วิธีการแบ่งนับแบบสมสัณฐานสำหรับการระบุผู้พูด

นายศราวุธ จันทร์สด

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2546
ISBN 974-17-3673-6

# AN APPROACH OF ISOMORPHIC QUANTIZATION FOR SPEAKER IDENTIFICATION

Mr. Sarawoot Junsod

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2003

ISBN 974-17-3673-6

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | วิธีการแบ่งนับแบบสมสัณฐานสำหรับการระบุผู้พูด |
| โดย | นายศราวุธ จันทร์สด |
| สาขาวิชา | วิทยาศาสตร์คอมพิวเตอร์ |
| อาจารย์ที่ปรึกษา | อาจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ |

---

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดี คณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ)

.....กรรมการ
(อาจารย์ ดร.วิษณุ โคตรจรัส)

ศราวุธ จันทร์สด : วิธีการแบ่งนับแบบสมสัณฐานสำหรับการระบุผู้พูด. (AN APPROACH OF ISOMORPHIC QUANTIZATION FOR SPEAKER IDENTIFICATION) อ. ที่ปรึกษา: ดร.อรรถสิทธิ์ สุรฤกษ์, 47 หน้า. ISBN 974-17-3673-6.

การแบ่งนับแบบสมสัณฐานเป็นวิธีการลดปริมาณข้อมูลเวกเตอร์ลักษณะเฉพาะซึ่งได้จากการสกัดจากเสียงพูด โดยพิจารณาความคล้ายกันของรูปแบบเวกเตอร์ วิธีการนี้วางอยู่บนหลักการของการสร้างฟังก์ชันวัดการเปลี่ยนแปลงค่าภายในเวกเตอร์ในแต่ละมิติเพื่อให้ได้เวกเตอร์ใหม่ในรูปของเลขฐานสอง จากนั้นเวกเตอร์ใหม่ที่ได้จะถูกนำมาทำการแบ่งส่วนตามมิติและถูกจัดกลุ่มตามความเหมือนของเวกเตอร์ในกลุ่มนั้น เวกเตอร์ที่ซ้ำกันมากที่สุดหนึ่งชุดจะถูกนำมาเป็นตัวแทนของเวกเตอร์ทั้งหมดและถูกเก็บเป็นตัวแบบผู้พูด จากผลการทดลองแสดงให้เห็นว่าวิธีการนี้สามารถให้ความถูกต้องเฉลี่ยในการระบุผู้พูดมากถึงร้อยละ 99.73 เมื่อทดสอบกับเสียงพูดต่อเนื่องความยาว 5 ถึง 8 วินาที นอกจากนั้นเรายังทำการเปรียบเทียบประสิทธิภาพกับวิธีการแบ่งนับแบบเวกเตอร์และวิธีการแบ่งนับแบบฐานสองด้วย



ภาควิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา 2546

## 4470559521: MAJOR COMPUTER SCIENCE

KEY WORD: SPEAKER IDENTIFICATION / FEATURE VECTOR / VECTOR QUANTIZATION / BINARY QUANTIZATION / CODEBOOK

SARAWOOT JUNSOD: AN APPROACH OF ISOMORPHIC QUANTIZATION FOR SPEAKER IDENTIFICATION. THESIS ADVISOR: ATHASIT SURARERKS, Ph.D., 47 pp. ISBN 974-17-3673-6.

Isomorphic quantization is a method for reducing amount of feature vectors by determining their similarity forms. The feature vectors are extracted from speech. This method is based on a function that measures internal changing of feature vectors to produce binary vectors. The binary vectors are partitioned and then clustered the same vectors into groups. A Set of groups that have maximum frequency is chosen to generate a codebook instead of using all binary vectors. Experimental results show the effective accuracy in speaker identification, especially in continuous speech length 5-8 seconds, the average accuracy is 99.73%. We also investigate its performance by comparing with vector quantization and binary quantization methods.



Department Computer Engineering Student's signature.....
Field of study Computer Science Advisor's signature.....
Academic year 2003

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์และความช่วยเหลืออย่างยิ่งจาก อาจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้แง่คิด ความรู้ และคำแนะนำต่างๆ ซึ่งเป็นประโยชน์อย่างยิ่งต่องานวิจัย เพื่อนๆ สมาชิก SE LAB ผู้ที่ให้คำแนะนำเพิ่มเติม จึงขอขอบพระคุณทุกท่านเป็นอย่างสูงมา ณ ที่นี้ด้วย

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา มารดา ครูอาจารย์ทุกท่าน ที่ได้อบรมสั่งสอน เป็นกำลังใจ และสนับสนุนผู้วิจัยเรื่อยมาจนสำเร็จการศึกษา

# สารบัญ

หน้า


บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฌ
สารบัญภาพ ..... ญ
บทที่
1 บทนำ ..... 1
1.1 ความสำคัญและความเป็นมาของปัญหา ..... 1
1.2 วัตถุประสงค์ของงานวิจัย ..... 2
1.3 ขอบเขตของงานวิจัย ..... 3
1.4 ขั้นตอนและวิธีดำเนินงานวิจัย ..... 3
1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย ..... 5
2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ..... 6
2.1 ทฤษฎีที่เกี่ยวข้อง ..... 6
2.1.1 การได้มาซึ่งข้อมูลเสียงพูด ..... 6
2.1.2 ขั้นตอนการระบุผู้พูด ..... 7
2.1.3 การวิเคราะห์สัญญาณเสียงพูด ..... 8
2.1.4 การสร้างตัวแบบผู้พูดและการระบุผู้พูด ..... 14
2.2 ทฤษฎีการแบ่งนับเวกเตอร์ ..... 15
2.3 ทฤษฎีการแบ่งนับแบบฐานสอง ..... 16
2.4 งานวิจัยที่เกี่ยวข้อง ..... 17
3 การระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน ..... 19
3.1 บทกล่าวนำ ..... 19
3.2 การวัดความเหมือน ..... 20
3.2.1 ฟังก์ชันวัดการเปลี่ยนแปลงความชัน ..... 20
3.2.2 ฟังก์ชันวัดการเปลี่ยนแปลงความสูง ..... 21
3.2.3 ไอโซมอร์ฟิคฟังก์ชัน ..... 22
3.3 การแบ่งส่วนและการจัดกลุ่ม ..... 24

# สารบัญ (ต่อ)


บทที่ หน้า

3.4 การสร้างตัวแบบผู้พูดและการระบุผู้พูด ..... 25

3.4.1 โครงสร้างระบบระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน ..... 25

3.4.2 การสร้างตัวแบบผู้พูด ..... 26

3.4.3 การระบุผู้พูด ..... 27

4 การทดลองและผลการทดลอง ..... 29

4.1 การออกแบบฐานข้อมูลและการวัดประสิทธิภาพ ..... 29

4.2 ขั้นตอนและวิธีการทดสอบ ..... 30

4.2.1 ผลกระทบจากขนาดของการแบ่งส่วน ..... 30

4.2.2 ผลกระทบจากอันดับของตัวแบบผู้พูด (ขนาดตัวแบบผู้พูด) ..... 31

4.2.3 ผลกระทบจากอันดับข้อมูลทดสอบ ..... 32

4.2.4 การเปรียบเทียบกับวิธีการแบ่งนับวิธีการอื่น .... ..... 33

5 สรุปการวิจัยและข้อเสนอแนะ . ..... 36

5.1 สรุป ..... 37

5.2 ข้อเสนอแนะ ..... 38

รายการอ้างอิง ..... 39

ภาคผนวก ..... 41

ภาคผนวก ก รูปแบบของไฟล์เสียงพูดนามสกุล .wav ..... 42

ภาคผนวก ข ตัวอย่างข้อมูลทดสอบ ..... 43

ภาคผนวก ค ตัวอย่างการกำจัดเสียงเงียบและดีซีออฟเซ็ต (DC Offset) ..... 44

ภาคผนวก ง การกำหนดค่าพารามิเตอร์สำหรับการสกัดลักษณะเฉพาะ ..... 45

ภาคผนวก จ ตัวอย่างไอโซมอร์ฟิคฟังก์ชัน เขียนโดยใช้ MATLAB..... 46

ประวัติผู้เขียนวิทยานิพนธ์ ..... 47

# สารบัญตาราง

ตารางที่ หน้า

4.1 แสดงความถูกต้องในการระบุผู้พูดในแต่ละวิธี ..... 33

4.2 แสดงขนาดของตัวแบบผู้พูดโดยประมาณของแต่ละวิธี
ด้วยข้อมูลชุดที่ 1 ถึง ชุดที่ 4 ..... 34

4.3 แสดงความเร็วในการสร้างตัวแบบผู้พูดในรูปของสัญกรณ์เชิงปริมาณ (บิกโอ) ..... 34

4.4 แสดงความเร็วในการเปรียบเทียบตัวแบบผู้พูดในรูปของสัญกรณ์เชิงปริมาณ (บิกโอ) .... 35

# สารบัญภาพ

รูปที่ หน้า
รูปที่ 2.1 ขั้นตอนการระบุผู้พูด ..... 6
รูปที่ 2.2 ขั้นตอนการสกัดลักษณะเฉพาะด้วยวิธีการวิเคราะห์เซ็พสตรัล ..... 8
รูปที่ 2.3 สัญญาณที่ได้จากการคูณด้วยวินโดว์สี่เหลี่ยมมุมฉาก ..... 11
รูปที่ 2.4 แฮมมิ่งวินโดว์ ..... 12
รูปที่ 2.5 แถบความถี่แบบเมล ..... 14
รูปที่ 3.1 แสดงการวาดกราฟของเวกเตอร์ลักษณะเฉพาะหนึ่งตัว ..... 21
รูปที่ 3.2 แสดงรูปแบบของเวกเตอร์ลักษณะเฉพาะ
ที่ไอโซมอร์ฟิคฟังก์ชันสามารถจำแนกได้ ..... 23
รูปที่ 3.3 การแบ่งนับโดยใช้ไอโซมอร์ฟิคฟังก์ชันในแต่ละลักษณะเฉพาะ ..... 23
รูปที่ 3.4 การแบ่งเวกเตอร์ฐานสองออกเป็น $P$ ส่วน ..... 24
รูปที่ 3.5 แสดงการจัดกลุ่มของเวกเตอร์ฐานสองในแต่ละส่วน (Partition) ..... 24
รูปที่ 3.6 โครงสร้างของการระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน ..... 25
รูปที่ 3.7 การสร้างตัวแบบผู้พูดย่อย $Cb_m$ จากเสียงพูด 5 ประโยค..... 27
รูปที่ 3.8 ตัวแบบผู้พูดที่มีลำดับเท่ากับ 3 ..... 27
รูปที่ 4.1 ผลกระทบของลำดับตัวแบบผู้พูด ..... 31
รูปที่ 4.2 ผลกระทบของลำดับข้อมูลทดสอบ ..... 32

# บทที่ 1

# บทนำ

## 1.1 ความสำคัญและความเป็นมาของปัญหา

ภาษาเป็นตัวขับเคลื่อนให้เกิดวัฒนธรรม เสียงพูดเป็นสิ่งที่มีพลังเป็นอย่างมากและมีรูปแบบที่เป็นธรรมชาติ การผสมผสานระหว่างตัวอักษรกับเสียงพูดทำให้เกิดเป็นภาษาพูด ซึ่งเป็นสิ่งสำคัญอย่างยิ่งในการดำรงชีวิตของมนุษย์ แต่สิ่งหนึ่งที่เสียงพูดแตกต่างจากอักษรคือ เสียงพูดนั้นได้ซ่อนรายละเอียดหลายอย่างที่อักษรไม่มี อย่างเช่น เสียงพูดสามารถบ่งบอกถึงสุขภาพของผู้พูด เพศ อารมณ์ และคุณลักษณะอื่นๆ ที่เป็นตัวระบุถึงตัวผู้พูด สามารถที่จะสื่อสารให้คนอื่นรับรู้ได้ทางเสียงพูด

ดังนั้นเจตนารมณ์ที่จะทำการระบุตัวบุคคลจากเสียงพูดโดยใช้คอมพิวเตอร์ สามารถทำได้โดยการตรวจสอบข้อมูลทางภาษาจากเสียงพูดของคนๆ นั้น ด้วยความหลากหลายและความซับซ้อนในการทำธุรกรรมในปัจจุบัน ความสามารถที่จะระบุตัวบุคคลโดยใช้เสียงพูดเป็นสิ่งที่ช่วยอำนวยความสะดวกได้มาก เนื่องจากข้อได้เปรียบของเสียงพูด 4 ประการคือ ประการที่หนึ่ง เสียงพูดสามารถผลิตได้ง่ายมาก โดยไม่ต้องการทักษะพิเศษเพิ่มเติมเหมือนกับการพิมพ์หรือการกดปุ่ม ประการที่สอง เสียงพูดสามารถผลิตได้เร็วกว่าการนำเข้าข้อมูลโดยใช้แป้นพิมพ์ 3-4 เท่า [5] ประการที่สาม เสียงพูดสามารถป้อนเป็นข้อมูลเข้าได้ โดยที่ผู้พูดกำลังเคลื่อนไหวหรือทำงานโดยใช้อวัยวะอื่นๆ และประการสุดท้าย เราสามารถทำการเข้าถึงระบบในระยะไกลผ่านทางไมโครโฟนและสายโทรศัพท์โดยใช้เสียงพูดได้

การระบุผู้พูด (Speaker Identification) เป็นวิธีการในการตัดสินข้อมูลเสียงพูดว่าเป็นเสียงของใครจากข้อมูลเสียงทั้งหมดที่มีอยู่ในฐานข้อมูล ซึ่งได้ทำการลงทะเบียนไว้ การตัดสินโดยใช้หลักการของความใกล้เคียงกันของข้อมูลเสียง โดยรายละเอียดจะกล่าวในลำดับถัดไป ในด้านของความถูกต้อง จำนวนข้อมูลทั้งหมดในฐานข้อมูล (หมายถึงจำนวนข้อมูลของเสียงพูดที่แตกต่างกัน) เป็นปัจจัยหนึ่งที่มีผลต่อความถูกต้องในการระบุผู้พูดเป็นอย่างมาก ประสิทธิภาพในการระบุจะดีมากเมื่อมีจำนวนผู้พูดที่ลงทะเบียนไม่กี่คน โดยเฉพาะอย่างยิ่งถ้าพวกเขาเหล่านั้นมีเสียงที่แตกต่างกันมากๆ ด้วย แต่เมื่อจำนวนผู้พูดในฐานข้อมูลเพิ่มขึ้น ความน่าจะเป็นในการที่จะระบุผู้พูดได้ถูกต้องนั้นย่อมจะน้อยลงหรือบางทีอาจจะไม่ถูกต้องเลยก็ได้

การระบุผู้พูดโดยไม่ขึ้นกับคำพูด (Text-Independent Speaker Identification) เป็นเป้าหมายสูงสุดของการระบุผู้พูด ซึ่งงานวิจัยที่ผ่านมาบ่งชี้ว่า ความถูกต้องโดยรวมของระบบลดลงเมื่อเทียบกับการระบุผู้พูดโดยขึ้นกับคำพูด (Text-Dependent Speaker Identification) อนึ่ง การสร้างข้อมูลใน

ฐานข้อมูลนั้นได้มาจากการรู้จำ (Recognition) เสียงของผู้พูดแต่ละบุคคล จากงานวิจัยของ Doddington [2] ระบุว่า การระบุผู้พูดโดยไม่ขึ้นกับคำพูดนั้นต้องการข้อมูลเสียงพูดมากกว่าการระบุผู้พูดโดยขึ้นกับคำพูด เพื่อใช้ในการสร้างตัวแบบในการรู้จำเสียงของบุคคลหนึ่งๆ

ด้วยเหตุผลข้างต้น การสร้างตัวแบบในการรู้จำเสียงของผู้พูดโดยไม่ขึ้นกับคำพูดจำเป็นต้องใช้ข้อมูลเสียงพูดจากผู้พูดคนนั้นจำนวนมาก ยิ่งมากยิ่งดี ดังนั้นเมื่อข้อมูลมากขึ้นการประมวลผลในขั้นตอนการสร้างตัวแบบเสียงพูด รวมถึงขั้นตอนการระบุผู้พูดจะมากมายมหาศาลจนไม่สามารถตอบสนองให้ทันต่อความต้องการของมนุษย์ได้ ตลอดทั้งพื้นที่ในการจัดเก็บจะมากขึ้นเมื่อมีการเพิ่มข้อมูลของผู้พูดคนใหม่เข้าไปในระบบ

โดยทั่วไปข้อมูลเสียงพูดจะถูกแสดงในรูปกลุ่มของเวกเตอร์ลักษณะเฉพาะ (Feature Vectors) การแบ่งนับเวกเตอร์ (Vector Quantization, VQ) เป็นวิธีการที่นำมาประยุกต์กับการลดข้อมูลเสียงพูด ซึ่งอยู่บนพื้นฐานของการวัดความคล้ายและการจัดกลุ่ม ลักษณะเฉพาะที่คล้ายกันจะถูกจัดอยู่ในกลุ่มเดียวกัน จากนั้นจะนำตัวแทนของแต่ละกลุ่มออกมาเป็นตัวแทนของข้อมูลทั้งหมด ผลลัพธ์คือทำให้ปริมาณข้อมูลลดลง แต่การวัดความคล้ายส่วนใหญ่จะวัดระยะห่างระหว่างลักษณะเฉพาะหรือระยะห่างของเวกเตอร์แต่ละตัวโดยไม่ได้คำนึงถึงธรรมชาติของข้อมูลอย่างอื่น เช่น การเปลี่ยนแปลงของค่าภายในลักษณะเฉพาะ ความสำคัญของค่าในแต่ละมิติของเวกเตอร์ [3] เป็นต้น นอกจากนั้นวิธีการในการวัดระยะห่างที่มีประสิทธิภาพมีการคำนวณที่ซับซ้อน ส่งผลให้ใช้เวลาในการประมวลผลมาก ตลอดจนปัญหาของการให้ความสำคัญกับตัวแทนข้อมูลแต่ละกลุ่มเท่าเทียมกัน ซึ่งเป็นการไม่สมเหตุสมผล เพราะบางกลุ่มมีปริมาณเวกเตอร์น้อยแต่บางกลุ่มมีปริมาณเวกเตอร์มาก หรือมีความหนาแน่นของปริมาณเวกเตอร์แตกต่างกัน [7] ดังนั้นในงานวิจัยนี้จึงมุ่งเน้นศึกษาวิเคราะห์เพื่อเป็นแนวทางในการกำหนดมีวิธีการใหม่ที่นำมาแก้ไขปัญหาข้างต้น ซึ่งจะได้กล่าวต่อไป

## 1.2 วัตถุประสงค์ของงานวิจัย

1. ศึกษากระบวนการในการระบุผู้พูดจากเสียงพูดรวมถึง กระบวนการในการแบ่งนับข้อมูลเพื่อประหยัดเนื้อที่ในการจัดเก็บ
2. เสนอแนวทางใหม่ในการแบ่งนับข้อมูลผู้พูดเพื่อความรวดเร็วในการระบุผู้พูดและประหยัดเนื้อที่ในการจัดเก็บข้อมูล โดยคงความถูกต้องในการระบุผู้พูดได้
3. ศึกษาเปรียบเทียบกระบวนการในการระบุผู้พูดในเชิงการทดลองของวิธีการที่นำเสนอใหม่และวิธีการเดิม

## 1.3 ขอบเขตการศึกษาวิจัย

1. กำหนดคำในการพูด เพื่อเป็นการกำหนดช่วงความถี่ให้แน่นอน โดยในงานวิจัยนี้คำที่กำหนดคือ ศูนย์ ถึง เก้าในภาษาไทย แต่ไม่กำหนดลำดับและความยาวในการพูดของคำเหล่านั้น
2. เก็บข้อมูลเสียงพูดสำหรับสร้างตัวแบบ ทดสอบตัวแบบ โดยการบันทึกเสียงพูดจากบุคคลต่างเพศต่างวัยในสภาพแวดล้อมแบบห้องทำงาน ซึ่งปกติจะมีเสียงรบกวนคือเสียงเครื่องปรับอากาศและเสียงคอมพิวเตอร์
3. วิเคราะห์และสกัดลักษณะสำคัญของเสียงพูดด้วยวิธีการแบบ MFCCs และเก็บลงในฐานข้อมูล
4. งานวิจัยนี้จะเปรียบเทียบ การแบ่งนับข้อมูลตามแนวคิดใหม่กับการแบ่งนับข้อมูลตามวิธีการแบ่งนับข้อมูลเวกเตอร์ (Vector Quantization) และการแบ่งนับข้อมูลแบบฐานสอง (Binary Quantization) แบบเดิม โดยพิจารณาถึงเปอร์เซ็นต์ความถูกต้องของประสิทธิภาพในการระบุตัวบุคคล ความเร็ว และพื้นที่ในการเก็บตัวแบบ เมื่อมีการเพิ่มจำนวนคนเข้าไปในระบบมากขึ้น โดยใช้ข้อมูลทดสอบชุดเดียวกัน
5. สรุปปัจจัยที่ส่งผลกระทบต่อระบบที่สร้างขึ้นจากแนวคิดใหม่

## 1.4 ขั้นตอนและวิธีดำเนินงานวิจัย

1. ศึกษาแนวคิด ทฤษฎี การระบุผู้พูดจากเสียงพูด
2. ศึกษาทฤษฎีที่เกี่ยวข้องที่สามารถทำการแบ่งนับข้อมูลตามสมติฐานเกี่ยวกับลักษณะทางกายภาพของข้อมูลที่ได้จากการศึกษาจากลักษณะเฉพาะของเสียงพูด
   2.1 ศึกษาวิธีการเพื่อหาจุดเปลี่ยนแปลงความชันในแนวแกนนอนและวิธีการในการตรวจสอบการเปลี่ยนแปลงจุดสูงสุดต่ำสุดในแนวแกนตั้งของลักษณะเฉพาะของเสียงพูด และนำมาทำการคำนวณค่าด้วยวิธีการแบ่งนับข้อมูลแบบฐานสอง เพื่อใช้ในการแยกกลุ่มของลักษณะเฉพาะของเสียงพูด (รายละเอียดในบทที่ 3)
   2.2 ศึกษาวิธีการทางสถิติเพื่อทำการคำนวณหาค่าน้ำหนักของกลุ่มลักษณะเฉพาะของเสียงพูดแต่ละกลุ่มที่มีจำนวนไม่เท่ากัน
3. เก็บข้อมูลเสียงพูด และสกัดลักษณะเฉพาะของเสียงพูด
   3.1 เก็บข้อมูลเสียงพูดของบุคคลต่างเพศต่างวัยโดยให้อ่านตามแบบร่างที่มีตัวเลขศูนย์ถึงเก้าภายในสภาพแวดล้อมแบบห้องทำงานและเก็บในรูปแบบของ .wav ผ่านทาง

ไมโครโฟน โดยมีอัตราการชักตัวอย่างเท่ากับ 8000 รอบต่อวินาที ความละเอียด 16 บิต โดยให้ผู้ทดสอบ

- อ่านตัวเลขศูนย์ถึงเก้าต่อเนื่องกัน 6 รอบ เพื่อใช้เป็นข้อมูลในการสร้างตัวแบบและข้อมูลในการตรวจสอบความถูกต้องของตัวแบบ
- อ่านตัวเลขศูนย์ถึงเก้าสลับที่กันไปมาโดยไม่คำนึงถึงลำดับและความยาว คนละ 2 นาทีเพื่อใช้เป็นข้อมูลสำหรับการทดสอบ (Testing)
- เก็บข้อมูลจากผู้ทดสอบทั้งหมดประมาณ 30 คน

3.2 นำข้อมูลเสียงที่ได้มาทำการสกัดลักษณะสำคัญของเสียงพูดของแต่ละคน โดยใช้วิธีการ MFCCs และทำการจัดเก็บลงฐานข้อมูล

4. สร้างตัวแบบของเสียงพูดศูนย์ถึงเก้าของแต่ละคน เพื่อใช้ในการรู้จำเสียงพูดตามวิธีการที่ได้จากข้อ 2 โดย

4.1 ออกแบบตัวแบบสำหรับเก็บค่าที่คำนวณได้จากข้อสอง เพื่อเป็นตัวแบบในการเปรียบเทียบของแต่ละคนโดยคำนึงถึง

- ความเร็วในการระบุผู้พูด
- พื้นที่ในการจัดเก็บข้อมูลเสียงพูดของแต่ละบุคคล
- ความถูกต้องในการค้นหา และประสิทธิภาพในการระบุตัวบุคคล

5. สร้างตัวแบบของวิธีการเดิม เพื่อใช้ในการเปรียบเทียบประสิทธิภาพในการระบุตัวบุคคล โดยใช้ข้อมูลเข้าเดียวกันกับตัวแบบในข้อ 4

6. เปรียบเทียบประสิทธิภาพของตัวแบบที่ได้จากข้อ 4 และข้อ 5 โดยที่

6.1 วัดความถูกต้องของการระบุผู้พูดในรูปแบบของร้อยละของความถูกต้อง โดยสามารถพูดสลับตัวเลขศูนย์ถึงเก้าได้และการทดสอบจะแบ่งช่วงเวลาของข้อมูลเข้า เพื่อใช้ในการทดสอบเช่น ข้อมูลทดสอบชุดที่หนึ่ง เป็นเสียงที่มีความยาว 1 วินาที ชุดที่สอง 3 วินาที เป็นต้น

6.2 วัดประสิทธิภาพในเชิงความเร็วในการลงทะเบียนผู้พูดคนใหม่และเวลาในการระบุผู้พูดเมื่อมีตัวอย่างเสียงที่ไม่รู้จักเข้ามา

6.3 วัดพื้นที่ในการเก็บข้อมูลเสียงพูด 1 คน

6.4 สรุปผลการทดลอง

7. ปรับแต่งตัวแบบ โดยเพิ่มสมมติฐานและทฤษฎีที่คาดว่าจะทำให้ตัวแบบมีประสิทธิภาพเพิ่มขึ้นจากนั้นกลับไปทำการทดสอบอีกครั้ง

8. ทำการสรุปข้อดีและข้อด้อยของวิธีการใหม่

## 1.5 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. เข้าใจถึงกระบวนการทั้งหมดในการทำการรู้จำเสียงพูด
2. เข้าใจถึงวิธีการในการวิเคราะห์สัญญาณเสียงและปัจจัยที่มีผลกระทบต่อความถูกต้องของระบบ
3. ได้ข้อสรุปของผลการทดสอบกระบวนการในการรู้จำเสียงพูดแต่ละวิธี ที่ถูกนำมาใช้กับประโยคที่เป็นภาษาไทย
4. ได้วิธีการในการรู้จำเสียงพูดอีกวิธีการหนึ่งที่คาดว่าสามารถนำไปใช้ในการทำการรู้จำเสียงพูดได้อย่างมีประสิทธิภาพ
5. สามารถนำผลการทดลองรวมถึงเครื่องมือที่ได้พัฒนาขึ้นมาไปเป็นแนวทางในการวิจัยต่อ และสามารถนำไปพัฒนาร่วมกับวิธีการอื่นเพื่อผลิตเป็นโปรแกรมประยุกต์ต่อไป

# บทที่ 2

# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง โดยส่วนแรกของบทนี้จะพูดถึงทฤษฎีที่เกี่ยวข้องกับงานวิจัย ซึ่งประกอบไปด้วย การได้มาซึ่งเสียงพูด การเตรียมเสียงพูด การวิเคราะห์สัญญาณเสียงพูดแบบไม่ต่อเนื่อง ทฤษฎีในการสร้างตัวแบบผู้พูด การสอน และการทดสอบ

ในส่วนท้ายของบทนี้จะกล่าวถึงงานวิจัยเกี่ยวกับการระบุผู้พูดซึ่งประกอบไปด้วยวิธีการแบ่งนับแบบต่างๆ เช่น การแบ่งนับแบบเวกเตอร์ (Vector Quantization) และวิธีการแบ่งนับแบบฐานสอง (Binary Quantization) ซึ่งมีหลักการที่สำคัญในการนำมาประยุกต์ใช้ในงานวิจัยนี้ รวมไปถึงวิธีการอื่นอีกหลายวิธีการที่มีประสิทธิภาพในการระบุผู้พูด

## 2.1 ทฤษฎีที่เกี่ยวข้อง

### 2.1.1 การได้มาซึ่งข้อมูลเสียงพูด (Speech Data Acquisition)

ในการระบุผู้พูดนั้นจำเป็นต้องนำเสียงพูดของแต่ละบุคคลมาใช้เป็นข้อมูลเข้า ดังนั้นในตอนต้นของบทนี้จะขอกล่าวถึงการได้มาซึ่งข้อมูลเสียงพูดก่อนเป็นอันดับแรก ซึ่งจะต้องผ่านกระบวนการดังต่อไปนี้

#### *2.1.1.1 กลไกในการผลิตเสียงพูด (Speech Production Mechanism)*

เสียงพูด (Speech) จะสามารถผลิตได้โดยผ่านขั้นตอนหลักๆ 3 ขั้นตอนคือ ต้นกำเนิดเสียง (Source Generation) ส่วนในการออกสียง (Articulation) และ ส่วนปล่อยเสียง (Radiation)

เมื่อคนเราตั้งใจจะพูด กล้ามเนื้อบริเวณกระบังลมจะบีบให้กระบังลมหนุนสูงขึ้นเพื่อบังคับให้ลมออกจากปอด เมื่อลมออกจากปอดผ่านท่อลม (Trachea) และผ่านไปถึงช่องเส้นเสียง (Glottis) ซึ่งเป็นช่องว่างอยู่ระหว่างซ้ายและขวาของสายเสียง (Vocal Cord) ซึ่งปกติจะเปิดตอนหายใจ ลมที่ผ่านเข้าไปในช่องเส้นเสียงจะถูกขัดจังหวะเป็นช่วงๆ ของคาบเวลา โดยการปิดและเปิดเพื่อให้ลมผ่านไปหาสายเสียง และให้เสียงออกมาเพื่อผ่านขั้นตอนสุดท้ายในการปล่อยเสียงซึ่งอาจจะเป็นทางปาก ทางจมูก หรือทั้งสองทาง

ช่องเสียง (Vocal Tract) สามารถที่จะเปลี่ยนแปลงรูปทรงได้หลากหลายรูปแบบโดยการเคลื่อนย้าย ฟัน ลิ้น ริมฝีปาก หรืออวัยวะส่วนที่อยู่ข้างใน เพื่อให้เสียงที่ส่งผ่านเกิดกำธร (Resonance)

ซึ่งการวิเคราะห์สัญญาณเสียงพูดนั้นทำได้โดยการจำลองรูปแบบของช่องเสียงขึ้นมา ซึ่งช่องเสียงของแต่ละคนจะมีรูปทรงและความยาวไม่เท่ากัน ส่งผลให้เสียงที่พูดออกมานั้นแตกต่างกัน [5]

#### *2.1.1.2 การบันทึกเสียงพูด (Speech Data Recording)*

สัญญาณเสียงนั้น เรานำเข้ามาประมวลผลโดยการวัดการเปลี่ยนแปลงของศักดิ์ไฟฟ้าผ่านทางไมโครโฟน สัญญาณที่ได้จะอยู่ในรูปของสัญญาณทางไฟฟ้าที่เปลี่ยนแปลงไปตามเวลาเรียกว่าสัญญาณแอนะล็อก เพื่อที่จะทำการแปลงสัญญาณเสียงจากสัญญาณแอนะล็อกไปเป็นสัญญาณดิจิทัลจะต้องทำการชักตัวอย่างสัญญาณ และทำการเข้ารหัสก่อนที่จะนำไปทำการวิเคราะห์ต่อไป

**1) ความถี่ในการชักตัวอย่าง (Sampling Frequency)**

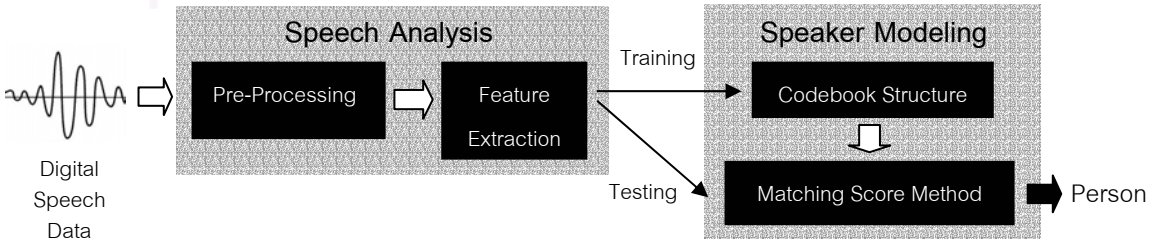
สัญญาณแอนะล็อกจะถูกกรองโดยการชักตัวอย่าง (Sampling) ซึ่งตามทฤษฎีแล้วความถี่ในการชักตัวอย่างจะมีค่ามากกว่าสองเท่าของความถี่สูงสุดของสัญญาณที่นำเสนอ [5] ตัวอย่างเช่นความถี่สูงสุดของสัญญาณเสียงโทรศัพท์เท่ากับ 3400 Hz. ดังนั้นถ้าต้องการให้ได้สัญญาณดิจิทัลครบจะต้องทำการชักตัวอย่างด้วยความถี่มากกว่า 2 x 3400 Hz.

**2) ความละเอียดของค่าที่ได้จากการชักตัวอย่าง (Sampling Resolution)**

เป็นจำนวนบิตของตัวเลขที่จะแสดงถึงระดับความดังของเสียง (Amplitude) ที่ได้จากการชักตัวอย่าง เช่น 8 บิตจะแสดงค่าความดังตั้งแต่ 0 – 127 และ 16 บิตจะแสดงค่าความดังตั้งแต่ – 32767…0…32768 เป็นต้น

### 2.1.2 ขั้นตอนการระบุผู้พูด

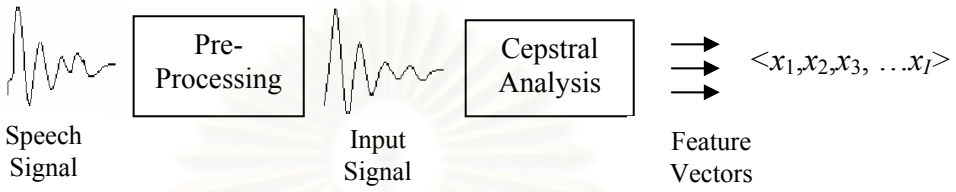
วิธีการระบุผู้พูดนั้นประกอบไปด้วยขั้นตอนหลัก 2 ขั้นตอน ดังรูปที่ 2.1 คือ ขั้นตอนการวิเคราะห์สัญญาณเสียงพูด (Speech Analysis) และขั้นตอนการสร้างตัวแบบผู้พูด (Speaker Modeling) ซึ่งมีรายละเอียดดังต่อไปนี้




รูปที่ 2.1 ขั้นตอนการระบุผู้พูด

### 2.1.3 **การวิเคราะห์สัญญาณเสียงพูด** (Speech Analysis)

เป็นขั้นตอนการวิเคราะห์สัญญาณเสียง โดยพิจารณาจากกระบวนการในการผลิตเสียงพูด เพื่อนำไปใช้เป็นข้อมูลในการสร้างตัวแบบสำหรับการรู้จำและระบุผู้พูดต่อไป ขั้นตอนนี้สามารถแบ่งออกได้เป็นสองขั้นตอนย่อยคือ ขั้นตอนการเตรียม (Pre-processing) และขั้นตอนการสกัดลักษณะเฉพาะ (Feature Extraction)




รูปที่ 2.2 ขั้นตอนการสกัดลักษณะเฉพาะด้วยวิธีการวิเคราะห์เซ็พสตรัล

#### *2.1.3.1 ขั้นตอนการเตรียม (Pre-processing)*

เนื่องจากข้อมูลเสียงพูดที่นำเข้ามาทำการวิเคราะห์ โดยส่วนใหญ่แล้วจะมีทั้งข้อมูลที่เป็นสัญญาณเสียงพูดและสัญญาณอื่นปะปนมา ซึ่งอาจส่งผลกระทบต่อการวิเคราะห์สัญญาณ ดังนั้นจึงจำเป็นต้องมีขั้นตอนในการเตรียมข้อมูลเสียงพูดเพื่อให้ได้ข้อมูลที่บริสุทธิ์ก่อนที่จะนำไปทำการวิเคราะห์ ขั้นตอนนี้ประกอบไปด้วย

1) **ขั้นตอนลดเสียงรบกวน** (Noise Reduction Process)

เสียงรบกวน (Noise) เป็นปัญหาที่สำคัญมากในการวิเคราะห์สัญญาณเพราะทำให้ระบบโดยรวมเกิดความคลาดเคลื่อนไปจากความเป็นจริง ประสิทธิภาพในการระบุผู้พูดลดลง

เสียงรบกวนอาจเกิดจากหลายปัจจัยด้วยกัน เช่น เกิดจากเสียงของเครื่องปรับอากาศ เสียงของแมลง เสียงของเครื่องจักร ซึ่งอาจจะเป็นเสียงที่มีลักษณะไม่เป็นคาบ เราสามารถตรวจจับได้ง่าย รวมไปถึงเสียงรบกวนที่ยากต่อการตรวจจับอย่างเช่น เสียงรบกวนจากคน เป็นต้น ดังนั้นระบบที่ดีจำเป็นต้องมีความคงทนต่อเสียงรบกวนหรือสามารถที่จะทำให้เสียงรบกวนนั้นหมดไป หรือเหลือน้อยที่สุด การหาวิธีการที่ดีที่สุดสำหรับลดทุกๆ เสียงรบกวนนั้นเป็นไปได้ยาก ด้วยเหตุนี้เราจึงต้องระบุสภาพแวดล้อมที่เราทำการเก็บข้อมูลเสียงพูดและสภาพแวดล้อมที่จะใช้งาน เพื่อให้ทราบว่าสภาพแวดล้อมนั้นๆ มีเสียงรบกวนอะไรบ้าง เพื่อประโยชน์สำหรับการเตรียมตัวแบบในการลดเสียงรบกวนนั้นๆ ได้ [11]

### 2) ขั้นตอนกำจัดออฟเซตของไฟฟ้ากระแสตรง (DC Offset Removal Process)

เพื่อให้ง่ายต่อการเข้าใจ ถ้านึกย้อนไปถึงประสบการณ์ในการใช้เครื่องวัดสัญญาณ (Oscilloscope) โดยครึ่งบนจะเป็นสัญญาณไฟฟ้าที่เป็นบวก ครึ่งล่างจะเป็นสัญญาณไฟฟ้าที่เป็นลบ และเส้นตรงกลางคือแกนศูนย์ เมื่อนำเครื่องวัดนี้ไปวัดสัญญาณที่มีลักษณะเป็นคาบ เช่น สัญญาณไฟฟ้ากระแสสลับ เป็นต้น ค่าเฉลี่ยของตำแหน่งส่วนบนและตำแหน่งล่างจะเท่ากับศูนย์ แต่เมื่อนำไปวัดกับไฟฟ้ากระแสตรงซึ่งไม่มีการเปลี่ยนแปลงไปตามเวลาหรือมีความถี่เท่ากับ 0 Hz. จะมีค่าเฉลี่ยเท่ากับค่าของไฟฟ้ากระแสตรงนั้น ดังนั้นดีซีออฟเซตก็คือ ส่วนของไฟฟ้ากระแสตรงที่ถูกบวกเข้าไปในสัญญาณที่เป็นคาบเวลา ในขั้นตอนของการแปลงสัญญาณจากแอนะล็อกไปเป็นสัญญาณดิจิทัล ส่งผลให้สัญญาณที่มีลักษณะเป็นคาบเวลายกขึ้นหรือลดระดับลงไปจากแกนศูนย์ทำให้การประเมินค่าของสัญญาณผิดเพี้ยนไปจากเดิมเมื่อยึดแกนศูนย์เป็นแกนอ้างอิง

ในการวัดค่าดีซีออฟเซตนั้นทำได้โดยการหาค่าเฉลี่ยค่าของสัญญาณเทียบกับแกนศูนย์ว่ามีค่าเท่าใด ถ้าค่าที่ได้เท่ากับศูนย์แสดงว่าสัญญาณไม่มีดีซีออฟเซตและถ้าค่าที่ได้มีค่ามากๆ หรือค่าดีซีออฟเซตสูงจะส่งผลให้จุดสูงต่ำของสัญญาณมีค่าเกินความเป็นจริง ซึ่งถ้าเป็นสัญญาณเสียงจะทำให้เสียงที่ได้เบาหรือดังเกินความเป็นจริง ดังนั้นเราควรกำจัดดีซีออฟเซตออกจากสัญญาณ ซึ่งทำได้โดยการกรองสัญญาณความถี่เท่ากับศูนย์ (DC) ออกไป [12]

### 3) ขั้นตอนกำจัดเสียงเงียบ (Silence Removal Process)

เสียงเงียบ (Silence) ในทางปฏิบัติแล้วสัญญาณของเสียงเงียบจะไม่เงียบจริง เพราะจะมีเสียงรบกวนที่อาจเกิดจากสภาพแวดล้อมหรือเกิดจากไมโครโฟนเอง ดังนั้น ในการพิจารณาว่าช่วงใดเป็นเสียงเงียบ จะต้องทำการตั้งค่า ระดับสูงต่ำไว้ ถ้าค่าความดังของเสียงอยู่ในช่วงที่กำหนดให้ถือว่าเป็นเสียงเงียบ

แม้ว่าเสียงเงียบจะไม่ได้ส่งผลกระทบต่อการวิเคราะห์สัญญาณโดยตรง แต่ถ้าพิจารณาดูแล้วเสียงเงียบไม่ใช่สัญญาณเสียงที่เราต้องการทำการวิเคราะห์ ถ้าเราไม่ทำการตัดออกจะทำให้เสียเวลาในการคำนวณโดยเปล่าประโยชน์ และนอกจากนั้นเสียงเงียบที่มีเสียงรบกวนมากๆ ถ้าเรานำมาร่วมในการวิเคราะห์จะส่งผลทำให้ประสิทธิภาพของระบบโดยรวมลดลง

### *2.1.3.2 ขั้นตอนการสกัดลักษณะเฉพาะ (Feature Extraction)*

การสกัดลักษณะเฉพาะ เป็นวิธีการที่ดึงลักษณะที่สำคัญออกมาจากข้อมูลเสียงพูด ทั้งนี้มีเหตุผลสองประการคือ ประการที่หนึ่ง เพื่อหลีกเลี่ยงการใช้ข้อมูลเสียงพูดทั้งหมด ประการที่สอง เพื่อลดการเปลี่ยนแปลงของข้อมูลเสียงพูด ซึ่งมีทฤษฎีที่เกี่ยวข้องดังนี้

#### 1) **โครงสร้างแถบความถี่ของเสียงพูด** (Spectral Structure of Speech)

คลื่นเสียงสามารถแปลงให้อยู่ในรูปแถบความถี่ (Spectrum) เพื่อทำการวิเคราะห์แทนการวิเคราะห์โดยตรงที่อยู่ในแกนเวลา มีเหตุผลสองประการคือ ประการแรก คลื่นเสียงเกิดจากการรวมกันของคลื่นพื้นฐาน (sin ($\omega$t), cos ($\omega$t)) หลายๆ ความถี่ที่มีการเปลี่ยนแปลงของแอมพลิจูดและเฟสอย่างช้าๆ ซึ่งองค์ประกอบพื้นฐานเหล่านี้แสดงอยู่ในรูปของความถี่ ประการที่สองคือ การรับรู้ของมนุษย์โดยผ่านการรับฟังจะต้องพิจารณาจากข้อมูลที่อยู่ในรูปของความถี่ [5]

ค่าความหนาแน่นของความถี่ (Power Spectral Density) ของเสียงช่วงสั้นๆ สามารถที่จะพิจารณาได้ว่า ประกอบจาก 2 ส่วนด้วยกันคือ ส่วนที่เปลี่ยนแปลงอย่างช้าๆ ไปตามสมการของความถี่เรียกว่าสเปกตรัลเอ็นเวลอป (Spectral Envelope) และส่วนที่มีการเปลี่ยนแปลงอย่างรวดเร็วเรียกว่าสเปกตรัลไฟน์สตรัคเจอร์ (Spectral Fine Structure) ซึ่งเป็นคาบของคลื่นที่เป็นเสียงพูดเท่านั้น [5] ส่วนของสเปกตรัลเอ็นเวลอปนั้น ไม่เพียงแต่จะบ่งบอกถึงเสียงที่เกิดจากการกำธรและที่ไม่กำธรจากอวัยวะภายในแล้ว ยังรวมไปถึงความถี่ของเสียงที่เกิดจากช่องเส้นเสียงและเสียงที่เกิดจากริมฝีปากรวมถึงช่องจมูกด้วย ในการที่จะทำการพิสูจน์เสียงหนึ่ง ๆ ว่าเป็นของใครนั้น เราจำเป็นต้องดึงเอาส่วนที่ไม่มีการเปลี่ยนแปลงหรือมีการเปลี่ยนแปลงอย่างช้าๆ ออกมา ซึ่งส่วนนั้นก็คือสเปกตรัลเอ็นเวลอป โดยใช้การวิเคราะห์เซฟสตรัล ซึ่งจะกล่าวถึงในหัวข้อต่อไป

#### 2) **การแปลงฟูริเยร์** (The Fourier Transform)

ในการวิเคราะห์ความหนาแน่นของความถี่นั้น เราจำเป็นต้องทำการแปลงจากคลื่นเสียงที่อยู่ในแกนของเวลาให้อยู่ในแกนของความถี่ ซึ่งสามารถทำได้โดยการแปลงฟูริเยร์

การแปลงฟูริเยร์จะอยู่บนพื้นฐานของการค้นหาว่าช่วงคลื่นเสียงหนึ่งๆ มีจำนวนผลบวกของฟังก์ชัน Sine และ Cosine อะไรบ้าง โดยมีความถี่เริ่มต้นจากศูนย์และจะเพิ่มขึ้นเป็นจำนวนเท่าของ $f_0 = 1/T$ จำนวนเต็มที่นำไปคูณกับความถี่มูลฐาน ซึ่ง $T$ เป็นคาบเวลาของของฟังก์ชัน $x(t)$ ที่ได้จากสมการที่ (1) ดังนี้

$$x(t) = a_0 + \sum_{k=1}^{\infty} (a_k \cos(2\pi k f_0 t) + b_k \sin(2\pi k f_0 t)) \quad \text{..... (1)}$$

จากสมการข้างต้น รูปแบบทางขวามือของสมการจะเรียกว่า อนุกรมฟูริเยร์ จุดประสงค์หลักของการแปลงฟูริเยร์ก็เพื่อจะหาสัมประสิทธิ์ $a_k$ และ $b_k$ ทั้งหมดที่เป็นไปได้ โดยให้ความถี่มูลฐาน และฟังก์ชัน $x(t)$ มา ค่า $a_o$ ของสมการจะเป็นค่าสัมประสิทธิ์ cosine สำหรับ $k = 0$ และจะไม่มีค่า $b_o$ ซึ่งเป็นค่าสัมประสิทธิ์ของ sine เพราะค่า sine ที่ $k = 0$ มีค่าเป็นศูนย์

แน่นอนว่าเราไม่สามารถที่จะทำการหาค่าสัมประสิทธิ์ที่จำกัดของสมการข้างต้นได้ เพราะค่า $k$ เข้าใกล้อนันต์ ดังนั้นเราจำเป็นต้องทำการจำกัดขอบเขตของช่วงเวลาที่เราทำการพิจารณา ซึ่งเรียกว่าการแปลงฟูริเยร์แบบไม่ต่อเนื่อง (Discrete Fourier Transform) โดยใช้วิธีการของวินโดว์วิง

**3) วินโดว์วิง** (Windowing)

ในการวิเคราะห์สัญญาณเสียงนั้น เราต้องทำการสมมติว่าสัญญาณเกิดขึ้น ณ ช่วงเวลาสั้นๆ และจากนั้นจะทำการแปลงฟูริเยร์ในช่วงนั้นๆ วิธีการก็คือ เราทำการคูณสัญญาณเสียงด้วยฟังก์ชันของวินโดว์ ซึ่งถ้าเกินขอบเขตของวินโดว์ ค่าของสัญญาณจะเท่ากับศูนย์

วินโดว์ที่เป็นสี่เหลี่ยมมุมฉากกำหนดดังสมการที่ (2) ดังนี้

$$w_{\mathrm{n}} = \begin{cases} 1 & 0 \le \mathrm{n} \le N \\ 0 & \text{otherwise} \end{cases} \quad \text{..... (2)}$$

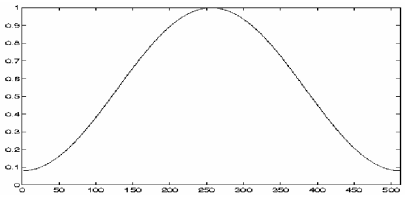
ซึ่งเมื่อนำไปคูณกับสัญญาณจะได้ลักษณะดังตัวอย่างรูปที่ 2.3

รูปที่ 2.3 สัญญาณที่ได้จากการคูณด้วยวินโดว์สี่เหลี่ยมมุมฉาก

แต่เมื่อพิจารณาดูแล้วจะเห็นถึงปัญหาการขาดตอนของสัญญาณ (Discontinuities) ในส่วนท้ายของวินโดว์ ซึ่งจะทำให้ช่วงท้ายของวินโดว์ หลังจากทำการแปลงฟูริเยร์จะมีส่วนที่เป็นลูกคลื่นติดมาด้วย (Ripple) [15] ซึ่งเป็นสิ่งที่เราไม่ต้องการ ดังนั้นหนทางหนึ่งที่จะหลีกเลี่ยงปัญหาการขาดตอนของสัญญาณคือการทำให้ส่วนท้ายมีลักษณะเรียวลงจนกระทั่งเป็นศูนย์หรือเข้าใกล้ศูนย์ ซึ่งวิธีการที่เป็นที่นิยมที่สุดคือ การใช้แฮมมิ่งวินโดว์ (Hamming Window)

$$w_n = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{(n-1)}\right) & , \ 0 \le n \le N \\ 0 & \text{otherwise} \end{cases} \quad \text{..... (3)}$$

ซึ่งเมื่อวาดค่าของ Hamming Window จากสมการที่ (3) จะได้กราฟดังรูปที่ 2.4



รูปที่ 2.4 แฮมมิ่งวินโดว์

เราจะทำการคูณช่วงของข้อมูลที่ได้จากวินโดว์แบบสี่เหลี่ยมด้วยแฮมมิ่งวินโดว์แล้วนำไปทำการวิเคราะห์ ต่อไปก็จะไม่เกิดปัญหาการขาดตอนของสัญญาณ

**4) การแปลงฟูริเยร์แบบเร็ว** (Fast Fourier Transform)

การแปลงฟูริเยร์แบบทั่วไปนั้น เมื่อนำมาใช้ในการประมวลผลสัญญาณเสียง ซึ่งข้อมูลมีจำนวนมากส่งผลให้มีการคำนวณที่มากในรูปแบบกำลังสองของข้อมูลที่เพิ่มขึ้นและทำให้ระบบโดยรวมช้ามาก ดังนั้น ขั้นตอนวิธีใหม่ที่ถูกนำเสนอขึ้นมาเพื่อแก้ปัญหาดังกล่าวก็คือการแปลงฟูริเยร์แบบเร็ว (Fast Fourier Transform, FFT) ซึ่งอยู่ในรูปแบบของผลคูณของค่าลอการิทึมของจำนวนข้อมูลที่เพิ่มขึ้น ซึ่งจะนำมาใช้ในงานวิจัยนี้

**5) การวิเคราะห์เซ็พสตรัล** (Cepstral Analysis)

เซ็พสตรัม (Cepstrum) หรือสัมประสิทธิ์เซ็พสตรัม $\mathbf{c(\tau)}$ ถูกค้นพบโดย Bogert ในปี ค.ศ.1963 โดยจากการทดลองเขาพบว่าค่าลอการิทึม (Logarithm) ของกำลังของความถี่ (Power Spectrum) ของสัญญาณเสียงที่มีการเพิ่มเสียงก้องที่มีลักษณะเป็นคาบเข้าไปจะมีจุดสูงสุดที่สูงขึ้นในส่วนของเสียงก้องที่มีการหน่วงเวลา (Echo Delay) เขาเรียกฟังก์ชันนี้ว่า เซ็พสตรัม ซึ่งได้จากการสลับอักษรคำว่า “Spectrum” และค่าพารามิเตอร์อิสระสำหรับเซ็พสตรัมจะเรียกว่าควีเฟรนซี่ (Quefrency) ซึ่งเป็นค่าพารามิเตอร์ที่อยู่ในแกนของเวลา คุณลักษณะพิเศษของเซ็พสตรัมคือ สามารถทำการแบ่งแยกส่วนของสเปกตรัลเอ็นเวลอปและสเปกตรัลไฟน์สตรัคเจอร์ออกจากกันได้ [5]

ตามทฤษฎีแล้วเสียงพูด $x(t)$ เกิดจากการผสมผสานกันระหว่างส่วนประกอบสองส่วนด้วยกัน คือ เสียงที่เกิดจากช่องการออกเสียง (Vocal Tract Articulation) $g(t)$ และเสียงที่เกิดจากช่องการตอบสนองต่ออิมพัลส์ (Vocal Tract Impulse Response) $h(t)$ [5] แสดงด้วยสมการที่ (4) ดังนี้

$$x(t) = \int g(\tau)h(t-\tau)d\tau \qquad \text{..... (4)}$$

ซึ่งเท่ากันทุกประการกับสมการที่ (5)

$$X(\omega) = G(\omega)\, H(\omega) \qquad \text{..... (5)}$$

ซึ่ง $X(\omega)$, $G(\omega)$ และ $H(\omega)$ ต่างก็คือการแปลงฟูริเยร์ของ *x(t), g(t)* และ *h(t)* ตามลำดับ

ค่าการแปลงฟูริเยร์ที่ได้จะเป็นจำนวนเชิงซ้อน ซึ่งประกอบไปด้วยส่วนจริงและส่วนจินตภาพ ในการวิเคราะห์เซ็พสตรัลนั้นเราไม่ได้คำนึงถึงเฟส ดังนั้นขนาด (Magnitude, $|X(\omega)|$) จึงถูกนำมาใช้แทนค่าจำนวนเชิงซ้อนที่ได้จากการแปลงฟูริเยร์นั้น

เมื่อทำการใส่ค่าลอการิทึมเข้าไปใน (5) ส่วนของ $G(\omega)$ และ $H(\omega)$ จะแยกออกจากกันในแกนของความถี่

$$\text{Log}\,|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \qquad \text{..... (6)}$$

เซ็พสตรัมคือค่าที่ได้จากการแปลงฟูริเยร์ย้อนกลับของ (6)

$$c(\tau) \quad = F^{-1}\,\text{Log}\,|X(\omega)| = F^{-1}\log|G(\omega)| + F^{-1}\log|H(\omega)| \qquad \text{..... (7)}$$

โดยที่ F คือการแปลงฟูเรียร์

ทางด้านขวามือของสมการ (6) ลำดับแรกและลำดับที่สอง จะบ่งบอกถึงส่วนของสเปกตรัลไฟน์สตรัคเจอร์และสเปกตรัลเอ็นเวลอปตามลำดับ

โดยทฤษฎีแล้ว ทางด้านขวามือของสมการที่ (7) ฟังก์ชันแรกจะบ่งบอกถึงโครงสร้างของเสียงที่อยู่ในส่วนของควีเฟรนซี่สูง (High Quefrency) และฟังก์ชันที่สองจะเป็นโครงสร้างของเสียงที่อยู่ในส่วนของควีเฟรนซี่ต่ำ (Low Quefrency) ซึ่งโดยทั้งไปส่วนของควีเฟรนซี่ต่ำจะอยู่ที่ประมาณ 0 ถึง 2 หรือ 4 มิลลิวินาที [5] จากนั้นเราสามารถทำการแยกสองส่วนออกจากกันได้และเรียกการแยกนั้นว่าลิฟเตอร์ริง (Liftering) ซึ่งมาจากคำว่า “Filtering” นั่นเอง

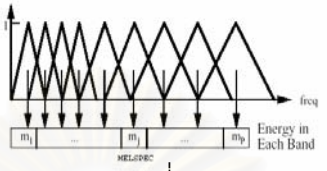
### 6) สัมประสิทธิ์เซ็พตรัลในแกนความถี่แบบเมล (Mel-Frequency Cepstral Coefficients, MFCCs)

เป็นวิธีการคำนวณค่าเซ็พตรัมวิธีการหนึ่ง ซึ่งจุดมุ่งหมายเป็นการนำเอาหลักการของความถี่แบบเมล (Mel- Frequency) มารวมเข้ากับการวิเคราะห์เซ็พตรัมแบบเดิมตามที่ได้กล่าวมาแล้วข้างต้น เป็นเพราะปัจจัยในการตอบสนองการรับฟังหรือการได้ยินของมนุษย์ในแต่ละช่วงความถี่ไม่เท่ากัน ดังนั้นจะต้องทำการแปลงแถบความถี่ รวมไปถึงการให้ความสำคัญและลดความความสำคัญของข้อมูลเสียงบางช่วงความถี่ให้อยู่ในช่วงที่มนุษย์สามารถรับรู้ได้มากที่สุด

การแปลงจากแกนความถี่เชิงเส้น (Linear) ไปยังแกนความถี่แบบเมล แปลงได้จากสมการ (8)

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \qquad \text{..... (8)}$$

จากรูปที่ 2.5 แสดงถึงการแปลงด้วยตัวกรองจำกัดช่วงความถี่ที่มีลักษณะเป็นสามเหลี่ยม โดยที่พื้นที่ทั้งหมดเป็น แถบความถี่แบบเมล โดยจำกัดความถี่อยู่ที่ 300-3400 Hz




รูปที่ 2.5 แถบความถี่แบบเมล [15]

MFCCs สามารถคำนวณได้จากสมการที่ (9)

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \qquad \text{..... (9)}$$

โดยที่ $N$ คือจำนวนของตัวกรอง, $m_j$ คือ ค่าลอการิทึมของแอมพลิจูดของผลลัพธ์ที่ได้จากตัวกรองจำกัดช่วงความถี่

MFCCs เป็นหนึ่งในวิธีการที่ได้รับความนิยมอย่างแพร่หลาย และถูกนำไปใช้ในงานวิจัยหลายแขนงเกี่ยวกับการวิเคราะห์เสียงพูด [13] การที่สามารถระบุช่วงความถี่จากเมลสเกล (Mel-scale) ได้นั้นยังทำให้ง่ายต่อการนำไปใช้งานเกี่ยวกับสัญญาณทางโทรศัพท์หรือประยุกต์ใช้งานอย่างอื่นที่มีการจำกัดช่วงความถี่ รวมไปถึงเป็นการกำจัดช่วงความถี่สูงที่ส่วนมากเป็นช่วงความถี่ของเสียงรบกวน [9]

### 2.1.4 การสร้างตัวแบบผู้พูดและการระบุผู้พูด

เป็นการสร้างตัวแบบผู้พูดจากข้อมูลที่ได้จากการสกัดลักษณะเฉพาะของเสียงพูด โดยคำนึงถึงพื้นที่ในการจัดเก็บข้อมูลให้น้อยที่สุด มีความรวดเร็วในการค้นคืนสูง และสามารถให้ความถูกต้องเมื่อนำไปใช้งานมากที่สุด ซึ่งขั้นตอนการสร้างตัวแบบผู้พูดประกอบไปด้วยสองขั้นตอนคือขั้นตอนการสอน (Training) และขั้นตอนการทดสอบ (Testing) ซึ่งมีรายละเอียดดังนี้

#### *2.1.4.1 ขั้นตอนการสอน*

เป็นขั้นตอนในการสร้างตัวแบบจากข้อมูลลักษณะเฉพาะและเก็บลงในฐานข้อมูลของแต่ละคน โดยจะเรียกตัวแบบนั้นว่าโค๊ตบุ๊ค โดยปรกติแล้วข้อมูลเสียงพูดของแต่ละคนจะมีขนาดหรือจำนวนเวกเตอร์มากจึงได้มีเทคนิคในการแบ่งนับเวกเตอร์ (Vector Quantization) ก่อนทำการจัดเก็บ ในการ

ระบุผู้พูดนั้นจำเป็นต้องมีการค้นหาข้อมูลตัวแบบของผู้พูดเพื่อนำมาเปรียบเทียบกับข้อมูลเสียงพูดที่เข้ามาใหม่ ดังนั้นโครงสร้างของตัวแบบจึงเป็นสิ่งสำคัญต่อประสิทธิภาพของการระบุผู้พูด ถ้าตัวแบบในฐานข้อมูลมีปริมาณมากเวลาในการค้นหาก็จะมากด้วย และถ้าโครงสร้างนั้นมีโอกาสในการค้นหาผิดพลาด ความถูกต้องในการระบุผู้พูดจะน้อยลง

#### *2.1.4.2 ขั้นตอนการทดสอบ*

วิธีการทดสอบระหว่างตัวแบบกับข้อมูลเสียงที่เข้ามาใหม่ทำได้โดยการเปรียบเทียบความคล้ายคลึงกันระหว่างตัวแบบในฐานข้อมูลกับตัวแบบที่เข้ามาใหม่ ซึ่งตัวแบบของผู้พูดนั้นได้จากการแบ่งนับ ตัวแบบของแต่ละคนที่ถูกเก็บลงในฐานข้อมูลเรียกว่าโค๊ดบุ๊ค (Codebook) จากนั้นเมื่อเสียงที่ไม่รู้จักเข้ามาจะทำการสร้างตัวแบบด้วยวิธีการแบ่งนับและเปรียบเทียบตัวแบบใหม่กับทุกๆ โค๊ดบุ๊ค จากนั้นจะเลือกโค๊ดบุ๊คของคนที่ให้ค่าใกล้เคียงกับข้อมูลใหม่มากที่สุดออกมาเป็นคำตอบ

### 2.2 **ทฤษฎีการแบ่งนับเวกเตอร์** (Vector Quantization, VQ)

การแบ่งนับแบบเวกเตอร์ [10] เป็นทฤษฎีการแบ่งนับที่ถูกคิดค้นขึ้นมาเพื่อบีบอัดข้อมูลแบบสูญเสีย (Lossy) และต่อมาได้ถูกนำมาใช้ในการลดปริมาณข้อมูลเสียงพูดเพื่อใช้ในการระบุผู้พูดกันอย่างแพร่หลายเพราะวิธีการนี้มีประสิทธิภาพในการระบุผู้พูดมาก

สมมติลำดับของเวกเตอร์ลักษณะเฉพาะที่สกัดได้จากเสียงพูดหนึ่งๆ คือ $x=\{x_1,x_2,...,x_I\}$ โดย $x_i=< x_{i1}, x_{i2},\ldots,x_{iJ} >$ คือเวกเตอร์ของจำนวนจริงขนาด $J$ มิติ แนวคิดหลักของวิธีการแบ่งนับเวกเตอร์คือ หาเซตของเวกเตอร์ที่ดีที่สุดสำหรับเป็นตัวแทนเวกเตอร์ทั้งหมด เรียกว่า*โค๊ตบุค*, $C = \{y_1,y_2,...,y_M\}$, โดยที่ $y_m=< y_{m1}, y_{m2},\ldots,y_{mJ} >$ และ $M$ คือขนาดของโค๊ตบุค ซึ่งสามารถกำหนดขนาดเองได้ โดยโค๊ตบุคที่ได้มีค่าการบิดเบี้ยว (Distortion) น้อยที่สุดเท่าที่จะเป็นไปได้ ซึ่งคำนวณได้จากค่าเฉลี่ยของความแตกต่างระหว่าง $x_i$ กับ $y_m$ที่ใกล้กันมากที่สุด การหาระยะทางระหว่างเวกเตอร์สองเวกเตอร์ทำได้โดยการหาค่ายูคลิเดียนดิสแทนซ์ (Euclidian distance) โดยนิยามดังสมการที่ (10)

$$d(y_m, x_i) = \sqrt{\sum_{j=1}^{J} \left(y_{mj} - x_{ij}\right)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

ขั้นตอนวิธีในการจัดหมวดหมู่ (Clustering algorithm) ถูกนำมาใช้ในการหาเซตของโค๊ตบุค ขั้นตอนวิธีที่มีประสิทธิภาพและถูกนำมาใช้อย่างแพร่หลายคือแอลบีจี (Linde-Buzo-Grey, LBG) ซึ่ง

ประกอบไปด้วย การกำหนดค่าเริ่มต้น (Initialization) การวัดค่าการบิดเบี้ยว (Distortion) และการวนซ้ำ (Iteration)

ในการเปรียบเทียบระหว่างเสียงของผู้พูดที่ไม่รู้จักกับตัวแบบผู้พูดหรือโค๊ตบุคของคน $N$ คน เพื่อระบุผู้พูด ทำได้โดยการวัดค่าโค๊ตบุคที่มีความเหมือนมากที่สุด $C_{match}$ จากฐานข้อมูล $C_1, C_2, ..., C_N$ โค๊ตบุคตัวที่เหมือนมากที่สุดหาได้จากการวัดค่าความแตกต่างเฉลี่ยจากสมการ (11)

$$C_{match} = \arg\max\left\{\frac{1}{I}\sum_{i}^{I}\frac{1}{\min\{d(x_i, y_{n,m})\}}\right\} \qquad \text{..... (11)}$$

โดยที่ $y_{n,m}$ เป็นเวกเตอร์ของโค๊ตบุคที่ใกล้เคียงกับ $x_i$ มากที่สุดในโค๊ตบุค $C_n$

## 2.3 ทฤษฎีการแบ่งนับแบบฐานสอง (Binary Quantization, BQ)

การนำวิธีการแบ่งนับแบบฐานสอง [17] มาใช้ในการระบุผู้พูด ทำได้โดยการนำเวกเตอร์ลักษณะเฉพาะ $x=\{x_1, x_2, ..., x_I\}$ มาทำการสร้างตัวแบบผู้พูด เริ่มจากการหาค่าเวกเตอร์เฉลี่ยของลักษณะเฉพาะทั้งหมด $x_c$ จากสมการที่ (12)

$$x_{cj} = \frac{1}{I}\sum_{i=1}^{I} x_{ij} \ , \ 1 \le j \le J \qquad \text{..... (12)}$$

โดยที่ค่า $J$ คือจำนวนมิติของเวกเตอร์ลักษณะเฉพาะ $x_{cj}$ คือ พารามิเตอร์ลำดับที่ $j$ ของ $x_c$ และ $I$ คือจำนวนเวกเตอร์ทั้งหมดในข้อมูลชุดนี้ จากนั้นแบ่งเวกเตอร์ลักษณะเฉพาะทั้งหมดออกเป็น $T$ ส่วนตามแกนของเวลาและหาเวกเตอร์เฉลี่ยของแต่ละส่วนจะได้ $S = <s_1, s_2, s_3, ..., s_T>$ เราสามารถหาค่าเวกเตอร์แบบฐานสองของแต่ละส่วนได้โดยการนำเวกเตอร์ลักษณะเฉพาะเฉลี่ยของแต่ละส่วนมาทำการเปรียบเทียบกับเวกเตอร์ค่าเฉลี่ยรวม $x_c$ ดังสมการที่ (13) คือ

$$b_{tj} = \begin{cases} 1 & s_{tj} - x_{cj} \ge 0 \\ 0 & s_{tj} - x_{cj} < 0 \end{cases} \qquad 1 \le j \le J \qquad \text{..... (13)}$$

ดังนั้นเวกเตอร์ $s_t \ , \ 1 \le t \le T$ , สามารถเขียนให้อยู่ในรูปเวกเตอร์ฐานสองคือ $b_t = < b_{i1}, b_{i2}, b_{i3}, ..., b_{iJ} >$ และตัวแบบผู้พูดฐานสองคือ $S' = < b_1, b_2, b_3, ..., b_T >$ ในขั้นตอนระบุผู้พูด เสียงพูดที่ไม่รู้จักจะถูกนำมาทำการหา $S'$ ด้วยวิธีการเดียวกัน จากนั้น $S'$ ใหม่จะถูกนำไปเปรียบเทียบกับตัวแบบผู้พูดทุกตัวในฐานข้อมูลด้วยตัวดำเนินการ XOR เพื่อหาค่าความเหมือน

**2.4 งานวิจัยที่เกี่ยวข้อง**

งานวิจัยที่เกี่ยวข้องกับการระบุผู้พูดโดยใช้วิธีการแบ่งนับ (Quantization) นั้นมีอยู่หลายงานวิจัยด้วยกัน ซึ่งบางครั้งงานวิจัยในวิธีการนี้ถูกนำไปเปรียบเทียบกับวิธีการอื่น เช่น Hidden Macov Model (HMM), Guassian Mixture Model (GMM), Artificial Neuron Network (ANN), Support Vector Machine (SVM) และวิธีการอื่นๆ

**YUE Xicai YE Datian, LIU Ming. "Text-independent speaker identification by genetic clustering radial basis function neural network": 2001** [16] ได้นำวิธีการของ Genetic Clustering Radial Basis Function Neural Network (GC-RBFNN) มาใช้ในการระบุผู้พูดโดยไม่ขึ้นอยู่กับคำพูด โดยสร้างโมเดลเพื่อทำการรู้จำผู้พูดจำนวน 20 คน และทำการเปรียบเทียบกับวิธีการ RBFNN แบบธรรมดา โดยทำการทดสอบกับเสียงความยาวต่างกันพบว่า วิธีการแบบ GC-RBFNN ให้ความถูกต้องมากกว่า คือ สำหรับช่วงความยาวของเสียงเท่ากับ 5, 10, 15 วินาที วิธีการแบบ RNFNN ธรรมดาให้ความถูกต้องเท่ากับ 84%, 89% และ 92 % ตามลำดับ และเมื่อใช้วิธีการแบบ GC-RNFNN จะให้ความถูกต้องเท่ากับ 90%, 93% และ 95 % ตามลำดับ

**HOU Fenglei, WANG Bingxi. "Text-independent speaker recognition using support vector machine" : 2001** [4] เป็นการนำเอาวิธีการของ Support Vector Machine (SVM) มาใช้ในการจำแนกกลุ่มของข้อมูลลักษณะเฉพาะของเสียงพูดของผู้พูด การทดลองได้ทำการจำแนกกลุ่มของลักษณะเฉพาะด้วย Kernel Function แบบ Polynomial, Radial basis function (RBF) และ Sigmiod function ทำการจำแนกลักษณะเฉพาะของเสียงพูดที่ได้จากการสกัดด้วยวิธีการ Linear Predictive Cepstral Coefficients (LPCCs) จากนั้นจะนำผลการทดลองนี้ไปเปรียบเทียบกับวิธีการแบบ MLP Neural Network ซึ่งให้ผลลัพธ์คือ 91.4% สำหรับ SVM และ 90.8 % สำหรับ MLP

**Tomi Kinnunen, Teemu Kilpelainen and Pasi Franti. "The Comparison of Vector Quantization Algotithm": 2000** [8] งานวิจัยนี้ได้ทำการทดสอบประสิทธิภาพของขั้นตอนวิธีการแบ่งนับ (Vector Quantization Algorithm) 5 วิธี โดยการเปรียบเทียบด้วยประสิทธภาพความถูกต้องในการระบุผู้พูด และวัดค่า Mean Square Error (MSE) โดยขั้นตอนวิธีทั้ง 5 ได้แก่ Generalized Lloyd Algorithm (GLA), Self-Organizing Maps (SOM), Pairwise Nearest Neighbor (PNN), Iterative spitting technique (Split), Randomized Local Search (RLS) โดยผลการทดลองกับการระบุผู้พูดโดยขึ้นกับคำพูด จำนวน 25 คนพบว่า ความถูกต้องจะเพิ่มขึ้นเมื่อขนาดของ Codebook เพิ่มขึ้น และขั้นตอนวิธีแบบ GLA,SPLIT,PNN,RLS ให้ความถูกต้อง 100 % เมื่อขนาดของ Codebook เท่ากับ 64

และค่า MSE ที่น้อยที่สุดคือขั้นตอนวิธีแบบ RLS ซึ่งหมายถึงเป็นวิธีการที่ให้ความถูกต้องมากที่สุด แต่เป็นขั้นตอนวิธีที่ใช้เวลานานที่สุดเช่นกัน

**Douglas A. Reynolds and Richard C. Rose. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models": 1995** [14] เป็นงานวิจัยที่นำเอา Gaussian Mixture Model (GMM) ซึ่งเป็นโมเดลทางสถิติมาประยุกต์ใช้ในการระบุผู้พูด ซึ่งให้ประสิทธิภาพในการระบุผู้พูดมาก ในงานวิจัยของพวกเขาได้ชี้ให้เห็นถึงประสิทธิภาพในเชิงความถูกต้องเทียบกับวิธีการอีกหลาย ๆ วิธีการและ GMM สามารถให้ประสิทธิภาพสูงสุดในการระบุผู้พูดโดยไม่ขึ้นกับคำพูด เมื่อทดสอบกับผู้พูดจำนวน 16 คน คือ 94.5% โดยทดสอบกับเสียงพูดความยาว 5 วินาที

**ศวิต กาสุริยา, สมชาย จิตะพันธ์กุล, วิศรุต อาขุบุตร, เอกฤทธิ์ มณีน้อย และ พงศ์ไท ทาสระคู. "ระบบการบ่งชี้ผู้พูดแบบขึ้นกับบทคำพูดโดยใช้การวิเคราะห์เซ็พตรอล"** : **1999** [18] ได้ใช้วิธีการ HMM ในการสร้างตัวแบบผู้พูดเพื่อใช้ในการระบุผู้พูดโดยขึ้นกับคำพูด ด้วยลักษณะเฉพาะแบบ Linear Prediction Coefficients (LPC) ก่อนที่จะทำการสร้างตัวแบบของ HMM ลักษณะเฉพาะจะถูกนำมาทำการลดปริมาณข้อมูลโดยการแบ่งนับแบบเวกเตอร์ (Vector Quantization, VQ) ผลการวิจัยในการระบุผู้พูดจำนวน 12 คน ชาย 6 หญิง 6 โดยกำหนดขนาดของโค๊ดบุ๊กเท่ากับ 50, 100, 150 และ 200 ให้ความถูกต้องที่ 100%

# บทที่ 3

# การระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน

จุดประสงค์หลักของงานวิจัยนี้คือ การนำวิธีการแบ่งนับแบบสมสัณฐานมาใช้ในการสร้างตัวแบบผู้พูดเพื่อใช้ในการระบุผู้พูด ซึ่งมุ่งเน้นในด้านของความเร็ว พื้นที่ในการจัดเก็บ โดยยังคงความถูกต้องในการระบุผู้พูดมากที่สุด

*นิยาม :* ***การแบ่งนับแบบสมสัณฐาน*** *(Binary Isomorphic Quantization) หรือบีไอคิว หมายถึง การแบ่งนับข้อมูลด้วยวิธีการประมาณค่าของข้อมูลด้วยตัวแทนข้อมูลที่มีความละเอียดลดลง โดยการประมาณจะคำนึงถึงลักษณะความคล้ายกันทางกายภาพของข้อมูล*

เนื้อหาในบทนี้จะพูดถึงแรงจูงใจของการนำวิธีการแบ่งนับแบบสมสัณฐานมาใช้ในการระบุผู้พูด การวัดค่าความเมือนของเวกเตอร์ลักษณะเฉพาะ การแบ่งนับแบบสมสัณฐานแบบฐานสอง การแบ่งส่วนและจัดกลุ่ม การสร้างตัวแบบผู้พูด และการระบุผู้พูดจากเสียงพูดที่ไม่รู้จัก ซึ่งรายละเอียดจะได้กล่าวต่อไป

## 3.1 บทกล่าวนำ

เวกเตอร์ลักษณะเฉพาะแต่ละตัวที่สกัดได้จากเสียงพูดด้วยวิธีการวิเคราะห์เซ็พตรัล (กล่าวในบทที่ 2) เกิดจากการนำวินโดว์ไปครอบสัญญาณเสียงในช่วงหนึ่งๆ และนำไปทำการวิเคราะห์ โดยมีการเลื่อนวินโดว์ไปตามแกนของเวลา ซึ่งมีความเป็นไปได้สูงที่ลักษณะเฉพาะของวินโดว์ที่อยู่ติดกัน (ลักษณะเฉพาะข้างเคียง) จะมีค่าพารามิเตอร์คล้ายคลึงกันมากกว่าลักษณะเฉพาะที่ได้จากวินโดว์ที่อยู่ห่างออกไป ทั้งนี้เป็นเพราะว่าทั้งสองวินโดว์อาจมาจากเสียงพูดของคำๆ เดียวกันที่มีความถี่ใกล้เคียงกัน ดังนั้นเราพยายามที่จะทำการประเมินความคล้ายกันและความเปลี่ยนแปลงของลักษณะเฉพาะที่อยู่ใกล้เคียงกันเพื่อที่จะหาฟังก์ชันที่สามารถจำแนกความคล้ายกันของเวกเตอร์ลักษณะเฉพาะได้

จากที่ได้กล่าวมาข้างต้น เราได้ทำการทดลองโดยนำเวกเตอร์ลักษณะเฉพาะแต่ละตัวมาทำการวาดลงไปในกราฟ โดยให้แกนตั้งคือแกนของค่าพารามิเตอร์ของเวกเตอร์ลักษณะเฉพาะและแกนนอนคือค่าของจำนวนมิติของลักษณะเฉพาะ จากนั้นลากเส้นเชื่อมค่าในมิติที่อยู่ติดกันของแต่ละลักษณะเฉพาะเพื่อสามารถดูการเปลี่ยนแปลงค่าพารามิเตอร์ภายในของเวกเตอร์ลักษณะเฉพาะหนึ่งๆ ได้ง่ายขึ้น และเราทำเช่นนี้กับลักษณะเฉพาะทุกตัวที่สกัดได้จากเสียงพูดนั้น เราพบว่าลักษณะเฉพาะที่

อยู่ใกล้เคียงกันหรือมาจากช่วงของเสียงพูดที่มีความถี่ใกล้เคียงกัน มีการเปลี่ยนแปลงค่าภายในของลักษณะเฉพาะคล้ายคลึงกันมาก ซึ่งหมายถึงมีการเปลี่ยนแปลงรูปแบบของลักษณะเฉพาะคล้ายคลึงกัน ดังนั้นเราจะหาฟังก์ชันที่มีประสิทธิภาพในการแบ่งนับเวกเตอร์ลักษณะเฉพาะตามรูปแบบของมัน และเรียกการแบ่งนับนี้ว่า “การแบ่งนับแบบสมสัณฐาน” หรือ “ไอโซมอร์ฟิคควอนไทเซชัน” (Isomorphic Quantization) ซึ่งคำว่า “Iso” หมายความว่าเหมือนกัน และคำว่า “morph” หมายถึงรูปทรงหรือรูปร่าง ซึ่งฟังก์ชันในการวัดค่าความเหมือนกันของรูปร่างนี้จะเป็นหัวใจสำคัญของวิธีการนี้

แรงจูงใจอีกประการหนึ่งคือ วิธีการแบ่งนับแบบฐานสอง (Binary quantization) เป็นวิธีการแบ่งนับที่มีประสิทธิภาพในด้านของความเร็ว อีกทั้งยังสามารถลดขนาดตัวแบบผู้พูดลงได้มาก ซึ่งข้อดีทั้งหมดเกิดจากการนำเสนอรูปแบบของลักษณะเฉพาะที่ได้จากการแบ่งนับให้อยู่ในรูปแบบของเลขฐานสอง ดังนั้นวิธีแบ่งนับแบบสมสัณฐานจะใช้ข้อดีของการนำเสนอในรูปแบบของฐานสองนี้มาเพิ่มประสิทธิภาพให้กับวิธีการ

## 3.2 การวัดค่าความเหมือน

ความเหมือนของเวกเตอร์ลักษณะเฉพาะในวิธีการนี้ เป็นความเหมือนกันของรูปทรงหรือจะกล่าวอีกนัยหนึ่งว่า มีการเปลี่ยนแปลงค่าพารามิเตอร์ภายในเหมือนกัน ดังนั้นฟังก์ชันที่สามารถวัดการเปลี่ยนแปลงค่าภายใน 2 ฟังก์ชันคือ ฟังก์ชันวัดการเปลี่ยนแปลงความชันและฟังก์ชันวัดการเปลี่ยนแปลงความสูง จะถูกนำมาใช้ในการจำแนกความเหมือนกันของเวกเตอร์สองเวกเตอร์เพื่อทำการยุบเวกเตอร์ที่มีการเปลี่ยนแปลงเหมือนกันเข้าด้วยกัน

กำหนดให้ $x = \{x_1, x_2, x_3, \ldots, x_I\}$ คือเวกเตอร์ลักษณะเฉพาะทั้งหมดที่สกัดได้จากเสียงพูดหนึ่งๆ โดยที่ $I$ คือจำนวนลักษณะเฉพาะทั้งหมด

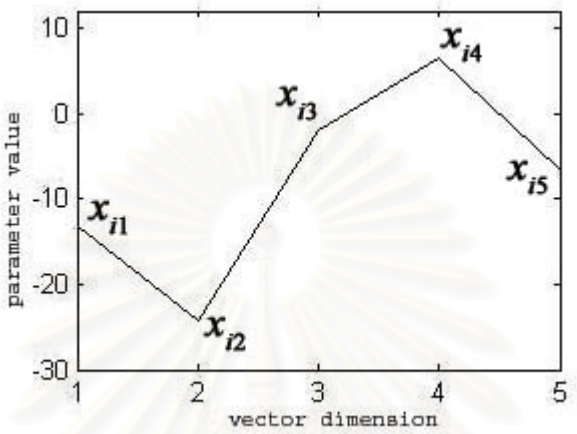
### *3.2.1 ฟังก์ชันวัดการเปลี่ยนแปลงความชัน (Curve changing function, $f_c$)*

อนุพันธ์ลำดับที่สองจะบ่งบอกถึงการเปลี่ยนแปลงความชันระหว่างก่อนและหลังจุดที่พิจารณา การเปลี่ยนแปลงความชันมี 2 สถานะและสามารถนำเสนอโดยใช้เลขฐานสองเพียงหนึ่งบิตเท่านั้น

เรากำหนดเวกเตอร์ลักษณะเฉพาะหนึ่งๆ $x_i = \langle x_{i1}, x_{i2}, \ldots, x_{iJ} \rangle$ โดยที่ $J$ คือจำนวนมิติของเวกเตอร์และ $x_{ij} \in \Re$, $1 \le j \le J$. จากรูปที่ 3.1 แสดงการวาดกราฟของลักษณะเฉพาะ $x_i$ ซึ่งมีจำนวนมิติเท่ากับ 5 โดยแกนนอนคือแกนของมิติและแกนตั้งคือแกนของค่าพารามิเตอร์$(x_{ij})$

วิธีการหาค่าการเปลี่ยนแปลงความชันของเส้นกราฟว่าจุดที่กำหนดมีการเปลี่ยนแปลงความชันในลักษณะที่เพิ่มขึ้นหรือลดลง ซึ่งค่านี้จะเป็นตัวบ่งบอกว่าจุดนั้นเป็นส่วนเว้าหรือส่วนนูนของเส้นกราฟ ตัวอย่างเช่น

จากรูปที่ 3.1 จากสมการความชัน : $m = \frac{\Delta y}{\Delta x} = \Delta y$ เนื่องจาก $\Delta x = 1$ ทุกกรณี




รูปที่ 3.1 การวาดกราฟของเวกเตอร์ลักษณะเฉพาะหนึ่งตัว ซึ่งมีจำนวนมิติเท่ากับ 5 โดยที่มี $x_{i1}, x_{i2,} x_{i3,}$ $x_{i4}$ และ $x_{i5}$ เป็นค่าในแต่ละมิติที่เราพิจารณา

การเปลี่ยนแปลงความชันที่จุด $x_{ij}$ คือ

$$f_c(x_{ij}) = \begin{cases} 1 & (x_{i,j+1} - x_{i,j}) - (x_{i,j} - x_{i,j-1}) \geq 0 \\ 0 & Otherwise \end{cases}$$

ในแต่ละมิติของลักษณะเฉพาะหนึ่งๆ จะถูกนำมาทำการหาค่าเปลี่ยนแปลงความชันด้วยฟังก์ชัน $f_c$ ยกเว้นจุดแรกกับจุดสุดท้าย ด้วยเหตุนี้เวกเตอร์ใหม่ที่ได้จะมีมิติ ลดลงจากเดิม 2 มิติ จากรูปที่ 3.1 เราสามารถหาการเปลี่ยนแปลงความชันจากสมการข้างต้นโดยเริ่มจากจุด $x_{i2}$ ไปจนถึงจุด $x_{i4}$ ผลลัพธ์ที่ได้จะเป็นลำดับของเลขฐานสอง คือ $<f_c(x_{i2}), f_c(x_{i3}), f_c(x_{i4})> = <1, 0, 0>$

*3.2.2 ฟังก์ชันวัดการเปลี่ยนแปลงความสูง (Height changing function, $f_h$)*

ฟังก์ชันวัดการเปลี่ยนแปลงความสูงเป็นวิธีการเปรียบเทียบด้วยตัวแปรชุดเดียวกันกับ $f_c$ โดยจะเป็นการพิจารณาความยาวของเส้นตรงของมิติทางซ้ายและขวาของมิติที่พิจารณาแทน

จากรูปที่ 3.1 การหาค่าความยาวของเส้นตรงระหว่างจุด 2 จุด $x_{i,j}$ และ $x_{i,j+1}$ สามารถคำนวณได้จากสมการวัดระยะทางแบบยูคลิเดียน $d(x,y) = \sqrt{\Delta x^2 + \Delta y^2}$ และ $\Delta x = 1$ ทุกกรณี

ดังนั้น

ระยะทาง $d(x_{i1}, x_{i2}) = \sqrt{(x_{i2} - x_{i1})^2 + 1^2}$

ระยะทาง $d(x_{i2}, x_{i3}) = \sqrt{(x_{i3} - x_{i2})^2 + 1^2}$

การเปลี่ยนแปลงความสูงสามารถหาได้จากการเปรียบเทียบค่าระยะทางของเส้นตรงสองเส้นที่อยู่ติดกัน ในตัวอย่างนี้คือ $d(x_{i1},x_{i2})$ และ $d(x_{i2},x_{i3})$ ถ้า $d(x_{i2},x_{i3}) > d(x_{i1},x_{i2})$ แล้ว

$$d(x_{i2}, x_{i3}) = \sqrt{(x_{i3} - x_{i2})^2 + 1^2} \quad > \quad d(x_{i1}, x_{i2}) = \sqrt{(x_{i2} - x_{i1})^2 + 1^2}$$

ยกกำลังสองทั้งสองข้างจะได้ $(x_{i3} - x_{i2})^2 + 1^2 > (x_{i2} - x_{i1})^2 + 1^2$

$= (x_{i3} - x_{i2})^2 + \cancel{1^2} > (x_{i2} - x_{i1})^2 + \cancel{1^2}$

$= (x_{i3} - x_{i2})^2 > (x_{i2} - x_{i1})^2$

ใส่รากที่สองทั้งสองข้างจะได้ $| x_{i3} - x_{i2} | > | x_{i2} - x_{i1} |$ (*อสมการนี้จะช่วยให้การเปรียบเทียบเร็วขึ้น*)
ดังนั้นฟังก์ชันวัดการเปลี่ยนแปลงความสูงที่จุด $x_{ij}$ คือ

$$f_h(x_{ij}) = \begin{cases} 1 & | x_{i,j+1} - x_{i,j} | - | x_{i,j} - x_{i,j-1} | \geq 0 \\ 0 & Otherwise \end{cases}$$

ทุกมิติของเวกเตอร์ลักษณะเฉพาะจะถูกนำมาทำการพิจารณาเพื่อหาค่าการเปลี่ยนแปลงความสูง ยกเว้นมิติแรกกับมิติสุดท้าย ผลลัพธ์สุดท้ายของการหาค่าการเปลี่ยนแปลงความสูงของเวกเตอร์ลักษณะเฉพาะในรูปที่ 3.1 จะได้ลำดับของเลข ฐานสองคือ $<f_h(x_{i2}), f_h(x_{i3}), f_h(x_{i4})> = <1, 0, 1>$

*3.2.3 ไอโซมอร์ฟิคฟังก์ชัน (Isomorphic function)*

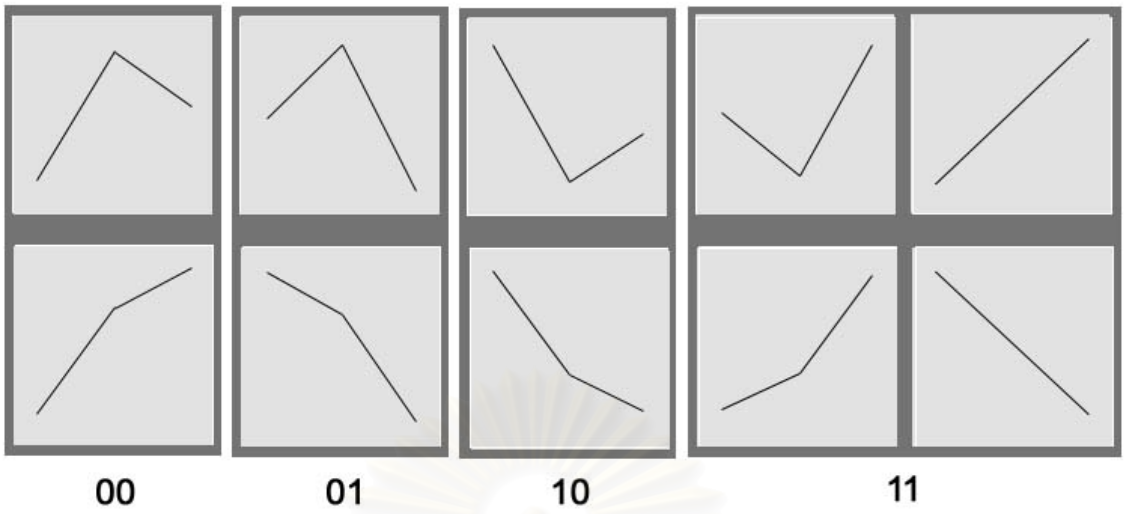
เป็นฟังก์ชันที่รวมฟังก์ชัน $f_c$ และ $f_h$ เข้าด้วยกันเพื่อใช้เป็นฟังก์ชันสำหรับวัดความเหมือนของเวกเตอร์ลักษณะเฉพาะ ซึ่งสามารถแบ่งแยกรูปแบบของลักษณะเฉพาะได้หลายรูปแบบ เราสามารถหาค่าของฟังก์ชัน $f_c$ และ $f_h$ ไปพร้อมๆ กันในทุกๆ มิติของแต่ละเวกเตอร์ลักษณะเฉพาะได้ โดยแต่ละมิติของลักษณะเฉพาะสามารถแสดงด้วยค่าเลขฐานสอง 2 บิตซึ่งมีค่าที่เป็นไปได้ 4 รูปแบบคือ

“00” บ่งบอกถึงลักษณะของเส้นกราฟตรงจุดนั้นมีลักษณะคว่ำ และเส้นที่สองสั้นกว่าเส้นแรก

“01”บ่งบอกถึงลักษณะของเส้นกราฟตรงจุดนั้นมีลักษณะคว่ำ และเส้นที่สองยาวกว่าเส้นแรก

“10” บ่งบอกถึงลักษณะของเส้นกราฟตรงจุดนั้นมีลักษณะหงาย และเส้นที่สองสั้นกว่าเส้นแรก

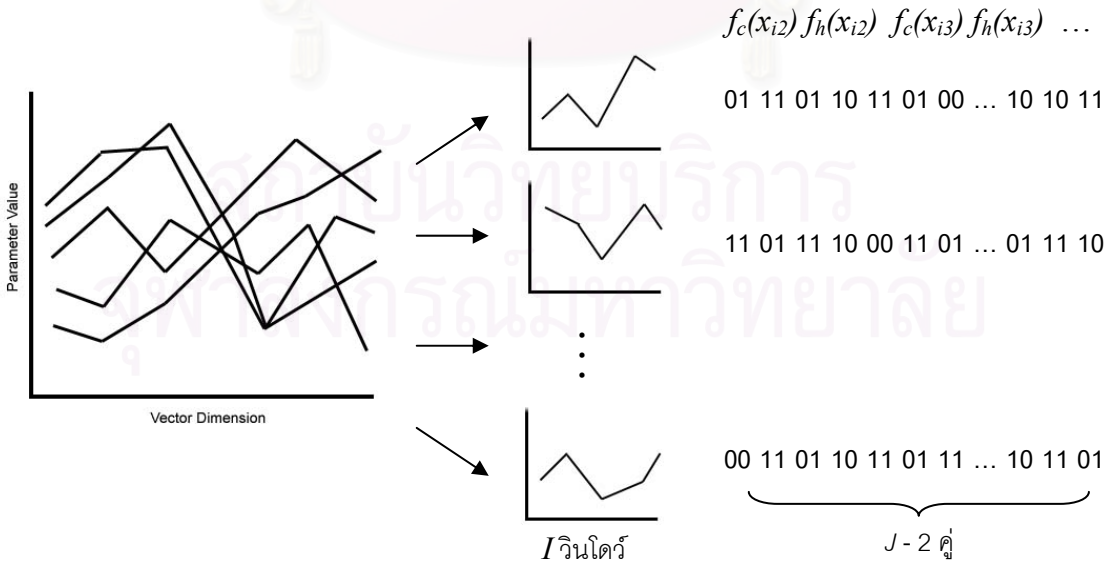
“11” บ่งบอกถึงลักษณะของเส้นกราฟตรงจุดนั้นมีลักษณะหงายและเส้นที่สองยาวกว่าเส้นแรก




รูปที่ 3.2 แสดงรูปแบบของเวกเตอร์ลักษณะเฉพาะที่ฟังก์ชันไอโซมอร์ฟิคสามารถจำแนกได้ โดยจะแทนรูปแบบทั้งสี่แบบด้วยเลขฐานสอง 2 บิต

โดยแต่ละมิติของลักษณะเฉพาะจะถูกแบ่งนับด้วยไอโซมอร์ฟิคฟังก์ชันยกเว้นมิติแรกและมิติสุดท้าย ดังนั้นผลลัพธ์สุดท้ายจะได้ลำดับของเลขฐานสองความยาว $(J-2) \times 2$ บิต โดยที่ $J$ คือจำนวนมิติของลักษณะเฉพาะ ตัวอย่าง ในรูปที่ 3.1 ค่าผลลัพธ์ที่ได้จากการแปลงเวกเตอร์ $x_i$ ด้วยไอโซมอร์ฟิคฟังก์ชันคือ $<f_c(x_{i2}), f_h(x_{i2}), f_c(x_{i3}), f_h(x_{i3}), f_c(x_{i4}), f_h(x_{i4})> = <1,1,0,0,0,1>$

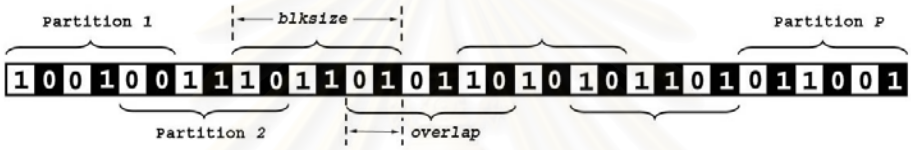
*ข้อสังเกต* ฟังก์ชันในการวัดการเปลี่ยนแปลงทั้งสองมีการคำนวณที่คล้ายกันคือ $(x_{i2} - x_{i1})$ และ $(x_{i3} - x_{i2})$ ซึ่งสามารถทำการคำนวณเพียงครั้งเดียวเท่านั้น




รูปที่ 3.3 การแบ่งนับโดยใช้ไอโซมอร์ฟิคฟังก์ชันในแต่ละลักษณะเฉพาะ

### 3.3 การแบ่งส่วนและการจัดกลุ่ม

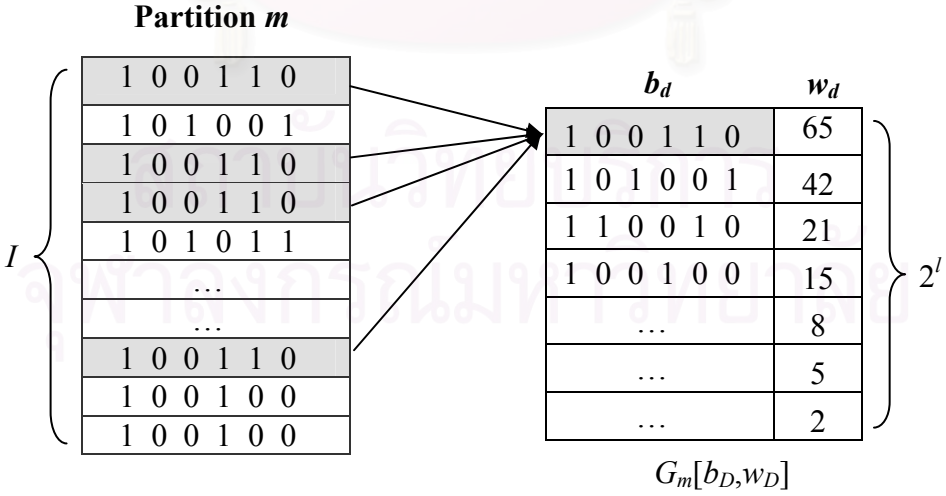
เนื่องจากในขั้นตอนการจัดกลุ่ม (Clustering) จะเป็นการพิจารณาความเหมือนกันของบิตในแต่ละมิติ ค่าเวกเตอร์ฐานสองแต่ละตัวที่ได้จากไอโซมอร์ฟิคฟังก์ชันจะถูกนำมาเปรียบเทียบความเหมือนแบบบิตต่อบิตและทำการจัดกลุ่มเวกเตอร์ที่มีลักษณะเหมือนกันให้อยู่กลุ่มเดียวกัน แต่ถ้าจำนวนมิติเวกเตอร์ลักษณะเฉพาะมีค่ามากขึ้น เวกเตอร์แต่ละตัวจะมีความเหมือนกันน้อยมาก ดังนั้นวิธีการแบ่งส่วน (Partitioning) สำหรับเวกเตอร์ทุกตัวตามมิติของมันจะทำให้สามารถจัดกลุ่มเวกเตอร์ได้ง่ายขึ้น จากนั้นจะทำการจัดกลุ่มในแต่ละส่วนแทนการจัดกลุ่มเวกเตอร์ทั้งหมด ตัวอย่างเช่นเวกเตอร์ฐานสอง $x_i$ มีจำนวนมิติเท่ากับ $J$ สามารถทำการแบ่งส่วนออกเป็น $P$ ส่วน แต่ละส่วนมี $l$ บิตและมีการซ้อนทับเท่ากับ $k$ บิต เนื่องจากบิตข้างเคียงแต่ละคู่มีความสัมพันธ์กัน ซึ่งแสดงในรูปที่ 3.4




รูปที่ 3.4 การแบ่งเวกเตอร์ฐานสองออกเป็น $P$ ส่วน โดยมีขนาดของแต่ละส่วนเท่ากับ $l$ บิต และมีการซ้อนทับเท่ากับ $k$ บิต ระหว่างส่วนที่อยู่ติดกัน

ตำแหน่งของบิตที่ 1 ของแต่ละส่วนคือ $p_m$ สามารถคำนวณได้จากสมการ

$$p_m = m(l\text{-}k) - (l\text{-}k\text{-}1), \text{ โดยที่ } m = 1,2,\ldots,P$$




รูปที่ 3.5 แสดงการจัดกลุ่มของเวกเตอร์ฐานสองในแต่ละส่วน (Partition) โดยค่า $b_d$ คือเวกเตอร์ฐานสองและ $w_d$ คือค่าจำนวนรูปแบบที่ซ้ำกันของเวกเตอร์ฐานสองหรือเรียกว่าค่าน้ำหนัก

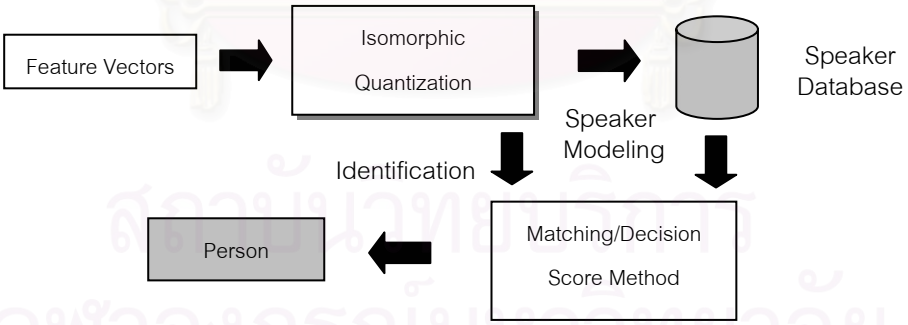
สำหรับทุกเวกเตอร์ฐานสองของเสียงพูดหนึ่งๆ จะถูกนำมาทำการแบ่งส่วนออกเป็น $P$ ส่วน และแต่ละส่วนจะถูกนำมาทำการจัดกลุ่มตามความเหมือน โดยที่จำนวนสมาชิกของเวกเตอร์ที่อยู่ในกลุ่มเรียกว่าน้ำหนัก (Weight) ซึ่งแต่ละส่วนที่ถูกจัดกลุ่มแล้วนิยามโดย $G_m[b_D,w_D]$, $D = 2^l$, $b_d$ คือ เวกเตอร์ฐานสองและ $w_d$ เป็นค่าน้ำหนักของเวกเตอร์นั้นๆโดยที่ $1 \leq d \leq D$ จากนั้นจะนำไปใช้ในขั้นตอนการสร้างตัวแบบผู้พูดต่อไป

## 3.4 การสร้างตัวแบบผู้พูดและการระบุผู้พูด

ในส่วนนี้จะกล่าวถึงโครงสร้างในการระบุผู้พูด ขั้นตอนการสร้างตัวแบบผู้พูดจากลักษณะเฉพาะแบบฐานสองที่ได้จากการแบ่งส่วนและจัดกลุ่ม $(G_m[b_d,w_d])$ และวิธีการในการเปรียบเทียบระหว่างตัวแบบผู้พูดกับเสียงพูดที่ไม่รู้จักเพื่อทำการระบุผู้พูด

### 3.4.1 โครงสร้างระบบระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน

หลักการสำคัญในการระบุผู้พูดคือ การเปรียบเทียบข้อมูล 2 ชุด ระหว่างตัวแบบผู้พูด (Speaker model) กับข้อมูลเสียงพูดที่ไม่รู้จัก (Unknown speech) เพื่อวัดระดับความคล้ายกันของข้อมูลทั้งสองและตัดสินว่าเสียงพูดที่ไม่รู้จักนั้นตรงกับตัวแบบของใครในฐานข้อมูล ซึ่งโครงสร้างหลักของการระบุผู้พูดแบบสมสัณฐานจะแสดงดังรูปที่ 3.6




รูปที่ 3.6 โครงสร้างของการระบุผู้พูดด้วยวิธีการแบ่งนับแบบสมสัณฐาน

จากรูปที่ 3.6 แสดงถึงโครงสร้างโดยรวมประกอบด้วยขั้นตอนการสร้างตัวแบบผู้พูดหรือการสอน (Speaker modeling: Training) และขั้นตอนการระบุผู้พูดหรือการทดสอบ (Speaker Identification: Testing) ซึ่งทั้งสองขั้นตอนนี้จะต้องผ่านขั้นตอนการสกัดลักษณะเฉพาะเป็นอันดับแรก กรณีของการสอนลักษณะเฉพาะจะถูกป้อนเป็นข้อมูลเข้าไปยังขั้นตอนการแบ่งนับแบบสมสัณฐานเพื่อให้ได้ตัวแบบของเสียงพูดนั้น ๆ และเก็บลงในฐานข้อมูลเป็นตัวแบบผู้พูด ในกรณีที่เป็นการทดสอบ

เสียงพูดที่ไม่รู้จักจะถูกนำมาทำการแบ่งนับด้วยวิธีการเดียวกันนี้ก่อนนำไปเปรียบเทียบกับตัวแบบผู้พูดทั้งหมดในฐานข้อมูล

### 3.4.2 การสร้างตัวแบบผู้พูด

ตัวแบบผู้พูดของคนหนึ่งคน สามารถสร้างได้ทั้งจากเสียงพูดหนึ่งเสียงหรือจากเสียงพูดหลายเสียงพูด และหนึ่งคนสามารถมีตัวแบบผู้พูดได้มากกว่าหนึ่งตัวแบบได้ ในงานวิจัยนี้ได้กำหนดให้ผู้พูดหนึ่งคนมีเพียงหนึ่งตัวแบบผู้พูดเท่านั้นและตัวแบบผู้พูดของแต่ละคนจะสร้างจากเสียงพูดประโยคเดียวกันจำนวน 5 ประโยค

ในการสร้างตัวแบบผู้พูดเป็นขั้นตอนที่รับค่าข้อมูลเข้าที่ได้จากขั้นตอนการแบ่งส่วนและการจัดกลุ่ม คือ $G_m[b_D\ ,w_D]$ ซึ่งในขณะโปรแกรมกำลังดำเนินการจะแสดงอยู่ในรูปของอาร์เรย์สองมิติ ซึ่งในขั้นตอนการสร้างตัวแบบผู้พูดนี้จะสามารถทำได้อย่างรวดเร็วเพราะสามารถอ้างอิงค่าในอาร์เรย์ทุกตัวได้โดยตรงผ่านค่าเลขดัชนีของอาร์เรย์และ $G_m[b_D\ ,w_D]$ มีขนาดที่แน่นอน

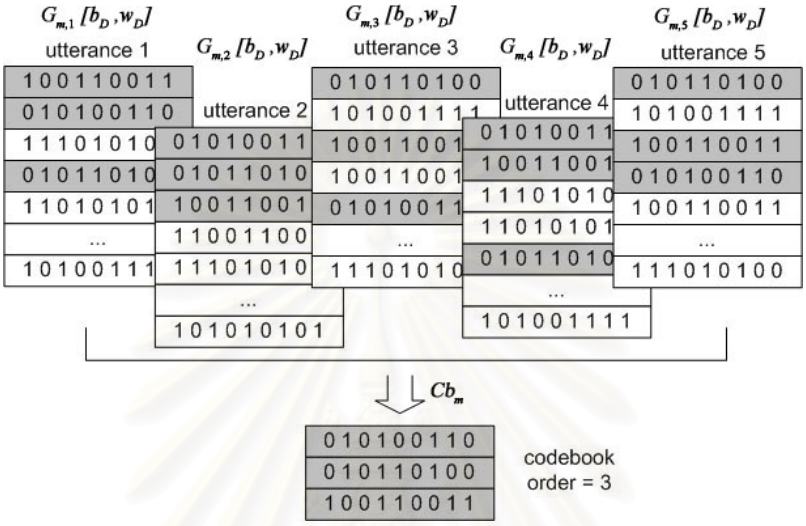
ในกรณีที่เราใช้เสียงพูดประโยคเดียวกัน $\delta$ ประโยคมาใช้ในการสร้างตัวแบบผู้พูด เวกเตอร์ลักษณะเฉพาะแต่ละตัวของทุกๆ เสียงพูดจะถูกแบ่งออกเป็น $P$ ส่วน โดยแต่ละส่วนมีขนาดเท่ากับ $l$ บิต และ $D = 2^l$ มีขั้นตอนวิธีดังนี้

Input: $\forall G_{m,r}[b_D\ ,w_D]\ ,\ 1 \le r \le \delta$
Output: $Cb$

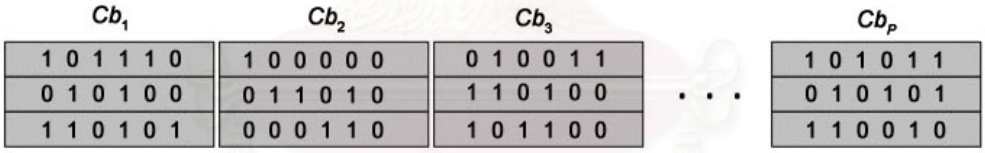
001: for $m$ = 1 to $P$ do
002: for $r$ = 1 to $\delta$ do
003: for $e$ = 1 to $D$ do
004: $votes_{m,r}$ [val($b_e$)]=$votes_{m,r}$ [val($b_e$)]+1
005: $wg_{m,r}$ [val($b_e$)]=$wg_{m,r}$ [val($b_e$)]+$w_e$
006: end
007: end
008: for $i$ = 1 to $2^l$ do
009: $score_m[i] = votes_m[i] \times wg_m[i]$
010: end
011: [$index$] = max($cr$, $score_m$)
012: $Cb = Cb$ . bin($index$)
013: end

โดยที่ $votes$, $wg$ และ $score$ เป็นตัวแปรอาร์เรย์ขนาด $2^l \times 1$ มิติ ฟังก์ชัน val() ใช้แปลงจากเลขฐานสองให้เป็นเลขฐานสิบเพื่อนำไปใช้เป็นเลขดัชนีอ้างอิงของอาร์เรย์ ในบรรทัดที่ 011 ฟังก์ชัน max() ใช้เพื่อหาเลขดัชนีของอาร์เรย์ $score$ ที่มากที่สุด $cr$ ลำดับ และบรรทัดที่ 012 จะเป็นการแปลงค่าเลขดัชนีของอาร์เรย์เป็นค่าเวกเตอร์ฐานสองและนำมาต่อกัน (.) ตามแกนมิติของเวกเตอร์

ในกระบวนการสร้างตัวแบบผู้พูด $Cb$ คือค่าของตัวแบบผู้พูดและเราเรียก $cr$ ว่า *ลำดับตัวแบบผู้พูด* (Codebook order) ในกระบวนการระบุผู้พูด $Cb$ จะสร้างจากเสียงพูดที่ไม่รู้จักเพียงหนึ่งเสียงเท่านั้นและจะเรียก $cr$ ว่า *ลำดับข้อมูลทดสอบ* (Testing order) ซึ่งต่างก็มีผลกระทบต่อความถูกต้องในการระบุผู้พูดของวิธีการบีไอคิวซึ่งจะได้กล่าวต่อไป




รูปที่ 3.7 การสร้างตัวแบบผู้พูดย่อย $Cb_m$ จากเสียงพูด 5 เสียงโดยมีลำดับตัวแบบผู้พูดเท่ากับ 3




รูปที่ 3.8 ตัวแบบผู้พูดที่มีลำดับเท่ากับ 3 ซึ่งเกิดจากการต่อกัน (Concatenate) ตั้งแต่ $Cb_1$ ถึง $Cb_P$

### 3.4.3 การระบุผู้พูด

การพิสูจน์ตัวจริงของผู้พูด (Speaker authentication) แบ่งออกเป็น 2 วิธีการคือ การตรวจสอบว่าเสียงพูดใหม่ที่นำเข้ามาตรวจสอบเป็นเสียงของคนที่อยู่ในฐานข้อมูลหรือไม่ ซึ่งจะเรียกวิธีการนี้ว่าการการทวนสอบผู้พูด (Speaker verification) และการบ่งบอกได้ว่าเสียงพูดนั้นเป็นเสียงของใคร เรียกว่าการระบุผู้พูด (Speaker identification) ซึ่งการระบุผู้พูดนั้นจะเป็นวิธีการค้นหาเสียงพูดที่คล้ายกับเสียงพูดที่นำเข้ามาตรวจสอบมากที่สุด ดังนั้นต้องให้แน่ใจว่าเสียงพูดที่นำเข้ามา

ตรวจสอบนี้ เป็นเสียงของคนที่อยู่ในฐานข้อมูลนั้นจริงๆ หรือเรียกอีกอย่างหนึ่งว่าเป็น *ระบบปิด* (Close system) และจะให้คำตอบที่ใกล้เคียงที่สุดอย่างน้อง 1 คน [1]

ในการเปรียบเทียบความเหมือนทำได้โดยการวัดค่าความคล้ายโดยจะบ่งบอกโดย *แต้ม (score)* ระหว่างคู่ของตัวแบบผู้พูดที่เข้ามาใหม่กับตัวแบบผู้พูดในฐานข้อมูลแต่ละตัว โดยใช้ตัวดำเนินการฐานสองคือ exclusive OR ถ้าค่าแต้มมากจะหมายถึงความคล้ายกันของข้อมูลทั้งสองมีน้อย และถ้าค่าแต้มน้อยจะหมายถึงค่าความคล้ายกันของข้อมูลทั้งสองมีมาก ซึ่งการที่จะตัดสินใจว่าเสียงพูดที่เข้ามาใหม่นี้เป็นเสียงของใคร ทำได้โดยการเลือกตัวแบบแบบผู้พูดที่ให้ค่า *score* น้อยที่สุดออกมาเป็นคำตอบ

กำหนดให้ $C = \{C_1, C_2, C_3, \ldots, C_N\}$ คือตัวแบบของผู้พูด $N$ คนที่เก็บอยู่ในฐานข้อมูล ซึ่งมีลำดับของตัวแบบผู้พูดเท่ากับ $R$ และตัวแบบของเสียงพูดใหม่ที่ต้องการตรวจสอบคือ $Y$ ซึ่งสร้างจากเสียงพูดที่มีเวกเตอร์ลักษณะเฉพาะทั้งหมด $I$ ตัวและมีลำดับข้อมูลทดสอบเท่ากับ $Z$ การเลือกตัวแบบที่ให้ค่าความเหมือนมากที่สุดสามารถอธิบายได้โดยสมการ

$$C_{match} = \arg\min_{1 \le n \le N} \left\{ \frac{\sum_{z=1}^{Z} \min_{1 \le r \le R} (C_{n,r} \oplus Y_z)}{I} \right\}$$

$\oplus$ = Exclusive OR operation

# บทที่ 4

# การทดลองและผลการทดลอง

ในส่วนของการทดลองจะเป็นการวัดประสิทธิภาพและผลกระทบในการระบุผู้พูดโดย เนื้อหาส่วนแรกจะอธิบายถึงข้อมูลที่ใช้ในการทดสอบและวิธีการวัดประสิทธิภาพของวิธีการบีไอคิว ส่วนที่สองแสดงถึงผลการทดลองเกี่ยวกับผลกระทบต่อประสิทธิภาพของวิธีการรวมถึงผลการทดลองเพื่อเปรียบเทียบประสิทธิภาพของวิธีการนี้กับอีกสองวิธีการคือ การแบ่งนับเวกเตอร์แบบแอลบีจีและวิธีการแบ่งนับแบบฐานสอง

## 4.1 การออกแบบฐานข้อมูลและการวัดประสิทธิภาพ

เราได้ทำการเก็บรวบรวมเสียงพูดจากบุคคลต่างเพศต่างวัยจำนวน 30 คนที่มีอายุระหว่าง 16-43 ปี แต่ละคนอ่านบทพูดที่กำหนดให้ 4 บทพูด บทพูดละ 6 ครั้ง ซึ่งแต่ละครั้งที่พูดจะแตกต่างกัน จากนั้นวิเคราะห์เซ็พสตรัลด้วยวิธีการเอ็มแอฟซีซี (MFCC) เพื่อสกัดลักษณะเฉพาะจำนวน 18 มิติต่อหนึ่งวินโดว์

ในงานวิจัยนี้มีการทดสอบกับข้อมูล 4 ชุดดังนี้

ชุดที่หนึ่ง เสียงพูดตัวเลขเดี่ยวคำว่า “เก้า” ความยาวประมาณ 0.6-1.5 วินาที

ชุดที่สอง เสียงพูดตัวเลขต่อเนื่อง คำว่า “ศูนย์ หนึ่ง สอง สาม สี่ ห้า” ความยาวประมาณ 2-5 วินาที

ชุดที่สาม เสียงพูดประโยคสนทนาต่อเนื่อง คำว่า “ขอผ่านเข้าสู่ระบบ รหัสผ่าน สี่ สอง เก้า เจ็ด” ความยาวประมาณ 4-6 วินาที

ชุดที่สี่ เสียงพูดประโยคสนทนาต่อเนื่อง คำว่า “ขอออกจากระบบ รหัสผ่าน เจ็ด สี่ สอง หนึ่ง สาม สอง หก ศูนย์” ความยาวประมาณ 6-8 วินาที

ซึ่งข้อมูลทั้งสี่ชุดที่นำมาทำการทดสอบถูกกำจัดเสียงเงียบเรียบร้อยแล้วด้วยวิธีการกำหนดค่าขีดแบ่ง (Amplitude threshold) และเสียงพูดชุดที่ 3 และชุดที่ 4 เป็นเสียงพูดต่อเนื่องหมายถึงประโยคที่เปล่งออกมาโดยไม่มีการเว้นวรรค ซึ่งเป็นลักษณะการพูดโดยทั่วไปของคนเรา

การวัดประสิทธิภาพจะใช้วิธีการครอสวาลิเดชัน (Cross validation) [6] โดยนำข้อมูลเสียงพูดแต่ละคนจำนวน 6 ครั้งมาเป็นข้อมูล $T_1, T_2, \ldots, T_6$ จากนั้นนำ 5 กลุ่มย่อยมาทำการสร้างตัวแบบผู้พูดและอีกหนึ่งกลุ่มมาใช้ในการทดสอบ ซึ่งทำให้แต่ละคนมีการทดสอบทั้งหมด 6 ครั้งและ

แต่ละครั้งจะทดสอบทำกับคนจำนวน 30 คน ซึ่งความถูกต้องของการทดสอบแต่ละครั้งสามารถหาได้จากสมการ

$$\%Correction\,(T_i) = \frac{num\,of\ matching(0-30)}{num\,of\ register\,speaker}$$

ค่าความถูกต้องเฉลี่ยโดยรวมในการระบุผู้พูดได้จากการหาค่าเฉลี่ยของเปอร์เซ็นต์ความถูกต้องทั้ง 6 ครั้ง

### 4.2 ขั้นตอนและวิธีการทดสอบ

เวกเตอร์ลักษณะเฉพาะจำนวนมิติเท่ากับ 18 จะถูกทำการแบ่งนับแบบสมสัณฐาน ในการแบ่งส่วนและการจัดกลุ่ม โดยจะแบ่งเวกเตอร์ลักษณะเฉพาะออกเป็น 7 ส่วน แต่ละส่วนมีจำนวนบิตเท่ากับ 6 บิตและมีการซ้อนทับเท่ากับ 2 บิต จากนั้นหา $G_m(b_D,w_D)$ ในแต่ละส่วนและนำไปทำการสร้างตัวแบบผู้พูดตามขั้นตอนวิธีในหัวข้อ 3.4.2

ในกระบวนการทดสอบ เราจะทำการวัดผลกระทบที่มีผลต่อประสิทธิภาพของการระบุผู้พูด ด้วยชุดข้อมูลเสียงพูดที่มีความยาวแตกต่างกัน รวมถึงการเปรียบเทียบประสิทธิภาพกับวิธีการแบ่งนับแบบอื่น

#### 4.2.1 ผลกระทบจากขนาดของการแบ่งส่วน

จากผลการทดลองสรุปได้ว่า เราไม่สามารถหาค่าที่ตายตัวในการกำหนดจำนวนบิตของแต่ละส่วน ($l$) ความถูกต้องในการระบุผู้พูดจะแตกต่างกันถ้าขนาดแตกต่างกัน ทั้งนี้ขึ้นอยู่กับปัจจัยดังนี้

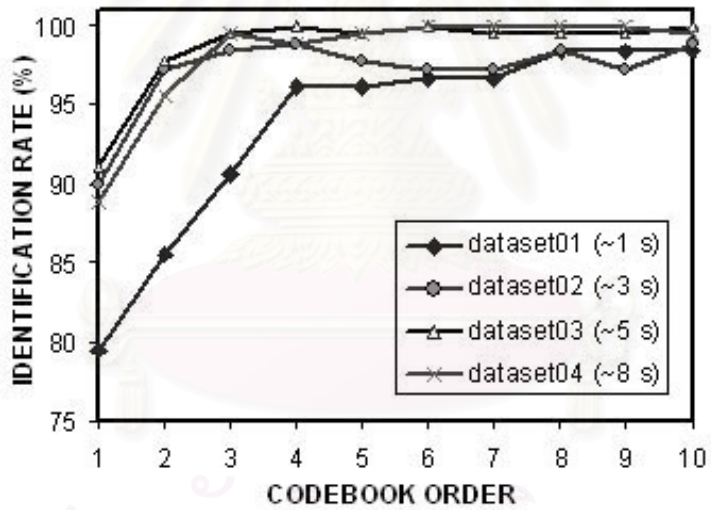
1) *จำนวนของเวกเตอร์ลักษณะเฉพาะที่นำมาสร้างตัวแบบ ($I$)*
   จุดประสงค์หลักของการจัดกลุ่มเวกเตอร์ที่มีลักษณะเหมือนกันเข้าไว้ในกลุ่มเดียวกัน วิธีการแบ่งกลุ่มทำเพื่อให้เวกเตอร์ที่มีปริมาณจำกัดมีการซ้ำกันมากขึ้น ถ้าขนาดของแต่ละส่วนมีขนาดใหญ่ จะเหมาะสมกับเสียงพูดที่ให้จำนวนเวกเตอร์ลักษณะเฉพาะมากๆ นั่นก็คือเสียงพูดที่มีความยาวมากๆ นั่นเองและเมื่อการนำไปใช้งานกับเสียงที่มีความยาวสั้นๆ ขนาดของส่วนควรจะมีขนาดเล็กลง
2) *จำนวนชุดของข้อมูลที่นำมาทำการสร้างตัวแบบผู้พูด ($\delta$)*
   เช่นเดียวกันเหตุผลข้อแรก การเพิ่มจำนวนชุดข้อมูลที่นำมาใช้สร้างตัวแบบก็เปรียบเสมือนการเพิ่มจำนวนเวกเตอร์ลักษณะเฉพาะ แต่วิธีการนี้มีส่วนที่เพิ่มเติมคือ

ถ้าชุดข้อมูลที่นำมาสร้างตัวแบบมาจากเสียงที่มีความถี่ใกล้เคียงกันจะมีเวกเตอร์ที่มีความคล้ายกันมาก ซึ่งการจัดกลุ่มก็ทำได้โดยง่าย

การพิจารณาความเหมาะสมในการจัดกลุ่มนั้น สามารถพิจารณาได้จากความซ้ำกันของเวกเตอร์ในแต่ละกลุ่ม ยิ่งซ้ำกันมากยิ่งดี และพิจารณาจากค่าน้ำหนักของแต่ละกลุ่มซึ่งเป็นตัวบ่งบอกว่า ถ้าน้ำหนักของกลุ่มที่มีปริมาณเวกเตอร์มากสุดแตกต่างจากกลุ่มอื่นมากๆ ย่อมจะหมายถึงเวกเตอร์กลุ่มนั้น เป็นตัวแทนที่ดีของเสียงพูดนั้นๆ

### 4.2.2 ผลกระทบจากอันดับของตัวแบบผู้พูด

ในการทดลองแรกเป็นการทดลอบถึงผลกระทบของอันดับตัวแบบผู้พูดหรือขนาดตัวแบบผู้พูด ว่ามีผลกระทบต่อความถูกต้องในการระบุผู้พูดอย่างไร เราได้ทำการทดสอบด้วยวิธีการที่ได้กล่าวมาแล้วข้างต้น ผลลัพธ์ที่ได้จะแสดงดังรูปที่ 4.1



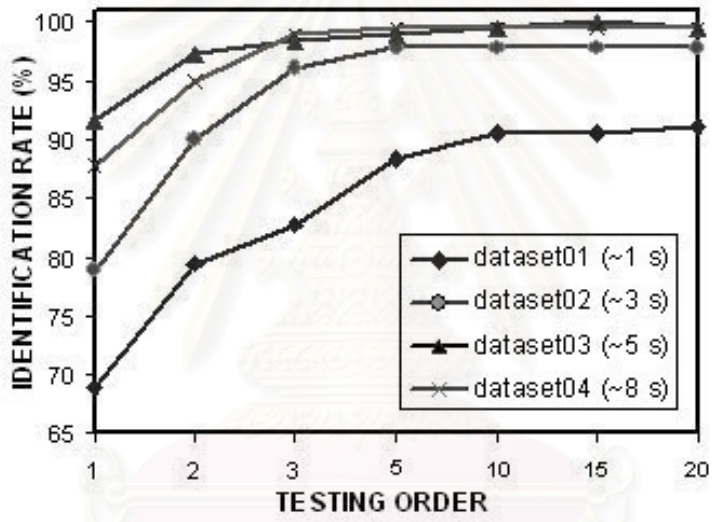

รูปที่ 4.1 ผลกระทบของอันดับตัวแบบผู้พูดที่เพิ่มขึ้นจาก 1 – 10 ในแต่ละช่วง ขนาดตัวแบบเพิ่มขึ้น 42 บิต โดยการทดลองนี้มีอันดับข้อมูลทดสอบเท่ากับ 10 และทดสอบกับข้อมูลทดสอบชุดที่ 1 ถึงชุดที่ 4 ซึ่งมีความยาวเสียงพูดประมาณ 1 วินาที 3 วินาที 5 วินาที และ 8 วินาที ตามลำดับ

จากรูปที่ 4.1 ความถูกต้องจะเพิ่มขึ้นอย่างรวดเร็วและจะเริ่มคงตัวที่อันดับตัวแบบผู้พูดเท่ากับ 3 ขึ้นไป ความถูกต้องในการระบุผู้พูดที่อันดับ 3 ($42 \times 3 = 126$ บิต) เท่ากับ 99.45 % ในชุดข้อมูลทดสอบที่ 3 ถึงแม้ว่าอันดับที่สูงขึ้นไปสามารถที่จะทำให้ความถูกต้องมากขึ้น แต่ก็ไม่มากนักเมื่อเทียบกับขนาดตัวแบบผู้พูดที่เพิ่มขึ้นตามมา ถึงแม้ว่าเราไม่สามารถกำหนดค่าที่ตายตัวของ

อันดับของตัวแบบผู้พูดสำหรับทุกๆ ชุดข้อมูลเสียงพูดได้ แต่จากการทดลองนี้สรุปได้ว่าเสียงพูดต่อเนื่องที่มีความยาวเพิ่มขึ้นจะทำให้ความถูกต้องเพิ่มขึ้น ความถูกต้องของการทดสอบในแต่ละชุดข้อมูลจะเริ่มคงตัวที่อันดับตัวแบบผู้พูดเท่ากับ 3 ดังนั้นเราสามารถนำค่าอันดับตัวแบบผู้พูดเท่ากับ 3 เป็นค่าโดยปริยาย (Default) ในการสร้างตัวแบบผู้พูดสำหรับการทดสอบครั้งต่อไปได้

### 4.2.3 ผลกระทบจากอันดับข้อมูลทดสอบ

ในส่วนที่สองของการทดลองจะแสดงถึงผลกระทบของอันดับข้อมูลทดสอบต่อความถูกต้องในการระบุผู้พูดโดยใช้ข้อมูลชุดเดียวกับการทดลองแรก ผลการทดลองจะแสดงดังรูปที่ 4.2




รูปที่ 4.2 ผลกระทบของอันดับข้อมูลทดสอบต่อความถูกต้องในการระบุผู้พูด จาก 1-20 โดยใช้อันดับตัวแบบผู้พูดเท่ากับ 3 และทดสอบกับข้อมูลทดสอบชุดที่ 1 ถึงชุดที่ 4

ในช่วงของอันดับข้อมูลทดสอบ 1 ถึง 10 ความถูกต้องจะเพิ่มขึ้นอย่างรวดเร็วและจะเริ่มคงที่ในอันดับที่สูงขึ้นไป ในการทดสอบนี้เราได้ทำการกำหนดอันดับตัวแบบผู้พูดเท่ากับ 3 ซึ่งจากผลการทดลองกับชุดข้อมูลเสียงที่มีความยาวที่สุด (ชุดที่ 3 และชุดที่ 4) จะให้ความถูกต้องในการระบุผู้พูดเท่ากับ 100% และ 99.45% ตามลำดับ ที่อันดับข้อมูลทดสอบเท่ากับ 15

แม้ว่าอันดับของข้อมูลทดสอบในอันดับสูง สามารถทำให้ความถูกต้องในการระบุผู้พูดเพิ่มมากขึ้น แต่ยังคงขึ้นอยู่กับความต่อเนื่องและความยาวของเสียงพูดด้วย เสียงพูดคำเดี่ยว (ชุดที่ 1) จะให้ความถูกต้องน้อย แม้ว่าอันดับของข้อมูลทดสอบจะเพิ่มขึ้นก็ตาม

### 4.2.4 การเปรียบเทียบกับวิธีการแบ่งนับวิธีการอื่น

การทดลองสุดท้ายจะพูดถึงการเปรียบเทียบประสิทธิภาพของวิธีการแบ่งนับแบบสมสัณฐานกับวิธีการแบ่งนับแบบอื่น จุดประสงค์หลักของการทดลองนี้คือ เพื่อที่จะแสดงให้เห็นถึงประสิทธิภาพในด้านของความถูกต้องในการระบุผู้พูด ขนาดของตัวแบบผู้พูด และความเร็วในการทดสอบ โดยทั้งสามวิธีการใช้ข้อมูลทดสอบชุดเดียวกัน

วิธีการแบ่งนับเวกเตอร์ ขั้นตอนวิธีแบบแอลบีจี ซึ่งถูกนำมาใช้กันอย่างแพร่หลายและถูกใช้เป็นเกณฑ์เปรียบเทียบสมรรถนะ (Benchmark) [8] เราทำการทดลองด้วยขนาดโค๊ดบุคเท่ากับ 4 และมีการวนซ้ำหนึ่งครั้ง สำหรับวิธีการแบ่งนับแบบฐานสอง (Binary Quantization) เราแบ่งส่วนเท่ากับ 12 ซึ่งเป็นค่าที่ให้ความถูกต้องมากที่สุดในการทดสอบเมื่อคำนึงถึงขนาดของตัวแบบผู้พูดด้วย ซึ่งทั้งสองวิธีการข้างต้นจะถูกนำมาทำการเปรียบเทียบกับวิธีการแบ่งนับแบบสมสัณฐานที่จำกัดขนาดของตัวแบบผู้พูดที่ประมาณ 16 ไบท์ ซึ่งเป็นค่าโดยปริยายของวิธีการนี้ (อันดับตัวแบบผู้พูดเท่ากับ 3 และอันดับข้อมูลทดสอบเท่ากับ 15)

ตารางที่ 4.1 แสดงความถูกต้องในการระบุผู้พูดในแต่ละวิธี

| วิธีการ | อัตราความถูกต้อง (%) | | | | ค่าเฉลี่ย |
|---|---|---|---|---|---|
| | ชุดที่ 1 | ชุดที่ 2 | ชุดที่ 3 | ชุดที่ 4 | |
| | (~1 sec.) | (~3 sec.) | (~5 sec.) | (~8 sec.) | |
| BIQ | 90.55 | 97.78 | 100 | 99.45 | 96.95 |
| BQ | 91.11 | 99.45 | 98.33 | 98.89 | 96.95 |
| VQ-LBG | 98.89 | 99.45 | 97.78 | 100 | 99.03 |

จากตารางที่ 4.1 แสดงให้เห็นว่า เมื่อความยาวของเสียงพูดเพิ่มขึ้นโดยเฉพาะอย่างยิ่งประโยคต่อเนื่องในชุดที่ 2 3 และ 4 ความถูกต้องของการระบุผู้พูดจะเพิ่มมากขึ้นและถูกต้อง 100% ในข้อมูลทดสอบชุดที่ 3 ซึ่งเราสามารถสรุปได้ว่าวิธีการบีไอคิวสามารถให้ความถูกในการระบุผู้พูดมากขึ้นเมื่อมีข้อมูลที่นำมาใช้สร้างตัวแบบผู้พูดมากเพียงพอหรือเมื่อทำการทดสอบกับประโยคต่อเนื่องที่มีความยาวเพียงพอ (ตัวแบบผู้พูด 1 ตัวแบบสร้างจาก 6 เสียงพูด และเสียงพูดที่มีความยาว 3 ถึง 8 วินาที จะทำให้วิธีการนี้มีประสิทธิภาพในการระบุผู้พูดมากขึ้น)

ตารางที่ 4.2 แสดงขนาดของตัวแบบผู้พูดโดยประมาณของแต่ละวิธีด้วยข้อมูลชุดที่ 1 ถึง ชุดที่ 4

| **วิธีการ** | **ขนาดของตัวแบบผู้พูด (BYTES)** |
|---|---|
| BIQ | 16 |
| BQ | 27 |
| VQ-LBG | 288 |

นอกจากนี้เรายังได้ทำการวัดประสิทธิภาพของวิธีการในด้านขนาดตัวแบบผู้พูด ในตารางที่ 4.2 แสดงถึงค่าโดยประมาณของขนาดตัวแบบผู้พูดของแต่ละวิธี ซึ่งวิธีการบีไอคิวให้ขนาดตัวแบบผู้พูดเล็กกว่าอีก 2 วิธีการ ในการทดลองนี้เราได้ทำการกำหนดขนาดตัวแบบผู้พูดของอีกสองวิธีการโดย

- กำหนดขนาดตัวแบบผู้พูดของวิธีการแบ่งนับเวกเตอร์แบบแอลบีจีให้มีขนาดเท่ากับ 4 เพราะว่าที่ขนาดตัวแบบผู้พูดลำดับต่อไป (เท่ากับ 8) จะให้ความถูกต้องมากที่สุดในทุกชุดข้อมูลทดสอบ (เฉลี่ยประมาณ 99.98%) แต่ขนาดตัวแบบผู้พูดจะเพิ่มขึ้นเป็นสองเท่าของขนาดตัวแบบผู้พูดเท่ากับ 4
- กำหนดจำนวนการแบ่งส่วนของวิธีการแบ่งนับแบบฐานสองเท่ากับ 12 เนื่องจากการทดลองบ่งบอกว่าเป็นค่าที่ความถูกต้องเริ่มคงตัว ถึงแม้ว่าการแบ่งส่วนให้มากขึ้นจะทำให้ความถูกต้องเพิ่มขึ้นแต่ก็น้อยมากเมื่อเทียบกับขนาดตัวแบบผู้พูดที่เพิ่มขึ้น (เพิ่มขึ้นประมาณ 2 ไบต์ต่อ 1 ส่วน)

ในด้านของความเร็วในการสร้างตัวแบบผู้พูด วิธีการแอลบีจีนั้นมีการคำนวณที่ซับซ้อนกว่าอีกสองวิธีการอย่างเห็นได้ชัด ซึ่งทำให้ใช้เวลาในการระบุผู้พูดมากที่สุด ส่วนวิธีการบีไอคิวและบีคิว การสร้างตัวแบบผู้พูดสามารถทำได้อย่างรวดเร็ว และในการระบุผู้พูด ถึงแม้ว่าวิธีการบีคิวจะใช้เวลาน้อยที่สุดแต่ก็ไม่แตกต่างกันมากนัก เพราะทั้งสองวิธีการอยู่บนพื้นฐานของเลขฐานสองและใช้ตัวดำเนินการพื้นฐานทั้งหมด

ตารางที่ 4.3 แสดงความเร็วในการสร้างตัวแบบผู้พูดในรูปของสัญกรณ์เชิงปริมาณ (บิกโอ)

| **วิธีการ** | **รายละเอียดของบิกโอ** | **คลาส** |
|---|---|---|
| BIQ | $O(4n) + O(\#P \times n) + O(\#P \times c \times \log(c))$ | $O(n)$ |
| BQ | $O(2n)+O(\log(n))$ | $O(n + \log(n))$ |
| VQ | N/A | $O(n \log(n))$ |

จากตารางที่ 4.3 ค่าที่แสดงในส่วนของรายละเอียดคำนวณได้จากแต่ละขั้นตอนของวิธีการแบ่งนับ โดยที่ค่า $n$ คือจำนวนลักษณะเฉพาะทั้งหมดที่ใช้สร้างตัวแบบผู้พูด $c$ คือค่าคงที่ และ $\#P$ คือจำนวนการแบ่งส่วนของวิธีการบีไอคิว (ในการทดลองนี้ $\#P = 7$) เมื่อเปรียบเทียบกับวิธีการบีคิวแล้ว ในด้านรายละเอียดของวิธีการ วิธีการบีไอคิวจะใช้ตัวดำเนินการมากกว่า ส่งผลให้ความเร็วในการสร้างตัวแบบผู้พูดช้ากว่าวิธีการบีคิว แต่เมื่อเปรียบเทียบความเร็วโดยรวมแล้ว ทั้งวิธีการบีคิวและบีไอคิวนั้นอยู่ในคลาสเดียวกัน ซึ่งหมายความว่า ถึงแม้ว่าข้อมูลที่ใช้ในการสร้างตัวแบบผู้พูดจะมากขึ้นแต่ทั้งสองวิธีจะมีประสิทธิภาพในด้านของความเร็วใกล้เคียงกัน

ตารางที่ 4.4 แสดงความเร็วในการเปรียบเทียบตัวแบบผู้พูดในรูปของสัญกรณ์เชิงปริมาณ (บิกโอ)

| **วิธีการ** | **รายละเอียดของบิกโอ** | **คลาส** |
|---|---|---|
| BIQ | $O(c \times \#cr \times \#P)$, $\#cr = 3$, $\#P = 7$ | $O(c)$ |
| BQ | $O(c \times \#T)$, $\#T = 12$ | $O(c)$ |
| VQ | $O(\#cbsize \times \#cbsize)$, $\#cbsize = \text{fixed}$ | $O(c)$ |

นอกจากนั้น เรายังได้ทำการเปรียบเทียบความเร็วในการระบุผู้พูดเมื่อมีเสียงพูดที่ไม่รู้จักเข้ามา จากตารางที่ 4.4 จำนวนตัวดำเนินการสามารถคำนวณได้จากค่าพารามิเตอร์ของแต่ละวิธีการได้แก่ $\#cr$ คืออันดับข้อมูลทดสอบของวิธีการบีไอคิว $\#T$ คือจำนวนการแบ่งส่วนของวิธีการบีคิว (รายละเอียดในบทที่ 2) และ $\#cbsize$ คือขนาดของตัวแบบผู้พูดของวิธีการแอลบีจี ซึ่งทั้ง 3 วิธีมีการกำหนดค่าพารามิเตอร์ที่ตายตัว และผลลัพธ์ของการเปรียบเทียบความเร็วโดยรวม สรุปได้ว่าทั้งสามวิธีการมีความเร็วในการระบุพูดอยู่ในคลาสเดียวกัน

# บทที่ 5

# สรุปการวิจัยและข้อเสนอแนะ

## 5.1 สรุป

งานวิจัยนี้ได้เสนอแนวคิดในการแบ่งนับแบบสมสัณฐานเพื่อใช้ในการระบุผู้พูด ซึ่งเป็นวิธีการที่เกิดจากแนวคิดในการหาฟังก์ชันเพื่อวัดค่าการเปลี่ยนแปลงรูปแบบของลักษณะเฉพาะและนำเสนอให้อยู่ในรูปเลขฐานสอง จากนั้นทำการเลือกลักษณะเฉพาะที่สามารถเป็นตัวแทนของลักษณะเฉพาะทั้งหมด ซึ่งผลลัพธ์คือทำให้ขนาดของตัวแบบผู้พูดมีขนาดเล็กกว่าวิธีการแบ่งนับแบบฐานสองและวิธีการแบ่งนับเวกเตอร์แบบแอลบีจีที่ให้ความถูกต้องในการระบุผู้พูดระดับเดียวกัน ความเร็วในการค้นหาและเปรียบเทียบเป็นไปอย่างรวดเร็ว โดยที่วิธีการบีไอคิวมีค่าสัญกรณ์เชิงปริมาณ (บิกโอ) อยู่ในคลาสเดียวกับวิธีการบีคิว นอกจากนี้ความถูกต้องในการระบุผู้พูด เมื่อเราทดสอบกับข้อมูล 4 ชุด สามารถสรุปได้ว่า วิธีการแบ่งนับแบบสมสัณฐานที่มีการกำหนดค่าพารามิเตอร์สำหรับการแบ่งส่วนเท่ากับ 6 บิต สามารถให้ความถูกต้องโดยรวมมาก เมื่อทดสอบกับข้อมูลเสียงพูดต่อเนื่องที่มีความยาวระหว่าง 3 ถึง 8 วินาที (ข้อมูลชุดที่ 2 3 และ 4) และให้ความถูกต้องเฉลี่ยถึงร้อยละ 99.73 เมื่อทดสอบกับเสียงพูดที่มีความยาว 5 ถึง 8 วินาที (ข้อมูลชุดที่ 3 และ 4) ซึ่งวิธีการนี้เป็นทางเลือกอีกทางหนึ่งในการระบุผู้พูดที่ต้องการความเร็วในการระบุผู้พูดและโดยเฉพาะอย่างยิ่งต้องการจำกัดพื้นที่จัดเก็บตัวแบบผู้พูด

## 5.2 ข้อเสนอแนะ

### 5.2.1 สภาวะแวดล้อมของการบันทึกเสียง

ดังที่ได้กล่าวมาตอนต้นบท เสียงรบกวนมีผลอย่างมากต่อความถูกต้องของระบบ ตัวแบบผู้พูดที่สร้างจากเสียงพูดภายใต้สภาวะแวดล้อมแบบหนึ่งจะมีประสิทธิภาพน้อยลงเมื่อนำมาทดสอบกับอีกภาพวะแวดล้อมอีกแบบหนึ่ง เช่น ตัวแบบผู้พูดที่สร้างจากเสียงพูดในสภาพแวดล้อมห้องทำงาน (Office environment) ซึ่งมีเสียงรบกวน เช่น เสียงแอร์ เสียงพัดลม จะแตกต่างกับตัวแบบที่ที่สร้างจากเสียงพูดภาพใต้สภาพแวดล้อมกลางแจ้ง (Outdoor environment) ซึ่งจะมีเสียงรบกวนหลายรูปแบบมากกว่า

ดังนั้นการกำหนดสภาพวะแวดล้อมของเสียงพูดที่จะใช้สร้างตัวแบบและสภาพวะแวดล้อมในการทดสอบหรือใช้งาน ต้องมีความใกล้เคียงกันหรือถ้าหลีกเลี่ยงไม่ได้ ควรจะเตรียมวิธีการในการลดเสียงรบกวนที่อาจเกิดขึ้นภายใต้สภาวะแวดล้อมนั้นๆ

### 5.2.2 การเลือกวิธีการสกัดลักษณะเฉพาะแบบอื่น

การสกัดลักษณะเฉพาะจากเสียงพูดมีหลายวิธีการ เช่น MFCC, LPC (Linear Predictive Coefficients), LPCC (Linear Predictive Cepstral Coefficients), LSP (Linear Spectrum Pair) เป็นต้น ซึ่งมีประสิทธิภาพแตกต่างกัน มีความทนทางต่อเสียงรบกวนต่างกัน ทั้งนี้ขึ้นอยู่กับขั้นตอนวิธีในการรู้จำหรือขั้นตอนวิธีในการสร้างตัวแบบผู้พูดด้วย

นอกจากวิธีการสกัดลักษณะเฉพาะแบบ MFCC ที่นำมาใช้กับงานวิจัยนี้ วิธีการสกัดลักษณะเฉพาะแบบอื่นก็เป็นทางเลือกที่เราสามารถนำมาใช้กับขั้นตอนวิธีนี้ได้ และประสิทธิภาพในการระบุผู้พูดจะแตกต่างกันในแต่ละวิธีการ

### 5.2.3 การประมวลผลขั้นต้นและการกำหนดค่าพารามิเตอร์ในการสกัดลักษณะเฉพาะ

ข้อมูลเข้า (Input data) เป็นสิ่งที่สำคัญเป็นอย่างยิ่งต่อการรู้จำและการทดสอบ และจะส่งผลต่อความถูกต้องของการระบุผู้พูด การประมวลผลขั้นต้นเป็นการทำให้ข้อมูลดิบกลายเป็นข้อมูลที่เหมาะสมในการนำไปใช้มากยิ่งขึ้น ในงานวิจัยนี้ผู้วิจัยได้นำขั้นตอนการ Pre-emphasis, การลดสัญญาณรบกวน, การตัดเสียงเงียบ มาใช้ในการเตรียมข้อมูลก่อนจะผ่านไปยังขั้นตอนการสกัดลักษณะเฉพาะ ไม่จำเป็นว่าจะเป็นขั้นตอนเหล่านี้เสมอไป เราสามารถทำการเพิ่มเติมหรือตัดขั้นตอนที่ไม่จำเป็นออกได้ ตามความเหมาะสม

ในขั้นตอนการสกัดลักษณะเฉพาะ เป็นวิธีการวิเคราะห์ความถี่ของเสียงพูด ซึ่งจะประกอบด้วยวิธีการย่อยภายในมากมาย การกำหนดค่าพารามิเตอร์สำหรับแต่ละวิธีแตกต่างกัน ทำให้ได้รูปแบบของลักษณะเฉพาะแตกต่างกันและจะมีผลต่อประสิทธิภาพของวิธีการแบ่งนับแบบสมสัณฐานด้วยเช่นกัน

### 5.2.4 การนำเสนอผู้พูดหนึ่งคนด้วยหลายตัวแบบผู้พูด

แน่นอนว่าตัวแบบผู้พูดในวิธีการนี้ถูกสร้างจากเสียงพูดภายใต้ความถี่หนึ่งๆ ดังนั้นเสียงพูดของคนๆ เดียวกันถึงแม้จะเป็นประโยคเดียวกันแต่พูดคนในช่วงเวลาแตกต่างกันหรือผู้พูดคนนั้นไม่

สบายในเวลาต่อมาจะทำให้ตัวแบบที่ได้แตกต่างกันด้วย ดังนั้นการสร้างตัวแบบผู้พูดของคนๆ เดียวกันหลายๆ ตัวแบบจากหลายๆ เสียงพูดที่แตกต่างกัน และเก็บลงในฐานข้อมูล จะเป็นการเพิ่มความถูกต้องในการระบุผู้พูดมากยิ่งขึ้น

### 5.2.5 การจัดเก็บโดยใช้โครงสร้างต้นไม้

ในงานวิจัยนี้การระบุผู้พูดจะเป็นการเปรียบเทียบข้อมูลระหว่างตัวแบบผู้พูดในฐานข้อมูลและตัวแบบผู้พูดใหม่ที่นำมาทดสอบ โดยวิธีการเปรียบเทียบแบบเชิงเส้น (Linear) ซึ่งจะต้องทำการเปรียบเทียบทุกๆ ตัวแบบ และเวลาในการระบุผู้พูดจะเพิ่มขึ้นเมื่อตัวแบบในฐานข้อมูลมากขึ้น ดังนั้นการจัดเก็บแบบต้นไม้จะช่วยให้การค้นคืนรวดเร็วขึ้นมาก แต่วิธีการจัดเก็บและค้นคืนก็จะยากขึ้นเช่นเดียวกัน

# รายการอ้างอิง


[1] J.P. Cambell, “Speaker recognition: A tutorial,” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 289-292,1997.

[2] G.R. Doddington, “*The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results, Perspective*,” Nation Institute of Standards and Technology (NIST) Gaithersburg USA, 1998.

[3] Z. Fang, W. Wenhu and F. Ditang, “A Log-Index Weighted Cepstral Distance Measure for Speech Recognition”, *Proceedings EI*, 12(2):177-184, 1997.

[4] H. Fenglei and W. Bingxi, ”Text-independent speaker recognition using support vector machine,”*Proceedings ,ICII 2001-Beijing*, Volume: 3, 29 Oct.-1 November, 2001

[5] S.Furui, *“Digital speech processing, synthesis, and recognition*,” 2nd Edition, Revised and Expanded; Handbook of Signal processing and communication, ISBN 0-8247-0452-5, 2001.

[6] T. Mitchell, “*Machine Learning*,” Handbook of Artificial Intelligence, New York; ISBN 0-07-042807-7, 1997.

[7] T. Kinnunen, I. Kärkkäinen, P. Fränti, "Is speech data clustered?-Statistical analysis of cepstral features," *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, vol. 4, pp. 2627-2630, Aalborg, Denmark, 2001.

[8] T. Kinnunen, T. Kilpeläinen, P. Fränti, "Comparison of clustering algorithms in speaker identification", *Proc. IASTED Int. Conf Signal Processing and Communications (SPC 2000)*, pp. 222-227, Marbella, Spain, 2000.

[9] J. Koolwaaij, “*Automatic Speaker Verification in Telephony: a probabilistic approach - Cepstral Analysis*”, website: http://www.ispeak.nl/prfhtm/node12.html, 2001.

[10] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.,* vol. 28, pp. 84–95, January, 1980.

[11] T.Matsui, T.Kanno and S.Furui, “Speaker Recognition Using HMM composition in noisy Environments,” *Proceedings Eurospeech-95*, Madrid, pp.621-624, 1995.

[12] M. Norris, “*Process of Sound Magic FX - Remove DC Offset*”, website: http://mnorris.wellington.net.nz/soundmagic/effects/Remove DC Offset.html, 2001.

[13] P.Premakanthan, W.B. Mikhael, “Speaker Verification/Recognition and The Importance of Selective Feature Extraction:REVIEW,” *Proc. IEEE Int. Acoustic, Speech, and Signal Processing*, 2002.

[14] D.A. Reynolds, R.C.Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Model,” *IEEE Transaction on Speech and Audio Processing*, Volumn: 3, NO. 1, January 1995.

[15] T.Robinson, “*Speech Analysis*,” The tutorial of signal processing, Cambridge University, England, 1998.

[16] Y. Xicai, Y. Datian and L. Ming, ”Text-independent speaker identification by genetic clustering radial basis function neural network,” *Proceedings of the 23rd Annual International Conference of the IEEE*, Volume: 2, 25-28 October, 2001.

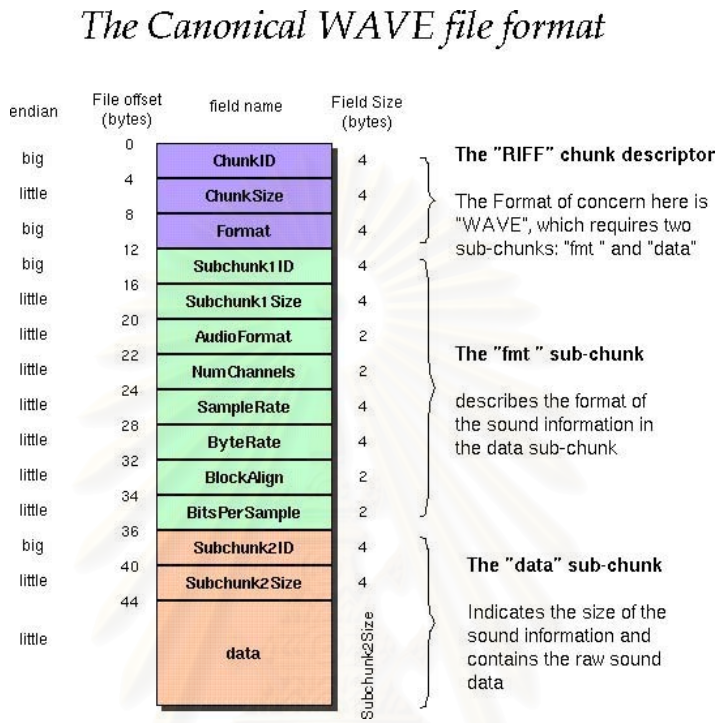
[17] Z. Yuan, B. Xu, and C.Yu, “Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification,” *IEEE Transaction on Speech and Audio Processing*, Volumn: 7, NO. 1, January 1999.

[18] ศวิต กาสุริยา, สมชาย จิตะพันธ์กุล, วิศรุต อาขุบุตร, เอกฤทธิ์ มณีน้อย และ พงศ์ไท ทาสระคู, “ระบบการบ่งชี้ผู้พูดแบบขึ้นกับบทคำพูดโดยใช้การวิเคราะห์เซ็พตรอล,” *การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 22*, 2-3 ธันวาคม 2542.

# ภาคผนวก

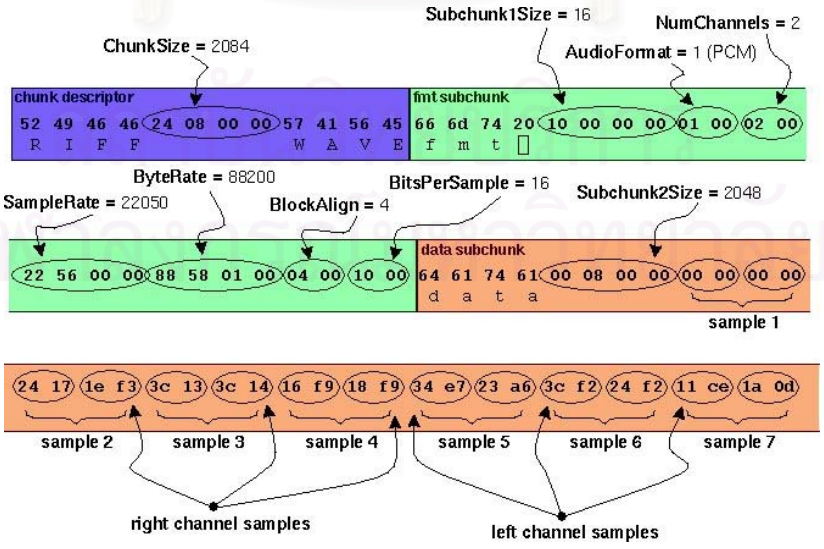
# ภาคผนวก ก

รูปแบบโครงสร้างของไฟล์เสียงนามสกุล .wav




ตัวอย่างข้อมูล

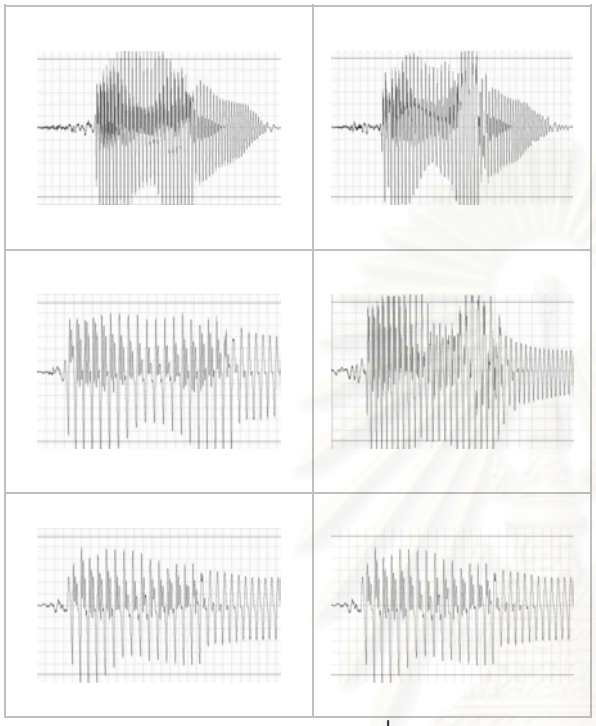
**52 49 46 46 24 08 00 00 57 41 56 45 66 6d 74 20 10 00 00 00 01 00 02 00**
**22 56 00 00 88 58 01 00 04 00 10 00 64 61 74 61 00 08 00 00 00 00 00 00**
**24 17 1e f3 3c 13 3c 14 16 f9 18 f9 34 e7 23 a6 3c f2 24 f2 11 ce 1a 0d**



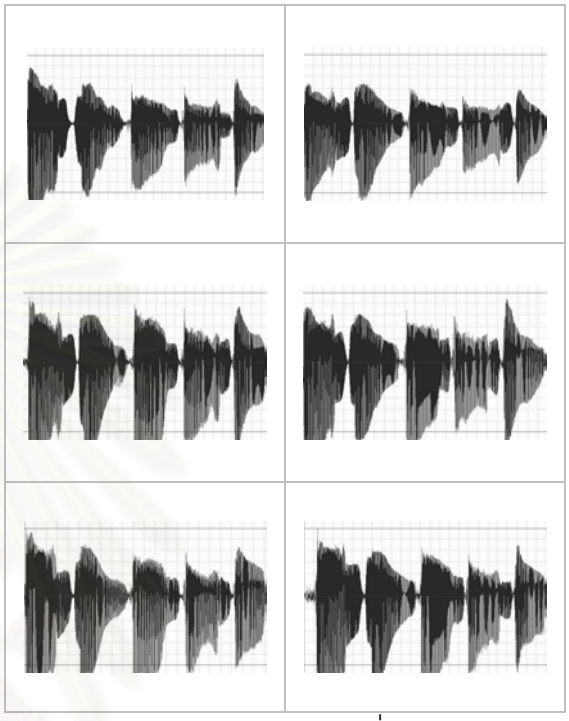

ที่มา: http://www.ora.com/centers/gff/formats/micriff/index.htm

# ภาคผนวก ข

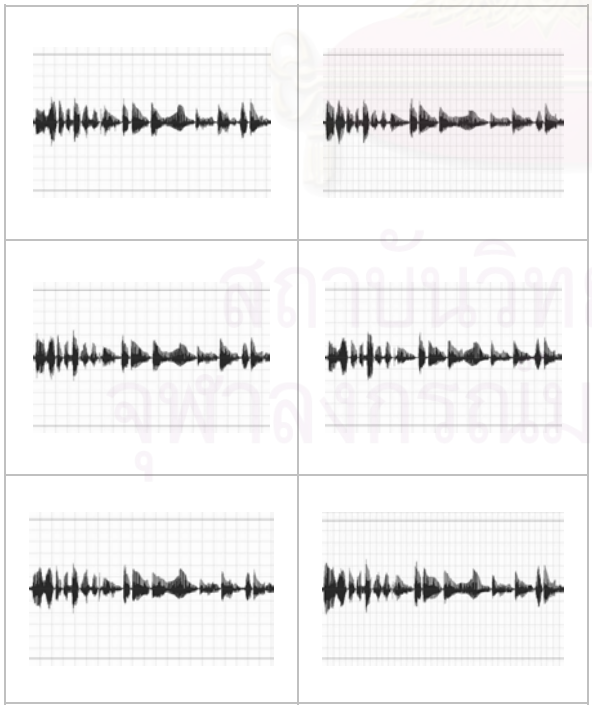
ตัวอย่างไฟล์เสียงพูดของข้อมูลที่ใช้ในการทดสอบ



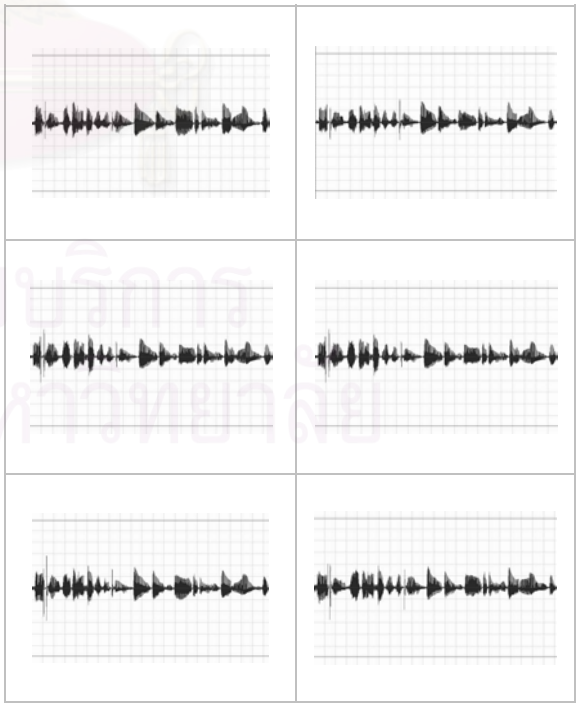
ข้อมูลทดสอบชุดที่ 1



ข้อมูลทดสอบชุดที่ 2



ข้อมูลทดสอบชุดที่ 3

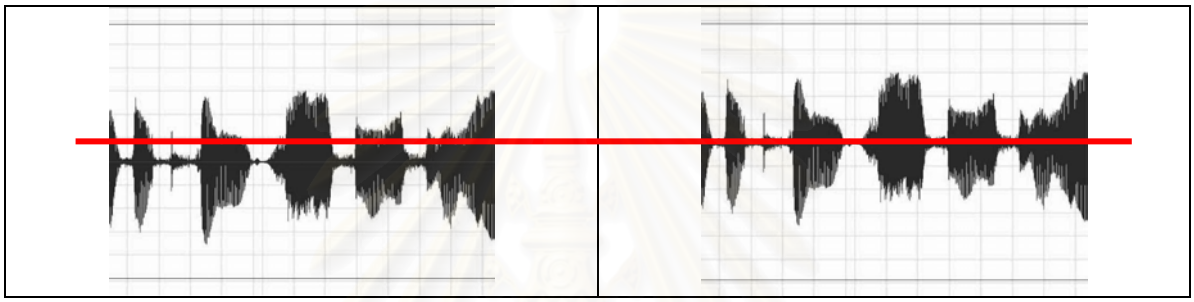


ข้อมูลทดสอบชุดที่ 4

# ภาคผนวก ค

## 1. DC Offset และการกำจัด DC Offset

สมมติให้เส้นตรงแนวนอนที่ลากผ่านในรูปแรก คือแกนศูนย์ ไฟล์เสียงพูดทางด้านซ้ายมือคือไฟล์เสียงพูดที่ประกอบไปด้วยสัญญาณเสียงพูดและสัญญาณไฟฟ้ากระแสตรง (DC) ที่เป็นลบ เราสามารถตัด DC Offset ออกได้โดยการกรองคลื่นความถี่เท่ากับศูนย์โดยโปรแกรมวิเคราะห์เสียงทั่วไป เช่น Cool Edit Pro 2.0 เป็นต้น ซึ่งผลลัพธ์จะได้ดังรูปทางด้านขวามือ

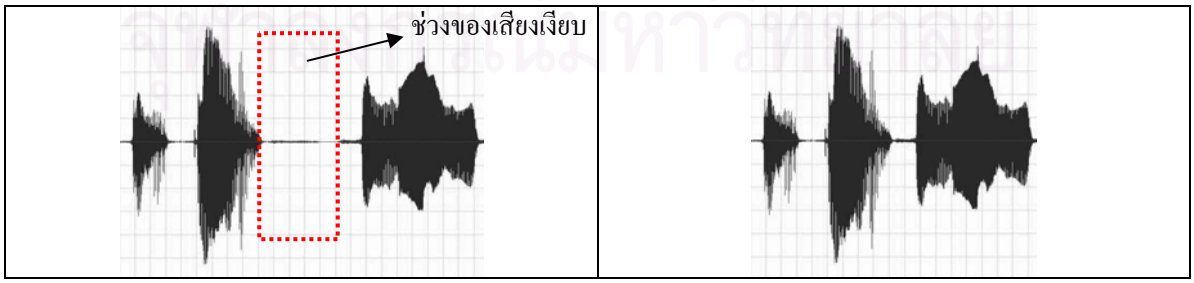


ก่อนกำจัด DC Offset

หลังกำจัด DC Offset

## 2. การกำจัดเสียงเงียบ (Silence Removal)

การกำจัดเสียงเงียบทำได้โดยการตั้งค่าขีดแบ่ง (Threshold value) เพื่อวัดแอมพลิจูดของคลื่นเสียง ซึ่งบ่งบอกถึงความดังของเสียงพูดว่าแอมพลิจูดช่วงไหถือว่าเป็นเสียงเงียบ เพราะในทางปฏิบัติแล้วเสียงเงียบจะไม่เงียบจริง และทำการกำหนดช่วงในการวิเคราะห์เสียงเงียบในแกนของเวลาเพื่อให้แน่ใจว่าเสียงเงียบนั้นเป็นเสียงเงียบที่เราไม่ต้องการจริงๆ




ก่อนกำจัดเสียงเงียบ

หลังกำจัดเสียงเงียบ

# ภาคผนวก ง

แสดงค่าพารามิเตอร์ที่จำเป็นสำหรับการสกัดลักษณะเฉพาะด้วยวิธีการแบบ MFCCs สามารถดาว์โหลดเครื่องมือที่ใช้ในการสกัดลักณะเฉพาะได้ทาง Internet เช่น HTK (Hidden Macov Model Toolkit) เป็นต้น โดยแต่ละเครื่องมืออยู่บนพื้นฐานของทฤษฎีเดียวกันแต่จะแตกต่างกันไปตามข้อปลีกย่อยและให้ความถูกต้องต่างกัน ซึ่งในงานวิจัยนี้ได้ทำการกำหนดค่าพารามิเตอร์ดังนี้

```
# Wave -> MFCC config file
SOURCEFORMAT = WAVE     # รูปแบบของไฟล์เสียงพูดที่จะทำการวิเคราะห์
SOURCERATE = 1250       # ความถี่ของไฟล์เสียงพูด
TARGETKIND = MFCC       # วิธีการในการสกัดลักษณะเฉพาะ
TARGETRATE = 100000     # อัตราความเร็วในการเลื่อนวินโดว์
                        # 10 ms frame rate = 8000*10 = 80 samples
WINDOWSIZE = 300000     # ขนาดของวินโดว์
                        # 30 ms window = 8000 * 30 ms = 240 samples
ZMEANSOURCE = T         # การกำจัดสัญญาณไฟกระแสตรงในแต่ละวินโดว์
USEHAMMING = T          # ใช้ Hamming window ในแต่ละวินโดว์
PREEMCOEF = 0.97        # มีการ Pre-emphasis ในแต่ละวินโดว์
NUMCHANS = 40           # จำนวนของตัวกรองของวิธีการ MFCCs
NUMCEPS = 18            # จำนวนมิติของเวกเตอร์ลักษณะเฉพาะ
LOFREQ = 20             # ค่าความถี่ต่ำสุดของตัวกรองความถี่
HIFREQ = 3600           # ค่าความถี่สูงสุดของตัวกรองความถี่
                        # Band Pass Filter 20-3600 Hz
```

# เป็นเครื่องหมายแสดงการอธิบายความหมายของการกำหนดค่าพารามิเตอร์

# ภาคผนวก จ

แสดงไอโซมอร์ฟิคฟังก์ชันที่เขียนโดยโปรแกรม Matlab โดยรับค่าข้อมูลเข้าคือ เวกเตอร์ลักษณะเฉพาะที่ได้จากการสกัดจากเสียงพูด โดยไม่จำกัดจำนวนมิติและจำนวนลักษณะเฉพาะ ผลลัพธ์ที่ได้คือลักษณะเฉพาะแบบฐานสอง ซึ่งสามารถอธิบายรูปแบบของลักษณะเฉพาะแต่ละตัว

```
function [countTb] =  Isomorphic(ceps)

[row,col] = size(ceps);
countTb = zeros(4,col-2);

for i = 1:row
   op02 = ceps(i,2) - ceps(i,1) ;
   for j = 2:col-1
      op01 = ceps(i,j+1) - ceps(i,j);
      abv01 = abs(op01);  % height  changing detection
      abv02 = abs(op02);
      if (op01 >= op02),
         if (abv01 >= abv02),
            countTb(1,j-1) = countTb(1,j-1) + 1;
         else
            countTb(2,j-1) = countTb(2,j-1) + 1;
         end
      else
         if (abv01 >= abv02),
            countTb(3,j-1) = countTb(3,j-1) + 1;
         else
            countTb(4,j-1) = countTb(4,j-1) + 1;
         end
      end
      op02 = op01;  % curve changing detection
   end
end

return
```

# ประวัติผู้เขียนวิทยานิพนธ์

นายศราวุธ จันทร์สด เกิดเมื่อวันที่ 23 พฤศจิกายน 2522 เรียนจบการศึกษาระดับมัธยมศึกษาที่โรงเรียนเรณูนครวิทยานุกูล อ.เรณูนคร จ.นครพนม เข้ารับการศึกษาต่อที่สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในคณะวิศวกรรมศาสตร์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ศึกษาและวิจัยเกี่ยวกับการวิเคราะห์สัญญาณเสียงพูด โมเดลการรู้จำที่เน้นทางด้านความเร็วและขนาดในการจัดเก็บ เคยทำวิจัยทางด้านการรู้จำเสียงพูดขนาดใหญ่โดยใช้ฮิดเดนมาร์คอฟโมเดล