การรู้จำเสียงตัวเลขต่อเนื่องภาษาไทยแบบไม่ขึ้นกับบุคคลโดยใช้ฮิดเดนมาคอร์ฟโมเดล

Speaker-Independent Thai Connected Digit Speech Recognition System using Hidden Markov Model

คำนำ

การติดต่อสื่อสารด้วยเสียงพูดเป็นวิธีการติดต่อสื่อสารที่สะดวกและเป็นธรรมชาติที่สุดวิธีหนึ่งสำหรับมนุษย์ เนื่องจากความก้าวหน้าด้านเทคโนโลยีคอมพิวเตอร์และระบบสื่อสาร ทำให้การติดต่อด้วยการพูดไม่เพียงแต่จะใช้ในโอกาสที่พบกันเท่านั้น นอกจากนี้ยังใช้ติดต่อผ่านทางโทรศัพท์มีสายและไร้สาย รวมถึงการติดต่อด้วยเสียงพูดผ่านอินเตอร์เน็ตซึ่งเป็นที่แพร่หลายในชีวิตประจำวัน ความต้องการในการสั่งงานด้วยเสียงพูดผ่านคอมพิวเตอร์และเครื่องมือสื่อสารต่างๆ จึงมีมากขึ้น การรู้จำเสียงพูด(Speech Recognition) เป็นงานวิจัยหลักที่จะทำให้คอมพิวเตอร์สามารถเข้าใจเสียงพูดของมนุษย์ได้ งานวิจัยด้านการรู้จำเสียงพูดเริ่มมีการพัฒนาตั้งแต่ปี ค.ศ. 1950 ซึ่งเป็นระบบรู้จำเสียงพูดแบบคำเดี่ยวสำหรับบุคคลเดียว ในช่วงทศวรรษที่ 80 ได้มีระบบรู้จำเสียงพูดที่ให้ประสิทธิภาพดีเกิดขึ้นหลายงานด้วยกัน รวมถึงเป็นช่วงเวลาที่สำคัญในการพัฒนาระบบรู้จำเสียงพูดแบบต่อเนื่อง (Rabiner, 1993) นอกจากนี้ในภาษาอื่นๆ เช่นภาษา ญี่ปุ่น ภาษาจีน ภาษาฝรั่งเศส ก็มีงานวิจัยด้านนี้มานานเช่นกัน ในปัจจุบันมีระบบรู้จำเสียงพูดที่สามารถนำไปใช้งานได้จริงในภาษาอังกฤษและมีการจำหน่ายในเชิงพาณิชเช่นผลิตภัณฑ์ของบริษัทไอบีเอ็ม(IBM) บริษัทดรากอน (Dragon) เป็นต้น ในการสร้างระบบรู้จำเสียงพูดให้มีประสิทธิภาพสูงนั้นมีปัญหาที่เกี่ยวข้องอยู่หลายประการโดยสามารถแบ่งได้ดังนี้

1) ผู้พูดที่ระบบสามารถรู้จำได้ แบ่งออกได้เป็นแบบขึ้นกับผู้พูด (Speaker Dependent) และแบบไม่ขึ้นกับผู้พูด (Speaker Independent) ระบบรู้จำเสียงพูดแบบขึ้นกับผู้พูดเป็นระบบที่สามารถรู้จำได้เฉพาะบุคคลคนเดียว ส่วนระบบการรู้จำเสียงพูดที่สามารถรู้จำเสียงพูดแบบไม่ขึ้นกับผู้พูดจะยากกว่าแบบขึ้นกับผู้พูดอย่างชัดเจนทั้งนี้เนื่องจากลักษณะโครงสร้างทางกายวิภาค(anatomical structures)ของทางเดินเสียงของแต่ละบุคคล และวิธีการพูด(speaking styles) ที่แตกต่างกันออกไป เช่น ในคำพูดเดียวกันบางคนอาจจะพูดเร็วแต่อีกคนอาจจะพูดช้า

2) จำนวนคำพูดที่ระบบสามารถรู้จำได้ (Vocabulary size) รู้จำสียงพูดบางระบบอาจจะสามารถรู้จำจำนวนคำพูดได้ไม่มากเช่น 10 คำ หรือ จำนวนมากเช่น 10,000 คำ เป็นต้น ระบบที่รู้จำได้ไม่เกิน100 คำถือว่าเป็นระบบที่รู้จำจำนวนคำได้น้อย (Small vocabulary) ส่วนระบบรู้จำที่สามารถรู้จำได้มากกว่า 10,000 คำเป็นระบบที่รู้จำได้จำนวนใหญ่มาก (Very large vocabulary) ระบบที่เพิ่มจำนวนคำศัพท์ที่สามารถรู้จำได้ให้มากขึ้นยิ่งจะมีโอกาสเกิดข้อผิดพลาดมากขึ้นนั่นคือประสิทธิภาพของระบบจะลดลง

3) รูปแบบของเสียงพูด (Speaking format) แบ่งได้เป็นเสียงพูดแบบคำเดี่ยว (Isolated Word) และคำพูดแบบต่อเนื่อง (Continuous Speech) การรู้จำเสียงพูดแบบคำเดี่ยวจะรับเสียงพูดทีละคำหรือพูดต่อกันไปโดยมีช่องว่างเป็นเสียงเงียบอย่างน้อย 0.2 วินาที ดังนั้นจะไม่ปัญหาขอบเขตของเสียงพูดแต่ละคำ ในส่วนของระบบรู้จำเสียงพูดแบบต่อเนื่องผู้พูดจะพูดเป็นธรรมชาติมากขึ้นและใช้งานสะดวกขึ้นโดยระบบสามารถรับคำพูดต่อเนื่องเป็นประโยคได้ สาเหตุที่การรู้จำเสียงพูดแบบต่อเนื่องมีความยากกว่าการรู้จำเสียงแบบคำเดี่ยวเนื่องจากขอบเขตที่ไม่แน่นอนของแต่ละคำ (Unclear word boundary) การเชื่อมต่อกันของคำพูด(Co-articulation effects) รวมถึงการลดทอนของหน่วยเสียง(Phoneme reduction) ซึ่งอาจเกิดจากความต่อเนื่องในการพูด

ในงานวิจัยนี้เป็นการรู้จำเสียงพูดตัวเลขภาษาไทยต่อเนื่องแบบไม่ขึ้นกับบุคคลซึ่งการรู้จำเสียงพูดตัวเลขแบบคำต่อเนื่องเป็นงานวิจัยที่สำคัญด้านหนึ่งของการรู้จำเสียงพูดเนื่องจากสามารถนำไปประยุกต์ใช้งานได้ในหลายเช่น การหมุนตัวเลขโทรศัพท์ด้วยเสียงพูด สำหรับงานวิจัยด้านนี้ในภาษาอังกฤษมีมานานและผลลัพธ์ที่ได้ให้ค่าความความถูกต้องสูง (Rabiner *et al.*, 1989; Wilpon *et al*., 1991 ; Ramesh *et al.*, 1992; Haeb-Umbach *et al*., 1993; Normandin *et al*., 1994). นอกจากนี้ยังมีงานวิจัยด้านการรู้จำเสียงพูดตัวเลขแบบต่อเนื่องในภาษาอื่นๆเช่น ภาษาญี่ปุ่น(Kawai and Higuchi, 1998) ภาษาจีนกลาง (Shyu *et al*., 1993), ภาษาเกาหลี (Choi *et al*., 1999), ภาษาเดนมาร์ก(Jacobsen and Wilpon, 1993) และภาษาอิตาลี (Cosi *et al*., 2000) เป็นต้น

ระบบรู้จำเสียงโดยทั่วไปประกอบด้วยส่วนสำคัญ 5 ส่วนได้แก่ การประมวลผลสัญญาณเสียงพูดเบื้องต้น (Speech Signal Preprocessing) การดึงค่าลักษณะสำคัญของเสียง (Speech Feature Extraction) การฝึกฝนรูปแบบ (Pattern training) การจำแนกรูปแบบ (Pattern classification) และการหาคำตอบ (Decoding process) การประมวลผลสัญญาณเสียงพูดเบื้องต้นทำหน้าที่ประมวลผล

เบื้องต้นเสียงโดยทำการปรับข้อมูลและทำการวางกรอบสัญญาณเสียง การดึงค่าคุณลักษณะสำคัญ ทำหน้าที่หาลักษณะสำคัญที่สามารถใช้เป็นตัวแทนของเสียงพูด การฝึกฝนรูปแบบทำหน้าที่ในการสอนระบบให้รู้จักลักษณะสำคัญของเสียงแต่ละเสียง การจำแนกรูปแบบทำหน้าที่วิเคราะห์เปรียบเทียบค่าคุณลักษณะสำคัญว่าเป็นเสียงใด และการหาคำตอบเป็นการหาลำดับที่ดีที่สุดเพื่อเป็นคำตอบ ในงานวิจัยนี้ได้เลือกใช้คุณลักษณะสำคัญคือ สัมประสิทธิ์เซปสตรัมบนสเกลเมล(Mel-frequency Cepstrum Coefficient-MFCC) ผลต่างสัมประสิทธิ์เซปสตรัมบนสเกลเมลทางเวลา (Differential Mel-frequency Cepstrum Coefficient) ในลำดับที่หนึ่ง ลำดับที่สอง และค่าพลังงาน (Energy) สำหรับการจำแนกรูปแบบเพื่อสร้างแบบจำลองของเสียงพูดใช้ทฤษฎีฮิดเดนมาร์คอฟแบบต่อเนื่อง(Continuous Density Hidden Markov Model-CDHMM)

วัตถุประสงค์

1. ศึกษาและพัฒนาระบบรู้จำเสียงพูดตัวเลขภาษาไทยต่อเนื่องแบบไม่ขึ้นกับบุคคล
2. หาจำนวนองค์ประกอบของลักษณะสำคัญของเสียงพูด จำนวนกลุ่มของเกาส์เซียนของแบบจำลองฮิดเดนมาคอร์ฟแบบต่อเนื่อง และจำนวนสถานะของแบบจำลองที่เหมาะสมกับการรู้จำ

ขอบเขตของโครงงาน

1. มีลักษณะเป็นเสียงพูดต่อเนื่อง พูดด้วยความเร็วปกติ บันทึกเสียงในสภาพแวดล้อมที่มีสัญญาณรบกวนน้อย เช่น ที่ทำงาน ที่บ้าน
2. ผู้พูดอายุระหว่าง 20-31 ปี
3. คำศัพท์ที่ระบบรู้จำได้คือตัวเลขภาษาไทย จำนวน 10 ตัวเลข

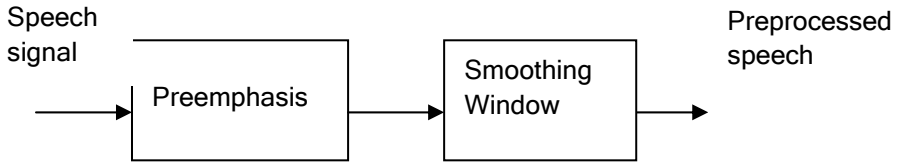
การตรวจเอกสาร

ในบทนี้อธิบายถึง ทฤษฎีพื้นฐานในการสร้างระบบรู้จำเสียง เพื่อทำให้เข้าใจกระบวนการรู้จำ ประกอบด้วย 4 หัวข้อ คือ การประมวลผลสัญญาณเสียงพูดเบื้องต้น การดึงค่าคุณลักษณะสำคัญและ ทฤษฎีสำหรับการรู้จำ และส่วนสุดท้ายกล่าวถึงงานวิจัยที่เกี่ยวข้อง

การประมวลผลสัญญาณเสียงพูดเบื้องต้น

การประมวลผลสัญญาณเสียงพูดเบื้องต้นเป็นขั้นตอนในการจัดเตรียมข้อมูลจากข้อมูลดิบของเสียงพูดที่ได้จากการบันทึกเสียงนำมาผ่านกรรมวิธีประมวลผลสัญญาณดิจิตอล เพื่อใช้ในการประมวลผลในขั้นตอนต่อไป เนื่องจากสัญญาณเสียงพูดโดยรวมจะเปลี่ยนตามเวลาและไม่คงตัว (Non-stationary) จึงต้องแบ่งสัญญาณเสียงพูดออกเป็นส่วนย่อย โดยส่วนย่อยแต่ละส่วนมีความยาวประมาณ 15-45 มิลลิวินาที (Rabiner,1993) ซึ่งสามารถนิยามให้ส่วนย่อยนี้เป็นสัญญาณที่ไม่แปรเปลี่ยนตามเวลาพร้อมที่จะคำนวณค่าทางสัญญาณเสียงพูดต่อไป

ขั้นตอนในการประมวลผลสัญญาณเสียงพูดเบื้องต้นประกอบด้วย 2 ขั้นตอนคือ ขั้นตอนการเน้นล่วงหน้า (Preemphasis) และ ขั้นตอนการวางกรอบขนาดสัญญาณ (Smoothing Window) ดังรูปที่ 1



ภาพที่ 1 การประมวลผลสัญญาณเสียงพูดเบื้องต้น

1. กรรมวิธีเน้นล่วงหน้า(Pre-emphasis)

มีผลทำให้อัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าสูงขึ้น(Signal to Noise Ratio) โดยนำสัญญาณเสียงพูดผ่านวงจรกรองแบบดิจิตอลอันดับหนึ่ง (First-Order Digital Filter) ค่าดังนี้

$$H(z) = 1 - az^{-1} \qquad 0 \le a \le 1$$

โดยที่ a คือค่าเป็นค่าสัมประสิทธิ์ของการเน้นล่วงหน้า

2. <u>การวางกรอบสัญญาณ(Windowing)</u>

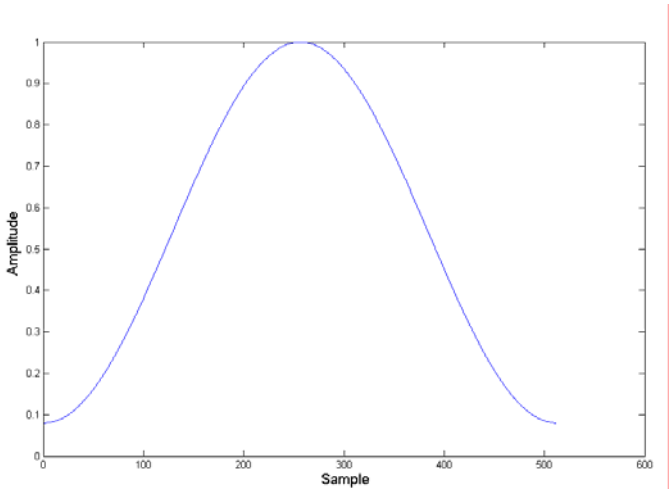
เสียงมีคุณสมบัติคือการเปลี่ยนแปลงของสัญญาณไม่ขึ้นกับเวลา และมีลักษณะการเปลี่ยนแปลงหลากหลาย จึงจำเป็นต้องวิเคราะห์สัญญาณเสียงทีละช่วงสั้น ๆ เฉพาะที่อยู่ภายในกรอบสัญญาณที่กำหนดขึ้น เรียกว่า การวางกรอบสัญญาณ ซึ่งผลของสัญญาณที่ได้ ($W(i)$) จะมีลักษณะต่างกันไป ขนาดของกรอบสัญญาณที่นิยมใช้ในการวิเคราะห์สัญญาณเสียงเพื่อการรู้จำมีค่าประมาณ 10-30 มิลลิวินาที และควรวางกรอบสัญญาณถัดไปให้มีความเหลื่อมกันประมาณ 1/3 (L. Rabiner, 1993) ของขนาดกรอบสัญญาณเพื่อให้ข้อมูลที่วิเคราะห์มีความต่อเนื่องกัน กรอบสัญญาณมีหลายชนิด เช่น แบบแฮนนิง(Hanning Window), แบบแฮมมิง(Hamming Window), แบบสามเหลี่ยม(Triangular Window), แบบสี่เหลี่ยม(Rectangular Window) เป็นต้น แต่กรอบสัญญาณที่นิยมใช้ คือ กรอบสัญญาณแฮมมิง วิธีการหาดังสมการ (1)

$$W(i) = 0.54 - 0.46\cos(\frac{2\pi i}{n-1}) \qquad (1)$$

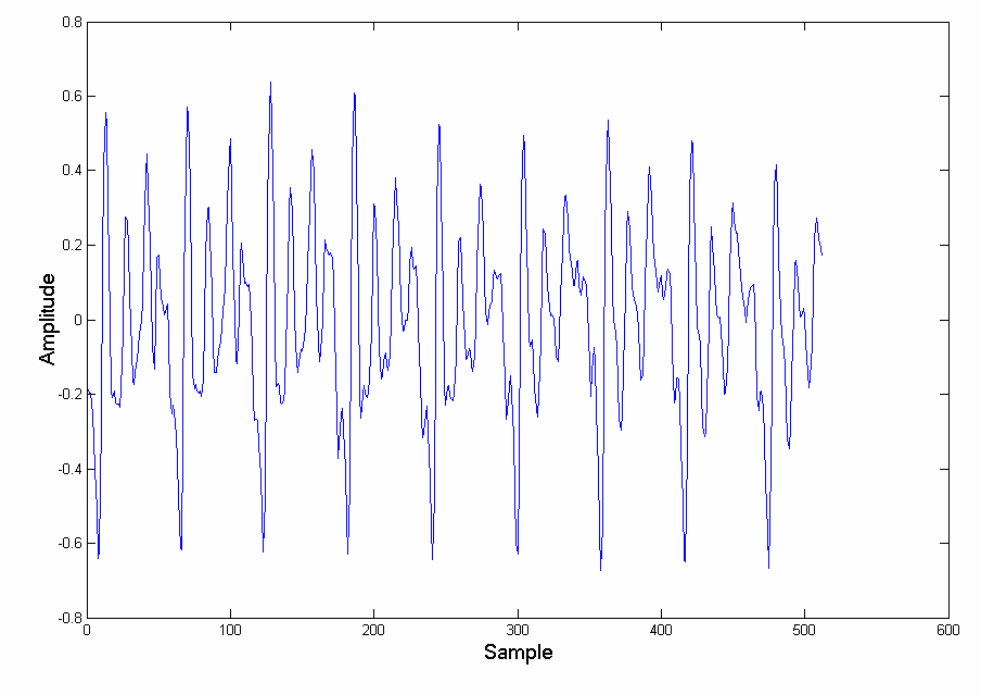
โดย $W(i)$ คือ กรอบสัญญาณที่ i

n คือ จำนวนสัญญาณเสียงที่ในหน้าต่างที่ i

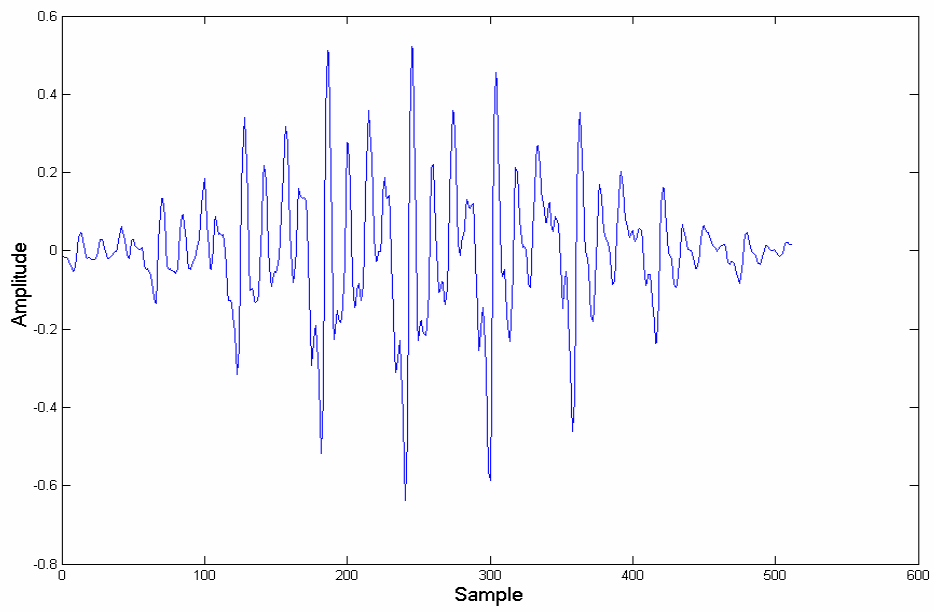
ซึ่งมีลักษณะดังภาพที่ 2



<u>ภาพที่ 2</u> กรอบสัญญาณแฮมมิง



ภาพที่ 3 สัญญาณเสียงเริ่มต้น



ภาพที่ 4 ผลลัพธ์ของการวางกรอบสัญญาณแฮมมิง

กรอบสัญญาณทั้งสองนี้เหมาะสำหรับการประมวลผลสัญญาณเสียง เพราะสามารถเน้นสัญญาณเสียงในเฟรมที่กำลังพิจารณาให้มีความสำคัญสูงสุด โดยลดความสำคัญของสัญญาณเสียงที่อยู่ในเฟรมรอบข้าง แต่ยังคงความต่อเนื่องของสัญญาณเสียงให้มีความต่อเนื่องกันทำให้ได้เสียงที่ผ่านการวางกรอบสัญญาณยังคงข้อมูลครบถ้วนอยู่ กรอบสัญญาณทั้งสองนี้มีความแตกต่างกันในส่วนปลายทั้งสองด้านของกรอบสัญญาณ กล่าวคือกรอบสัญญาณแบบแฮมมิงนั้นมีความต่อเนื่องของสัญญาณเสียงมากกว่ากรอบสัญญาณแบบแฮนนิง ความต่อเนื่องของสัญญาณนั้นช่วยลดปัญหาในกรณีที่เสียงพูดมีการเปลี่ยนแปลงอย่างรวดเร็ว

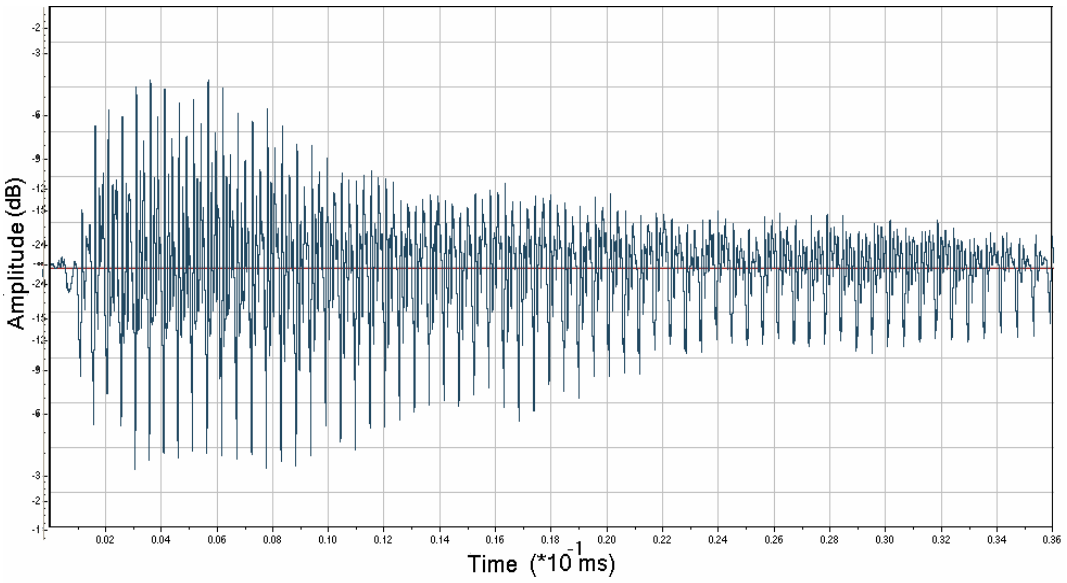
1.1 การแปลงสัญญาณฟูริเยร์(Fourier Transform)

การแปลงสัญญาณโดยใช้สมการฟูริเยร์นั้นเพื่อเปลี่ยนสัญญาณเสียงจากโดเมนของเวลา x(n) ให้อยู่ในโดเมนของความถี่ ดังสมการ (3)

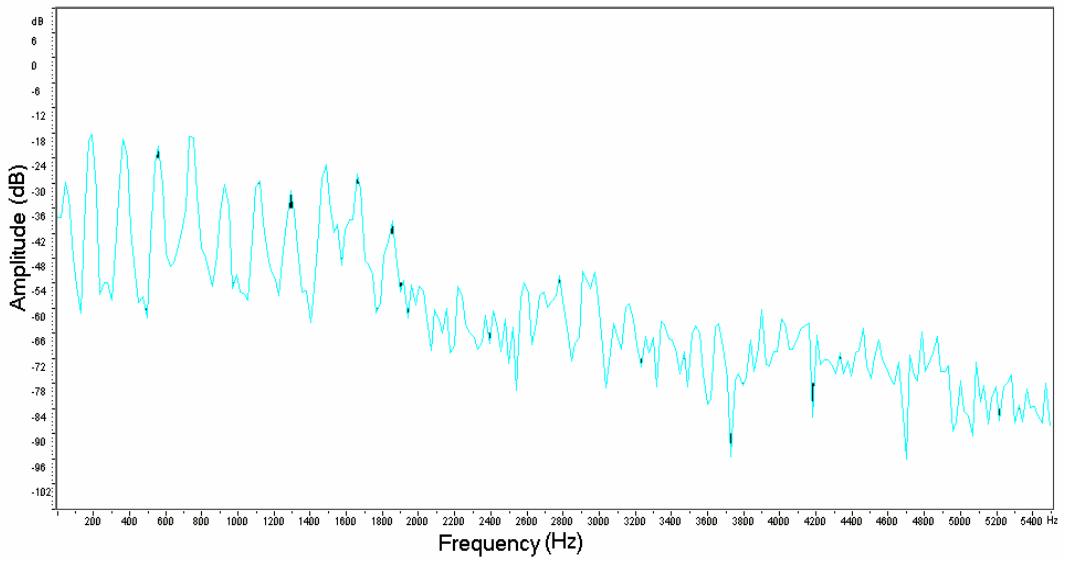
$$x(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\Pi nk/N} \tag{3}$$

โดย N คือ จำนวนตัวอย่างในหนึ่งกรอบสัญญาณ
k คือ ลำดับของกรอบสัญญาณ k = 1, 2, …, K

ผลลัพธ์จากการแปลงสัญญาณฟูริเยร์ มีลักษณะดังภาพที่ 5



(A) สัญญาณเสียงในโดเมนเวลา



(B) สัญญาณเสียงในโดเมนความถี่

<u>ภาพที่ 5</u> สัญญาณเสียงที่ผ่านกระบวนการแปลงสัญญาณฟูริเยร์

การดึงค่าคุณลักษณะสำคัญของเสียง

2. การดึงค่าคุณลักษณะสำคัญของเสียงพูด(Speech Feature Extraction)

การหาลักษณะสำคัญของเสียงเป็นการดึงเอาลักษณะเด่นเพื่อเป็นตัวแทนของเสียงพูด ทำให้สามารถช่วยในการแยกแยะความแตกต่างของแต่ละคำในการรู้จำเสียง ขั้นตอนนี้จะทำการกำจัดองค์ประกอบที่ไม่เกี่ยวข้องกับลักษณะเด่นออกไป ค่าลักษณะสำคัญของเสียงอย่างง่ายอาจจะเป็น ค่าพลังงาน (Energy) ก็ได้ (Roe and Wilpon, 1993) วิธีการหาลักษณะสำคัญของเสียงที่นิยมใช้แบ่งได้เป็นสองด้านคือ และให้ผลการรู้จำที่ดีเช่น การประมาณพันธะเชิงเส้นซึ่งเป็นการจำลองทางเดินของเสียง หรือ การสร้างเสียงพูด (Speech Production) วิธีการอีกด้านหนึ่งใช้หลักของการรับรู้เสียงพูดของมนุษย์ (Speech Perception) เช่น สัมประสิทธิ์เซปสรัลบนความถี่เมล (Mel Frequency Cepstral Coefficient-MFCC) การประมาณการรับรู้เชิงเส้น (Perceptual Linear Prediction-PLP) วิธีการที่ใช้ในงานวิจัยนี้สัมประสิทธิ์เซปสรัลบนความถี่เมลซึ่งเป็นวิธีที่ให้ผลการรู้จำที่ดีวิธีหนึ่งดังในงานของ Davis, 1989

2.1 สัมประสิทธิ์การประมาณพันธะเชิงเส้น(Linear Prediction Coefficient-LPC)

สัมประสิทธิ์การประมาณพันธะเชิงเส้น(LPC) (Sadaoki Furui, 1989) คือการคำนวณค่าสัมประสิทธิ์การทำนาย (a_k) ที่ทำให้ค่าผิดพลาดกำลังสองเฉลี่ยของค่าจากการทำนายมีค่าต่ำที่สุดที่กรอบสัญญาณใด ๆ โดยกำหนดให้สัญญาณเสียงพูด $x(n)$ ประมาณด้วยผลรวมเชิงเส้นของสัญญาณเสียงในอดีต p ตัว ดังสมการ (4)

$$x(n) = a_1 x(n-1) + a_2 x(n-2) + \ldots + a_p x(n-p) \tag{4}$$

หรืออยู่ในรูป

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) \tag{5}$$

โดยค่าสัมประสิทธิ์ $a_1, a_2, \ldots, a_k$ มีค่าคงที่ตลอดกรอบสัญญาณเสียงพูด

ค่าผิดพลาดจากการประมาณสัญญาณเสียงดังสมการ (5) ($e(n)$) คำนวณได้ดังสมการ

(6)

$$e(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k) \tag{6}$$

การนำสัมประสิทธิ์การประมาณพันธะเชิงเส้น มาประยุกต์ใช้เพื่อหาค่าคุณลักษณะสำคัญนั้น สัญญาณเสียงจะถูกแบ่งเป็นเฟรมด้วยการวางกรอบสัญญาณ ในแต่ละกรอบมีทั้งหมด m สัญญาณ เมื่อพิจารณาค่าเฟรมที่ n จะได้ค่าผิดพลาดรวม ดังสมการ (7)

$$E_n = \sum_m [x_n(m) - \sum_{k=1}^{p} a_k x_n(m-k)]^2 \tag{7}$$

หลังจากนั้นคำนวณค่าสัมประสิทธิ์การทำนาย (a_k) ที่ทำให้ค่าการผิดพลาด (E_n) ต่ำที่สุดได้โดยการหาค่าอนุพันธ์ ดังสมการ (8)

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad k = 1, 2, ..., p \tag{8}$$

แก้สมการจำนวน p สมการ ด้วยวิธีอัตสหสัมพันธ์(Autocorrelation) โดยแปลงสมการข้างต้นให้อยู่ในรูปของเมตริกซ์ของค่าอัตสหสัมพันธ์ ซึ่งจะได้เป็นเมตริกซ์ Toeplitz ซึ่งมีลักษณะสมมาตร และทุก ๆ สมาชิกในแนวทแยงมุมมีค่าเท่ากันดังนี้

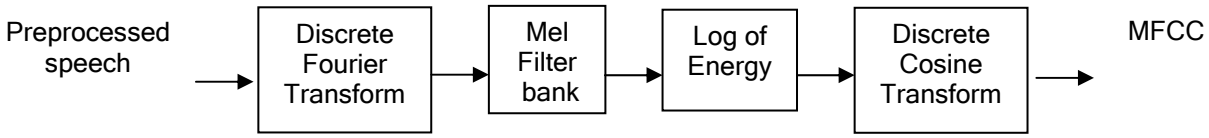
$$\begin{bmatrix} R(0) & R(1) & R(p-1) \\ R(1) & R(0) & R(p-2) \\ \vdots & \vdots & \vdots \\ R(p-1) & R(p-2) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix}$$

โดย $R(p\text{-}1)\ R(p\text{-}2)\ \ldots\ R(0)$ คือ เวกเตอร์อัตสหสัมพันธ์

สัมประสิทธิ์การประมาณพันธะเชิงเส้น จะเป็นค่าคุณลักษณะสำคัญที่สามารถแทนลักษณะของสัญญาณเสียงพูดในโดเมนความถี่ได้ดีเพียงใดขึ้นอยู่กับอันดับของพหุนาม(p) การประมาณพันธะเชิงเส้นที่กำหนด

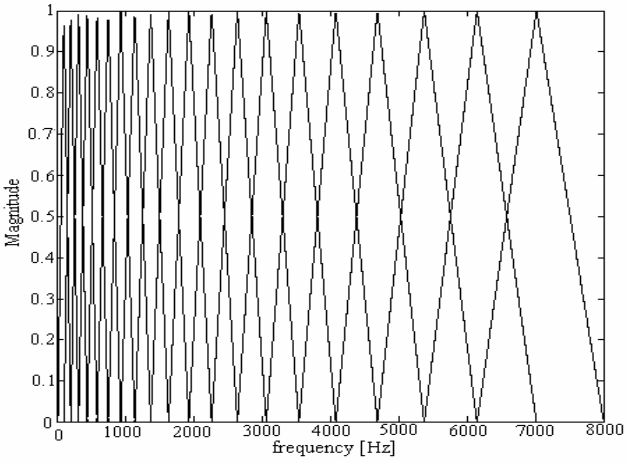
2.2 สัมประสิทธิ์เซปสตรัมบนสเกลเมล(Mel Frequency Cepstral Coefficient : MFCC)

เซปสตรัม(cepstrum) คือ การแปลงฟูริเยร์กลับ(Inverse Fourier Transformation) ของลอการิธึม(logarithm) ของสเปกตรัมสัญญาณเสียงในช่วงเวลาสั้น ๆ สัมประสิทธิ์เซปสตรัมบนสเกลเมล(Sadaoki Furui, 1989) เป็นเทคนิคที่ปรับปรุงจากเซปสตรัมด้วยการปรับสเกลของสเปกตรัมให้อยู่บนสเกลที่เหมาะสมสำหรับการรับฟังของมนุษย์ โดยสังเกตจากลักษณะของสัญญาณเสียงพูด สัญญาณเสียงพูดในช่วงความถี่ต่ำจะมีความสำคัญมากกว่าช่วงความถี่สูง จึงได้มีการออกแบบสเกลของสเปกตรัมให้สามารถเก็บรายละเอียดของสัญญาณเสียงช่วงความถี่ต่ำได้มากกว่า เรียกว่า สเกลเมล(Mel scale) มีขั้นตอนดังภาพที่ 6



ภาพที่ 6 ขั้นตอนการหาค่า MFCC

จากภาพที่ 6 ขั้นตอนการหาค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล(MFCC) เริ่มต้นจากนำสัญญาณเสียงมาผ่านการประมวลผลสัญญาณเสียงพูดเบื้องต้น (Preprocessed speech) ผ่านการแปลง หลังจากนั้นส่งเสียงไปผ่านชุดตัวกรองสามเหลี่ยม (Filter Bank) เพื่อเน้นความสำคัญของความถี่ที่อยู่ในช่วงกลางของชุดตัวกรองแต่ละตัวกรอง ชุดตัวกรองสามเหลี่ยมมีลักษณะดังภาพที่ 7

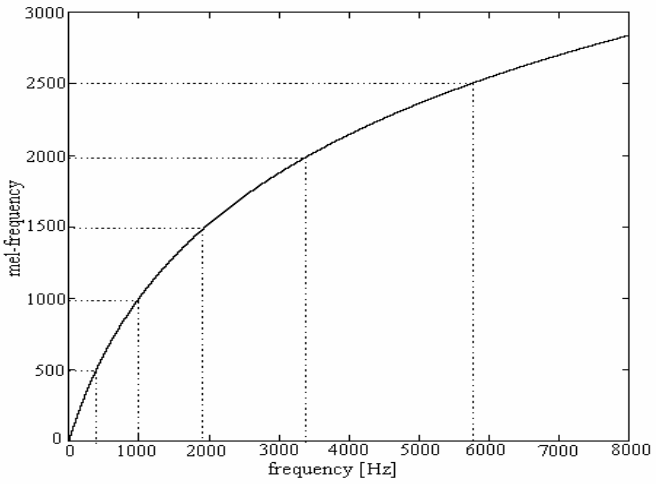


ภาพที่ 7 ชุดตัวกรองแบบสามเหลี่ยม

โดยที่ความถี่กลางของตัวกรองแต่ละอันนั้นเกิดจากการแปลงค่าความถี่ปรกติ(f) ให้อยู่บนสเกลเมล (f_{mel}) ดังสมการ (9)

$$f_{mel} = 2595 * \log 10(1 + \frac{f}{700}) \tag{9}$$

ความสัมพันธ์ระหว่างความถี่ของเสียงกับความถี่ที่แปลงให้อยู่บนสเกลเมล มีลักษณะดังภาพที่ 8



ภาพที่ 8 ความสัมพันธ์ระหว่างความถี่ปกติและความถี่สเกลเมล

หลังจากนั้นจึงคำนวณค่าพลังงานเสียง(E_j) สำหรับตัวกรองที่ j ดังสมการ (10)

$$E_j = \sum_{k=0}^{K-1} \Phi_j(k) \left| \tilde{x}(k) \right|^2 \tag{10}$$

โดย ϕ_j คือ ค่าประจำตัวกรองที่ j คำนวณจากความสูงของตัวกรองที่ตำแหน่งที่ j

$\tilde{x}(k)$ คือ สัญญาณเสียงที่ผ่านกรอบสัญญาณ

และเมื่อนำค่าลอการิธึมของพลังงานนี้มาผ่านการแปลงโคซายน์แบบไม่ต่อเนื่อง (Discrete Cosine Transformation) จะได้ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล(MFCC) ลำดับที่ n ดังสมการ (11)

$$c_{mel}(n) = \sum_{j=1}^{J} \log(E_j) \cos(n(j-0.5)\frac{\pi}{J}) \quad (11)$$

2.3 <u>การเปลี่ยนแปลงเซปสตรัมทางเวลา(Temporal Cepstral Derivative)</u>

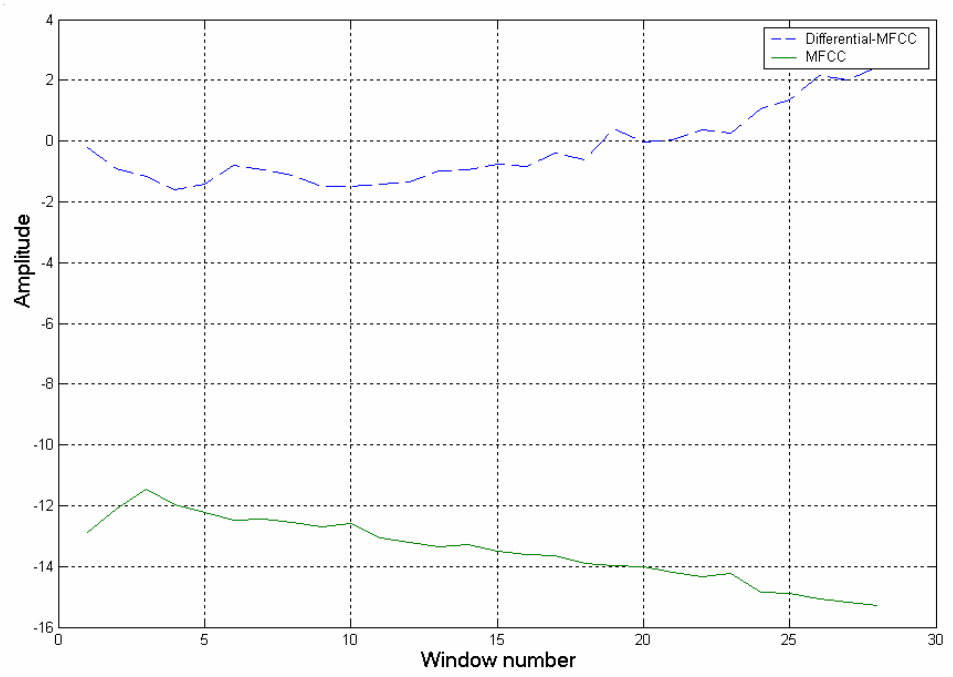
การเปลี่ยนแปลงของเซปสตรัมทางเวลา(temporal cepstrum derivative) (Sadaoki Furui, 1989) เป็นการคำนวณเพื่อวิเคราะห์คุณลักษณะของเสียงที่ไม่ขึ้นกับเวลา และลักษณะการเปลี่ยนแปลงของเสียงทั้งช่วงของการพูดครั้งหนึ่ง ๆ การคำนวณการเปลี่ยนแปลงของเซปสตรัมทางเวลา(temporal cepstrum derivative) (Sadaoki Furui, 1989) เป็นดังสมการ (12)

$$\frac{\partial c_{mel}(t)}{\partial t} = \Delta c_{mel}(t) \approx \mu \sum_{k=-K}^{K} k c_{mel}(t+k) \quad (12)$$

โดย $c_{mel}(t)$ คือ สัมประสิทธิ์เซปสตรัมลำดับที่ m ณ เวลา t

μ คือ ค่าคงที่

จำนวนเฟรมทั้งหมด ที่ใช้ในการคำนวณค่าการเปลี่ยนแปลงของสัมประสิทธิ์เซปสตรัมในแต่ละครั้งเท่ากับ 2K + 1 เฟรม ค่า K ที่เหมาะสมที่ใช้ในการคำนวณทั่วไปเท่ากับ 3 การเปลี่ยนแปลงเซปสตรัมทางเวลามีลักษณะดังภาพที่ 7



ภาพที่ 9 การเปลี่ยนแปลงเซปสตรัมทางเวลา

2.4 เวกเตอร์ควอนไตเซชัน

เวกเตอร์ควอนไตเซชัน(Robert M. Gray, 1984; Teuvo Kohonen, 1995) เป็นขั้นตอนหลังจากดึงค่าคุณลักษณะสำคัญของเสียงแล้ว เวกเตอร์ควอนไตเซชัน เป็นกระบวนการที่ทำหน้าที่ลดขนาดของข้อมูล(เช่น ค่าคุณลักษณะสำคัญ) และปรับข้อมูลให้อยู่ในรูปที่เหมาะสมกับการฝึกฝนแบบจำลองของเสียง ตัวแทนของกลุ่มข้อมูล เช่น ค่าเฉลี่ย กระบวนการเวกเตอร์ควอนไตเซชันได้ผลลัพธ์เป็นชุดรหัส(codebook) มีลักษณะเป็นตารางที่ใช้เก็บตัวแทนของกลุ่มข้อมูลทั้งหมดในระบบ ชุดรหัสหนึ่งค่าประกอบด้วย ค่าคุณลักษณะสำคัญซึ่งเป็นตัวแทนของกลุ่ม และสัญลักษณ์ประจำกลุ่ม นิยมกำหนดให้มีขนาดเท่ากับ 2^n เช่น 64, 128 เป็นต้น ขั้นตอนเวกเตอร์ควอนไตเซชันประกอบด้วย 2 ขั้นตอนคือ

2.4.1 ขั้นตอนการเรียนรู้

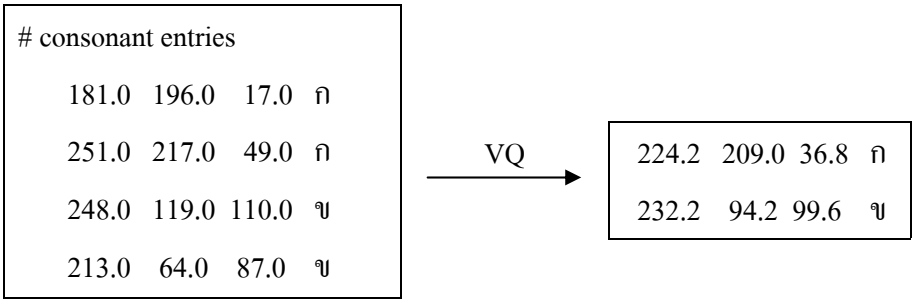
การเรียนรู้เพื่อหาตัวแทนของกลุ่มข้อมูลมีวิธีการดังต่อไปนี้

1) กำหนดตัวแทนเริ่มต้นของกลุ่ม

2) ใส่ค่าคุณลักษณะสำคัญของเสียงพูดแต่ละกลุ่ม เพื่อให้เวกเตอร์ควอนไตเซชัน จัดค่าคุณลักษณะสำคัญที่มีค่าใกล้เคียงกันอยู่เป็นกลุ่มเดียวกัน

3) เมื่อกลุ่มนั้นมีสมาชิกใหม่เพิ่มขึ้น หาค่าตัวแทนที่ใกล้เคียงกับสมาชิกมากที่สุด และปรับค่าตัวแทนของกลุ่มโดยรวมสมาชิกใหม่เข้าไปด้วย

หลังจากการเรียนรู้ จะได้ตัวแทนของกลุ่มทั้งหมดเก็บไว้ในชุดรหัส เพื่อนำไปใช้เทียบกับค่าคุณลักษณะสำคัญในขั้นตอนการรู้จำต่อไป ตัวอย่างข้อมูลเสียงแบบง่ายและชุดรหัสที่ผ่านกระบวนการเวกเตอร์ควอนไตเซชันมีลักษณะดังภาพที่ 12



ภาพที่ 12 ตัวอย่างข้อมูลเสียงอย่างง่าย(ด้านซ้าย) และชุดรหัสของข้อมูลที่ผ่านการเรียนรู้ด้วยเวกเตอร์ควอนไตเซชัน(ด้านขวา)

จากภาพที่ 12 เป็นข้อมูลเสียงสมมติ(ภาพด้านซ้าย) เพื่อประกอบการอธิบายกระบวนการเวกเตอร์ควอนไตเซชัน การเรียนรู้ใช้วิธีการหาค่าเฉลี่ยของข้อมูลในการปรับค่าตัวแทนของกลุ่มและเก็บตัวแทนไว้ในชุดรหัส(ภาพด้านขวา)

3.2 ขั้นตอนการรู้จำ

ในการรู้จำ ต้องนำค่าคุณลักษณะสำคัญของเสียงพูดที่คำนวณได้มาลดขนาดโดยแทนค่าคุณลักษณะสำคัญนั้นด้วยสัญลักษณ์ในชุดรหัส สัญลักษณ์นั้นหาได้จากตัวแทนที่มีค่าใกล้เคียงกับค่าคุณลักษณะสำคัญที่นำมารู้จำมากที่สุดเมื่อเทียบกับชุดรหัสทั้งหมด การวัดความใกล้เคียงใช้สมการวัดระยะห่างระหว่างเวกเตอร์ ตัวอย่างเช่น การวัดระยะห่างแบบยูคลิเดียน(Euclidean

Distance)(Robert M. Gray, 1984) ดังสมการ (13) และการวัดระยะห่างแบบอิตากูระไซโต (Itakura-Saito)(Robert M. Gray, 1984) ดังสมการ (14)

สมการ Euclidean Distance (d)

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + ... + (x_{n-1} - x_n)^2 + (y_{n-1} - y_n)^2} \quad (13)$$

สมการ Itakura saito (d_{IS})

$$d_{IS}(f, \hat{f}) = \int_{-\pi}^{\pi} \frac{dw}{2\pi} \left[\frac{f}{\hat{f}} - \ln \frac{f}{\hat{f}} - 1 \right] \quad (14)$$

ทฤษฎีสำหรับการรู้จำ

การรู้จำโดยทั่วไป เริ่มจากการฝึกฝนคอมพิวเตอร์ให้เรียนรู้สิ่งที่ฝึกฝนได้ด้วยการใส่ตัวอย่างจำนวนหนึ่ง หลังจากนั้นจึงทดสอบการรู้จำของคอมพิวเตอร์ เรียกว่า ขั้นตอนการรู้จำ เรียนรู้และทดสอบคอมพิวเตอร์วนซ้ำจนกระทั่งคอมพิวเตอร์ให้ความถูกต้องในการรู้จำสูง เทคนิคที่สามารถทำให้คอมพิวเตอร์รู้จำได้มีหลายเทคนิค เช่น โครงข่ายประสาทเทียม(Artificial Neural Network) แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model) เป็นต้น ซึ่งจะขออธิบายเฉพาะทฤษฎีแบบจำลองฮิดเดนมาร์คอฟ มีรายละเอียดดังต่อไปนี้

3.1 แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model)

แบบจำลองฮิดเดนมาร์คอฟเป็นทฤษฎีที่ใช้ฝึกฝนและรู้จำเสียงพูดด้วยวิธีการสร้างแบบจำลอง ดังจะอธิบายรายละเอียดเกี่ยวกับความหมายและลักษณะของแบบจำลอง และอัลกอริธึมของแบบจำลองดังนี้

3.1.1 ความหมายและลักษณะของแบบจำลอง

แบบจำลองฮิดเดนมาร์คอฟ สามารถจำลองเสียงในระดับคำหรือระดับหน่วยเสียงก็ได้ แต่เหตุผลของการที่จะสร้างแบบจำลองให้เป็นคำหรือหน่วยเสียง จะขึ้นอยู่กับการใช้งานของระบบการรู้จำเสียงเช่น ถ้าระบบรู้จำเสียงมีลักษณะคือให้รู้จำคำศัพท์ที่น้อยและรู้จำคำพูดที่ไม่ติดต่อกัน การจำลองเสียงในระดับคำจะดีกว่าการจำลองเสียงในระดับหน่วยเสียงเนื่องจากให้ความแม่นยำมากกว่า การรู้จำเสียงในระดับหน่วยเสียงนั้น ผลของการรู้จำนั้นได้จากการประกอบกันของแบบจำลองหลายหน่วยเสียง ถ้ามีการรู้จำหน่วยเสียงใดหน่วยเสียงหนึ่งผิดไปก็จะทำให้รู้จำคำคำนั้นผิดได้ แต่สำหรับการจำลองเสียงในระดับคำ ผลของการรู้จำจะขึ้นกับความแม่นยำของเสียงที่ได้ฝึกฝนมาและไม่ต้องสร้างคำขึ้นจากหน่วยเสียง ดังนั้นวิธีการของระบบรู้จำเสียงในระดับคำจะให้อัตราการรู้จำที่สูงกว่าการจำลองเสียงในระดับหน่วยเสียง

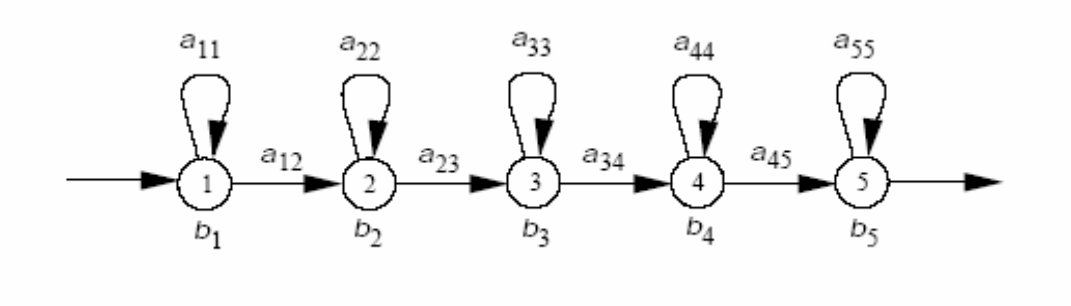
ตามทฤษฎีแบบจำลองฮิดเดนมาร์คอฟ คือ การเก็บรวบรวมความน่าจะเป็นของสถานะหลาย ๆ สถานะที่ถูกเชื่อมโยง โดยแต่ละสถานะจะให้ผลลัพธ์(O) คือ ลักษณะของสัญญาณเสียงหรือเครื่องหมายแทนสัญญาณเสียงนั้น ณ เวลาหนึ่ง ๆ เช่น แบบจำลองเสียงคำว่า “กา” สถานะที่หนึ่งให้ผลลัพธ์เป็นเสียงของตัวอักษร “ก” สถานะที่สองให้ผลลัพธ์เป็นเสียงสระ “อา” ผลลัพธ์ทั้งหมดของแบบจำลองฮิดเดนมาร์คอฟในช่วงเวลาทั้งหมด(T) เรียกรวมว่า ลำดับของผลลัพธ์($O=O_1, O_2, \ldots, O_T$)

องค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ(λ) มี 5 ข้อคือ

1) จำนวนสถานะของแบบจำลอง(N)
2) จำนวนผลลัพธ์ของสถานะ(M)
3) ค่าความน่าจะเป็นของการเปลี่ยนสถานะ(a)
4) ค่าความน่าจะเป็นของผลลัพธ์เมื่อมีการเปลี่ยนสถานะ(b)
5) ค่าความน่าจะเป็นของสถานะเริ่มต้น(π)

ตัวอย่างแบบจำลองฮิดเดนมาร์คอฟอย่างง่าย มีลักษณะดังภาพที่ 8

ภาพที่ 10 เป็นแบบจำลองฮิดเดนมาร์คอฟที่มี 5 สถานะ คือ สถานะ 1 ถึง สถานะ 5 เป็นแบบซ้ายไปขวา (left-to-right) ค่าความน่าจะเป็นแต่ละสถานะให้ผลลัพธ์ 2 แบบคือ A และ B และมีค่าความน่าจะเป็นของ a และ b



ภาพที่ 10 ตัวอย่างแบบจำลองฮิดเดนมาร์คอฟอย่างง่าย

3.1.2 อัลกอริธึม

อัลกอริธึมในการสร้างแบบจำลองฮิดเดนมาร์คอฟมี 3 อัลกอริธึม ได้แก่

3.1.2.1 อัลกอริธึมฟอร์เวิร์ด(The Forward Algorithm)

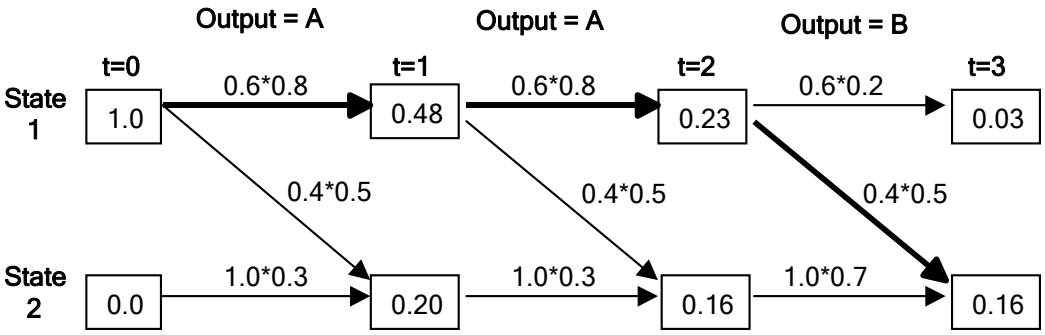
อัลกอริธึมฟอร์เวิร์ด เป็นอัลกอริธึมที่ใช้ในการรู้จำ โดยการหาค่าความน่าจะเป็นที่แบบจำลอง(λ) ให้ลำดับของผลลัพธ์($O=O_1, O_2, \ldots, O_T$) $P(O \mid \lambda)$ ดังสมการ (15)

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{15}$$

โดย

$$\alpha_{t+1}(i) = \begin{cases} \pi_i b_i(O_1), & 1 \le i \le N, t = 0 \\ \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right] b_i(O_{t+1}), & 1 \le j \le N, 1 \le t \le T-1 \end{cases} \tag{16}$$

ตัวอย่างการคำนวณตามอัลกอริธึมฟอร์เวิร์ด โดยใช้แบบจำลองฮิดเดนมาร์คอฟตามภาพที่ 9 และสมมุติว่าลำดับการสังเกตเป็น A A B มีลักษณะดังภาพที่ 14



ภาพที่ 11 ตัวอย่างการคำนวณตามอัลกอริธึมฟอร์เวิร์ด

ภาพที่ 11 เป็นตัวอย่างการคำนวณตามอัลกอริธึมฟอร์เวิร์ด ที่ 3 สถานะแรก มีความน่าจะเป็นที่ให้ผลลัพธ์ A A B เท่ากับ 0.16

3.1.2.2 อัลกอริธึมไวเทอร์บิ(The Viterbi Algorithm)

อัลกอริธึมไวเทอร์บิ เป็นอัลกอริธึมที่ใช้ในการหาลำดับของสถานะ ($Q=\{q_1, q_2, \ldots, q_T\}$) ที่ให้ค่าความน่าจะเป็นสูงที่สุด(P^*) ของแบบจำลอง มีประโยชน์ในการรู้จำเสียงที่ประกอบด้วยแบบจำลองหลาย ๆ แบบจำลองต่อกัน ลำดับของสถานะที่ให้ค่าความน่าจะเป็นสูงที่สุดหาได้จากสมการ (17)

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \ldots, 1 \tag{17}$$

และค่าความน่าจะเป็นสูงที่สุดหาได้จากสมการ (18)

$$P^* = \max_{1 \le i \le N}\left[\delta_T(i)\right] \tag{18}$$

โดย Ψ_t คือ สถานะที่เวลาใด ๆ หาได้จากสมการ (19)

$$\psi_t(j) = \begin{cases} 0, & 1 \le j \le N, t = 1 \\ \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], & 1 \le j \le N, 2 \le t \le T \end{cases} \quad (19)$$

และ δ_t คือ ค่าความน่าจะเป็นที่เวลาใด ๆ หาได้จากสมการ (20)

$$\delta_t(j) = \begin{cases} \pi_j b_j(O_1), & 1 \le j \le N, t = 1 \\ \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), & 1 \le j \le N, 2 \le t \le T \end{cases} \quad (20)$$

จากภาพที่ 12 การคำนวณตามอัลกอริธึมไวเทอร์บิให้ผลลัพธ์เป็นลำดับของสถานะที่ให้ค่าความน่าจะเป็นสูงที่สุดในเวลา *t* ต่าง ๆ ดังนี้คือ สถานะ 1 ที่เวลา $t = 0$ สถานะ 1 ที่เวลา $t = 1$ สถานะ 1 ที่เวลา $t = 2$ และสถานะ 2 ที่เวลา $t = 3$ (ตามเส้นหนา)

3.1.2.3 <u>อัลกอริธึมฟอร์เวิร์ดแบคเวิร์ด(The Forward-Backward Algorithm)</u>

อัลกอริธึมฟอร์เวิร์ดแบคเวิร์ด ใช้สำหรับสอนแบบจำลองฮิดเดนมาร์คอฟโดยวิธีการปรับค่าพารามิเตอร์ต่าง ๆ เพื่อให้แบบจำลองสามารถเป็นตัวแทนข้อมูลได้ดีที่สุด ในการอธิบายอัลกอริธึมนี้จะขอนิยามตัวแปรดังต่อไปนี้

1) $\xi_t(i, j)$ คือ ค่าความน่าจะเป็นที่อยู่ในสถานะ *i* ที่เวลา *t* และอยู่ในสถานะ *j* ที่เวลา *t*+1

2) $\gamma_t(i)$ คือ ค่าความน่าจะเป็นที่อยู่ในสถานะ *i* ที่เวลา *t* หรือหาได้จากสมการ (21)

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \quad (21)$$

ค่าพารามิเตอร์ของแบบจำลอง *a, b,* π เป็นค่าความน่าจะเป็นซึ่งคำนวณด้วยวิธีการนับ ค่าพารามิเตอร์ที่ถูกปรับค่าแล้ว($\bar{a}, \bar{b}, \bar{\pi}$) สามารถคำนวณได้ดังนี้

$\bar{\pi}$ เท่ากับความถี่ที่อยู่ในสถานะ *i* ที่เวลา $t = 1$ หาได้จากสมการ (22)

$$\bar{\pi} = \gamma_1(i) \tag{22}$$

$\bar{a}_{ij}$ เท่ากับอัตราส่วนระหว่างจำนวนการเปลี่ยนสถานะจาก i ไปสถานะ j และจำนวนการเปลี่ยนสถานะทั้งหมดของสถานะ i หาได้ดังสมการ (23)

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{23}$$

$\bar{b}_j(k)$ เท่ากับอัตราส่วนระหว่างจำนวนครั้งที่สถานะ j ให้ผลลัพธ์เป็น k และจำนวนครั้งที่อยู่ในสถานะ j หาได้ดังสมการ (24)

$$\bar{b}_j(k) = \frac{\sum_{\substack{t=1 \\ O_t=k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{24}$$

ถ้าเราให้แบบจำลองเดิมเป็น $\lambda = (a, b, \pi)$ แบบจำลองที่ผ่านการสอนด้วยการปรับค่าพารามิเตอร์ตามอัลกอริธึมฟอร์เวิร์ดแบคเวิร์ดคือ $\bar{\lambda} = (\bar{a}, \bar{b}, \bar{\lambda})$ สามารถให้ค่าความน่าจะเป็นสูงกว่าแบบจำลองเดิม($P(O \mid \bar{\lambda}) > P(O \mid \lambda)$) แทนผลลัพธ์ได้ดีขึ้น เมื่อสอนแบบจำลองเป็นจำนวนรอบที่สูงขึ้นค่าความน่าจะเป็นของแบบจำลองที่ผ่านการสอนจะสูงขึ้นเช่นนี้เรื่อยไป แต่ไม่ได้มีการกำหนดจำนวนรอบในการสอนที่เหมาะสมไว้ในทฤษฎี จำนวนรอบของการสอนแบบจำลองจึงขึ้นอยู่กับข้อมูลที่นำมาสอนและผลต่างของค่าความน่าจะเป็นในแต่ละรอบที่ระบบได้กำหนดไว้

3.2 แบบจำลองฮิดเดนมาคอร์ฟแบบต่อเนื่อง

แบบจำลองฮิดเดนมาร์คอฟมีทั้งแบบต่อเนื่อง (Continuous Density HMM) และแบบไม่ต่อเนื่อง (Discrete HMM) โดยหัวข้อก่อนหน้านี้เป็นแบบจำลองฮิดเดนมาคอร์ฟแบบไม่ต่อเนื่อง

ในงานวิจัยนี้ใช้แบบจำลองฮิดเดนมาร์คอฟแบบต่อเนื่องซึ่งมีเด่นกว่าฮิดเดนมาคอร์ฟแบบไม่ต่อเนื่องในแง่ของความสามารถในการแทนข้อมูลเสียงพูดที่ต่อเนื่องกันไปได้ดีกว่า นอกจากนี้ยังไม่ต้องทำเวกเตอร์ควอนไตซ์ (Vector Quantize) ข้อแตกต่างของแบบจำลองฮิดเดนมาคอร์ฟแบบไม่ต่อเนื่องและแบบต่อเนื่องที่ชัดเจนคือค่าความน่าจะเป็นของผลลัพธ์เมื่อมีการเปลี่ยนสถานะดังสมการที่ (25)

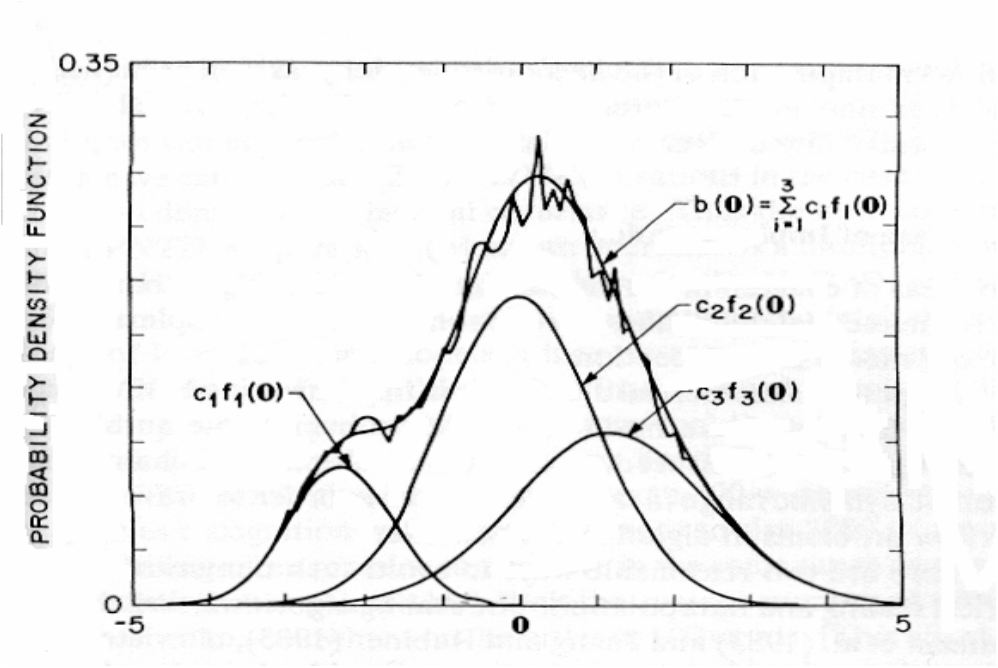
$$b_j(O) = \sum_{k=1}^{M} c_{jk} F(O, m_{jk}, U_{jk}) \qquad (25)$$

โดยที่ c_{jk} เป็นค่าสัมประสิทธ์ผสมของกลุ่มเกาส์เซียน โดยมีเงื่อนไขคือ

$$\sum_{k=1}^{M} c_{jk} = 1$$

และ ค่า F เป็นฟังก์ชั่นเกาส์เซียนที่ใช้ มีค่าพารามิเตอร์ในฟังก์ชั่นคือค่าเฉลี่ย และค่าความแปรปรวน ในการเรียนรู้

ภาพที่ 12 เป็นภาพความน่าจะเป็นในฮิดเดนมาร์คอร์ฟแบบต่อเนื่องที่มีสัมประสิทธิ์ผสมของกลุ่มเกาส์เซียนจำนวน 3 ค่านั่นคือ $c_1f_1(O)$, $c_2f_2(O)$ และ $c_3f_3(O)$ ตามสมการที่ 25 โดยผลรวมของค่า b ที่เกิดจากกลุ่มเกาส์เซียนสามกลุ่ม รวมกันเพื่อประมาณเป็นสัญญาณเสียงจริง



ภาพที่12 ภาพความน่าจะเป็นในฮิดเดนมาร์คอร์ฟแบบต่อเนื่อง (Furui, 1991)

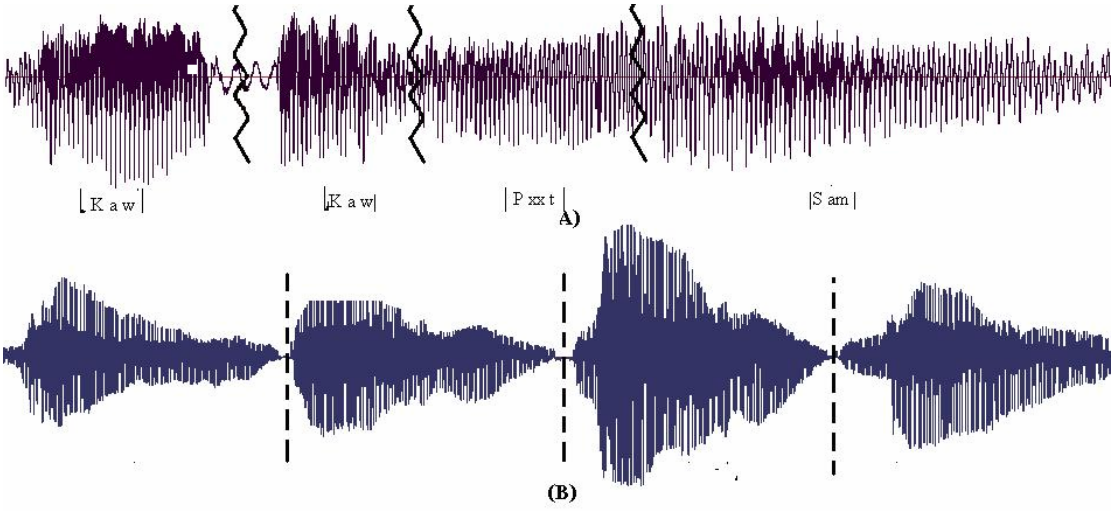
ปัญหาการรู้จำเสียงพูดต่อเนื่อง

เสียงพูดต่อเนื่อง คือการพูดคำหลาย ๆ คำต่อเนื่องกันไปโดยไม่มีการเว้นวรรค เช่น การพูดเป็นประโยค ซึ่งต่างจากเสียงพูดคำเดี่ยว ที่มีการเว้นวรรคระหว่างคำแต่ละคำ จากลักษณะการพูดต่อเนื่อง ทำให้เสียงที่เปล่งออกมามีการซ้อนทับหรือเหลื่อมกันของเสียง ส่งผลให้การรู้จำเสียงผิดพลาด ปัญหาการรู้จำเสียงพูดต่อเนื่องและสาเหตุที่ทำให้เกิดปัญหามีดังต่อไปนี้

1. ความคลุมเครือของขอบเขตพยางค์

การรู้จำเสียงโดยทั่วไป มักนิยมสร้างแบบจำลองของเสียงในระดับหน่วยย่อย เช่น คำ, พยางค์ หรือหน่วยย่อยของเสียง ดังนั้นจึงมีการนำส่วนที่ทำหน้าที่แบ่งเสียงมาใช้เพื่อแบ่งเสียงต่อเนื่องออกเป็นส่วนย่อยที่อยู่ในระดับเดียวกับแบบจำลองของเสียงก่อนนำไปรู้จำ แต่การพูดเป็นประโยคหรือเป็นคำต่อเนื่องกันไป จะเกิดกรณีที่บางพยางค์ในประโยคซ้อนทับหรือเหลื่อมกันกับพยางค์ถัดไป(ดังภาพที่ 11) ซึ่งลักษณะนี้เกิดจากอัตราเร็วในการพูด ยิ่งพูดด้วยอัตราเร็วสูงจะทำให้

เสียงกลืนกันมากขึ้น การซ้อนทับหรือเหลื่อมกันดังกล่าวทำให้เกิดความคลุมเครือในการแบ่งเสียง ซึ่งถ้าการแบ่งเสียงเกิดความผิดพลาดขึ้นจะทำให้การรู้จำผิดพลาดตามไปด้วย



ภาพที่ 12 การเปรียบเทียบลักษณะเสียงพูดต่อเนื่อง(A) และเสียงพูดไม่ต่อเนื่อง(B)

จากภาพที่ 11 เป็นเสียงของตัวเลข "9983"จะเห็นว่า เสียงแต่ละพยางค์กลืนกันหรือซ้อนทับกันจนกระทั่งไม่สามารถหาขอบเขตที่แน่นอนได้

หมายเหตุ สำหรับภาพ (A) เส้นขอบเขตของแต่ละพยางค์( ) $\gtrsim$ กำหนดด้วยการฟังและดูรูปคลื่นเสียงประกอบกัน

2. ความยืดหยุ่นของเสียงพูด

ลักษณะโดยธรรมชาติของเสียงพูดคือ มีความยืดหยุ่นสูง ซึ่งเกิดขึ้นในการพูดแบบคำเดี่ยวเช่นเดียวกัน แต่สำหรับเสียงต่อเนื่องนั้นยิ่งทำให้เกิดความแตกต่างมากขึ้นเนื่องจากท่วงทำนองของเสียงพูด และลักษณะการเน้นหนักของพยางค์ซึ่งการเน้นหนักของคำแต่ละคำจะมีความแตกต่างกัน ดังนั้นเสียงที่เกิดขึ้นในแต่ละช่วงดังกล่าวจึงมีรูปร่างแตกต่างกันไป เป็นเหตุให้เสียงของคำๆ หนึ่งมีลักษณะหลากหลายมากยิ่งขึ้น เกิดความคลุมเครือสูงขึ้น เป็นการยากที่จะทำให้ระบบสามารถรู้จำเสียงได้ครบทุกลักษณะ ดังนั้นประสิทธิภาพในการรู้จำจะลดลงในภาพที่ 11 จะเห็นว่าแม้แต่เลข 9 เหมือนกันก็มีความแตกต่างกันเมื่อพูดออกมาในแต่ละตำแหน่ง

3. แนวทางการแก้ปัญหาเสียงต่อเนื่อง

จากปัญหาที่กล่าวมาแล้วข้างต้นมีแนวทางแก้ไขปัญหาหลักๆ อยู่ 2 วิธีด้วยกัน คือ การรู้จำโดยมีการตัดแบ่งพยางค์ และการรู้จำโดยไม่ต้องมีการแบ่งพยางค์ ดังรายละเอียดต่อไปนี้

1. การรู้จำโดยไม่ต้องมีการแบ่งพยางค์

1.1 เป็นวิธีที่มีการประยุกต์ใช้ในภาษาต่างประเทศ อัลกอริธึมที่รู้จำได้โดยไม่ต้องมีส่วนการแบ่งพยางค์ เช่น อัลกอริธึมการสร้างระดับ (Level Building algorithm) เป็นต้น

1.2 การใช้ด้วยอัลกอริธึมการเรียนรู้ในการเรียนรู้จากตัวอย่างเสียงพูดที่ผ่านการกำกับคำแล้ว วิธีการที่นิยมใช้เช่น วิธีโครงข่ายประสาทเทียม หรือ แบบจำลองฮิดเดนมาคอร์ฟ

2. การรู้จำโดยมีการตัดแบ่งพยางค์

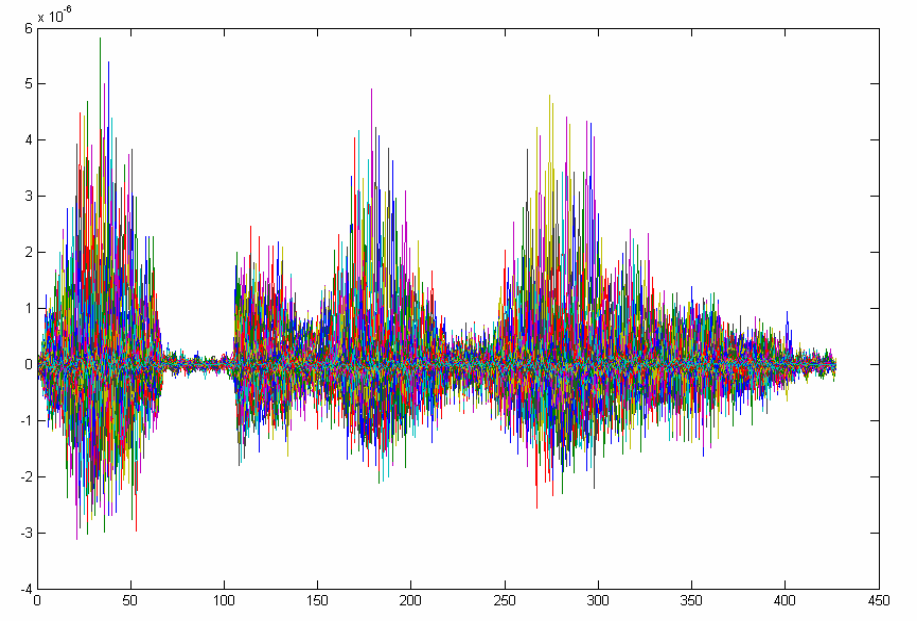
การแบ่งพยางค์มีหลายวิธี วิธีที่นิยมนำมาใช้คือ การดูกราฟพลังงานเทียบกับระดับพลังงานอ้างอิง 2 ระดับ โดยจุดแบ่งพยางค์ คือตำแหน่งที่พลังงานเริ่มมีค่ามากกว่าระดับอ้างอิงที่ 1 จะเป็นตำแหน่งเริ่มต้นของคำศัพท์ และตำแหน่งที่พลังงานเริ่มมีค่าน้อยกว่าระดับอ้างอิงที่ 2 จะเป็นตำแหน่งสิ้นสุดของคำศัพท์ การสร้างกราฟพลังงานทำได้โดยการวางกรอบสัญญาณ แล้วคำนวณพลังงานของเสียงที่ตำแหน่งใด ๆ (E_i) ดังสมการ (26)

$$E_i = \sum_{n=1}^{N_w} |x_i(n)| \tag{26}$$

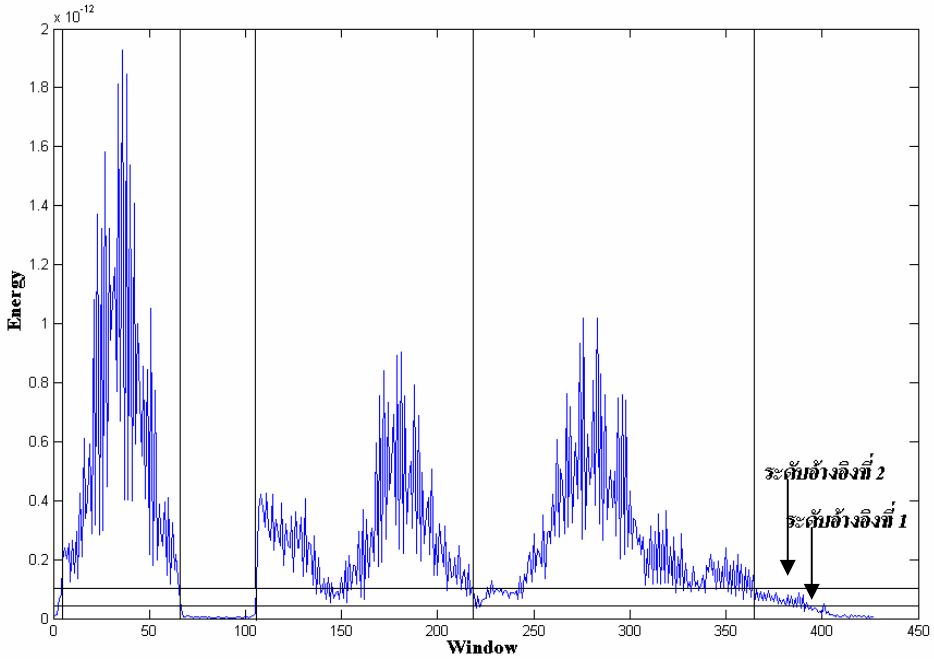
โดย N_w คือ ความกว้างของหน้าต่าง

$x_i(n)$ คือ ตัวอย่างที่ n ของเสียงพูดในเฟรมที่ i

นำค่าพลังงานสัมบูรณ์ที่คำนวณได้มาสร้างกราฟพลังงาน การแบ่งพยางค์ด้วยกราฟพลังงานดังภาพที่ 12



(A) รูปคลื่นเสียง



(B) กราฟพลังงานของเสียง

ภาพที่ 12 การแบ่งพยางค์ด้วยวิธีการสร้างกราฟพลังงาน

4. หน่วยเสียงภาษาไทยของตัวเลขแต่ละตัว

ในหัวข้อนี้แสดงข้อมูลของหน่วยเสียงในแต่ตัวเลขภาษาไทยดังตารางที่ 1

	หน่วยเสียง			
ตัวเลข	พยัญชนะต้น	เสียงสระ	พยัญชนะสะกด	วรรณยุกต์
ศูนย์	/ s /	/ uu /	/ n /	จัตวา
หนึ่ง	/ n /	/ vv /	/ ng /	เอก
สอง	/ s /	/ @@ /	/ ng /	จัตวา
สาม	/ s /	/ aa /	/ m /	จัตวา
สี่	/ s /	/ ii /		เอก
ห้า	/ h /	/ aa /		โท
หก	/ h /	/ o /	/ k /	เอก
เจ็ด	/ c /	/ e /	/ t /	เอก
แปด	/ p /	/ae/	/t/	เอก
เก้า	/ k /	/ aa /	/ w /	โท

งานวิจัยเกี่ยวกับการรู้จำเสียงภาษาไทยที่ผ่านมาสรุปได้ดังตารางที่ 1 และ 2 มีรายละเอียดเกี่ยวกับปัญหาที่สนใจ เทคนิคที่ใช้แก้ปัญหา และผลที่ได้ของงานวิจัย