

การรู้จำเสียงคำไทยเฉพาะบุคคล

SPEAKER DEPENDENT VOICE RECOGNITION OF THAI WORD



ทัตเวท วีระวัฒน์

TASAWATE WEERAWAT

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2541

เลขที่.....

ISBN 974-622-155-8

เลขทะเบียน... 30949

วัน, เดือน, ปี... ๕ ๔ ก.ย. 2541

สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

SPEAKER DEPENDENT VOICE RECOGNITION OF THAI LANGUAGE

TASAWATE WEERAWAT



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

1998

ISBN 974-622-155-8

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



COPYRIGHT 1998

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อวิทยานิพนธ์	การรู้จำเสียงคำไทยเฉพาะบุคคล
นักศึกษา	นายพัศเวช วีระวัฒน์
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร.กอบชัย เดชหาญ
หลักสูตร	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
พ.ศ.	2541

บทคัดย่อ

เนื่องจากเสียงพูดในภาษาไทยมีระดับเสียงสูงต่ำของวรรณยุกต์แตกต่างจากระดับเสียงในภาษาอังกฤษ ดังนั้นการศึกษาวิธีทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจภาษาที่มนุษย์พูดได้โดยเฉพาะกับภาษาไทยเพื่อทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจเสียงคำพูดภาษาไทยได้ถูกต้องมากยิ่งขึ้น จึงจำเป็นที่จะต้องหาวิธีแยกความแตกต่างของเสียงวรรณยุกต์ทั้ง 5 เสียง คือ สามัญ เอก โท ตรี และจัตวา ที่มีอยู่ในภาษาไทยให้ได้และประยุกต์ใช้กับวิธีในการรู้จำเสียงพูดในภาษาอังกฤษที่มีอยู่เดิม วิทยานิพนธ์ฉบับนี้จำกัดการทดลองอยู่เฉพาะกับเสียงผู้พูดคำสั้นๆ ที่เป็นคนคนเดียวกับที่ได้ฝึกฝน (trained) ให้เครื่องรับรู้ในตอนแรกเท่านั้น (speaker dependent) ซึ่งสามารถนำไปประยุกต์ให้รู้จำเสียงพูดต่อเนื่องและไม่ขึ้นกับผู้พูดได้

Thesis Title	Speaker Dependent Voice Recognition of Thai Word
Student	Mr.Tasawate Weerawat
Thesis Advisor	Assoc.Prof.Dr. Kobchai Dejhan
Degree	Master of Engineering in Electrical Engineering
Year	1998

ABSTRACT

Every language has particular character same as English and Thai. The difference between English and Thai which concerned is the level of tone in Thai. Nowadays computer system is developed to recognize only English language which has no level of tone whereas Thai has 5 level of tone (Saman, Eke, Tho, Tree and Jattava). Thus this thesis is performed for developing the old methodology which is for English language only to be suitable for Thai Language. The limitation of this study is dependent voice recognition for only short word of one speaker who is trained. Finally, this is the concept of speaker dependent voice recognition of Thai word which can be applied to develop continue word and speaker independent voice recognition.

กิตติกรรมประกาศ

วิทยานิพนธ์ ฉบับนี้สำเร็จลงด้วยได้รับความช่วยเหลือจากหลายๆด้าน โดยเฉพาะ ดร. กอบชัย เดชหาญ อาจารย์ที่ปรึกษาที่ได้ช่วยแนะนำและแก้ปัญหาในการดำเนินการทำวิทยานิพนธ์ ฉบับนี้เป็นอย่างดี

ขอขอบคุณ ดร. ไกรสิน ส่วงวัฒนา และ ดร. อภิชาติ ตั้งทรงธรรม เป็นอย่างสูงที่ได้ให้คำแนะนำที่เป็นประโยชน์ในการทำวิทยานิพนธ์เรื่องนี้เป็นอย่างยิ่ง รวมทั้งขอขอบคุณน้องๆ ที่ทำวิจัยด้านการรู้จำเสียงพูดกับ ดร. ไกรสิน ซึ่งให้ความช่วยเหลือเรื่องเอกสารและซอฟต์แวร์เป็นอย่างดี

สุดท้ายขอกราบขอบพระคุณ บิดา มารดา ที่ให้การสนับสนุนและคอยให้กำลังใจ จนวิทยานิพนธ์นี้สำเร็จลงด้วยดี

ทศเวท วีระวัฒน์

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....	I
บทคัดย่อภาษาอังกฤษ.....	II
กิตติกรรมประกาศ.....	III
สารบัญ.....	IV
สารบัญตาราง.....	VII
สารบัญภาพ.....	VIII
บทที่ 1 บทนำ.....	1
1.1 บทนำ.....	1
1.2 การจัดเนื้อเรื่องของวิทยานิพนธ์.....	2
บทที่ 2 ระบบการพูดและเสียงคำพูดในภาษาไทย.....	4
2.1 กายวิภาคของระบบการเปล่งเสียงของมนุษย์.....	4
2.2 กระบวนการผลิตเสียงพูด.....	6
2.3 เสียงพูดของมนุษย์.....	7
2.4 ความถี่ฟอร์แมนต์.....	8
2.5 แบบจำลองระบบกำเนิดเสียงพูด.....	8
2.6 เสียงพูดในภาษาไทย.....	9
2.7 เสียงสระ.....	10
2.7.1 ลักษณะของเสียงสระ.....	10
2.7.2 อวัยวะที่ทำให้เกิดความแตกต่างกันของเสียงสระ.....	10
2.7.3 การแบ่งประเภทของเสียงสระ.....	11

สารบัญ

หน้า

2.8 เสียงพยัญชนะ.....	12
2.8.1 ลักษณะของเสียงพยัญชนะ.....	12
2.8.2 อวัยวะที่ทำให้เกิดความแตกต่างกันของเสียงพยัญชนะ.....	12
2.8.3 การแบ่งประเภทของเสียงพยัญชนะ.....	14
2.8.4 ไตรยางค์.....	16
2.9 เสียงวรรณยุกต์.....	16
2.9.1 ลักษณะของเสียงวรรณยุกต์.....	17
2.9.2 ระดับเสียงวรรณยุกต์.....	17
2.9.3 การเทียบเสียงวรรณยุกต์.....	17
บทที่ 3 การคำนวณหาพิตช์.....	19
3.1 คุณลักษณะของสัญญาณเสียง.....	19
3.2 การหาจุดเริ่มและจุดท้ายของเสียง.....	20
3.3 การคำนวณออกโศคอร์รีเรชัน.....	21
3.4 การปรับแต่งค่าพิตช์.....	23
บทที่ 4 การประมาณเชิงเส้น.....	25
4.1 บทนำ.....	25
4.2 การฟรีเอมฟาซิส.....	25
4.3 การแบ่งช่วงสัญญาณ.....	28
4.4 การวินโดว์.....	29
4.5 การหาค่าสัมประสิทธิ์ด้วยวิธีออกโศคอร์รีเรชัน.....	30
4.6 การหาค่าสัมประสิทธิ์ α และอัตราขยาย G	34
4.7 การเปลี่ยนพารามิเตอร์ LPC เป็น สัมประสิทธิ์เชปสตรัม.....	35
4.8 การเวทค่าพารามิเตอร์.....	36

สารบัญ

หน้า

บทที่ 5 การควอนไทซ์.....	37
5.1 บทนำ.....	37
5.2 การควอนไทซ์แบบเวกเตอร์.....	37
5.3 ส่วนประกอบของ การควอนไทซ์แบบเวกเตอร์.....	39
5.3.1 การทำงานของการควอนไทซ์แบบเวกเตอร์.....	40
5.3.2 เซตเทรนนิ่ง (Training) ของการควอนไทซ์แบบเวกเตอร์.....	40
5.4 หลักการของการควอนไทซ์แบบเวกเตอร์.....	41
5.5 การคำนวณการควอนไทซ์แบบเวกเตอร์.....	41
5.6 การวัดความคลาดเคลื่อน.....	42
5.7 การจัดกลุ่มเพื่อสร้างแบบอ้างอิงโดยใช้ K-means Algorithm.....	42
5.7.1 การหาจุดศูนย์กลางของกลุ่ม (Center Cluster).....	43
5.7.2 การจัดกลุ่มใหม่ (Reclassification).....	43
5.8 การคำนวณจุดศูนย์กลางสำหรับการออกแบบตารางรหัสการควอนไทซ์แบบเวกเตอร์.....	44
5.9 กฎของการตัดสินใจ.....	45
5.9.1 กฎ Nearest-Neighbor (NN).....	45
5.9.2 กฎ K Nearest-Neighbor (KNN).....	45
5.10 การควอนไทซ์แบบเวกเตอร์สำหรับระบบการจดจำเสียงพูด	46
บทที่ 6 แบบจำลองมาร์คอฟ.....	49
6.1 บทนำ.....	49
6.2 กระบวนการ แบบจำลองมาร์คอฟ แบบไม่ต่อเนื่องทางเวลา.....	49
6.3 ส่วนประกอบของ แบบจำลองมาร์คอฟ.....	50
6.4 ชนิดของ HMM.....	51

สารบัญ

หน้า

6.5 ปัญหาพื้นฐานของแบบจำลองมาร์คอฟ.....	53
6.5.1 ปัญหาที่ 1.....	53
6.5.2 ปัญหาที่ 2.....	54
6.5.3 ปัญหาที่ 3.....	54
6.6 การแก้ปัญหาค่าพื้นฐานของแบบจำลองมาร์คอฟ.....	55
6.6.1 การแก้ปัญหาค่าที่ 1 การคำนวณหาค่าความน่าจะเป็น.....	55
6.6.1.1 กระบวนการไปหน้า (Forward Procedure).....	56
6.6.1.2 กระบวนการย้อนกลับ (Backward Procedure).....	58
6.6.2 การแก้ปัญหาค่าที่ 2 การหาลำดับสเตตที่เหมาะสม.....	59
6.6.3 การแก้ปัญหาค่าที่ 3.....	60
6.7 การปรับปรุงค่าพารามิเตอร์ของ HMM.....	63
6.7.1 การสเกล (Scaling).....	63
6.7.2 ลำดับของเหตุการณ์หลายลำดับ (Multiple Observation Sequence).....	66
6.8 ระบบแบบจำลองมาร์คอฟ.....	68
บทที่ 7 การทดลองและผลการทดลอง.....	70
7.1 แบบจำลองการรู้จำเสียงภาษาไทย.....	70
7.2 ค่าพารามิเตอร์ต่างๆ ที่ใช้ในการสร้างแบบจำลอง.....	71
7.3 การเลือกกลุ่มตัวอย่างของเสียงและผลการทดลอง.....	74
7.3.1 แบบจำลองส่วนที่ 1.....	74
7.3.2 แบบจำลองส่วนที่ 2 (ส่วนที่ใช้ในการรู้จำเสียงวรรณยุกต์).....	81
7.3.3 แบบจำลองส่วนที่ 1 และส่วนที่ 2 รวมกัน.....	81
บทที่ 8 สรุปผลการทดลอง.....	83

สารบัญ

หน้า

บรรณานุกรม.....85

ภาคผนวก.....87

ประวัติผู้เขียน..... 97



สารบัญตาราง

ตารางที่

หน้า

2.1	เสียงพ้องขณะต้น.....	14
2.2	เสียงพ้องขณะท้ายพยางค์.....	15
7.1	แสดงผลการทดสอบคำพูดของเสียงตัวเลขศูนย์ถึงเก้า.....	75
7.2	แสดงผลการทดสอบคำพูดของเสียงที่มีเสียงพื้นสามัญเสียงเดียวกัน.....	75
7.3	ผลการทดสอบการรู้จำของเสียงที่ใช้ในการฝึกจำของแบบจำลอง HMM.....	78
7.4	ผลการทดสอบการรู้จำของเสียงที่ไม่เคยใช้ในการฝึกจำของแบบจำลอง HMM.....	79
7.5	ผลสรุปการทดสอบการรู้จำของเสียงทดสอบกับแบบจำลอง HMM ของเสียง วรรณยุกต์.....	79
7.6	ผลการทดสอบการรู้จำของคำที่มีพื้นเสียงเดียวกัน.....	82

สารบัญภาพ

ภาพที่

หน้า

1.1 แผนภาพแสดงระบบรู้จำเสียงในภาษาไทย.....	2
2.1 ภาพตัดขวางแสดงอวัยวะในระบบการพูดของมนุษย์.....	4
2.2 รูปกล่องเสียงขณะ (ก) หายใจปกติ (ข) หายใจลึกๆ (ค) กำลังส่งเสียง (ง) ส่งเสียงกระซิบหรือเสียงแผ่ว.....	5
2.3 แผนภาพระบบเสียงพูดของมนุษย์.....	6
2.4 (ก) ตัวอย่างรูปคลื่นของสัญญาณเสียงก้อง /a/ (ข) ตัวอย่างรูปคลื่นของสัญญาณเสียงไม่ก้อง /sh/.....	7
2.5 สเปกตรัมของเสียง /a/ ที่ได้จากสัมประสิทธิ์ LPC.....	8
2.6 บล็อกไดอะแกรมจำลองระบบกำเนิดเสียงเริ่มต้น.....	9
2.7 โครงสร้างของเสียงพูดภาษาไทย.....	9
2.8 ระดับเสียงวรรณยุกต์ (ระดับการเปลี่ยนแปลงของพิตช์).....	18
3.1 ตัวอย่างคาบเวลาพิตช์ของเสียงพูด.....	19
3.2 เฟรมเริ่มต้น(T1) และเฟรมสุดท้าย(T2) ของพลังงานของเสียงพูดที่ใช้ในการคำนวณ.....	20
3.3 ออโตคอร์รีเรชันของสัญญาณเสียงพูด.....	22
3.4 ตัวอย่างเสียงพูดของคำว่า ‘จิม’.....	23
4.1 แสดงสเปกตรัมของสัญญาณเสียงเทียบกับสเปกตรัมที่ได้จากการประมาณเชิงเส้นที่มี อันดับ (LPC order) เท่ากับ p ต่างๆ (ก) p=4 (ข) p=8 (ค) p=12 (ง) p=16 (จ) p=20.....	26
4.2 แสดงขั้นตอนการเตรียมสัญญาณในการวิเคราะห์.....	27
4.3 แสดงขนาดสเปกตรัมของฟังก์ชันถ่ายโอนของการพรีเอมฟาสิส.....	28
4.4 แสดงการแบ่งช่วงของสัญญาณที่ใช้ในการวิเคราะห์.....	28
4.5 แสดงวินโดว์สี่เหลี่ยม (ก) ในโดเมนเวลา (ข) ในโดเมนความถี่.....	30
4.6 แสดงวินโดว์แบบแฮมมิง (ก) ในโดเมนเวลา (ข) ในโดเมนความถี่.....	30
4.7 บล็อกไดอะแกรมแสดงโมเดลการสร้างสัญญาณเสียงพูดอย่างง่าย.....	31

สารบัญภาพ

ภาพที่

หน้า

5.1 แสดงข้อมูลที่จะทำการควอนไทซ์.....	38
5.2 แสดงการจับกลุ่มและครรชนีของตารางรหัส (code book) ที่ได้.....	39
5.3 แสดงบล็อกไดอะแกรมของการควอนไทซ์แบบเวกเตอร์.....	39
5.4 แสดงเวกเตอร์ตัว ควอนไทซ์ที่ใช้ในระบบการรับรู้เสียงพูด.....	46
6.1 ตัวอย่างแบบจำลองมาร์คอฟ 5 สเตต.....	50
6.2 แสดงแบบจำลองต่าง ๆ ของ HMM.....	52
6.3 กระบวนการไปข้างหน้า.....	57
6.4 กระบวนการถอยหลัง.....	58
6.5 แสดงลำดับการคำนวณเหตุการณ์ร่วมซึ่งระบบจะอยู่ในสเตต i ที่เวลา t และอยู่ที่สเตต j ที่เวลา $t+1$	61
6.6 บล็อกไดอะแกรม ของการรู้จำคำโคคด้วยแบบจำลองของมาร์คอฟ.....	69
7.1 แผนภาพแสดงการรู้จำเสียงในภาษาไทยด้วยแบบจำลอง HMM.....	70
7.2 แสดงการแบ่งช่วงเฟรมข้อมูล.....	72
7.3 แสดงความสัมพันธ์ระหว่างค่าความคลาดเคลื่อนกับจำนวนสเตต.....	73
7.4 ค่าเซปตัมของคำ ก) โป ข) โป่ ค) โป้ ง) โป๊ จ) โป๋.....	76
7.5 ค่าเซปตัมของคำ ก) คา ข) ข่า ค) คำ-ข้า ง) คำ จ) ขา.....	77
7.6 ค่าพิตช์ของคำในตารางที่ 7.3 * ก) เสียงสามัญ ข) เสียงเอก ค) เสียง โท ง) เสียงตรี จ) เสียงจัตวา.....	80

บทที่ 1

บทนำ

1.1 บทนำ

การที่เครื่องคอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้นั้น มาซึ่งประโยชน์ต่อมนุษย์มากมาย เช่น การสั่งงานด้วยเสียง, การป้องกันความปลอดภัยโดยใช้เทคโนโลยีการจำเสียงผู้ที่เป็นเจ้าของ หรือการสั่งให้คอมพิวเตอร์พิมพ์ตามที่มนุษย์พูดแทนการพิมพ์คีย์ ในปัจจุบันจะพบกับเครื่องมือที่สามารถรู้จำภาษามนุษย์ได้ซึ่งส่วนใหญ่จะใช้ได้กับภาษาอังกฤษเป็นหลัก แต่เนื่องจากเสียงพูดในภาษาไทยเป็นเสียงพูดที่มีระดับโทนเสียงสูงต่ำหรือภาษาเสียงวรรณยุกต์ (tonal language) ซึ่งแตกต่างจากภาษาอังกฤษที่มีเพียงแต่เพียงเสียงสระและพยัญชนะเท่านั้น เพื่อให้เครื่องคอมพิวเตอร์สามารถเข้าใจภาษาไทยได้ถูกต้องมากยิ่งขึ้นจึงจำเป็นต้องหาวิธีแยกความแตกต่างของเสียงวรรณยุกต์ทั้ง 5 เสียง คือ สามัญ เอก โท ตรี และ จัตวา ที่มีอยู่ในภาษาไทย

ในภาษาที่ไม่ใช่ภาษาเสียงวรรณยุกต์ คำแต่ละคำจะเปลี่ยนระดับเสียงไปอย่างไรก็ยังคงความหมายเดิม เช่น คำ *คาร์ (car)* ในภาษาอังกฤษ ถึงแม้จะออกเสียงเพี้ยนเป็น ข้า คำ (ข้า) คำ ขา ก็ยังคงความหมายว่า *รถยนต์* เหมือนกันทุกคำ ซึ่งวิธีการรู้จำเสียงพูดที่มีใช้กับภาษาอังกฤษไม่เหมาะที่จะใช้แยกความแตกต่างนี้ เช่น เสียงคำว่า โอ โอ้ โอ้อ โอ้อ โอ้อ ในภาษาอังกฤษจะแทนเสียงทั้ง 5 เสียงเป็นคำว่า *oh* เพียงคำเดียว เป็นต้น

ดังนั้นการที่จะทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจภาษาไทยได้นั้นจึงจำเป็นต้องมีการพัฒนาวิธีการเพิ่มเติมนอกเหนือจากวิธีการรู้จำเสียงในภาษาอังกฤษ เพื่อให้แบบจำลอง (model) ที่ใช้ในการรู้จำมีความสามารถในการรู้จำภาษาไทยได้ถูกต้องมากยิ่งขึ้น โดยมีวัตถุประสงค์ในการทำวิทยานิพนธ์ฉบับนี้ดังนี้

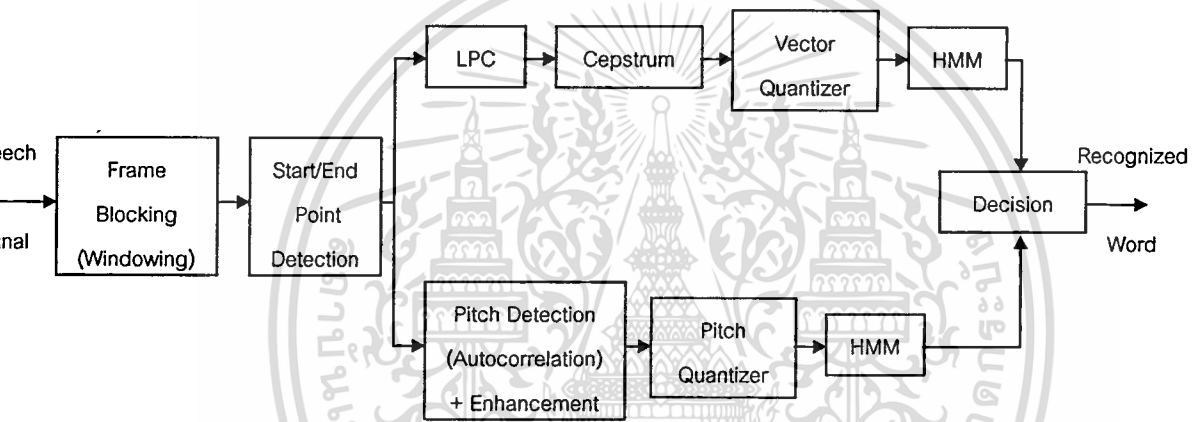
1. ศึกษาวิธีการที่จะทำให้เครื่องคอมพิวเตอร์สามารถรู้จำเสียงพูดคำสั้นๆ ได้
2. ศึกษาวิธีเฉพาะที่จะนำมาประยุกต์ใช้กับการรู้จำเสียงพูดในภาษาไทย
3. เพื่อเป็นพื้นฐานในการพัฒนาให้เครื่องคอมพิวเตอร์สามารถรู้จำเสียงที่เป็นประโยคต่อเนื่อง
4. เพื่อศึกษาและเป็นพื้นฐานในการรู้จำ เสียงในแบบแยกองค์ประกอบของเสียงในภาษาไทยทั้ง 4 ประเภทคือ เสียงพยัญชนะต้น เสียงสระ เสียงวรรณยุกต์ และเสียงพยัญชนะท้ายพยางค์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การศึกษาวิจัยในวิทยานิพนธ์นี้กำหนดขอบเขตไว้ที่การรู้จำเสียงพูดสั้นๆ หรือคำโดด (คำที่มี 1 พยางค์) ในภาษาไทยเฉพาะบุคคลเท่านั้น ซึ่งสามารถนำไปพัฒนาต่อให้สามารถรู้จำเสียงคำต่อเนื่องมากกว่า 1 พยางค์ และไม่ขึ้นกับบุคคลได้

การเก็บตัวอย่างข้อมูลเสียงที่ใช้ในการวิจัยนี้กระทำได้โดยการสุ่มตัวอย่างโดยใช้เครื่องคอมพิวเตอร์ส่วนบุคคลพร้อมการ์ดเสียงเพื่อใช้เป็นตัวอย่างข้อมูลเสียงในการทดลอง และใช้โปรแกรม MATLAB เป็นเครื่องมือหลักในการเขียนโปรแกรมเพื่อทดลองวิธีการรู้จำเสียงตามรูปที่ 1.1

รูปที่ 1.1 แผนภาพแสดงระบบรู้จำเสียงในภาษาไทย



1.2 การจัดเนื้อเรื่องของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 8 บท ดังนี้

บทที่ 1 บทนำ

บทที่ 2 กล่าวถึงระบบการพูดของมนุษย์ อวัยวะที่ใช้ในการออกเสียง รวมทั้งระบบโครงสร้างของเสียงคำพูดในภาษาไทย

บทที่ 3 เป็นการหาค่าพิตช์ ซึ่งก็คือค่าความถี่ในการสั่นของเส้นเสียง โดยใช้วิธีในการตัดข้อมูลเสียงส่วนที่ไม่จำเป็นออก รวมทั้งศึกษาวิธีการลดความผิดพลาดของค่าพิตช์ ที่คำนวณได้เนื่องจากการออกเสียงพยัญชนะ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

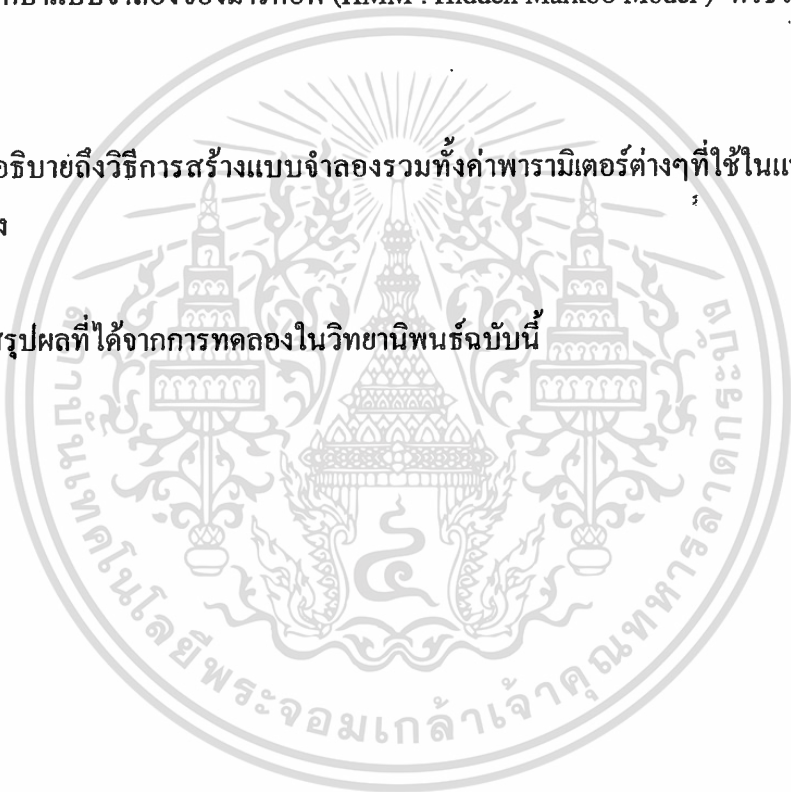
บทที่ 4 กล่าวถึงการประมารเชิงเส้นซึ่งค่าที่ได้มีผลเสมือนเป็นค่าประมาณของสเปคตรัมของเสียง แต่ปริมาณข้อมูลที่ได้น้อยกว่า ซึ่งทำให้สามารถลดข้อมูลการคำนวณในกระบวนการถัดไปได้

บทที่ 5 กล่าวถึงการควอนไดซ์แบบเวกเตอร์ ซึ่งจะช่วยให้ช่วยลดปริมาณข้อมูลให้ลดน้อยลงแต่ยังคงระดับความผิดพลาดในการควอนไดร์ไว้ในระดับต่ำเมื่อเทียบกับจำนวนระดับที่เท่ากันของการควอนไดซ์แบบธรรมดา (PCM : Pulse Code Modulation)

บทที่ 6 เป็นการศึกษาแบบจำลองของมาร์คอฟ (HMM : Hidden Markov Model) ที่ใช้ในการเรียนรู้และรู้จำเสียงพูด

บทที่ 7 เป็นการอธิบายถึงวิธีการสร้างแบบจำลองรวมทั้งค่าพารามิเตอร์ต่างๆที่ใช้ในแบบจำลอง และผลการทดลอง

บทที่ 8 เป็นการสรุปผลที่ได้จากการทดลองในวิทยานิพนธ์ฉบับนี้



บทที่ 2

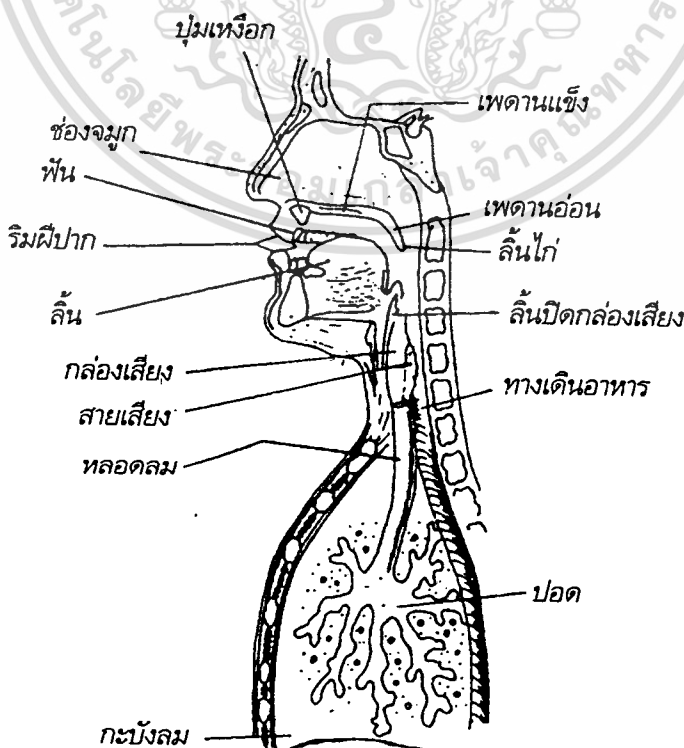
ระบบการพูดและเสียงพูดในภาษาไทย

2.1 กายวิภาคของระบบการเปล่งเสียงของมนุษย์

จากการศึกษาด้าน กายวิภาคศาสตร์ของมนุษย์ (human anatomy) วิชาที่ว่าด้วย เสียงของ ภาษา (phonetics) และศาสตร์ทางด้านเสียง (acoustics) ช่วยให้เข้าใจ ขั้นตอนการทำงานร่วมกัน ของอวัยวะต่างๆ ในการเปล่งเสียงพูด ตลอดจนลักษณะทางกายภาพของเสียงพูด เพื่อนำมาวิเคราะห์ และสร้างแบบจำลองเลียนแบบเสียงพูดของมนุษย์ [10]

การทำให้เกิดเสียงเป็นหน้าที่หนึ่งของระบบหายใจ การออกเสียงหรือการพูดของมนุษย์แต่ละครั้งจะต้องมีการทำงานร่วมกันของอวัยวะต่าง ๆ ของร่างกาย ดังรูปที่ 2.1 อันประกอบด้วย

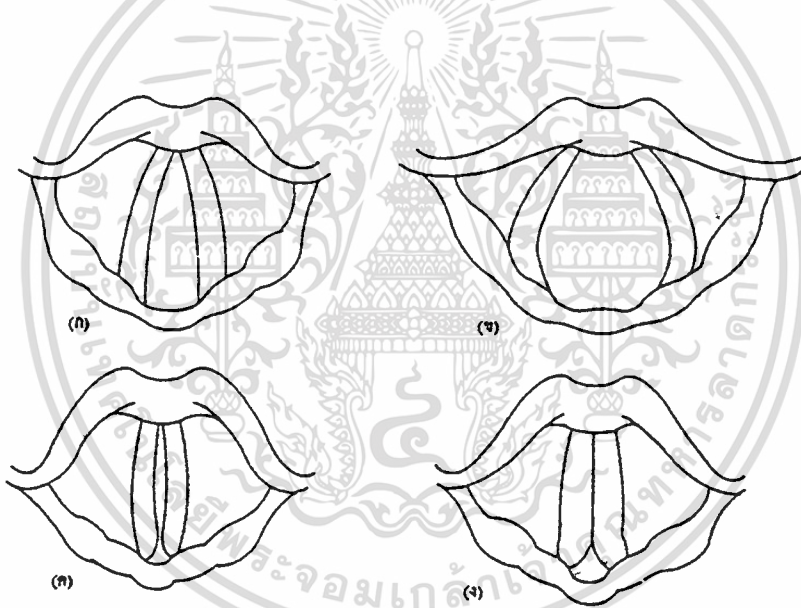
รูปที่ 2.1 ภาพตัดขวางแสดงอวัยวะในระบบการพูดของมนุษย์



1. ปอดและกระบังลม ทำหน้าที่สำคัญในการหายใจ และเป็นต้นกำเนิดการไหลของอากาศในกระบวนการผลิตเสียง
2. หลอดลม ทำหน้าที่นำอากาศจากปอดผ่านกล่องเสียง และเป็นอวัยวะที่อยู่ด้านหน้าของหลอดอาหาร
3. กล่องเสียง เป็นอวัยวะพิเศษที่ทำหน้าที่เป็นทางเดินอากาศเวลาหายใจ และเป็นตัวผลิตพัลส์ (pulse) ของอากาศขณะเปล่งเสียง ซึ่งประกอบด้วยเส้นเสียง (vocal cords) และช่องสายเสียง (glottis) รูปร่างของกล่องเสียงและเส้นเสียงในลักษณะต่าง ๆ แสดงในรูปที่ 2.2

รูปที่ 2.2 รูปกล่องเสียงขณะ (ก) หายใจปกติ (ข) หายใจลึกๆ

(ค) กำลังส่งเสียง (ง) ส่งเสียงกระซิบหรือเสียงแผ่ว



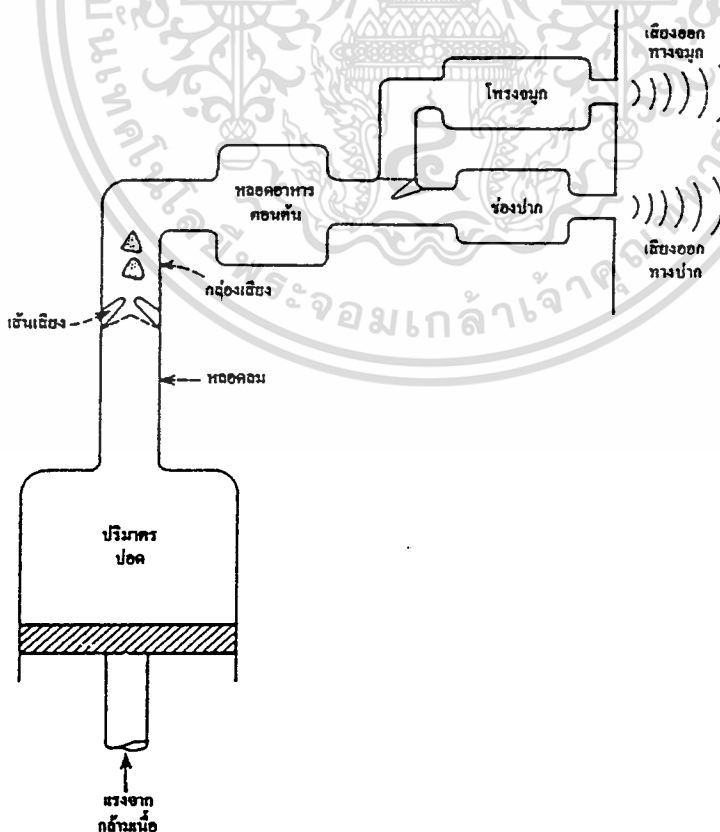
4. ช่องปากและส่วนของหลอดอาหารตอนต้น อวัยวะกลุ่มนี้อยู่ต่อจากกล่องเสียงอาจเรียกว่าอวัยวะกำทอนเสียง (vocal tract) ทำหน้าที่กำทอนเสียง โดยให้กำทอนทั้งเสียงที่เกิดจากกล่องเสียงและเสียงที่เกิดภายในช่องปาก ขนาดของอวัยวะกำทอนเสียงขึ้นอยู่กับตำแหน่งของลิ้น ริมฝีปาก ขากรรไกร และ เพดานอ่อน และเปลี่ยนแปลงไปตามการออกเสียง
5. โพรงจมูก เริ่มจากเพดานอ่อนจนถึงรูจมูกทั้งสอง ทำหน้าที่กำทอนเสียงร่วมกับช่องปาก เมื่อมีการเปล่งเสียงที่ออกทางจมูก (nasal sounds) เช่นเสียง /ม/ ,/น/ และ /ง/ เป็นต้น

2.2 กระบวนการผลิตเสียงพูด

จากระบบเสียงพูด สามารถแสดงเป็นแผนภาพของระบบกำเนิดเสียง ดังรูปที่ 2.3 ซึ่งสามารถจำแนกกลไกการสร้างเสียงพูดของมนุษย์ได้ 3 แบบ ดังนี้

1. อากาศไหลจากปอดจะถูกมอดูเลท (modulate) โดยการสั่นของเส้นเสียงทำให้เกิดคลื่นเสียงลักษณะคล้ายพัลส์ที่มีความยาวแบบควอไซ (quasi-periodic pulse-like excitation)
2. อากาศที่ไหลจากปอดถูกทำให้ปั่นป่วน ด้วยการบังคับให้ไหลผ่านช่องแคบอันเกิดจากการบีบตัวของอวัยวะในช่องปากทำให้เกิดเสียงลักษณะคล้ายเสียงรบกวน (noise-like excitation)
3. อากาศที่ไหลถูกกัก และเกิดแรงดันอยู่ภายในส่วนของช่องปากที่ปิด จากนั้นจึงปล่อยให้ อากาศที่มีแรงดัน พุ่งออกไปอย่างรวดเร็ว ทำให้เกิดการกระตุ้นเป็นเสียงในช่วงเริ่มต้น (transient excitation)

รูปที่ 2.3 แผนภาพระบบเสียงพูดของมนุษย์

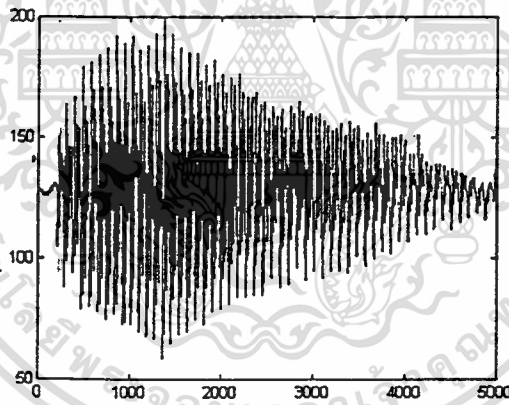


2.3 เสียงพูดของมนุษย์

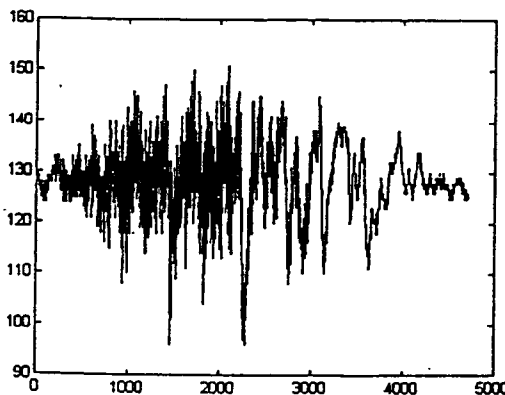
เสียงพูด เป็นคลื่นตามยาว (longitudinal wave) เกิดจากการสั่นของอนุภาคตัวกลางนั้นคือ อากาศ และทิศทางการสั่นของอนุภาคจะอยู่ในทิศเดียวกันกับทิศทางของการเคลื่อนที่ คลื่นเสียงเป็น คลื่นที่เปลี่ยนแปลงไปตามเวลา เสียงพูดแบ่งออกได้เป็น 2 ชนิดตามการกำเนิดเสียงหรือโหมด (mode) การกระตุ้น คือ

1. เสียงก้องหรือเสียงโฆษะ (voiced) เกิดจากการบังคับอากาศให้ผ่านช่องสายเสียงทำให้เกิดการเปลี่ยนแปลงความตึงหย่อนของเส้นเสียง โดยเส้นเสียงจะสั่นและเกิดเป็นพัลส์ (pulse) ของ อากาศไปกระตุ้นอวัยวะคำทอนเกิดเป็นเสียงก้อง ตัวอย่างเสียงก้องได้แก่ เสียงสระ เสียงพยัญชนะ ที่ต้องออกเสียงจากลำคอ (voiced consonants) เช่น เสียง /a/ ดังรูป 2.4

รูปที่ 2.4 (ก) ตัวอย่างรูปคลื่นของสัญญาณเสียงก้อง /a/



รูปที่ 2.4 (ข) ตัวอย่างรูปคลื่นของสัญญาณเสียงไม่ก้อง /sh/

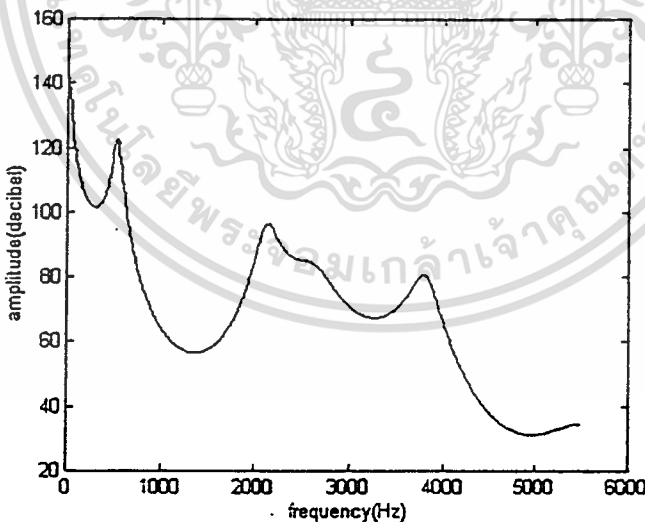


2. เสียงไม่ก้องหรือโห่ (unvoiced หรือ voiceless) เป็นเสียงที่เกิดในช่องปากหรือโพรงจมูก โดยอวัยวะภายในช่องปาก ริมฝีปาก ขว้างการไหลของอากาศให้ผ่านได้เป็นช่องเล็กๆ อากาศจึงไหลผ่านอย่างรวดเร็วและปั่นป่วนจนกระทั่งสร้างเป็นเสียงรบกวนช่วงความถี่กว้าง (broad-spectrum noise) ตัวอย่างเสียงไม่ก้องได้แก่ เสียงพยัญชนะที่ไม่ได้เกิดจากลำคอ (voiceless consonants) เช่น เสียง /sh/ ดังรูปที่ 2.4 (ข)

2.4 ความถี่ฟอร์แมนท์ (formant frequency)

ความถี่ฟอร์แมนท์ คือความถี่กำทอน (resonance frequency) ของกลุ่มอวัยวะกำทอนเสียง ความถี่ฟอร์แมนท์ของเสียงจะมีค่าเท่าใดขึ้นอยู่กับ การบังคับขนาดและรูปร่างของอวัยวะส่วนนี้ อันเกิดเป็นเสียงความถี่ที่แตกต่างกันนั้นคือเสียงพูดต่าง ๆ นั่นเอง ความถี่ฟอร์แมนท์สำหรับเสียงพูดคำหนึ่งๆ อาจมีได้หลายค่า ดังรูปที่ 2.5 เป็นสเปกตรัมของเสียง /a/ ซึ่งมีความถี่ฟอร์แมนท์ที่ 1,2,3 และ 4 อยู่ ณ ตำแหน่งความถี่ ที่ 550, 2150, 2600 และ 3750 เฮิรตซ์ตามลำดับ

รูปที่ 2.5 สเปกตรัมของเสียง /a/ ที่ได้จากสัมประสิทธิ์ LPC



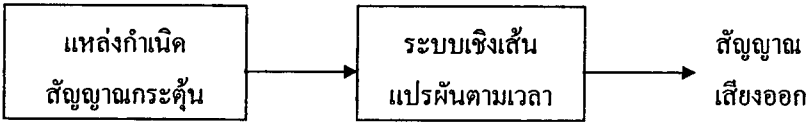
2.5 แบบจำลองระบบกำเนิดเสียงพูด

จากภาพรูปที่ 2.3 สามารถแสดงแผนภาพกรอบจำลองระบบกำเนิดเสียงเบื้องต้น ดังรูปที่

2.6 จากรูปมีการแยกภาคแหล่งกำเนิดสัญญาณกระตุ้น ออกจากส่วนกำทอนเสียง ซึ่งแทนด้วยระบบเชิงเส้นแปรผันตามเวลา

เอกสารนี้เป็นเอกสารสำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 2.6 แผนภาพกรอบจำลองระบบกำเนิดเสียงเริ่มต้น



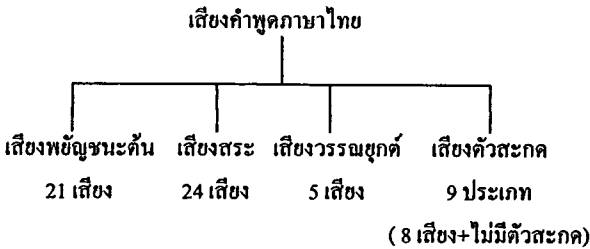
จากรูปที่ 2.6 แหล่งกำเนิดสัญญาณกระตุ้นทำหน้าที่แทนการทำงานของปอด และกล่องเสียงส่วนนี้จะผลิตขบวนพัลส์ที่มีคาบเวลาพิช้ขณะเปล่งเสียงว้อยซ์ และให้กำเนิดเสียงซึ่งคล้ายเสียงรบกวนขณะเปล่งเสียงอโห้หะ (unvoiced) ส่วนที่สองเป็นท่อกำทอนเสียง จะแทนการทำงานของช่องปากและโพรงจมูก ทำหน้าที่เสมือนตัวกรองสัญญาณ (filter) ที่ยอมให้ความถี่ฟอร์แมนท์ผ่านได้ ซึ่งสามารถแทนด้วยระบบเชิงเส้นแปรผันตามเวลา (time-varying linear system)

2.6 เสียงพูดในภาษาไทย

ภาษาส่วนมากประกอบด้วยเสียงสระและเสียงพยัญชนะ บางภาษาเท่านั้น เช่น ภาษาไทย และภาษาจีนที่มีเสียงวรรณยุกต์เพิ่มขึ้น เสียงวรรณยุกต์ คือ ระดับสูงต่ำของเสียงสระที่ทำให้ความหมายของคำเปลี่ยนไป กล่าวคือคำที่มีเสียงสระและเสียงพยัญชนะอย่างเดียวกันแต่ระดับเสียงของเสียงสระไม่เหมือนกัน จะมีความหมายต่างกันไปเป็นค่นละคำ เสียงคำพูดต่างๆ ในภาษาไทยที่เปล่งออกมาหนึ่งพยางค์จะประกอบไปด้วยการผสมเสียง 3 ชนิดคือ เสียงพยัญชนะ, เสียงสระ และเสียงวรรณยุกต์ โดยมีรูปแบบที่เป็นไปได้ 2 รูปแบบ [11] คือ

- 1. เสียงพยัญชนะต้นพยางค์ + เสียงสระ + เสียงวรรณยุกต์
- 2. เสียงพยัญชนะต้นพยางค์ + เสียงสระ + เสียงวรรณยุกต์ + เสียงพยัญชนะท้ายพยางค์

รูปที่ 2.7 โครงสร้างของเสียงพูดภาษาไทย



2.7 เสียงสระ

2.7.1 ลักษณะของเสียงสระ

เสียงสระเกิดจากกระแสลมจากปอดผ่านเส้นเสียงและเคลื่อนออกไปโดยสะดวก ไม่ถูกอวัยวะในช่องปากกีดขวางแต่อย่างใด ในขณะที่ลมผ่านนั้นเส้นเสียงจะเกร็งตัวตันสะบัดทำให้เกิดความกังวาน เสียงสระทุกเสียงจึงมีลักษณะกังวาน หรือที่เรียกว่าเสียงก้องหรือเสียงโฆษะ (voiced)

2.7.2 อวัยวะที่ทำให้เกิดความแตกต่างกันของเสียงสระ

เสียงสระเป็นเสียงที่เปล่งออกมาโดยตรง ไม่ถูกอวัยวะในช่องปากสกัดกั้นกั๊กก็ตาม แต่ก็ยังถูกอวัยวะบางส่วนในช่องปากกล่อมเกลา ทำให้เกิดเสียงแตกต่างกันไป อวัยวะสำคัญที่ทำหน้าที่กล่อมเกลาเสียงสระ ได้แก่ ลิ้นและริมฝีปาก การที่เสียงสระแตกต่างกันได้หลายเสียง เกิดจากการกล่อมเกลาของลิ้นกับริมฝีปาก

ลิ้น ในการกล่อมเกลาเสียงสระ ลิ้นมีส่วนเกี่ยวข้องกับสองลักษณะ คือ ส่วนของลิ้น และระดับความสูงของลิ้น ส่วนของลิ้น ได้แก่ ลิ้นส่วนหน้า (blade) ลิ้นส่วนกลาง (center) และลิ้นส่วนหลัง (dorsum)

ระดับความสูงของลิ้น มี 3 ระดับ คือ ระดับสูง (high) ระดับกลาง (mid) และระดับต่ำ (low) การกระดกลิ้นแต่ละส่วนขึ้นตามระดับต่างๆ จะเกิดเสียงสระแตกต่างกันเป็น 3 กลุ่ม คือ ถ้ากระดกลิ้นส่วนหน้าจะเกิดสระหน้า กระดกลิ้นส่วนกลางจะเกิดสระกลาง และกระดกลิ้นส่วนหลังจะเกิดสระหลัง ในขณะที่ลิ้นขึ้นเมื่อลิ้นกระดกขึ้นสูงระดับใดก็จะเกิดเสียงสระผิดแผกกันออกไปอีก 3 กลุ่ม เป็น สระระดับสูง สระระดับกลาง และสระระดับต่ำ โดยนับนี้ลิ้นทำหน้าที่กล่อมเกลาเสียงสระให้แตกต่างกันคือ

สระหน้าระดับสูง	สระกลางระดับสูง	สระหลังระดับสูง
สระหน้าระดับกลาง	สระกลางระดับกลาง	สระหลังระดับกลาง
สระหน้าระดับต่ำ	สระกลางระดับต่ำ	สระหลังระดับต่ำ

ริมฝีปาก ในการกล่อมเกลາเสียงสระให้แตกต่างกันคือ

1. ริมฝีปากเหยียดปิดหรือเกือบปิด ลิ้นส่วนหน้าหรือส่วนกลางจะอยู่ในระดับสูง เกิดสระหน้าระดับสูงหรือสระกลางระดับสูง
2. ริมฝีปากเหยียดกึ่งปิดหรือกึ่งเปิด ลิ้นส่วนหน้าหรือส่วนกลางจะอยู่ในระดับกลาง เกิดสระหน้าระดับกลางหรือสระกลางระดับกลาง
3. ริมฝีปากเหยียดเปิดกว้าง ลิ้นส่วนหน้าหรือส่วนกลางจะอยู่ในระดับต่ำ เกิดสระหน้าระดับต่ำหรือสระกลางระดับต่ำ
4. ริมฝีปากห่อกลมปิดหรือเกือบปิด ลิ้นส่วนหลังจะอยู่ในระดับสูง เกิดสระหลังระดับสูง
5. ริมฝีปากห่อกลมกึ่งปิดหรือกึ่งเปิด ลิ้นส่วนหลังจะอยู่ในระดับกลาง เกิดสระหลังระดับกลาง
6. ริมฝีปากห่อกลมเปิดกว้าง ลิ้นส่วนหลังจะอยู่ในระดับต่ำ เกิดสระหลังระดับต่ำ

2.7.3 การแบ่งประเภทของเสียงสระ

เสียงสระในภาษาไทยมี 24 เสียง ถ้าแบ่งตามช่วงเวลาจะ ได้เป็น 2 พวก คือ

1. สระเสียงสั้นใช้เวลาในการเปล่งเสียงน้อย มี 12 เสียง คือ อะ อี อี อู เอะ แอะ โอะ เอาะ เออะ เอียะ เอื้อะ อัวะ
2. สระเสียงยาวใช้เวลาในการเปล่งเสียงนานกว่า มี 12 เสียงเช่นกันคือ อา อี อือ อุ เอ แอ โอ ออ เออ เอีย เอื้อ อัว

สระทั้งสองพวกนี้มีที่ก่เกิดเสียงและใช้อวัยวะกล่อมเกล่าเสียงอย่างเดียวกัน แตกต่างที่ช่วงเวลาในการเปล่งเสียง เสียงสระยังแบ่งได้ตามส่วนประกอบของเสียงเป็น 2 ประเภทคือ

1. สระเดี่ยว เป็นสระที่เปล่งออกมาเป็นเสียงเดี่ยว มี 18 เสียง คือ อะ อา อี อี อือ อุ เอะ เอ แอะ แอ โอะ โอ เอาะ ออ เออะ เออ
2. สระประสม คือ สระที่ประกอบด้วยสระเดี่ยวสองเสียง มี 6 เสียง คือ เอียะ(อี+อะ) เอีย (อี+อา) เอื้อะ(อี+อะ) เอื้อ(อี+อา) อัวะ(อุ+อะ) อัว(อุ+อา)

ในการเปล่งเสียงสระเดี่ยวใช้ลิ้นและริมฝีปากกล่อมเกล่าเสียงในลักษณะใดลักษณะหนึ่งเพียงอย่างเดียว ส่วนการเปล่งสระประสมใช้ลิ้นและริมฝีปากกล่อมเกล่าเสียงร่วมกันสองลักษณะ สระเดี่ยวถึงแม้เกิดจากลมผ่านสายเสียงซึ่งมีการสั่นสะบัดแล้วผ่านเลยไปทางช่องปากโดยไม่ถูกกัก ณ อวัยวะส่วนใดส่วนหนึ่งก็ตามแต่ก็ยังถูกลิ้นและริมฝีปากกล่อมเกล่าในลักษณะต่างๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.8 เสียงพยัญชนะ

2.8.1 ลักษณะของเสียงพยัญชนะ

เสียงพยัญชนะเกิดจากกระแสลมจากปอดผ่านเส้นเสียงเช่นเดียวกับเสียงสระ แต่กระแสลมไม่อาจเคลื่อนไปโดยสะดวกอย่างเสียงสระ ถูกสกัดกั้นชั่วคราวไว้ทั้งหมดหรือเพียงบางส่วนแล้วปล่อยให้ออกไปทางปาก หรือทางจมูก ขณะที่กระแสลมเคลื่อนผ่านเส้นเสียงซึ่งถ้าอยู่ในสภาพคล้ายตัวตามปกติจะเกิดเสียงพยัญชนะธรรมดา หรือ พยัญชนะไม่ก้อง หรือ อโฆษะ(voiceless) หากเส้นเสียงเกร็งตัวตามปกติจะเกิดเสียง พยัญชนะก้อง หรือ โฆษะ (voiced) โดยทั่วไปเสียงพยัญชนะจะผ่านออกทางปาก แต่มีพยัญชนะบางเสียงผ่านออกทางช่องจมูก เสียงพยัญชนะจำพวกนี้เรียกว่า เสียงพยัญชนะนาสิก (nasal)

2.8.2 อวัยวะที่ทำให้เกิดความแตกต่างกันของเสียงพยัญชนะ

เสียงพยัญชนะเป็นเสียงที่เปล่งออกมาแล้วถูกอวัยวะหนึ่งในช่องปากกีดกันไว้ชั่วขณะ ณ ที่ใดที่หนึ่งก่อนปล่อยให้ออกไปทางช่องปากหรือช่องจมูก พยัญชนะจึงมีเสียงแตกต่างกันตามชนิดของอวัยวะที่เป็นฐานกักเสียงรวม 21 เสียงดังนี้

1. คอหอย (glottis) เส้นเสียง (vocal cords) ใช้เส้นเสียงทั้งสองกักเสียง ได้แก่ เสียง /อ/ /ฮ/
2. เพดานอ่อน (velar) ใช้โคนลิ้นกับเพดานอ่อนกักเสียง ได้แก่ เสียง /ก/ /ค/ /ง/ /ว/
3. ปุ่มเหงือก-เพดานแข็ง ใช้ปลายลิ้นกับปุ่มเหงือก-เพดานแข็งกักเสียง ได้แก่ /จ/ /ช/
4. เพดานแข็ง (palatal) ใช้ปลายลิ้นกับเพดานแข็งกักเสียง ได้แก่ เสียง /ย/
5. ปุ่มเหงือก (alveolar) ใช้ปลายลิ้นกับปุ่มเหงือกกักเสียง ได้แก่ เสียง /ซ/ /ด/ /น/ /ร/ /ล/
6. ฟัน-ปุ่มเหงือก (dental-alveolar) ใช้ปลายลิ้นกับฟัน-ปุ่มเหงือกกักเสียง ได้แก่ เสียง /ต/ /ท/
7. ริมฝีปากและฟัน (labio-dental) ใช้ริมฝีปากล่างกับฟันบนกักเสียง ได้แก่ /ฟ/
8. ริมฝีปาก (bilabial) ใช้ริมฝีปากบนกับริมฝีปากล่างกักเสียง ได้แก่ เสียง /บ/ /ป/ /พ/ /ม/ /ว/

เสียง /ว/ อาจจัดเป็นเสียงเกิดจากฐาน ริมฝีปาก หรือเพดานอ่อนก็ได้

เสียง /ย/ อาจเกิดจากฐานปุ่มเหงือกหรือเพดานแข็งก็ได้

เสียงพยัญชนะที่เกิดแต่ฐานกักเสียงเดียวกัน ถึงแม้จะมีฐานที่เกิดเสียงเดียวกันแต่ยังมีความแตกต่างกันในส่วนปลีกย่อยอื่นอีกสุดแต่วิธีการปล่อยลมจากฐานที่เกิดเสียงผิดแผกกันอย่างไร ความแตกต่างของลักษณะการปล่อยลมจากฐานที่เกิดเสียง ทำให้เสียงพยัญชนะแต่ละฐานที่เกิดเสียงต่างกันออกไปมีหลายลักษณะ ดังนี้

1. เสียงกัก หรือ เสียงระเบิด เกิดจากการที่ลมจากปอดถูกปิดกั้นสนิทไว้ชั่วขณะหนึ่ง ฐานที่เกิดเสียงค่อจากนั้นอวัยวะที่ปิดกั้นทางเดินของลมเคลื่อนที่ออกจากกันอย่างรวดเร็วปล่อยลมจากที่กักเสียงออกไปทันทีทางปาก ได้แก่ เสียง /ป/ /พ/ /ต/ /ท/ /ก/ /ค/
2. เสียงเสียดแทรก เกิดจากการที่ลมผ่านอวัยวะในการเปล่งซึ่งอยู่ใกล้กันมากๆ เหลือเพียงช่องแคบๆ ลมต้องแทรกตัวออกไป ได้แก่ เสียง /ฟ/ /ข/ /ฮ/
3. เสียงกักเสียดแทรก หรือ เสียงกึ่งเสียดแทรก เกิดจากการที่ลมผ่านอวัยวะในการเปล่งเสียงซึ่งปิดสนิทแล้วเปิดออกช้าๆ ทำให้เสียงที่เกิดขึ้นมีลักษณะเสียดแทรกผสม ได้แก่ /จ/ /ช/
4. เสียงกระทบ (tap) หรือ เสียงลิ้นกระดก (flap) เกิดจากการที่ลมผ่านอวัยวะในการเปล่งเสียงซึ่งกระทบกันอย่างรวดเร็วเพียงครั้งเดียว โดยที่ปลายลิ้นยกขึ้นไปกระทบปุ่มเหงือก แล้วสะบัดลงอย่างรวดเร็วเพียงครั้งเดียว ได้แก่ เสียง /ร/ ถ้าลมผ่านอวัยวะในการเปล่งเสียงซึ่งกระทบกันอย่างต่อเนื่องหลายครั้ง เรียกว่า เสียงร่ว ซึ่งไม่มีในภาษาไทย
5. เสียงข้างลิ้น เกิดจากการที่ลมผ่านออกมาทางข้างลิ้น โดยที่ปลายลิ้นยกไปแตะที่ปุ่มเหงือกปล่อยให้ลมออกมาทางสองข้างลิ้น ได้แก่ เสียง /ล/
6. เสียงนาสิก เกิดจากการที่ลมถูกผันให้ออกทางช่องจมูกพร้อมกับทางช่องปาก โดยลิ้นไก่และเพดานอ่อนลง ทำให้ช่องปากและช่องจมูกติดต่อดังกัน เช่น เสียง /ม/ /น/ /ง/
7. เสียงกึ่งสระ เกิดจากการที่ลมผ่านอวัยวะในการเปล่งเสียงซึ่งอยู่ใกล้กันแต่ไม่มากลมผ่านออกมาได้สะดวก เช่นเดียวกับการออกเสียงสระทั่วไป

เสียงกึ่งสระ ซึ่งเกิดจากฐานเพดานแข็ง ได้แก่ เสียง /ย/ ซึ่งมีลักษณะใกล้เคียงกับสระ อี อี

เสียงกึ่งสระ ซึ่งเกิดจากฐานเพดานอ่อน ได้แก่ เสียง /ว/ ซึ่งมีลักษณะใกล้เคียงกับสระ อุ อุ

นอกจากเสียงพยัญชนะจะแตกต่างเพราะความต่างกันของฐานที่เกิดเสียงและลักษณะการปล่อยลมจากฐานที่เกิดเสียงตามที่กล่าวมาแล้ว ความแตกต่างของเสียงพยัญชนะยังเกิดจากการคัดแปลงลมที่เส้นเสียงและปริมาณของลมดังนี้

1. เสียงก้อง (เสียงโหมะ) เกิดจากการที่ลมผ่านเส้นเสียงที่เกร็งตัวชิดกัน แต่ไม่ถึงกับติดกันสนิทพอให้ลมผ่านได้ โดยที่เส้นเสียงสั่นสะบัดได้แก่ เสียง /บ/ /ค/ /ร/ /ล/ /ว/ /ย/ /ม/ /น/ /ง/
2. เสียงไม่ก้อง (เสียงอโหมะ) เกิดจากการที่ลมผ่านเส้นเสียงที่หย่อนตัวห่างกัน ลมผ่านได้สะดวกโดยที่เส้นเสียงไม่สั่นสะบัด ได้แก่ เสียง /ป/ /พ/ /ต/ /ท/ /ก/ /ค/ /จ/ /ช/ /ฟ/ /ข/ /อ/ /ฮ/
3. เสียงมีกลุ่มลมหรือเสียงหนัก (เสียงชนิด) เกิดจากการที่เส้นเสียงไม่สั่นทันทีเมื่ออวัยวะที่ปิดกั้นทางเดินของลมเคลื่อนตัวออกจากกัน คือหยุดนิ่งอยู่ชั่วครู่ แล้วจึงเริ่มสั่นตัวพร้อมกับปล่อยลมออกมาปริมาณมาก ได้แก่ เสียง /พ/ /ท/ /ค/ /ช/
4. เสียงไม่มีกลุ่มลมหรือเสียงเบา (เสียงชนิด) เกิดจากการที่เส้นเสียงเริ่มสั่นตัวทันทีที่อวัยวะซึ่งปิดกั้นทางเดินของลมในช่องปากเคลื่อนตัวออกจากกัน และปล่อยลมออกมาปริมาณน้อยมาก ได้แก่ เสียง /ป/ /ค/ /ก/ /จ/

2.8.3 การแบ่งประเภทของเสียงพยัญชนะ

เสียงพูดในภาษาทั้งสามชนิด คือ เสียงสระ เสียงพยัญชนะ และเสียงวรรณยุกต์ จะต้องเปล่งออกมาเกือบพร้อมๆ กัน เสียงที่เกิดจากการเปล่งตัวกล่วครั้งหนึ่งๆ เรียกว่า พยางค์ ในการเปล่งเสียงพยางค์หนึ่งๆ เสียงพยัญชนะเป็นเสียงที่ได้ยินเด่นชัดนำเสียงสระและเสียงวรรณยุกต์ เสียงพยัญชนะที่อยู่ต้นพยางค์เรียกว่า เสียงพยัญชนะต้น เช่น ฟี มา ดู เสียง /ฟ/ /ม/ /ค/ เป็นเสียงพยัญชนะต้น โดยเสียงพยัญชนะต้นมีทั้งหมด 21 เสียงดังตารางที่ 2.1 บางพยางค์นอกจากมีเสียงพยัญชนะต้นแล้วยังปรากฏเสียงพยัญชนะตรงท้ายพยางค์ เสียงพยัญชนะที่อยู่ข้างท้ายพยางค์เรียกว่า เสียงพยัญชนะท้าย หรือ เสียงพยัญชนะสะกด เช่น ลม พัด แร่ง มี /ล/ /พ/ /ร/ เป็นเสียงพยัญชนะต้น เสียง /ม/ /ค/ /จ/ เป็นเสียงพยัญชนะท้าย หรือ เสียงพยัญชนะสะกด พยัญชนะท้ายพยางค์มีทั้งหมด 8 เสียง (9 ประเภท) ดังตารางที่ 2.2

ตารางที่ 2.1 เสียงพยัญชนะต้น

ที่	เสียงพยัญชนะต้น	รูปพยัญชนะ
1	/ก/	ก
2	/ค/	ข ก ฦ

3	/ง/	ง หง (หงาย)
4	/จ/	จ
5	/ช/	ฉ ช ฌ
6	/ซ/	ซ ศ ษ ส ทร (ทรง)
7	/ย/	ญ ย หญ (หญ้ำ) หย (หยิบ) หย (อย่า)
8	/ด/	ฎ ฑ (บัณฑิต) ด
9	/ต/	ฏ ต
10	/ท/	ฐ ฑ ฒ ถ ท ธ
11	/น/	ณ น
12	/บ/	บ
13	/ป/	ป
14	/พ/	ผ พ ภ
15	/ฟ/	ฝ ฟ
16	/ม/	ม หม (หมู)
17	/ร/	ร ฤ ทร (หฺร)
18	/ล/	ล ฬ หล (หลา)
19	/ว/	ว หว (หฺวี)
20	/ฮ/	ฮ ห
21	/อ/	อ

ตารางที่ 2.2 เสียงพยัญชนะท้ายพยางค์

ที่	เสียงพยัญชนะท้าย	รูปพยัญชนะท้าย
1	/ก/	แม็กก มี ก ข ค ฌ
2	/ง/	แม็กง มี ง
3	/ต/	แม็กค มี ค จ ช ฌ ฎ ฏ ฐ ฑ ฒ ติ ต ตร ตี ตุ ถ ท ทร ฐ ธิ ษ ศ ทร ฐ
4	/น/	แม็กน มี น ญ ณ ร ล ฬ
5	/ม/	แม็กม มี ม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับใช้เฉพาะในวงจำกัดเท่านั้น ไม่อนุญาตให้นำไปใช้เพื่อประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

6	/ป/	แม่กบ มี บ ป ปร พ ฟ ภ
7	/ย/	แม่เกย มี ย
8	/ว/	แม่เกว มี ว
9	ไม่มีเสียงพยัญชนะสะกด	แม่กา

2.8.4 ไตรยางค์

นอกจากนี้พยัญชนะยังสามารถแบ่งออกได้เป็น 3 พวก โดยพิจารณาถึงการใช้อุปวรรณยุกต์ และการผันเสียงวรรณยุกต์เป็นสำคัญ เรียกว่า ไตรยางค์ หรือ อักษรสามหมู่ ดังนี้

1. อักษรสูง มีพื้นเสียงสูงหรือเสียงวรรณยุกต์จัตวา ผันเสียงวรรณยุกต์ได้ไม่ครบ 5 เสียง มี 11 ตัว คือ ข ฃ ฉ ฐ ถ ผ ฝ ศ ช ส ห
2. อักษรกลาง มีพื้นเสียงกลางๆ หรือเสียงวรรณยุกต์สามัญและผันเสียงวรรณยุกต์ได้ครบ 5 เสียง มี 9 ตัว คือ ก จ ด ต ถ ฎ บ ป อ
3. อักษรต่ำ มีพื้นเสียงวรรณยุกต์สามัญเช่นเดียวกับอักษรกลาง แต่ผันเสียงวรรณยุกต์ได้ไม่ครบ 5 เสียง มี 24 ตัว คือ ค ฅ ฆ ง ช ฌ ญ ฑ ฒ ท ฐ น พ ฟ ภ ม ย ร ล พ ษ

2.9 เสียงวรรณยุกต์

ภาษาไทยเป็นภาษาเสียงวรรณยุกต์ (tonal language) คือ คำทุกคำต้องมีระดับเสียงวรรณยุกต์อย่างใดอย่างหนึ่ง คำที่ประกอบไปด้วยเสียงพยัญชนะและเสียงสระอย่างเดียวกัน เมื่อระดับเสียงวรรณยุกต์เปลี่ยนไปความหมายของคำนั้นๆ จะแตกต่างกันไปหรือไม่ก็กลายเป็นเสียงที่ไม่มีความหมายเลย เช่น ปา ป่า ป้า ป๊า ป๋า แต่ละคำมีความหมายต่างกัน ส่วนคำ กา ซึ่งมีความหมายหลายอย่าง ถ้าเปลี่ยนเสียงวรรณยุกต์เป็น ก่า กั๊ว กั๊ว กั๊ว แต่ละพยางค์เหล่านี้ไม่มีความหมาย

ภาษาที่มีลักษณะเป็นภาษาเสียงวรรณยุกต์เหมือนภาษาไทยมีไม่กี่ภาษา ที่เห็นได้ชัดเจนคือภาษาจีน นอกจากนี้ยังมีภาษาของชาวทวีปแอฟริกาบางเผ่าตอนใต้และตะวันตกของทวีป และภาษาอินเดียนแดงบางภาษาในทวีปอเมริกา

ภาษาอื่นๆ แม้ภาษาของชนชาติที่อยู่ใกล้เคียงกับไทย เช่น ภาษาเขมร มอญ พม่า เมเลย์ ไม่ใช่ภาษาเสียงวรรณยุกต์ ในภาษาที่ไม่ใช่ภาษาเสียงวรรณยุกต์คำแต่ละคำจะเปลี่ยนระดับไปอย่างไรก็ยังคงความหมายเดิม เช่น คำ คาร์ (car) ในภาษาอังกฤษ ถึงแม้จะออกเสียงเพี้ยนเป็น ข้า-ค่า(ข้า) คำ ขา ก็ยังคงความหมายว่า รถยนต์ เหมือนกันทุกคำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.9.1 ลักษณะของเสียงวรรณยุกต์

เสียงวรรณยุกต์เกิดจากกระแสลมจากปอดผ่านเส้นเสียงพร้อมๆ กับเสียงสระแต่ไม่มีการกล่อมเกลาเสียงอย่างเสียงสระ และไม่มีการกักกั้นและกล่อมเกลาของเสียงอย่างเสียงพยัญชนะ การที่เสียงวรรณยุกต์แตกต่างกัน เนื่องจากระดับความสั้นสะท้อนของเส้นเสียงในขณะเปล่งเสียงวรรณยุกต์แต่ละเสียงไม่เท่ากัน

2.9.2 ระดับเสียงวรรณยุกต์

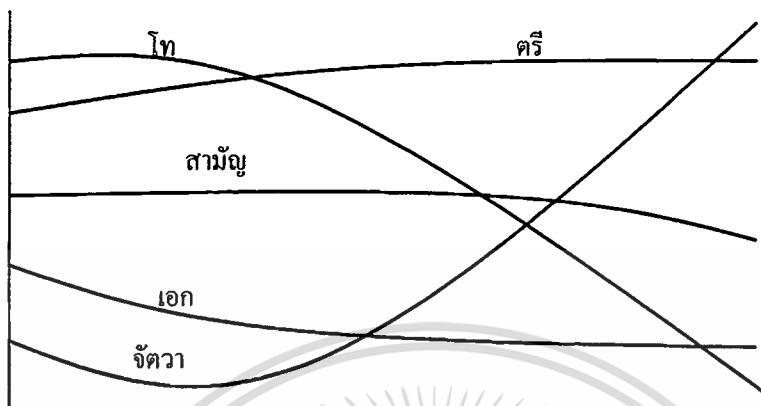
เสียงวรรณยุกต์ในภาษาไทยมี 5 ระดับดังนี้

1. เสียงสามัญ เริ่มต้นเสียงจากระดับกลาง เสียงคงระดับอยู่ชั่วระยะหนึ่งแล้วลดระดับลงเล็กน้อย เช่น คำ กา ตี ปู ถ้าเปล่งเสียงคำเหล่านี้แต่ละคำนานๆ จะเห็นว่าระดับเสียงคงที่ชั่วระยะหนึ่งแล้วลดลงเล็กน้อย
2. เสียงเอก เริ่มต้นระดับเดียวกับเสียงสามัญ แล้วลดระดับต่ำกว่าเสียงสามัญ เช่น เด็ก จะกวาด เศษ
3. เสียงโท เริ่มต้นระดับสูงกว่าเสียงสามัญแล้วตกลงต่ำกว่าเสียงเอก เช่น ที่ ว่า แยก ได้
4. เสียงตรี เริ่มต้นสูงกว่าเสียงสามัญเล็กน้อยแล้วขึ้นไปเรื่อยๆ จนเสมอดูจุดเริ่มต้นของเสียงโทแล้วตกลงเล็กน้อย เช่น รู้ แล้ว นะ จะ
5. เสียงจัตวา เริ่มต้นต่ำกว่าเสียงสามัญแล้วสูงขึ้น ไปเรื่อยๆ จนเสมอดูเสียงตรี เช่น หนู เห็น เขา ไหม

2.9.3 การเทียบเสียงวรรณยุกต์

เสียงวรรณยุกต์ 5 เสียงมีเครื่องหมายแทน 4 รูป คือ ไม้เอก ไม้โท ไม้ตรี และ ไม้จัตวา ตามกระบวนการเขียนเสียงกับรูปวรรณยุกต์อาจไม่ตรงกันเสมอไป เช่น คำที่ใช้รูปวรรณยุกต์เอกอาจมีเสียงโท คำที่ใช้รูปโทอาจมีเสียงตรี หรือคำที่ไม่มีรูปวรรณยุกต์กำกับอาจมีเสียงวรรณยุกต์ได้ครบเสียง คือ สามัญ เอก โท ตรี จัตวา ส่วนคำที่มีเสียงและรูปวรรณยุกต์ตรงกันเสมอเป็นเฉพาะในคำที่ใช้รูปวรรณยุกต์ตรีและจัตวา กล่าวคือ คำที่ใช้รูปวรรณยุกต์ตรีจะมีเสียงตรี คำที่ใช้รูปจัตวาจะมีเสียงจัตวาทุกคำ อักษรกลางเป็นพยัญชนะต้น มีเสียงและรูปวรรณยุกต์ตรงกันเป็นส่วนมาก

รูปที่ 2.8 ระดับเสียงวรรณยุกต์ (ระดับการเปลี่ยนแปลงของพิตช์)



ดังนั้นถ้าจะหาจำนวนเสียงทั้งหมดที่เป็นไปได้ในภาษาไทยในหนึ่งพยางค์จะมีความเป็นไปได้ในการผสมเสียงทั้งหมด $21 \times 24 \times 5 \times 9 = 22,680$ เสียงโดยประมาณ (ถ้าไม่คำนึงถึงการผสมอักขระบางแบบที่ไม่สามารถออกเสียงได้ และเสียงคำพูดที่ไม่มีเสียงออกมา เช่น การกระซิบหรือเสียงที่พูดโดยใช้อลมที่ไม่มีการสั่นของเส้นเสียง)

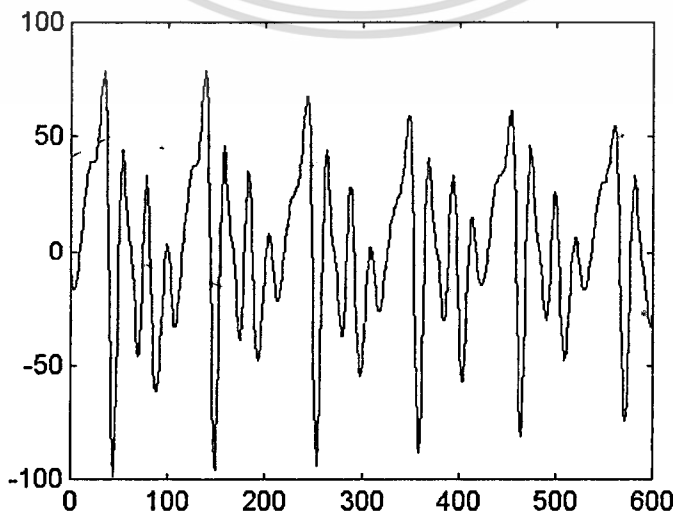
บทที่ 3

การคำนวณหาพิตช์

3.1 คุณลักษณะของสัญญาณเสียง

เสียงพูดเป็นสัญญาณที่มีคุณลักษณะที่ไม่คงที่ (Nonstationary Signal) มีการเปลี่ยนแปลงอย่างช้าๆ แต่อย่างไรก็ตามเมื่อพิจารณาเฉพาะส่วนย่อยของสัญญาณเสียงพูด นั่นคือในช่วงเวลาไม่เกิน 30 มิลลิวินาทีแล้ว สามารถอนุมานได้ว่า เสียงพูดเป็นสัญญาณที่มีคุณลักษณะคงที่ ดังนั้นสัญญาณเสียงพูดจึงถูกแบ่งเป็นส่วนย่อยๆ เรียกว่าเฟรม (frame) จากนั้นจึงทำการประมวลผลสัญญาณแต่ละเฟรม คุณลักษณะที่สำคัญอย่างหนึ่งของเสียงพูดคือพิตช์ (pitch) ซึ่งเป็นความถี่พื้นฐานที่เกิดจากการสั่นในหลอดเสียง ดังตัวอย่างในรูปที่ 3.1 พิตช์เป็นตัวกำหนดระดับความสูงต่ำของเสียง เนื่องจากว่าภาษาไทยเป็นภาษาที่มีเสียงวรรณยุกต์ (tonal language) คือมีระดับ โทนเสียง สามัญ เอก โท ตรี และจัตวา ซึ่งแต่ละโทนเสียงก็จะให้ความหมายที่แตกต่างกันออกไป ดังนั้นพิตช์จึงเป็นสิ่งที่สำคัญในการแยกแยะคำในภาษาไทย

รูปที่ 3.1 ตัวอย่างคาบเวลาพิตช์ของเสียงพูด



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.2 การหาจุดเริ่มและจุดท้ายของเสียง

ในระบบการรู้จำเสียงแบบคำโคความถูกต้องในการหาจุดเริ่มและจุดท้ายของเสียงมีความสำคัญอยู่ 2 ประการคือ ความคลาดเคลื่อนในการรู้จำ และปริมาณข้อมูลที่ใช้ในการคำนวณ ถ้ากำหนดให้สัญญาณเสียง $s(t)$ ถูกแปลงเป็นเชิงเลข (digitize) เป็นลำดับของข้อมูลเสียง $x(m)$ โดยที่

$$\{ x[m] \} = \{ x[1], x[2], x[3], \dots, x[m], \dots, x[K] \} \quad (3.1)$$

จะสามารถหาค่าพลังงานของสัญญาณได้คือ

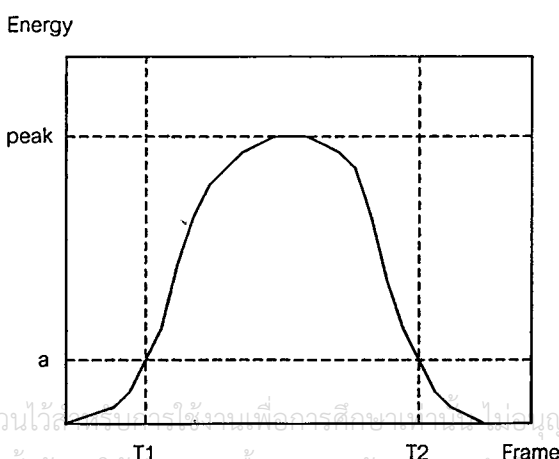
$$E_l(0) = \sum_{m=0}^{N-1} x^2(m), \quad l=0,1,\dots,L-1 \quad (3.2)$$

และเพื่อลดการคำนวณสามารถใช้ขนาดของสัญญาณแทนค่ากำลังสอง ดังสมการ

$$M_l(0) = \sum_{m=0}^{N-1} |x[m]| \quad (3.3)$$

โดยที่ $M_l(0)$, N , และ l คือ ค่าพลังงาน, ขนาดของเฟรม (frame size), และลำดับที่ของเฟรม หลังจากนั้นจะกำหนดระดับของพลังงานขีดเริ่มเปลี่ยน (threshold) a เพื่อใช้ในการหาจุดเริ่มและจุดท้ายของสัญญาณเสียง ดังแสดงในรูปที่ 3.2

รูปที่ 3.2 เฟรมเริ่มต้น($T1$) และเฟรมสุดท้าย($T2$) ของพลังงานของเสียงพูดที่ใช้ในการคำนวณ



โดยกำหนดให้ a มีค่า 10% ของค่าของเฟรมข้อมูลที่มีพลังงานมากที่สุด โดยค่า 10% นี้ได้มาจากการทดลองที่ให้ผลในการตัดข้อมูลที่ไมต้องการออกไปได้ดีที่สุด เขียนได้ดังสมการ

$$a = 0.1 M_{\max} \quad (3.4)$$

ซึ่งก็คือการตัดเฟรมที่มีพลังงานต่ำกว่า 10% ของเฟรมที่มีพลังงานมากที่สุด แล้วคำนวณค่า $T1$ และ $T2$ ใช้เป็นเฟรมเริ่มต้นและเฟรมสุดท้ายของข้อมูลที่ใช้ในการคำนวณ [9]

3.3 การคำนวณออโตคอร์รีเรนซ์

เสียงวรรณยุกต์แต่ละเสียงที่เกิดขึ้นเกิดจากการเปลี่ยนแปลงของพิตช์ ซึ่งการคำนวณหาพิตช์นั้นทำได้หลายวิธี วิธีหนึ่งซึ่งให้ผลดีคือการใช้ออโตคอร์รีเรนซ์หรืออัตสหสัมพันธ์ (Autocorrelation) ดังสมการ

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \quad (3.5)$$

โดยที่ $x(m)$ เป็นสัญญาณอินพุต ถ้า $x(m)$ เป็นสัญญาณที่มีคาบเวลาคงที่เท่ากับ P จะได้ว่า

$$\phi(k) = \phi(k+P) \quad (3.6)$$

นอกจากนี้ ออโตคอร์รีเรนซ์ยังมีคุณสมบัติที่สำคัญคือ

$$\phi(k) = \phi(-k) \quad (3.7)$$

และมีค่าสูงสุดเมื่อ $k = 0$ นั่นคือ

$$|\phi(k)| \leq |\phi(0)| \quad (3.8)$$

สำหรับสัญญาณที่มีคาบเวลา P ออโตคอร์รีเรนซ์จะมีค่าสูงสุดที่แซมเปิล $0, P, 2P, \dots$ ดังนั้นจึงสามารถคำนวณหาคาบเวลาของสัญญาณได้โดยการค้นหาค่า a แห่งแรกที่ให้ออโตคอร์รีเรนซ์มี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ค่าสูงสุด ในการประมวลสัญญาณเสียงพูดนั้นใช้ออโตคอร์รีเรชันสำหรับช่วงเวลาสั้นๆ โดยให้ค่านิยามไว้ว่า

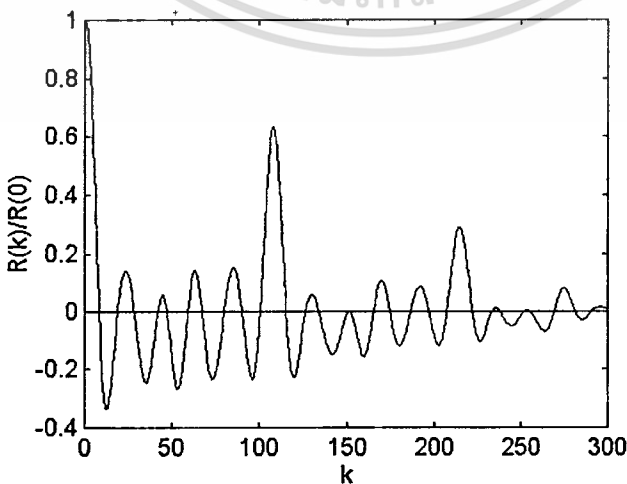
$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)][x(m+k)w(n-m-k)] \quad (3.9)$$

โดยที่ $w(n)$ เป็นฟังก์ชันหน้าต่าง ซึ่งอาจจะเป็นแบบสามเหลี่ยม สี่เหลี่ยม หรือแบบแฮมมิง ทำการคำนวณค่าของ $R_n(k)$ ในช่วงของ k จากน้อยที่สุดเท่าที่เป็นไปได้คือช่วงคาบเวลาของพิคซ์ของเสียงผู้หญิงซึ่งมีค่าเท่ากับ 3 มิลลิวินาที จนถึงค่า k สูงที่สุดเท่าที่เป็นไปได้คือช่วงคาบเวลาพิคซ์ของเสียงผู้ชายซึ่งมีค่าเท่ากับ 20 มิลลิวินาที ถ้าใช้อัตราการซีกตัวอย่าง (sampling rate) 10 kHz ก็จะใช้ k อยู่ในช่วง 30 ถึง 200 ถ้าเลือกให้ $w(n)$ มีความยาว N ตัวอย่าง โดยที่ N จะต้องมียค่ามากกว่า k มากๆ ($N \gg k$) ก็จะได้นิยามของออโตคอร์รีเรชันสำหรับช่วงเวลาสั้นๆ ดังนี้ [8]

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(m)w(n-m)][x(m+k)w(n-m-k)] \quad (3.10)$$

รูปที่ 3.3 แสดงให้เห็นถึงออโตคอร์รีเรชันแบบช่วงเวลาสั้นๆ ของสัญญาณเสียงพูด จะเห็นว่ามียค่าสูงสุดเมื่อ $k = 0$ และมีค่าสูงสุดอีกครั้งเมื่อ $k = 110$ ดังนั้นคาบเวลาของพิคซ์จึงเป็น 110 แซมเปิ้ล หรือเท่ากับ 11 มิลลิวินาทีเมื่ออัตราการซีกตัวอย่าง คือ 10 kHz นั่นคือความถี่ของพิคซ์เท่ากับ 91 Hz

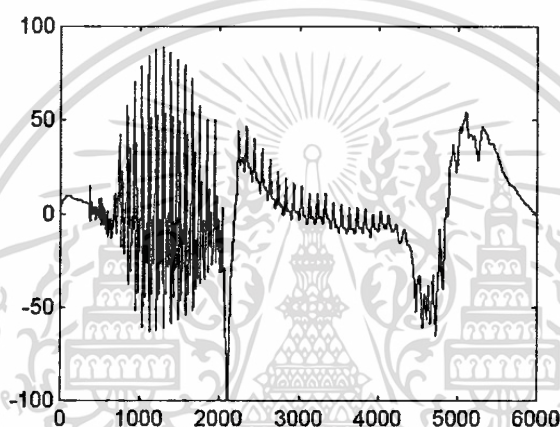
รูปที่ 3.3 ออโตคอร์รีเรชันของสัญญาณเสียงพูด



3.4 การปรับแต่งค่าพิตช์

เนื่องจากสัญญาณเสียงพูดแต่ละคำ พูดย่อมมีความแตกต่างกันของค่าพิตช์และความต่อเนื่องของพิตช์ โดยการเปลี่ยนแปลงของพิตช์จะขึ้นอยู่กับเสียงวรรณยุกต์ ส่วนความต่อเนื่องของพิตช์จะขึ้นอยู่กับองค์ประกอบของเสียงพยัญชนะเป็นหลัก ดังตัวอย่างในรูปที่ 5 ซึ่งจะเห็นได้ว่าข้อมูลบางช่วงจะไม่มีคาบพิตช์อยู่

รูปที่ 3.4 ตัวอย่างเสียงพูดของคำว่า 'จิม'



การคำนวณค่าออกโตคอร์รีเรนซ์จากทุกเฟรมของสัญญาณเสียงพูดเพื่อใช้ในการหาค่าพิตช์ จะทำให้ได้ข้อมูลพิตช์ที่มีความผิดพลาดปะปนอยู่อันเนื่องมาจากเสียงพยัญชนะดังกล่าว จึงทำให้ต้องมีการตัดข้อมูลพิตช์ที่ปะปนนี้ออก โดยใช้วิธีเทียบกับค่าเฉลี่ยพิตช์รวมของเสียงทั้งหมดที่ใช้เป็นต้นแบบในการฝึกจำ และค่ามัธยฐานของเสียงนั้นเฉลี่ยกัน ดังนี้

$$P_i = (\text{Mean_of_Pitch}_{\text{all words}} + \text{Median_of_Pitch}_i) / 2$$

$$k = 1$$

For j = 1 to length_of_Pitch

$$\text{If } |\text{Pitch}(j) - P_i| \leq 0.3 P_i$$

$$\text{NewPitch}(k) = \text{Pitch}(j)$$

$$k = k + 1$$

Endif

Next j

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

โดย P_i เป็นค่าเฉลี่ยของค่าเฉลี่ยรวมกับค่ามัธยฐานของเสียงที่ i ซึ่งพิตช์ที่มีขนาดมากกว่า 30% ของค่าเฉลี่ย P_i จะถูกตัดออก (ค่า 30% นี้ได้มาจากการทดลองที่ทำให้ค่าพิตช์มีความต่อเนื่องมากที่สุด) ทำให้ค่าพิตช์ใหม่ที่ได้นั้นมีความต่อเนื่องดีขึ้น การตัดค่าพิตช์ออกในลักษณะนี้จะให้ค่าพิตช์ที่ได้จากคำพูดที่มีเสียงพยัญชนะผสมอยู่มีความเหมาะสมที่จะใช้ในการฝึกจำ ซึ่งจะทำให้แบบจำลองของเสียงวรรณยุกต์ที่ได้มีความสามารถในการจดจำ เสียงวรรณยุกต์ที่มีเสียงพยัญชนะต่างๆ ผสมอยู่ได้



บทที่ 4

การประมาณเชิงเส้น

4.1 บทนำ

การประมาณเชิงเส้น (LPC : Linear Predictive Coding) เป็นการประมาณค่าของสัญญาณจากผลรวมเชิงเส้นของสัญญาณก่อนหน้า ซึ่งผลที่ได้จะเหมือนกันเป็นการประมาณค่าสเปกตรัมของสัญญาณนั่นเองดังในรูปที่ 4.1 ส่วนวิธีในการหาค่าสัมประสิทธิ์ LPC (α) และอัตราขยาย (G) ซึ่งต้องมีการเตรียมข้อมูลก่อนการวิเคราะห์เพื่อเตรียมสัญญาณเข้าเป็นดังรูป 4.2 [2], [7]

4.2 การพรีเอมฟาสิส

เนื่องจากสัญญาณเสียงพูดของมนุษย์ จะมีองค์ประกอบส่วนใหญ่อยู่บริเวณความถี่ต่ำ เมื่อเทียบกับแถบความถี่ที่ปฏิบัติงาน (bandwidth) ไม่เกิน 5 kHz ดังนั้น เพื่อให้อัตราส่วนสัญญาณเสียงต่อสัญญาณรบกวน (signal to noise ratio : SNR) มีค่าค่อนข้างคงที่ ตลอดช่วงความถี่ที่ปฏิบัติงานนี้ จึงต้องมีการพรีเอมฟาสิสหรือการเน้นล่วงหน้า (pre-emphasis) โดยเน้นความถี่สูงให้มีขนาดสูงขึ้น นั่นคือ การพรีเอมฟาสิสก็คือการกรองสัญญาณด้วยวงจรกรองความถี่สูงผ่าน (high pass filter) ซึ่งมักนิยมใช้วงจรกรองอันดับหนึ่ง มีฟังก์ชันถ่ายโอนเป็น

$$H(z) = 1 - a \cdot z^{-1} \quad , \quad 0.9 \leq a \leq 1.0$$

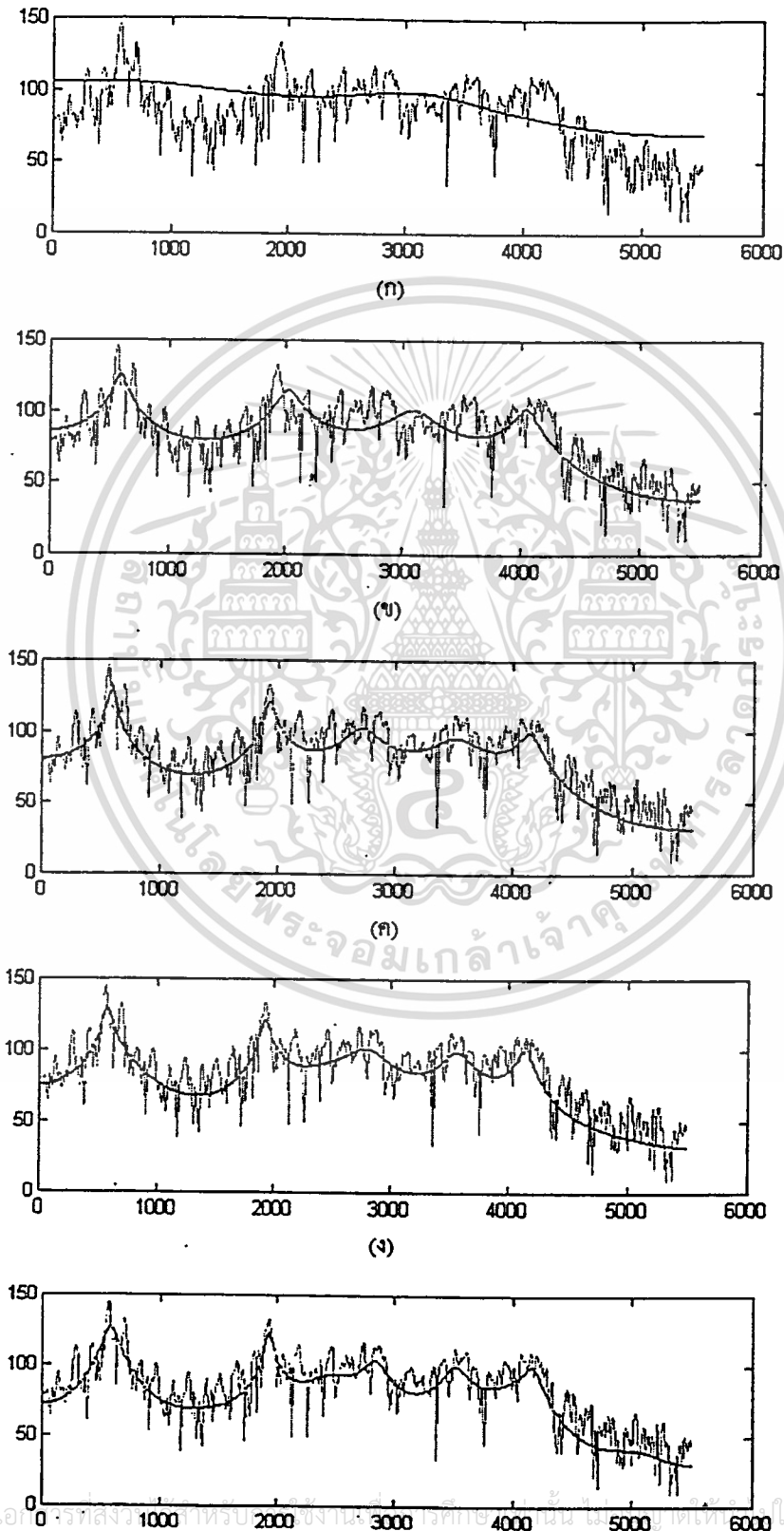
กรณีนี้เอาต์พุตการพรีเอมฟาสิส $s'(n)$ มีความสัมพันธ์กับอินพุต $s(n)$ ดังสมการผลต่างสืบเนื่อง

$$s'(n) = s(n) - a \cdot s(n-1)$$

ยิ่งค่า a ใกล้เคียง 1 เท่าใด ความถี่สูงก็จะถูกขยายมากขึ้นเท่านั้น ค่า a ที่นิยมสำหรับใช้ในการหาพารามิเตอร์ของ LPC คือ 0.95 ([2], หน้า 113) เมื่อนำฟังก์ชันถ่ายโอนมาพล็อตกราฟของขนาดเทียบกับความถี่จะได้ดังรูปที่ 4.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 4.1 แสดงสเปกตรัมของสัญญาณเสียงเทียบกับสเปกตรัมที่ได้จากการประมาณเชิงเส้นที่มี
อันดับ (LPC order) เท่ากับ p ต่างๆ (ก) $p=4$ (ข) $p=8$ (ค) $p=12$ (ง) $p=16$ (จ) $p=20$



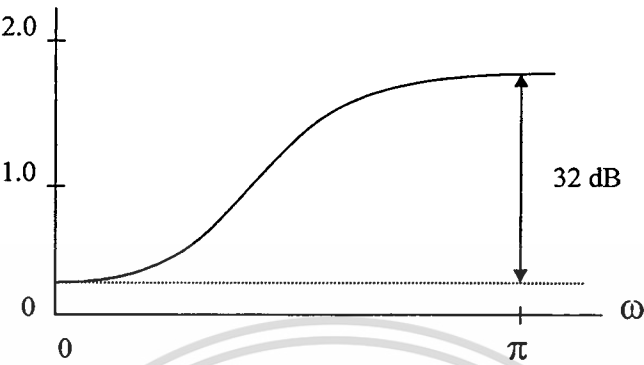
เอกสารนี้เป็นเอกสารสงวนลิขสิทธิ์ของมหาวิทยาลัยราชภัฏวชิรเวศน์บุรีรัมย์ ไม่สามารถนำเอกสารนี้ไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 4.2 แสดงขั้นตอนการเตรียมสัญญาณในการวิเคราะห์



เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

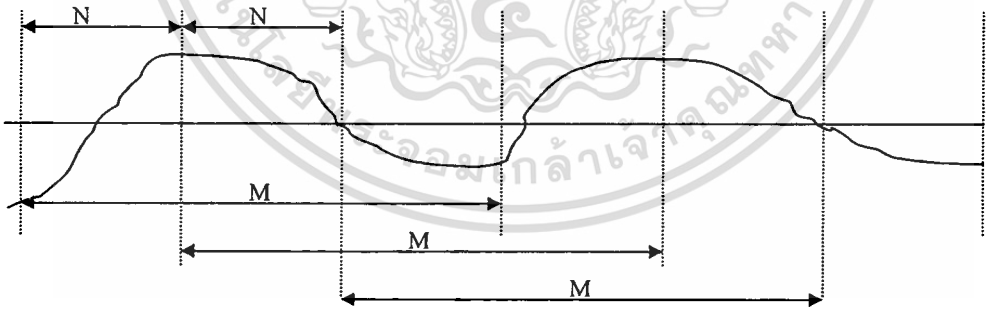
รูปที่ 4.3 แสดงขนาดสเปกตรัมของฟังก์ชันถ่ายโอนของการพรีเอมฟาสซิส



4.3 การแบ่งช่วงสัญญาณ

สัญญาณที่ผ่านการพรีเอมฟาสซิสแล้ว $s'(n)$ จะถูกตัดแบ่งออกเป็นช่วงๆ หรือ เฟรม ช่วงละ N ตัวอย่างสัญญาณ การวิเคราะห์จะวิเคราะห์ทีละช่วงของแต่ละ N ตัวอย่างสัญญาณ ดังรูปที่ 4.4

รูปที่ 4.4 แสดงการแบ่งช่วงของสัญญาณที่ใช้ในการวิเคราะห์



โดยช่วงในการวิเคราะห์แต่ละช่วงจะถูกเลื่อนไปเป็นระยะ M ช่วงสัญญาณ จะเห็นได้ว่า ถ้าค่า M โตกว่าค่า N ในการเลื่อนของช่วงในการวิเคราะห์จะทำให้บางสัญญาณไม่ถูกใช้ในการวิเคราะห์ ก็จะเป็นการสูญเสียส่วนหนึ่งทำให้ผลที่ได้ไม่ถูกต้องเท่าที่ควร ถ้าค่า M เล็กกว่า N ก็จะทำให้ตัวอย่างสัญญาณทุกตัวถูกนำมาวิเคราะห์ ยิ่งค่า M เล็กเท่าใด ความแม่นยำในการวิเคราะห์ก็จะยิ่งสูงขึ้นเท่านั้น แต่ก็จะทำให้การคำนวณช้าลง

หลักการกำหนดขนาดของเฟรม

1. เฟรมจะต้องสั้นพอที่คุณสมบัติของเสียงที่สนใจจะวิเคราะห์นั้น ยังไม่เปลี่ยนแปลงในเฟรม
2. เฟรมจะต้องยาวพอที่ จำนวนตัวอย่างสัญญาณในเฟรม สามารถนำมาคำนวณหาคุณสมบัติที่ต้องการได้
3. เฟรมที่ติดกัน ไม่ควรจะสั้นจนกระโดดข้ามข้อมูลบางส่วนไป แต่ควรเลื่อนเฟรมให้น้อยกว่าขนาดของเฟรม

เนื่องจาก ใช้ความถี่ในการสุ่มสัญญาณ 11.025 กิโลเฮิร์ตซ์ ในการวิเคราะห์นี้ เลือกใช้ค่า $N = 300$ และค่า $M = 100$ นั่นคือ ช่วงในการวิเคราะห์ คือ 27.21 มิลลิวินาที และระยะในการเลื่อน(เฟรม) ประมาณ 9 มิลลิวินาที

4.4 การวินโดว์

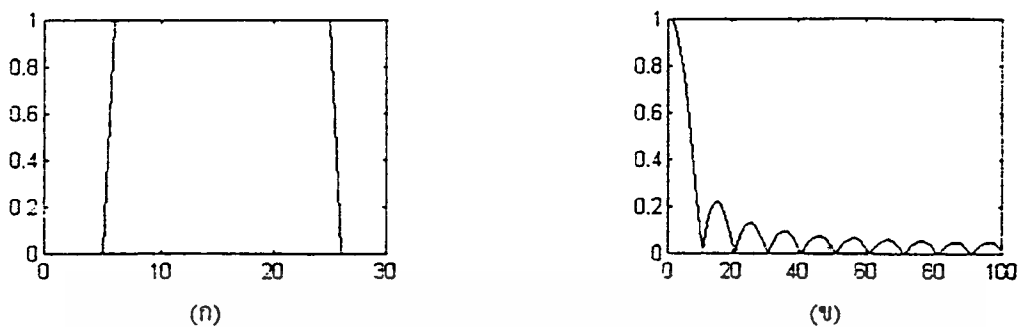
ถ้าพิจารณาช่วงสัญญาณ N ตัวอย่างสัญญาณของช่วงใด ๆ ที่ตัดมาวิเคราะห์ จะเห็นว่าที่ขอบของเฟรมที่ตัดมานี้มีความไม่ต่อเนื่องของสัญญาณ ถ้ามองในโดเมนความถี่สูง ก็จะมีค่าสูงเกิดขึ้น ดังนั้นเพื่อที่จะลดองค์ประกอบทางความถี่ที่สูงเหล่านี้ จะคูณด้วยฟังก์ชันวินโดว์เพื่อลดความไม่ต่อเนื่องของสัญญาณที่ขอบ และไม่ทำให้สเปกตรัมของสัญญาณในช่วงความถี่ต่ำเปลี่ยนไปมากนัก ในที่นี่จะใช้ฟังก์ชันวินโดว์แฮมมิง (Hamming window function) ซึ่งนิยามโดยสมการดังนี้คือ

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad \text{เมื่อ } n = 0, 1, 2, \dots, N-1 \quad (4.1)$$

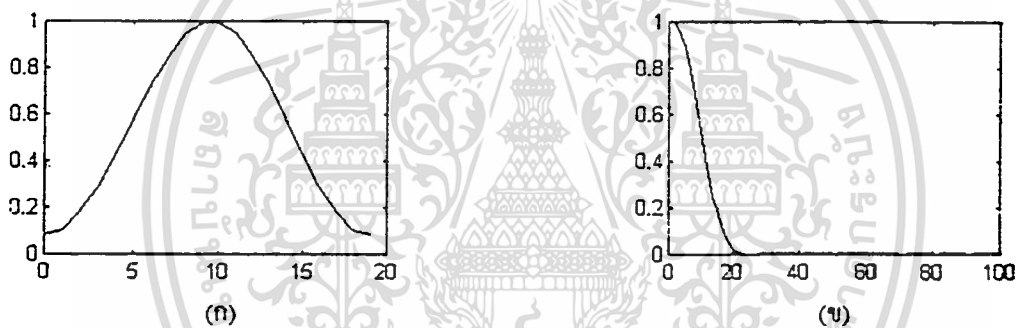
รูปที่ 4.5 และ 4.6 แสดงองค์ประกอบทางเวลา และทางความถี่ของฟังก์ชันวินโดว์แบบสี่เหลี่ยม และแฮมมิง ตามลำดับ จะเห็นว่าสเปกตรัมของวินโดว์แฮมมิงมีความพลัว (ripple) น้อยกว่าของวินโดว์สี่เหลี่ยม ดังนั้น จากรูปที่ 4.1 จะได้ว่า

$$x'(n) = x(n) \cdot w(n) \quad (4.2)$$

รูปที่ 4.5 แสดงวินโดว์สี่เหลี่ยม (ก) ในโดเมนเวลา (ข) ในโดเมนความถี่



รูปที่ 4.6 แสดงวินโดว์แบบแฮมมิง (ก) ในโดเมนเวลา (ข) ในโดเมนความถี่



4.5 การหาค่าสัมประสิทธิ์ด้วยวิธีออกโตคอร์รีเรชัน

จากหลักการพื้นฐานของการประมาณเชิงเส้นคือ การประมาณค่าสัญญาณจากผลรวมเชิงเส้นของสัญญาณก่อนหน้านี้ สมมติว่าสัญญาณเดิมเป็น $s(n)$ การประมาณค่าสัญญาณเป็น $s'(n)$ ดังนั้นสามารถอธิบายการประมาณเชิงเส้นได้ด้วยสมการต่อไปนี้

$$s'(n) = \sum_{k=1}^p \alpha_k s(n-k) \quad (4.3)$$

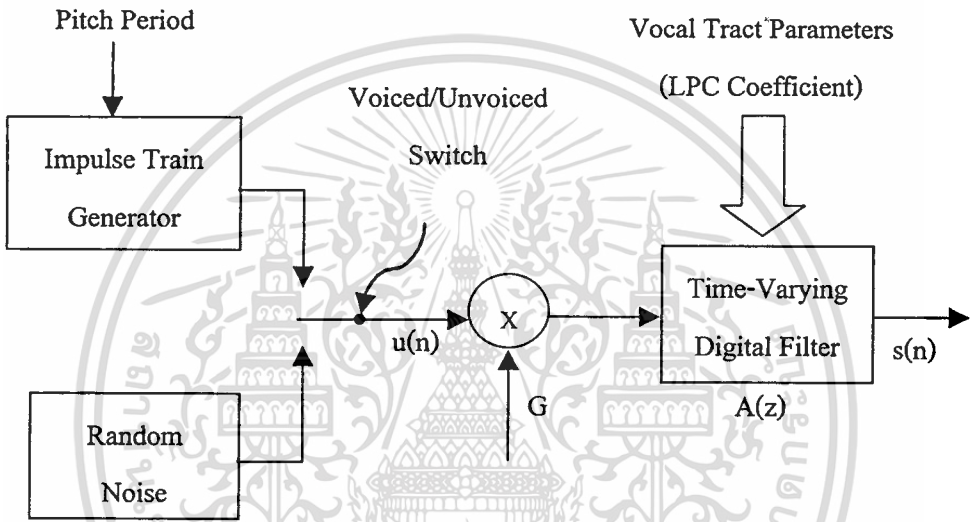
เมื่อ α_k เป็นค่าคงที่ เรียกวิธีการนี้ว่าการประมาณเชิงเส้นอันดับ p โดยมีเงื่อนไขว่า ค่า α_k ที่ใช้ในการประมาณจะต้องทำให้ ผลรวมของกำลังสองของความคลาดเคลื่อน $\{s(n)-s'(n)\}^2$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

มีค่าน้อยที่สุด นั่นคือ $\sum e^2(n) = \sum \{s(n) - s'(n)\}^2$ มีค่าต่ำที่สุด ซึ่งจะใช้การประมาณเชิงเส้นวิธี
ออโตคอร์รีเรชัน (Autocorrelation Method) หรือ วิธีอัตราสัมพันธ์

จากหลักการพื้นฐานของการประมาณเชิงเส้นและแบบจำลองระบบสร้างสัญญาณเสียง
สามารถเขียน บล็อกไดอะแกรมการทำการประมาณเชิงเส้น มาสร้างสัญญาณเสียงพูดได้ดังรูปที่ 4.7

รูปที่ 4.7 บล็อกไดอะแกรมแสดงแบบจำลองการสร้างสัญญาณเสียงพูดอย่างง่าย



จากรูป 4.7 สามารถเขียนสมการได้เป็น

$$s(n) = G \cdot u(n) + \sum_{k=1}^p a_k s(n-k) \tag{4.4}$$

การประมาณเชิงเส้นโดยใช้สัมประสิทธิ์ $\{\alpha_k\}$ คือ

$$s'(n) = \sum_{k=1}^p \alpha_k s(n-k) \tag{4.5}$$

ดังนั้น ความคลาดเคลื่อน คือ

$$e(n) = s(n) - s'(n)$$

$$= s(n) - \sum_{k=1}^p \alpha_k s(n-k) \quad (4.6)$$

ฟังก์ชันถ่ายโอนระหว่าง $e(n)$ และ $s(n)$ คือ

$$A(z) = E(z) / S(z)$$

$$= 1 - \sum_{k=1}^p \alpha_k z^{-k} \quad (4.7)$$

จากสมการ (4.4) - (4.6) จะเห็นได้ว่า ถ้า $\{\alpha_k\} = \{a_k\}$ แล้ว

$$e(n) = G * u(n) \quad (4.8)$$

ดังนั้น ค่าผลรวมของกำลังสองของความคลาดเคลื่อน

$$E_n = \sum_m e_n^2(m)$$

$$E_n = \sum_m [s(m) - s'(m)]^2 \quad (4.9)$$

โดยที่ n คือ ช่วงที่ n ของสัญญาณที่ใช้คำนวณ เพราะฉะนั้นเพื่อให้ได้ค่า E_n ต่ำที่สุดจะต้องมีเงื่อนไขว่า

$$\frac{\partial E_n}{\partial \alpha_i} = 0 \quad \text{เมื่อ } i=1,2,3,\dots,p$$

จากสมการ (4.9)

$$\frac{\partial E_n}{\partial \alpha_i} = -2s_n(m-i) \sum_m [s_n(m) - \sum_{k=1}^p \alpha_k s_n(m-k)] \quad \text{เมื่อ } i=1,2,3,\dots,p$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$=-2\left[\sum_m s(m)s(m-i) - \sum_{k=1}^p \sum_m \alpha_k s(m-k)s(m-i)\right]$$

$$\frac{\partial E_n}{\partial \alpha_i} = 0 \quad \text{ก็ต่อเมื่อ}$$

$$\sum_{k=1}^p \alpha_k \sum_m s_n(m-k)s_n(m-i) = \sum_m s_n(m)s_n(m-i) \quad \text{เมื่อ } i=1,2,3,\dots,p \quad (4.10)$$

ถ้ากำหนดให้ $\phi_n(i,j) = \sum_m s_n(m-k)s_n(m-i)$ เพราะฉะนั้น

$$\sum_{k=1}^p \alpha_k \phi_n(i,j) = \phi_n(i,0) \quad (4.11)$$

โดยสมการ (4.9) - (4.10) จะได้ว่า

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^p \alpha_k \sum_m s_n(m)s_n(m-k)$$

และจาก $\phi_n(i,j) = \sum_m s_n(m-k)s_n(m-i)$

$$E_n = \phi_n(0,0) - \sum_{k=1}^p \alpha_k \phi_n(0,k) \quad (4.12)$$

สมมุติว่าใน 1 เฟรมของสัญญาณ ที่ตัดมาจำนวนมี N ตัวอย่าง คือ $s_n(0), s_n(1), s_n(2), \dots, s_n(N-1)$ ในที่นี้ให้ $s_n(m) = 0$ เมื่อ $m < 0$ หรือ $m > N-1$ เพราะฉะนั้น

$$\begin{aligned} \phi_n(i,j) &= \sum_m s_n(m-k)s_n(m-i) \\ &= \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \quad 0 \leq k \leq p, 1 \leq i \leq p \end{aligned}$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ให้
$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k) \quad \text{เมื่อ } k=0,1,2,\dots,p \quad (4.13)$$

ดังนั้น จากสมการ (4.2) และ (4.13) จะได้ว่า

$$R_n(k) = \sum_{m=0}^{N-1-k} x'(m)x'(m+k) \quad (4.14)$$

จากสมการที่ (4.11) จะได้ว่า

$$\sum_{k=1}^p \alpha_k R_n(i-k) = R_n(i) \quad \text{เมื่อ } i=1,2,3,\dots,p \quad (4.15)$$

จากสมการ (4.15) เขียนให้อยู่ในรูปเมทริกซ์ได้เป็น

$$\begin{bmatrix} R_n(0) & R_n(1) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & \cdots & R_n(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_n(p-1) & R_n(p-2) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(p) \end{bmatrix} \quad (4.16)$$

หรือ
$$R_n \cdot \alpha = r_n \quad (4.17)$$

เมื่อ $R_n = \begin{bmatrix} R_n(0) & R_n(1) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & \cdots & R_n(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_n(p-1) & R_n(p-2) & \cdots & R_n(0) \end{bmatrix}$, $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix}$ และ $r_n = \begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(p) \end{bmatrix}$

4.6 การหาสัมประสิทธิ์ α และอัตราขยาย G

เมื่อได้ค่า $R_n(0), R_n(1), R_n(2), \dots, R_n(p)$ แล้ว ก็สามารถหาค่า α ได้จากสมการ (4.17)

นั่นคือ

$$\alpha = R_n^{-1} \cdot r_n \quad (4.18)$$

และจากสมการ (4.8) จะได้ว่า

$$e(n) = G * u(n)$$

$$\therefore E_n = \sum_{m=0}^{N-1} e^2(m) = G \sum_{m=0}^{N-1} u^2(m)$$

จาก สมการ (4.12) จะได้ว่า

$$\begin{aligned} E_n &= \phi_n(0,0) - \sum_{k=1}^p \alpha_k \phi_n(0,k) \\ &= R_n(0) - \sum_{k=1}^p \alpha_k R_n(k) \end{aligned} \quad (4.19)$$

และจากสมการ (4.19) สามารถหาค่า G โดยตรงจาก

$$G^2 = \frac{R_n(0) - \sum_{k=1}^p \alpha_k R_n(k)}{\sum_{m=0}^{N-1} u^2(m)} \quad (4.20)$$

4.7 การเปลี่ยนพารามิเตอร์ LPC เป็น สัมประสิทธิ์เซปสตรัม (Cepstrum)

ในการรู้จำเสียงพูดนั้น สัมประสิทธิ์เซปสตรัมนี้เป็นพารามิเตอร์ที่มีลักษณะน่าเชื่อถือได้ดีกว่าสัมประสิทธิ์ LPC ทั้งยังมีความสัมพันธ์ใกล้ชิดกับการรู้จำเสียง ตามความรู้สึกของมนุษย์โดยแท้จริง สัมประสิทธิ์เซปสตรัมสามารถหาได้โดยตรงจากสัมประสิทธิ์ LPC ดังนี้

$$C_0 = \ln G \quad (4.21)$$

$$C_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k}, 1 \leq m \leq p \quad (4.22)$$

$$C_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k}, m \geq p \quad (4.23)$$

$$Q \approx \frac{3}{2} p \quad (4.24)$$

4.8 การถ่วงน้ำหนักค่าพารามิเตอร์ (Parameter Weighting)

เนื่องจากสัมประสิทธิ์เซปสตรีมที่ได้นั้นช่วงลำดับต้นๆและลำดับท้ายๆของเฟรมที่นำมาวิเคราะห์จะเกิดความคลาดเคลื่อนมากกว่าบริเวณส่วนอื่น เพราะฉะนั้นจึงทำการถ่วงน้ำหนักเพื่อลดค่าความคลาดเคลื่อนดังกล่าวนี้ ด้วยฟังก์ชันเวทดัง ดังนี้คือ

$$W_m = \left[1 + \frac{Q}{2} \sin\left(\frac{\pi m}{Q}\right)\right], 1 \leq m \leq Q \quad (4.25)$$

จะได้

$$C'_m = C_m * W_m \quad (4.26)$$

บทที่ 5

การควอนไทซ์ (Quantization)

5.1 บทนำ

การจดจำเสียงพูด เป็นลักษณะหนึ่งของการจดจำรูปแบบ (Pattern Recognition) ก็จะเป็นการเปรียบเทียบระหว่าง แบบทดสอบ (Test Pattern) กับ แบบอ้างอิง (Reference Pattern) ซึ่งเป็นรูปแบบที่ทราบและเก็บไว้ล่วงหน้า

ขั้นตอนในการจดจำแบ่งเป็น 2 ขั้นตอน ดังนี้

ก. ขั้นเรียนรู้ (Learning) จะเป็นการสร้างกลุ่มของแบบอ้างอิง ในการจดจำเสียงพูด ในขั้นนี้จะทำการวิเคราะห์เสียงพูดก่อน แล้วเก็บลักษณะของเสียงในรูปของพารามิเตอร์ และใช้การควอนไทซ์แบบเวกเตอร์ ในการสร้างแบบอ้างอิง เพื่อใช้เปรียบเทียบในขั้นตอนต่อไป

ข. ขั้นจดจำ (Recognition) จะเป็นการทดสอบการจดจำ ระหว่างแบบอ้างอิง กับแบบทดสอบ โดยจะทำการเปรียบเทียบพารามิเตอร์ของแบบทดสอบ กับแบบอ้างอิงทั้งหมด แบบอ้างอิงที่เลือก คือ แบบอ้างอิงที่มีพารามิเตอร์ใกล้เคียงกับแบบทดสอบที่สุด

5.2 การควอนไทซ์แบบเวกเตอร์ (Vector Quantization)

ในทฤษฎีของการสื่อสารนั้น แหล่งเข้ารหัสจะใช้เทคนิคการเปลี่ยนค่าสัญญาณ เอาต์พุตไปเป็นรหัสแบบไบนารี ซึ่งเป้าหมายของเครื่องเข้ารหัสคือ การเป็นไปได้อย่างมีประสิทธิภาพของข้อมูลน้อยสุด ด้วยอัตราความเร็ว(bit rate) ที่กำหนด

การควอนไทซ์แบบเวกเตอร์เป็นเทคนิคเข้ารหัสเทคนิคหนึ่งที่มีคุณภาพ ซึ่งจะทำการเข้ารหัสเวกเตอร์พารามิเตอร์ของสัญญาณที่แทนสเปกตรัมของเสียง ที่เป็นเวกเตอร์อินพุต ให้เป็นเลขดัชนี (index) โดยนำไปเปรียบเทียบกับตารางรหัส (code book) ที่ได้สร้างเก็บไว้ก่อนแล้ว และตารางรหัส ที่ถูกเลือกนั้นจะเป็นตัวที่ใกล้เคียงที่สุดกับเวกเตอร์อินพุต ผลที่ได้จากการควอนไทซ์แบบเวกเตอร์ที่แสดงออกมานั้นขึ้นกับชุดของ ตารางรหัสที่สร้างเก็บไว้ก่อนหน้านั้นเอง

ถ้าเปรียบเทียบอัตราข้อมูลของการแทนเวกเตอร์สำหรับเสียงที่ยังไม่ได้เข้ารหัส จะเห็นว่า การวิเคราะห์เสียงนั้นต้องการการลดของอัตราข้อมูล พิจารณาดังนี้ เสียงพูดที่ชักตัวอย่าง (sampling) ด้วยความถี่ 10 kHz และมีขนาด 16 บิต จะต้องการอัตราข้อมูล 160,000 bps ในการจัดเก็บ ซึ่งเป็นรูปแบบที่ยังไม่ได้มีการบีบอัดข้อมูล สำหรับการวิเคราะห์เสียงพูดพิจารณาเวกเตอร์ที่มี 10 มิติ ใช้ 100 สเปกตรัมต่อวินาที ถ้าใช้ 16 บิตแทนแต่ละเวกเตอร์ อัตราที่ใช้ในการเก็บ คือ $100 \times 10 \times 16$ bps หรือ 16,000 bps สามารถลดไปได้ถึง 10 เท่าของสัญญาณที่ยังไม่ได้บีบอัด ยิ่งไปกว่านั้นวิธีการการควอนไทซ์แบบเวกเตอร์ยังสามารถลดอัตราข้อมูลลงไปได้อีก โดยการสร้าง ตารางรหัส เพื่อใช้ในการชี้ว่าเวกเตอร์ตารางรหัสไหนที่เหมือนกับเวกเตอร์เสียงอินพุตที่เข้ามามากที่สุด สมมุติ ตารางรหัสแทนด้วยตัวเลข 10 บิต และมีอัตราข้อมูล 100 สเปกตรัมต่อวินาที จะได้ว่า อัตราข้อมูล 1,000 bps ซึ่งลดอัตราข้อมูลลงไปได้ถึง 16 เท่า

เวกเตอร์สเปกตรัมที่จะนำไปทำการควอนไทซ์ เป็นเวกเตอร์ที่ได้จากการประมาณเชิงเส้น และผ่านการปรับปรุงให้ดีขึ้นโดยการทำ เซปสตรัมและการเวทค่าพารามิเตอร์ ซึ่งจะแสดงโดย $V_l, l = 1, 2, \dots, L$ ซึ่งแต่ละเวกเตอร์เป็นเวกเตอร์ที่มี p มิติ (p คือ order หลังจากผ่านการเวทค่า เรียบร้อยแล้ว) เช่นในรูปตัวอย่างที่ 5.1 และ 5.2 คือ $p = 2$ (order = 2) ของ LPC Model

ข้อดีของวิธีการ การควอนไทซ์แบบเวกเตอร์ คือ

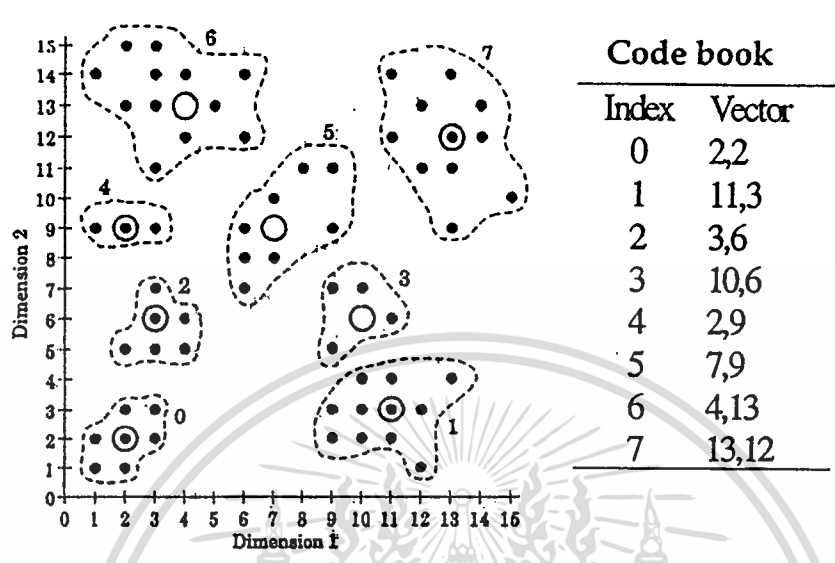
- ลดความจุในการเก็บข้อมูลสเปกตรัม ที่ได้จากการวิเคราะห์หลัง
- ลดการคำนวณในการเปรียบเทียบความเหมือนของเวกเตอร์สเปกตรัม ที่ได้จากการวิเคราะห์หลัง ซึ่งเป็นการคำนวณหลักที่ใช้ในการจดจำเสียงพูด

รูปที่ 5.1 แสดงข้อมูลที่จะทำการควอนไทซ์

Training data

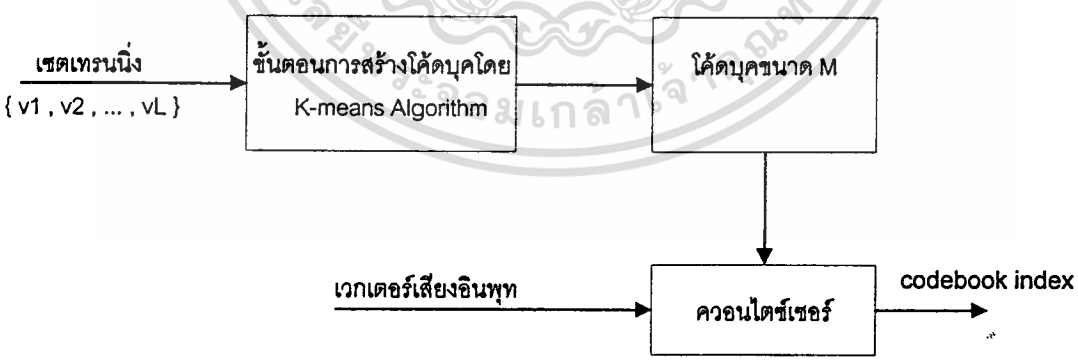
1,1	3,5	6,7	9,9	12,1
1,2	3,6	6,8	9,11	12,3
1,9	3,7	6,9	10,2	12,11
1,14	3,9	6,12	10,3	12,13
2,1	3,11	6,14	10,4	13,4
2,2	3,13	7,8	10,7	13,9
2,3	3,14	7,10	11,2	13,11
2,5	3,15	8,11	11,3	13,12
2,9	4,5	9,2	11,4	13,14
2,13	4,6	9,3	11,6	14,12
2,15	4,12	9,5	11,12	14,13
3,2	4,14	9,7	11,14	15,10
3,3	5,13			

รูปที่ 5.2 แสดงการจับกลุ่มและดัชนีของตารางรหัส (code book) ที่ได้



5.3 ส่วนประกอบของการควอนไทซ์แบบเวกเตอร์

รูปที่ 5.3 แสดงบล็อกไดอะแกรมของการควอนไทซ์แบบเวกเตอร์



5.3.1 การทำงานของการควอนไทซ์แบบเวกเตอร์

1. เซตของเวกเตอร์สเปกตรัมที่ได้จากการวิเคราะห์ $v_1, v_2, v_3, \dots, v_L$ ซึ่งเป็นเซตขนาดใหญ่ จะถูกนำมาใช้ในการสร้างเซตของ ตารางรหัส เพื่อใช้เป็นตัวแทนเวกเตอร์สเปกตรัมของเสียงที่สังเกตได้ ขนาดตารางรหัสของการควอนไทซ์แบบเวกเตอร์ คือ $M = 2^B$ เวกเตอร์ (เรียกว่า B-bit ตารางรหัส) ซึ่งต้องการให้ $L \gg M$ เพื่อที่จะหาเซตของ M ตารางรหัส ที่ดีที่สุด ในทางปฏิบัตินั้นเพื่อให้ตารางรหัสของการควอนไทซ์แบบเวกเตอร์ สามารถทำงานได้ดีนั้น ขนาด L อย่างน้อยที่สุดที่สามารถใช้ได้ คือ 10M

2. การวัดความเหมือน หรือ ระยะทางระหว่าง เวกเตอร์สเปกตรัมที่ได้จากการวิเคราะห์ 2 เวกเตอร์ ดังนั้นสามารถจัดกลุ่มเซตของเวกเตอร์ เทอร์มิ่ง เพื่อสร้าง ตารางรหัส แสดงระยะทางสเปกตรัม $d(v_i, v_j)$ ระหว่าง 2 เวกเตอร์ v_i และ v_j คือ d_{ij}

3. ฟังก์ชันย่อยในการคำนวณหาจุดศูนย์กลาง เพื่อแบ่งแยกเซตของ เทอร์มิ่ง จำนวน L เวกเตอร์ ไปเป็น M กลุ่ม และ M ตารางรหัส ที่ได้คือจุดศูนย์กลางของแต่ละกลุ่ม

4. ฟังก์ชันย่อยในการแบ่งแยกสำหรับเวกเตอร์สเปกตรัมที่ได้จากการวิเคราะห์ ซึ่งจะเลือก ตารางรหัส ที่ใกล้เคียงที่สุดให้กับเวกเตอร์ อินพุต และผลลัพธ์ที่ได้คือ วรรณิตารายรหัสซึ่งเป็นตัวแทนของสเปกตรัม มักจะใช้กฎ Nearest Neighbour ในการพิจารณา ส่วนนี้เป็นหัวใจหลักของ ตัวควอนไทซ์ ซึ่ง อินพุต คือ เวกเตอร์สเปกตรัมของเสียง และ เอาต์พุต คือ วรรณิตารายรหัส ที่มีความเหมือนกับ เวกเตอร์เสียง อินพุตมากที่สุด

5.3.2 เซตเทอร์มิ่ง (Training) ของ การควอนไทซ์แบบเวกเตอร์

เซตเทอร์มิ่งของการควอนไทซ์แบบเวกเตอร์ซึ่งเป็นเสียงพูดที่จะนำมาใช้ในการสร้างตารางรหัส ซึ่งเซตเทอร์มิ่งจะต้องมีขนาดใหญ่มาก ๆ ถึงจะสามารถสร้างตารางรหัสที่ดีได้ เซตเทอร์มิ่งจะแปรตามสถานะต่าง ๆ ดังนี้

- ผู้พูด รวมทั้งช่วงของอายุ เพศ ความเร็ว ช้าในการพูด ระดับเสียง ฯลฯ
- สถานะแวดล้อมขณะพูด เช่น ในห้องเงียบ หรือในรถ
- ระบบการแปลงการส่ง รวมทั้ง ไมโครโฟน หูโทรศัพท์
- หน่วยของเสียงรวมถึง คำศัพท์ต่าง ๆ และข้อความที่สนทนา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

5.4 หลักการของการควอนไทซ์แบบเวกเตอร์

การควอนไทซ์แบบเวกเตอร์ เป็นวิธีการควอนไทซ์คลื่นเสียง ในการควอนไทซ์แบบเวกเตอร์ มีวิธีการควอนไทซ์ดังนี้ คือ จะมีเวกเตอร์รหัส หรือ ตารางรหัสเก็บไว้ อินพุตที่เข้ามาจะถูกทำการเปรียบเทียบกับตารางรหัสที่มีอยู่ โดยจะพิจารณาว่าอินพุตที่เข้ามานั้นห่างจากตารางรหัสใดน้อยที่สุด อินพุตดังกล่าวจะถูกแทนด้วยเวกเตอร์รหัสนั้น

ตารางรหัสที่ใช้จะเหมาะสมเพียงใด ขึ้นกับความคลาดเคลื่อนรวมทั้งหมดของต้นแบบที่ใช้หา เมื่อต้นแบบที่พิจารณามีจำนวนจำกัด ตารางรหัสที่มีความคลาดเคลื่อนน้อยที่สุด จะได้จากการทำซ้ำจนกว่าจะได้ค่าที่เหมาะสม โดยเหตุนี้จึงขึ้นกับ ค่าแรมคอมที่สุ่มขึ้นมาจำนวนเท่ากับขนาดของตารางรหัสเพื่อใช้เป็นค่าเริ่มต้น ในวิธีการนี้ข้อมูลต้นแบบที่ต้องการหาถูกแยกไปอยู่ในแต่ละกลุ่มของค่าเริ่มต้น จะทำการหาค่าเฉลี่ยของแต่ละกลุ่มเพื่อเป็นค่ากลางของกลุ่ม ทำวิธีการดังกล่าวซ้ำจนกว่าความคลาดเคลื่อนรวมจะต่ำกว่าค่า ๆ หนึ่ง หรือการลดลงของความคลาดเคลื่อนรวมน้อยกว่าค่า ๆ หนึ่ง โดยค่ากลางดังกล่าวของกลุ่มจะถูกเก็บเป็นเวกเตอร์รหัส แต่แน่นอนว่าค่าความคลาดเคลื่อนรวมจะลดลงทุกครั้งที่มีการคำนวณซ้ำใหม่ จึงขึ้นกับค่าที่กำหนดความต้องการให้ความคลาดเคลื่อนรวมน้อยเพียงใด

5.5 การคำนวณการควอนไทซ์แบบเวกเตอร์

ถ้าสมมุติ $x = [x_1, x_2, \dots, x_p]$ มี p มิติ จะทำการหาควอนไทซ์ของเวกเตอร์ x กับตารางรหัส สามารถเขียนว่า y เป็นควอนไทซ์ของค่า x

$$y = q(x) \quad (5.1)$$

โดย $q(.)$ เป็นโอเปอเรเตอร์ของควอนไทซ์ y ถูกเรียกว่าเอาต์พุตเวกเตอร์ของค่า x โดย y เป็นค่าใดค่าหนึ่งใน $Y = \{y_i, 1 \leq i \leq M\}$ โดย $y_i = [y_{i1} y_{i2} \dots y_{ip}]$ Y เป็นเซตของตารางรหัส M เป็นขนาดของตารางรหัส และ $\{y_i\}$ เป็นเซตของเวกเตอร์รหัส y_i อาจเรียกว่าเป็นรหัสอ้างอิง และ M อาจเรียกว่าจำนวนระดับขั้นจะทำการแบ่งเวกเตอร์ x ไปใน M เซล $\{C_i, 1 \leq i \leq M\}$ เมื่อ x อยู่ในเซลล์ C_i

$$q(x) = Y_i \text{ ถ้า } x \in C_i$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถ้า x ถูกควอนไทซ์ได้ค่า y ค่าความคลาดเคลื่อนจากการควอนไทซ์สามารถแสดงได้โดย ระยะห่างของทั้งสอง $d(x,y)$ โดยความคลาดเคลื่อนรวมคือ

$$D = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{n=1}^M d[x(n), y(n)] \quad (5.2)$$

การคำนวณหาตารางรหัสมีหลายวิธี แต่ที่จะนำมาใช้คือ Lloyd Algorithm (K-means Algorithm)

5.6 การวัดความคลาดเคลื่อน

การวัดความคลาดเคลื่อนเป็นส่วนจำเป็นและเป็นประโยชน์ต่อการออกแบบ ตารางรหัส ซึ่งวัดโดยการวัดระยะทางซึ่ง สมการพีชคณิตที่ใช้ในการหาระยะทางในปัจจุบันมีวิธีการหลายแบบ แต่ที่จะนำมาใช้ในปริยญาณิพนธ์นี้ คือ square error distortion.

square error distortion : ถ้าสัญญาณที่เข้ามามี P มิติ สามารถหาระยะห่างระหว่าง สัญญาณเข้ากับเวกเตอร์รหัส

$$d(v_1, v_2) = \|v_1 - v_2\|^2 = \sum_{i=0}^{k-1} (x_i - y_i)^2 \quad (5.3)$$

ซึ่งเป็นวิธีการคำนวณที่ง่าย และรวดเร็วในการวัดความคลาดเคลื่อน จึงนิยมใช้วิธีนี้ แต่ก็ยังไม่เหมาะสมเพียงพอ ควรใช้ Itakura distance มากกว่า แต่เป็นวิธีที่มีความซับซ้อนมากและยุ่งยากในการคำนวณ

5.7 การจัดกลุ่มเพื่อสร้างแบบอ้างอิงโดยใช้ K-means Algorithm

การสร้างกลุ่มของแบบอ้างอิงเป็นขั้นตอนในส่วนของการเรียนรู้ โดยเริ่มต้นจากการนำข้อมูลสัญญาณเสียงที่ได้จากส่วนของการวิเคราะห์ มาทำการสร้างเป็นกลุ่มของแบบอ้างอิง ในระบบการจดจำเสียงพูดแบบต่างบุคคล จะใช้แบบอ้างอิงของคำหนึ่ง ๆ จากผู้พูดจำนวนมาก เพื่อที่จะได้

ครอบคลุมถึงความแปรปรวนต่าง ๆ ที่จะเกิดขึ้นระหว่างผู้พูดแต่ละคน แต่อย่างไรก็ตามผลที่เกิดขึ้นตามมา เมื่อจำเป็นจะต้องมีข้อมูลจำนวนมากในกลุ่มของแบบอ้างอิง คือ

1. เวลาที่ใช้ในการตอบสนอง ในการจดจำเสียงพูด ขั้นตอนการเปรียบเทียบแบบทดสอบกับแบบอ้างอิงทั้งหมดที่มีเป็นขั้นตอนที่สำคัญในการจดจำ เมื่อแบบอ้างอิงมีจำนวนมาก การเปรียบเทียบจึงจำเป็นที่จะต้องอาศัยเวลาเพิ่มขึ้นด้วย

2. เนื้อหาในหน่วยความจำสำรองที่ใช้ในการเก็บแบบอ้างอิง เมื่อมีแบบอ้างอิงจำนวนมากที่จำเป็นต้องใช้ในการเปรียบเทียบ เนื้อหาในหน่วยจำสำรองก็จำเป็นต้องมีเพิ่มขึ้นด้วย

3. ความถูกต้องในการจดจำ เนื่องจากจำนวนแบบอ้างอิงในแต่ละคำไม่ได้เป็นส่วนโดยตรงกับความถูกต้องในการจดจำ กล่าวคือเมื่อเพิ่มจำนวนแบบอ้างอิงในแต่ละคำไปจนถึงระดับหนึ่ง ความถูกต้องในการจดจำจะเริ่มคงที่ และในบางครั้งอาจจะมีค่าลดลงด้วย

จากเหตุผลดังกล่าว จึงทำให้มีความพยายามที่จะจัดกลุ่มของแบบอ้างอิงในแต่ละคำใหม่ เพื่อให้ได้แบบอ้างอิงจำนวนที่พอเหมาะ และสามารถใช้เป็นตัวแทนของแบบอ้างอิงที่มีอยู่ทั้งหมดได้ อัลกอริทึมที่ใช้ในการแบ่งกลุ่มมีอยู่เป็นจำนวนมาก สำหรับอัลกอริทึมที่จะกล่าวถึงคือ K-means Algorithm ตามขั้นตอนโดยสรุป ดังนี้

5.7.1 การหาจุดศูนย์กลางของกลุ่ม (Center Cluster)

ในขั้นตอนนี้จะทำการกำหนดจำนวนกลุ่มของแบบอ้างอิงให้มีค่าเท่ากับ M และเริ่มต้นจากการสุ่ม (random) ค่ามาจำนวนเท่ากับขนาดของตารางรหัส เพื่อเป็นค่าจุดศูนย์กลางของแต่ละกลุ่มก่อน จากนั้นจะแบ่งเวกเตอร์ต่าง ๆ เข้าไปตามกลุ่มต่าง ๆ โดยพิจารณาจากระยะทาง หรือความต่างกันน้อยที่สุดกับกลุ่มไหนก็จะให้อยู่กลุ่มนั้น จากนั้นก็จะทำการเฉลี่ยค่าของเวกเตอร์ที่อยู่ในกลุ่มเพื่อให้เป็นจุดศูนย์กลางใหม่

5.7.2 การจัดกลุ่มใหม่ (Reclassification)

แบบอ้างอิงที่เหลืออยู่ในแต่ละคำ จะถูกจัดให้เข้าไปในแต่ละกลุ่ม โดยมีเงื่อนไขคือ ระยะทางของแบบอ้างอิงนี้กับแบบอ้างอิงในกลุ่มจะมีค่าเฉลี่ยน้อยที่สุด หลังจากที่มีแบบอ้างอิงทุกแบบถูกจัดเข้ากลุ่มเรียบร้อยแล้ว จะทำการหาจุดศูนย์กลางของแต่ละกลุ่มใหม่ ขั้นตอนนี้จะทำให้จำแนกว่าจุดศูนย์กลางของแต่ละกลุ่มจะไม่มีเปลี่ยนแปลง อย่างไรก็ตามจากขั้นตอนดังกล่าวจะเห็นว่า การกำหนดค่าจำนวนกลุ่มที่เหมาะสมจะทำได้ลำบาก ดังนั้นจึงได้มีการปรับปรุงเทคนิคนี้เป็นการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กำหนดค่าระยะทางสูงสุดที่จะทำให้เกิดกลุ่มใหม่ขึ้น คือเมื่อแบบอ้างอิงใดที่ถูกเลือกให้เข้ากลุ่มด้วยระยะทางที่น้อยที่สุดแล้ว แต่ระยะทางนี้ยังคงมีค่ามากกว่าระดับที่กำหนด แบบอ้างอิงนี้จะถูกนำไปสร้างเป็นจุดศูนย์กลางของกลุ่มใหม่ทันที

5.8 การคำนวณจุดศูนย์กลางสำหรับการออกแบบตารางรหัสการควอนไทซ์แบบเวกเตอร์

การสร้าง ตารางรหัส โดยใช้ K-means สามารถสรุปสั้น ๆ ได้ดังนี้

1. ความเพี้ยนน้อยสุด สำหรับแต่ละเวกเตอร์ อินพุต x_i เมื่อเปรียบเทียบกับเวกเตอร์ y_i ซึ่งเป็น ตารางรหัส C ที่เข้าไป กลุ่มของเวกเตอร์ เทรนนิ่ง จะเป็นไปตาม Code words

2. การคำนวณหาจุดศูนย์กลาง แต่ละกลุ่มของเวกเตอร์ จะคำนวณหาจุดศูนย์กลางใหม่ ซึ่งมีค่าความผิดเพี้ยนเฉลี่ยต่ำ สำหรับสมาชิกกลุ่มที่ถูกแทนโดยเวกเตอร์จุดศูนย์กลาง

ทำ 2 ขั้นตอนซ้ำจนกว่าจะเกิดการลู่เข้า (convergent) ขั้นแรกนั้นเป็นการทำโดยตรง คือ นำเข้าไปเพื่อหาความผิดเพี้ยน ส่วนขั้นที่ 2 นั้นเป็นการหาจุดศูนย์กลาง ซึ่งเป็นขั้นตอนที่สำคัญ

พิจารณา $\{x_i\}_{i=1}^L$ และค่าความเพี้ยน $d(x,y)$ โดยทั่วไป สมมติว่าเวกเตอร์เหล่านี้ถูกกำหนดไว้ในกลุ่มป้ายเดียวกัน (หรือ Code word) จุดศูนย์กลางของ $\{x_i\}_{i=1}^L$ ถูกจำกัดความเป็นเหมือนเวกเตอร์ $\bar{Y}$ ซึ่งเป็นค่าความเพี้ยนเฉลี่ย น้อยที่สุด

$$\bar{Y} = \Delta \arg \min \frac{1}{L} \sum_{i=1}^L d(x_i, y) \quad (5.4)$$

เมื่อ x_i และ y เป็นเวกเตอร์ โดยที่ $x_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{ik})$ และ $y = (y_1, y_2, \dots, y_k)$ วัดในพื้นที่ k มิติ กับระยะทางตามแบบยูคลิด จุดศูนย์กลางคือค่าเฉลี่ยของเซตของ เทรนนิ่ง

$$\bar{Y} = \frac{1}{L} \sum_{i=1}^L x_i \quad (5.5)$$

$\bar{Y}$ เป็นจุดศูนย์กลางซึ่งเป็นเวกเตอร์ที่อยู่ตรงกลางของ $\{x_i\}_{i=1}^L$ ซึ่งแต่ละมิติจะไม่ขึ้นแก่กัน หมายความว่า แต่ละ y_k เป็นค่ากลางของ $\{x_i\}_{i=1}^L$

5.9 กฎของการตัดสินใจ (Decision Rules)

ผลที่ได้จากการเปรียบเทียบระยะทางระหว่างแบบทดสอบกับแบบอ้างอิงแต่ละแบบ ระยะที่ได้ทั้งหมดจะนำมาผ่านขั้นตอนการตัดสินใจ เพื่อหาผลลัพธ์ของการจดจำ กฎการตัดสินใจ (Decision Rules) ที่ใช้ในการจดจำเสียงพูด มีดังนี้ คือ

5.9.1 กฎ Nearest-Neighbor (NN)

ผลลัพธ์ของการจดจำจากกฎการตัดสินใจนี้ ได้แก่ แบบอ้างอิงที่มีระยะทางจากแบบทดสอบน้อยที่สุด กฎนี้เหมาะสำหรับระบบการจดจำเสียงพูดแบบบุคคลเดียว

5.9.2 กฎ K Nearest-Neighbor (KNN)

เป็นกฎที่ดัดแปลงมาจากกฎ NN อีกทีหนึ่ง เพื่อให้เหมาะกับการจดจำเสียงพูดแบบต่างบุคคล ซึ่งมีแบบอ้างอิงในแต่ละคำมากกว่า 1 แบบ กฎนี้แบ่งออกเป็น 2 แบบ คือ

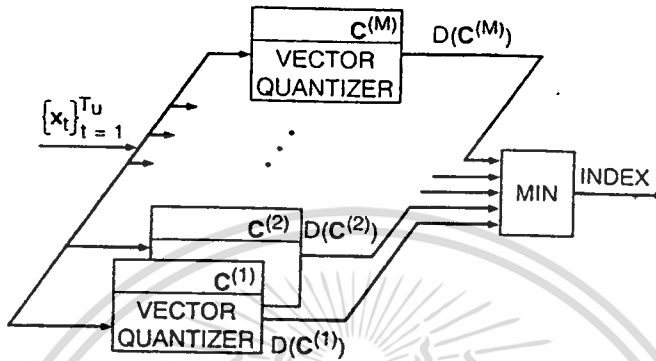
1. แบบที่ 1 จะทำการหาแบบอ้างอิงจำนวน K แบบ ที่มีระยะทางห่างจากแบบทดสอบน้อยที่สุด ผลลัพธ์ของการจดจำจะได้จาก แบบอ้างอิงที่มีจำนวนมากที่สุด ในจำนวน K แบบ ที่ได้รับการเลือก

2. แบบที่ 2 จะทำการหาแบบอ้างอิงจำนวน K แบบ ในแต่ละคำ โดยที่ทั้ง K แบบมีระยะทางห่างจากแบบทดสอบน้อยที่สุด ผลลัพธ์ของการเรียนรู้จะได้จาก แบบอ้างอิงที่ค่าเฉลี่ยของระยะทางในจำนวน K แบบ ที่ได้รับการเลือกมาน้อยที่สุด วิธีนี้จะให้ผลลัพธ์ที่ถูกต้องมากขึ้น โดยกำหนดค่า K เป็น 2 หรือ 3 เมื่อเทียบกับกฎ NN อย่างไรก็ตามวิธีนี้จะทำให้เวลาที่ใช้ในการจดจำเพิ่มขึ้นด้วย

ในระบบเสียงพูดทั่วไป จะมีการกำหนดค่าระยะทางมากที่สุดที่ยอมรับได้ (Reject Value) เพื่อเพิ่มความถูกต้องของการจดจำในกรณีที่ผลลัพธ์ที่ได้จากการตัดสินใจมีค่าใกล้เคียงกันมาก วิธีการกำหนดค่านี้ต้องใช้ความระมัดระวังเป็นอย่างมาก เนื่องจากถ้ากำหนดค่านี้สูงเกินไป หรือ ไม่ได้กำหนด โอกาสที่จะทำให้ผลลัพธ์ของการจดจำผิดพลาดไปจะมีมาก แต่ถ้ากำหนดค่านี้ต่ำเกินไป จะทำให้ผลลัพธ์ของการจดจำสำหรับแบบอ้างอิงที่ถูกต้องเกิดขึ้นได้ยาก ดังนั้นโดยทั่วไปค่าระยะทางนี้มักจะได้มาจากการทดลอง เพื่อหาค่าที่เหมาะสมสำหรับกลุ่มของแบบอ้างอิงหนึ่ง ๆ เท่านั้น

5.10 การควอนไทซ์แบบเวกเตอร์สำหรับระบบการจดจำเสียงพูด

รูปที่ 5.4 แสดงเวกเตอร์ตัวควอนไทซ์ (vector quantizer) ที่ใช้ในระบบการจดจำเสียงพูด



การที่จะสร้างตารางรหัสจำเป็นต้องใช้เวกเตอร์ทดสอบเป็นจำนวนมาก ซึ่งจะจัดเวกเตอร์ออกเป็นกลุ่มๆ และแต่ละกลุ่ม i แทนจุดศูนย์กลางด้วย y_i ซึ่งหาจุดศูนย์กลางได้จากการเฉลี่ยค่าเวกเตอร์ทั้งหมดในกลุ่ม ซึ่ง y_i จะเป็นเวกเตอร์รหัสที่เก็บไว้ หรือเป็นตารางรหัสเก็บไว้สำหรับการควอนไทซ์เวกเตอร์เสียงต่อไป เวกเตอร์เสียงอินพุตของตัวควอนไทซ์ $= x_i$ จะถูกเปรียบเทียบกับ

y_i
 $d(x_i, y_i)$ ค่า y_m ที่เลือกคือ ระยะทาง d น้อยที่สุด ซึ่ง y_m จะแทน x_i ที่เข้ามา จากนั้นจะส่งครรชนี m ไปยังเครื่องรับ จากเมตริกซ์ระยะทาง ได้ว่า ถ้า x ใกล้ y_1 มากกว่า y_2 สรุปได้ว่า x เสียงเหมือน y_1 มากกว่า y_2

การควอนไทซ์แบบเวกเตอร์ที่ใช้เพื่อการออกแบบการจดจำเสียงพูดนั้น มีจำนวนตัวควอนไทซ์ M ตัว นั่นหมายถึง มี M ระดับของเสียงเพื่อการจดจำ แต่ละระดับเสียงสามารถพิจารณาจากเซตของข้อมูล เทรนนิ่ง $\{x_i^{(u)}\}$ เมื่อ $i=1, 2, \dots, M$ ซึ่ง M เป็นครรชนีระดับ แต่ละเซตของเทรนนิ่ง ในแต่ละระดับจะเก็บเสียงที่อยู่ในระดับเดียวกัน เมื่อมีเสียงที่ไม่รู้ $\{x_i\}^{T_u}$ เข้ามา จะเป็นอินพุตเข้าไปยังทุก ๆ ตัวควอนไทซ์ ค่าครรชนีระดับ (index) ที่เลือก จะเป็นระดับที่มีความผิดพลาดเฉลี่ยน้อยที่สุด $D(c^{(u)})$ เมื่อ $i=1, 2, \dots, M$ ซึ่งความผิดพลาดเฉลี่ยนี้ดูได้จากการวัดระยะทางกลุ่มที่ i ในจำนวน j กลุ่ม ซึ่งจะทำให้การวนซ้ำ k ครั้งแสดงโดย $\{w_j^{(u)}\}$ ซึ่ง $i=1, 2, \dots, j$ และ $k=0, 1, \dots, k_{\max}$ ($k_{\max}$ คือจำนวนการนับวนซ้ำที่มากที่สุด) $j=1$ (กลุ่มเดียว) จนถึง $j_{\max}$ (จำนวนกลุ่มที่มากที่สุด) โดยมีรายละเอียดดังนี้

1. เริ่มโดย $j = 1, k = 1, i = 1$ คำนวณเมตริกซ์ระยะทาง $D = \hat{\delta}(x_i, x_m), 1 \leq i, m \leq N$
2. คำนวณจุดศูนย์กลางกลุ่ม คือ $c(\Omega)$ ของ Ω จุดศูนย์กลางกลุ่มของข้อมูล เทรนนิ่ง อาจถูกคำนวณได้โดย แบบจุดศูนย์กลางที่สุด หรือ ศูนย์กลางเฉลี่ย
3. ให้ $(w_1^{(0)}) = \Omega, c[(w_1^{(0)})] = c(\Omega)$
4. แบ่งแยกแต่ละ pattern x_i ของ Ω ซึ่งเป็นของกลุ่ม 1 กลุ่ม $(\Omega_j^{(k)})$, $i = 1, 2, \dots, j$ โดยเลือกกลุ่มซึ่งมีระยะทางน้อยสุดจาก pattern x_i ไปยังจุดศูนย์กลางของกลุ่ม $c[(w_j^{(k)})]$ เช่น $x_i \in (w_j^{(k)})$ iff $\hat{\delta}(x_i, c[(w_j^{(k)})]) \leq \hat{\delta}(x_i, c[(w_{\hat{i}}^{(k)})])$ สำหรับทุก $\hat{i} \neq i$
5. สำหรับแต่ละกลุ่มที่ได้ $(w_j^{(k)})$, $i = 1, 2, \dots, j$ คำนวณหาจุดศูนย์กลางกลุ่ม $c[(w_j^{(k)})]$ อีกครั้ง โดยใช้ จุดศูนย์กลางที่สุด หรือ จุดศูนย์กลางเฉลี่ย
6. ตรวจสอบการบรรจบกัน โดยดูจากมีบางกลุ่มถูกเปลี่ยนไปจากการทำซ้ำครั้งก่อนหรือไม่ (เช่น $(w_j^{(k)}) = (w_j^{(k-1)})$, $i = 1, 2, \dots, j$) ถ้ามีบางกลุ่มเปลี่ยนไป ตัวนับ k จะเพิ่มค่าขึ้น และตรวจสอบกับค่าที่มากที่สุดที่กำหนด ถ้าค่าการวนซ้ำยังไม่ถึงค่ามากที่สุด K-means จะทำการวนซ้ำอีก (ขั้นตอนที่ 4 - 5 จะถูกทำซ้ำ)
7. คำนวณหา $\bar{D}_j$ ซึ่งเป็นระยะทางภายในกลุ่มเฉลี่ย ทุก ๆ กลุ่ม $|(w_j^{(k)})|$ เป็นจำนวน pattern ของ Ω ที่รวมอยู่ในกลุ่ม $(w_j^{(k)})$ ที่จุดสิ้นสุดของการทำซ้ำของ K-means
8. เมื่อการบรรจบกันได้รับ (ทุกกลุ่มไม่เปลี่ยนไปจากการทำซ้ำในครั้งก่อน หรือได้รับค่าสูงสุดของการทำซ้ำแล้ว) กลุ่ม $(w_j^{(k)})$ ถูกเก็บไว้เป็น j กลุ่มที่ดีที่สุด j จะเพิ่มขึ้นและตรวจสอบขนาดกลุ่ม ถ้าถึง $j_{\max}$ แล้ว กระบวนการจะสิ้นสุด และศูนย์กลางกลุ่มสุดท้ายจะถูกสร้างขึ้น สำหรับทุกกลุ่ม ($j = 1, 2, \dots, j_{\max}, i = 1, 2, \dots, j$)
9. ถ้าขั้นตอนอื่นของ K-means ถูกทำงานนั้น อันดับแรกต้องหากกลุ่ม $(\lambda_{i,j})$ ซึ่งมีระยะทางภายในกลุ่มมากที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$\overline{D}_i' \geq \overline{D}_i \quad i \neq i'$$

กลุ่มที่มีขนาดใหญ่ที่สุดนี้ ถูกแบ่งแยกไปเป็น 2 กลุ่ม โดยการหากรณี i_1, i_2 ของทั้งสองแบบใน (λ_{j-1}') ดังนั้นระยะทางระหว่างแบบเหล่านี้จะมากที่สุด

$$\hat{d}(x_{i_1}, x_{i_2}) \geq \hat{d}(x_{i_3}, x_{i_4}) \quad \text{สำหรับทุก } i_3, i_4 \in (\lambda_{j-1}')$$

ศูนย์กลางกลุ่มเริ่มต้นใหม่ สำหรับกลุ่มที่ i ถูกตั้งเป็น แบบ x_{i_1} และ ศูนย์กลางกลุ่มเริ่มต้นใหม่ สำหรับกลุ่มที่ j (เช่นอันใหม่ที่ถูกรวมเข้าไปที่การทำซ้ำนี้) เป็นแบบ x_{i_2} ทุก ๆ ศูนย์กลางกลุ่มเริ่มต้นอื่น ๆ $c[(w_j^{(0)})]$ ถูกเลือกเป็นศูนย์กลางกลุ่ม $c[(w_{j-1}^i)]$ ที่จุดสิ้นสุดของการวนซ้ำของ K-means ในครั้งก่อนนั้น ตัวนับการวนซ้ำจะถูกรีเซ็ตไปเป็น $k = 1$

10. ทำการวนซ้ำอีก ขั้นตอนที่ 4 - 5

Template สุดท้ายถูกสร้างโดยการใช้จุดศูนย์กลางของการคำนวณกลุ่ม หรือโดยการหาค่าเฉลี่ยทุก pattern ในกลุ่ม ลักษณะเด่นของ K-means Algorithm คือมันจะคำนวณการแบ่งกลุ่มซึ่งครอบคลุม 100 % สำหรับทุกขนาดการแบ่งกลุ่มจาก 1 กลุ่มต่อ 1 คำ จนถึง j กลุ่มต่อ 1 คำ

บทที่ 6

แบบจำลองมาร์คอฟแบบซ่อนเร้น (Hidden Markov Model)

6.1 บทนำ

แบบจำลองมาร์คอฟเป็นแบบจำลอง (model) ทางสถิติซึ่งพัฒนามาเพื่อแบ่งกลุ่มของอนุกรมทางเวลา หรือสัญญาณที่ไม่คงที่ นั่นคือ ใช้สำหรับจัดกลุ่มของสัญญาณที่ไม่รู้จัก (Unknown signal) ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่งของสัญญาณ ซึ่ง แบบจำลองมาร์คอฟ ได้ถูกนำมาประยุกต์ใช้ในการรู้จำเสียงพูด และเป็นวิธีการที่วิทยานิพนธ์นี้เลือกใช้

แบบจำลองมาร์คอฟ แบ่งออกเป็น 2 ประเภทคือแบบต่อเนื่อง (Continuous) และแบบไม่ต่อเนื่อง (Discrete-time) ในวิทยานิพนธ์นี้ได้เลือกใช้แบบไม่ต่อเนื่อง เพราะเป็นวิธีการที่ซับซ้อนน้อยกว่า และใช้งานได้ดีกับคำพูดสั้นๆ [5] ที่ทำการทดลองในวิทยานิพนธ์นี้

6.2 กระบวนการแบบจำลองมาร์คอฟแบบไม่ต่อเนื่องทางเวลา

รูปที่ 6.1 เป็นรูปที่แสดงระบบที่มี 5 สถานะ (state) การอธิบายระบบ ณ เวลาหนึ่งๆ จะอธิบายด้วยเซตของกรณีสถานะ $\{1, 2, \dots, N\}$ โดยในแต่ละช่วงเวลา สามารถเปลี่ยนสถานะไป ด้วยเซตของความน่าจะเป็นที่เกี่ยวข้องกับสถานะ

จะแทนค่าเวลาที่เกี่ยวข้องกับการเปลี่ยนสถานะไป ด้วย $t = 1, 2, \dots, T$ และแทนสถานะที่เวลา t ด้วย q_t ในห่วงโซ่มาร์คอฟ (Markov Chain) นั้นจะแสดงความน่าจะเป็นในการย้ายสถานะไป โดยให้ความสนใจเพียงสเตตก่อนหน้านั้น เท่านั้น คือ

$$P[q_t=j \mid q_{t-1}=i, q_{t-2}=k, \dots] = P[q_t=j \mid q_{t-1}=i] \quad (6.1)$$

และ a_{ij} จะแทนความน่าจะเป็นของการย้ายสเตต จาก i ไป j

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$a_{ij} = P[q_t = j \mid q_{t-1} = i] \quad 1 \leq i, j \leq N \quad (6.2)$$

โดยมีคุณสมบัติว่า

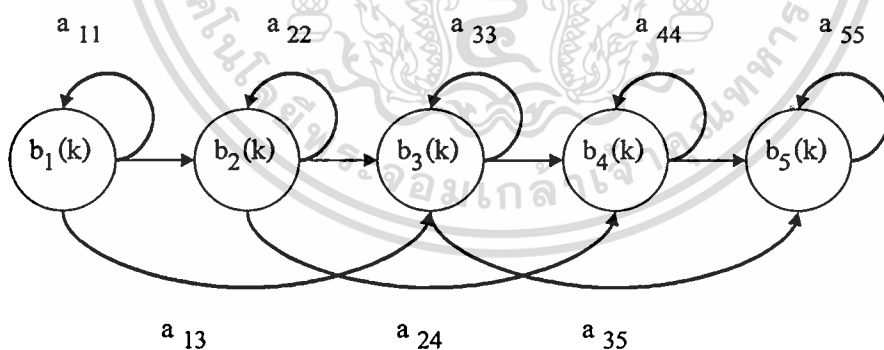
$$a_{ij} \geq 0 \quad \forall j, i \quad (6.3)$$

$$\sum_{j=1}^N a_{ij} = 1 \quad \forall i \quad (\text{ผลรวมความน่าจะเป็นทั้งหมดที่ออก} = 1) \quad (6.4)$$

ถ้าหากเอาที่พู่ (output) ของกระบวนการเป็นเซตของสแตตที่แต่ละเวลาหนึ่ง ซึ่งแต่ละสแตต หมายถึง เหตุการณ์ที่สังเกตได้แล้วจะเรียกมาร์คอฟแบบจำลองนี้ว่า มาร์คอฟแบบจำลองที่สังเกตเห็นได้ (Observable Markov Model)

ส่วน แบบจำลองมาร์คอฟ นั้น เหตุการณ์จะเป็นฟังก์ชันของความน่าจะเป็นของสแตต คือจะไม่ทราบเหตุการณ์ที่จะเกิดในแต่ละสแตตโดยตรง แต่จะรู้เหตุการณ์ได้โดยวิธีทางสถิติ ซึ่งจะทำให้ได้ลำดับของเหตุการณ์ออกมา [2], [7]

รูปที่ 6.1 ตัวอย่างแบบจำลองมาร์คอฟ 5 สแตต



6.3 ส่วนประกอบของ แบบจำลองมาร์คอฟ

1. N : จำนวนสแตตในแบบจำลอง โดยทั่วไปแล้วสแตตจะเกี่ยวข้องซึ่งกันและกัน ในแง่ที่ว่าสแตตจะสามารถย้ายมาได้จากสแตตอื่นๆ จะให้เซตของสแตตเป็น $\{1, 2, \dots, N\}$ และแทนสแตตที่เวลา t เป็น q_t

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. M : จำนวนของเหตุการณ์ต่างๆ ต่อหนึ่งสแตต จะแทนแต่ละสัญลักษณ์ เป็น $V = \{v_1, v_2, \dots, v_M\}$

3. ความน่าจะเป็นของการย้ายสแตต $A = \{a_{ij}\}$ เมื่อ $a_{ij} = P[q_i = j | q_{i-1} = i]$ $1 \leq i, j \leq N$
 สำหรับกรณีพิเศษที่สแตตใดๆ จะย้ายไปยังสแตตอื่นๆ จะได้ $a_{ij} \geq 0$ สำหรับทุกๆ i, j สำหรับ แบบ
 จาลองมาร์คอฟ ชนิดอื่นๆ จะได้ $a_{ij} = 0$ สำหรับคู่ (i, j) หนึ่งคู่ หรือมากกว่า

4. การกระจายความน่าจะเป็นของสัญลักษณ์ที่สังเกตได้

$$B = \{b_j(k)\}$$

$$\text{โดย } b(k) = P[o_i = v_k | q_i = j], 1 \leq k \leq m \quad (6.5)$$

เป็นนิยามการกระจายสัญลักษณ์ในสแตต $j, j=1, 2, \dots, N$

จะเห็นได้ว่า แบบจาลองมาร์คอฟ ต้องการพารามิเตอร์ของแบบจาลอง 2 ตัว คือ N และ M
 และกลุ่มของค่าความน่าจะเป็น A, B, π และ เพื่อความสะดวกจะใช้สัญลักษณ์เป็น (A, B, π) เพื่อ
 แสดงเซตของพารามิเตอร์ที่สมบูรณ์ของแบบจาลอง

6.4 ชนิดของ HMM

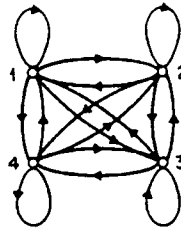
แบ่งชนิดตามลักษณะของการย้ายสแตต (Transition) ของเมตริกซ์ A

1. HMM แบบ Egordic Model หรือ Fully Connected Model : การย้ายสแตตสามารถย้าย
 ไปยังทุกๆ สแตตของแบบจาลอง ดังรูปที่ 6.2(a) เป็นตัวอย่างของแบบจาลองที่มี $N=4$

ซึ่งจากรูปนี้มี

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

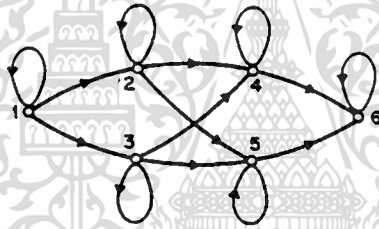
รูปที่ 6.2 แสดงแบบจำลองต่าง ๆ ของ HMM



(a)



(b)



(c)

2. HMM แบบ Left - Right Model หรือ Bakis Model : การย้ายสแตตจะย้ายจากซ้ายไปขวา ซึ่งจะมีคุณสมบัติของสัมประสิทธิ์การย้ายสแตตดังนี้

$$w_j = 0, \quad j < i$$

คือจะไม่มีมีการย้ายสแตตที่ต่ำกว่าสแตตปัจจุบัน และนอกจากนี้ก็ยังมีที่น่าจะเป็นของสแตต เริ่มต้นดังนี้

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

คือลำดับของสแตตต้องเริ่มที่สแตต 1 และ Left - Right Model นี้มักจะมีกฎบังคับการย้ายสแตต เพื่อไม่ให้เกิดการเปลี่ยนแปลงกรณีของสแตตมากนัก กล่าวคือ ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

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

ดังรูปที่ 6.2 (b) ค่าของ $\Delta i = 2$ คือจะไม่กระโดด (ย้าย) ไปเกิน 2 สเตต จะได้เมตริกซ์การย้ายสเตตเป็น

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}$$

จะเห็นว่าสเตตสุดท้าย สัมประสิทธิ์การย้ายสเตตจะเป็น

$$a_{NN} = 1$$

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

แบบจำลองแบบนี้จะเหมาะสมกับสัญญาณที่มีลักษณะเปลี่ยนแปลงตามเวลาอย่างต่อเนื่อง เช่น เสียงพูด

3. HMM แบบ parallel Left - Right Model : เป็นแบบจำลอง ที่มีความยืดหยุ่นมากกว่าแบบที่ 2 Left - Right Model ดังรูปที่ 6.2 (c)

6.5 ปัญหาพื้นฐานของ แบบจำลองมาร์คอฟ

6.5.1 ปัญหาที่ 1

เมื่อกำหนดลำดับเหตุการณ์ $O = \{o_1, o_2, \dots, o_T\}$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะคำนวณหา $P(O, \lambda)$ ซึ่งเป็นความน่าจะเป็นของลำดับเหตุการณ์นี้ได้อย่างไร การแก้ปัญหานี้จะทำให้สามารถเลือกแบบจำลองที่สอดคล้องกับเหตุการณ์ที่ดีที่สุดได้ จากแบบจำลองหลายอันที่มีอยู่

6.5.2 ปัญหาที่ 2

เมื่อกำหนดลำดับเหตุการณ์ $O = \{o_1, o_2, \dots, o_T\}$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะเลือกลำดับสแตต $q = \{q_1, q_2, \dots, q_T\}$ ที่เหมาะกับเหตุการณ์ที่สุดได้อย่างไร (อธิบายเหตุการณ์ได้ดีที่สุด) การแก้ปัญหานี้ช่วยหาลำดับของสแตตที่ถูกต้อง

6.5.3 ปัญหาที่ 3

จะปรับพารามิเตอร์ของแบบจำลอง $\lambda = (A, B, \pi)$ เพื่อให้ได้ค่าสูงสุดของ $P(O|\lambda)$ ได้อย่างไร การแก้ปัญหานี้ช่วยให้ได้ λ ที่เหมาะสมที่สุด เพื่ออธิบายการได้มาซึ่งลำดับเหตุการณ์ที่กำหนดให้ได้ดีที่สุด

ลำดับเหตุการณ์ที่ใช้ปรับค่าพารามิเตอร์ของแบบจำลอง เรียกว่าลำดับเทรนนิ่ง (training sequence) เนื่องจากเป็นลำดับเหตุการณ์ที่ถูกใช้เพื่อเทรนนิ่ง แบบจำลองมาร์คอฟ และด้วยการแก้ปัญหานี้จะได้แบบจำลองที่ดีที่สุดสำหรับคำนั้น

ตัวอย่าง สำหรับตัวรู้จำเสียงพูด (isolate word speech recognizer) ง่ายๆ เช่น สำหรับคำว่า ‘หนึ่ง’ ต้องการออกแบบ แบบจำลองมาร์คอฟ แบบ N สแตต จะแทนสัญญาณคำพูดของคำด้วยลำดับเวลาของเวกเตอร์สเปกตรัม สมมติว่าข้อมูลได้ถูกเข้ารหัสมาแล้วโดยใช้รหัสตารางสเปกตรัม (spectral codebook) ที่มีเวกเตอร์สเปกตรัมอยู่ M แบบ นั่นคือ แต่ละเหตุการณ์แทนด้วยดัชนีของเวกเตอร์สเปกตรัมที่ใกล้เคียงกับสัญญาณเสียงเดิมที่สุด ดังนั้นแต่ละคำศัพท์จะแทนด้วยลำดับเทรนนิ่ง ซึ่งประกอบด้วยลำดับของดัชนีไค์ดนูของคำจำนวนมาก สิ่งแรกที่จะต้องทำคือการสร้างแบบจำลองของแต่ละคำ โดยใช้ การแก้ปัญหานี้ 3 พารามิเตอร์ λ ของแบบจำลอง ที่เหมาะสมที่สุด สำหรับแต่ละแบบจำลองของคำ จากนั้นใช้การแก้ปัญหานี้ 2 เพื่อแบ่งแต่ละลำดับเทรนนิ่งของคำไปยังสแตตต่างๆ และเรียนรู้ถึงคุณสมบัติของเวกเตอร์สเปกตรัม ซึ่งจะช่วยให้ทราบถึงเหตุการณ์ที่เกิดขึ้นในแต่ละสแตต จุดประสงค์ก็เพื่อให้แบบจำลองมีความสามารถในการจำลองเสียงที่พูดเข้าไบนั่นเอง

เมื่อได้ แบบจำลองมาร์คอฟ ของคำว่า ‘หนึ่ง’ ที่เหมาะสมแล้ว ก็จะใช้ กรรแก้ปัญหานี้ 1 ในการหาความน่าจะเป็นไปได้ของแบบจำลองในการเกิดเหตุการณ์ ตามลำดับเหตุการณ์ของเสียง ที่ไม่รู้ว่าเป็นเสียงใดเข้ามา ถ้าแบบจำลองใดของคำว่า ‘หนึ่ง’ ให้ค่าความน่าจะเป็นสูงกว่าแบบ

จำลองของคำอื่นๆ ('สอง', 'สาม', 'สี่', ...) แสดงว่า คำที่ไม่ทราบว่าเป็นเสียงใด นี่คือคำว่า 'หนึ่ง' นั่นเอง

6.6 การแก้ปัญหาพื้นฐานของ แบบจำลองมาร์คอฟ

6.6.1 การแก้ปัญหาที่ 1 การคำนวณหาความน่าจะเป็น

เมื่อต้องการคำนวณหาความน่าจะเป็นของลำดับเหตุการณ์ $O = \{o_1, o_2, \dots, o_T\}$ เมื่อกำหนดแบบจำลองที่มีค่าพารามิเตอร์ λ ให้เช่น $P(O, \lambda)$ วิธีหนึ่งที่จะทำได้คือระบุสแตตที่เวลาต่างๆ จนถึงเวลา T ซึ่งเป็นไปได้ถึง N^T แบบ พิจารณาลำดับที่กำหนดไว้

$$q = \{q_1, q_2, \dots, q_T\} \quad (6.6)$$

เมื่อ q_1 เป็นสแตตเริ่มต้น ความน่าจะเป็นของลำดับเหตุการณ์ O ที่กำหนดไว้ คือ

$$P(O | q, \lambda) = \prod_{t=1}^T P(o_t | q_t, \lambda) \quad (6.7)$$

ความน่าจะเป็นในการเกิดเหตุการณ์คือ

$$P(O | q, \lambda) = b_{q_1}(o_1) \cdot b_{q_2}(o_2) \dots b_{q_T}(o_T) \quad (6.8)$$

ความน่าจะเป็นในการย้ายสแตต q จะเป็น

$$P(q | \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \quad (6.9)$$

เมื่อเอาความน่าจะเป็นของ O และ q มารวมกัน นั่นคือ ความน่าจะเป็นที่ O และ q จะเกิดพร้อมกัน จะได้

$$P(O, q | \lambda) = P(O | q, \lambda) P(q | \lambda) \quad (6.10)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ความน่าจะเป็นของ O ได้มาจากการรวมกันของความน่าจะเป็นร่วม โดยคิดจากลำดับสแตต q ที่เป็นไปได้ ดังนี้

$$P(O, q | \lambda) = \sum_{all q} P(O | q, \lambda) P(q | \lambda) \quad (6.11)$$

$$= \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) a_{q_2 q_3} \dots a_{q_{T-1} q_T} b_{q_T}(o_T) \quad (6.12)$$

เนื่องจากสมการดังกล่าวมีการคำนวณที่ยุ่งยาก ในรูปลำดับ $2T \cdot N^T$ ซึ่งแม้ค่า N และ T จะเป็นค่าน้อยๆ ดังนั้นจึงได้มีวิธีที่ง่ายกว่า มาช่วยแก้ปัญหาคือ 1 นี้ได้

6.6.1.1 กระบวนการไปหน้า (Forward Procedure)

นิยาม $\alpha_i(i) = P(o_1 o_2 \dots o_T, q_i = i | \lambda) \quad (6.13)$

เป็นความน่าจะเป็น ของการเกิดลำดับเหตุการณ์ $o_1 o_2 \dots o_T$ (จนกระทั่งเวลา t) และอยู่ที่สแตต i ณ เวลา t โดยมีแบบจำลองเป็น λ จะสามารถหา $\alpha_i(i)$ ได้ดังนี้

1. การเริ่มต้น (Initialization)

$$\alpha_i(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \quad (6.14)$$

เริ่มต้นด้วยการกำหนดความน่าจะเป็นไปหน้า เป็นความน่าจะเป็นร่วมของสแตต i และเหตุการณ์เริ่มต้น o_1

2. การเหนี่ยวนำ (Induction)

$$\alpha_{i+1}(j) = \left[\sum_{i=1}^N \alpha_i(i) a_{ij} \right] b_j(o_{t+1}), \quad 1 \leq t \leq T-1 \quad (6.15)$$

$$1 \leq j \leq N$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

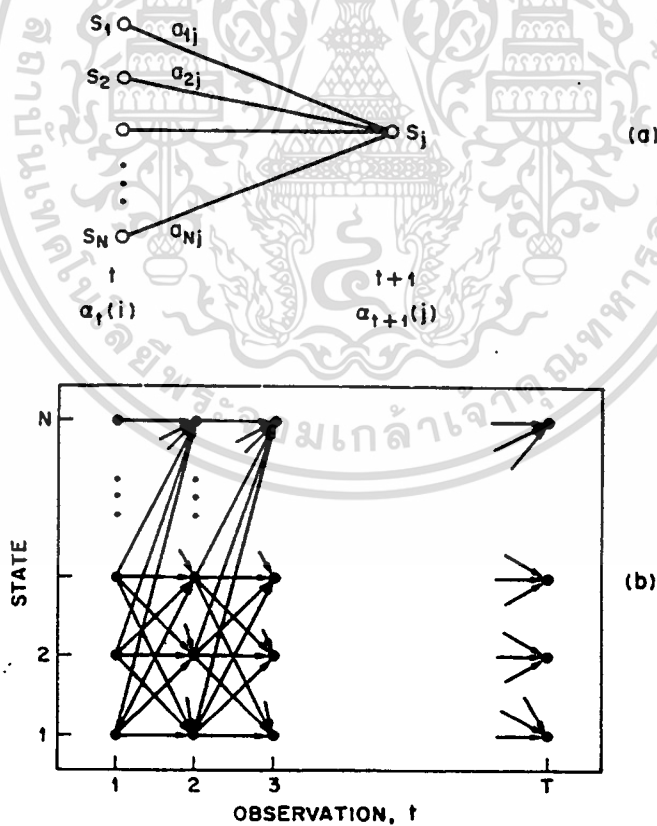
แสดงถึงว่า สเตต j ที่เวลา $t+1$ มาได้จากสเตต i ที่เป็นไปได้ถึง N สเตต, $1 \leq i \leq N$ ที่เวลา t ดังรูปที่ 6.3

3. การสิ้นสุด (Termination)

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i) \quad (6.16)$$

หา $P(O|\lambda)$ ได้จากผลรวมของ $\alpha_T(i)$ จากทุกๆ i ($1 \leq i \leq N$)

รูปที่ 6.3 กระบวนการไปหน้า



6.6.1.2 กระบวนการย้อนกลับ (Backward Procedure)

นิยาม $\beta_T(i) = P(o_{t+1}o_{t+2}\dots o_T, q_i | \lambda)$ (6.17)

คือความน่าจะเป็นของลำดับเหตุการณ์ส่วนหลัง จากเวลา $t+1$ ไปจนจบ โดยกำหนดว่าต้องอยู่ที่สแตต i ที่เวลา t และมีแบบจำลองเป็น λ จะคำนวณหา $\beta_T(i)$ ได้ดังนี้

1. การเริ่มต้น (Initialization)

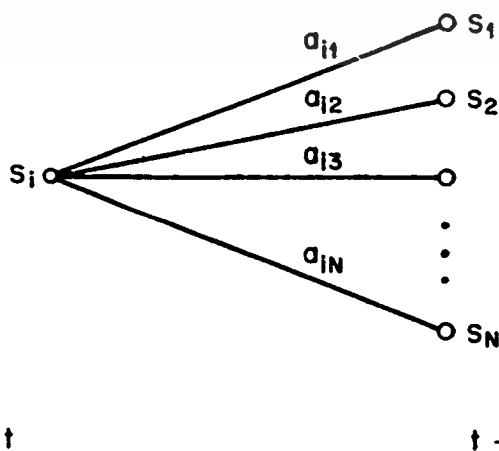
$$\beta_T(i) = 1, 1 \leq i \leq N \quad (6.18)$$

2. การเหนี่ยวนำ (Induction)

$$\beta_T(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \dots, 1, 1 \leq i \leq N \quad (6.19)$$

รูปที่ 6.4 แสดงว่า เพื่อที่เหตุการณ์ จะอยู่ที่สแตต i ที่เวลา t โดยคำนึงถึงลำดับเหตุการณ์จากเวลา $t+1$ ต้องพิจารณาสแตต j ที่จะเป็นไปได้ทั้งหมด ณ เวลา $t+1$ โดยจะขึ้นอยู่กับค่า a_{ij} และ $b_j(o_{t+1})$

รูปที่ 6.4 กระบวนการย้อนกลับ



6.6.2 การแก้ปัญหาที่ 2 การหาลำดับสแตตที่เหมาะสม

จะใช้วิธี วิทเทอร์บี อัลกอริทึม (Viterbi Algorithm) เพื่อหาลำดับสแตตที่ดีที่สุดใน เวลา t หนึ่ง ๆ เมื่อกำหนดลำดับเหตุการณ์ $O = (o_1 o_2 \dots o_t)$ โดยนิยาม

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P[q_1 q_2 \dots q_{t-1}, q_t = i, o_1 o_2 \dots o_t | \lambda] \quad (6.20)$$

$\max_i \delta_t(i)$ คือ $\delta_t(i)$ เป็นความน่าจะเป็นสูงสุดของเส้นทาง (path) ณ เวลา t ซึ่งเริ่มนับจากเหตุการณ์ที่เวลาเริ่มต้นจนถึงเวลา t ที่สแตต i และ โดยการเหนี่ยวนำ (induction) จะได้

$$\delta_{t+1}(i) = [\max_j \delta_t(j) a_{ji}] \cdot b_j(o_{t+1}) \quad (6.21)$$

จะใช้กระบวนการต่อไปนี้ ในการหาลำดับสแตตที่ดีที่สุด โดยให้ $\psi_t(i)$ เป็น อาร์เรย์ (array)

1. การเริ่มต้น (Initialization)

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \quad (6.22)$$

$$\psi_1(i) = 0 \quad (6.23)$$

2. การย้อนกลับ (Recursion)

$$\delta_t(j) = [\max_{1 \leq i \leq N} \delta_{t-1}(i) a_{ij}] \cdot b_j(o_t), \quad 2 \leq t \leq T \quad (6.24)$$

$$1 \leq j \leq N$$

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \quad 2 \leq t \leq T \quad (6.25)$$

$$1 \leq j \leq N$$

3. การสิ้นสุด (Termination)

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (6.26)$$

$$q_T^* = \max_{1 \leq i \leq N} \delta_T(i) \quad (6.27)$$

4. เส้นทางเดินย้อนกลับ (Path backtracking)

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \quad t=T-1, T-2, \dots, 1 \quad (6.28)$$

6.6.3 การแก้ปัญหาที่ 3

จะเลือกค่า $\lambda = (A, B, \pi)$ ที่ทำให้ $P(O|\lambda)$ มีค่าสูงสุด โดยใช้ กระบวนการทำซ้ำ (Iterative) ซึ่งวิธีที่เลือกใช้ คือวิธี บาม-เวลช์ (Baum-Welch) หรือ EM (Expectation-Maximization)

นิยาม

$$1. \gamma_t(i) = P(q_t = i | O, \lambda) \quad (6.29)$$

เป็นค่าความน่าจะเป็นที่จะอยู่ที่สแตต i ณ เวลา t โดยกำหนดลำดับเหตุการณ์ O และแบบจำลอง λ ให้ จะแสดงค่า $\gamma_t(i)$ ได้ดังนี้

$$\gamma_t(i) = P(q_t = i | O, \lambda) \quad (6.30)$$

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

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

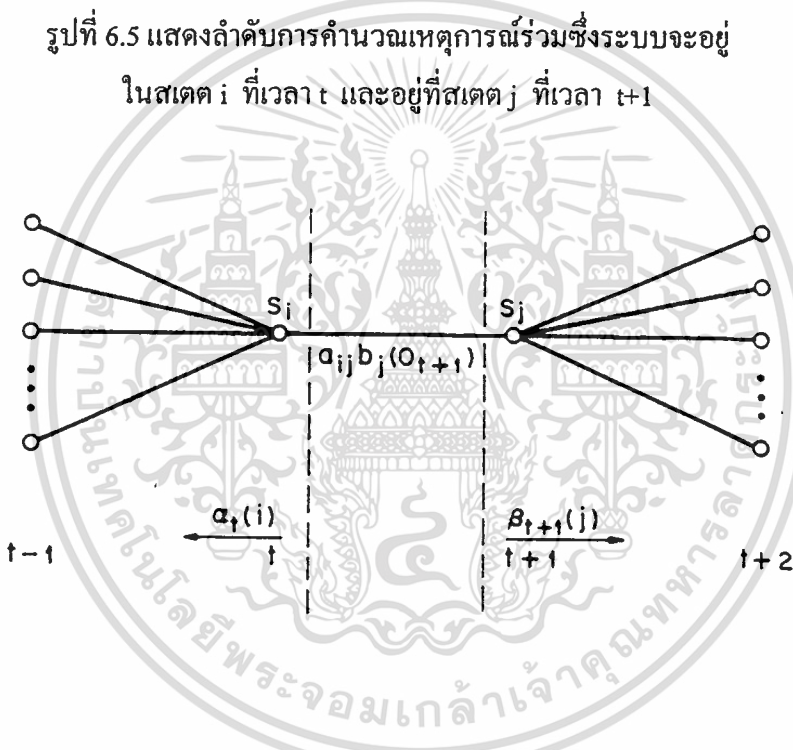
เนื่องจาก $P(O, q_t = i | \lambda)$ มีค่าเท่ากับ $\alpha_t(i)\beta_t(i)$ ดังนั้นจะได้ว่า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$\gamma_i(i) = \frac{\alpha(i)\beta_r(i)}{\sum_{i=1}^N \alpha(i)\beta_r(i)} \quad (6.31)$$

$$2. \mathcal{E}_t(i,j) = P(q_t=i, q_{t+1}=j | O, \lambda) \quad (6.32)$$

เป็นความน่าจะเป็น ที่จะอยู่ที่สแตต i ที่เวลา t และสแตต j ที่เวลา $t+1$ เมื่อกำหนดแบบจำลอง และลำดับเหตุการณ์ให้ ดังแสดงในรูปที่ 6.5



ซึ่งจากนิยามของตัวแปรไปหน้าและตัวแปรย้อนกลับ จะเขียน $\mathcal{E}_t(i,j)$ ได้ดังนี้

$$\begin{aligned} \mathcal{E}_t(i,j) &= \frac{P(q_t=i, q_{t+1}=j, O | \lambda)}{P(O | \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O | \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \end{aligned} \quad (6.33)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากที่ได้นิยาม $\gamma_t(i)$ แล้ว จะนำมาสัมพันธ์กับ $\varepsilon_t(i,j)$ ได้

$$\gamma_t(i) = \sum_{j=1}^N \varepsilon_t(i,j) \quad (6.34)$$

และ

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{จำนวนของการย้ายสแตตจากสแตต } i \text{ ในเหตุการณ์ } O \quad (6.35)$$

$$\sum_{t=1}^{T-1} \varepsilon_t(i,j) = \text{จำนวนของการย้ายสแตตจากสแตต } i \text{ ไป } j \text{ ในเหตุการณ์ } O \quad (6.36)$$

ดังนั้นจะสามารถหาเซตของพารามิเตอร์ได้ดังนี้

$$\begin{aligned} \pi' &= \text{จำนวนครั้งในการอยู่ที่สแตต } i \text{ ที่เวลา } t=1 \\ &= \gamma_t(i) \end{aligned} \quad (6.37)$$

$$\begin{aligned} a_{ij}' &= \frac{\text{จำนวนครั้งที่คาดว่าจะย้ายจากสแตต } i \text{ ไป } j}{\text{จำนวนครั้งที่คาดว่าจะย้ายจากสแตต } i} \\ &= \frac{\sum_{t=1, O_t=V_k}^T \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \end{aligned} \quad (6.38)$$

$$\begin{aligned} b_j(k) &= \frac{\text{จำนวนครั้งที่คาดว่าจะอยู่ในสแตต } j \text{ และเกิดเหตุการณ์ } V_k}{\text{จำนวนครั้งที่คาดว่าจะอยู่ที่สแตต } j} \end{aligned}$$

$$= \frac{\sum_{t=1, O_t=V_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (6.39)$$

จะให้ $\lambda=(A,B,\pi)$ เป็นแบบจำลองปัจจุบัน และจะใช้ λ นี้ คำนวณในด้านขวาของสมการ และให้แบบจำลองที่ได้จากการคำนวณซ้ำๆ เป็น $\lambda'=(A',B',\pi')$ เป็นแบบจำลองที่ได้จาก
 ไม่่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้างซ้ายของสมการ จะได้จุดวิกฤตของฟังก์ชันความน่าจะเป็นในกรณีที่ $\lambda' = \lambda$ หรือถ้า λ' มีความน่าจะเป็นมากกว่าแบบจำลอง λ ในลักษณะที่ $P(O|\lambda') > P(O|\lambda)$ นั่นคือ ก็จะได้ แบบจำลอง λ' ใหม่ ที่น่าจะทำให้เกิดลำดับเหตุการณ์ 0 ได้มากกว่า

จากกระบวนการข้างต้น ถ้าคำนวณซ้ำๆ โดยใช้ λ' แทน λ แล้ว จะทำให้ความน่าจะเป็นของการเกิดเหตุการณ์ 0 จะดีขึ้น จนกระทั่งถึงจุดวิกฤต ซึ่งทั้งหมดนี้ก็คือการประมาณค่าความน่าจะเป็นสูงสุด (ML estimate) ของแบบจำลองมาร์คอฟ นั่นเอง

6.7 การปรับปรุงค่าพารามิเตอร์ของ HMM

6.7.1 การสเกล (Scaling)

เนื่องจาก $\alpha_t(i)$ จะประกอบด้วยผลรวมของเทอมจำนวนมาก ซึ่งแต่ละเทอม คือ

$$\left(\prod_{s=1}^{t-1} a_{q_s, q_{s+1}} \prod_{s=1}^t b_{q_s}(O_s) \right)$$

เมื่อ $q_t = i$ และ b เป็นความน่าจะเป็นในสมการ (6.5) เนื่องจากแต่ละเทอม a และ b มีค่าน้อยกว่า 1 แสดงว่าเมื่อ t มากขึ้น แต่ละเทอมของ $\alpha_t(i)$ จะเข้าสู่ศูนย์ ซึ่งถ้า t มากขึ้นเรื่อย ๆ Dynamic Range ของการคำนวณ $\alpha_t(i)$ เกิน Range ของคอมพิวเตอร์ ทำให้ค่าที่ได้ไม่ถูกต้อง จึงได้มีการสเกลขึ้น

การสเกลทำได้โดยการคูณ $\alpha_t(i)$ โดยสัมประสิทธิ์การสเกล ซึ่งไม่ขึ้นกับ i เพื่อให้ $\alpha_t(i)$ อยู่ภายใน Dynamic Range ของคอมพิวเตอร์ $1 \leq t \leq T$ การสเกล $B_t(i)$ ก็เช่นเดียวกัน หลังการคำนวณค่า การสเกล ก็จะตัดกันหมดไปเอง

พิจารณาดังนี้ จาก สมการ (6.38)

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^T \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad (6.40)$$

ในการคำนวณ $\alpha_{\tau}(i)$ ใช้ $\alpha_{\tau-1}(i)$ แทน α ที่ยังไม่ได้สเกล

$\hat{\alpha}_{\tau-1}(i)$ เป็น α ที่สเกลแล้ว

และ $\hat{\alpha}_{\tau-1}(i)$ แทนเวอร์ชันของ α ก่อนการสเกล

เมื่อ $t = 1$ จะได้ $\hat{\alpha}_1(i) = c_1 \alpha_{\tau}(i)$

$$\text{เมื่อ } c_1 = \frac{1}{\sum_{i=1}^N \alpha_{\tau}(i)}$$

เมื่อ $2 \leq t \leq T$ คำนวณ $\hat{\alpha}_{\tau}(i)$ จาก สมการ 6.15 ในเทอมของ $\hat{\alpha}_{\tau-1}(i)$ ค่าก่อน

$$\hat{\alpha}_{\tau}(i) = \sum_{j=1}^N \hat{\alpha}_{\tau-1}(j) a_{ji} b_j(O_{\tau}) \quad (6.41a)$$

เมื่อ c_{τ} สัมประสิทธิ์ การสเกล เป็น

$$c_{\tau} = \frac{1}{\sum_{i=1}^N \hat{\alpha}_{\tau}(i)} \quad (6.41b)$$

$$\text{โดยให้ } \alpha_{\tau}(i) = c_{\tau} \hat{\alpha}_{\tau}(i) \quad (6.41c)$$

จากสมการ (6.41) จะเขียนได้ว่า

$$\alpha_{\tau}(i) = \frac{\sum_{j=1}^N \hat{\alpha}_{\tau-1}(j) a_{ji} b_j(O_{\tau})}{\sum_{i=1}^N \sum_{j=1}^N \hat{\alpha}_{\tau-1}(j) a_{ji} b_j(O_{\tau})} \quad (6.42)$$

โดย การเหนี่ยวนำ จะได้

$$\hat{\alpha}_{\tau-1}(j) = \left(\prod_{\tau=1}^{t-1} c_{\tau} \right) \alpha_{\tau-1}(j) \quad (6.43a)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จะเขียนได้ว่า

$$\hat{\alpha}_i(i) = \frac{\sum_{j=1}^N \alpha_{i-1}(j) \left(\prod_{\tau=1}^{i-1} c_\tau \right) a_{ji} b_i(O_i)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_{i-1}(j) \left(\prod_{\tau=1}^{i-1} c_\tau \right) a_{ji} b_i(O_i)} = \frac{\alpha_i(i)}{\sum_{i=1}^N \alpha_i(i)} \quad (6.43b)$$

นั่นก็จะสเกล $\alpha_i(i)$ ได้โดยหารด้วยผลรวมของ $\alpha_i(i)$ ทั้งหมด และสเกล $\beta_i(i)$ ด้วยค่าเดียวกันนี้ในเทอมของการสเกลนี้สมการ (6.4) จะเป็น

$$\bar{a}_{ij} = \frac{\sum_{i=1}^{T-1} \hat{\alpha}_i(i) a_{ij} b_j(O_{i+1}) \hat{\beta}_{i+1}(j)}{\sum_{i=1}^{T-1} \sum_{j=1}^N \hat{\alpha}_i(i) a_{ij} b_j(O_{i+1}) \hat{\beta}_{i+1}(j)} \quad (6.44)$$

โดยแต่ละ $\hat{\alpha}_i(i), \hat{\beta}_{i+1}(j)$ เขียนได้เป็น

$$\hat{\alpha}_i(i) = \left[\prod_{s=1}^i c_s \right] \alpha_i(i) = C_i \alpha_i(i) \quad (6.45)$$

$$\hat{\beta}_{i+1}(j) = \left[\prod_{s=1}^T c_s \right] \beta_{i+1}(j) = D_{i+1} \beta_{i+1}(j) \quad (6.46)$$

ดังนั้นสมการ (6.44) จะเขียนได้เป็น

$$\bar{a}_{ij} = \frac{\sum_{i=1}^{T-1} C_i \alpha_i(i) a_{ij} b_j(O_{i+1}) D_{i+1} \beta_{i+1}(j)}{\sum_{i=1}^{T-1} \sum_{j=1}^N C_i \alpha_i(i) a_{ij} b_j(O_{i+1}) D_{i+1} \beta_{i+1}(j)} \quad (6.47)$$

ซึ่งเทอม $C_i D_{i+1}$ จะเขียนได้ในเทอม

$$C_i D_{i+1} = \prod_{s=1}^i c_s \prod_{s=i+1}^T c_s = \prod_{s=1}^T c_s = C_T \quad (6.48)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ซึ่งไม่ขึ้นกับเวลา t ดังนั้น $C_{D_{t+1}}$ จะถูกตัดทิ้ง ทั้งเศษและส่วนของสมการ (6.47) ซึ่งทำให้ได้สูตรการคำนวณซ้ำ ๆ (re-estimate) เดิมกลับคืนมา

กระบวนการ สเกล ดังกล่าวนี้นี้ก็ใช้ได้กับสัมประสิทธิ์ β และ π ในการสเกลจะทำให้การคำนวณค่า $P(O|\lambda)$ เปลี่ยนไป จะไม่สามารถหาได้จากการรวมกับของเทอม $\alpha_T(i)$ แต่จะหาจากคุณสมบัติ

$$\prod_{i=1}^T c_i \sum_{i=1}^N \alpha_T(i) = C_T \sum_{i=1}^N \alpha_T(i) = 1 \quad (6.49)$$

ดังนั้นจะได้

$$\prod_{i=1}^T c_i P(O|\lambda) = 1 \quad (6.50)$$

$$P(O|\lambda) = \frac{1}{\prod_{i=1}^T c_i} \quad (6.51)$$

$$\log[P(O|\lambda)] = \sum_{i=1}^T \log c_i \quad (6.52)$$

ซึ่งจะได้ค่า $\log$ ของ P ไม่ใช่ค่าของ P เพราะมันจะเกิน dynamic range ของคอมพิวเตอร์

6.7.2 ลำดับของเหตุการณ์หลายลำดับ (Multiple Observation Sequence)

ในการใช้แบบจำลองแบบ Left-Right นั้น ไม่สามารถใช้ลำดับเหตุการณ์เพียงลำดับเดียวในการเทรนแบบจำลองได้ ต้องใช้หลาย ๆ เหตุการณ์เข้ามาเทรน เพื่อให้ได้ค่าพารามิเตอร์ที่ถูกต้องมากขึ้น ถ้าให้เซตของ v ลำดับ เหตุการณ์ เป็น

$$O = [O^{(1)}, O^{(2)}, \dots, O^{(k)}]$$

เมื่อ $O^{(v)} = O_1^{(v)} O_2^{(v)} \dots O_{T_v}^{(v)}$ เป็นลำดับเหตุการณ์ลำดับที่ v สมมติว่าแต่ละลำดับ เหตุการณ์ เป็นอิสระต่อกัน จุดมุ่งหมายก็คือปรับค่าพารามิเตอร์ของ λ ให้ดีที่สุด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$P(O|\lambda) = \prod_{v=1}^V P \quad (6.53)$$

$$P(O|\lambda) = \prod P_v \quad (6.54)$$

เนื่องจากสูตรการคำนวณซ้ำๆ มาจากพื้นฐานของความน่าจะเป็นของการเกิดของเหตุการณ์ต่างๆ สูตรการคำนวณซ้ำของลำดับเหตุการณ์หลายๆ ลำดับ ก็จะสามารถหาได้จากการเอาความถี่ของการเกิดเหตุการณ์ต่างๆ แต่ละเหตุการณ์มารวมกัน ดังนั้นจะได้สูตรหา $a_{ij}, b_i(i)$ เป็น

$$\overline{a_{ij}} = \frac{\sum_{v=1}^V \frac{1}{P_v} \sum_{t=1}^{T_v-1} \alpha_t^v(i) a_{ij} b_j(O_{t+1}^v) \beta_{t+1}^v(j)}{\sum_{v=1}^V \frac{1}{P_v} \sum_{t=1}^{T_v-1} \alpha_t^v(i) \beta_t^v(i)} \quad (6.55)$$

$$\overline{b_j(i)} = \frac{\sum_{v=1}^V \frac{1}{P_v} \sum_{t=1}^{T_v-1} \alpha_t^v(i) \beta_t^v(i)}{\sum_{v=1}^V \frac{1}{P_v} \sum_{t=1}^{T_v-1} \alpha_t^v(i) \beta_t^v(i)} \quad (6.56)$$

ส่วน π_i ไม่ต้องคำนวณเนื่องจาก $\pi_1=1$, $\pi_i=1, i \neq 1$ การสเกลที่เหมาะสมของสมการ (6.55)-(6.56)

$$\overline{a_{ij}} = \frac{\sum_{v=1}^V \sum_{t=1}^{T_v-1} \hat{\alpha}_t^v a_{ij} b_j(O_{t+1}^v) \hat{\beta}_{t+1}^v(j)}{\sum_{v=1}^V \sum_{t=1}^{T_v-1} \sum_{j=1}^N \hat{\alpha}_t^v(i) a_{ij} b_j(O_{t+1}^v) \hat{\beta}_{t+1}^v(j)} \quad (6.57)$$

$$\overline{b_j(k)} = \frac{\sum_{v=1}^V \sum_{t=1, O_t=k}^{T_v-1} \hat{\alpha}_t^v a_{ij} b_j(O_{t+1}^v) \hat{\beta}_{t+1}^v(j)}{\sum_{v=1}^V \sum_{t=1}^{T_v-1} \sum_{j=1}^N \hat{\alpha}_t^v(i) a_{ij} b_j(O_{t+1}^v) \hat{\beta}_{t+1}^v(j)} \quad (6.58)$$

6.8 ระบบแบบจำลองมาร์คอฟ

เมื่อมีคำศัพท์อยู่ V คำ ในการทำการรู้จำได้ จะต้องทำการสร้างแบบจำลองของคำ แต่ละคำ ด้วย แบบจำลองมาร์คอฟ ที่แตกต่างกัน คำแต่ละคำจะมี ลำดับเทรนนิ่งที่ได้มาจากคุณลักษณะ เฉพาะของคำศัพท์นั้นๆ

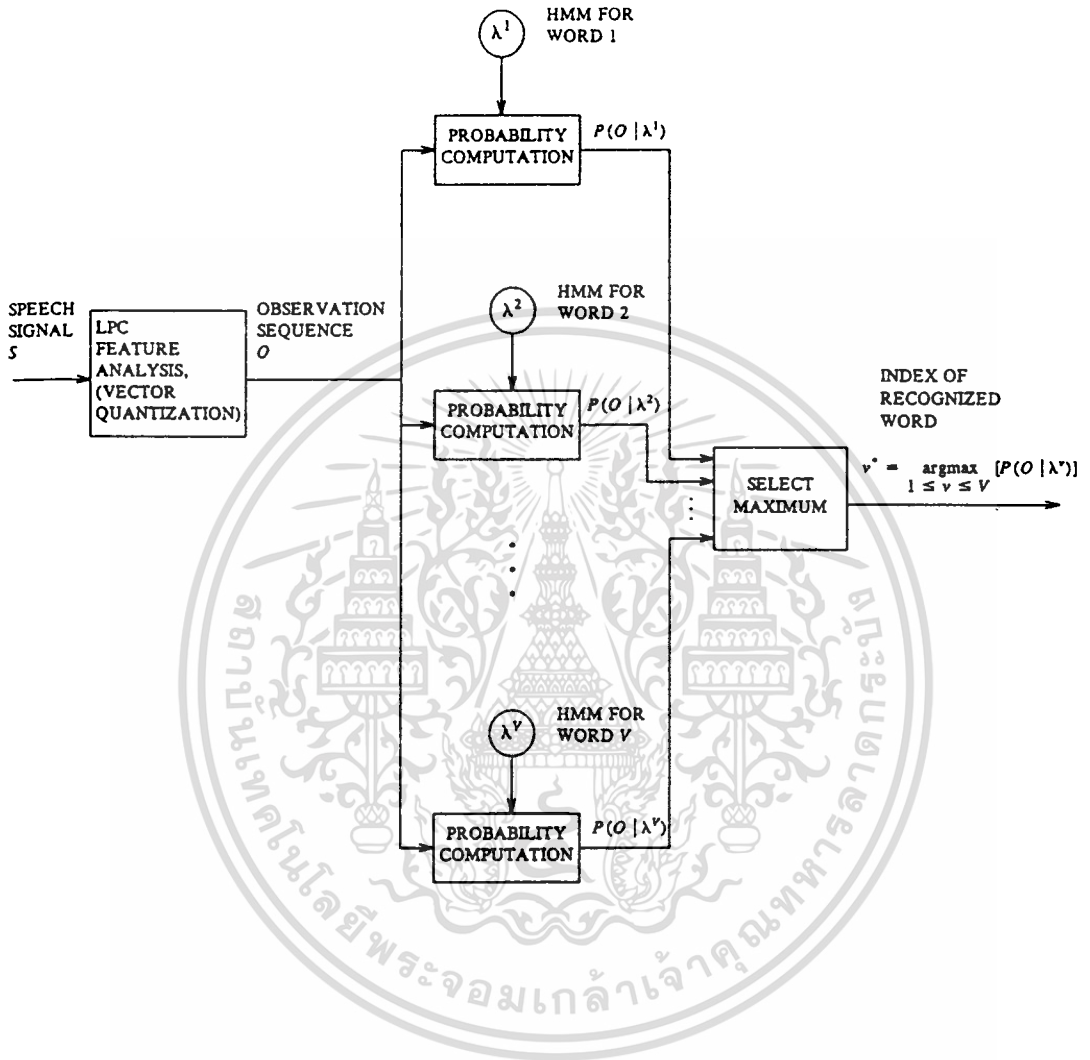
การที่จะรู้จำคำพูดได้ต้องทำได้ ดังนี้

1. สำหรับแต่ละคำศัพท์ V ต้องสร้าง แบบจำลองมาร์คอฟ λ_v นั่นคือการหาค่า (A, B, π) ที่เหมาะสมกับลำดับเทรนนิ่งของคำ V นั้น
2. ในการจะรู้จำคำศัพท์แต่ละคำ (ดูรูป ที่ 6.6) หลังจากที่มีลำดับเหตุการณ์ของคำพูดเข้ามา (เสียงพูดที่ผ่าน LPC และ การจัดระดับเวกเตอร์มาแล้ว) จะทำการหาค่า $P(O|\lambda)$ ของทุกๆ แบบจำลอง แล้วเลือกแบบจำลองที่มีค่าความน่าจะเป็นในการเกิดเหตุการณ์สูงสุด คือ

$$V^* = \arg \max P(O|\lambda_v) \quad (6.59)$$

คำศัพท์ที่สอดคล้องกับแบบจำลองดังกล่าวนี้ จะเป็นคำเดียวกับคำศัพท์ที่ต้องการจะรู้จำนั่นเอง โดยขั้นตอนการคำนวณหาค่าความน่าจะเป็น จะใช้วิธีวิทเทอร์บี ดังที่ได้กล่าวไว้ในตอนต้น

รูปที่ 6.6 บล็อกไดอะแกรม ของการรู้จำคำโคคด้วยแบบจำลองของมาร์คอฟ



บทที่ 7

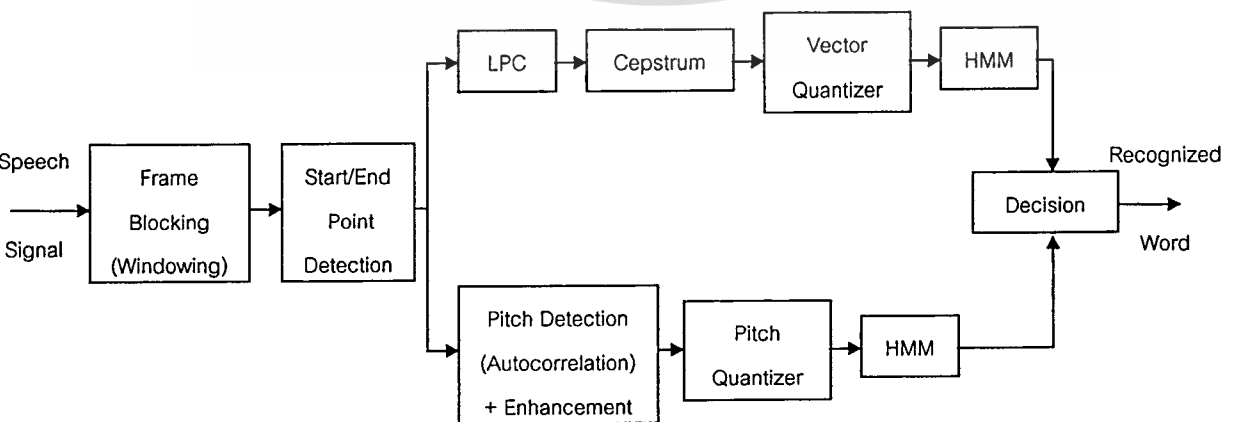
การดำเนินการทดลองและผลการทดลอง

7.1 แบบจำลองการรู้จำเสียงภาษาไทย

เนื่องจากเสียงพูดในภาษาไทยเป็นภาษาเสียงวรรณยุกต์ (tonal language) ซึ่งแตกต่างจากเสียงพูดในภาษาอังกฤษที่มีเฉพาะเสียงพยัญชนะและเสียงสระเท่านั้น ในภาษาที่ไม่ใช่ภาษาเสียงวรรณยุกต์คำแต่ละคำจะเปลี่ยนระดับเสียงไปอย่างไรก็ยังคงความหมายเดิม เช่น คำ *คาร์* (car) ในภาษาอังกฤษ ถึงแม้จะออกเสียงเพี้ยนเป็น ข้า คำ (ข้า) คำ ขา ก็ยังคงความหมายว่า รถยนต์ เหมือนกันทุกคำ ซึ่งวิธีการรู้จำเสียงพูดที่มีใช้กับภาษาอังกฤษไม่เหมาะที่จะใช้แยกความแตกต่างนี้ เช่น เสียงคำว่า โอ โอ โอ้ โอ้ โอ้ ในภาษาอังกฤษจะแทนเสียงทั้ง 5 เสียงเป็นคำว่า oh เพียงคำเดียว เป็นต้น ดังนั้นแบบจำลอง (model) ในการรู้จำเสียงจึงถูกแบ่งแยกออกเป็น 2 ส่วน คือ

1. ส่วนที่ใช้ในการรู้จำเสียงพูดโดยทั่วไปที่ใช้สำหรับภาษาอังกฤษ (เสียงพยัญชนะ+เสียงสระ)
2. ส่วนที่ใช้ในการรู้จำเสียงวรรณยุกต์

รูปที่ 7.1 แผนภาพแสดงการรู้จำเสียงในภาษาไทยด้วยแบบจำลอง HMM



โดยในการรู้จำเสียงคำพูดในภาษาไทยนั้นจะใช้แบบจำลองทั้งสองในการตัดสินใจว่าเป็นคำพูดคำใด ซึ่งแบบจำลองในส่วนแรกจะใช้สัมประสิทธิ์ LPC เป็นข้อมูลในการสร้างแบบจำลอง HMM (Hidden Markov Model) สำหรับส่วนที่สองจะใช้ค่าพิตช์ (pitch) เป็นข้อมูลในการสร้างแบบจำลอง HMM โดยแบบจำลองทั้งสองส่วนจะใช้ในการแปลความหมายของเสียงพูดว่าเป็นคำพูดคำใดร่วมกัน ดังในรูปที่ 7.1

7.2 ค่าพารามิเตอร์ต่างๆ ที่ใช้ในการสร้างแบบจำลอง

การเก็บตัวอย่างข้อมูลเสียงพูด (sound sample) ใช้เครื่องคอมพิวเตอร์ส่วนบุคคลและการ์ดเสียง และเก็บข้อมูลอยู่ในรูปไฟล์ข้อมูลแบบ '.wav' ข้อมูล 1 ตัวอย่างของเสียงจะถูกแทนด้วยข้อมูลขนาด 8 บิต ซึ่งใช้ความถี่ในการชักตัวอย่าง (sampling frequency) เท่ากับ 11.025 kHz ไฟล์ข้อมูล '.wav' นี้จะถูกใช้เป็นอินพุตสำหรับการคำนวณของโปรแกรมที่เขียนขึ้น ในวิทยานิพนธ์นี้เลือกใช้โปรแกรม MATLAB ในการทดลองและพัฒนาวิธีการต่างๆ ที่ใช้ในการรู้จำเสียงพูด

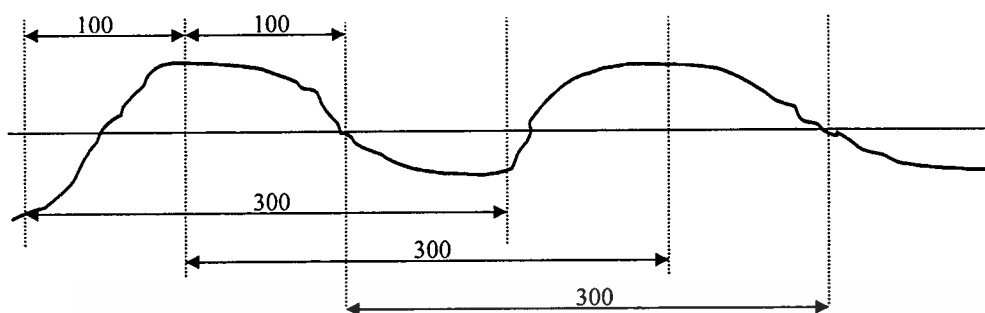
หลังจากข้อมูลเสียงถูกโหลดเข้าในโปรแกรม จะถูกปรับขนาดเพื่อให้อัตราส่วนของเสียงที่ได้ในกรณีไม่มีเสียงมีค่าเป็นเทียบเป็น 0 และจะถูกทำให้เป็นบรรทัดฐาน (normalize) โดยเทียบกับขนาดแอมพลิจูดที่มากที่สุดของเสียงนั้นและเทียบให้ค่าแอมพลิจูดมากสุดนั้นมีค่าขนาดเป็น 100 เพื่อให้เสียงทุกเสียงที่ใช้ในการทดสอบอยู่ในระดับความดังของเสียงใกล้เคียงกัน

ข้อมูลเสียงที่ได้จากการทำบรรทัดฐาน (normalize) แล้วจะถูกนำมาทำพรีเอมฟาซิส (pre-emphasis) โดยวงจรกรองความถี่สูง (high pass filter) ที่มีฟังก์ชันถ่ายโอน (transfer function) เพื่อหาค่าจํากัดของตัวเลขทศนิยมในการประมวลผลสัญญาณ ดังนี้

$$H(z) = 1 - 0.9375 z^{-1}$$

ค่า 0.9375 นี้ได้มาจากตัวเลขเศษส่วน 15/16 ซึ่งเป็นค่าที่นิยมใช้ในการทำพรีเอมฟาซิสแบบ fixed-point ([2], หน้า 113) จากนั้นจึงแบ่งช่วงข้อมูลให้เป็นเฟรม (Frame Blocking) โดยในแต่ละเฟรมมีขนาด 300 ข้อมูล และเฟรมแต่ละเฟรมจะถูกเลื่อนไปที่ละ 100 ข้อมูล ดังแสดงในรูปที่ 7.2

รูปที่ 7.2 แสดงการแบ่งช่วงเฟรมข้อมูล



ข้อมูลที่ถูกแบ่งเป็นเฟรมแล้วจะถูกนำมาหาค่าพลังงานของสัญญาณในเฟรมข้อมูลนั้นตามสมการ

$$M_f(0) = \sum_{m=0}^{N-1} |x[m]|$$

โดยที่ $M_f(0)$, N , และ L คือ ค่าพลังงาน, ขนาดของเฟรม (frame size), และลำดับที่ของเฟรม และหาเฟรมที่มีพลังงานสูงสุดซึ่งจะใช้เป็นตัวเทียบในการตัดเฟรมข้อมูลที่มีค่าพลังงานน้อยกว่า 10% ของค่าพลังงานสูงสุด เพื่อตัดข้อมูลให้เหลือในส่วนที่เป็นเสียงที่ต้องการใช้ในการวิเคราะห์จริงๆ

ข้อมูลในแต่ละเฟรมที่จะใช้ในการหาค่าสัมประสิทธิ์ LPC จะถูกคูณด้วยฟังก์ชันวินโดว์แฮมมิง (hamming window function) เพื่อลดความไม่ต่อเนื่องของข้อมูลที่ขอบผ่านการประมวลผลเชิงตัวเลข ซึ่งนิยามโดยสมการดังนี้ ([2], หน้า 114)

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad \text{เมื่อ } n = 0, 1, 2, \dots, N-1$$

จากนั้นข้อมูลที่ถูกคูณด้วยฟังก์ชันวินโดว์จะถูกนำไปหาค่า สัมประสิทธิ์ LPC และค่า Gain ซึ่งก็คือค่าโดยประมาณของสเปกตรัมของสัญญาณเสียงในแต่ละเฟรมนั่นเอง โดยใช้จำนวนลำดับ (LPC order) ของสัมประสิทธิ์ LPC เท่ากับ 12 แล้วนำไปแปลงให้เป็นค่าเซปตรัมที่มีลำดับ (Cepstrum order) เท่ากับ 19 ([2], หน้า 115) เพื่อให้ข้อมูลที่ได้นี้มีค่าใกล้เคียงกับการรับรู้เสียงตามความรู้สึกของมนุษย์ รวมทั้งทำการเวทค่าพารามิเตอร์หรือถ่วงน้ำหนัก เพื่อลดค่าความไวต่อเสียงรบกวน (noise) ที่จะเกิดขึ้นกับข้อมูลในเฟรมที่มีลำดับต่ำ (low-order) และสูง (high-order) ด้วย ข้อมูลที่ได้ต่อเฟรมของสัญญาณเสียงจะเหลือเพียง 19 ข้อมูล ซึ่งทำให้ลดปริมาณข้อมูลที่จะต้องนำไปใช้ในการคำนวณในกระบวนการถัดไปเป็นจำนวนมาก (19:300)

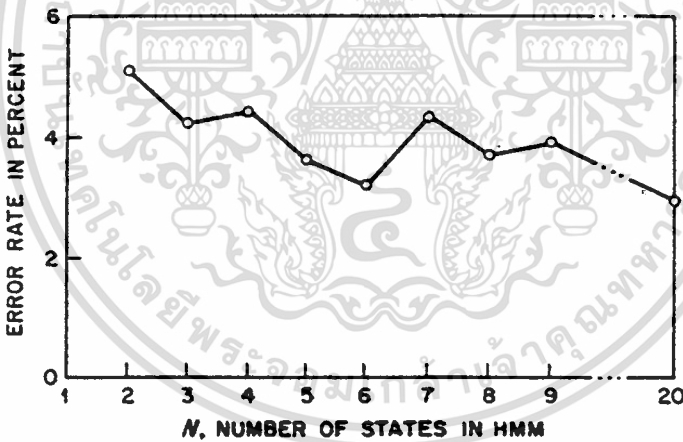
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ถึงแม้ว่าข้อมูลที่ได้เมื่อเทียบกับข้อมูลตัวอย่างของเสียงจะลดลงเหลือเพียงเฟรมละ 19 ค่า จากข้อมูลเดิมที่มีจำนวนเฟรมละ 300 ค่า แต่ถ้าใช้การควอนไทซ์แบบเวกเตอร์ (Vector Quantization) ก็ยังสามารถที่จะลดปริมาณข้อมูลต่อเฟรมลงไปได้อีก โดยจะเหลือเพียง 1 ค่าต่อหนึ่งเฟรมเท่านั้น ในวิทยานิพนธ์นี้จะใช้ค่าในการทำเวกเตอร์ควอนไทเซชันเป็น 64 ระดับ (สร้างโค้ดบุคที่มีขนาดเป็น 64 กลุ่ม) เพื่อนำไปใช้ในการสร้างแบบจำลอง HMM ต่อไป

ส่วนการหาค่าพิตช์นั้นจะใช้วิธีฮอโดคอร์รีเรชันและนำมาหาระยะทางจากตำแหน่งที่มีค่าฮอโดคอร์รีเรชันสูงสุดถัดจากค่าฮอโดคอร์รีเรชันที่ตำแหน่ง 0 เทียบกับที่ตำแหน่ง 0 ตามหัวข้อ 3.2 ระยะทางที่ได้ก็คือค่าพิตช์ และนำค่าพิตช์ที่ได้ไปปรับแต่งตามที่กล่าวถึงในหัวข้อ 3.3 ซึ่งจะตัดค่าพิตช์ที่มีค่ามากกว่า 30% (ได้จากการทดลองที่ให้ผลของความต่อเนื่องของพิตช์ดีที่สุด) ของค่าพิตช์เฉลี่ย P_i ทำให้ค่าพิตช์ที่ได้มีความต่อเนื่องและเหมาะสมที่จะนำไปใช้ในการสร้างแบบจำลอง

รูปที่ 7.3 แสดงความสัมพันธ์ระหว่างค่าความคลาดเคลื่อนกับจำนวนสเตต



ในวิทยานิพนธ์นี้ได้เลือกสร้างแบบจำลอง HMM เป็นแบบ left-right เพราะเป็นแบบจำลองที่เหมาะสมสำหรับการรู้จำรูปแบบของคำ ประเภทคำศัพท์เดี่ยว (Isolated word) เนื่องจากจะสามารถนำเวลาเข้ามาเกี่ยวข้องกับสเตต (state) ของแบบจำลองได้โดยตรง และอาจตีความหมายทางกายภาพของสเตตเป็นเสียงที่แตกต่างกันของคำได้ เนื่องจากความคลาดเคลื่อนในการรู้จำไม่แปรผันโดยตรงกับจำนวนสเตตที่มากพอ ([2], หน้า 380) ดังนั้นการเลือกจำนวนสเตตจะใช้จำนวนสเตตเท่า ๆ กันทุก ๆ คำ ซึ่งจำนวนสเตตที่เลือกใช้คือ 6 เนื่องจากเป็นค่าที่มีค่าความคลาดเคลื่อนเฉพาะบริเวณต่ำสุด (local minimum at $N = 6$ state) ดังแสดงในรูปที่ 7.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้าไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7.3 การเลือกกลุ่มตัวอย่างของเสียงและผลการทดลอง

เนื่องจากเสียงในภาษาไทยที่เปล่งออกมาเป็นคำ 1 โศกหนึ่งพยางค์โดยทั่วไปจะประกอบด้วยรูปแบบที่เป็นไปได้ 2 รูปแบบดังที่กล่าวถึงในข้างต้นคือ

1. พยางค์ที่ประกอบไปด้วยเสียงพยัญชนะต้นพยางค์, เสียงสระ และ เสียงวรรณยุกต์
2. พยางค์ที่ประกอบไปด้วยเสียงพยัญชนะต้นพยางค์, เสียงสระ, เสียงวรรณยุกต์ และเสียงพยัญชนะท้ายพยางค์

ดังนั้นการเลือกกลุ่มเสียงตัวอย่างที่จะใช้ในการทดสอบแบบจำลอง สามารถเลือกได้โดยการเลือกเสียงตัวอย่างจากเสียงทั้ง 4 ประเภท คือ เสียงพยัญชนะต้น, เสียงสระ, เสียงวรรณยุกต์ และเสียงพยัญชนะท้ายพยางค์ แล้วนำมาประกอบเป็นเสียงของคำพูดที่ใช้ในการทดสอบ โดยขึ้นอยู่กับแบบจำลองทั้งสองส่วนว่าจะต้องการทดสอบอะไรดังนี้

7.3.1 แบบจำลองส่วนที่ 1

สำหรับแบบจำลองแรกจะแบ่งกลุ่มเสียงตัวอย่างออกเป็น 2 กลุ่มคือ กลุ่มที่มีเสียงแตกต่างกัน และกลุ่มที่มีเสียงใกล้เคียงกัน โดยกลุ่มที่มีเสียงแตกต่างกันจะใช้คำ 10 คำของตัวเลขในภาษาไทยตั้งแต่ศูนย์ถึงเก้าเพื่อหาความถูกต้องในการรู้จำ ในกรณีที่เสียงพูดนั้นมาจากเสียงพยัญชนะและเสียงสระแตกต่างกันอย่างชัดเจน ส่วนกลุ่มที่สองจะใช้คำที่มีเสียงพื้นสามัญเสียงเดียวกันแต่มีการผันวรรณยุกต์ครบทั้งห้าเสียงคือ สามัญ เอก โท ตรี จัตวา เพื่อใช้ทดสอบแบบจำลองว่าจะมีความถูกต้องเพียงไร ถ้าใช้แบบจำลองเดิมในการรู้จำเสียงที่มีการผันวรรณยุกต์แตกต่างกันทั้ง 5 เสียงวรรณยุกต์ ซึ่งคำต้นแบบที่ใช้คือคำว่า ‘โบ (bo)’ และ ‘คา (car)’ ร่วมกับการผันวรรณยุกต์รวมเป็น 10 เสียง คือ โบ โป โป้ โป๊ โป๋ และ คา ข่า ค่า(ข้า) ค้า ขา โดยทั้ง 2 กลุ่มเสียงทดสอบพูดเสียงละ 5 ครั้งซึ่งได้ผลดังตารางที่ 7.1 และ 7.2

จะเห็นได้ว่าคำที่มีเสียงพื้นต่างกันความสามารถในการรู้จำของแบบจำลองจะมีค่าสูง แต่ถ้าคำที่ใช้ทดสอบมีเสียงพื้นเสียงเดียวกันจะให้ค่าความถูกต้องลดลง เนื่องจากค่าเซปตรัมของคำที่มีเสียงพื้นเสียงเดียวกัน (ต่างกันที่วรรณยุกต์) มีค่าใกล้เคียงกันมากดังรูปที่ 7.4 และ 7.5 การรู้จำที่ผิดพลาดของแบบจำลองนี้ส่วนใหญ่จะอยู่ในกลุ่มเสียงพื้นเดียวกัน ดังนั้นจึงยังอาจใช้แบบจำลองส่วนที่ 1 นี้ในการรู้จำคำที่มีเสียงพื้นต่างกันได้

ตารางที่ 7.1 แสดงผลการทดสอบคำพูดของเสียงตัวเลขศูนย์ถึงเก้า

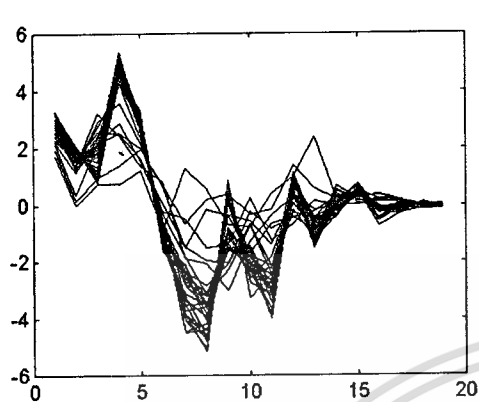
เสียงต้นแบบ	0	1	2	3	4	5	6	7	8	9
ทดสอบครั้งที่ 1	0	1	2	3	4	5	6	7	8	2
ทดสอบครั้งที่ 2	0	1	2	2	4	2	6	7	8	4
ทดสอบครั้งที่ 3	0	1	2	3	4	5	6	7	8	9
ทดสอบครั้งที่ 4	0	1	2	3	4	5	6	7	8	9
ทดสอบครั้งที่ 5	0	1	2	4	4	5	6	7	8	3
ถูกเฉลี่ย (%)	100	100	100	60	100	80	100	100	100	40
เฉลี่ยรวม	88									

ตารางที่ 7.2 แสดงผลการทดสอบคำพูดของเสียงที่มีเสียงพื้นสามัญเสียงเดียวกัน

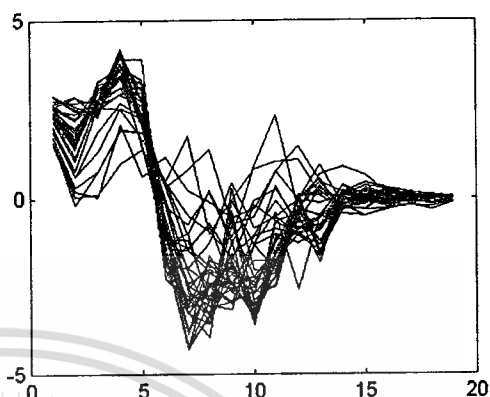
สัญลักษณ์	1	2	3	4	5	6	7	8	9	10
เสียงต้นแบบ	โบ	โป	ป๋	ป๊	ป๋	กา	ข่า	ค่า(ข้า)	ค้ำ	ขา
ทดสอบครั้งที่ 1	3	2	3	4	2	6	7	8	9	10
ทดสอบครั้งที่ 2	3	2	3	3	5	6	7	9	7	10
ทดสอบครั้งที่ 3	3	2	3	4	3	6	7	8	9	7
ทดสอบครั้งที่ 4	3	2	3	4	3	6	8	8	9	10
ทดสอบครั้งที่ 5	3	2	3	4	5	6	7	8	9	7
ถูกเฉลี่ย (%)	0	100	100	80	40	100	80	80	80	60
เฉลี่ยรวม (%)	72									
ถูกเฉลี่ยต่อคำพูดที่มีเสียงพื้นเดียวกัน (%)	100					100				

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

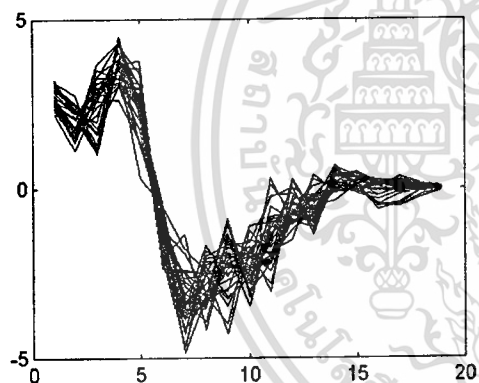
รูปที่ 7.4 ค่าเซปตริ่มของค่า ก) โบ ข) โป ค) โป้ ง) โป๊ จ) โป้



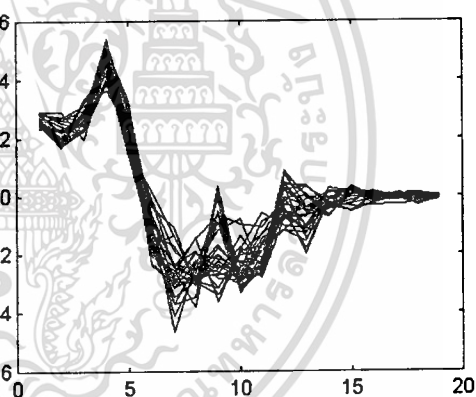
ก)



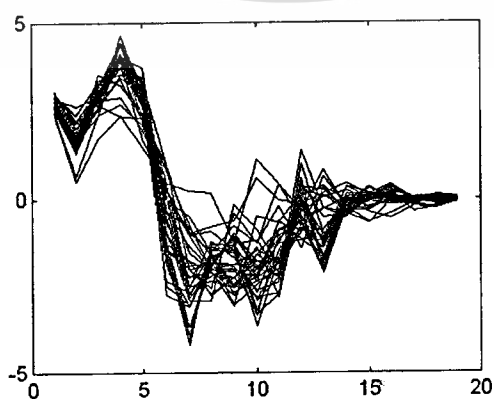
ข)



ค)



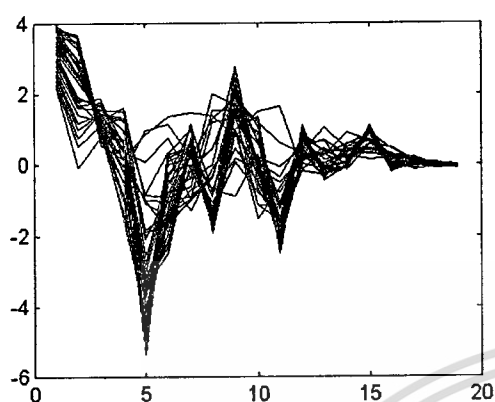
ง)



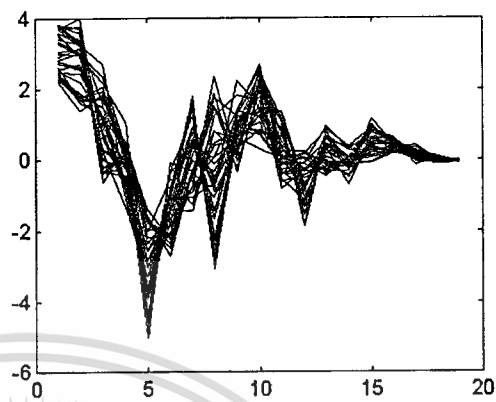
จ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

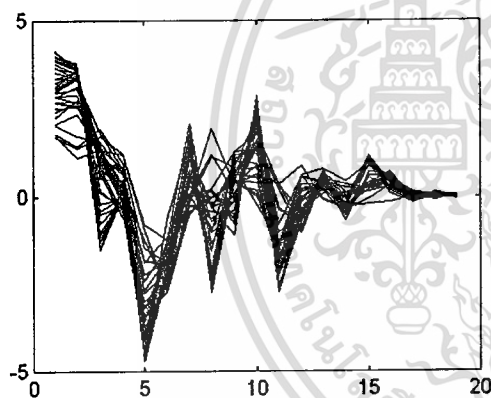
รูปที่ 7.5 ค่าเซปตรัมของค่า ก) คา ข) ข่า ค) ค่า-ข่า ง) ค้า จ) ขา



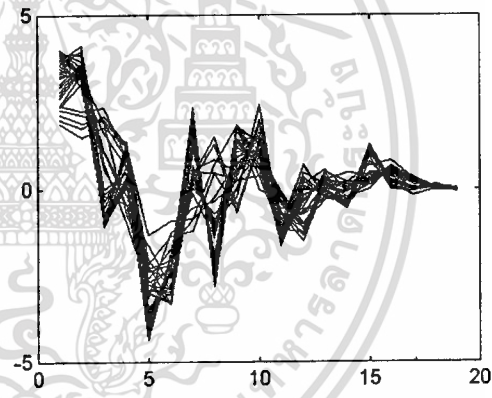
ก)



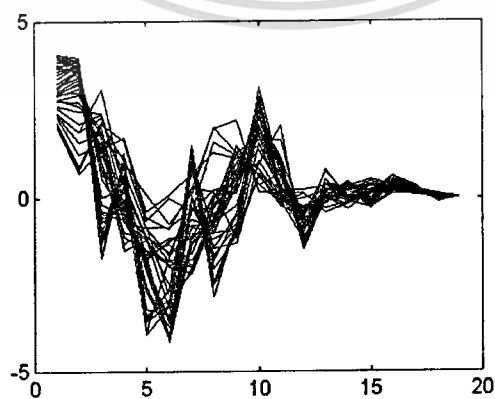
ข)



ค)



ง)



จ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ตัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 7.3 ผลการทดสอบการรู้จำของเสียงที่ใช้ในการฝึกจำของแบบจำลอง HMM

	ผลลัพธ์จากการรู้จำ				
สัญลักษณ์	1	2	3	4	5
เสียงพูด	สามัญ	เอก	โท	ตรี	จัตวา
โบ (bo)	1	2	3	4	5
กา (car)	1	2	3	4	5
เจียง (cheng)	1	2	3	4	5
ดาว (dow)	1	2	3	4	5
อิง (eng)	3	3	3	3	4
ฮัว (hua)	1	2	3	4	5
จา (ja)	1	2	3	4	5
คอ (kor)	1	2	3	4	5
นู (nu)	1	2	3	4	4
น่อง (nong)	1	2	3	4	5
แพ (pair)	1	2	3	4	5
ไช (chai)	2	2	3	4	5
ทีน (teen)	1	2	3	4	4
ทาย (try)	2	2	3	4	5
แยม (yam)	5	2	3	4	5
ถูกเฉลี่ย (%)	73.3	93.3	100	93.3	80
เฉลี่ยรวม (%)	87.9				

ตารางที่ 7.4 ผลการทดสอบการรู้จำของเสียงที่ไม่เคยใช้ในการฝึกจำของแบบจำลอง HMM

	ผลลัพธ์จากการรู้จำ				
สัญลักษณ์	1	2	3	4	5
เสียงพูด	สามัญ	เอก	โท	ตรี	จัตวา
อา (ar)	1	2	3	4	5
ฮาง (hang)	1	2	3	4	5
จิม (jim)	1	2	3	4	4
นา (na)	1	2	3	4	5
โอ (oh)	4	1	3	3	5
จาบ (jarb)	1	2	3	4	5
มู (muu)	1	2	3	4	5
อิง (eng)	3	2	3	4	5
เปา (pao)	1	2	3	4	5
มวย (muay)	1	2	3	3	5
ถูกเฉลี่ย (%)	80	90	100	80	90
เฉลี่ยรวม (%)	88				

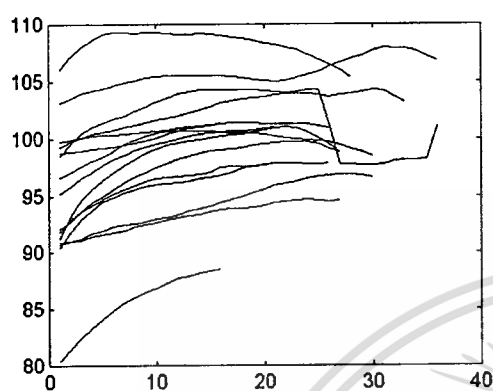
ตารางที่ 7.5 ผลสรุปการทดสอบการรู้จำของเสียงทดสอบกับแบบจำลอง HMM ของเสียงวรรณยุกต์

เสียงวรรณยุกต์	เสียงที่เคยใช้ฝึก แบบจำลอง	เสียงที่ไม่เคยใช้ ฝึกแบบจำลอง	ถูก เฉลี่ย
สามัญ	73.33	80	76.67
เอก	93.33	90	91.67
โท	100	100	100
ตรี	93.33	80	91.67
จัตวา	80	90	85
ถูกเฉลี่ย	87.9	88	87.95

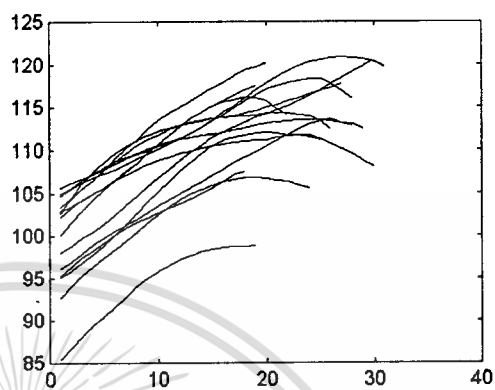
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูปที่ 7.6 ค่าพิชช์ของคำในตารางที่ 7.3 ก) เสียงสามัญ ข) เสียงเอก

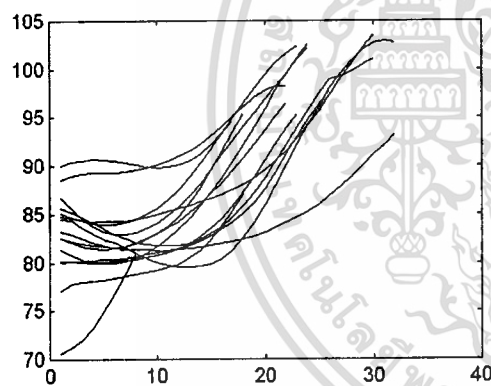
ค) เสียง โท ง) เสียงตรี จ) เสียงจัตวา



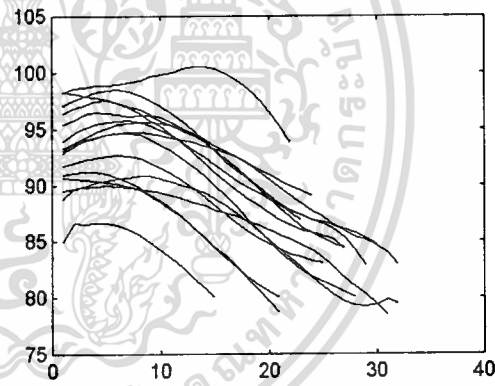
ก)



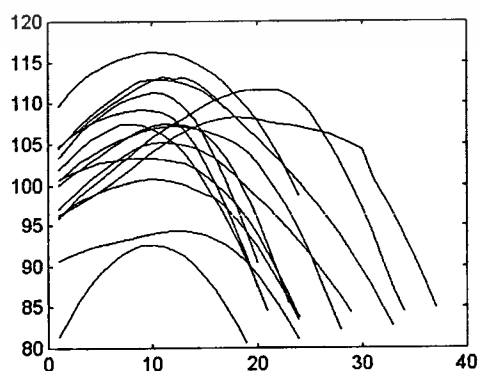
ข)



ค)



ง)



จ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

7.3.2 แบบจำลองส่วนที่ 2 (ส่วนที่ใช้ในการรู้จำเสียงวรรณยุกต์)

เพื่อให้การทดลองรู้จำเสียงวรรณยุกต์ครอบคลุมเสียงทั้งหมดในภาษาไทย จึงเลือกตัวอย่างจากเสียงทั้ง 3 ชนิดคือ เสียงพยัญชนะ(ต้น และ ท้ายพยางค์) เสียงสระ และเสียงวรรณยุกต์ เพื่อให้จำนวนรูปแบบการผสมเสียง (combination) ที่เป็นไปได้ทั้งหมดต่อเสียงวรรณยุกต์หนึ่งเสียงเหมาะสมต่อการทดลอง จึงคัดเลือกเสียงจากกลุ่มตัวอย่างให้กระจายครอบคลุมเสียงประเภทต่างๆ ในเสียงแต่ละชนิดมาเป็นจำนวน 15 คำพูด และเมื่อผสมรวมกับเสียงวรรณยุกต์ทั้ง 5 เสียง จะเกิดเป็นกลุ่มตัวอย่างจำนวน 75 เสียง ที่นำมาใช้ในการเทรน (training) แบบจำลอง ส่วนการทดสอบการรู้จำเสียงวรรณยุกต์ของแบบจำลองนั้น จะใช้เสียงทั้งที่ใช้ในการฝึกจำ (15 x 5 เสียง) และเสียงพูดที่ไม่เคยมีการฝึกจำมาก่อน (10 x 5 เสียง) ให้ผลดังตาราง 7.3 (สัญลักษณ์ 1-5 แทนเสียง สามัญ, เอก, โท, ตรี และ จัตวา ตามลำดับ)

จากรูปที่ 7.6 จะเห็นได้ว่าค่าพิชิตซ์ของเสียงวรรณยุกต์ทั้ง 5 เสียง มีลักษณะที่ค่อนข้างจะแตกต่างกันอย่างชัดเจน จึงทำให้ค่าความถูกต้องในการรู้จำของเสียงวรรณยุกต์มีค่าค่อนข้างสูง

7.3.3 แบบจำลองส่วนที่ 1 และส่วนที่ 2 รวมกัน

เมื่อนำแบบจำลองส่วนที่ 1 (ส่วนที่ใช้ในการรู้จำเสียงพูดโดยทั่วไป) และส่วนที่ 2 (ส่วนที่ใช้ในการรู้จำเสียงวรรณยุกต์) มารวมกันเพื่อใช้ในการรู้จำความหมายของเสียงพูด โดยใช้แบบจำลองส่วนที่ 1 ในการรู้จำเสียงพื้นสามัญ(เสียงพยัญชนะและเสียงสระ) ซึ่งให้ผลการรู้จำในตารางที่ 7.6 และใช้แบบจำลองที่ 2 ในการรู้จำเสียงวรรณยุกต์ ซึ่งให้ผลดังตารางที่ 7.3 และนำผลจากการทดสอบของแบบจำลองทั้งสองมาแปลความหมายใหม่ จะทำให้ระบบแบบจำลองสามารถรู้จำเสียงได้ถูกต้องมากกว่าการใช้เพียงแบบจำลองส่วนที่ 1 เพียงส่วนเดียว (ให้ผลการรู้จำตามตารางที่ 7.2) เพราะเสียงที่มีพื้นเสียงเดียวกันแต่มีเสียงวรรณยุกต์ต่างกันค่าเซปตริมจะมีความใกล้เคียงกันมาก (ดังตัวอย่างในรูป 7.3 และ 7.4) ทำให้ความสามารถในการแยกแยะความแตกต่างของคำที่มีวรรณยุกต์ต่างกันจากค่าเซปตริมทำได้ยากขึ้น

ดังนั้นในการรู้จำเสียงภาษาไทยจึงควรจะต้องใช้แบบจำลองทั้ง 2 ส่วนเพื่อให้สามารถแยกความแตกต่างของโทนเสียงในภาษาไทยได้ชัดเจนขึ้นและให้ค่าการรู้จำถูกต้องมากยิ่งขึ้น

ตารางที่ 7.6 ผลการทดสอบการรู้จำของคำที่มีพื้นเสียงเดียวกัน

พื้นเสียง	เสียงวรรณยุกต์					ถูกเฉลี่ย (%)
	สามัญ	เอก	โท	ตรี	จัตวา	
โบ (bo) = 1	1	1	1	1	1	100
กา (car) = 2	2	2	2	2	2	100
เซียง (cheng) = 3	3	3	3	3	3	100
ดาว (dow) = 4	4	4	4	4	4	100
อิง (eng) = 5	5	5	5	5	5	100
ฮัว (hua) = 6	6	6	6	6	6	100
จา (ja) = 7	7	7	7	7	7	100
กอ (kor) = 8	8	8	8	8	8	100
นู (nu) = 9	9	9	9	9	9	100
นอง (nong) = 10	10	10	10	10	10	100
แพ (pair) = 11	11	11	11	11	11	100
ไช (chai) = 12	12	12	12	12	12	100
ทีน (teen) = 13	13	13	13	13	13	100
ทาย (try) = 14	14	14	14	14	14	100
แยม (yam) = 15	15	15	15	15	15	100

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 8

สรุปผลและข้อเสนอแนะในการวิจัย

เนื่องจากระบบแบบจำลองที่ใช้ในการรู้จำเสียงที่ใช้ค่าสเปกตรัมหรือตัวแทนในโดเมนความถี่ เช่น ค่าสัมประสิทธิ์ LPC เป็นหลักในการแยกแยะความแตกต่างของเสียงแต่ละเสียง จะให้ความคลาดเคลื่อนที่ค่อนข้างสูงเมื่อใช้กับภาษาที่มีเสียงวรรณยุกต์ (tonal language) เช่น ภาษาไทย เพราะการเปลี่ยนแปลงของคาบพิทช์ของเสียงที่มีวรรณยุกต์ต่างกัน เมื่อพิจารณาในโดเมนความถี่ จะเห็นความแตกต่างของความถี่พิทช์นี้น้อยมาก และเมื่อนำไปเป็นข้อมูลในการสร้างแบบจำลองในการรู้จำ จึงทำให้แบบจำลองของเสียงที่มีพื้นเสียงเดียวกันแต่วรรณยุกต์ต่างกันมีลักษณะใกล้เคียงกันมาก ทำให้การรู้จำเสียงพื้นที่มีวรรณยุกต์ต่างกันเกิดความคลาดเคลื่อนได้ง่าย ดังนั้นการสร้างแบบจำลองเพื่อหาความแตกต่างของการเปลี่ยนแปลงคาบพิทช์ของเสียงที่มีวรรณยุกต์ต่างกันเพื่อรู้จำเสียงวรรณยุกต์โดยตรง ร่วมกับแบบจำลองที่สร้างขึ้นจากข้อมูลในโดเมนความถี่จะช่วยให้การพิจารณารู้จำเสียงพูดในภาษาที่มีเสียงวรรณยุกต์ถูกต้องมากยิ่งขึ้น

เนื่องจากจำนวนเสียงที่เป็นไปได้โดยประมาณในหนึ่งพยางค์ของภาษาไทยมีถึง 22,680 ($21 \times 24 \times 5 \times 9$) เสียง ถ้าต้องการจะสร้างแบบจำลองของแต่ละเสียงที่มีทั้งหมดจะต้องใช้แบบจำลองที่ขึ้นต่อกัน (dependent sound-type model) ของเสียงแต่ละประเภทจำนวนมาก แต่ถ้าจะสร้างแบบจำลองของเสียงทั้ง 4 ชนิด (เสียงพยัญชนะต้น, เสียงสระ, เสียงวรรณยุกต์ และเสียงพยัญชนะตัวสะกด) ที่ไม่ขึ้นต่อกัน (independent sound-type model) คือ เสียงพยัญชนะต้น 21 แบบจำลอง, เสียงสระ 24 แบบจำลอง, เสียงวรรณยุกต์ 5 แบบจำลอง และเสียงพยัญชนะตัวสะกด 9 แบบจำลอง ก็จะทำให้มีแบบจำลองทั้งหมดจำนวนเพียง 59 ($21 + 24 + 5 + 9$) แบบจำลองเท่านั้น โดยในการจดจำความหมายในภาษาไทยอาจจะใช้แบบจำลองที่แยกจากกันในแต่ละประเภท เพื่อหาองค์ประกอบของคำซึ่งจะได้ผลเป็นเสมือนคำที่เขียนแบบเสียงตัวสะกด (เช่น คำว่า วรรณยุกต์ สามารถเขียนแบบเสียงตัวสะกดได้เป็น วัน-นะ-ยุก) และใช้วิธีนี้ร่วมกับการวิเคราะห์ไวยากรณ์ในภาษาไทย เพื่อให้เกิดเป็นคำที่ถูกต้องตามหลักการเขียนของภาษาไทย, ก็จะทำให้สามารถทำให้เครื่องคอมพิวเตอร์สามารถที่จะแปลความหมายของเสียงพูดออกเป็นคำในภาษาไทยได้ โดยใช้แบบจำลองของเสียงจำนวนน้อยซึ่งจะมีผลทำให้การประมวลผลการรู้จำเสียงสามารถกระทำได้รวดเร็ว

การรู้จำโดยวิธีการแยกองค์ประกอบของเสียงทั้ง 4 ชนิดที่ไม่ขึ้นต่อกัน (เสียงพยัญชนะต้น, เสียงสระ, เสียงวรรณยุกต์ และเสียงพยัญชนะตัวสะกด) ในปัจจุบันนี้มีผู้ทำการวิจัยการรู้จำเสียงคำไทย ประเภทเสียงวรรณยุกต์และเสียงสระ [12] แล้ว แต่ยังไม่มีการทำวิจัยในประเภทการรู้จำเสียงพยัญชนะ ดังนั้นการวิจัยการรู้จำเสียงพยัญชนะทั้งพยัญชนะต้นและพยัญชนะตัวสะกดจึงเป็นแนวทางหนึ่งที่สามารถจะทำการดำเนินการวิจัยต่อ รวมทั้งการวิจัยการรู้จำเสียงแบบเป็นคำต่อเนื่อง, การวิจัยการแก้คำผิดในภาษาไทย และการวิจัยการแก้คำผิดตามหลักไวยากรณ์ในภาษาไทย เมื่อนำมารวมกันก็จะทำให้เครื่องคอมพิวเตอร์หรือเครื่องจักรสามารถทำความเข้าใจเสียงคำพูดที่เป็นประโยคในภาษาไทยได้

วิทยานิพนธ์นี้เป็นการเสนอการจดจำเสียงวรรณยุกต์เฉพาะบุคคล ที่สามารถใช้ได้กับเสียงวรรณยุกต์สำหรับคำไทยได้เป็นอย่างดี โดยที่สามารถใช้ได้ดีกับทั้งคำที่ใช้และไม่ได้ใช้ในการเรียนรู้ (training) มาก่อนได้ ซึ่งวิธีการในวิทยานิพนธ์นี้สามารถประยุกต์ให้สามารถจดจำเสียงโดยไม่ขึ้นกับบุคคลได้ โดยการเพิ่มจำนวนเสียงของผู้พูดที่ใช้ในการสร้างแบบจำลองให้มาจากบุคคลต่างๆ เพื่อให้แบบจำลองเรียนรู้และสร้างแบบจำลองขึ้นเพื่อให้สามารถรู้จำเสียงโดยไม่ขึ้นกับผู้พูด แต่ควรจะต้องมีการศึกษาถึงผลกระทบอันเนื่องมาจากผู้พูดที่มีลักษณะโทนเสียงต่างกันมากๆ เพิ่มเติม เช่น ผู้ชาย, ผู้หญิง, เด็ก, ผู้ใหญ่ เป็นต้น

บรรณานุกรม

- [1] J.R. Deller, Jr., J.G. Proakis, and J.H.L Hansen, Discrete-Time Processing of Speech Signals, MacMillan, 1993
- [2] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993
- [3] L.R. Rabiner and others, A Comparative Performance Study of Several Pitch Detection Algorithms, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-24, October 1976
- [4] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi, An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, Bell System Technical Journal 62, April 1983
- [5] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi, On the application of Vector Quantization and Hidden Markov Model to speaker independent isolated word recognition, The Bell System Technical Journal 62, April 1983
- [6] The Student Edition of MATLAB version 4 User's Guide, Prentice Hall, 1995
- [7] ชวดี อิศรปริดา และคนอื่นๆ, การรู้จำเสียงพูด, ปริญญาานิพนธ์สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, พ.ศ. 2538
- [8] ผศ.ดร.อภิชาติ ตั้งทางธรรม, การเปลี่ยนความเร็วของเสียงพูด, การประชุมวิชาการทางวิศวกรรมไฟฟ้าครั้งที่ 17, 2537
- [9] ระพีพัฒน์ เพ็ญศิริม, รศ.ดร.สมชาย จิตะพันธ์กุล, การรู้จำเสียงตัวเลขภาษาไทยโดยไม่ขึ้นต่อผู้พูดโดยการใช้ไดนามิกไทม์วาร์ปปีง, การประชุมวิชาการทางวิศวกรรมไฟฟ้าครั้งที่ 18, พ.ศ. 2538
- [10] สุนทร อรอินทร์, อัฐ เครือฟัก, การประมวลเสียงพูดโดยการประมาณเชิงเส้น, ปริญญาานิพนธ์สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, พ.ศ. 2537
- [11] เสนีย์ วัฒวรรณ, หนังสือเรียนภาษาไทย ท101-ท102, สำนักพิมพ์วัฒนาพานิช, พ.ศ. 2533
- [12] เสาวลักษณ์ อารีย์พงศา, การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธี ฮิดเดน มาร์คอฟ โมเดล และเวกเตอร์ควอนไดซ์เซชัน, วิทยานิพนธ์จุฬาลงกรณ์มหาวิทยาลัย, พ.ศ. 2538

- [13] ชีระ ภัทรการนันท์, ดร.สมชาย จิตะพันธ์กุล, การรู้จำเสียงสระภาษาไทยใดๆไม่ขึ้นกับผู้พูด โดยการวัดสเปกตรัมดิสแตนซ์และใช้ไดนามิกไทม์วาร์ปปีง, การประชุมวิชาการทางวิศวกรรมไฟฟ้าครั้งที่ 18, พ.ศ. 2538
- [14] คณะอนุกรรมการปรับปรุงศัพท์เทคนิคทางวิศวกรรมไฟฟ้าในคณะกรรมการวิชาการสาขาวิศวกรรมไฟฟ้า พ.ศ. 2527-2534, ศัพท์เทคนิควิศวกรรมไฟฟ้าสื่อสาร, สมาคมวิศวกรรมสถานแห่งประเทศไทย ในพระบรมราชูปถัมภ์, มีนาคม 2539
- [15] คณะอนุกรรมการปรับปรุงศัพท์วิศวกรรมคอมพิวเตอร์และสื่อสารข้อมูลในคณะกรรมการวิชาการสาขาวิศวกรรมไฟฟ้า พ.ศ. 2535-2536, ศัพท์เทคนิควิศวกรรมคอมพิวเตอร์และสื่อสารข้อมูล, สมาคมวิศวกรรมสถานแห่งประเทศไทย ในพระบรมราชูปถัมภ์, พิมพ์ครั้งที่ 3, ตุลาคม 2540





เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การรู้จำเสียงวรรณยุกต์ของคำไทยเฉพาะบุคคล

Tonal Recognition for Speaker-Dependent Thai Word

ทิศเวช วีระวัฒน์

รศ.ดร. กอบชัย เคชหาญ

คณะวิศวกรรมศาสตร์ และ สำนักวิจัยการสื่อสารและเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

เนื่องจากเสียงพูดในภาษาไทยเป็นเสียงพูดที่มีระดับโทนเสียงสูงต่ำหรือภาษาเสียงวรรณยุกต์ (tonal language) ซึ่งแตกต่างจากภาษาอังกฤษที่มีเพียงแต่เพียงเสียงสระและพยัญชนะเท่านั้น เพื่อให้เครื่องคอมพิวเตอร์สามารถเข้าใจภาษาไทยได้ถูกต้องมากยิ่งขึ้นจึงจำเป็นต้องหาวิธีแยกความแตกต่างของเสียงวรรณยุกต์ทั้ง 5 เสียง คือ สามัญ เอก โท ตรี และ จัตวา ที่มีอยู่ในภาษาไทย ซึ่งวิธีการที่น่าเสนอในบทความนี้ใช้วิธีการตรวจสอบการเปลี่ยนแปลงของคาบเวลาพิตช์ (pitch period) ของเสียงพูดโดยวิธีออโตคอร์เรลชัน (Autocorrelation) ร่วมกับการปรับปรุงค่าพิตช์ และใช้แบบจำลอง HMM (Hidden Markov Model) ในการสร้างแบบจำลองของเสียงวรรณยุกต์ เพื่อใช้ในการรู้จำเสียงวรรณยุกต์ทั้ง 5 เสียง โดยผลที่ได้จากการทดสอบแสดงให้เห็นว่าแบบจำลองที่สร้างขึ้น สามารถรู้จำเสียงวรรณยุกต์ของทั้งคำที่ใช้ในการฝึกจำ (trained word) และของคำที่ไม่เคยได้ใช้ในการฝึกจำ (untrained word) ของแบบจำลองมาก่อนได้เป็นอย่างดี

Abstract

Thai language is a tonal language which differs from English language which phonetic sound derives from only consonants and vowels. To enhance the computer to capture Thai language more precisely, there must be a way to distinguish this 5 different tonal levels of Thai language. This article proposes the method of pitch period detection of speech by using autocorrelation combining with pitch enhancement technique and using Hidden Markov Model (HMM) to build a model for the recognition of tonal level differences. The results of the test of the model presented show that the machine can practically recognize the tonal level differences of both the trained word and the untrained word.

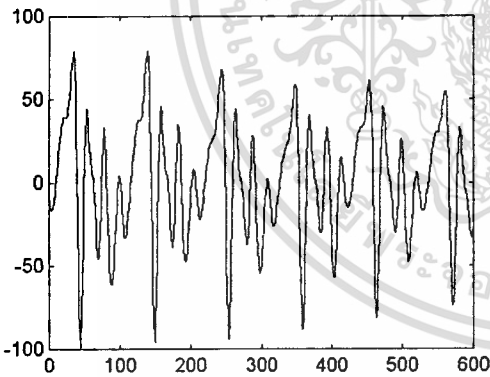
1. บทนำ

ในภาษาที่ไม่ใช่ภาษาเสียงวรรณยุกต์ คำแต่ละคำจะเปลี่ยนระดับเสียงไปอย่างไรก็ยังคงความหมายเดิม เช่น คำ คาร์ (car) ในภาษาอังกฤษ ถึงแม้จะออกเสียงเพี้ยนเป็น ข้า คำ (ข้า) คำ ขา ก็ยังคงความหมายว่า รถยนต์ เหมือนกันทุกคำ ซึ่งวิธี

การรู้จำเสียงพูดที่มีใช้กับภาษาอังกฤษไม่เหมาะที่จะใช้แยกความแตกต่างนี้ เช่น เสียงคำว่า โอ โอ โอ โอ โอ ในภาษาอังกฤษจะแทนเสียงทั้ง 5 เสียงเป็นคำว่า oh เพียงคำเดียว เป็นต้น

2. ลักษณะเสียงพูดในภาษาไทย

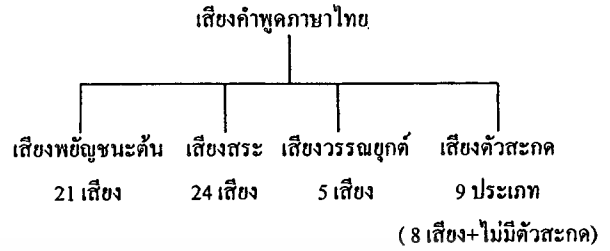
เสียงพูดเป็นสัญญาณที่มีคุณลักษณะที่ไม่คงที่ (Nonstationary Signal) มีการเปลี่ยนแปลงอย่างช้าๆ แต่อย่างไรก็ตามเมื่อพิจารณาเฉพาะส่วนย่อยของสัญญาณเสียงพูด นั่นคือในช่วงเวลาไม่เกิน 30 มิลลิวินาทีแล้ว สามารถอนุมานได้ว่า เสียงพูดเป็นสัญญาณที่มีคุณลักษณะคงที่ ดังนั้นสัญญาณเสียงพูดจึงถูกแบ่งเป็นส่วนย่อยๆ เรียกว่าเฟรม (frame) จากนั้นจึงทำการประมวลสัญญาณแต่ละเฟรม คุณลักษณะที่สำคัญอย่างหนึ่งของเสียงพูดคือ พิตช์ (pitch) ซึ่งเป็นความถี่พื้นฐานที่เกิดจากการสั่นในหลอดเสียง ดังตัวอย่างในรูปที่ 1 พิตช์เป็นตัวกำหนดระดับความสูงต่ำของเสียง เนื่องจากว่าภาษาไทยเป็นภาษาที่มีเสียงวรรณยุกต์ (tonal language) คือมีระดับโทนเสียง สามัญ เอก โท ตรี และจัตวา ซึ่งแต่ละโทนเสียงก็จะให้ความหมายที่แตกต่างกันออกไป ดังนั้นพิตช์จึงเป็นสิ่งที่สำคัญในการแยกแยะคำในภาษาไทย



รูปที่ 1 ตัวอย่างคาบเวลาพิตช์ของเสียงพูด

เสียงคำพูดต่างๆในภาษาไทยที่เปล่งออกมาหนึ่งพยางค์โดยทั่วไปจะประกอบไปด้วยการผสมเสียง 3 ชนิดคือ เสียงพยัญชนะ, เสียงสระ และเสียงวรรณยุกต์ [5] โดยมีรูปแบบที่เป็นไปได้ 2 รูปแบบคือ

1. เสียงพยัญชนะต้นพยางค์ + เสียงสระ + เสียงวรรณยุกต์
2. เสียงพยัญชนะต้นพยางค์ + เสียงสระ + เสียงวรรณยุกต์ + เสียงพยัญชนะท้ายพยางค์



รูปที่ 2 โครงสร้างของเสียงพูดภาษาไทย

เสียงพยัญชนะ : ในภาษาไทยมี 21 เสียง โดยสามารถแบ่งตามตำแหน่งได้เป็น

- เสียงพยัญชนะต้นพยางค์ (พยัญชนะต้น) มี 21 เสียงแบ่งเป็น

1. อักษรตัวเดียวโดดๆ คือ พยัญชนะ 44 ตัว
2. อักษรควบ คือ พยัญชนะสองตัว ตัวหลังเป็น ร ล หรือ ว เช่น กรวด คลุม ขวาง
3. อักษรนำ-อักษรตาม คือ พยัญชนะสองตัวที่ใช้เสียง ห เป็นเสียงนำในพยางค์ เช่น หงาย หนู จลาด แผนก อ๋าย อ๋ยาก

- เสียงพยัญชนะท้ายพยางค์ (ตัวสะกด) มี 9 ประเภท (8 เสียง) คือ

1. ไม่มีเสียงพยัญชนะสะกด หรือ แม่ ก-กา เช่น มือ หู หน้า ตา
2. เสียง ก หรือ แม่กก มี ก ข ค จ เช่น นก
3. เสียง ง หรือ แม่กง มี ง เช่น เก่ง จริง
4. เสียง ด หรือ แม่กด มี ด จ ช ซ ฎ ฏ ฐ ฑ ฒ ต ถ ท ธ ษ ศ ส เช่น พุด กิจ
5. เสียง น หรือ แม่กน มี น ญ ณ ร ล พ เช่น คน ชาญ
6. เสียง บ หรือ แม่กบ มี บ ป ฟ ภ เช่น กับ รูป
7. เสียง ม หรือ แม่กม มี ม เช่น ชาม เทียม
8. เสียง ย หรือ แม่เกย มี ย เช่น กาย เลย
9. เสียง ว หรือ แม่เกว มี ว เช่น กาว เขียว

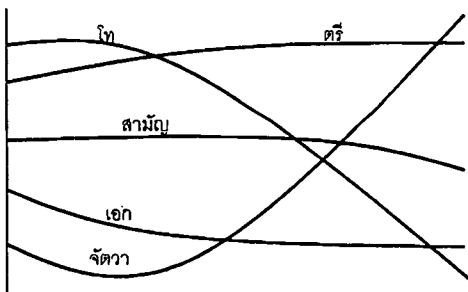
เสียงสระ : ในภาษาไทยมี 24 เสียงแบ่งตามช่วงเวลา
ได้เป็น 2 พวก คือ

1. สระเสียงสั้นใช้เวลาในการเปล่งเสียงน้อย มี 12 เสียง คือ อะ อี อื อุ เอะ แอะ โอะ เอาะ เออะ เอียะ เอือะ อัวะ
2. สระเสียงยาวใช้เวลาในการเปล่งเสียงนานกว่า มี 12 เสียงเช่นกันคือ อา อี อือ อุ เอ แอ โอ ออ เออ เอีย เอือ อัว

สระทั้งสองพวกนี้มีที่เกิดเสียงและใช้ไวยากรณ์กลมกล่อม
เกลาเสียงอย่างเดียวกัน ต่างกันที่ช่วงเวลาในการ
เปล่งเสียง

เสียงวรรณยุกต์ : ในภาษาไทยมี 5 ระดับดังนี้

1. เสียงสามัญ คือเสียงที่มีระดับปานกลางและมีระดับคงที่ตลอดไป เช่น กานอน ใน รัง
2. เสียงเอก คือเสียงที่เริ่มต้นระดับเดียวกับเสียงสามัญ แล้วลดระดับต่ำกว่าเสียงสามัญ เช่น เด็ก จะ กวาด เศษ
3. เสียงโท คือเสียงที่เริ่มต้นระดับสูงกว่าเสียงสามัญ แล้วตกลงต่ำกว่าเสียงเอก เช่น ที่ ว่า แยก ได้
4. เสียงตรี คือเสียงที่เริ่มต้นสูงกว่าเสียงสามัญ เล็กน้อยแล้วขึ้นไปเรื่อยๆ จนเสมอจุดเริ่มต้นของเสียงโทแล้วตกลงเล็กน้อย เช่น รู้ แล้ว นะ
5. เสียงจัตวา คือเสียงที่เริ่มต้นต่ำกว่าเสียงสามัญ แล้วสูงขึ้นไปเรื่อยๆ จนเสมอเสียงตรี เช่น หนู เห็น เขา ไหม



รูปที่ 3 ระดับเสียงวรรณยุกต์ [4]

(ระดับการเปลี่ยนแปลงของความถี่พิช)

ดังนั้นถ้าจะหาจำนวนเสียงทั้งหมดที่เป็น
ไปได้ในภาษาไทยในหนึ่งพยางค์จะมีความเป็นไปได้
ได้ในการผสมเสียงทั้งหมด $21 \times 24 \times 5 \times 9 = 22,680$ เสียงโดยประมาณ (ถ้าไม่คำนึงถึงการผสม
อักษรบางแบบที่ไม่สามารถออกเสียงได้ และเสียง
คำพูดที่ไม่มีเสียงออกมา เช่น การกระซิบหรือเสียง
ที่พูดโดยใช้ลมที่ไม่มีกระแสลมของเส้นเสียง)

3. การคำนวณหาพิช

3.1. การหาจุดเริ่มและจุดท้ายของเสียง

ในระบบการรู้จำเสียงแบบคำโคความ
ถูกต้องในการหาจุดเริ่มและจุดท้ายของเสียงมีความ
สำคัญอยู่ 2 ประการคือ ความคลาดเคลื่อนในการรู้จำ
และปริมาณข้อมูลที่ใช้ในการคำนวณ ถ้ากำหนดให้
สัญญาณเสียง $s(t)$ ถูกแปลงเป็นเชิงเลข (digitize)
เป็นลำดับของข้อมูลเสียง $x(m)$ โดยที่

$$\{ x[m] \} = \{ x[1], x[2], x[3], \dots, x[m], \dots, x[K] \}$$

จะสามารถหาค่าพลังงานของสัญญาณได้คือ

$$E_l(0) = \sum_{m=0}^{N-1} x^2(m), \quad l=0,1,\dots,L-1$$

และเพื่อลดการคำนวณสามารถใช้ขนาดของ
สัญญาณแทนค่ากำลังสอง ดังสมการ

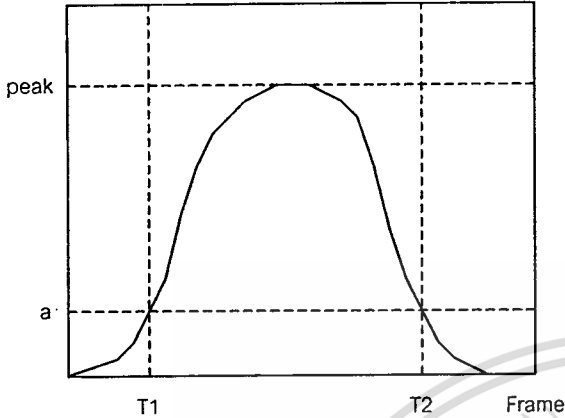
$$M_l(0) = \sum_{m=0}^{N-1} |x[m]|$$

โดยที่ $M_l(0)$, N , และ l คือ ค่าพลังงาน, ขนาดของ
เฟรม (frame size), และลำดับที่ของเฟรม หลังจาก
นั้นจะกำหนดระดับของพลังงาน (threshold) a เพื่อ
ใช้ในการหาจุดเริ่มและจุดท้ายของสัญญาณเสียง ดัง
แสดงในรูปที่ 4 โดยในที่นี้กำหนดให้ a มีค่า 10%
ของค่าของเฟรมข้อมูลที่มีพลังงานมากที่สุด ซึ่งก็
คือการตัดเฟรมที่มีพลังงานต่ำกว่า 10% ของเฟรมที่
มีพลังงานมากที่สุด

$$a = 0.1 M_{\max}$$

แล้วคำนวณหาค่า $T1$ และ $T2$ ใช้เป็นเฟรมเริ่มต้น
และเฟรมสุดท้ายของข้อมูลที่ใช้ในการคำนวณ [6]

Energy



รูปที่ 4 เฟรมเริ่มต้น($T1$) และเฟรมสุดท้าย($T2$)
ของพลังงานของเสียงพูดที่ใช้ในการคำนวณ

3.2. การคำนวณออโตคอร์รีเรชัน

เสียงวรรณยุกต์แต่ละเสียงที่เกิดขึ้นเกิด
จากการเปลี่ยนแปลงของพิตช์ ซึ่งการคำนวณหา
พิตช์นั้นทำได้หลายวิธี วิธีหนึ่งซึ่งให้ผลคือการใช้ออโตคอร์รีเรชัน หรือ วิธีอัตโนมัติสัมพันธ์
(Autocorrelation) ซึ่งสามารถเขียนได้ดังสมการ [4]

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

โดยที่ $x(m)$ เป็นสัญญาณอินพุต ถ้า $x(m)$ เป็น
สัญญาณที่มีคาบเวลาคงที่เท่ากับ P จะได้ว่า

$$\phi(k) = \phi(k+P)$$

นอกจากนี้ ออโตคอร์รีเรชันยังมีคุณสมบัติที่สำคัญ
คือ

$$\phi(k) = \phi(-k)$$

และจะมีค่าสูงสุดเมื่อ $k = 0$ นั่นคือ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$|\phi(k)| \leq |\phi(0)|$$

สำหรับสัญญาณที่มีคาบเวลา P ออโตคอร์รีเรชันจะมีค่าสูงสุดที่ตัวอย่าง $0, P, 2P, \dots$ ดังนั้นจึงสามารถคำนวณหาคาบเวลาของสัญญาณได้โดยการค้นหาตำแหน่งแรกที่ทำให้ออโตคอร์รีเรชันมีค่าสูงสุด

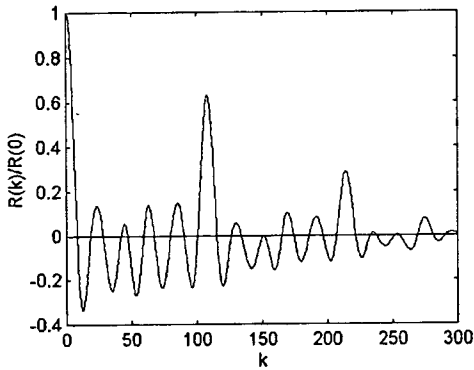
ในการประมวลสัญญาณเสียงพูดนั้นใช้ออโตคอร์รีเรชันสำหรับช่วงเวลาสั้นๆ โดยให้นิยามไว้ว่า

$$R_n(k) = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)][x(m+k)w(n-m-k)]$$

โดยที่ $w(n)$ เป็นฟังก์ชันหน้าต่าง ซึ่งอาจจะเป็นแบบสามเหลี่ยม สี่เหลี่ยม หรือแบบแฮมมิง ทำการคำนวณค่าของ $R_n(k)$ ในช่วงของ k จากน้อยที่สุดเท่าที่เป็นไปได้คือช่วงคาบเวลาของพิตช์ของเสียงผู้หญิงซึ่งมีค่าเท่ากับ 3 มิลลิวินาที จนถึงค่า k สูงที่สุดเท่าที่เป็นไปได้คือช่วงคาบเวลาพิตช์ของเสียงผู้ชายซึ่งมีค่าเท่ากับ 20 มิลลิวินาที ถ้าใช้ความถี่ในการแซมปลิง 10 kHz ก็จะใช้ k อยู่ในช่วง 30 ถึง 200 ถ้าเลือกให้ $w(n)$ มีความยาว N แซมเปิล โดยที่ N จะต้องมีค่ามากกว่า k มากๆ ($N \gg k$) ก็จะได้นิยามของออโตคอร์รีเรชันสำหรับช่วงเวลาสั้นๆ ดังนี้ [4]

$$R_n(k) = \sum_{m=0}^{N-1-k} [x(m)w(n-m)][x(m+k)w(n-m-k)]$$

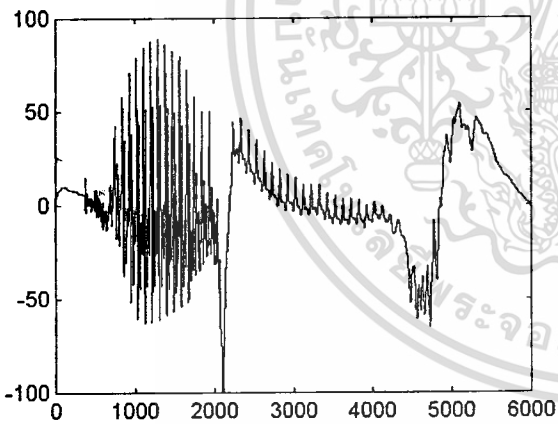
รูปที่ 4 แสดงให้เห็นถึงออโตคอร์รีเรชันแบบช่วงเวลาสั้นๆ ของสัญญาณเสียงพูด จะเห็นว่า มีค่าสูงสุดเมื่อ $k = 0$ และมีค่าสูงสุดอีกครั้งเมื่อ $k = 110$ ดังนั้นคาบเวลาของพิตช์จึงเป็น 110 แซมเปิล หรือเท่ากับ 11 มิลลิวินาทีเมื่ออัตราการซีกตัวอย่าง (sampling rate) คือ 10 kHz นั่นคือความถี่ของพิตช์เท่ากับ 91 Hz



รูปที่ 4 ออโตคอร์เรชันของสัญญาณเสียงพูด

3.3. การปรับแต่งคำพิตช์

เนื่องจากสัญญาณเสียงพูดแต่ละคำพูดจะมีความแตกต่างกันของคำพิตช์และความต่อเนื่องของพิตช์ โดยการเปลี่ยนแปลงของพิตช์จะขึ้นอยู่กับเสียงวรรณยุกต์ ส่วนความต่อเนื่องของพิตช์จะขึ้นอยู่กับองค์ประกอบของเสียงพยัญชนะเป็นหลัก ดังตัวอย่างในรูปที่ 5 ซึ่งจะเห็นได้ว่าข้อมูลบางช่วงจะไม่มีคาบพิตช์อยู่



รูปที่ 5 ตัวอย่างเสียงพูดของคำว่า 'จิม'

การคำนวณค่าออโตคอร์เรชันจากทุกเฟรมของสัญญาณเสียงพูดเพื่อใช้ในการหาคำพิตช์จะทำให้ได้ข้อมูลพิตช์ที่มีความผิดพลาดปะปนอยู่ อันเนื่องมาจากเสียงพยัญชนะดังกล่าว จึงทำให้ต้องมีการตัดข้อมูลพิตช์ที่ปะปนนี้ออก โดยใช้วิธีเทียบกับค่าเฉลี่ยพิตช์รวมของเสียงทั้งหมดที่ใช้เป็นต้นแบบในการฝึกจำ และค่ามัธยฐานของเสียงนั้นเฉลี่ยกัน ดังนี้

$$P_i = (\text{Mean_of_Pitch}_{\text{all words}} + \text{Median_of_Pitch}_i) / 2$$

$$k = 1$$

For j = 1 to length_of_Pitch

$$\text{If } |\text{Pitch}(j) - P_i| \leq 0.3 P_i$$

$$\text{NewPitch}(k) = \text{Pitch}(j)$$

$$k = k + 1$$

Endif

Next j

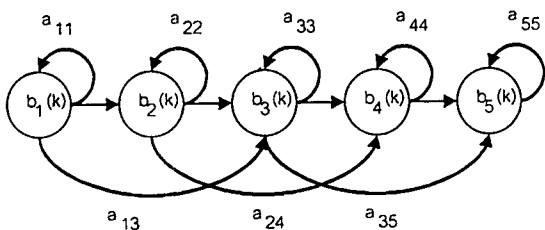
โดย P_i เป็นค่าเฉลี่ยของค่าเฉลี่ยรวมกับค่ามัธยฐานของเสียงที่ i ซึ่งพิตช์ที่มีขนาดมากกว่า 30% ของค่าเฉลี่ย P_i จะถูกตัดออก ซึ่งค่า 30% นี้ได้มาจากการทดลอง ทำให้คำพิตช์ใหม่ที่ได้นั้นมีความต่อเนื่องดีขึ้น การตัดคำพิตช์ออกในลักษณะนี้จะให้คำพิตช์ที่ได้จากคำพูดที่มีเสียงพยัญชนะผสมอยู่มีความเหมาะสมที่จะใช้ในการฝึกจำ ซึ่งจะทำให้แบบจำลองของเสียงวรรณยุกต์ที่ได้มีความสามารถในการรู้จำเสียงวรรณยุกต์ที่มีเสียงพยัญชนะต่างๆผสมอยู่ได้

4. แบบจำลองมาร์คอฟ

แบบจำลองมาร์คอฟ เป็นแบบจำลอง (model) ทางสถิติซึ่งพัฒนามาเพื่อแบ่งกลุ่มของอนุกรมทางเวลา หรือสัญญาณที่ไม่คงที่ นั่นคือ ใช้สำหรับจัดกลุ่มของสัญญาณที่ไม่รู้จัก (Unknown signal) ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่งของสัญญาณ ซึ่ง แบบจำลองมาร์คอฟ ได้ถูกนำมาประยุกต์ใช้ในการรู้จำเสียงพูด [7]

4.1. แบบจำลอง HMM (Hidden Markov Model)

รูปที่ 5 เป็นรูปที่แสดง HMM (แบบจำลองมาร์คอฟแบบซ่อนเร้น) แบบอนุกรมที่มี 5 สถานะ (state) การอธิบายระบบ ณ เวลาหนึ่งๆจะอธิบายด้วยเซตของกรณีสถานะ $\{1, 2, \dots, N\}$ โดยในแต่ละช่วงเวลาสามารถเปลี่ยนสถานะไปด้วยเซตของความน่าจะเป็นที่เกี่ยวข้องกับสถานะ



รูปที่ 5 ตัวอย่างแบบจำลองมาร์คอฟ 5 สเตต

จะแทนค่าเวลาที่เกี่ยวข้องกับการเปลี่ยนสเตตไป ด้วย $t = 1, 2, \dots, T$ และแทนสเตตที่เวลา t ด้วย q_t ในห่วงโซ่มาร์คอฟ (Markov Chain) นั้นจะแสดงความน่าจะเป็นในการย้ายสเตตไป โดยให้ความสนใจเพียงสเตตก่อนหน้านั้น เท่านั้น คือ

$P[q_t=j | q_{t-1}=i, q_{t-2}=k, \dots] = P[q_t=j | q_{t-1}=i]$ และ a_{ij} จะแทนความน่าจะเป็นของการย้ายสเตต จาก i ไป j

$$a_{ij} = P[q_t=j | q_{t-1}=i] \quad 1 \leq i, j \leq$$

โดยมีคุณสมบัติว่า

$$\begin{aligned} a_{ij} &\geq 0 & \forall j, i \\ \sum_{j=1}^N a_{ij} &= 1 & \forall i \end{aligned}$$

แบบจำลองนั้น เหตุการณ์จะเป็นฟังก์ชันของความน่าจะเป็นของสเตต คือ จะไม่ทราบเหตุการณ์ที่จะเกิดในแต่ละสเตตโดยตรง แต่จะรู้เหตุการณ์ได้โดยวิธีทางสถิติ ซึ่งจะทำได้ลำดับของเหตุการณ์ออกมา และเนื่องจากความคลาดเคลื่อนในการรู้จำไม่แปรผันโดยตรงกับจำนวนสเตตที่มากพอ [3] ในบทความนี้เลือกใช้ HMM ที่มีโครงสร้างแบบอนุกรมที่มี 6 สเตต เนื่องจากเป็นค่าที่มีความคลาดเคลื่อนเฉพาะบริเวณต่ำสุด (local minimum at $N = 6$ state)

4.2. ส่วนประกอบของแบบจำลอง HMM

T = ความยาวของค่าปรากฏ

N = จำนวนสเตตในแบบจำลอง

M = จำนวนค่าที่ปรากฏ

$Q = \{q_1, q_2, \dots, q_N\}$ สเตต

q_t = สเตตที่เวลา t

$V = \{v_1, v_2, \dots, v_M\}$ ค่าปรากฏ

$A = \{a_{ij}\}$ เมื่อ $a_{ij} = P[q_t=j | q_{t-1}=i] \quad (1 \leq i, j \leq N)$

คือความน่าจะเป็นของการเปลี่ยนสเตต

$B = \{b_j(k)\}$ โดย $b_j(k) = P[o_t=v_k | q_t=j]$

$(1 \leq k \leq m)$ คือ ความน่าจะเป็นของการเกิดค่าปรากฏในสเตต $j \quad (j=1, 2, \dots, N)$

$\pi = \{\pi_i\}$ เมื่อ $\pi_i = P[q_1=i]$ คือ ความน่าจะเป็นที่แต่ละสเตตจะเป็นสเตตเริ่มต้น

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

4.3. ปัญหาพื้นฐานของ แบบจำลอง HMM

ปัญหาที่ 1 : เมื่อกำหนดลำดับเหตุการณ์

$O = \{o_1, o_2, \dots, o_T\}$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะคำนวณหา $P(O, \lambda)$ ซึ่งเป็นความน่าจะเป็นของลำดับเหตุการณ์นี้ได้อย่างไร การแก้ปัญหานี้จะทำให้สามารถเลือกแบบจำลองที่สอดคล้องกับเหตุการณ์ที่ดีที่สุดได้จากแบบจำลองหลายอันที่มีอยู่ ซึ่งสามารถหาคำตอบได้โดยวิธีการคำนวณไปข้างหน้าและกลับหลัง (Forward & Backward Procedure)

ปัญหาที่ 2 : เมื่อกำหนดลำดับเหตุการณ์

$O = \{o_1, o_2, \dots, o_T\}$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะเลือกลำดับสเตต $q = \{q_1, q_2, \dots, q_T\}$ ที่เหมาะสมกับเหตุการณ์ที่สุดได้อย่างไร (อธิบายเหตุการณ์ได้ดีที่สุด) การแก้ปัญหานี้ช่วยหาลำดับของสเตตที่ถูกต้องสามารถหาคำตอบได้โดยการใช้วิธเทอริบีอัลกอริทึม (Viterbi Algorithm)

ปัญหาที่ 3 : จะปรับพารามิเตอร์ของแบบ

จำลอง $\lambda = (A, B, \pi)$ เพื่อให้ได้ค่าสูงสุดของ $P(O|\lambda)$ ได้อย่างไร การแก้ปัญหานี้ช่วยให้ได้ λ ที่เหมาะสมที่สุด โดยใช้เพื่ออธิบายการได้มาซึ่งลำดับเหตุการณ์ที่กำหนดให้อย่างดีที่สุด สำหรับการแก้ปัญหานี้สามารถทำได้โดยกระบวนการทำซ้ำ (iterative) โดยวิธีแบม-เวลช์ (Baum-Welch) หรือ EM (Expectation Maximization)

โดยในรายละเอียดของการแก้ปัญหานี้ 3

จะไม่ขอกล่าวในที่นี้ สามารถหาได้จากเอกสารอ้างอิง [1] และ [2]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า

ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3 ผลสรุปการทดสอบการรู้จำของเสียง
ทดสอบกับแบบจำลอง HMM ของเสียงวรรณยุกต์

เสียงวรรณยุกต์	ความถูกต้องในการรู้จำ (%)		
	เสียงที่ไม่เคยใช้ฝึกแบบจำลอง	เสียงที่เคยใช้ฝึกแบบจำลอง	เฉลี่ยรวม
สามัญ	73.33	80	76.67
เอก	93.33	90	91.67
โท	100	100	100
ตรี	93.33	80	91.67
จัตวา	80	90	85
เฉลี่ย	87.9	88	87.95

6. สรุป

เนื่องจากจำนวนเสียงที่เป็นไปได้โดยประมาณในหนึ่งพยางค์ของภาษาไทยมีถึง 22,680 ($21 \times 24 \times 5 \times 9$) เสียง ถ้าต้องการจะสร้างแบบจำลองของแต่ละเสียงที่มีทั้งหมดจะต้องใช้แบบจำลองที่ขึ้นต่อกัน (dependent sound-type model) ของเสียงแต่ละประเภทจำนวนมาก แต่ถ้าจะสร้างแบบจำลองของเสียงทั้ง 4 ชนิด (เสียงพยัญชนะต้น, เสียงสระ [8], เสียงวรรณยุกต์ และเสียงพยัญชนะตัวสะกด) ที่ไม่ขึ้นต่อกัน (independent sound-type model) คือ เสียงพยัญชนะต้น 21 แบบจำลอง, เสียงสระ 24 แบบจำลอง, เสียงวรรณยุกต์ 5 แบบจำลอง และเสียงพยัญชนะตัวสะกด 9 แบบจำลอง ก็จะทำให้มีแบบจำลองทั้งหมดจำนวนเพียง 59 ($21 + 24 + 5 + 9$) แบบจำลองเท่านั้น โดยในการรู้จำความหมายในภาษาไทยจะใช้แบบจำลองที่แยกจากกันในแต่ละประเภทนี้เพื่อหาค่าประกอบของคำซึ่งจะได้ผลเป็นเสมือนคำที่เขียนแบบเสียงตัวสะกด (เช่น คำว่า วรรณยุกต์ สามารถเขียนแบบเสียงตัวสะกดได้เป็น วัน-นะ-ยุก) และใช้วิธีนี้ร่วมกับการวิเคราะห์ไวยากรณ์ในภาษาไทยเพื่อให้เกิดเป็นคำที่ถูกต้องตามหลักการเขียนของภาษาไทย

บทความนี้เป็นการเล่นการเสนอการรู้จำเสียงวรรณยุกต์เฉพาะบุคคลที่สามารถใช้ได้กับเสียงคำวรรณยุกต์ที่ไม่ได้รับการฝึกมาก่อนได้เป็นอย่างดีซึ่งสามารถประยุกต์ให้สามารถรู้จำเสียงโดยไม่ขึ้นกับบุคคลได้ โดยการเพิ่มจำนวนเสียงของผู้พูดที่ใช้ในการสร้างแบบจำลองให้มากขึ้น

7. เอกสารอ้างอิง

- [1] J.R. Deller, Jr., J.G. Proakis, and J.H.L. Hansen, Discrete-Time Processing of Speech Signals, MacMillan, 1993
- [2] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993
- [3] เสาวลักษณ์ อริย์พงศา, การรู้จำเสียงพูดตัวเลขเป็นภาษาไทยแบบไม่ขึ้นกับผู้พูดโดยวิธีฮิดเดน มาร์คอฟ โมเดล และเวกเตอร์ควอนไทซ์ เซชัน, วิทยานิพนธ์ จุฬาลงกรณ์มหาวิทยาลัย, พ.ศ. 2538
- [4] ผศ.ดร.อภิชาติ ตั้งทางธรรม, การเปลี่ยนความเร็วของเสียงพูด, การประชุมวิชาการทางวิศวกรรมไฟฟ้าครั้งที่ 17, 2537
- [5] เสนีย์ วิลาวรรณ, หนังสือเรียนภาษาไทย ท101-ท102, สำนักพิมพ์วัฒนาพานิช, 2533
- [6] ระพีพัฒน์ เพ็ญศิริ, รศ.ดร.สมชาย จิตะพันธ์กุล, การรู้จำเสียงตัวเลขภาษาไทยโดยไม่ขึ้นต่อผู้พูดโดยการใช้ไดนามิกไทม์วอร์ปปีง, การประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 18, 2538
- [7] ชวดี อิศรปริดา และคนอื่นๆ, การรู้จำเสียงพูด, ปรินูญานิพนธ์สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง, พ.ศ. 2538
- [8] ชีระ ภัทธกรนันท์, ดร.สมชาย จิตะพันธ์กุล, การรู้จำเสียงสระภาษาไทยโดยดัดแปลงไม่ขึ้นกับผู้พูดโดยการใช้ไดนามิกไทม์วอร์ปปีง, การประชุมวิชาการทางวิศวกรรมไฟฟ้าครั้งที่ 18, พ.ศ. 2538

ประวัติผู้เขียน

นาย ทศเวท วีระวัฒน์ เกิดเมื่อวันที่ 28 กรกฎาคม 2513 ที่กรุงเทพมหานคร จบการศึกษาระดับปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมโทรคมนาคม ภาควิชาวิศวกรรมโทรคมนาคม คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (KMITL : King Mongkut Institute of Technology at Ladkrabang) ปี พ.ศ. 2534 และจบการศึกษาระดับปริญญาวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมอุตสาหกรรม ภาควิชาวิศวกรรมอุตสาหกรรมและการจัดการ สถาบันเทคโนโลยีแห่งเอเชีย (AIT : Asian Institute of Technology) ปี พ.ศ. 2536

