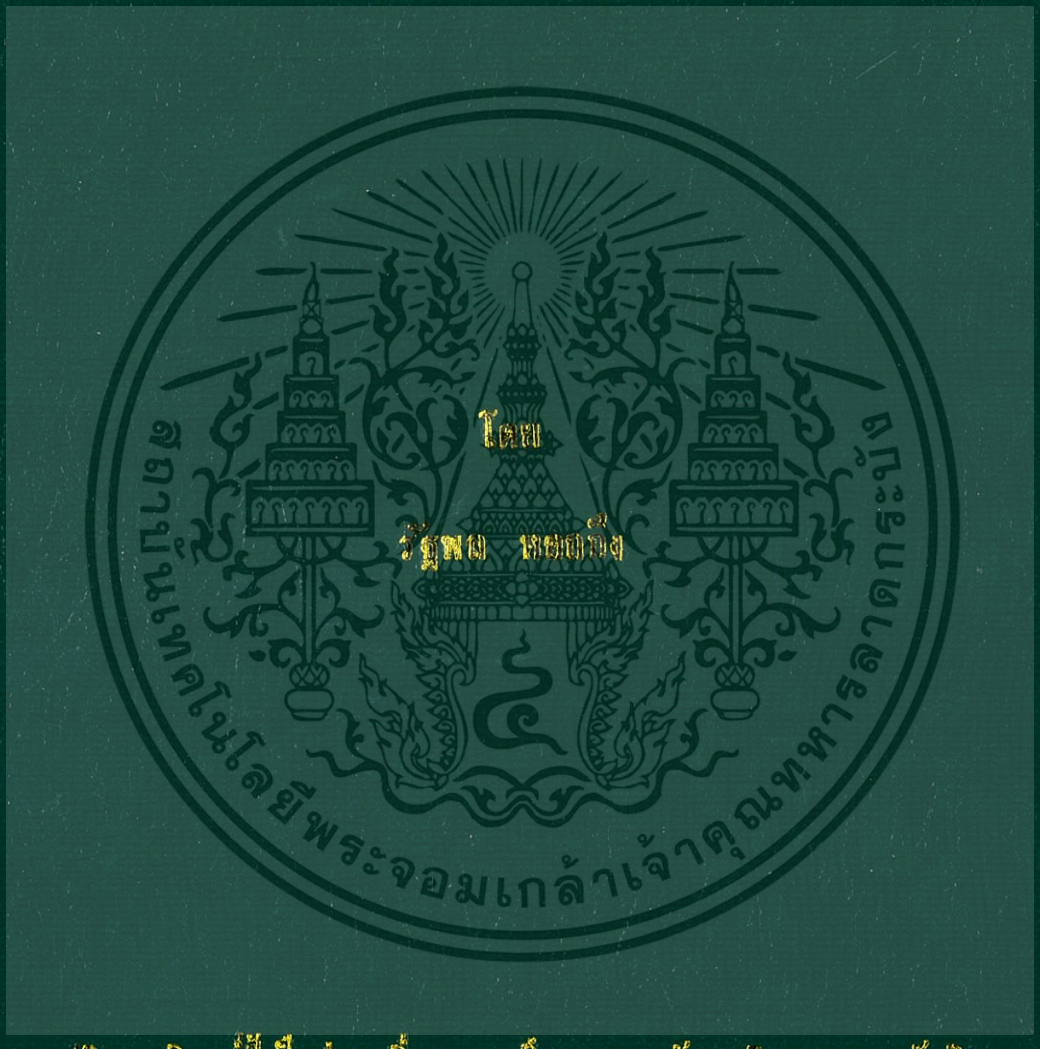
การปรับปรุงประสิทธิภาพของการสกัดคุณลักษณะต้นสำหรับการรู้จำเสียงพูด

AN IMPROVED FEATURE EXTRACTION FOR SPEECH RECOGNITION



โดย

วัฒพล หยกขาว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรดุษฎีบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2557

การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นสำหรับการรู้จำเสียง

AN IMPROVED FEATURE EXTRACTION FOR SPEECH RECOGNITION

โดย

รัฐพล หอกกิ่ง

RATTAPHON HOKKING

อาจารย์ที่ปรึกษา

ผู้ช่วยศาสตราจารย์ ดร.กันต์พงษ์ วรรัตน์ปัญญา

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2557

การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นสำหรับการรู้จำเสียง

AN IMPROVED FEATURE EXTRACTION FOR SPEECH RECOGNITION

โดย

รัฐพล หอกกิ่ง

RATTAPHON HOKKING

อาจารย์ที่ปรึกษา

ผู้ช่วยศาสตราจารย์ ดร.กันต์พงษ์ วรรัตน์ปัญญา

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2557

AN IMPROVED FEATURE EXTRACTION FOR SPEECH RECOGNITION

RATTAPHON HOKKING

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/2014

ใบรับรองปริญญานิพนธ์ ประจำปีการศึกษา 2557

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นสำหรับการรู้จำเสียง

AN IMPROVED FEATURE EXTRACTION FOR SPEECH RECOGNITION

ผู้จัดทำ

1. นายรัฐพล หอกกิ่ง รหัสนักศึกษา 54070081

.....อาจารย์ที่ปรึกษา

(ผู้ช่วยศาสตราจารย์ ดร. กันต์พงษ์ วรรัตน์ปัญญา)

หัวข้อโครงงาน	การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นสำหรับการรู้จำเสียง
นักศึกษา	นายรัฐพล หอกกิ่ง รหัสนักศึกษา 54070081
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
ปีการศึกษา	2557
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.กันต์พงษ์ วรรัตน์ปัญญา

บทคัดย่อ

ในปัจจุบันการรู้จำเสียงได้ถูกนำไปประยุกต์ใช้ในหลายงาน ตัวอย่างเช่น การใช้ในส่วนติดต่อผู้ใช้กับคอมพิวเตอร์โดยสั่งงานคอมพิวเตอร์ได้ด้วยเสียงพูด การจดบันทึกเอกสารโดยการพูด และการประยุกต์ใช้ในโปรแกรมแปลภาษา เป็นต้น กระบวนการรู้จำเสียงนิยมใช้ลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients ร่วมกับแบบจำลองฮิดเดนมาร์คอฟ แต่ประสิทธิภาพของลักษณะเด่นดังกล่าวขึ้นอยู่กับอัตราการสุ่มตัวอย่างของเสียงพูดที่แตกต่างกัน

งานวิจัยนี้มุ่งเน้นไปที่การปรับปรุงประสิทธิภาพของการสกัดลักษณะเด่นให้มีความทนทานต่ออัตราการสุ่มตัวอย่างที่แตกต่างกัน โดยประยุกต์ใช้เทคนิค Fractal Code ผลการทดลองพบว่าประสิทธิภาพของการรู้จำเสียงพูดมีประสิทธิภาพและไม่แตกต่างกัน ในขณะที่ข้อมูลเสียงพูดมีอัตราการสุ่มตัวอย่างที่แตกต่างกัน

Project Title	An Improved Feature Extraction for Speech Recognition
Student	Mr. Rattaphon Hokking Student ID 54070081
Degree	Bachelor of Science
Program	Information Technology
Academic Year	2014
Advisor	Assistant Professor Dr. Kuntpong Woraratpanya

ABSTRACT

Currently, speech recognition systems have many applications such as voice command applications, documentation by voice and real-time translation. Mel Frequency Cepstral Coefficients and Hidden Markov Model are widely applied to this system. However, Mel Frequency Cepstral Coefficient features are not robust to different sampling rates of speech signals.

This project focuses on improving feature extraction in order to obtain robust features that are invariant to sampling rate. Thus, feature extraction method is improved by using fractal code. The results show that this method archives effective recognition rate and tolerance to different sampling rates.

กิตติกรรมประกาศ

ขอขอบคุณศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ สำหรับข้อมูลเสียงพูดที่ใช้สำหรับสร้างแบบจำลอง และ Audacity team สำหรับโปรแกรมที่ใช้บันทึกและตัดต่อเสียงพูดที่ใช้ในงานนี้

ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.กันต์พงษ์ วรรัฒน์ปัญญา และสมาชิก Pattern Recognition and Image Processing Lab ทุกท่านและคณะเทคโนโลยสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ให้คำปรึกษาและช่วยเหลือจนกระทั่งสำเร็จโครงการ รวมทั้งให้การสนับสนุนเครื่องคอมพิวเตอร์สำหรับการทดลอง\

และขอขอบคุณ คุณพ่อและคุณแม่และพี่ๆ ที่คอยช่วยเหลือและให้กำลังใจในช่วงเวลาที่ทำโครงการนี้อย่างสม่ำเสมอ รวมทั้งเพื่อนๆ ที่คอยช่วยเหลือตลอดเวลา

รัฐพล หอกกิ่ง

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญรูป

หน้า

รูปที่

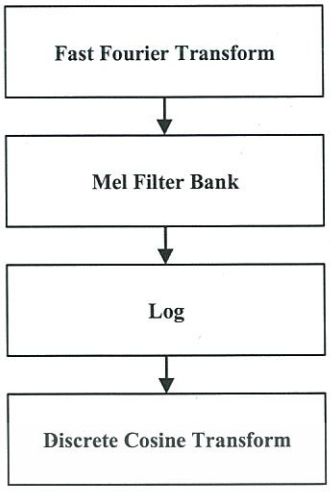
บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

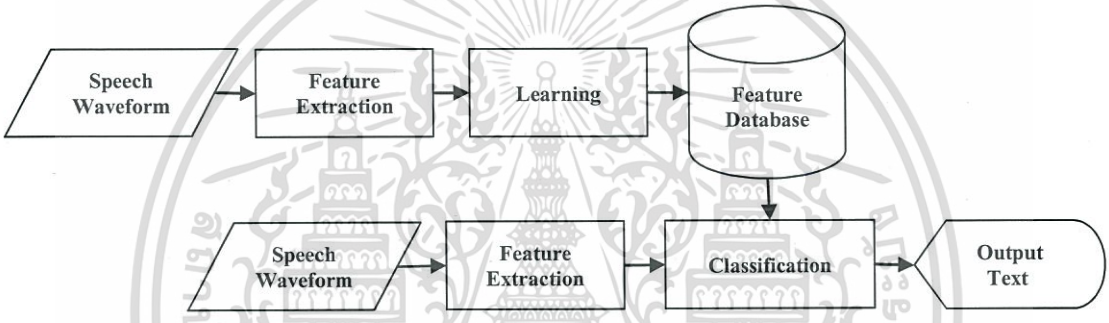
ธรรมชาติในการติดต่อสื่อสารกันของมนุษย์ คือ การใช้คำพูด โดยมนุษย์สามารถรับรู้และเข้าใจเสียงพูดอย่างเป็นธรรมชาติจากการประมวลผลที่ซับซ้อนของสมอง ส่วนคอมพิวเตอร์สามารถแปลความหมายและนำคำพูดมนุษย์ไปใช้ได้จากกระบวนการรู้จำเสียงพูด ในปัจจุบันการรู้จำเสียงได้ถูกนำไปประยุกต์ใช้งานอย่างแพร่หลาย ได้แก่ ใช้ในส่วนติดต่อผู้ใช้กับคอมพิวเตอร์โดยส่วนใหญ่จะใช้แป้นพิมพ์และตัวชี้ ซึ่งเป็นวิธีการที่ไม่เป็นธรรมชาติ การรู้จำเสียงจะทำให้ผู้ใช้สามารถสั่งงานคอมพิวเตอร์ได้ด้วยเสียงพูด ซึ่งเป็นการติดต่อสื่อสารที่เป็นธรรมชาติมากขึ้น `การใช้การจดบันทึกเอกสารโดยการพูดแทนการพิมพ์ผ่านแป้นพิมพ์ด้วยมือ และประยุกต์ใช้ในโปรแกรมแปลภาษาจากเสียงพูด

ขั้นตอนวิธีการรู้จำเสียงพูดประกอบ 2 ขั้นตอนหลัก คือ ขั้นตอนแรกเป็นการสกัดลักษณะเด่นเพื่อเป็นค่าตัวแทนที่สำคัญของเสียงพูด ซึ่งลักษณะเด่นที่นิยมในปัจจุบัน คือ Mel Frequency Cepstral Coefficients (MFCC) มีความแม่นยำในการรู้จำเสียงสูง ดังรูปที่ 1.1 และขั้นตอนที่สองเป็นการเรียนรู้ (Learning) เพื่อการรู้จำเสียงพูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ [1] [2] [3] ดังรูปที่ 1.2 ในขั้นตอนการสกัดลักษณะเด่นเริ่มจากการแปลงข้อมูลเสียงพูดในรูปคลื่นให้เป็นระดับพลังงานของความถี่ โดยใช้การแปลงแบบฟูเรียร์ จากนั้นทำการกรองเอาเฉพาะความถี่เสียงพูดของมนุษย์โดยใช้ตัวกรองแบบ Mel Filter Bank แล้วใช้ฟังก์ชันลอการิทึมแปลงข้อมูลให้มีความต่างของระดับพลังงานเพิ่มขึ้น และทำการสกัดค่าสัมประสิทธิ์ของระดับพลังงานด้วยฟังก์ชัน Discrete Cosine Transform แต่ลักษณะเด่นที่ได้มีข้อเสีย คือ ไม่ทนทานต่ออัตราการสุ่มตัวอย่างที่แตกต่างกันของข้อมูลนำเข้า เมื่อมีการสุ่มตัวอย่างข้อมูลนำเข้าในอัตราที่แตกต่างกันทำให้ได้เสียงที่แตกต่างกัน [4] [5] [6] [7]

ดังนั้น โครงงานนี้จึงนำเสนอวิธีการปรับปรุงการสกัดลักษณะเด่นของเสียงโดยใช้ Mel Frequency Cepstral Coefficients ร่วมกับการใช้ Fractal Code [8] แล้วนำลักษณะเด่นของเสียงที่ได้ไปสู่ขั้นตอนการเรียนรู้เพื่อใช้ในการจำแนกกลุ่มของเสียงด้วยแบบจำลองฮิดเดนมาร์คอฟ โดยมีข้อมูลนำเข้าของเสียงที่มีการสุ่มแตกต่างกัน ซึ่งจะทำให้ได้กระบวนการรู้จำเสียงที่สามารถสร้างแบบจำลองเสียงพูดเพียงแบบจำลองเดียวแล้วใช้จำแนกกลุ่มของเสียงพูดที่มีอัตราสุ่มต่างๆ กันทั้งหมดได้ โดยมีความแม่นยำไม่แตกต่างกัน



รูปที่ 1. 1 กระบวนการแปลงสัญญาณเสียงเป็น MFCC



รูปที่ 1. 2 กระบวนการรู้จำเสียง

1.2 ความมุ่งหมายและวัตถุประสงค์ของการวิจัย

เพื่อพัฒนาขั้นตอนวิธีการสกัดลักษณะเด่นของเสียงพูดให้มีประสิทธิภาพเพิ่มขึ้น เมื่อมีอัตราการสุ่มตัวอย่างข้อมูลนำเข้าแตกต่างกัน

1.3 ขอบเขตการวิจัย

1.3.1 ใช้กระบวนการการสกัดลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients

1.3.2 ใช้แบบจำลองฮิดเดนมาร์คอฟในการสร้างแบบจำลองเสียง

1.3.3 มีการเปรียบเทียบประสิทธิภาพของลักษณะเด่นที่สกัดได้จากขั้นตอนที่พัฒนาแล้วกับลักษณะเด่นของคู่เทียบประสิทธิภาพเพื่อวัดผล

1.3.4 เป็นการรู้จำเสียงแบบคำเดี่ยว (Isolated Word) แบบพยางค์เดียว (Mono Phone) ตัวอย่างเช่น คำว่า “ครับ” หรือ “ค่ะ” ซึ่งข้อมูลเสียงที่ใช้ในการเรียนรู้และทดสอบเป็นเสียงของบุคคลเดียว และมีสัญญาณรบกวนเล็กน้อย

1.3.5 ข้อมูลนำเข้ามีอัตราการสุ่มตัวอย่าง (Sampling Rate) ที่แตกต่างกัน ประกอบด้วย 11,025 22,050 และ 44,100 ครั้งต่อวินาที

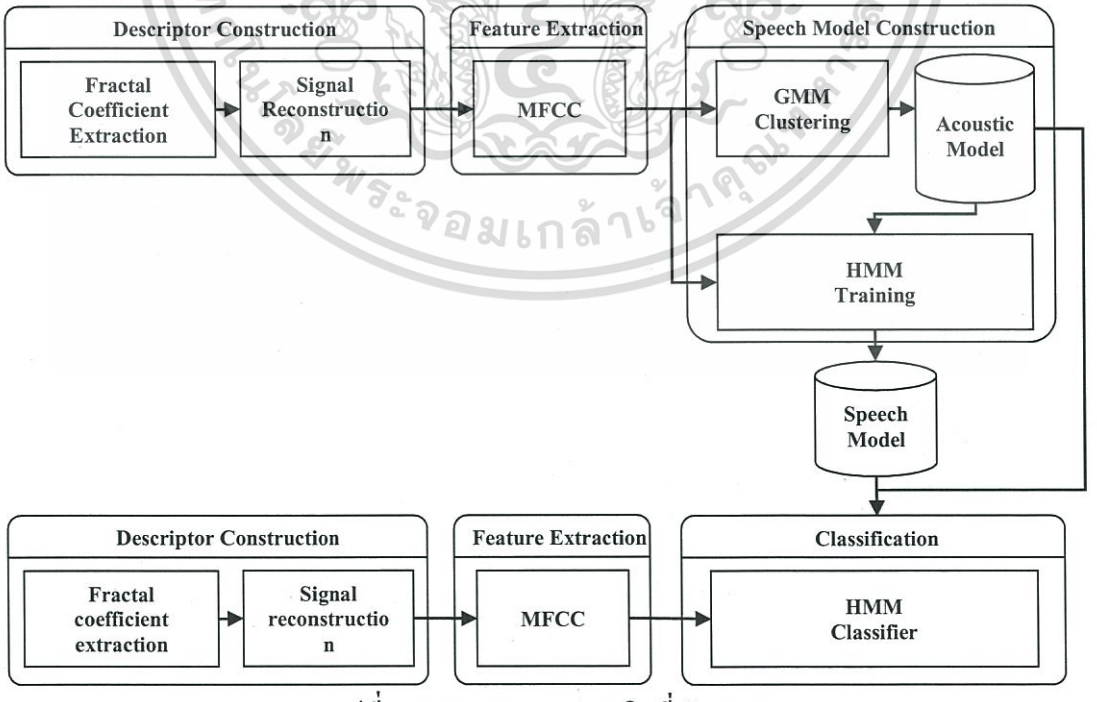
1.3.6 ข้อมูลที่ใช้ทดสอบแบบจำลองจะเป็นข้อมูลคนละชุดกับที่ใช้สร้างแบบจำลอง แต่เป็นคำพูดเดียวกัน

1.4 กรอบแนวความคิดที่นำเสนอ

ในการรู้จำเสียงพูด ประกอบด้วย 2 กระบวนการหลัก คือ กระบวนการสร้างแบบจำลองเสียงพูด และการจำแนกกลุ่ม ดังภาพที่ 1.3

ในกระบวนการสร้างแบบจำลองเสียงพูดมี 3 ขั้นตอน คือ ขั้นตอนการสร้างตัวกลางมาตรฐานจาก Fractal Code ของข้อมูลเสียงพูดที่มีอัตราสุ่มตัวอย่างต่างๆ ขั้นตอนการสกัดลักษณะเด่น Mel Frequency Cepstral Coefficients จากตัวกลาง ขั้นตอนการสร้างแบบจำลองลำดับเสียงพูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟร่วมกับแบบจำลองเชิงเสียงโดยใช้แบบจำลองส่วนผสมเกาส์เซียน (Gaussian Mixture Model)

ในกระบวนการจำแนกกลุ่ม จะประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการสร้างตัวกลางมาตรฐานจาก Fractal Code ของข้อมูลเสียงพูดที่มีอัตราสุ่มต่างๆ ขั้นตอนการสกัดลักษณะเด่น Mel Frequency Cepstral Coefficients และการจำแนกกลุ่มโดยใช้ขั้นตอนวิธี Viterbi ค้นหาลำดับเสียงที่มีความน่าจะเป็นมากที่สุดจากแบบจำลองฮิดเดนมาร์คอฟ



รูปที่ 1. 3 กรอบแนวความคิดที่นำเสนอ

1.5 ข้อตกลงเบื้องต้น

เสียงที่ใช้ในงานวิจัยจะมีสองชุด ชุดแรกเป็นเสียงพูดแบบคำเดี่ยว (Isolated Word) ภาษาไทยและใช้ผู้พูดเพียงผู้เดียว โดยสัญญาณเสียงที่ใช้มีสัญญาณรบกวนเล็กน้อย โดยมีรูปแบบไฟล์เป็น .wav 16 บิต ซึ่งถูกอัดโดยโปรแกรม Audacity ด้วยอัตราสุ่มตัวอย่าง 44,100 ครั้งต่อวินาที แล้วทำการลดอัตราสุ่มของตัวอย่างเป็น 22,050 ครั้งต่อวินาที 11,025 ครั้งต่อวินาทีเพื่อสร้างเป็นชุดข้อมูลเทียม (Pseudo Dataset) เสียงที่ใช้สร้างแบบจำลองและทดสอบแบบจำลองเป็นเสียงพูดคนละครั้งกัน

เสียงชุดที่สองเป็นไฟล์เสียงมาตรฐานจากฐานข้อมูล NECTEC - Thai Voice Command Corpus ซึ่งเลือกใช้เฉพาะเสียงที่เป็นพยางค์เดี่ยว (Isolated Word)

1.6 ประโยชน์ที่คาดว่าจะได้รับ

องค์ความรู้ที่ได้จากโครงการวิจัยจะนำไปพัฒนาขั้นตอนวิธีการสกัดลักษณะเด่นของเสียงพูดที่มีประสิทธิภาพ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การทบทวนวรรณกรรมที่เกี่ยวข้อง (Literature Review)

ผู้วิจัยได้ศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับการรู้จำเสียงด้วยวิธีการต่างๆ เพื่อใช้เป็นพื้นฐานในการวิจัยซึ่งพอสรุปได้ดังนี้

ได้มีงานวิจัย [1] แสดงถึงการประยุกต์ใช้แบบจำลองฮิดเดนมาร์คอฟ ในการรู้จำเสียงซึ่งมีการทดลองหาจำนวนสถานะของแบบจำลองที่เหมาะสม และมีการประยุกต์ใช้แบบจำลองกับข้อมูลที่มีค่าต่อเนื่อง ซึ่งงานวิจัยนี้ได้เสนอพื้นฐานการใช้งานแบบจำลองที่หลากหลาย

ในปี 1990 ได้มีการศึกษา [2] การรู้จำเสียงโดยใช้แบบจำลองฮิดเดนมาร์คอฟ และมีการใช้ Lincoln Robust Isolated Word Recognition ซึ่งช่วยในการกรองสัญญาณของสภาพแวดล้อมที่มีสัญญาณรบกวนออกไปจากเสียงพูด และนำไปปรับใช้กับการรู้จำเสียงแบบ Continuous Speech Recognition ที่มีคลังศัพท์จำนวน 1,000 คำ ซึ่งจากการทดลอง พบว่ามีความผิดพลาดในการรู้จำประโยคเพียง 0.1 เปอร์เซ็นต์

มีการศึกษา [3] ระบบรู้จำเสียงโดยมีการกล่าวถึงขั้นตอนในการแปลงสัญญาณเสียงพูดเป็นข้อมูลโดยแปลงเป็น Mel Frequency Cepstral Coefficients แล้วใช้แบบจำลองทางสถิติในการรู้จำเสียงพูดจากข้อมูลที่ได้ และแนะนำเกี่ยวกับการประยุกต์ใช้ระบบรู้จำเสียง

ได้มีงานวิจัย [4] ที่แสดงถึงประสิทธิภาพที่แตกต่างในการรู้จำเสียงเมื่อใช้ข้อมูลที่มีอัตราสุ่มตัวอย่าง (Sampling Rate) ที่แตกต่างกัน ทำให้ประสิทธิภาพในการรู้จำแตกต่างกัน โดยงานวิจัยนี้ได้ทดลองใช้ลักษณะเด่น Mel Frequency Cepstral Coefficients และ Linear Prediction derived Cepstral Coefficients ในการรู้จำด้วยอัตราสุ่มตัวอย่างที่ 6,000-16,000 ครั้งต่อวินาที ซึ่งพบว่าประสิทธิภาพจะเพิ่มขึ้นแปรผันตามอัตราการสุ่มตัวอย่างจนกระถึงอัตราสุ่มประมาณ 14,000 ครั้งต่อวินาที ประสิทธิภาพจะเริ่มลดลง นอกจากนี้ งานวิจัยนี้ยังกล่าวถึงความสัมพันธ์ระหว่างประสิทธิภาพในการรู้จำเสียงกับขนาดของลักษณะเด่น (จำนวนสัมประสิทธิ์) อีกด้วย โดยผลลัพธ์ของงานวิจัยนี้ พบว่าเมื่อใช้ลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients ที่มีสัมประสิทธิ์สูงกว่า 12 จะทำให้ประสิทธิภาพด้านความแม่นยำในการรู้จำเสียงพูดลดลงอย่างเห็นได้ชัด

ได้มีงานวิจัย [5] เกี่ยวกับการแก้ปัญหาในการรู้จำเสียง เมื่อมีอัตราสุ่มตัวอย่างที่แตกต่างกัน โดยวิธีการเดิมใช้การสร้างแบบจำลองแยกแต่ละอัตราสุ่มตัวอย่างไว้สำหรับข้อมูลแต่ละอัตราสุ่มตัวอย่าง มีข้อเสีย คือ ต้องทำการสร้างแบบจำลองหลายครั้ง ดังนั้น งานวิจัยนี้จึงมุ่งเน้นไปที่การสกัดลักษณะเด่นสำหรับแบบจำลองเดียว โดยการสกัดลักษณะเด่นจากเสียงแต่ละอัตราสุ่มตัวอย่างจะได้ลักษณะเด่นที่มีขนาดแตกต่างกัน ในงานวิจัยนี้จะสกัดข้อมูลช่วงความถี่ 0 – 8,000 ครั้งต่อ

วินาที มีขนาดสัมประสิทธิ์เท่ากับ 13 สำหรับ ข้อมูลช่วงความถี่ 0 – 11,000 ครั้งต่อวินาที จะขนาดสัมประสิทธิ์เท่ากับ 14 สำหรับข้อมูลช่วงความถี่ 0 – 16,000 ครั้งต่อวินาที ออกมา มีขนาดสัมประสิทธิ์เท่ากับ 15 ผลลัพธ์ของการรู้จำ คือ สามารถรู้จำเสียงโดยมีความแม่นยำที่วัดด้วยค่า Word Error Rate ใกล้เคียงกัน

ได้มีงานวิจัย [6] เกี่ยวกับการแก้ปัญหาเรื่องอัตราสุ่มตัวอย่างที่แตกต่างกัน โดยมุ่งเน้นไปที่การปรับค่าพารามิเตอร์ของ Mel Frequency Filter Bank ซึ่งในการรู้จำเสียงพูดนั้น จะเป็นการประยุกต์ใช้เทคนิคที่เกิดจากการทำ Time Scale Modification โดยจะทำการพิจารณาข้อมูลเสียงเป็นช่วงๆ ในกรอบเวลาที่กำหนดไว้ ซึ่งในกรอบเวลาที่กำหนดจะมีจำนวนข้อมูลมากน้อยขึ้นอยู่กับอัตราสุ่มตัวอย่างว่าจะมากน้อยเพียงใด และจำนวนข้อมูลที่ต่างกันส่งผลให้การสกัดลักษณะเด่นออกมาได้ต่างกัน งานวิจัยนี้จึงเสนอวิธีการปรับค่า Time Scale ให้เหมาะกับอัตราสุ่มตัวอย่างของข้อมูล ซึ่งจะทำให้ลักษณะเด่นที่ได้มีลักษณะใกล้เคียงกัน โดยมีการวัดประสิทธิภาพด้วยค่า Pearson correlation ระหว่างข้อมูลเดิมกับข้อมูลที่ถูกลดอัตราสุ่มตัวอย่างลง

ปี 1993 มีการวิจัย [9] ซึ่งมุ่งเน้นไปที่การประยุกต์การรู้จำเสียงแบบ Speaker-Independent Recognition ร่วมกับ Speaker-Dependent Recognition เพื่อทำให้สามารถรู้จำเสียงพูดเมื่อฐานข้อมูลไม่พบข้อมูลของผู้พูด และลดความผิดพลาดในการรู้จำเมื่อพบข้อมูลเสียงของผู้พูดในฐานข้อมูล ซึ่งเรียกว่า Speaker-Adaptive Speech Recognition โดยการรู้จำแบบ Speaker-Independent Recognition มีฐานข้อมูลเสียง 3,990 ประโยคมีความผิดพลาดในการรู้จำเฉลี่ย 4.3 เปอร์เซ็นต์ การรู้จำแบบ Speaker-Independent Recognition มีฐานข้อมูลเสียง 600 ประโยคมีความผิดพลาดในการรู้จำเฉลี่ย 2.6 เปอร์เซ็นต์ และเมื่อใช้ฐานข้อมูลที่มีขนาดเพิ่มขึ้นเป็น 2,400 ประโยค จะทำให้ความผิดพลาดลดลงมาเหลือ 1.4 เปอร์เซ็นต์ และการรู้จำแบบ Speaker-Adaptive Speech Recognition เมื่อใช้ฐานข้อมูลเสียง 2,400 ประโยคมีความผิดพลาด 1.4 เปอร์เซ็นต์

ในปี 1998 ได้มีการศึกษา [10] การแบ่งเสียงพูดเป็นพยางค์ ด้วยการใช้เทคนิค Local Maximum and Minimum Energy Contour ซึ่งทำการทดลองกับ Energy Algorithm 5 แบบ โดยใช้ข้อมูลเสียงพูดติดกัน 396 ประโยค ที่มี 2,992 พยางค์ บันทึกเสียงในสภาพแวดล้อมแบบ Office เป็นเสียงผู้ชาย 7 คน เสียงผู้หญิง 4 คน อายุระหว่าง 25–35 ปี ได้ความถูกต้องโดยเฉลี่ยดีกว่าเทคนิค Lamel LF ซึ่งความถูกต้องขึ้นอยู่กับ Energy Algorithm ที่เลือกใช้

ในปี 2001 ได้มีการศึกษา [11] การรู้จำเสียงสูงต่ำและตรวจจับการเน้นเสียงในภาษาไทย ซึ่งมี 5 ระดับเสียงโดยใช้วิธี Separated Stress Method (SSM) ซึ่งจะตรวจจับการเน้นเสียงของพยางค์แล้วทำการแยกกลุ่มของพยางค์เป็นกลุ่มเน้นเสียง และไม่เน้นเสียง และใช้ร่วมกับวิธี Incorporated Stress Feature Method (ISFM) ร่วมกับ Assimilation Feature ซึ่งทำให้ความถูกต้องในการรู้จำเสียงสูงต่ำเพิ่มขึ้นจาก 80.02 เปอร์เซ็นต์ เป็น 89.52 เปอร์เซ็นต์

มีงานวิจัย [12] นำเสนอวิธีการลบเสียงสัญญาณรบกวนจากสภาพแวดล้อมออกจากเสียงพูด โดยใช้วิธี Three-Step-Gain-Factor ซึ่งเป็นการใช้อัลกอริทึม Two-Step-Decision-Directed ร่วมกับการใช้ Median Filter เพื่อลดเสียงสัญญาณรบกวนจากสภาพแวดล้อมหรือเสียงดนตรีออกจากเสียงพูด โดยทำซ้ำหลายๆ ครั้ง เพื่อลบเสียงรบกวนสภาพแวดล้อมออกให้มากที่สุด โดยคงคุณภาพของเสียงพูด

มีงานวิจัย [13] เกี่ยวกับการรู้จำอารมณ์จากเสียงพูด โดยใช้ลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients และจำแนกด้วยแบบจำลองส่วนผสมเกาส์เซียน และแบบจำลองฮิดเดนมาร์คอฟ โดยอารมณ์ของมนุษย์เกิดจากอารมณ์พื้นฐาน ทั้ง 7 แบบ คือ อารมณ์โกรธ เกลียด กลัว มีความสุข ปกติ ดูถูก และแปลกใจ ใช้ฐานข้อมูลเสียงจาก IITKGP-SESC และ IITKGP-SEHSC ซึ่งมีทั้งหมด 1,2000 เสียง ซึ่งเป็นเสียงของผู้ชาย 5 คน และผู้หญิง 5 คน พูด 15 ประโยค มีการเพิ่มอารมณ์ที่ 8 จากอารมณ์พื้นฐาน คือ อารมณ์เศร้า ซึ่งอารมณ์แต่ละแบบจะถูกอัดด้วยช่วงเวลาต่างๆ กัน 10 ช่วงเวลา พบว่าจำนวน Gaussian Mixtures นั้นมีผลต่อประสิทธิภาพในการจำแนกกลุ่มของอารมณ์จากสียงพูดมาก และไม่สามารถระบุจำนวน Gaussian Mixtures ที่เหมาะสมได้ จึงทำการทดลองด้วยจำนวน Gaussian Mixtures และจำนวน States ของแบบจำลองฮิดเดนมาร์คอฟที่ต่างกันกับสองภาษา ผลที่ได้ คือ สามารถรู้จำอารมณ์กับภาษา Hindi ได้ถูกต้อง โดยเฉลี่ย 18.52%, 19.38% และ 17.95% ด้วยการใช้จำนวน Observer 8, 16 และ 32 จุดตามลำดับ และสามารถรู้จำอารมณ์กับภาษา Telugu ได้ถูกต้องโดยเฉลี่ย 18.58%, 17.95% และ 19% ด้วยการใช้จำนวน Observer 8, 16 และ 32 จุด ตามลำดับ

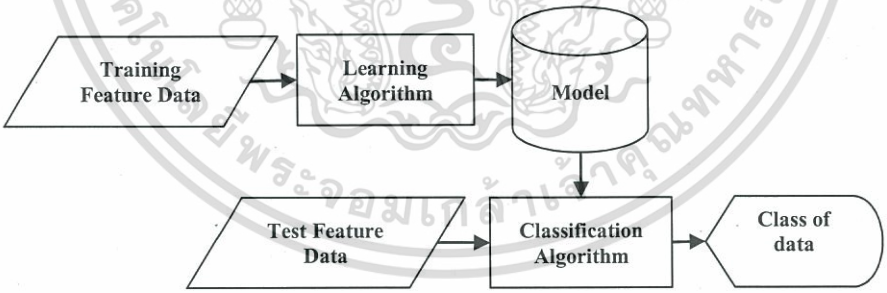
ได้มีงานวิจัย [14] เป็นการศึกษาจำนวนองค์ประกอบของแบบจำลองส่วนผสมกาส์เซียน ที่เหมาะสมโดยอัตโนมัติสำหรับการรู้จำเสียง ซึ่งงานวิจัยนี้ได้นำเสนอวิธีการที่ชื่อว่า Bayesian Ying-Yang (BYY) ซึ่งเป็นการใช้หลักการ Best Harmony Principle เป็นวิธีการทางสถิติ เพื่อหาแบบจำลอง Gaussian ที่เหมาะสม ซึ่งสามารถหาได้ขณะที่ทำการปรับค่าพารามิเตอร์ในกระบวนการเรียนรู้ด้วย Baum-Welch Algorithm วิธีการนี้จะถูกนำไปใช้เป็นส่วนเสริมในกระบวนการกล่าว ซึ่งวิธีการ BYY จะมีลักษณะที่คล้ายกับวิธีการทำ EM Algorithm ในการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับแบบจำลอง GMM โดยทำการทดลองกับฐานข้อมูลเสียง Hub4 Mandarin Broadcast News ที่มีจำนวนทั้งหมด 1,997 ประโยค ที่มีความยาวประมาณ 30 ชั่วโมง พบว่ากระบวนการ BYY นอกจากจะสามารถลดจำนวนองค์ประกอบ (Components) ลงแล้ว ยังเพิ่มความถูกต้องในการจำแนกอีกด้วย โดยเปรียบเทียบประสิทธิภาพด้วยค่า Word Error Rate (WER) และจำนวน Gaussian Component โดยเฉลี่ย เปรียบเทียบค่าที่ได้กับกระบวนการ EM-ML (Expectation Maximization - Maximum Likelihood), Bayesian Information Criterion (BIC) และ Akaike Information Criterion (AIC) ในการเปรียบเทียบผู้วิจัยได้แบ่งแนวทางการทดลองกับ BYY ไว้เป็น 2 แบบ คือ แบบ Strategy A และ แบบ Strategy B ซึ่งค่าที่ได้จากการทดลอง คือ EM-ML ใช้จำนวน Component

เฉลี่ยที่ 32 ได้ค่า WER 21.89, BYY - strategy A ใช้จำนวน Component เฉลี่ยที่ 19.96 ได้ค่า WER 20.80, BYY-Strategy B ใช้จำนวน Component เฉลี่ยที่ 22.79 ได้ค่า WER 21.21, BIC ใช้จำนวน Component เฉลี่ยที่ 10.39 ได้ค่า WER 23.48, AIC ใช้จำนวน Component เฉลี่ยที่ 25.29 ได้ค่า WER 22.21

ได้มีการศึกษา [15] การใช้ Fractal Image Coding ร่วมกับการ Down sampling และ Interpolation เพื่อลดเวลาที่ใช้ในการเข้ารหัส โดยทำการ Down sample รูปภาพตัวอย่างก่อนทำการเข้ารหัส และทำ Interpolation ข้อมูลที่ถอดรหัสแล้วให้มีขนาดเท่าเดิม ซึ่งผลที่ได้ทำให้ลดเวลาที่ใช้ในการเข้ารหัสลง โดยที่ยังคงความสมบูรณ์ของข้อมูลหลังจากถอดรหัสได้ ซึ่งทำการวัดคุณภาพของข้อมูลที่ถอดรหัสแล้วทำการ Interpolation โดยใช้ค่า Peak to Noise Ratio (PSNR) ซึ่งภาพที่ได้จะมีค่า PSNR อยู่ที่ประมาณ 22.5 dB ซึ่งเป็นความละเอียดที่ยอมรับได้โดยทั่วไปสำหรับรูปภาพโดยพบว่าแม้จะทำการลดคุณภาพของข้อมูลก่อนทำการเข้ารหัสด้วย Fractal Code แต่ยังสามารถถอดรหัสกลับคืนได้โดยที่คุณภาพของข้อมูลยังคงมีคุณภาพสมบูรณ์อยู่

2.2 การเรียนรู้ด้วยเครื่อง

กระบวนการเรียนรู้ด้วยเครื่อง มี 2 ขั้นตอนหลัก คือ การ Training ให้คอมพิวเตอร์เรียนรู้จากข้อมูลที่ป้อนเข้าไปเพื่อสร้างแบบจำลอง และกระบวนการในการจำแนกข้อมูลด้วยแบบจำลองที่ได้เรียนรู้แล้ว ดังรูปที่ 2.1



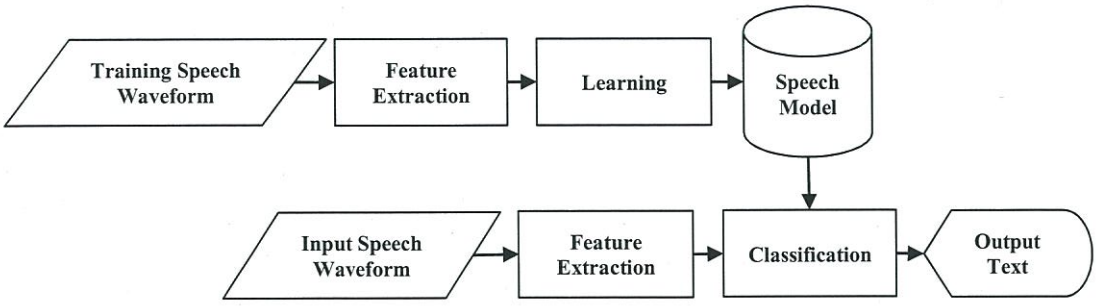
รูปที่ 2. 1 กระบวนการเรียนรู้ด้วยเครื่อง

2.3 การรู้จำเสียงพูด

กระบวนการรู้จำเสียงพูดมี 2 ขั้นตอนหลัก คือ ส่วนสกัดลักษณะเด่นของเสียง ส่วนการเรียนรู้ด้วยเครื่อง (Training) กับขั้นตอนการจำแนกกลุ่มของเสียง (Classification) ดังรูปที่ 2.2 มีขั้นตอนดังนี้

1) การ Training หรือการทำให้เครื่องเรียนรู้จากข้อมูลเสียงที่มีอยู่ที่ป้อนเข้าไป และ

2) การ Classification หรือการจำแนกกลุ่มของเสียงที่ต้องการจะทราบกลุ่มจากฐานข้อมูลลักษณะเด่นที่มี



รูปที่ 2. 2 แบบจำลองการรู้จำเสียงพูด

2.4 ตัวกรองแบบ Mel Scale Filter Bank

Melody scale คือ ระดับเสียงพูดที่ผู้รับสามารถได้ยินโดยระดับเสียง Mel จากความถี่ f สามารถคำนวณได้โดยใช้สมการที่ 2.1

$$m = 2595 \log_{10}\left(1+\frac{f}{700}\right) \qquad (2.1)$$

ส่วน Mel Scale Filter Bank คือ การเปลี่ยนระดับเสียง Mel ร่วมกับการใช้ Triangular Overlapping Windows ซึ่งทำการเลือกข้อมูลที่ต้องการที่อยู่ในช่วงของ window มาแปลงข้อมูล ในโครงงานนี้จะใช้การแบ่ง Window แบบ Hamming Window

2.5 การแปลงโคไซน์ไม่ต่อเนื่อง (Discrete Cosine Transform)

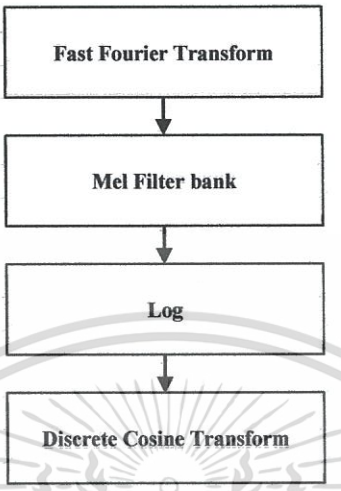
การแปลงโคไซน์แบบไม่ต่อเนื่อง เป็นการแปลงที่ใช้แปลงสัญญาณไม่ต่อเนื่อง เพื่ออธิบายลักษณะของสัญญาณในรูปสัมประสิทธิ์จากผลรวมของฟังก์ชันโคไซน์ จำแนกตามความถี่ของข้อมูล ถูกประยุกต์ใช้อย่างหลายหลายในเชิงวิศวกรรม และการคำนวณทางคณิตศาสตร์ต่างๆ ซึ่งสามารถมีฟังก์ชันโคไซน์พื้นฐานได้หลากหลาย โดยในโครงงานนี้ใช้การแปลงโคไซน์แบบไม่ต่อเนื่องแบบ DCT-III ซึ่งมีรูปฟังก์ชันดังสมการที่ 2.2

$$X_k = \frac{1}{2}x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N} n\left(k+\frac{1}{2}\right)\right] \qquad k = 0, \ldots, N-1 \qquad (2.2)$$

2.6 ลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients เป็นลักษณะเด่นของเสียงพูดมีขั้นตอนในการสร้างดังนี้ โดยแปลงข้อมูลเสียงพูดในรูปคลื่นให้เป็นระดับพลังงานของความถี่ โดยใช้การแปลงแบบฟูเรียร์ จากนั้นทำการกรองเอาเฉพาะความถี่เสียงพูดของมนุษย์โดยใช้ตัวกรองแบบ Mel filter bank

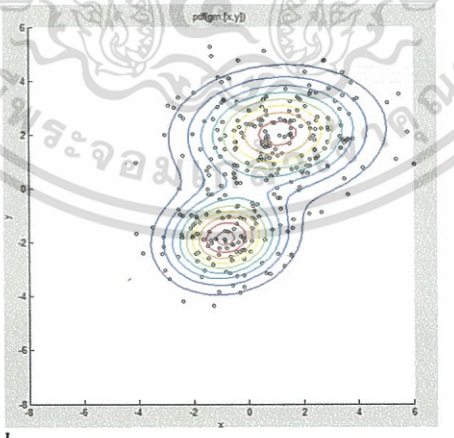
แล้วใช้ฟังก์ชันลอการิทึมแปลงข้อมูลให้มีความต่างของระดับพลังงานเพิ่มขึ้น และทำการสกัดค่าสัมประสิทธิ์ของระดับพลังงานด้วยฟังก์ชัน Discrete Cosine Transform ดังรูปที่ 2.3



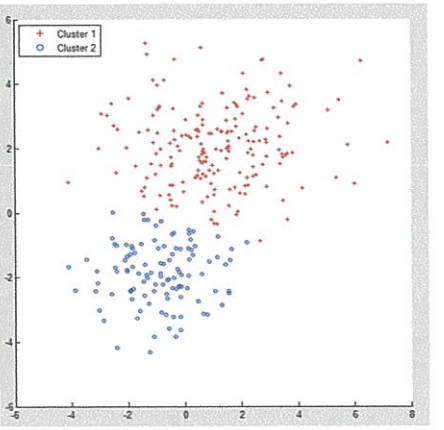
รูปที่ 2. 3 กระบวนการแปลงสัญญาณเสียงเป็น Mel Frequency Cepstral Coefficients

2.7 แบบจำลองส่วนผสมเกาส์เซียน (Gaussian Mixture Model)

กระบวนการในการจัดกลุ่มลักษณะเด่นด้วยแบบจำลองเชิงสถิติ โดยใช้ค่าเฉลี่ย (μ) และค่าความแปรปรวนของตัวอย่าง (σ^2) เป็นการแบ่งกลุ่มแบบ Cluster ซึ่งใช้การแบ่งกลุ่มแบบนี้เมื่อไม่รู้ว่าข้อมูลแต่ละชุดนั้นอยู่ในกลุ่ม (Class) ใด ดังรูปที่ 2.4 และรูปที่ 2.5



รูปที่ 2. 4 การกระจายของข้อมูลแบบ Gaussian



รูปที่ 2. 5 การแบ่งกลุ่มของข้อมูลด้วยแบบจำลองส่วนผสมเกาส์เซียน

แบบจำลองเกาส์เซียนจะถูกใช้เป็นเครื่องมือสำหรับหาค่าความคล้ายกัน (Likelihood) เพื่อแสดงความน่าจะเป็นที่ผู้สังเกต (Observer) จะมีสัดส่วนความน่าจะเป็นที่อยู่รอบๆ จุดที่ใช้เป็นตัวแทนขององค์ประกอบ (μ) ที่มีลักษณะการกระจายแบบเกาส์เซียนด้วยค่าความแปรปรวน (σ^2) ซึ่งสามารถหาค่าได้จากสมการที่ 2.3

$$p(x) = N(x;\ u,\ \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \tag{2.3}$$

โดยค่า μ และ σ^2 สามารถหาค่าประมาณได้จากสมการที่ 2.4 และ 2.5 ตามลำดับ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^i \tag{2.4}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x^i - \mu)^2 \tag{2.5}$$

สำหรับข้อมูลที่มีหลายมิติ จะหาค่า Probability Density ได้จากเวกเตอร์ค่าเฉลี่ย μ (สมการที่ 2.7) และเมตริกซ์ความสัมพันธ์ $\sum$ (สมการที่ 2.8) ได้จากสมการที่ 2.6

$$p(x \mid \mu, \textstyle\sum) = \frac{1}{(2\pi)^{d/2}\ |\sum|^{1/2}}\, e^{\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)} \tag{2.6}$$

โดยที่

$$\mu = E[x] \tag{2.7}$$

$$\sum = E\left[(x-\mu)(x-\mu)^T\right] \tag{2.8}$$

แบบจำลองส่วนผสม

สำหรับการประมาณค่าความหนาแน่นด้วยความยืดหยุ่นสามารถหาได้จากผลรวมของ Probability Density ขององค์ประกอบ ดังสมการที่ 2.9

$$p(x) = \sum_{j=1}^{m} p(x|j)P(j) \qquad (2.9)$$

p(x|j) คือ ความหนาแน่นขององค์ประกอบ

P(j) คือ ค่าน้ำหนัก

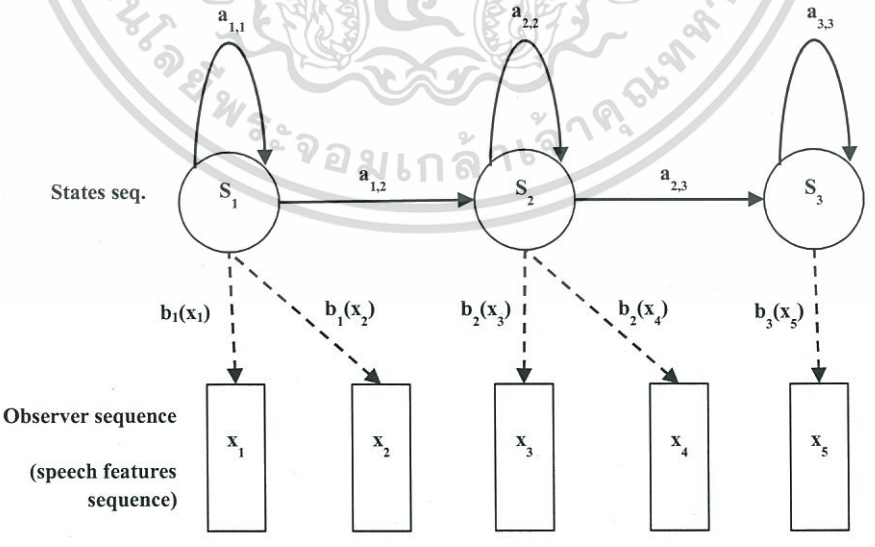
m คือ จำนวนองค์ประกอบ

และค่าความคล้ายกันของข้อมูล $X = \{x^1, x^2, x^3, \dots, x^n\}$ สามารถหาได้จากสมการที่ 2.10

$$L = \prod_{n=1}^{N} \sum_{j=1}^{M} p(x^n|j)\, P(j) \qquad (2.10)$$

2.8 แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model)

แบบจำลองฮิดเดนมาร์คอฟ คือ แบบจำลองทางสถิติที่ใช้พยากรณ์เหตุการณ์ที่จะเกิด โดยใช้ข้อมูลของผู้สังเกต (Observer) ที่อยู่ในอดีต โดยการหาความน่าจะเป็นแบบเงื่อนไข ซึ่งใช้สมมติฐาน Markov ความน่าจะเป็นของสถานะปัจจุบัน จะขึ้นอยู่กับความน่าจะเป็นของสถานะก่อนหน้าเท่านั้น ดังรูปที่ 2.6



รูปที่ 2. 6 แบบจำลองฮิดเดนมาร์คอฟ

ความน่าจะเป็นของเหตุการณ์ ณ สถานะที่ n ดังสมการที่ 2.11

$$P(q_n|q_{n-1}, q_{n-2}, \ldots, q_1) \tag{2.11}$$

จากสมมติฐานของมาร์คอฟดังสมการที่ 2.12

$$P(q_n|q_{n-1}) \tag{2.12}$$

เมื่อกำหนดให้เหตุการณ์แต่ละสถานะเป็นอิสระต่อกัน ก็จะได้ความน่าจะเป็นของลำดับสถานะ ดังสมการที่ 2.13

$$P(q_n| q_{n-1}, q_{n-2}, \ldots q_1) = \prod_{i=1}^{n} P(q_i|q_{i-1}) \tag{2.13}$$

แต่ในทางปฏิบัติไม่สามารถรู้ลำดับของสถานะที่แน่นอนได้ แต่จะทราบว่าผู้สังเกต นั้นเกี่ยวข้องกับสถานะแต่ละสถานะอย่างไร ซึ่งสามารถใช้กฎของ Bayes ในการหาค่าความน่าจะเป็นจากลำดับของผู้สังเกต ดังสมการที่ 2.14 และ 2.15

$$P(q_1, \ldots,q_n|x_1,\ldots, x_n) = P(x_1,\ldots, x_n|q_1, \ldots,q_n)P(q_1, \ldots,q_n) \tag{2.14}$$

$$= \prod_{i=1}^{n} P(x_i|q_i) \cdot \prod_{i=1}^{n} P(q_i|q_{i-1}) \tag{2.15}$$

โดย $P(x_i|q_j)$ คือ ค่าความน่าจะเป็นที่จะเป็น Observer x_i เมื่อมีสถานะเป็น q_i สำหรับ Observer ที่เป็นแบบ Discrete Value จะหาได้จาก Transition Matrix ดังสมการที่ 2.16

$$P(x_j|q_i) = b(i,j) \tag{2.16}$$

สำหรับค่าความน่าจะเป็นบนข้อมูลที่มีลักษณะแบบ Continuous Value จะใช้ค่าการกระจายของข้อมูล (Probability Density Function) แทนโดยใช้สมการที่ 2.17

$$P\left(x|q_j\right) = p(x|\mu^j, \Sigma^j) \tag{2.17}$$

โดยที่ μ^j, Σ^j คือ ค่าเฉลี่ยและค่าความแปรปรวนที่ใช้เป็นตัวแทนข้อมูลสถานะที่ j ซึ่ง p คือ Probability Density Function และ P คือค่าความน่าจะเป็น

2.9 ขั้นตอนวิธี Viterbi (Viterbi Algorithm)

ขั้นตอนวิธี Viterbi คือ การหาลำดับของสถานะที่ให้ความน่าจะเป็นสูงสุด เมื่อให้ลำดับของผู้สังเกตและแบบจำลองโดยใช้หลักการเรียกตัวเองซ้ำ (Recursion) ซึ่งผลลัพธ์จะได้ค่าความน่าจะเป็นของลำดับของสถานะที่มากที่สุด และลำดับของสถานะนั้น โดย $\delta_n(i)$ คือ ค่าความน่าจะเป็นสูงสุดของเส้นทางเพียงเส้นทางเดียวที่สถานะสุดท้ายเป็น สถานะ s_i ณ เวลาที่ n และ $\psi_n(i)$ คือ เส้นทางที่มีค่าความน่าจะเป็นสูงสุดของเส้นทางเพียงเส้นทางเดียวที่สถานะสุดท้ายเป็น สถานะ s_i ณ เวลาที่ n โดยในการเรียกตัวเองซ้ำ มีขั้นตอนดังนี้

1. การตั้งค่าความน่าจะเป็นเมื่อเริ่มแรก (Initialization) ในแบบจำลองดังสมการที่ 2.18 และตั้งลำดับเริ่มต้นของสถานะในแบบจำลองดังสมการที่ 2.19

$$\delta_1(i) = \pi_i \cdot b_j(x_1) \quad ; \ i = 1, \ldots, N_s \qquad (2.18)$$

$$\varphi_1(i) = 0 \qquad (2.19)$$

π_i คือ ความน่าจะเป็นที่จะเป็นสถานะที่ i ณ เวลา n =1

2. การเรียกตัวเองซ้ำ (Recursion) เป็นการเรียกตัวเองซ้ำเพื่อหาความน่าจะเป็นของลำดับการเปลี่ยนสถานะเมื่อสถานะสุดท้ายเป็น j ณ เวลาที่ n (สมการที่2.20) และสามารถหาสถานะที่มีความน่าจะเป็นสูงสุดได้จากสมการที่ 2.21

$$\delta_n(j) = \max (\delta_{n-1}(i) \cdot a_{ij}) \cdot b_j(x_n) \quad ; \ \begin{matrix} 1 \le i \le N_s, \\ 2 \le n \le N_s, \\ 1 \le j \le N_s \end{matrix} \qquad (2.20)$$

$$\varphi_n(i) = \arg\max (\delta_{n-1}(i) \cdot a_{ij}) \quad ; \ \begin{matrix} 1 \le i \le N_s, \\ 2 \le n \le N_s, \\ 1 \le j \le N_s \end{matrix} \qquad (2.21)$$

3. การหยุดเรียกตัวเองซ้ำ (Termination) เมื่อทำการคำนวนมาจนกระทั่งเวลาที่ N ค่าความน่าจะเป็นที่ N คือความน่าจะเป็นของลำดับสถานะที่เป็นไปได้ (สมการที่ 2.22)

$$p^*(X|\theta) = \max \delta_N(i) \quad ; 1 \le i \le N_s \qquad (2.22)$$

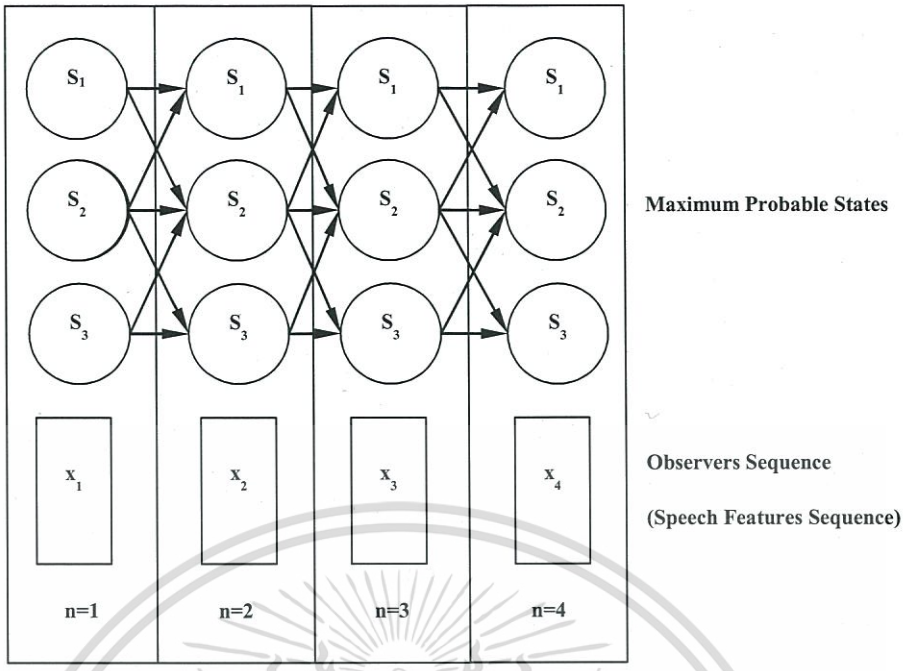
$$q_N^* = \arg\max \delta_N(i) \quad ; 1 \le i \le N_s \qquad (2.23)$$

4. การเดินย้อนกลับ (Backtracking) คือ การหาลำดับสถานะที่ให้ความน่าจะเป็นสูงที่สุด (สมการที่ 2.24) จากการเรียกตัวเองซ้ำ (สมการที่ 2.25)

$$Q^* = \{q_1^*, q_2^*, q_3^*, \ldots q_n^*\} \quad ; n = N-1, N-2, \ldots, 1 \qquad (2.24)$$

$$q_n^* = \varphi_{n+1}(q_{n-1}^*) \quad ; n = N-1, N-2, \ldots, 1 \qquad (2.25)$$

เพื่อหาลำดับของสถานะที่ให้ค่าความน่าจะเป็นสูงสุด



รูปที่ 2. 7 เส้นทางการเปลี่ยนสถานะโดย Viterbi Algorithm

2.10 การเข้ารหัสแบบ Fractal Code

การบีบอัดแบบ Fractal coding เป็นการบีบอัดข้อมูล ซึ่งมีโดยแบ่งชุดข้อมูลออกเป็นส่วนเล็กๆ และนำไปเปรียบเทียบกับแต่ละส่วนในชุดข้อมูล เพื่อหาตำแหน่งของส่วนข้อมูลที่มีความคล้ายกันและเก็บค่ารหัสไว้สำหรับใช้ในการประกอบข้อมูลกลับ (Reconstruction) ซึ่งมีขั้นตอนการทำงานดังนี้

1. กำหนดขนาดของ Block ของส่วนข้อมูลที่จะใช้แบ่งเพื่อทำการบีบอัด (Encode Frame Size) โดยแบ่งให้มีขนาดเป็น 2^n ด้วยสมการที่ 2.26

$$EFS = 2^n \qquad n = 1,2,\ldots \qquad (2.26)$$

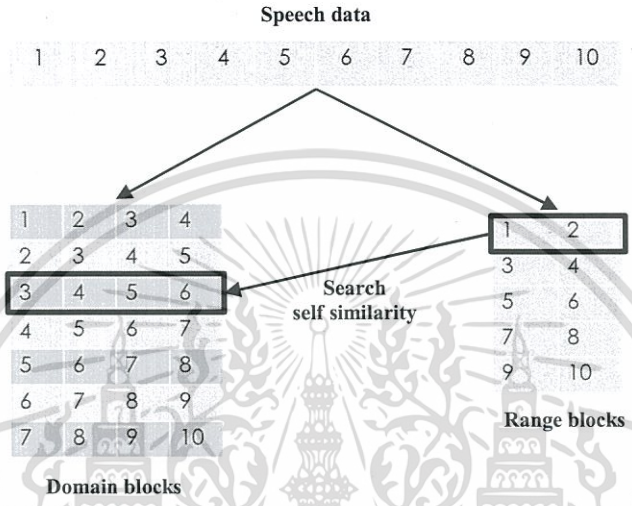
2. แบ่งส่วนของข้อมูลเป็นส่วนๆที่มีขนาดเท่ากันในแต่ละส่วน และมีขนาดเท่ากับ EFS จะรียกว่า Range Block

3. แบ่งส่วนของข้อมูลเป็นส่วนๆที่มีขนาดเท่ากันในแต่ละส่วน และมีขนาดเท่ากับสองท่าของ EFS เพื่อสร้างเป็น Domain Block

4. นำ Range Block แต่ละ Block ไปทำการเปรียบเทียบหาข้อมูลกับ Domain Block ที่มีค่าความคล้ายกันมากที่สุด

5. ทำการคำนวณหาค่า Code Block f_i ของแต่ละ Range Block ซึ่งแต่ละประกอบด้วยค่า Contrast (s), Brightness (o) และตำแหน่ง Domain Block ที่ Range Block มีความคล้ายคลึงที่สุด ดังสมการที่ 2.27

$$f_i = \{s, o, \text{Domain Index}\} \quad (2.27)$$



รูปที่ 2. 8 การบีบอัดข้อมูลแบบ Fractal Code

2.11 การประกอบข้อมูลใหม่ (Reconstruction)

การประกอบข้อมูลใหม่ เป็นการใช้ Fractal Code ที่หาได้มาทำการสร้างข้อมูลใหม่ด้วย Fractal Code ที่คำนวณได้ร่วมกับ Iterative Function ซึ่ง Fractal Code มีคุณสมบัติพิเศษ คือ สามารถสร้างข้อมูลที่มีความละเอียดไม่เท่ากับข้อมูลต้นฉบับก่อนทำการบีบอัดก็ได้ (Resolution Independence)

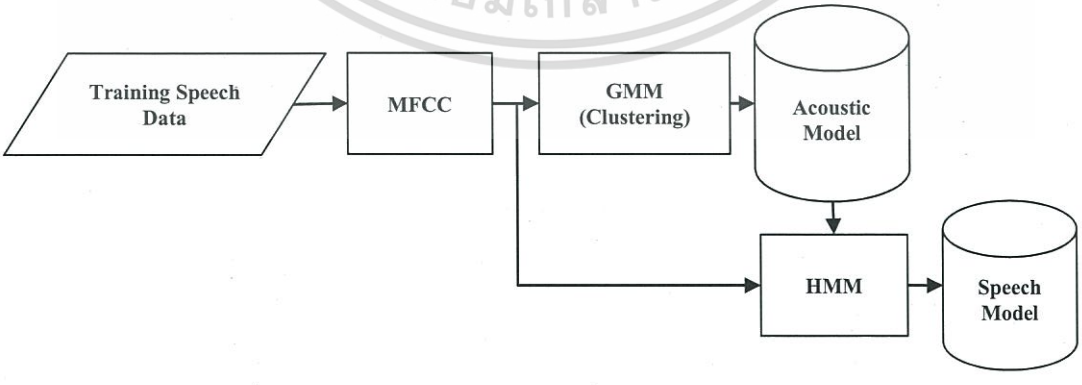
บทที่ 3
การดำเนินงานวิจัย

3.1 ปัญหาของการรู้จำเสียงพูด

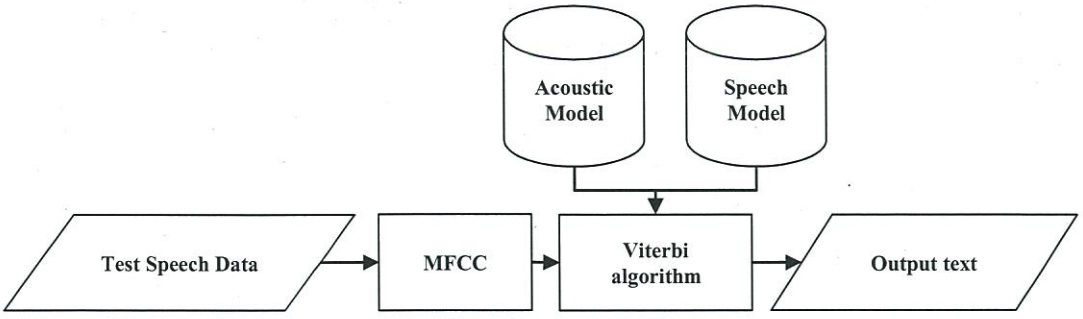
ในการรู้จำด้วยแบบจำลองฮิดเดนมาร์คอฟ นั้น ต้องการอนุกรม (Sequence) ของลักษณะเด่นของเสียง โดยประสิทธิภาพของแบบจำลองจะขึ้นอยู่กับลักษณะเด่นที่สกัดได้จากกระบวนการสกัดลักษณะเด่น จากการทบทวนวรรณกรรม [4] [5] [6] [7] พบว่าลักษณะเด่นของเสียงแบบ Mel Frequency Cepstral Coefficients (MFCC) ของข้อมูลเสียงเดียวกันที่มีอัตราสุ่มตัวอย่างที่แตกต่างกันจะทำให้ลักษณะเด่นที่สกัดได้มีความแตกต่างกันด้วย ทำให้การสกัดลักษณะเด่นด้วย MFCC ไม่ทนทานต่อความแตกต่างของอัตราสุ่มตัวอย่างที่แตกต่างกัน ส่งผลให้ประสิทธิภาพในการจำแนกกลุ่มเสียงพูดแตกต่างกันตามอัตราสุ่มตัวอย่างที่แตกต่างกัน

3.2 แนวทางการดำเนินงานวิจัย

โครงการนี้ใช้แบบจำลองฮิดเดนมาร์คอฟ ในการฝึกให้เครื่องเรียนรู้ลำดับของลักษณะเด่นที่ได้มาจากสัญญาณเสียง โดยใช้ลักษณะเด่นที่ได้จากกระบวนการ Mel Frequency Cepstral Coefficients และใช้แบบจำลองส่วนผสมเกาส์เซียน ในการบอกกลุ่มของลักษณะเด่นด้วยป้ายชื่อ (Label) และใช้ป้ายชื่อนี้เป็นข้อมูลนำเข้าของแบบจำลองฮิดเดนมาร์คอฟ ซึ่งผลลัพธ์จะได้แบบจำลองเสียงพูด (Speech Model) (รูปที่ 3.1) และจำแนกกลุ่มของเสียงโดยใช้อัลกอริทึม Viterbi ในการหาค่าความเหมือนของเสียงเปรียบเทียบกับแบบจำลองเสียงได้จากแบบจำลองฮิดเดนมาร์คอฟ (รูปที่ 3.2)



รูปที่ 3. 1 กระบวนการ Training เพื่อสร้างแบบจำลองเชิงเสียง



รูปที่ 3. 2 กระบวนการจำแนกเสียงพูด

3.3 การเตรียมข้อมูลก่อนประมวลผล

3.3.1 การแบ่งกลุ่มสำหรับทดสอบ

ข้อมูลเสียงที่ใช้ทดสอบแบ่งเป็น 2 ชุด

1) ชุดข้อมูลที่ใช้ทดสอบความแตกต่างที่เกิดจากการ Sampling ด้วยความถี่ที่แตกต่างกัน ข้อมูลชุดนี้จะมีผู้พูดเพียงคนเดียว มี 18 คำ เป็นคำพูดแบบพยางค์เดี่ยว แต่ละคำมี 10 เสียงสำหรับการ Training และ 10 เสียงสำหรับการ Classification ข้อมูลทั้งหมดจะมี 3 ชุดที่ถูกอัดเสียงด้วยอัตราสุ่มตัวอย่างที่แตกต่างกัน คือ 44,100 22,050 และ 11,025 ครั้งต่อวินาที รวมจำนวนไฟล์เสียงทั้งหมดคือ 1,080 ไฟล์ (18×(10+10)×3) ซึ่งจะใช้ทดสอบด้วยแบบจำลองส่วนผสมเกาส์เซียนและแบบจำลองฮิดเดนมาร์คอฟ

ตารางที่ 3. 1 ลักษณะของชุดข้อมูลในการทดลองรู้จำด้วยอัตราสุ่มตัวอย่างที่แตกต่างกัน

ลักษณะของชุดข้อมูล	อัตราการสุ่มตัวอย่าง (ครั้งต่อวินาที)		
	11,025	**22,050**	**44,100**
จำนวนผู้พูด	1 คน		
จำนวนพยางค์	1 พยางค์		
จำนวนคำ	18 คำ		
จำนวนตัวอย่างในการสร้างแบบจำลอง (ไฟล์)	180	180	180
จำนวนตัวอย่างในการทดสอบแบบจำลอง (ไฟล์)	180	180	180

จากตารางที่ 4.2 คือ ลักษณะของข้อมูลที่ใช้ในการสร้างแบบจำลอง โดยใช้ข้อมูลนำเข้าที่มีอัตราสุ่มตัวอย่างที่แตกต่างกัน คือ 11,000 ครั้งต่อวินาที 22,050 ครั้งต่อวินาที 44,100 ครั้งต่อวินาที โดยคำพูดแต่ละคำจะถูกอัดเสียงในช่วงเวลาเดียวกันที่สภาพแวดล้อมเดียวกัน จำนวนสิบครั้ง สำหรับใช้ในการสร้างแบบจำลอง และอัดเสียงช่วงเวลาเดียวกันที่สภาพแวดล้อมเดียวกันอีกสิบครั้ง สำหรับใช้เป็นข้อมูลทดสอบ

2) ชุดข้อมูลสำหรับการพัฒนากระบวนการสกัดลักษณะเด่น เป็นชุดข้อมูลที่มาจากฐานข้อมูล NECTEC - Thai Voice Command Corpus ซึ่งเลือกเฉพาะข้อมูลที่เป็นเสียงพูดพยางค์เดียว ในข้อมูลชุดนี้ประกอบด้วย เสียงพูดจากผู้ชาย 5 คนและเสียงผู้หญิง 7 คน แต่ละคนจะพูดทั้งหมด 67 คำ จำนวนไฟล์เสียงรวมทั้งหมด 804 ไฟล์ ถูกบันทึกด้วยอัตราสุ่มตัวอย่างที่ 16,000 ครั้งต่อวินาที โดยในการ Training และการ Classification จะใช้ข้อมูลชุดเดียวกัน

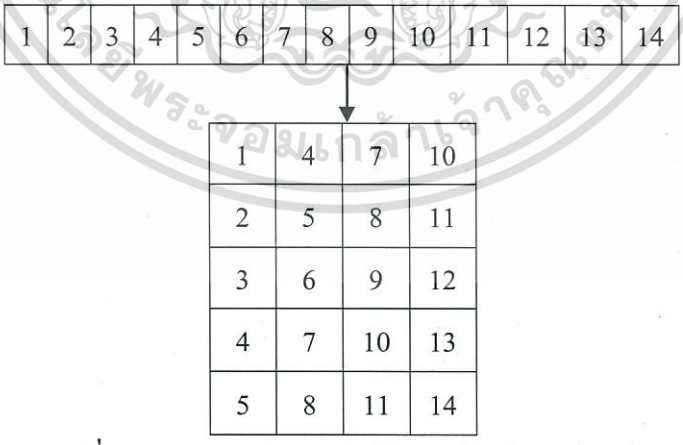
3.3.2 การแปลงข้อมูลเสียงเป็นเวกเตอร์ข้อมูล

งานวิจัยนี้จะทำเลือกตัวอย่าง (Sampling) ข้อมูลเสียงที่เข้ามาทุกๆ Ts มิลลิวินาทีเป็น 1 เวกเตอร์ข้อมูล (รูปที่ 3.3) ซึ่งความยาวเสียงที่ถูกแบ่งจะมีขนาดเท่ากับ Tw มิลลิวินาที ซึ่งถ้าหาก Tw > Ts จะเกิดการ shift ของข้อมูลโดยการเลือกความยาวของเสียงที่จะตัด ซึ่งข้อมูลเสียงที่เก็บไว้จะมีข้อมูลตัวอย่าง (sample) เท่ากับความถี่ที่ใช้อุปกรณ์บันทึก มีหน่วยเป็นจำนวนครั้งต่อวินาที ซึ่งสามารถแปลงความยาวของเสียงจาก Tw ที่มีหน่วยเป็นวินาทีให้เป็น Nw (จำนวน sample ในหนึ่งหน่วยเวลา) ที่มีหน่วยเป็นจำนวน sample ได้ด้วยสมการที่ 3.1

$$Nw = \frac{Tw \times (\text{Sampling Rate})}{1000} \tag{3.1}$$

ซึ่งการแปลงหน่วยของความถี่ในการเลือกตัวอย่างจาก จำนวนครั้งต่อวินาทีเป็นจำนวนตัวอย่างต่อครั้ง ได้ด้วยสมการ

$$Ns = \frac{Ts \times (\text{Sampling Rate})}{1000} \tag{3.2}$$

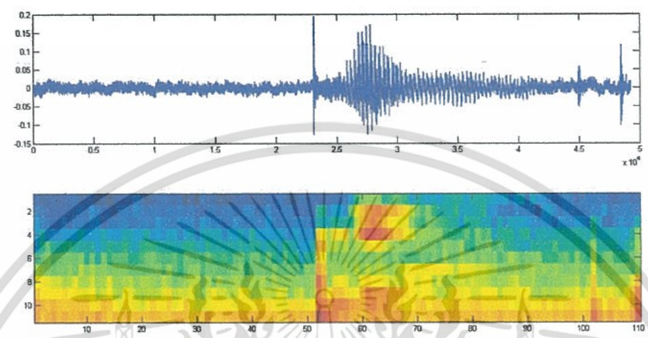


รูปที่ 3. 3 กระบวนการแปลงข้อมูลเสียงเป็นเวกเตอร์

3.4 การสกัดลักษณะเด่น

3.4.1 การสกัดลักษณะเด่นด้วย Mel Frequency Cepstral Coefficients

ลักษณะเด่นจะถูกสกัดจากข้อมูลเสียงที่ผ่านการแปลงเป็นเมตริกซ์ข้อมูลแล้ว ซึ่งผลลัพธ์ที่ได้จากกระบวนการนี้เป็นเมตริกซ์ของลักษณะเด่นที่แทนข้อมูลของผู้สังเกต (Observer) ซึ่งแต่ละคอลัมน์จะแทนผู้สังเกตของแต่ละเสียงในช่วงเวลาหนึ่ง ดังรูปที่ 3.4



รูปที่ 3. 4 ลักษณะเด่น Mel Frequency Cepstral Coefficient

3.5 การเรียนรู้ด้วยเครื่อง

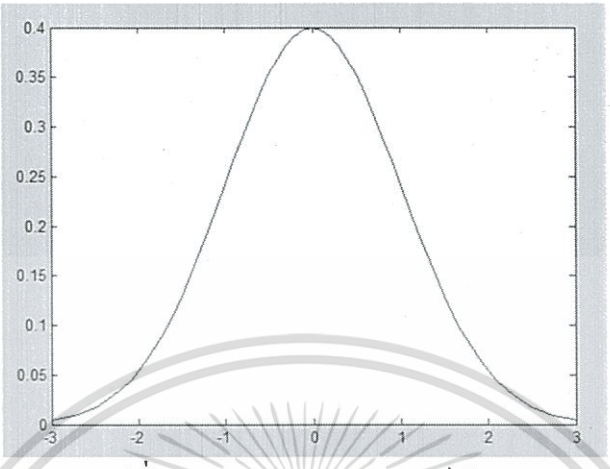
3.5.1 การสร้างแบบจำลองส่วนผสมเกาส์เซียน (Gaussian Mixture Model)

แบบจำลองเกาส์เซียนจะถูกใช้เป็นข้อมูลนำเข้าสำหรับการคำนวณหาค่า Output Distribution สำหรับกระบวนการสร้างแบบจำลองและทดสอบด้วยแบบจำลองฮิดเดนมาร์คอฟ ด้วยฟังก์ชันหาค่าความน่าจะเป็นแบบต่อเนื่อง (Probability Density Function)

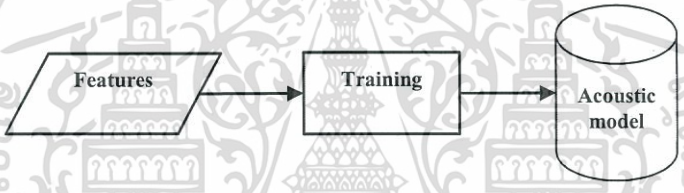
ในการสร้างแบบจำลองส่วนผสมเกาส์เซียน จะใช้ข้อมูลนำเข้าที่ได้จากกระบวนการสกัดลักษณะเด่นซึ่งมีลักษณะเป็นเวกเตอร์ของข้อมูลเสียงทั้งหมดที่ถูกนำมารวมกันและถูกทำให้แสดงผลในรูปแบบของเมตริกซ์ข้อมูล ในการสร้างแบบจำลองเกาส์เซียนจำเป็นต้องรู้ค่าพารามิเตอร์สำหรับการปรับแบบจำลองที่สำคัญคือ จำนวนองค์ประกอบของเกาส์เซียน ซึ่งค่าที่ดีที่สุดนั้นสามารถหาได้จากการทดลองหรือกระบวนการค้นหาค่าที่ดีที่สุดเท่านั้น [13] [14] ซึ่งในงานวิจัยนี้มีข้อสันนิฐานว่าเสียงพูด 1 พยางค์จะมีตัวแทนหรือผู้สังเกต (Observer) ของสัญญาณจำนวน 3 ตัวแทน ซึ่งจะใช้ตัวแทนสัญญาณที่แตกต่างกันทั้งหมด สำหรับทุกเสียงพูด และทุกเสียงพูดจะมีตัวแทนสัญญาณเงียบ (Silent Signal) เพิ่มขึ้นอีกหนึ่งตัวแทน ดังนั้นแบบจำลองจะมีจำนวนองค์ประกอบเกาส์เซียนคือ 55 องค์ประกอบสำหรับข้อมูลชุดที่ 1 และ 403 สำหรับข้อมูลชุดที่สอง ซึ่งในข้อมูลชุดที่สองจะแบ่งชนิดของผู้สังเกตเป็นสองชนิดคือ ผู้สังเกตที่เป็นของผู้หญิง และผู้สังเกตที่เป็นของผู้ชาย

วิธีการหาแบบจำลองเกาส์เซียนที่ดีที่สุดสำหรับข้อมูลที่ใช้นั้น จะหาได้จากขั้นตอนวิธี Expectation Maximization – Maximum Likelihood ซึ่งผลลัพธ์จะได้เป็น แบบจำลองที่เหมาะกับ

ข้อมูลที่จะถูกนำไปใช้เป็นข้อมูลนำเข้าสำหรับคำนวณหาค่าความน่าจะเป็นแบบต่อเนื่องสำหรับแบบจำลองฮิดเดนมาร์คอฟ ดังรูปที่ 3.5 และ 3.6



รูปที่ 3. 5 การกระจายแบบเกาส์เซียน



รูปที่ 3. 6 ขั้นตอนการสร้างแบบจำลองเกาส์เซียน

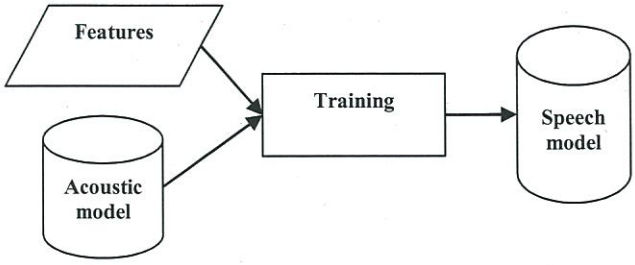
3.5.2 การสร้างแบบจำลองฮิดเดนมาร์คอฟ

แบบจำลองฮิดเดนมาร์คอฟ จะถูกใช้เป็นแบบจำลองสำหรับจำแนกลำดับของผู้สังเกต (Observer Sequence) เพื่อหาค่าความคล้ายกัน (Likelihood) ของลำดับผู้สังเกต

โดยผู้สังเกตสำหรับข้อมูลแบบต่อเนื่อง จะเป็นเวกเตอร์ของลักษณะเด่นที่สกัดได้ ในการสร้างแบบจำลอง จำเป็นต้องรู้จำนวนสถานะที่แน่นอนในเสียงพูดแต่ละคำ ซึ่งงานวิจัยนี้สมมติฐานว่าเสียงพูดแต่ละคำ มีสถานะของสัญญาณอยู่ 3 สถานะ

ในการสร้างแบบจำลอง จะใช้ข้อมูลนำเข้า คือ แบบจำลองส่วนผสมเกาส์เซียน (รูปที่ 3.7) เพื่อใช้หาค่า Output Distribution ที่ได้จากการหาค่า pdf และค่าความน่าจะเป็นในการเปลี่ยนสถานะที่ถูกแสดงอยู่ในรูปเมตริกซ์ความสัมพันธ์ a_{ij}

สำหรับการเรียนรู้ด้วยเครื่องของแบบจำลองฮิดเดนมาร์คอฟ จะทำการเรียนรู้ด้วยขั้นตอนวิธี Viterbi ซึ่งผลลัพธ์จะได้แบบจำลองที่เหมาะกับเสียงที่ใช้เรียนรู้มากที่สุด ในงานวิจัยนี้จะใช้แบบจำลองฮิดเดนมาร์คอฟ แตกต่างกันไปตามแต่ละเสียง



รูปที่ 3. 7 การสร้างแบบจำลองฮิดเดนมาร์คอฟ

3.6 การบีบอัดข้อมูลแบบแฟรคทอล

การบีบอัดข้อมูลด้วย Fractal Code สามารถทำได้ดังรูปที่ 3.8

รูปที่ 3. 8 กระบวนการบีบอัดข้อมูลแบบแฟรคทอล

3.7 การแปลงรหัส

การแปลงรหัส(Transcoding) คือ การเปลี่ยนแปลงข้อมูลที่อยู่ในรหัสหนึ่งด้วยการเข้ารหัสข้อมูลและถอดรหัส ซึ่งสามารถแปลงได้ด้วยการใช้ Fractal Code สามารถใช้แปลงข้อมูลให้มีความละเอียดต่างๆ ตามที่ต้องการได้ตามรูปที่ 3.9 โดยมีวิธีการดังนี้

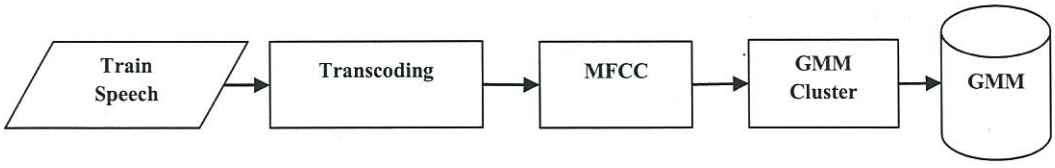
1. ทำการคำนวณหาขนาดของ Frame ที่จะแปลงไป
2. ทำการบีบอัดข้อมูล
3. ทำการประกอบข้อมูลกลับด้วยขนาดของ Frame ที่ได้คำนวณมา

รูปที่ 3. 9 การแปลงรหัสข้อมูล

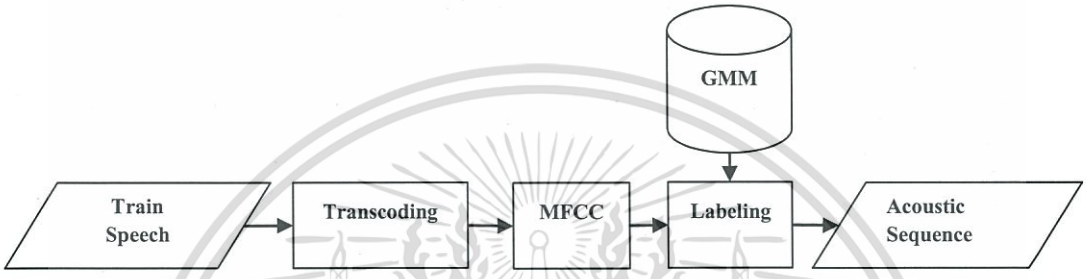
3.8 การรู้จำเสียงพูดร่วมกับการใช้การแปลงรหัสแบบทั้งไฟล์

การรู้จำเสียงพูดร่วมกับการใช้การแปลงรหัสแบบทั้งไฟล์ เป็นการนำกระบวนการแปลงรหัสไปใช้แปลงข้อมูลเสียงซึ่งจะทำการแปลงข้อมูลทั้งไฟล์โดยไม่ทำการแปลงเป็นเวกเตอร์ข้อมูลก่อน เพื่อให้มีอัตราสุ่มตัวอย่างที่ต้องการก่อนทำการรู้จำเสียงพูด มีกระบวนการหลักๆ คือ การ Training และ การ Classification โดยมีกระบวนการทำงานดังนี้

1. กระบวนการ Training ด้วยแบบจำลอง HMM ร่วมกับ GMM เพื่อสร้างแบบจำลองเชิงเสียง ดังรูปที่ 3.10 และแบบจำลองเชิงภาษา ดังรูปที่ 3.11 และ 3.12



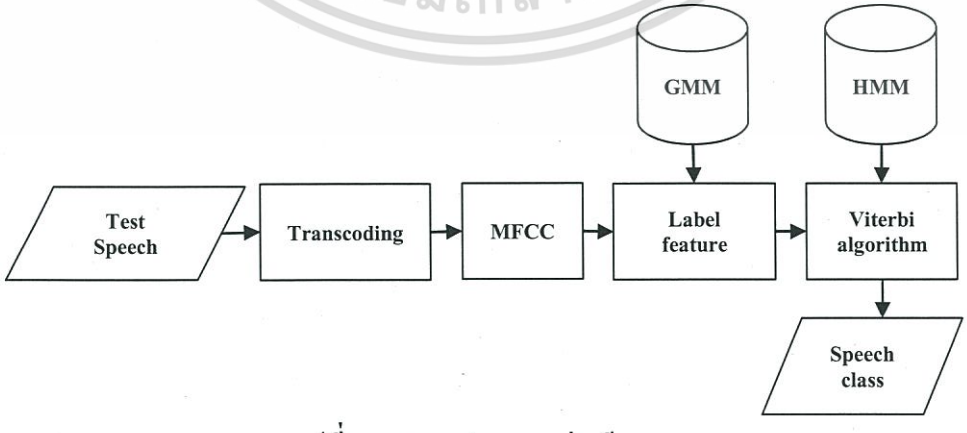
รูปที่ 3. 10 การสร้างแบบจำลองเกาส์เซียนเพื่อใช้เป็นฐานข้อมูลลักษณะเด่นของเสียง



รูปที่ 3. 11 การจำแนกลำดับของเสียงพูด

รูปที่ 3. 12 การสร้างแบบจำลองฮิดเดนมาร์คอฟ

2. กระบวนการ Classification เพื่อจำแนกกลุ่มของเสียงพูด โดยมีขั้นตอนการทำงานดังรูปที่ 3.13



รูปที่ 3. 13 การจำแนกกลุ่มเสียงพูด

3.9 การจำแนกกลุ่ม

ในการจำแนกกลุ่มของเสียงสามารถหาได้โดยการนำลักษณะเด่นของเสียงที่ต้องการทดสอบไปเข้าขั้นตอนวิธี Viterbi โดยมีตัวแปรนำเข้าคือ ลักษณะเด่นของเสียงที่จะทดสอบ และแบบจำลองเสียงของทุกเสียง ซึ่งผลลัพธ์จะได้เป็นค่าความน่าจะเป็นที่เสียงที่ใช้ทดสอบ จะเป็นเสียงเดียวกับแบบจำลองที่ใช้ ซึ่งค่าที่ได้จากแบบจำลองที่มีค่ามากที่สุด จะหมายถึงเสียงที่ใช้ทดสอบมีความคล้ายแบบจำลองนั้นที่สุด และเสียงนั้นจะเป็นเสียงเดียวกันกับแบบจำลองเสียงด้วยสมการที่ 3.3

$$W^* = \arg\ \max_W\ P(W|X) \tag{3.3}$$

โดยที่ W คือ แบบจำลองเสียงพูดของแต่ละคำ

และ X คือ ลำดับของลักษณะเด่นที่จะใช้ทดสอบ

3.10 การวัดผลลัพธ์

การวัดผลจะวัดความแม่นยำในการจำแนกกลุ่ม โดยคำนวณได้จากสมการที่ 3.4

$$\text{ความแม่นยำ} = \frac{\text{จำนวนเสียงที่จำแนกกลุ่มถูกต้อง}}{\text{จำนวนเสียงทั้งหมด}} \tag{3.4}$$

3.11 ตัวแปรที่ใช้ในการทดลอง

ตารางที่ 3.2 ค่าตัวแปรต่างๆ ที่ใช้ในการสร้างแบบจำลองเสียงพูดของคู่เทียบ

ตัวแปรที่ใช้ในการทดลอง	ค่าตัวแปร
จำนวน Cepstral Coefficients (MFCC)	12 มิติ
จำนวน Mixture Component (GMM)	96 ประกอบ
จำนวนสถานะในแบบจำลอง (HMM)	8 สถานะ
จำนวนไฟล์เสียงที่ใช้สร้างแบบจำลอง	180 ไฟล์
จำนวนไฟล์เสียงที่ใช้ทดสอบ	180 ไฟล์
ความยาวของเสียงต่อหนึ่งลักษณะเด่น (Speech Duration / Feature)	25 มิลลิวินาที

จากตารางที่ 3.2 โครงงานนี้ได้ใช้ค่าตัวแปรต่างๆสำหรับการสร้างแบบจำลองสียงพูดและการจำแนกกลุ่มโดยลักษณะเด่นแต่ละเวกเตอร์ คือ เสียงพูดที่ถูกตัดมาด้วยความยาวเสียง 10 มิลลิวินาที เพื่อสร้างเป็นชุดข้อมูลนำเข้าที่มีขนาดเวกเตอร์เท่ากับ 12 สำหรับแบบจำลอง และชุดข้อมูลนำเข้าที่มีขนาดเวกเตอร์เท่ากับ 12 สำหรับการจำแนกกลุ่ม ซึ่งแบบจำลองจะมีจำนวนองค์ประกอบผสม (Mixture Component) เท่ากับ 96 องค์ประกอบในแต่ละแบบจำลองเสียง และมีจำนวนสถานะการเปลี่ยนของเสียงในแต่ละกลุ่มของเสียงเป็น 8 สถานะ โดยแบบจำลองจะถูกสร้างขึ้นจากข้อมูลเสียงจำนวน 180 ไฟล์และถูกทดสอบด้วยข้อมูลเสียงคนละชุดกับที่ใช้สร้างแบบจำลองจำนวน 180 ไฟล์

ตารางที่ 3.3 ค่าตัวแปรต่างๆ ที่ใช้ในการสร้างแบบจำลองเสียงพูดของวิธีการที่นำเสนอ

ตัวแปรที่ใช้ในการทดลอง	ค่าตัวแปร
จำนวน Cepstral Coefficients (MFCC)	12 มิติ
จำนวน Mixture Component (GMM)	96 ประกอบ
จำนวนสถานะในแบบจำลอง (HMM)	31 สถานะ
จำนวนไฟล์เสียงที่ใช้สร้างแบบจำลอง	180 ไฟล์
จำนวนไฟล์เสียงที่ใช้ทดสอบ	180 ไฟล์
ความยาวของเสียงต่อหนึ่งลักษณะเด่น (Speech Duration / Feature)	25 มิลลิวินาที

จากตารางที่ 3.3 โครงงานนี้ได้ใช้ค่าตัวแปรต่างๆ สำหรับการสร้างแบบจำลองเสียงพูดและการจำแนกกลุ่มโดยลักษณะเด่นแต่ละเวกเตอร์ คือ เสียงพูดที่ถูกตัดมาด้วยความยาวเสียง 10 มิลลิวินาที เพื่อสร้างเป็นชุดข้อมูลนำเข้าที่มีขนาดเวกเตอร์เท่ากับ 12 สำหรับแบบจำลอง และชุดข้อมูลนำเข้าที่มีขนาดเวกเตอร์เท่ากับ 12 สำหรับการจำแนกกลุ่ม ซึ่งแบบจำลองจะมีจำนวนองค์ประกอบผสม (Mixture Component) เท่ากับ 96 องค์ประกอบในแต่ละแบบจำลองเสียง และมีจำนวนสถานะการเปลี่ยนของเสียงในแต่ละกลุ่มของเสียงเป็น 31 สถานะ โดยแบบจำลองจะถูกสร้างขึ้นจากข้อมูลเสียงจำนวน 180 ไฟล์และถูกทดสอบด้วยข้อมูลเสียงคนละชุดกับที่ใช้สร้างแบบจำลองจำนวน 180 ไฟล์

บทที่ 4
ผลการดำเนินงานวิจัย

4.1 การแปลงรหัสด้วย Fractal Code

ตารางที่ 4.1 อัตราสัญญาณต่อสัญญาณรบกวนของข้อมูลที่ได้จากการแปลงรหัส

อัตราสุ่มของชุดข้อมูล (ครั้งต่อวินาที)	อัตราสัญญาณต่อสัญญาณรบกวน (dB)
11,025	21.6
22,050	28.4
44,100	30.1

ในการทดลองนี้ใช้ข้อมูลชุดที่ 1 ซึ่งอัดเสียงเองในการทดลองหาค่าอัตราสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio, SNR) ในหน่วยเดซิเบล โดยหาค่าจากสัญญาณที่ถูกประกอบข้อมูลใหม่เทียบกับสัญญาณเดิมก่อนทำการเข้ารหัส ซึ่งค่ายิ่งมาก หมายความว่า สัญญาณที่ได้จากการประกอบข้อมูลกลับมีความสมบูรณ์มาก จากตารางจะเห็นได้ว่าค่า SNR ของข้อมูลที่มีอัตราสุ่มตัวอย่าง 22,050 และ 44,100 มีค่าใกล้เคียงกันมาก

4.2 ผลการทดลองรู้จำเสียงพูดด้วยข้อมูลที่มีอัตราสุ่มตัวอย่างแตกต่างกัน

4.2.1 แบบจำลองที่สร้างด้วย MFCC, GMM และ HMM

ในการทดลองนี้จะใช้ข้อมูลชุดที่ 1 ซึ่งมีสามแบบย่อยแตกต่างกันตามอัตราการสุ่มตัวอย่าง (Sampling Rate) ซึ่งข้อมูลทั้งสามแบบจะสุ่มด้วยความถี่ 44,100 22,050 และ 11,025 ครั้งต่อวินาที แต่ละแบบจะประกอบด้วยเสียงพูด 18 คำ คำละ 1 พยางค์ แต่ละคำจะถูกบันทึก 20 ครั้ง สำหรับใช้สร้างแบบจำจองจำนวน 10 ไฟล์ และใช้ทดสอบแบบจำลองอีก 10 ไฟล์ โดยใช้กระบวนการรู้จำด้วย MFCC ร่วมกับ GMM และ HMM

ตารางที่ 4.2 ผลลัพธ์การรู้จำโดยใช้แบบจำลองที่เรียนรู้ด้วยอัตราสุ่มตัวอย่างที่ต่างกัน เมื่อถูกทดสอบด้วยข้อมูลที่มีอัตราสุ่มตัวอย่างต่างกัน

อัตราการสุ่มตัวอย่างของข้อมูลที่ใช้สร้างแบบจำลอง (ครั้งต่อวินาที)	ข้อมูลทดสอบที่มีอัตราสุ่มตัวอย่างแตกต่างกัน (ครั้งต่อวินาที)		
	11,025	**22,050**	**44,100**
11,025	82.2%	78.3%	77.8%
22,050	79.4%	81.7%	86.1%
44,100	67.8%	70.6%	80.0%

จากตารางที่ 4.2 แสดงถึงประสิทธิภาพของแบบจำลองสียงพูด เมื่อใช้จำแนกกลุ่มเสียงพูดที่อัตราสุ่มแตกต่างกัน มีแนวโน้มเพิ่มขึ้นเมื่ออัตราสุ่มตัวอย่างของข้อมูลเสียงที่ถูกจำแนกมีอัตราสุ่มตัวอย่างใกล้เคียงกัน แต่มีค่าแตกต่างกันชัดเจน โดยเสียงพูดที่ใช้สร้างแบบจำลองแต่ละครั้งจะมีอัตราสุ่มแตกต่างกัน

4.2.2 แบบจำลองที่สร้างด้วย Fractal Code, MFCC, GMM และ HMM

ตารางที่ 4.3 ผลลัพธ์การทดลองเพิ่มประสิทธิภาพด้วย Fractal Code แบบบีบอัดทั้งไฟล์

อัตราสุ่มตัวอย่างตัวกลางของแบบจำลอง (ครั้งต่อวินาที)	ความถูกต้องกับข้อมูลชุด Test ด้วยอัตราสุ่มตัวอย่างแตกต่างกัน (ครั้งต่อวินาที)		
	11,025	**22,050**	**44,100**
11,025	56.7%	46.7%	32.2%
22,050	53.9%	79.4%	74.4%
44,100	35.6%	72.8%	76.7%

จากตารางที่ 4.3 แสดงถึงประสิทธิภาพของแบบจำลองสียงพูด เมื่อใช้จำแนกกลุ่มเสียงพูดที่อัตราสุ่มแตกต่างกัน มีแนวโน้มเพิ่มขึ้นเมื่ออัตราสุ่มตัวอย่างของข้อมูลเสียงที่ถูกจำแนกมีอัตราสุ่มตัวอย่างใกล้เคียงกัน โดยเสียงพูดที่ใช้สร้างแบบจำลองแต่ละครั้งจะมีอัตราสุ่มแตกต่างกัน

4.2.3 การเปรียบเทียบผลลัพท์ของวิธีการที่นำเสนอกับคู่เทียบ

ตารางที่ 4.4 ผลลัพธ์ของวิธีการที่นำเสนอเมื่อเทียบกับคู่เทียบ

อัตราสุ่มตัวอย่างเสียงของแบบจำลอง (ครั้งต่อวินาที)	ความถูกต้องกับข้อมูลชุด Test ด้วยอัตราสุ่มตัวอย่าง (ครั้งต่อวินาที)					
	11,025		22,050		44,100	
	คู่เทียบ	วิธีการที่นำเสนอ	คู่เทียบ	วิธีการที่นำเสนอ	คู่เทียบ	วิธีการที่นำเสนอ
11,025	82.2%	56.7%	78.3%	46.7%	77.8%	32.2%
22,050	79.4%	53.9%	81.7%	79.4%	86.1%	74.4%
44,100	67.8%	35.6%	70.6%	72.8%	80.0%	76.7%

จากตารางที่ 4.4 จะแสดงถึงความแม่นยำในการรู้จำเสียงพูดเมื่อใช้วิธีการของคู่เทียบ เทียบกับวิธีการที่นำเสนอ โดยใช้แบบจำลองที่สร้างด้วยข้อมูลดั้งเดิมสำหรับการทดสอบคู่เทียบ และใช้แบบจำลองที่สร้างด้วยข้อมูลที่ผ่านการทำการแปลงรหัสด้วย Fractal Code แล้วสำหรับทดสอบวิธีการที่นำเสนอ โดยจะใช้แบบจำลองที่สร้างจากข้อมูลที่มีอัตราสุ่มตัวอย่างแต่ละค่า กับ ข้อมูลที่มีอัตราสุ่มตัวอย่าง ต่างๆกัน คือ 11,025 ครั้งต่อวินาที, 22,050 ครั้งต่อวินาที และ 44,100 ครั้งต่อวินาที ความแม่นยำในการรู้จำเสียงพูดมีค่าใกล้กันมากขึ้นเมื่อใช้ Fractal Code กับข้อมูลที่มีอัตราสุ่มตัวอย่าง 22,050 และ 44,100 ครั้งต่อวินาที

บทที่ 5
สรุปผลและข้อเสนอแนะ

5.1 สรุปผล

การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นสำหรับการรู้จำเสียงมีเป้าหมายเพื่อลดความแตกต่างของความแม่นยำในการรู้จำเสียงพูดสำหรับข้อมูลเสียงพูดที่มีอัตราสุ่มตัวอย่างแตกต่างกัน โดยใช้เทคนิคการแปลงรหัสแบบ Fractal Code มาประยุกต์ใช้แก้ปัญหา พบว่าก่อนใช้เทคนิคการแปลงรหัส การรู้จำเสียงด้วยแบบจำลองที่สร้างจากข้อมูลที่มีอัตราสุ่มตัวอย่าง 44,100 ครั้งต่อวินาที มีความแม่นยำ 80.0% เมื่อใช้จำแนกกลุ่มข้อมูลเสียงที่มีอัตราสุ่มตัวอย่างเดียวกัน และมีความแม่นยำเป็น 70.6% เมื่อใช้จำแนกกลุ่มข้อมูลเสียงที่มีอัตราสุ่มตัวอย่าง 22,050 ครั้งต่อวินาที และมีความแม่นยำ 67.8% เมื่อใช้จำแนกกลุ่มเสียงพูดที่มีอัตราสุ่มตัวอย่าง 11,025 ครั้งต่อวินาที จะเห็นได้ว่าความแม่นยำในการรู้จำเสียงพูดแตกต่างกัน

แต่หลังจากที่ใช้เทคนิคการแปลงโค้ดด้วย Fractal Code แล้วพบว่า การรู้จำเสียงด้วยแบบจำลองที่สร้างจากข้อมูลที่มีอัตราสุ่มตัวอย่าง 44,100 ครั้งต่อวินาที มีความแม่นยำ 76.7% เมื่อใช้จำแนกกลุ่มข้อมูลเสียงที่มีอัตราสุ่มตัวอย่างเดียวกัน และมีความแม่นยำเป็น 72.8% เมื่อใช้จำแนกกลุ่มข้อมูลเสียงที่มีอัตราสุ่มตัวอย่าง 22,050 ครั้งต่อวินาที และมีความแม่นยำ 35.6% เมื่อใช้จำแนกกลุ่มเสียงพูดที่มีอัตราสุ่มตัวอย่าง 11,025 ครั้งต่อวินาที จะเห็นได้ว่า ความแตกต่างของความแม่นยำในการรู้จำเสียงพูดจะใกล้เคียงกันเมื่อจำแนกข้อมูลเสียงที่มีอัตราสุ่มตัวอย่าง 44,100 และ 22,050 ครั้งต่อวินาที แต่เมื่อจำแนกกลุ่มข้อมูลที่มีอัตราสุ่มตัวอย่าง 11,025 ครั้งต่อวินาที ความแม่นยำจะแตกต่างกันมาก

5.2 ข้อเสนอแนะ

ในการวิจัยนี้จะใช้วิธีการแบ่งข้อมูลในขั้นตอนการเข้ารหัสที่แบ่งแบบเท่ากันทั้งสัญญาณ ส่งผลทำให้ประสิทธิภาพต่ำ เนื่องจากการแบ่งข้อมูลแต่ละบล็อกให้เท่ากัน อาจจะมีค่าความแปรปรวนของข้อมูลแตกต่างกันมากตามข้อมูลที่มี ในโครงการนี้เลือกขนาดของบล็อคที่จะแบ่งแบบค่าคงที่ซึ่งไม่อาจจะใช้กับข้อมูลได้ทุกชุด ซึ่งจะเห็นได้จากการทดลองกับข้อมูลที่มีอัตราสุ่มตัวอย่าง 11,025 ครั้งต่อวินาที จะได้ผลลัพธ์น้อยกว่าการทดลองกับข้อมูลที่มีอัตราสุ่มตัวอย่าง 22,050 และ 44,100 ครั้งต่อวินาทีอย่างมาก ซึ่งการเข้ารหัสแบบ Fractal Code ที่ดีควรจะเข้ารหัสด้วยบล็อคที่แบ่งให้มีขนาดแตกต่างกันตามความแปรปรวนของข้อมูล โดยใช้วิธีการต่างๆ เช่น การแบ่งข้อมูลโดยใช้ Quadtree Decomposition

บรรณานุกรม

[1] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," **Proceedings of the IEEE**, vol. 77, no. 2, pp. 257-286, February 1989.

[2] D. B. Paul, "Speech recognition using hidden markov models," **The Lincoln Laboratory Journal**, vol. 3, no. 1, pp. 41-62, 1990.

[3] K. Samudravijaya, "Automatic Speech Recognition," 2009.

[4] C. Ssnderson and K. K. Paliwal, "Effect of different sampling rates and feature vector sizes on speech recognition performance," in , **Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications**, 1997.

[5] H.-G. Hirsch, K. Hellwig and S. Dobler, "Speech recognition at multiple sampling rates," in **EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event**, 2001, Aalborg, Denmark, September 3-7, 2001.

[6] S. K. Kopparapu and M. Laxminarayana, "Choice of Mel filter bank in computing MFCC of a resampled speech," in **2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)**, 2010.

[7] S. K. Kopparapu and K. K. Bhuvanagiri, "Recognition of subsampled speech using a modified Mel filter bank," **Computers \& Electrical Engineering**, vol. 39, no. 2, pp. 655-662, February 2013.

[8] A. Jacquin, "Fractal image coding: a review," **Proceedings of the IEEE**, vol. 81, no. 10, pp. 1451-1465, October 1993.

[9] X. Huang and K.-F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," **Speech and audio processing, ieee transactions on**, vol. 1, no. 2, pp. 150-157, 1993.

[10] N. Jittiwarangkul, S. Jitapunkul, S. Luksaneeyanavin, V. Ahkuputra and C. Wutiwiwatchai, "Thai syllable segmentation for connected speech based on energy," in **Circuits and Systems, 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on**, 1998.

[11] N. Thubthong, B. Kijsirikul and S. Luksaneeyanawin, "Stress and tone recognition of polysyllabic words in Thai speech," in **International Conference on Intelligent Technologies (INTECH)**, Bangkok, 2001.

[12] C.-T. Lu, "Noise reduction using three-step gain factor and iterative-directional-median filter," **Applied Acoustics**, vol. 76, pp. 249-261, February 2014.

[13] M. Bhaykar, J. Yadav and K. S. Rao, "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM," in **Communications (NCC), 2013 National Conference on**, 2013.

[14] D. Su, X. Wu and L. Xu, "GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection," in **Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on**, 2010.

[15] I. Ismail, A. Hamdy and R. Frig, "Studying the effect of down sampling and spatial interpolation on fractal image compression," in **2010 International Conference on Computer Engineering and Systems (ICCES)**, 2010.

[16] S. Poobal and G. Ravindran, "Analysis on the effect of tolerance criteria in fractal image compression," in **IEEE International Workshop on Imaging Systems and Techniques**, 2005, 2005.

[17] R. Guido, L. Vieira, S. Barbon Junior, F. Sanchez, M. Guilherme, K. Sergio, T. Scarpa, E. Fonseca, J. Pereira and M. Monteiro, "A Fractal and Wavelet-Based Approach for Audio Coding.," in **Eighth IEEE International Symposium on Multimedia, 2006. ISM'06**, 2006.

[18] F. Li and H. Hermansky, "Effect of filter bandwidth and spectral sampling rate of analysis filterbank on automatic phoneme recognition," in **2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2013.

[19] T. L. Nwe, S. W. Foo and L. C. De Silva, "Speech emotion recognition using hidden Markov models," **Speech communication**, vol. 41, no. 4, pp. 603-623, 2003.

[20] K. Nosirov, I. Gavrilov and A. Abduazizov, "The fractal method of compression of broadband audio signals," in **2010 4th International Conference on Application of Information and Communication Technologies (AICT)**, 2010.

[21] L. Bu and T.-D. Church, "Perceptual speech processing and phonetic feature mapping for robust vowel recognition," **IEEE Transactions on Speech and Audio Processing**, vol. 8, no. 2, pp. 105-114, March 2000.

[22] G. Davis, "A wavelet-based analysis of fractal image compression," **IEEE Transactions on Image Processing**, vol. 7, no. 2, pp. 141-154, February 1998.

ประวัติผู้เขียน

ผู้เขียน	นายรัฐพล หอกกิ่ง
วันเดือนปีเกิด	14 มิถุนายน 2535
สถานที่เกิด	จังหวัด นครราชสีมา
ปริญญา	วิทยาศาสตรบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประวัติการทำงาน	ไม่มี

การเพิ่มประสิทธิภาพ
การสกัดลักษณะเด่นสำหรับการรู้จำเสียงพูด

รัฐพล หอกกิ่ง

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ

Emails: ratthapon.h@gmail.com

บทคัดย่อ

ปัจจุบันมีโปรแกรมประยุกต์ต่างๆ เช่น เลขาส่วนบุคคล โปรแกรมค้นหาและโปรแกรมแปลภาษา เริ่มรับข้อมูลนำเข้าจากเสียงพูดของผู้ใช้เพื่อทำให้ง่ายต่อการใช้งานและสามารถสั่งงานได้อย่างรวดเร็วและสะดวกสบาย แต่คอมพิวเตอร์ไม่สามารถเข้าใจข้อมูลเสียงพูดที่ป้อนให้ได้ทันที ดังนั้นกระบวนการรู้จำเสียงพูดจึงถูกพัฒนาเพื่อทำให้คอมพิวเตอร์สามารถเข้าใจและจำแนกเสียงพูดของมนุษย์ได้ ซึ่งโปรแกรมประยุกต์จำนวนมากทำงานบนอุปกรณ์ที่หลากหลายและมีคุณภาพในการบันทึกเสียงพูดต่างกันขึ้นอยู่กับอัตราสุ่มของเสียงที่บันทึก ส่งผลให้ประสิทธิภาพในการรู้จำเสียงพูดแตกต่างกัน [1] [2] โครงงานนี้จึงมุ่งเน้นไปที่การเพิ่มประสิทธิภาพของการสกัดลักษณะเด่นจากเสียงพูดซึ่งเป็นข้อมูลตัวแทนสียงสำหรับกระบวนการรู้จำเสียงพูดจากเสียงที่มีอัตราสุ่มแตกต่างกัน ซึ่งจะส่งผลให้สามารถรู้จำเสียงพูดจากข้อมูลจากข้อมูลนำเข้าที่มีอัตราสุ่มแตกต่างกันได้อย่างมีประสิทธิภาพ

คำสำคัญ – การรู้จำเสียงพูด; การสกัดลักษณะเด่น; การเพิ่มประสิทธิภาพ; อัตราสุ่มแตกต่างกัน

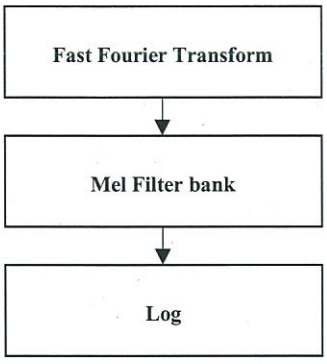
1. บทนำ

การรู้จำเสียงพูดจะมีสองกระบวนการหลัก คือ การสร้างแบบจำลองเสียงพูด และการจำแนกกลุ่มเสียงพูด ในการสร้างแบบจำลองมีขั้นตอนการสร้าง คือ 1.สกัดลักษณะเด่น และ 2.การฝึกแบบจำลอง ในการสกัดลักษณะเด่นจะใช้ลักษณะเด่นของเสียงแทนการใช้ข้อมูลเสียงโดยตรงซึ่งลักษณะเด่นที่นิยมใช้กันอย่างแพร่หลายในการรู้จำเสียงพูด คือ ลักษณะเด่น แบบ Mel Frequency Cepstral Coefficients (MFCC) [3]โดยใช้ร่วมกับแบบจำลอง Hidden Markov ร่วมกับแบบจำลอง Gaussian Mixture (HMM-GMM) [4][5] ในการสร้างแบบจำลองเสียงพูดและทำการจำแนกกลุ่มสียงพูดด้วยขั้นตอนวิธี Viterbi ซึ่งการสกัดลักษณะเด่นจะถูกเพิ่มประสิทธิภาพด้วยการใช้ Fractal code [6] ร่วมกับวิธีเดิม ซึ่งจะทำให้ได้ลักษณะเด่นที่ทนทานต่ออัตราสุ่มตัวอย่างที่แตกต่างกันได้

ในการรู้จำเสียงพูด จะใช้ลักษณะเด่นเพื่อเป็นตัวแทนของข้อมูลเสียงพูดเป็นข้อมูลนำเข้าสำหรับทำการรู้จำ ลักษณะเด่นที่ใช้คือ MFCC ซึ่งเป็นลักษณะเด่นที่ถูกใช้อย่างแพร่หลายในงานด้านการรู้จำเสียงพูด ในการสร้างลักษณะเด่น จะมีขั้นตอนดังนี้ ทำการแบ่งข้อมูลเสียงพูดที่อยู่ในรูปคลื่นออกเป็นกรอบข้อมูล (frame) เพื่อพิจารณาข้อมูลสียงทีละส่วน โดยในแต่ละส่วนจะมีข้อมูลสียงที่มีความยาวเท่าๆกันขนาด 25 มิลลิวินาทีซึ่งข้อมูลจะถูกเลือกมาทุกๆ 10 มิลลิวินาทีแล้วทำการแปลงข้อมูลทีละกรอบด้วยการแปลงฟูเรียแบบเร็วจะได้ค่าพลังงานของความถี่ต่างๆ แล้วทำการกรองด้วยตัวกรอง Mel filter bank ซึ่งเป็นตัวกรองแบบสามเหลี่ยม กรองเอาข้อมูลที่อยู่ในย่านเสียงพูดออกมา และทำการปรับค่าความต่าง (contrast) ด้วยการแปลงแบบลอการิทึม ดังรูปที่ 1 ผลลัพธ์ที่ได้คือลักษณะเด่นที่เป็นตัวแทนเสียง

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การสกัดลักษณะเด่นแบบ Mel Frequency Cepstral Coefficients (MFCC)

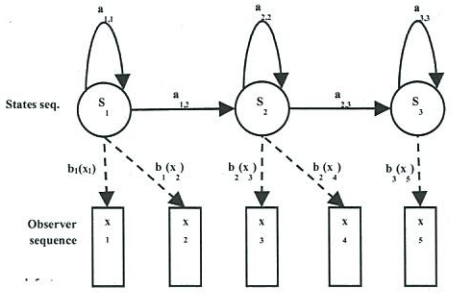


รูปที่ 1. ขั้นตอนการสกัด MFCC

2.2 แบบจำลอง Hidden Markov (HMM)

สำหรับการสร้างแบบจำลองเพื่ออธิบายลักษณะของเสียงพูด จะใช้แบบจำลอง Hidden Markov ร่วมกับแบบจำลอง Gaussian Mixture คำพูดแต่ละคำจะประกอบขึ้นจากการออกเสียงช่วงสั้นๆ ที่แตกต่างกันต่อเนื่องตามลำดับจึงเกิดเป็นคำพูด ซึ่งสามารถใช้แบบจำลอง Markov อธิบายลำดับของการออกเสียงสำหรับแต่ละคำได้ โดยแบบจำลองจะประกอบไปด้วยสถานะ ที่ใช้แทนการออกสียงแต่ละแบบ และความน่าจะเป็นในการเปลี่ยนสถานะของแต่ละสถานะ

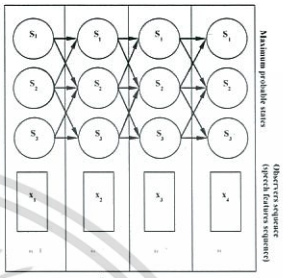
ในการอธิบายลำดับของการออกเสียง แบบจำลองจะไม่สามารถรู้ลำดับของสถานะและความน่าจะเป็นในการเปลี่ยนสถานะได้ จึงต้องค้นหาลำดับสถานะและความน่าจะเป็นในการเปลี่ยนสถานะจากการสังเกตลำดับของผู้สังเกต (observer) ในการรู้จำเสียงลำดับของผู้สังเกต คือ ลำดับของลักษณะเด่น จึงเรียกแบบจำลองมาร์คอฟที่ไม่สามารถรู้สถานะในแบบจำลองว่า แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov) และสถานะของลักษณะเด่นที่มีค่าแบบต่อเนื่องสามารถหาได้จากแบบจำลองส่วนผสมเกาส์เซียน (Gaussian Mixture Model)



รูปที่ 2. แบบจำลองฮิดเดนมาร์คอฟ

2.3 ขั้นตอนวิธี Viterbi

ในการค้นหาความน่าจะเป็นของลำดับของสถานะจะใช้ขั้นตอนวิธี Viterbi (Viterbi Algorithm) หาลำดับความน่าจะเป็นสูงที่สุดของลำดับลักษณะเด่นเทียบกับแบบจำลองฮิดเดนมาร์คอฟ ในการสร้างแบบจำลองฮิดเดนมาร์คอฟที่ให้ความน่าจะเป็นสูงสุดสำหรับเสียงพูดแต่ละเสียงจะใช้ขั้นตอนวิธี Viterbi ในการฝึกฝนแบบจำลองด้วย



รูปที่ 3. เส้นทางการเปลี่ยนสถานะโดย Viterbi Algorithm

2.4 การเข้ารหัสแบบ Fractal code

การเข้ารหัสด้วย Fractal code ถูกนำมาใช้เป็นเครื่องมือสำหรับการเพิ่มประสิทธิภาพในการสกัดลักษณะเด่น ซึ่งการข้ารหัสเสียงด้วยวิธีนี้มีคุณสมบัติอย่างหนึ่ง คือ ไม่ขึ้นอยู่กับความละเอียดของข้อมูล และสามารถถอดรหัสข้อมูลกลับให้มความละเอียดเท่าที่ต้องการได้

ในการเข้ารหัสแบบ Fractal code มีขั้นตอนดังนี้

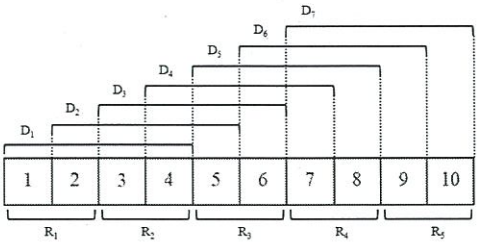
1.แบ่งข้อมูลออกเป็น Domain block และ Range block

2.นำ Range block แต่ละ block ไปทำการคำนวณหาค่า self-similarity จาก Domain block ทั้งหมด

3.ทำการคำนวณหาค่า Contrast, Amplitude, ตำแหน่งของ Domain ที่ให้ค่าความคล้ายมากที่สุดสำหรับการแปลงข้อมูลกลับ

2.5 การประกอบข้อมูลใหม่ (Reconstruction)

การประกอบข้อมูลใหม่ เป็นการใช้ Fractal code ที่หาได้มาทำการสร้างข้อมูลใหม่ด้วย Fractal code ที่คำนวณได้ร่วมกับ Iterative Function ซึ่ง Fractal code มีคุณสมบัติพิเศษ คือ สามารถสร้างข้อมูลที่มีความละเอียดไม่เท่ากับข้อมูลต้นฉบับก่อนทำการบีบอัดก็ได้ (Resolution Independence)



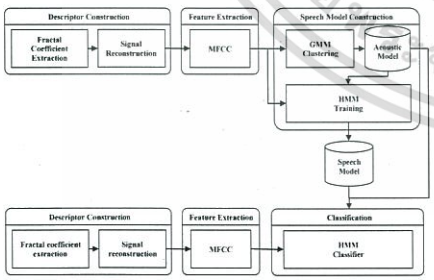
รูปที่ 4. การแบ่งข้อมูลออกเป็น block ก่อนทำการเข้ารหัส

3. การดำเนินงานวิจัย

ในการรู้จำเสียงพูด ประกอบด้วย 2 กระบวนการหลัก คือ กระบวนการสร้างแบบจำลองเสียงพูด และการจำแนกกลุ่ม

ในกระบวนการสร้างแบบจำลองเสียงพูดมี 3 ขั้นตอน คือ ขั้นตอนการสร้างตัวกลางมาตรฐานจาก Fractal code ของข้อมูลเสียงพูดที่มีอัตราสุ่มตัวอย่างต่างๆ ขั้นตอนการสกัดลักษณะเด่น Mel Frequency Cepstral Coefficients จากตัวกลาง ขั้นตอนการสร้างแบบจำลองลำดับเสียงพูดโดยใช้แบบจำลองฮิดเดนมาร์คอฟ ร่วมกับแบบจำลองเชิงเสียงโดยใช้แบบจำลองส่วนผสมเกาส์เซียน (Gaussian Mixture Model)

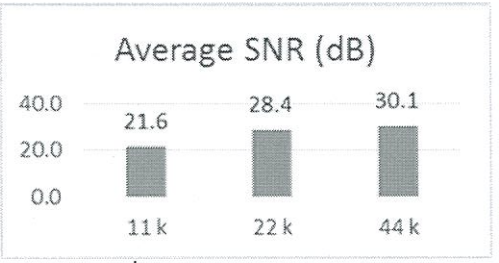
ในกระบวนการจำแนกกลุ่ม จะประกอบด้วย 3 ขั้นตอน คือ ขั้นตอนการสร้างตัวกลางมาตรฐานจาก Fractal code ของข้อมูลเสียงพูดที่มีอัตราสุ่มต่างๆ ขั้นตอนการสกัดลักษณะเด่น Mel Frequency Cepstral Coefficients และการจำแนกกลุ่มโดยใช้ขั้นตอนวิธี Viterbi ค้นหาลำดับเสียงที่มีความน่าจะเป็นมากที่สุดจากแบบจำลองฮิดเดนมาร์คอฟ



รูปที่ 5. กรอบแนวความคิด

4.ผลการดำเนินงานวิจัย

จากการใช้ Fractal code เพื่อสร้างตัวแทนของเสียง พบว่าสามารถแปลงข้อมูลเสียงกลับคืนโดยมีค่า Signal to noise ratio ที่ค่อนข้างสูง สำหรับข้อมูลนำเข้า 22 kHz และ 44 kHz ซึ่งมีค่าประมาณ 30 dB แสดงว่า Fractal code สามารถทำงานได้มีประสิทธิภาพ



รูปที่ 6. ความสมบูรณ์ของข้อมูลตัวกลาง

ในการรู้จำเสียงพูดจะทดสอบโดยวัดความแม่นยำจากค่า accuracy rate ซึ่งได้จากจำนวนเสียงที่จำแนกกลุ่มถูกหารด้วยจำนวนเสียงทั้งหมด

ตารางที่ 1. ความแม่นยำในการรู้จำเสียงพูดแบบเดิม

แบบจำลองข้อมูล	ข้อมูลที่ถูกนำมาทดสอบ		
	11 kHz	22 kHz	44 kHz
model 11 kHz	82.2%	78.3%	77.8%
model 22 kHz	79.4%	81.7%	86.1%
model 44 kHz	67.8%	70.6%	80.0%

ตารางที่ 2. ความแม่นยำในการรู้จำเสียงพูดที่ถูกเพิ่มประสิทธิภาพ

แบบจำลองข้อมูล	ข้อมูลที่ถูกนำมาทดสอบ		
	11 kHz	22 kHz	44 kHz
model 11 kHz	56.7%	46.7%	32.2%
model 22 kHz	53.9%	79.4%	74.4%
model 44 kHz	35.6%	72.8%	76.7%

5.สรุปผลการวิจัย

จากการวิจัยพบว่า สามารถใช้ Fractal code สร้างสัญญาณที่เป็นตัวกลางที่มีคุณภาพเดียวกันได้ โดยมีความสมบูรณ์ของข้อมูลอยู่ในระดับที่รับได้ที่ SNR มากกว่า 20 dB จากรูปที่ 6. ในการรู้จำเสียงพบว่า ความถูกต้องในการรู้จำเสียงเมื่อใช้ Fractal code ร่วมด้วย ทำให้ความต่างของความแม่นยำของข้อมูลที่มีอัตราสุ่ม 22 kHz และ 44 kHz โดยเฉลี่ยมีค่าใกล้เคียงกันไม่ว่าจะสร้างแบบจำลองด้วย

อัตราสุ่มใดก็ตามและทดสอบแบบจำลองด้วยข้อมูลที่มีอัตราสุ่มใดก็ตาม แต่อย่างไรก็ตาม ผลการทดลองใช้ Fractal code กับข้อมูลที่มีอัตราสุ่มต่ำ 11 kHz กลับพบว่าไม่สามารถสร้างสัญญาณที่เป็นตัวกลางได้สมบูรณ์เท่าส่งผลให้ความแม่นยำในการรู้จำต่ำ จึงสรุปได้ว่าสามารถใช้กระบวนการสร้างตัวแทนเสียงเพื่อเพิ่มประสิทธิภาพในการสกัดลักษณะเด่นได้

เอกสารอ้างอิง

[1] C. Ssnderson and K. K. Paliwal, "Effect of different sampling rates and feature vector sizes on speech recognition performance," in , *Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications*, 1997.

[2] H.-G. Hirsch, K. Hellwig and S. Dobler, "Speech recognition at multiple sampling rates," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001*, 2001.

[3] K. Samudravijaya, "Automatic Speech Recognition," 2009.

[4] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, no. 2, pp. 257-286, February 1989.

[5] D. B. Paul, "Speech recognition using hidden markov models," *The Lincoln Laboratory Journal,* vol. 3, no. 1, pp. 41-62, 1990.

[6] A. Jacquin, "Fractal image coding: a review," *Proceedings of the IEEE,* vol. 81, no. 10, pp. 1451-1465, October 1993.