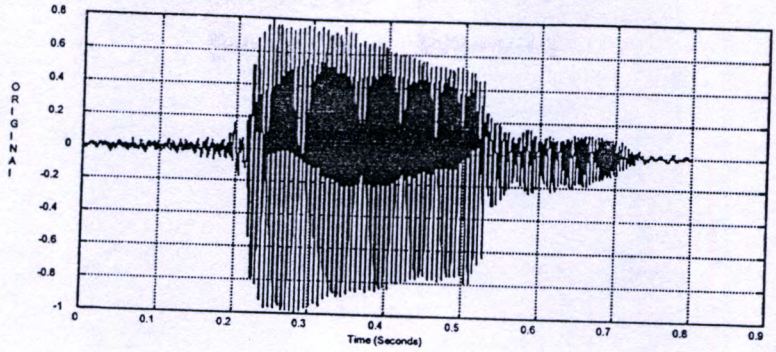
4.1 ข้อกำหนดในการทดลอง

ในการทดลองนี้จะมีข้อกำหนดดังนี้

4.1.1 ลักษณะของข้อมูลเสียงพูด

สัญญาณเสียงพูดที่ใช้ในวิทยานิพนธ์นี้เป็นคำพยางค์เดี่ยว หรือคำโดดที่ได้จากการเก็บตัวอย่างข้อมูลเสียง โดยใช้เครื่องคอมพิวเตอร์ส่วนบุคคล และการ์ดเสียง ซึ่งข้อมูลจะถูกเก็บอยู่ในรูปไฟล์ '.wav' ข้อมูล 1 ตัวอย่างเสียงจะถูกแทนด้วยข้อมูลขนาด 16 บิท โดยใช้ความถี่ในการซักตัวอย่างเท่ากับ 8 KHz รูปที่ 4.1 แสดงลักษณะของสัญญาณเสียงพูดที่ใช้ในการทดลอง

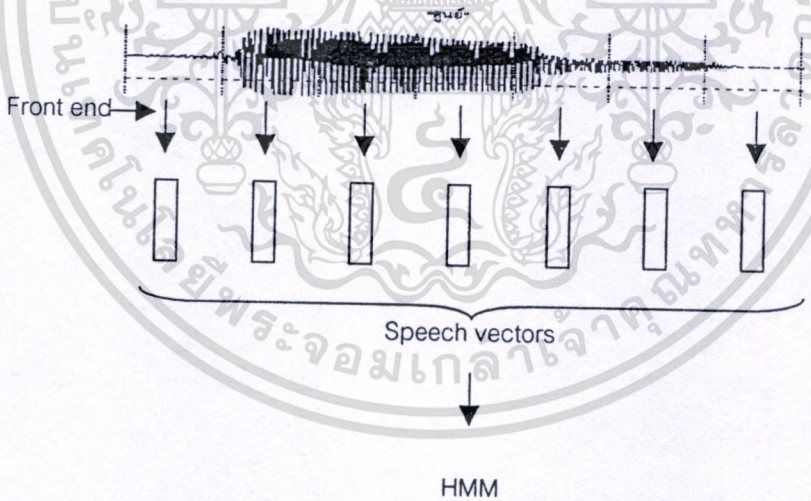
ในระบบของเราจะสามารถทำการรู้จำได้ทั้งหมด 20 คำ เป็นคำที่มีความจำเป็นในการควบคุมอุปกรณ์แล้วแต่จะนำไปประยุกต์ใช้ คือ "ศูนย์", "หนึ่ง", "สอง", "สาม", "สี่", "ห้า", "หก", "เจ็ด", "แปด", "เก้า", "เปิด", "ปิด", "ยก", "วาง", "ซ้าย", "ขวา", "หน้า", "หลัง", "ไป" และ "มา" ตัวอย่างเสียงที่ใช้ในการทดลองนี้ สำหรับนำไปฝึกสอน มีทั้งหมด 900 เสียง ได้จากเพศชาย 15 คน แต่ละคนพูด 3 ครั้ง ต่อเสียง 1 คำ และสำหรับนำไปทดสอบ มีทั้งหมด 300 เสียง ได้จากเพศชาย 5 คน (ผู้พูดคนละกลุ่ม) แต่ละคนพูด 3 ครั้ง ต่อเสียง 1 คำ เช่นเดียวกัน ทำการเก็บตัวอย่างเสียงในห้องทดลองโดยควบคุมไม่ให้มีสัญญาณรบกวนเกิดขึ้น (clean speech)



รูปที่ 4.1 ลักษณะของสัญญาณเสียง "ศูนย์"

4.1.2 การประมวลผลเบื้องต้น (Front Ends)

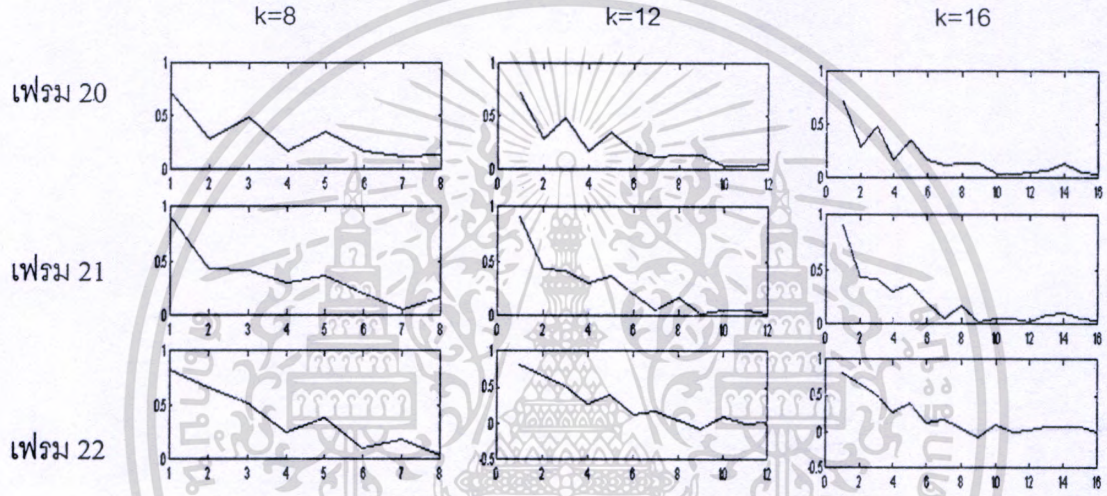
การประมวลผลเบื้องต้นจัดเป็นขั้นตอนในการลดจำนวนข้อมูลโดยการแสดงลักษณะของสัญญาณเสียงพูดด้วยพารามิเตอร์เพียงไม่กี่ค่าได้อย่างมีประสิทธิภาพ หาค่าสัมประสิทธิ์เฉพาะแต่ละเฟรมของข้อมูลเสียงพูดแทนด้วยเวกเตอร์ของสัมประสิทธิ์ เพื่อเป็นตัวแทนเข้าสู่ระบบการฝึกสอนและรู้จำ โดยใช้แบบจำลองฮิดเดนมาร์คอฟ จากรูปที่ 4.2 แสดงเวกเตอร์สัมประสิทธิ์แต่ละเฟรมของเสียงพูดที่ได้จากการประมวลผลเบื้องต้น เพื่อเป็นตัวแทนสัญญาณเสียงให้กับแบบจำลองฮิดเดนมาร์คอฟ ในวิทยานิพนธ์นี้เราใช้การประมวลผลเบื้องต้นทั้งหมด 3 วิธีดังนี้



รูปที่ 4.2 แสดงเวกเตอร์สัมประสิทธิ์ของแต่ละเฟรมใช้เป็นตัวแทนเข้าสู่ระบบการฝึกสอนและรู้จำ

4.1.2.1 การประมวลผลเบื้องต้นแบบ LPCC

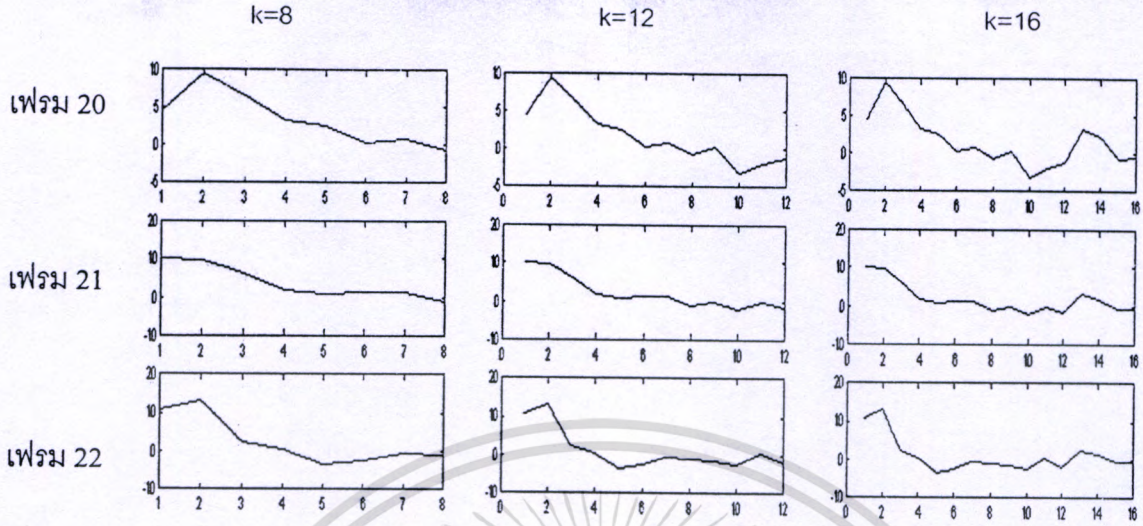
สำหรับในวิทยานิพนธ์นี้จะอาศัยวิธีการทางอัตสหสัมพันธ์ และวิธีการของ Levinson-Durbin ในการหาค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นสำหรับแต่ละเฟรมของเสียงพูด โดยเราจะใช้การวิเคราะห์ลำดับของสัมประสิทธิ์ที่ 10 เนื่องจากเป็นลำดับที่ใช้เป็นตัวแทนเสียงพูดที่ดีที่สุด [5] นำค่าสัมประสิทธิ์ของการประมาณพันธะเชิงเส้นหาค่าสัมประสิทธิ์เซ็ปสตรัม (Cepstrum coefficient) กำหนดให้มีขนาดเท่ากับ 8, 12 และ 16 ที่ใช้ในการทดลอง รูปที่ 4.3 แสดงตัวอย่างของสัมประสิทธิ์เซ็ปสตรัมของเสียง "ศูนย์" เฟรมที่ 20, 21 และ 22 มีขนาดของสัมประสิทธิ์เซ็ปสตรัมแต่ละเฟรมเท่ากับ 8, 12 และ 16 ตามลำดับ



รูปที่ 4.3 แสดงตัวอย่างสัมประสิทธิ์เซ็ปสตรัม LPCC ที่มีจำนวนข้อมูลขนาดต่างๆ

4.1.2.2 การประมวลผลเบื้องต้นแบบ MFCC

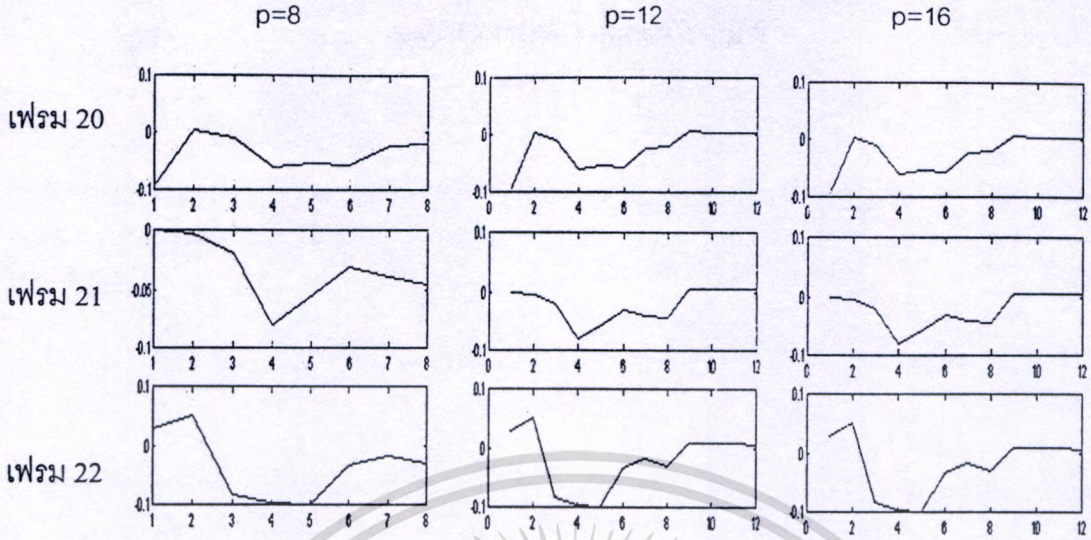
นำข้อมูลเสียงพูดแต่ละเฟรมผ่าน Fast Foureir Transfrom (FFT) จะได้สเปกตรัมของแต่ละเฟรมนำไปผ่าน Mel filter bank ซึ่งเป็นการปรับสเปกตรัมของเสียงพูดให้มีความราบเรียบ (smooting) เนื่องจากค่าสเปกตรัมอยู่ในช่วงความถี่ 0-4000 Hz (อัตราการชักตัวอย่าง = 8,000 Hz) ดังนั้นจะใช้ตัวกรองสัญญาณ 18 ช่อง (ตามตารางที่ 3.1) จะได้เมลสเปกตัม (Mel spectrum) ของเสียงพูดแต่ละเฟรมมีจำนวนข้อมูลเท่ากับ 18 นำไปหาสัมประสิทธิ์เซ็ปสตุรัม ในวิทยานิพนธ์นี้เราจะทดลองใช้จำนวนสัมประสิทธิ์เซ็ปสตรัมเท่ากับ 8, 12 และ 16 รูปที่ 4.4 เป็นตัวอย่างของสัมประสิทธิ์เซ็ปสตรัมของเสียง "ศูนย์" เฟรมที่ 20, 21 และ 22 มขนาดของสัมประสิทธิ์เซ็ปสตรัมแต่ละเฟรมเท่ากับ 8, 12 และ 16 ตามลำดับ



รูปที่ 4.4 แสดงตัวอย่างสัมประสิทธิ์เซ็ปสตรัม MFCC ที่มีจำนวนข้อมูลขนาดต่างๆ

4.1.2.3 การประมวลผลเบื้องต้นแบบ PLP

การประมวลผลเบื้องต้นแบบนี้พยายามที่จะทำการวิเคราะห์เสียงพูดให้ใกล้เคียงกับระบบการรับฟังของหูมนุษย์มากกว่า 2 วิธีที่กล่าวมาแล้ว โดยนำข้อมูลเสียงพูดแต่ละเฟรมผ่านการ Fast Foureir Transfrom จะได้สเปกตรัมของสัญญาณเสียงพูด นำไปผ่าน Bark filter bank จะได้ข้อมูลอยู่ในรูป บาร์คสเปกตรัม มีขนาดของสเปกตรัม แต่ละเฟรมเท่ากับ 17 (ตามตารางที่ 3.1 ช่วงความถี่ที่ทำการวิเคราะห์ 0-4000 Hz) นำข้อมูลที่ได้ผ่าน Equal-Loudness Pre-emphasis และ intensity-loudness power-law วิธีการดังกล่าวจะทำให้ข้อมูลสเปกตรัมของการพูดมีความสัมพันธ์ใกล้เคียงกับระบบการรับฟังเสียงของหูมนุษย์มากขึ้น ผ่านขั้นตอน autoregressive modelling ซึ่งเป็นการวิเคราะห์สัญญาณบนโดเมนเวลา ก็คือการประมาณพันธะเชิงเส้นมีขนาดเท่ากับ 8 [9] สุดท้ายปรับข้อมูลที่ได้ ให้อยู่ในรูปของสัมประสิทธิ์เซ็ปสตรัม โดยในที่นี้เราจะทำการทดลอง เปลี่ยนขนาดของสัมประสิทธิ์ เท่ากับ 8 12 และ 16 รูปที่ 4.5 เป็นตัวอย่างของสัมประสิทธิ์เซ็ปสตรัมที่ได้จากการประมวลผลเบื้องต้นแบบ PLP



รูปที่ 4.5 แสดงตัวอย่างสัมประสิทธิ์เซ็ปสตรัม PLP ที่มีจำนวนข้อมูลขนาดต่างๆ

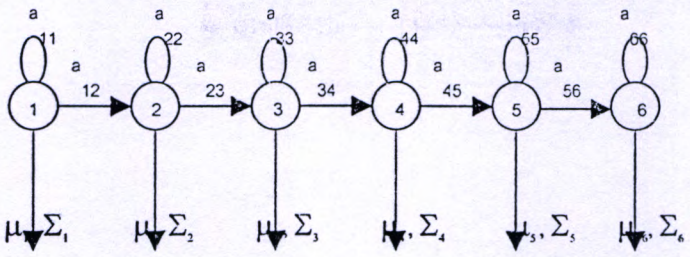
จากที่กล่าวมาเราจะได้สัมประสิทธิเซ็ปสตรัม C(1), C(2), C(3), ……., C(k) เป็นตัวแทนของแต่ละเฟรมแสดงคุณสมบัติของเสียงพูด ซึ่งเป็นการแสดงคุณสมบัติที่เรียกว่า static นอกจากนี้แล้วเรายังสามารถหาคุณสมบัติที่เรียกว่า dynamic ได้จากการหาความแตกต่างสัมประสิทธิเซ็ปสตรัมของเฟรมข้างเคียง ตามสมการ 4.1 นี้

$$\Delta C(n) \equiv \frac{d}{dt} C(n) \approx C(n+1) - C(n-1)$$
$$\Delta\Delta C(n) \equiv \frac{d}{dt} \Delta C(n) \approx \Delta C(n+1) - \Delta C(n-1) \qquad (4.1)$$

4.1.3 แบบจำลองฮิดเดนมาร์คอฟ

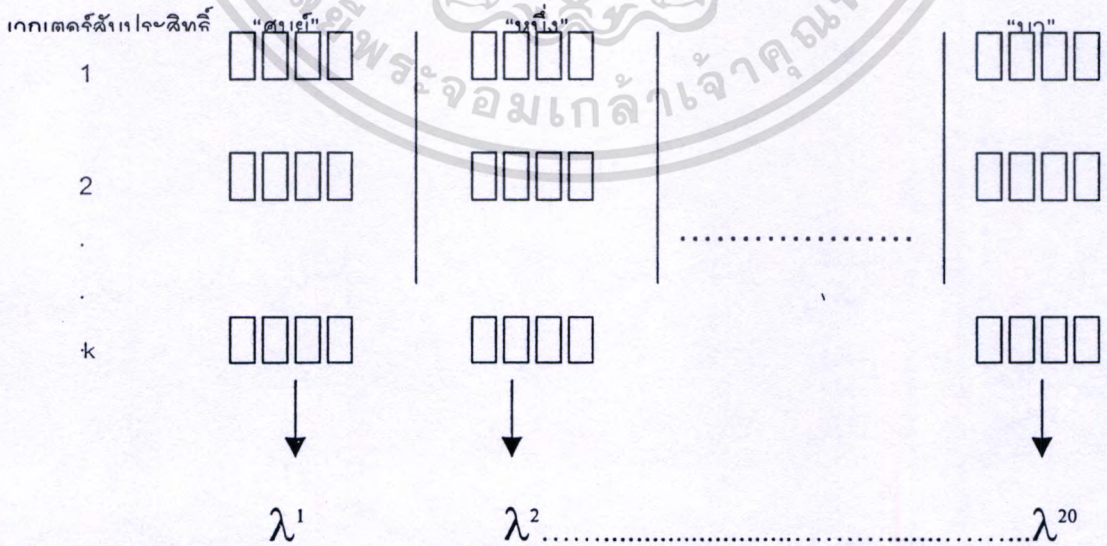
4.1.3.1 การฝึกสอนด้วย แบบจำลองฮิดเดน มาร์คอฟ

ลักษณะของแบบจำลองฮิดเดนมาร์คอฟ เป็นแบบ Left-Right Model มีจำนวน state 6 state มีการข้ามย้าย state สูงสุดได้ไม่เกิน 1 state แต่ละ state ให้ค่า mean (μ_j) และ covariance (Σ_j) ตามรูปที่ 4.6



รูปที่ 4.6 แสดงลักษณะของแบบจำลองฮิดเดนมาร์คอฟ

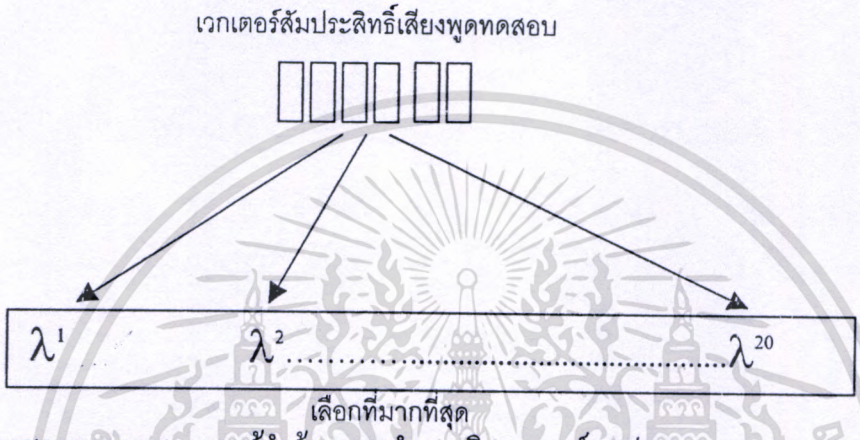
ขั้นตอนการฝึกสอนของแบบจำลอง ฮิดเดนมาร์คอฟนี้ เพื่อจำแนกความแตกต่างของเสียงพูดแต่ละคำโดยทำการปรับเปลี่ยนค่าพารามิเตอร์ของแบบจำลองฮิดเดนมาร์คอฟ $\lambda = (A,\mu_j\Sigma_j,\pi)$ ให้เป็นไปตามเสียงพูดแต่ละคำ ในวิทยานิพนธ์นี้ต้องสร้างแบบจำลองฮิดเดนมาร์คอฟ ทั้งหมด 20 แบบจำลอง λ^1, λ^2, λ^3, ...λ^{20} ตามคำพูดที่ต้องการแยกแยะในระบบนี้ สิ่งที่ต้องมีก็คือกลุ่มเสียงต้นแบบหรือลำดับค่าปรากฏทั้ง 20 กลุ่ม เพื่อใช้เป็นข้อมูลในการฝึกสอน (Train data) ก่อนอื่นจะต้องกำหนดค่า $\lambda = (A,\mu_j\Sigma_j,\pi)$ เริ่มต้น แล้วทำการปรับค่าพารามิเตอร์เหล่านี้อาศัยการแก้ปัญหาพื้นฐานข้อที่ 1 และข้อที่ 3 (ในวิทยานิพนธ์นี้แบบจำลองฮิดเดนมาร์คอฟ เป็นแบบต่อเนื่อง ดังนั้น $b_j(o_t)$ หาได้จากสมการที่ 2.53 และการปรับพารามิเตอร์ของ μ_j Σ_j หาได้จากสมการที่ 2.51 และ 2.52 ตามลำดับ) รูปที่ 4.7 แสดงการหาแบบจำลอง ฮิดเดนมาร์คอฟ ทั้ง 20 แบบจำลอง โดยแต่ละแบบจำลองจะมีลำดับของค่าปรากฏหรือเวกเตอร์สัมประสิทธิ์ที่ได้จากการประมวลผลเบื้องต้นเป็นข้อมูลสำหรับการฝึกสอน



รูปที่ 4.7 แสดงการฝึกสอนแบบจำลองฮิดเดนมาร์คอฟ

4.1.3.2 การทดสอบรู้จำด้วยแบบจำลอง ฮิดเดน มาร์คอฟ

รูปที่ 4.8 แสดงขั้นตอนในการทดสอบระบบการรู้จำโดยใช้แบบจำลอง ฮิดเดน มาร์คอฟ นำเวกเตอร์สัมประสิทธิ์ที่ได้จากการประมวลผลเบื้องต้นทำการทดสอบกับแบบจำลองทุกแบบจำลอง $\lambda^1, \lambda^2, \lambda^3, \ldots\ldots\ldots\ldots\lambda^{20}$ โดยแบบจำลองใดที่ให้ค่าความน่าจะเป็นสูงสุดจะถือว่าเสียงพูดที่นำมาทดสอบก็คือคำพูดเดียวกับแบบจำลองนั้นนั่นเอง โดยขั้นตอนการคำนวณหาค่าความน่าจะเป็นจะใช้การแก้ปัญหาพื้นฐานที่ 2 (Viterbi algorithm)



รูปที่ 4.8 แสดงการทดสอบการรู้จำด้วยแบบจำลองฮิดเดนมาร์คอฟ

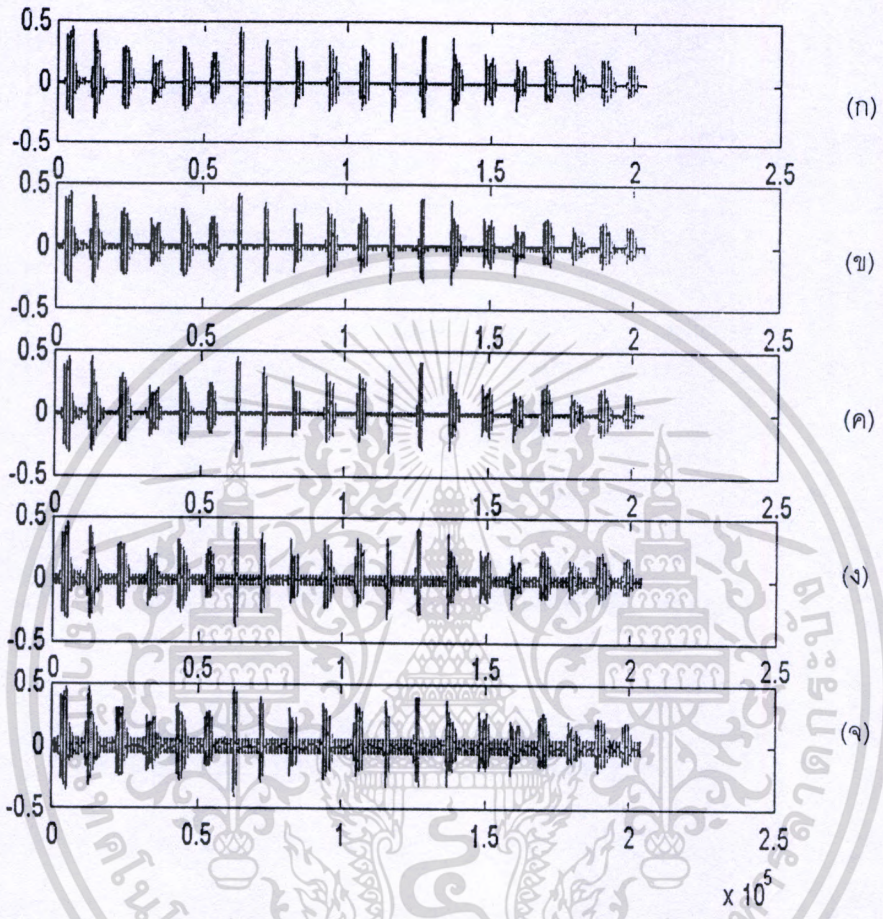
4.1.4 สัญญาณรบกวน (noise)

การนำระบบการรู้จำเสียงพูดไปประยุกต์ใช้กับอุปกรณ์ เพื่อใช้ในชีวิตประจำวัน เช่น การหมุนหมายเลขโทรศัพท์ด้วยเสียงพูด จะหลีกเลี่ยงสัญญาณเสียงที่มีสัญญาณรบกวนไม่ได้ ในวิทยานิพนธ์นี้ทำการทดลองระบบการรู้จำเสียงพูดทั้งสัญญาณเสียงที่ปราศจากสัญญาณรบกวน ได้จากการเก็บตัวอย่างเสียงในห้องทดลอง (clean speech) และสัญญาณเสียงที่มีสัญญาณรบกวน (noisy speech) โดยสัญญาณรบกวนที่เราจะใช้ในการทดลอง จากสถานที่ที่ทำการบันทึก และลักษณะของสัญญาณรบกวนดังนี้

1. เกาส์เซียน : เป็นสัญญาณรบกวนได้จากการสร้างขึ้นมาเอง
2. เสียงพูดแทรก : เป็นสัญญาณรบกวนที่มีผู้คนมากมายกำลังสนทนากันอยู่ บันทึกเสียงที่ตลาดสด
3. เสียงรถ : เป็นสัญญาณรบกวนที่มีเสียงรถวิ่งผ่านตลอด บันทึกเสียงตามสี่แยก

สำหรับเสียงพูดที่มีสัญญาณรบกวน (Noise speech) ในการทดลองนี้เราจะนำสัญญาณเสียงที่ปราศจากสัญญาณรบกวน (clean speech) บวกกับสัญญาณรบกวน (noise) โดยเรากำหนดให้ระดับของสัญญาณรบกวน เป็นค่าเฉลี่ยรากของรากที่สองของค่าเฉลี่ยพลังงานยกกำลังสอง (Average RMS Power) 4 ระดับ คือ –45dB, –40dB, –35dB และ –30dB หรือถ้าเทียบเป็นอัตราส่วนของสัญญาณต่อสัญญาณรบกวน (SNR) คือ 25dB, 20dB, 15dB และ 10dB

รูปที่ 4.9(ก) เป็นสัญญาณเสียง "ศูนย์-มา" ที่ปราศจากสัญญาณรบกวน 4.9(ข-จ) เป็นสัญญาณเสียงที่มีสัญญาณรบกวน (Noise speech) ตามระดับของสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB



รูปที่ 4.9 เสียงพูด "ศูนย์-มา" ที่ปราศจากสัญญาณรบกวน และมีสัญญาณรบกวนแบบ Gaussian ที่ระดับต่างๆ

4.1.5 รูปแบบของข้อมูลที่ใช้ในการฝึกสอน

การทดลองในวิทยานิพนธ์นี้เราจะแบ่งการทดลองเป็น 2 รูปแบบ ตามข้อมูลเสียงพูดที่ใช้ในการฝึกสอนดังนี้

1. การฝึกสอนแบบมาตรฐาน (Standard Training) เราจะใช้สัญญาณเสียงพูดที่ปราศจากสัญญาณรบกวนทำการฝึกสอนให้กับแบบจำลองฮิดเดนมาร์คอฟ
2. การฝึกสอนแบบหลากหลาย (Multistyle training) เราจะใช้สัญญาณเสียงพูดที่ปราศจากสัญญาณรบกวนและเสียงพูดที่มีสัญญาณรบกวนมีระดับของสัญญาณรบกวนต่างๆ กัน ทำการฝึกสอนให้กับแบบจำลองฮิดเดนมาร์คอฟ

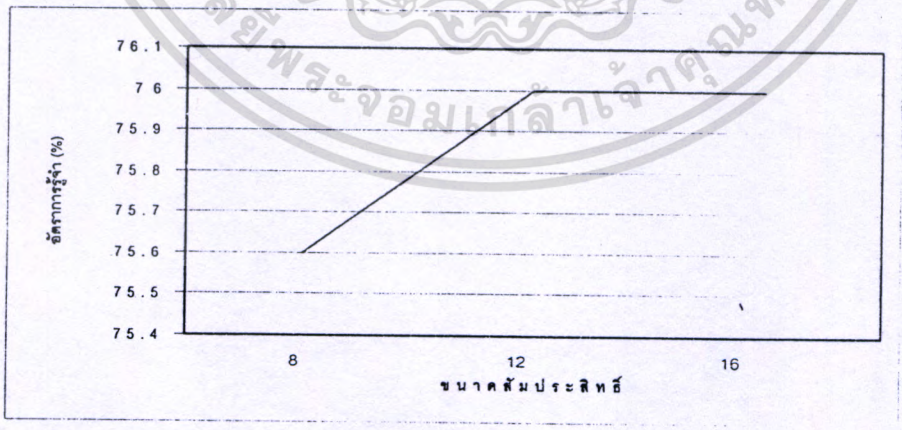
4.2 การทดลองและผลการทดลอง

ในวิทยานิพนธ์นี้เราจะทำการศึกษาเชิงเปรียบเทียบประสิทธิภาพของการประมวลผลเบื้องต้นแบบต่างๆ โดยเสียงพูดที่ใช้ในการทดสอบจะมีทั้งเสียงพูดที่ปราศจากสัญญาณรบกวนและเสียงพูดที่มีสัญญาณรบกวน แต่ก่อนที่เราจะทำการเปรียบเทียบประสิทธิภาพเราจำเป็นต้องทำทดลองโดยการเปลี่ยนขนาดสัมประสิทธิ์ที่ใช้เป็นตัวแทนของเสียงพูดแต่ละเฟรม เพื่อหาขนาดสัมประสิทธิ์ที่เหมาะสมที่สุด

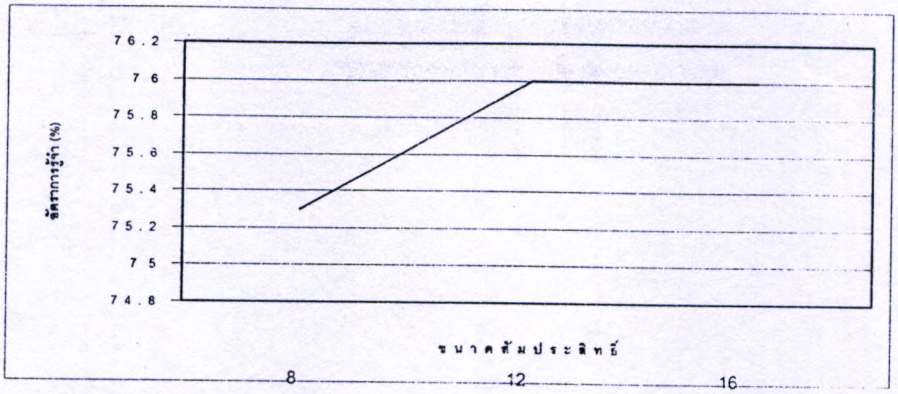
4.2.1 การทดลองหาขนาดสัมประสิทธิ์ที่เหมาะสม

ในการทดลองเบื้องต้นเราจะกำหนดให้ขนาดข้อมูลของสัมประสิทธิ์ ที่ได้จากการประมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP ตามนี้ 8, 12 และ 16 เสียงที่เราใช้สำหรับการฝึกสอนมีทั้งหมด 900 เสียง (15 คนพูด พูดคนละ 3 ครั้ง ทั้งหมด 20 คำพูด) ทำการฝึกสอนด้วยแบบจำลองฮิดเดนมาร์คอฟ และเสียงที่ใช้สำหรับการทดสอบทั้งหมด 300 เสียง (5 คนพูด พูดคนละ 3 ครั้ง ทั้งหมด 20 คำพูด) จะได้ผลการทดลองเป็นเปอร์เซนต์ (%) การรู้จำถูกต้อง ตามรูปที่ 4.10, 4.11 และ 4.12 เสียงพูดที่ใช้สำหรับการฝึกสอนและทดสอบเป็นเสียงพูดที่ปราศจากการรบกวนทั้งหมด

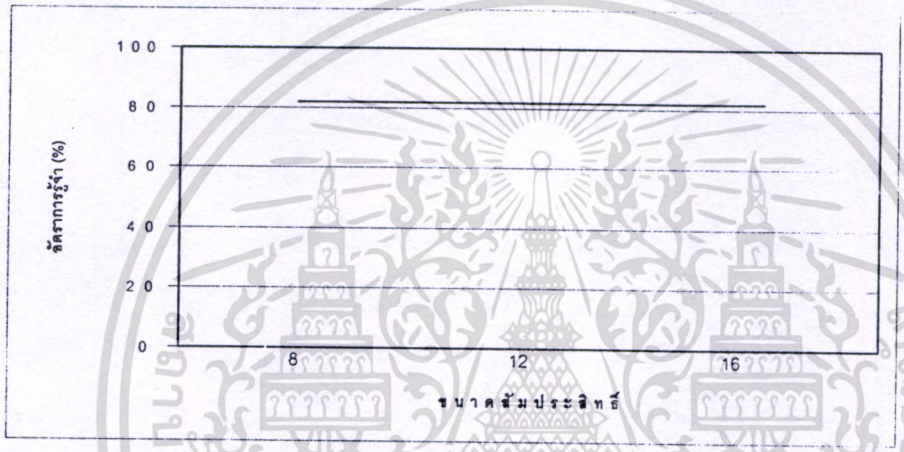
จากรูปที่ 4.10 แสดงให้เห็นว่าในระบบการรู้จำเสียงพูดภาษาไทยในวิทยานิพนธ์นี้ที่ใช้การประมวลผลเบื้องต้นแบบ LPCC จำนวนสัมประสิทธิ์ที่เหมาะสมที่สุด (มีอัตราการรู้จำถูกต้องมากที่สุด) เท่ากับ 12 ซึ่งมีอัตราการรู้จำถูกต้องเท่ากับ 76.0 % รูปที่ 4.11 จำนวนสัมประสิทธิ์ที่เหมาะสมที่สุดในการประมวลผลแบบ MFCC คือ 12 มีอัตราการรู้จำถูกต้องเท่ากับ 76.0 % และสุดท้าย การประมวลผลเบื้องต้นแบบ PLP ตามรูปที่ 4.12 จำนวนสัมประสิทธิ์ที่เหมาะสมที่สุดคือ 12 เนื่องจากมีอัตราการรู้จำถูกต้องสูงสุดเท่ากับ 82.0 %



รูปที่ 4.10 ผลการทดลองอัตราการรู้จำถูกต้อง LPCC



รูปที่ 4.11 ผลการทดลองอัตราการรู้จำถูกต้อง MFCC



รูปที่ 4.12 ผลการทดลองอัตราการรู้จำถูกต้อง PLP

4.2.2 การทดสอบระบบการรู้จำเสียงด้วยเสียงที่มีสัญญาณรบกวน

จากการทดลองข้างต้นทำให้เราทราบว่าจำนวนสัมประสิทธิ์ที่ได้จากการปะมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP มีขนาดสัมประสิทธิ์ที่เหมาะสมที่สุดคือ 12, 12 และ 12 ตามลำดับ ดังนั้นแต่ละเฟรมมีขนาดสัมประสิทธิ์ที่ได้จากการประมวลผลเบื้องต้นแบบต่างๆ เท่ากับ 12 ซึ่งเป็นคุณสมบัติเชิง static นอกจากนี้เรายังได้ทดลองโดยนำคุณสมบัติเชิง dynamic มาเป็นตัวแทนของเสียงพูดด้วยแต่ละเฟรมจะมีขนาดสัมประสิทธิ์เทากับ 36 (C(n) ΔC(n) $\Delta\Delta$C(n)) การทดลองส่วนนี้เราจะแบ่งการทดลองเป็น 2 รูปแบบ ตามรูปแบบของข้อมูลเสียงพูดที่ใช้ในการฝึกสอน

4.2.2.1 การฝึกสอนแบบมาตรฐาน (Standard training)

เสียงพูดที่ใช้ในการฝึกสอนแบบจำลองฮิดเดนมาร์คอฟ นั้นจะเป็นเสียงพูดที่ปราศจากสัญญาณรบกวน ทั้งหมด 900 เสียง และทดสอบการรู้จำด้วยเสียงพูดทั้งที่ปราศจากสัญญาณรบกวน และที่มีสัญญาณรบกวนด้วยระดับสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB อย่างละ 300 เสียง จากการทดลองการประมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP เมื่อทดสอบกับ

สัญญาณเสียงพูดที่มีสัญญาณรบกวน แบบเกาส์เซียน แบบเสียงพูดแทรก และแบบเสียงรถ ได้อัตราการรู้จำตามตารางที่ 4.1(ก) ตารางที่ 4.1(ข) และตารางที่ 4.1(ค) ตามลำดับ

ตารางที่ 4.1 การฝึกสอนแบบมาตรฐาน ผลการรู้จำถูกต้อง (%) ทดสอบกับเสียงพูดที่มีสัญญาณรบกวน (ก) เกาส์เซียน (ข) เสียงพูดแทรก (ค) เสียงรถ

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	76.0%	76.0%	82.0%	94.0%	94.0%	96.3%
-45	49.6%	62.6%	70.3%	72.3%	86.6%	89.6%
-40	30.3%	48.3%	52.6%	48.3%	74.0%	83.0%
-35	21.0%	27.0%	33.3%	46.0%	57.6%	66.3%
-30	14.3%	19.0%	22.6%	27.6%	32.3%	35.3%

(ก)

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	76.0%	76.0%	82.0%	94.0%	94.0%	96.3%
-45	64.0%	66.3%	71.6%	84.3%	86.0%	89.6%
-40	55.0%	48.6%	57.0%	76.0%	77.6%	84.0%
-35	37.3%	38.0%	45.6%	55.6%	64.0%	67.0%
-30	23.0%	23.6%	26.6%	35.6%	44.0%	44.6%

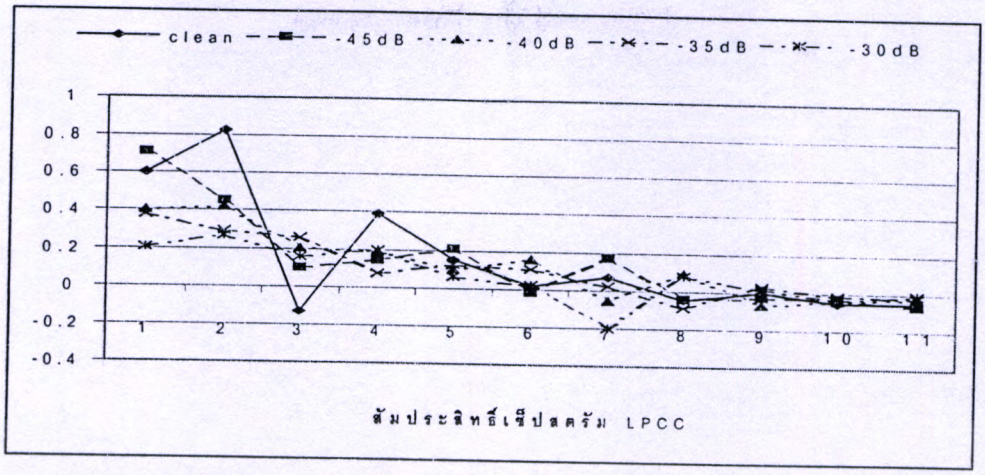
(ข)

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	76.0%	76.0%	82.0%	94.0%	94.0%	96.3%
-45	69.3%	70.0%	72.6%	86.0%	88.0%	91.0%
-40	59.3%	55.0%	63.0%	79.0%	82.6%	85.3%
-35	44.0%	41.3%	49.3%	62.0%	68.3%	72.3%
-30	30.3%	31.0%	34.0%	43.6%	48.6%	51.2%

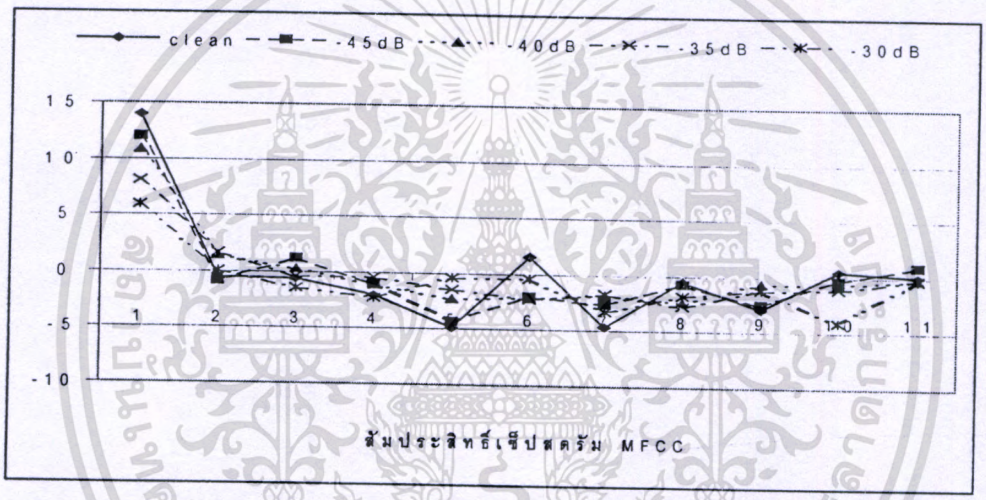
(ค)

จากผลการทดลองในส่วน static freture สำหรับเสียงพูดที่ปราศจากสัญญาณรบกวน จะได้อัตราการรู้จำสูงใกล้เคียงกัน คือ 76.0% (LPCC), 76.0% (MFCC) และ 82.0% แต่เมื่อทดสอบกับเสียงพูดที่มีสัญญาณรบกวนที่ระดับต่างๆ ประสิทธิภาพในการรู้จำจะลดลงอย่างเห็นได้ชัดเจนขึ้นอยู่กับระดับของสัญญาณรบกวน ตัวอย่างเช่น สัญญาณรบกวนแบบเกาส์เซียนที่ระดับ –30dB มีอัตราการรู้จำ 14.3% (LPCC), 19.0% (MFCC) และ 22.6% (PLP) อย่างไรก็ตามการประมวลผลเบื้องต้นแบบ PLP ก็จะให้ประสิทธิภาพการรู้จำสูงกว่าการประมวลผลเบื้องต้นแบบ MFCC และ LPCC ของสัญญาณรบกวนทุกชนิด และจากผลการทดลองในส่วนของ static and dynamic freture ทำให้เห็นว่าการนำส่วนของ dynamic freture มาใช้เป็นตัวแทนของเสียงพูดด้วยนั้น ทำให้ประสิทธิภาพของระบบการรู้จำดีขึ้นมาก ตัวอย่างเช่น สัญญาณรบกวนแบบเกาส์เซียนที่ระดับ –35dB มีอัตราการรู้จำ 46.0% (LPCC), 57.6% (MFCC) และ 66.3% (PLP) มีอัตราการรู้จำสูงกว่าการประมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP ที่ใช้แต่ส่วน static freture อย่างเดียวถึง 25.0% 30.6% และ 33.0% ตามลำดับ

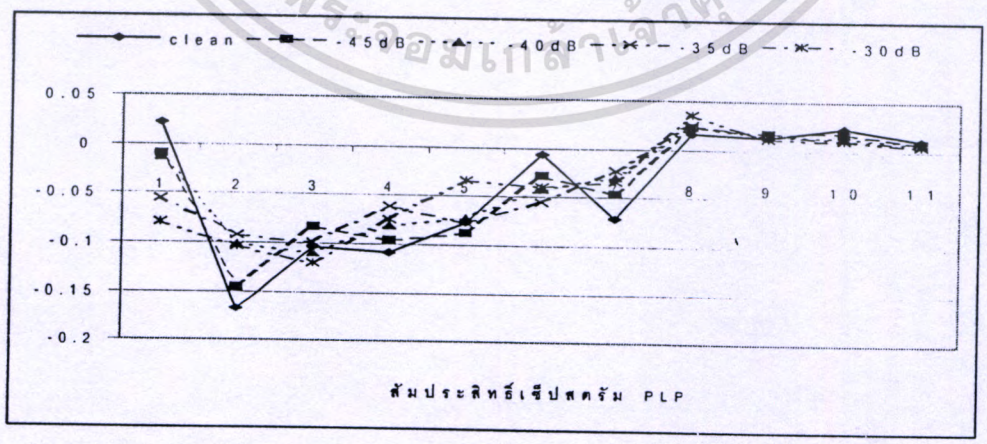
เมื่อเรานำสัมประสิทธิ์เซ็ปสตรัมเฟรมที่ 20 เสียง “ศูนย์” ที่ระดับสัญญาณรบกวนต่างๆ ที่ได้จากการประมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP มาพล็อตกราฟจะได้ตามรูปที่ 4.13 4.14 และ 4.15 ตามลำดับ จากรูปที่ 4.13 ค่าสัมประสิทธิ์ในช่วงแรกที่ระดับสัญญาณรบกวนระดับต่างๆ จะมีค่าสัมประสิทธิ์ต่างกันกับเสียงที่ปราศจากสัญญาณรบกวนมาก และจากรูปที่ 4.14 และ 4.15 ที่ผ่านการปรมวลผลเบื้องต้นแบบ MFCC และ PLP จะเห็นได้ว่าค่าสัมประสิทธิ์เซ็ปสตรัมในช่วงแรก มีความใกล้เคลียงกันระหว่างเสียงที่ปราศจากสัญญาณรบกวน กับเสียงที่มีสัญญาณรบกวน แต่ที่ช่วงหลังๆ ค่าสัมประสิทธิ์เซ็ปสตรัมที่ได้จากการประมวลผลเบื้องต้นแบบ PLP ตามรูปที่ 4.15 มีความใกล้เคลียงกันระหว่างเสียงที่ปราศจากสัญญาณรบกวน กับเสียงที่มีสัญญาณรบกวน มากกว่าการประมวลผลเบื้องต้นแบบ MFCC ตามรูปที่ 4.14 ที่กล่าวมามีความสอดคล้องกับผลการทดลอง คือ การประมวลผลเบื้องต้นแบบ PLP จะมีประสิทธิภาพในการรู้จำสูงกว่าการประมวลผลเบื้องต้นแบบอื่นๆ โดยเฉพาะทดสอบกับเสียงที่มีสัญญาณรบกวน



รูปที่ 4.13 สัมประสิทธิ์เซ็ปสตรัมจากการประมวลผลเบื้องต้นแบบ LPCC ที่ระดับสัญญาณรบกวนต่างๆ



รูปที่ 4.14 สัมประสิทธิ์เซ็ปสตรัมจากการประมวลผลเบื้องต้นแบบ MFCC ที่ระดับสัญญาณรบกวนต่างๆ



รูปที่ 4.15 สัมประสิทธิ์เซ็ปสตรัมจากการประมวลผลเบื้องต้น PLP ที่ระดับสัญญาณรบกวนต่างๆ

4.2.2.2 การฝึกสอนแบบหลากหลาย (Multistyle training)

เสียงพูดที่ใช้ในการฝึกสอนแบบจำลองฮิดเดนมาร์คอฟนั้นจะเป็นเสียงพูดที่ปราศจากสัญญาณรบกวนและเสียงพูดที่มีสัญญาณรบกวน ทั้งหมด 1500 เสียง (ได้จากเสียงที่ปราศจากสัญญาณรบกวน ดึงมา 300 เสียง จาก 900 เสียง และ เสียงพูดที่มีสัญญาณรบกวน ซึ่งมีระดับของสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB อย่างละ 300 เสียง) จะทดสอบการรู้จำด้วยเสียงพูด ที่ปราศจากสัญญาณรบกวน และเสียงพูดที่มีสัญญาณรบกวนด้วยระดับสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB อย่างละ 300 เสียง ผลการทดลองของการประมวลผลเบื้องต้นแบบ LPCC MFCC และ PLP เมื่อทดสอบกับสัญญาณเสียงพูดที่มีสัญญาณรบกวน แบบเกาส์เซียนแบบเสียงพูดแทรก และแบบเสียง ได้อัตราการรู้จำตามตารางที่ 4.2(ก) ตารางที่ 4.2(ข) และตารางที่ 4.2(ค) ตามลำดับ

ตารางที่ 4.2 การฝึกสอนแบบหลากหลาย ผลการรู้จำถูกต้อง (%)เมื่อทดสอบกับเสียงพูดทดสอบกับเสียงพูดที่มี สัญญาณรบกวน (ก) เกาส์เซียน (ข) เสียงพูดแทรก (ค) เสียงรถ

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	75.3%	74.6%	70.6%	92.6%	88.6%	94.6%
-45	77.0%	75.3%	78.0%	95.6%	91.0%	96.3%
-40	75.0%	73.3%	79.0%	90.3%	87.0%	93.3%
-35	63.6%	68.6%	73.0%	80.6%	81.0%	86.3%
-30	45.3%	53.0%	54.3%	55.0%	52.6%	64.0%

(ก)

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	68.6%	77.3%	71.0%	82.6%	89.6%	94.6%
-45	72.6%	76.6%	75.3%	88.3%	92.2%	95.0%
-40	71.0%	73.6%	75.0%	82.0%	88.3%	93.6%
-35	66.0%	68.0%	74.3%	82.0%	86.3%	87.6%
-30	57.3%	56.0%	64.6%	65.0%	70.3%	79.3%

(ข)

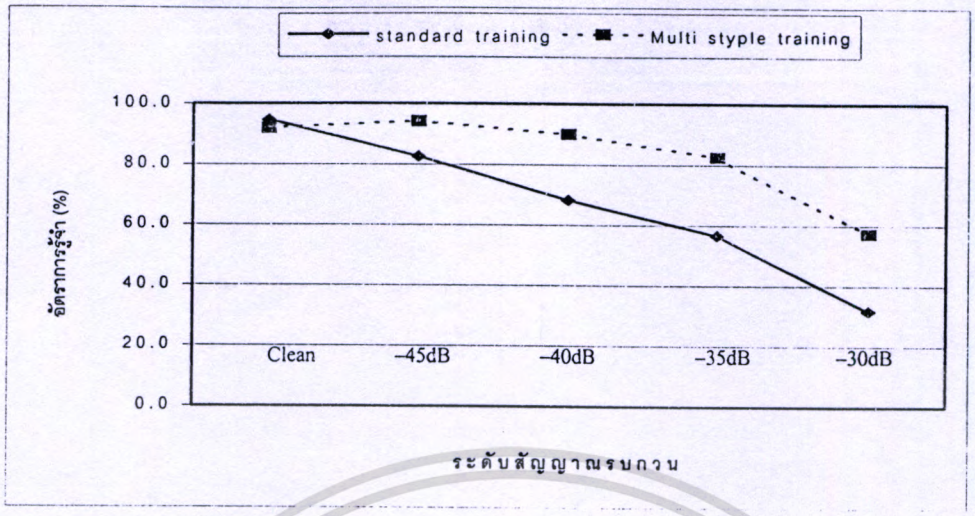
ตารางที่ 4.2 (ต่อ)

Standard training DB	Static Features			Static and Dynamic Features		
	LPCC	MFCC	PLP	LPCC	MFCC	PLP
Clean	71.0%	71.0%	72.3%	82.6%	88.6%	92.0%
-45	74.3%	76.3%	75.6%	90.0%	92.6%	94.0%
-40	73.3%	74.0%	76.3%	89.0%	90.3%	92.6%
-35	65.6%	69.6%	71.3%	81.6%	84.0%	87.6%
-30	57.0%	58.6%	64.0%	69.3%	75.6%	79.0%

(ค)

จากผลการทดลองแบบนี้ ผลการรู้จำมีลักษณะเดียวกับการทดลองตามหัวข้อที่ผ่านมา คือ เมื่อทดสอบกับเสียงพูดที่ปราศจากสัญญาณรบกวนได้อัตราการรู้จำสูง แต่เมื่อทดสอบกับเสียงพูดที่มีสัญญาณรบกวนอัตราการรู้จำจะลดลง ขึ้นอยู่กับระดับของสัญญาณรบกวน เมื่อเรานำอัตราการรู้จำทั้งแบบ LPCC MFCC และ PLP ในส่วนของ Static and Dynamic Features มีสัญญาณรบกวนแบบเกาส์เซียน หาค่าเฉลี่ย แล้วนำมาทำกราฟตามรูปที่ 4.16 โดยแกน x เป็นระดับของสัญญาณรบกวน จากกราฟเป็นการเปรียบเทียบประสิทธิภาพการรู้จำระหว่างการฝึกสอนแบบมาตรฐาน (Standard) และการฝึกสอนแบบหลากหลาย (Multi style) ที่เสียงพูดที่ปราศจากสัญญาณรบกวน (clean) จะให้อัตราการรู้จำพอๆ กัน เมื่อทดสอบเสียงพูดที่มีสัญญาณรบกวนการฝึกสอนแบบหลากหลาย จะมีประสิทธิภาพสูงกว่าการฝึกสอนแบบมาตรฐาน ถึง 4%-20% ขึ้นอยู่กับระดับสัญญาณรบกวน

จากผลการทดลองตามตารางที่ 4.1 และ 4.2 โดยรวมแล้วจะเห็นได้ว่าสัญญาณรบกวนแบบเกาส์เซียน จะมีผลกระทบกับระบบการรู้จำมากกว่าสัญญาณรบกวนแบบอื่น คือทำให้มีการลดลงของอัตราการรู้จำสูงกว่า และสัญญาณรบกวนแบบเสียงรถ มีผลกระทบกับระบบน้อยที่สุด



รูปที่ 4.16 เปรียบเทียบประสิทธิภาพการรู้จำ ระหว่างที่ฝึกสอนแบบมาตรฐาน (standard traning) กับฝึกสอนแบบหลากหลาย (Multistyle training)

4.3 การทดลอง และผลการทดลองโดยใช้โปรแกรม HTK

จากการทดลองตามหัวข้อที่ 4.2 ทั้งหมด ได้จากการเขียนโปรแกรมของผู้ทำวิทยานิพนธ์เอง ซึ่งในปัจจุบันนี้ได้มีโปรแกรมสำเร็จรูป HTK [14] เป็นโปรแกรมเกี่ยวกับระบบการรู้จำเสียง ซึ่งถูกพัฒนามาจากการรู้จำเสียงภาษาอังกฤษจนเป็นโปรแกรมที่มาตราฐาน และมีประสิทธิภาพสูงสุดในปัจจุบันนี้ ดังนั้นในวิทยานิพนธ์นี้จึงได้นำโปรแกรม HTK มาทดลองใช้ กับเสียงภาษาไทย โดยใช้ข้อมูลเสียงพูดเดียวกับการทดลองในหัวข้อที่ 4.2

4.3.1 ข้อกำหนดเบื้องต้น

ในส่วนของการประมวลผลเบื้องต้น หรือส่วนที่ดึงลักษณะที่สำคัญของเสียง เราใช้ 2 วิธี คือ LPCC กับ MFCC เนื่องจากว่าโปรแกรม HTK ยังไม่มีการประมวลผลเบื้องต้นแบบ PLP ขนาดของสัมประสิทธิ์เซ็ปสตรัมแต่ละเฟรมเท่ากับ 12 เป็นส่วน static และเราจะใช้ทั้งส่วนที่เป็น static และ dynamic ดังนั้นตัวแทนของเสียงพูด แต่ละเฟรมจะมีขนาดเท่ากับ 36 (C(n) ΔC(n) $\Delta\Delta$C(n)) สำหรับส่วนของระบบการเรียนรู้ และรู้จำ ด้วยแบบจำลองฮิดเดนมาร์คอฟ กำหนดให้มีจำนวน satae เท่ากับ 5 state

4.3.2 การฝึกสอนแบบมาตรฐาน (standard training)

ข้อมูลเสียงพูดที่ใช้สำหรับการฝึกสอน และทดสอบ เหมือนกับหัวข้อที่ 4.2.2.1 ได้ผลการทดลองตามตารางที่ 4.3 ดังนี้

ตารางที่ 4.3 การฝึกสอนแบบมาตรฐาน ผลการรู้จำถูกต้อง (%) ทดสอบกับเสียงพูดที่มีสัญญาณรบกวน (ก) เกาส์เซียน (ข) เสียงพูดแทรก (ค) เสียงรถ

Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	100%	100%	100%	100%
-45	83.3%	89.3%	94.3%	97.6%
-40	56.3%	74.6%	74.6%	92.3%
-35	22.6%	50.3%	38.3%	74.0%
-30	16.0%	23.3%	20.0%	38.6%

(ก)

Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	100%	100%	100%	100%
-45	83.6%	89.6%	94.6%	97.3%
-40	74.3%	83.0%	91.3%	95.0%
-35	54.0%	70.3%	79.6%	90.6%
-30	26.0%	41.3%	52.0%	67.6%

(ข)

Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	100%	100%	100%	100%
-45	83.6%	89.3%	94.3%	97.6%
-40	74.0%	84.6%	91.6%	96.0%
-35	56.6%	72.0%	81.0%	89.3%
-30	31.0%	47.3%	53.6%	70.3%

(ค)

4.3.3 การฝึกสอนแบบหลากหลาย (Multistyle training)

ข้อมูลเสียงพูดที่ใช้สำหรับการฝึกสอน และทดสอบ เหมือนกับหัวข้อที่ 4.2.2.2 ได้ผลการทดลองตามตารางที่ 4.4 ดังนี้

ตารางที่ 4.4 การฝึกสอนแบบหลากหลาย ผลการรู้จำถูกต้อง (%)เมื่อทดสอบกับเสียงพูดทดสอบกับเสียงพูดที่มี สัญญาณรบกวน (ก) เกาส์เซียน (ข) เสียงพูดแทรก (ค) เสียงรถ

Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	97.0%	97.0%	98.6%	100%
-45	97.0%	96.3%	100%	100%
-40	95.6%	96.6%	98.6%	97.6%
-35	80.6%	82.3%	86.0%	87.0%
-30	73.0%	73.3%	76.3%	76.6%

(ก)

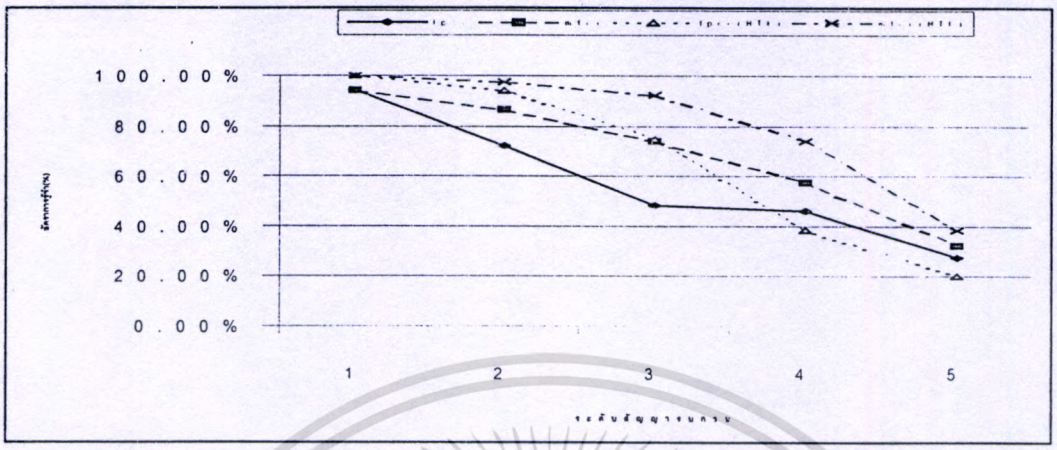
Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	95.6%	95.0%	100%	100%
-45	96.6%	97.6%	100%	100%
-40	95.0%	96.6%	98.6%	97.0%
-35	83.3%	81.3%	92.6%	90.0%
-30	76.0%	75.0%	85.3%	83.6%

(ข)

Standard	Static Features		Static and Dynamic Features	
training	LPCC	MFCC	LPCC	MFCC
Clean	96.6%	96.3%	100%	100%
-45	98.0%	98.6%	100%	100%
-40	97.0%	97.3%	99.0%	96.3%
-35	82.6%	85.6%	90.6%	90.6%
-30	74.6%	78.3%	86.6%	85.6%

(ค)

เมื่อนำผลการทดลองจากตารางที่ 4.1 และตารางที่ 4.3 ส่วนของการประมวลผลเบื้องต้นแบบ LPCC และ MFCC นำมาพล็อตกราฟรวมกันจะได้ตามรูปที่ 4.17



รูปที่ 4.17 เปรียบเทียบประสิทธิภาพการรู้จำ ระหว่างโปรแกรมที่สร้างขึ้น กับโปรแกรม HTK

จากการทดลองตามตารางที่ 4.3 และ 4.4 เมื่อเทียบกับผลการทดลองจากโปรแกรมที่ผู้ทำวิทยานิพนธ์สร้างขึ้นเอง นั้นจะเห็นได้ว่า มีประสิทธิภาพต่ำกว่าโปรแกรม HTK โดยดูจากผลการทดลอง ที่นำเสียงพูดที่ปราศจากสัญญาณรบกวนทำการทดสอบระบบ จะได้อัตราการรู้จำสูงถึง 100% เมื่อพิจารณาผลการทดลองโดยรวมดูรูปที่ 4.17 จากการใช้โปรแกรม HTK สามารถสรุปได้แบบเดียวกับการทดลองในหัวข้อที่ 4.2 ก็คือ เมื่อนำเสียงพูดที่มีสัญญาณรบกวนมาทดสอบกับระบบการรู้จำเสียง ก็จะทำให้ประสิทธิภาพในการรู้จำลดลงขึ้นอยู่กับระดับของสัญญาณรบกวน

จากผลการทดลองโดยใช้โปรแกรม HTK จะเห็นได้ว่าสามารถนำมาประยุกต์ใช้เสียงพูดภาษาไทยแบบคำโดดได้เป็นอย่างดี แต่เนื่องจากว่าเสียงที่เป็นภาษาไทยมีความแตกต่างจากภาษาอังกฤษตรงที่เสียงภาษาไทยมีระดับของเสียงอย่างชัดเจน คือเสียงวรรณยุกต์ สามัญ เอก โท ตรี และจัตวา ดังนั้นเพื่อทดสอบโปรแกรม HTK ว่าสามารถแยกเสียงวรรณยุกต์ได้หรือไม่ โดยทำการนำเสียงเป็นเสียงเดียวกัน แต่มีวรรณยุกต์ต่างกัน คือ กา ก่า ก้า ก๊า ก๋า อัดเสียงจากผู้พูดคนเดียว คำละ 30 ครั้ง สำหรับฝึกสอน และ 20 ครั้ง สำหรับทดสอบ จะได้ผลการทดลองตามตารางที่ 4.5

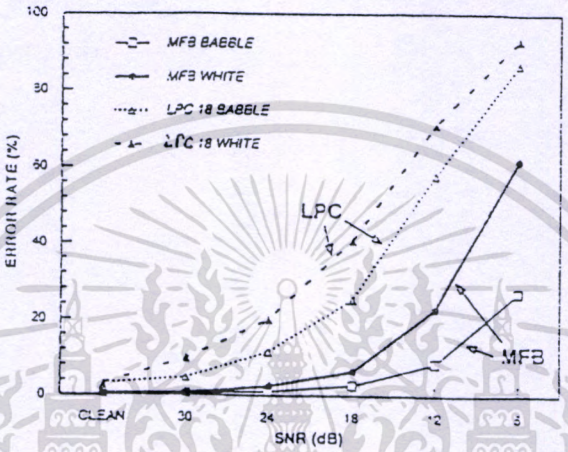
ตารางที่ 4.5 ผลการทดลองใช้โปรแกรม HTK แยกเสียงวรรณยุกต์

Test SNR	กา	ก่า	ก้า	ก๊า	ก๋า
LPCC	100%	100%	100%	100%	100%
MFCC	100%	100%	100%	100%	100%

จากผลการทดลองแสดงให้เห็นว่าโปรแกรม HTK สามารถนำมาประยุกต์ใช้แยกเสียงภาษาไทยได้ แต่การทดลองนี้ยังไม่สามารถยืนยันได้แน่ชัด เพราะเสียงที่ใช้ในการทดลองมีน้อยเกินไป

4.4 เปรียบเทียบผลการทดลอง กับผลงานวิจัยอื่น

จากงานวิจัย [7] เป็นการทดลองระบบการรู้จำเสียงภาษาอังกฤษแบบคำโดดทั้งหมด 105 คำ ผลการทดลอง ตามรูปที่ 4.18 แสดงผลการทดลองของการรู้จำระหว่างวิธีการประมวลผลเบื้องต้นแบบ LPCC และ MFCC โดยทำการทดสอบกับเสียงที่มีสัญญาณรบกวนแบบ White Noise และแบบ Babble Noise ที่ระดับต่างๆกัน



รูปที่ 4.18 ประสิทธิภาพของการรู้จำระหว่างวิธีการประมวลผลเบื้องต้นแบบ LPCC และ MFCC

และจาก [9] เป็นการทดลองวัดประสิทธิภาพของระบบการรู้เสียงพูดภาษาชาวฮอลแลน ทั้งหมด 26 คำ มีการประมวลผลเบื้องต้นแบบ MFCC และ PLP โดยทำการทดสอบกับเสียงที่มีสัญญาณรบกวนแบบ Gaussian Noise ได้ผลการทดลองตามตารางที่ 4.5

ตารางที่ 4.5 ประสิทธิภาพของการรู้จำระหว่างวิธีการประมวลผลเบื้องต้นแบบ MFCC และ PLP

Test SNR	Clean	22dB	16dB	10dB
MFCC	99.83	97.57	76.95	51.36
PLP	99.72	95.82	79.94	66.05

จากผลการทดลองของงานวิจัยทั้งสอง เมื่อมาเทียบกับผลการทดลองที่ทำในวิทยานิพนธ์นี้ มีลักษณะคล้ายๆ กัน คือการประมวลผลเบื้องต้นแบบ MFCC จะมีประสิทธิภาพมากกว่า การประมวลผลเบื้องต้นแบบ LPCC เห็นได้ชัดจาก รูปที่ 4.17 และอัตราการรู้จำของระบบจะลดลง เมื่อนำเสียงที่มีสัญญาณรบกวนทดสอบกับระบบ และจากผลการทดลองตามตารางที่ 4.5 ที่ระดับของสัญญาณรบกวนสูงขึ้น การประมวลผลเบื้องต้นแบบ PLP จะมีประสิทธิภาพในการรู้จำได้ดีกว่าการประมวลผลเบื้องต้นแบบ MFCC

บทที่ 5

สรุปผลการทดลอง

จากการทดลองโดยเปรียบเทียบประสิทธิภาพการประมวลผลเบื้องต้น ทั้ง 3 วิธี คือ Linear predictive cepstrum coefficience (LPCC), Mel Frequency Cepstral Coefficients (MFCC) และ Perceptual Linear Predictive (PLP) สองวิธีแรกเป็นวิธีที่รู้จักกันดี และนิยมใช้กันมากสำหรับระบบการรู้จำเสียงพูดโดยทั่วไป แต่อีกวิธีหนึ่ง เป็นวิธีที่รู้จักกันในระบบการรู้จำเสียงพูด ที่เสียงมีความแปรปรวน (Robust Speech Recognition) ในวิทยานิพนธ์นี้สร้างระบบการรู้จำเสียงพูดแบบคำโดดไม่ขึ้นอยู่กับผู้พูดโดยเสียงพูดที่สามารถรู้จำได้ทั้งหมด 20 คำ "ศูนย์", "หนึ่ง", "สอง", "สาม", "สี่", "ห้า", "หก", "เจ็ด", "แปด", "เก้า", "เปิด", "ปิด", "ยก", "วาง", "ซ้าย", "ขวา", "หน้า", "หลัง", "ไป" และ "มา" โดยเสียงพูดที่ใช้ในการทดลองมีทั้งเสียงที่ปราศจากสัญญาณรบกวน และเสียงที่มีสัญญาณรบกวนแบบเกาส์เซียน แบบเสียงพูดแทรก และแบบเสียงรถ ที่มีระดับของสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB

การทำงานของระบบมี 4 ขั้นตอนคือ ขั้นตอนการเตรียมข้อมูลเบื้องต้น ขั้นตอนการวิเคราะห์และดึงลักษณะที่สำคัญ หรือเรียกว่าการประมวลผลเบื้องต้น ขั้นตอนการจำแนกรูปแบบ และขั้นตอนการตัดสินใจ

ขั้นตอนการเตรียมข้อมูลเบื้องต้น เป็นการแบ่งข้อมูลเสียงออกเป็นเฟรม เพื่อเตรียมข้อมูลก่อนที่จะทำการวิเคราะห์ แบ่งออกเป็นส่วนได้ดังนี้ การพรีเอมฟาซิส การแบ่งช่วงสัญญาณ การวินโดว์ การหาจุดสิ้นสุดของเสียงพูด

ขั้นตอนการจำแนกรูปแบบ และขั้นตอนการตัดสินใจ ในวิทยานิพนธ์เลือกใช้แบบจำลองฮิดเดนมาร์คอฟแบบต่อเนื่อง มี state 6 state มีการข้ามย้าย state แบบ Left – Right Model

ขั้นตอนการประมวลผลเบื้องต้น จากการทดลองเบื้องต้นเพื่อหาขนาดสัมประสิทธิ์ที่เหมาะสมของแต่วิธีการประมวลผลเบื้องต้นแบบต่างๆ ได้ขนาดสัมประสิทธิ์เท่ากับ 12 เหมือนกันทั้งหมด เพราะให้ผลการทดสอบความถูกต้องดีที่สุด วิธีการประมวลผลเบื้องต้นที่เราใช้ทดลองเพื่อทำการเปรียบเทียบมี 3 วิธี ดังนี้

1. LPCC : นำแต่ละเฟรมของเสียงพูดหาค่าสัมประสิทธิ์การพันธะเชิงเส้น (LPC) โดยใช้วิธีอัตสหสัมพันธ์ (autocorrelation) ได้ขนาดเท่ากับ 10 เปลี่ยนเป็นสัมประสิทธิ์เซ็ปสตรัมขนาดเท่ากับ 12
2. MFCC : หาสเปกตรัมของเสียงพูดแต่ละเฟรมด้วย FFT ผ่าน Mel Filter Bank 18 ช่อง หาสัมประสิทธิ์เซ็ปสตรัมมีขนาดเท่ากับ 12 โดยใช้ logarithms และ IDFT

3. PLP : หาสเปกตรัมของเสียงพูดแต่ละเฟรมด้วย FFT ผ่าน Bark Filter Bank 17 ช่อง ปรับคุณภาพโดยยึดหลักการรับรู้ของหูมนุษย์ด้วย equal-loudness และ intensity-loudness หาค่าสัมประสิทธิ์การพันธะเชิงเส้น(LPC) โดยใช้วิธีอัตสหสัมพันธ์ (autocorrelation) จะได้สัมประสิทธิ์มีขนาดเท่ากับ 8 เปลี่ยนเป็นสัมประสิทธิ์เซ็ปสตรัม ขนาดเท่ากับ 12

แต่ละเฟรมมีขนาดสัมประสิทธิ์ที่ได้จากการประมวลผลเบื้องต้นแบบต่างๆ เท่ากับ 12 ซึ่งเป็นคุณสมบัติเชิง static นอกจากนี้เรายังได้ทดลองโดยนำคุณสมบัติเชิง dynamic มาเป็นตัวแทนของเสียงพูดด้วยแต่ละเฟรมจะมีขนาดสัมประสิทธิ์เท่ากับ 36 (C(n) ΔC(n) $\Delta\Delta$C(n)) จากการทดสอบจะให้อัตราการรู้จำสูงขึ้น มากกว่าการใช้เพียงคุณสมบัติเชิง static อย่างเดียว

ระบบการรู้จำเสียงพูดที่ใช้เสียงที่ปราศจากสัญญาณรบกวนฝึกสอนให้กับระบบ เรียกว่าการฝึกสอนแบบมาตรฐาน (Standard Training) ประสิทธิภาพในการรู้จำที่ทดสอบกับเสียงที่ปราศจากสัญญาณรบกวน (clean speech) การประมวลผลเบื้องต้นทั้ง 3 วิธีให้ผลการรู้จำถูกต้องสูงใกล้เคลียงกัน แต่เมื่อทดสอบกับสัญญาณเสียงพูดที่มีสัญญาณรบกวน (noise speech) แบบเกาส์เซียนมีประสิทธิภาพในการรู้จำลดลง สำหรับการประมวลผลเบื้องต้นแบบ LPCC มีอัตราการรู้จำระหว่าง 94.0% ถึง 27.6% ประมวลผลเบื้องต้นแบบ MFCC มีอัตราการรู้จำระหว่าง 94.0% ถึง 32.3% และการประมวลผลเบื้องต้นแบบ PLP มีอัตราการรู้จำระหว่าง 96.3% ถึง 35.3% ขึ้นอยู่กับระดับของสัญญาณรบกวน แสดงให้เห็นว่าระบบการรู้จำเสียงพูดที่พัฒนากันอยู่ทุกวันนี้ให้ผลการรู้จำถูกต้องสูง เฉพาะระบบที่เสียงพูดที่ใช้ฝึกสอน กับเสียงพูดที่ใช้ทดสอบการรู้จำ เป็นเสียงที่ไม่มีความแปรปรวน อย่างไรก็ตามการประมวลผลเบื้องต้นแบบ PLP ก็ยังให้ผลการรู้จำถูกต้องสูงเป็นที่น่าพอใจกว่าวิธีอื่น

ระบบการรู้จำเสียงพูดที่ใช้ทั้งเสียงที่ปราศจากสัญญาณรบกวน และเสียงที่มีสัญญาณรบกวนที่มีระดับของสัญญาณรบกวน –45dB, –40dB, –35dB และ –30dB ฝึกสอนให้กับระบบ เรียกว่าการฝึกสอนแบบหลากหลาย (Multistyle Training) ระบบการรู้จำเสียงพูดที่ใช้การฝึกสอนแบบนี้ทำให้ระบบสามารถรู้จำเสียงมีประสิทธิภาพดีขึ้น โดยเฉพาะอย่างยิ่งใช้การประมวลผลเบื้องต้นแบบ PLP ซึ่งให้ผลการรู้จำถูกต้องสูงขึ้น 28.7% เมื่อเทียบกับระบบที่ฝึกสอนแบบมาตราฐาน ที่ระดับสัญญาณรบกวนแบบเกาส์เซียน –30dB

เราได้ทดสอบกับเสียงพูดที่มีสัญญาณรบกวน 3 แบบ คือแบบเกาส์เซียน แบบเสียงพูดแทรก และแบบเสียงรถ จากผลการทดลองแสดงให้เห็นว่าสัญญาณรบกวนรบกวนแบบเกาส์เซียนมีผลกระทบกับประสิทธิภาพของการรู้จำมากที่สุด และสัญญาณรบกวนรบกวนแบบเสียงรถ มีผลกระทบกับประสิทธิภาพของการรู้จำน้อยที่สุด

จากที่กล่าวมาทั้งหมดแสดงให้เห็นว่าการที่จะนำระบบการรู้จำเสียงพูดแบบคำโดดไม่ขึ้นอยู่กับผู้พูดไปประยุกต์ใช้ในชีวิตประจำวันนั้น ตอนนี้ระบบการรู้จำเสียงพูดที่พัฒนาขึ้นในประเทศ

ไทย ยังไม่สามารถนำออกไปใช้งานได้จริง เพราะระบบที่งานวิจัยส่วนใหญ่ทำขึ้นจะมีประสิทธิภาพดีสำหรับทดลองในห้องปฏิบัติงานเท่านั้น เป็นเหตุให้วิทยานิพนธ์นี้ได้ทำการศึกษาให้เห็นถึงผลกระทบเมื่อเสียงที่ทดสอบ มีคุณภาพต่างจากเสียงที่ใช้ฝึกสอน ถึงแม้จะได้ทดลองใช้การประมวลผลเบื้องต้นแบบ PLP และได้ทดลองทำการฝึกสอนแบบหลากหลาย ก็ยังมีประสิทธิภาพดีกับระดับสัญญาณรบกวนต่ำๆ เท่านั้น แต่เมื่อเสียงพูดมีสัญญาณรบกวนมากๆ ประสิทธิภาพการรู้จำก็จะลดลง ดังนั้นถ้าเราจะนำระบบการรู้จำเสียงพูดไปประยุกต์ใช้เราจำเป็นต้องทำการศึกษาเพิ่มเติมโดยเฉพาะอย่างยิ่งระบบการรู้จำที่เรียกว่า Robust Speech Recognition

ความผิดพลาดในการรู้จำที่เกิดขึ้นส่วนใหญ่เป็นแบบสุ่ม ซึ่งคาดว่าเกิดจากสาเหตุดังนี้ จากวิธีการหาจุดเริ่มต้น และจุดสิ้นสุดของเสียงพูดยังไม่ดีพอ ทำให้ข้อมูลเสียงที่สำคัญถูกตัดทิ้งไปได้ จากการผิดพลาดของผู้พูด ที่อาจทำให้สัญญาณเสียงพูดผิดเพี้ยนได้ จากการดึงลักษณะที่สำคัญของเสียงยังไม่เหมาะกับเสียงที่เป็นภาษาไทย เพราะเสียงภาษาไทยประกอบไปด้วยหลายส่วน ที่แตกต่างจากภาษาอื่น คือเสียงวรรณยุกต์ อาจทำให้ข้อมูลส่วนนี้ขาดหายไป

เอกสารอ้างอิง

[1] จิตรลดา จารุมิศร์, "การออกแบบ แบบจำลองในการรู้จำเสียงวรรณยุกต์สำหรับภาษาไทย โดยใช้เทคนิคการควอนไตซ์พิทช์ และ Hidden Markov Modeling.", วิทยานิพนธ์ปริญญามหาบัณฑิต, สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง. พ.ศ. 2542

[2] ณัฐกร ทับทอง, "การรู้จำคำพูดภาษาไทยโดยใช้ลักษณะบ่งความต่างของหน่วยเสียง.", วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. พ.ศ. 2538

[3] ระพีพัฒน์ อารีย์พงศา, "การรู้จำเสียงพูดตัวเลขไทยโดยไม่ขึ้นอยู่กับผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปิง.", วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. พ.ศ. 2538

[4] สมชาย จิตะพันธ์กุล. "การรู้จำเสียงพูดภาษาไทย ระยะที่หนึ่ง : การรู้จำเสียงพูดคำไทยโดดๆ โดยไม่ขึ้นกับผู้พูด." โครงการวิจัยเลขที่ 45G-EE-2538 สถาบันวิจัยและพัฒนาของคณะวิศวกรรมศาสตร์ คณะวิศวกรรมศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย. 2540

[5] เสาวลักษณ์ อารีย์พงศา, " การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดนมาร์คอฟโมเดล และเวกเตอร์ควอนไทซ์เซชัน. ", วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. พ.ศ. 2538

[6] วุฒิพงษ์ พรสุขจันทรา, " การรู้จำเสียงตัวเลขภาษาไทยแบบไม่ขึ้นอยู่กับผู้พูดโดยใช้แอลพีซีและนิวรอลเน็ตเวิร์กแบบแบ็กพรอพาเกชัน. ", วิทยานิพนธ์ปริญญามหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย. พ.ศ. 2539

[7] Charles R., Jankowski Jr., Hoang-Doan H. Vo, Richard P. Lippmann, "A Comparrison of Signal Processing Front Ends for Automatic Word Recognition." IEEE Transactions on Speech and Audio Processing., Vol. 3, No. 4, July 1995. pp. 286-293

[8] Chen, S.H., and Wang, Y.R., "Tone Recognition of Continuous Mandarin Speech Based on Neural networks." IEEE Transactions on Speech and Audio Processing., Vol. 3, March 1995. pp. 146-150

[9] H. Hermansky, " Perceptual linear predictive (PLP) analysis for speech ", J. Acoust. Soc. Amer., pp: 1738-1752, 1990

[10]L.R. Rabiner and B.H. Juang, "Fundamental of Speech Recognition", New jersey : Prentice Hall, 1993

[11] Rebiner, L.R. and Levinson, S.E. "Isolated and Connected Word Recognition-Theory and Selected Applications." IEEE Trans. On Comm., Vol. COM-29, No. 5, May 1981. pp. 621-659

[12] Rebiner, L.R. and Schmidt, C.E. "Application of Dynamic Time Warping to Connected Digit Recognition." IEEE Trans. Acoustic. Speech. Signal Processing., Vol. ASSP-28, No. 4, August 1980. pp. 377-388

[13] Steven F. Boll., " Suppresion of acoustic noise in speech using spectral subtraction ", IEEE Transaction on Acoustics, Speech and Signal Processing, Vol.ASSP27, no 12, pp: 113-120, April 1979

[14] Steve .Y,Dan K.,Julian O.,Dave O.,Valtcho V.,Phil W., "The htk book (for htk version 2.2)", 1995-1999 Entropic

[15] R. Kongkachandra, K. Tamee and C. Kimpan., "Using Karhunen-Loeve Transformation for Feature Reduction and Tones Analysis in Thai Harmonic-Frequency Speech." IEEE Internatoional Symposium . On Intelligent Signal Processing and Communication Systems.,Dec 1999. pp. 793-796

[16] R. Kongkachandra, K. Tamee and C. Kimpan., "Improving Thai Isolated Word Recognition by Using Karhunen-Loeve Transformationand Learning Vector Quantization." IEEE Internatoional Symposium . On Intelligent Signal Processing and Communication Systems.,Dec 1999. pp. 777-780

ประวัติผู้เขียน

นายเกรียงศักดิ์ เตมีย์ เกิดเมื่อวันที่ 10 พฤศจิกายน 2518 สำเร็จการศึกษาวิทยาศาสตร์บัณฑิต (ฟิสิกส์) จากมหาวิทยาลัยเชียงใหม่ ปีการศึกษา 2540 ประวัติการทำงาน ลูกจ้างชั่วคราว ตำแหน่งผู้ช่วยวิจัย สังกัดสำนักวิจัยและบริการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ลูกจ้างชั่วคราว ตำแหน่งผู้ช่วยวิจัย สังกัด Information Science โครงการสำนักสื่อสารและเทคโนโลยีสารสนเทศ (ReCCIT) สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง