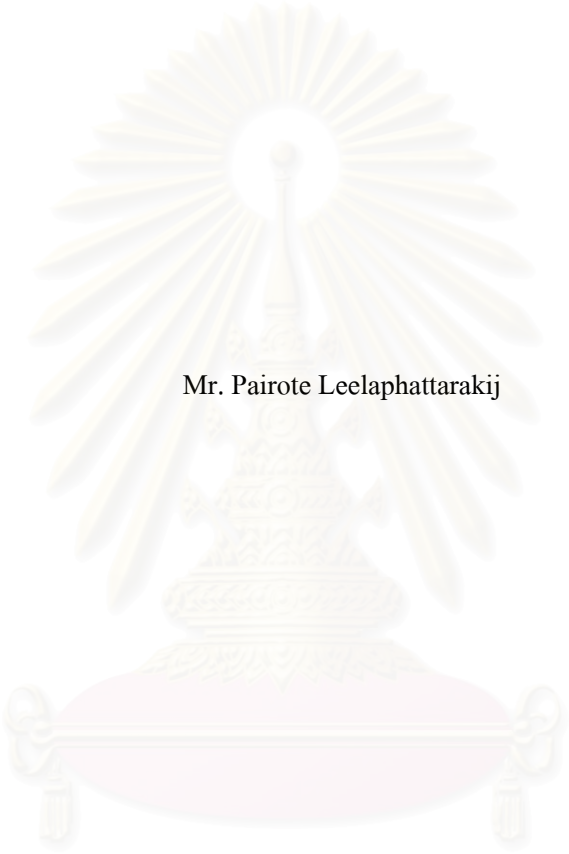
การแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดภาษาไทยแบบอาศัยเซกเมนต์ โดยใช้สารสนเทศสวนสัทศาสตร์

นายไพโรจน์ ลีลาภัทรกิจ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549

SPEECH SEGMENTATION FOR THAI SEGMENT-BASED SPEECH RECOGNITION USING ACOUSTIC-PHONETIC INFORMATION

Mr. Pairote Leelaphattarakij



A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2006

หัวข้อวิทยานิพนธ์	การแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดภาษาไทยแบบอาศัยเซกเมนต์โดยใช้สารสนเทศสวนสัทศาสตร์
โดย	นายไพโรจน์ ลีลาภัทรกิจ
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.อติวงศ์ สุชาโต
อาจารย์ที่ปรึกษาร่วม	อาจารย์ ดร. โปรดปราน บุณยพุกกณะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร.อติวงศ์ สุชาโต)

..... อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร. โปรดปราน บุณยพุกกณะ)

..... กรรมการ
(อาจารย์ ดร.พิษณุ คนองชัยยศ)

..... กรรมการ
(ดร.ชัย วุฒิวิวัฒน์ชัย)

ไพโรจน์ ลีลาภัทรกิจ : การแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดภาษาไทยแบบอาศัยเซกเมนต์โดยใช้สารสนเทศสวนสัทศาสตร์. (SPEECH SEGMENTATION FOR THAI SEGMENT-BASED SPEECH RECOGNITION USING ACOUSTIC-PHONETIC INFORMATION) อ. ที่ปรึกษา : อ.ดร.อติวงศ์ สุชาโต, อ.ที่ปรึกษาร่วม : อ.ดร.โปรดปราน บุณยพุกกณะ, 98 หน้า.

ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จะต้องแบ่งเสียงพูดออกเป็นเซกเมนต์ โดยมีวัตถุประสงค์เพื่อค้นหาขอบเขตของหน่วยเสียงหรือตำแหน่งบอกเวลาเริ่มต้นและสิ้นสุดของหน่วยเสียง แล้วนำไปสร้างเป็นกราฟของเซกเมนต์ ซึ่งจะถูกใช้เป็นข้อมูลขาเข้าของขั้นตอนการรู้จำเสียงพูด เพื่อค้นหาลำดับของหน่วยเสียงที่ดีที่สุดออกมา เป้าหมายของวิทยานิพนธ์นี้ ต้องการพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่มีประสิทธิภาพและสามารถทำงานได้อย่างรวดเร็ว เพื่อนำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ วิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบเดิมนั้นจะอาศัยเครื่องรู้จำเสียงพูดในระดับหน่วยเสียงมาค้นหาขอบเขตของหน่วยเสียง แต่เนื่องจากประสิทธิภาพนั้นยังห่างไกลกับประสิทธิภาพของวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยคน อีกทั้งกราฟของเซกเมนต์ที่ได้มีขนาดใหญ่และใช้เวลาในการทำงานนาน จึงไม่เหมาะกับระบบรู้จำเสียงพูดที่ต้องการความรวดเร็ว วิทยานิพนธ์นี้จึงเสนอวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่มีประสิทธิภาพดีกว่า โดยมีขั้นตอนการทำงานสองขั้นตอน คือ ขั้นตอนการหาขอบเขตของหน่วยเสียงจากตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียง โดยอาศัยลักษณะสำคัญของเสียงที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์และอาศัยซัพพอร์ตเวกเตอร์แมชชีนมาจำแนกเสียงพูดตามลักษณะการออกเสียง ขั้นตอนต่อมาจะสร้างกราฟของเซกเมนต์โดยใช้วิธีการสร้างกราฟแบบหลายระดับ รวมถึงมีการคิดคะแนนให้กับขอบเขตของหน่วยเสียงที่หามาได้จากค่าการเปลี่ยนแปลงสเปกตรัม การทดลองทั้งหมดจะทดสอบโดยใช้ฐานข้อมูลเสียงพูดภาษาไทย โดยเมื่อยอมให้ขอบเขตของหน่วยเสียงที่หามาได้คลาดเคลื่อนไปจากขอบเขตของหน่วยเสียงอ้างอิงได้ไม่เกิน 20 มิลลิวินาที วิธีการแบ่งเสียงพูดเป็นเซกเมนต์นี้จะสามารถค้นหาขอบเขตของหน่วยเสียงได้ความแม่นยำและความครอบคลุมเพิ่มขึ้น 8.3% (จาก 68.0% เป็น 76.3%) และ 5.1% (จาก 82.1% เป็น 87.2%) ตามลำดับ และสามารถลดขนาดกราฟของเซกเมนต์ได้ประมาณ 14 เท่าโดยที่ยังรักษาระดับความครอบคลุมไว้ได้ที่ 77.4% เมื่อเปรียบกับวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ภาควิชา.....วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต.....ไพโรจน์.....
สาขาวิชา.....วิศวกรรมคอมพิวเตอร์	ลายมือชื่ออาจารย์ที่ปรึกษา.....อติวงศ์.....
ปีการศึกษา.....2549	ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....[illegible].....

497 04919 21 : MAJOR COMPUTER ENGINEERING

KEY WORD: AUTOMATIC SPEECH RECOGNITION / SPEECH SEGMENTATION / SEGMENT-BASED SPEECH RECOGNITION

PAIROTE LEELAPHATTARAKIJ : SPEECH SEGMENTATION FOR THAI SEGMENT-BASED SPEECH RECOGNITION USING ACOUSTIC-PHONETIC INFORMATION. THESIS ADVISOR : ATIWONG SUCHATO, Ph.D., THESIS COADVISOR : PROADPRAN PUNYABUKKANA, Ph.D., 98 pp.

Segment-based speech recognition systems must explicitly hypothesize segment start and end times. The purpose of a segmentation algorithm is to hypothesize those times and to compose a graph of segments from them. During recognition, this graph is an input to a search that finds the optimal sequence of sound units through the graph. The goal of this thesis is to create a high-quality, real-time phonetic segmentation algorithm for segment-based speech recognition. The baseline algorithm makes use of frame-based phonetic recognizer to hypothesize possible phonetic segments but its performance was still far from human's ability to perform such a task. This thesis addresses the quality and computational requirements by employing more efficient phonetic segmentation algorithm, and by shrinking the search space. The algorithm is done in two stages. Boundaries are detected in the first stage via manner-of-articulation changes by using acoustic features obtained from acoustic-phonetic information and applying multiple Support Vector Machines for the classification of manner features. Multi-Level Segmentation is used to compose a graph from the boundary list for the graph size reduction. In addition, it includes a landmark scoring by utilizing the spectral transition measurement. Experiments reported were done on Thai continuous speech corpus. Allowing at most 20 ms. deviation from the actual boundaries, the algorithm detects boundaries that have over 8.3% and 5.1% improvement in precision (from 68.0% to 76.3%) and recall rate (from 82.1% to 87.2%) and produces a segment-graph that has over 14 times fewer segments while still maintaining a 77.4% in recall rate over a baseline speech segmentation algorithm.

Department .....Computer Engineering...	Student's signature .....
Field of study.....Computer Engineering...	Advisor's signature .....
Academic year 2006	Co-advisor's signature .....

กิตติกรรมประกาศ

กิติกรรมประกาศนี้ของอุทิศให้กับผู้ที่มีส่วนให้การช่วยเหลือจนสามารถข้ามผ่านปัญหาและอุปสรรคต่างๆในการทำวิทยานิพนธ์นี้ไปได้ด้วยดี

วิทยานิพนธ์นี้สำเร็จด้วยดีเพราะแรงบัลดาลใจจาก อ.ดร.อติวงศ์ สุชาโต และ อ.ดร.โปรดปราน บุณยพุกกณะ ที่เป็นผู้ชี้แนะแนวทาง และจุดประกายให้เกิดความสนใจและความหลงใหลในเทคโนโลยีเสียงพูด นอกจากนี้ อยากขอขอบพระคุณเป็นพิเศษ สำหรับคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ รศ. ดร. บุญเสริม กิจศิริกุล ผู้ซึ่งเป็นประธาน อ. ดร.พิษณุ คนองชัยยศ และ ดร. ชัย วุฒิวิวัฒน์ชัย ที่สละเวลาอันมีค่ามาชี้ให้เห็นถึงข้อบกพร่อง รวมทั้งข้อแนะนำที่น่าสนใจ

ขอขอบคุณสมาชิกของห้องปฏิบัติการ SLS ทุกคน รวมทั้งพี่ ๆ และเพื่อน ๆ ระดับบัณฑิตศึกษาที่สร้างบรรยากาศที่ดีในการทำงาน

ขอขอบคุณจุฬาลงกรณ์มหาวิทยาลัย และโรงเรียนเทพศิรินทร์ ที่ได้ประสิทธิประสาทวิชาความรู้

สุดท้ายนี้ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ และญาติ ๆ ที่ได้ผลักดันจนผู้เขียนประสบความสำเร็จอย่างเช่นทุกวันนี้

งานวิจัยนี้ได้รับเงินทุนสนับสนุนจากคณะวิศวกรรมศาสตร์ในโครงการจัดการศึกษาสาขาวิชาวิศวกรรมศาสตร์ เพื่อเพิ่มศักยภาพทางด้านวิทยาศาสตร์เทคโนโลยี และอุตสาหกรรม หมวดเงินอุดหนุนการศึกษาประจำปีการศึกษา 2549

สารบัญ

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

หน่วยเสียง คือ ส่วนย่อยของเสียงพูดที่มีขนาดเล็กที่สุดที่ทำให้เสียงของคำในภาษามีความหลากหลายแตกต่างกัน และขอบเขตของหน่วยเสียงก็คือตำแหน่งบอกเวลาเริ่มต้นและสิ้นสุดของหน่วยเสียงนั้นๆ โดยส่วนใหญ่ระบบรู้จำเสียงพูดจะแยกความแตกต่างระหว่างเสียงของคำในภาษาด้วยการพิจารณารูปแบบการประกอบกันของหน่วยเสียง เช่นเดียวกันกับระบบสังเคราะห์เสียง ที่นำหน่วยเสียงย่อยๆมาประกอบกลับเป็นเสียงของคำ ด้วยเหตุนี้สิ่งที่จำเป็นต่อการสร้างระบบการรู้จำและการสังเคราะห์เสียงในภาษา คือการมีข้อมูลเสียงที่มีการกำกับขอบเขตของหน่วยเสียง หรือจัดวางตำแหน่งทางเวลาของแต่ละหน่วยเสียงไว้ก่อนแล้ว

ในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ (Segment-based Speech Recognition) [1] จะพิจารณาหน่วยเสียงให้เป็นเซกเมนต์ โดยขั้นตอนการทำงานจะประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนที่ทำงานต่อเนื่องกัน คือ ขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ และขั้นตอนการรู้จำเสียงพูด จุดประสงค์ของขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ ก็เพื่อสันนิษฐานหาขอบเขตของเซกเมนต์แล้วนำมาประกอบกันเป็นกราฟของเซกเมนต์ โดยในขั้นตอนการรู้จำเสียงพูด กราฟนี้จะถูกใช้เป็นข้อมูลอินพุตเพื่อค้นหาลำดับของหน่วยเสียงที่ดีที่สุดออกมา ด้วยเหตุนี้สิ่งที่จำเป็นและส่งผลโดยตรงต่อประสิทธิภาพในการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ คือประสิทธิภาพของกระบวนการแบ่งเสียงพูดเป็นเซกเมนต์ ซึ่งจะต้องมีความถูกต้องแม่นยำและทำงานได้อย่างรวดเร็ว

การแบ่งเสียงพูดเป็นเซกเมนต์สามารถกระทำได้ด้วยคนหรือด้วยวิธีแบบอัตโนมัติ แม้ว่าการกระทำด้วยคนจะได้ผลลัพธ์ที่มีความแม่นยำสูงกว่า แต่ก็อาศัยเวลามากเมื่อมีข้อมูลเป็นจำนวนมาก โดยในการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยคนจะใช้เวลาประมาณ 11 ถึง 30 วินาทีต่อหนึ่งหน่วยเสียง [2] ในระหว่างที่การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยกระบวนการแบบอัตโนมัติใช้เวลาอยู่ระหว่าง 0.001 ถึง 0.009 วินาทีต่อหนึ่งหน่วยเสียง บนเครื่องคอมพิวเตอร์ที่มีหน่วยประมวลผลกลางเพนเทียมเอ็ม 1.4 กิกะเฮิรต์ ความแตกต่างนี้จะมากยิ่งขึ้นเมื่อประสิทธิภาพและความเร็วในการประมวลผลบนคอมพิวเตอร์ได้รับการพัฒนามากขึ้น ในขณะที่ความเร็วของคนในการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยมือจะยังคงเท่าเดิม

วิธีการแบ่งเสียงพูดเป็นเซกเมนต์ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ในปัจจุบันยังมีประสิทธิภาพไม่ดีพอ โดยระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ SUMMIT ของ MIT [3] ซึ่งใช้วิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด มีประสิทธิภาพแย่กว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยคนอยู่ประมาณ 20 ถึง 30% เมื่อทดลองกับฐานข้อมูลเสียง TIMIT [4]

วิทยานิพนธ์นี้เสนอวิธีการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์แบบใหม่ โดยนำสารสนเทศสวนสัทศาสตร์ (Acoustic-Phonetic Information) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) มาประยุกต์ใช้เพื่อให้ประสิทธิภาพทั้งในด้านคุณภาพกราฟของเซกเมนต์และความเร็วในการแบ่งเสียงพูดเป็นเซกเมนต์ สูงกว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

วัตถุประสงค์ของการวิจัย

เป้าหมายของวิทยานิพนธ์นี้ ต้องการที่จะพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์เพื่อนำไปใช้งานในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยอาศัยเทคนิคการเรียนรู้ของเครื่องแบบซัพพอร์ตเวกเตอร์แมชชีน และการสกัดลักษณะสำคัญโดยใช้สารสนเทศสวนสัทศาสตร์เพื่อเพิ่มความแม่นยำในหาขอบเขตของหน่วยเสียง ซึ่งส่งผลให้ได้กราฟของเซกเมนต์ที่มีคุณภาพดีกว่า รวมถึงสามารถทำงานได้รวดเร็วกว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ขอบเขตของการวิจัย

ในขอบเขตของงานวิจัยนี้จะพัฒนาเฉพาะส่วนของขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์เพียงอย่างเดียวเท่านั้น ไม่ได้พัฒนาในส่วนของการรู้จำเสียงพูดด้วยการค้นหาและให้คะแนนในขอบข่ายงานของการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ส่วนการวัดประสิทธิภาพจะทำการทดลองเปรียบเทียบโดยใช้ฐานข้อมูลเสียงที่มีข้อมูลทางเวลากำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว

รูปที่ 1.1 แผนภาพส่วนประกอบในระบบการรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ขั้นตอนการวิจัย

1. ขั้นตอนเตรียมตัว
    1. ศึกษาข้อมูลและทฤษฎีต่างๆเกี่ยวกับงานวิจัย เช่น ทฤษฎีทางภาษาศาสตร์ การวิเคราะห์เสียงพูด แบบจำลองทางภาษา การรู้จำเสียงแบบอาศัยเซกเมนต์ การตรวจหาขอบเขตของหน่วยเสียงในเสียงพูด และความรู้อื่นๆ

2. ศึกษาเกี่ยวกับสารสนเทศสวนสัทศาสตร์
3. ศึกษางานวิจัยที่เกี่ยวข้อง

2. ขั้นตอนออกแบบวิธีการในการแบ่งเสียงพูดเป็นเซกเมนต์
   1. ศึกษาคุณสมบัติของหน่วยเสียงและออกแบบเซตของค่าลักษณะสำคัญที่เหมาะสมต่อการนำไปใช้ในกระบวนการตรวจหาขอบเขตของหน่วยเสียงในเสียงพูด
   2. ออกแบบขั้นตอนในการแบ่งเสียงพูดเป็นเซกเมนต์
3. ขั้นตอนพัฒนาและทดสอบระบบการแบ่งเสียงพูดเป็นเซกเมนต์
   1. จัดทำโปรแกรมในแต่ละส่วนที่ได้ออกแบบไว้สำหรับใช้ทำการแบ่งเสียงพูดเป็นเซกเมนต์
   2. ทดสอบประสิทธิภาพของการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยการทดสอบกับฐานข้อมูลเสียงภาษาไทย
4. ขั้นตอนการวิเคราะห์และสรุปผล
   1. วิเคราะห์และสรุปผลการทดลอง
   2. เรียบเรียงวิทยานิพนธ์

ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถปรับปรุงและพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์บนโดเมนภาษาไทยไปใช้สร้างระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่มีประสิทธิภาพได้
2. นำไปใช้ประโยชน์ในการเตรียมข้อมูลเสียงสำหรับระบบสังเคราะห์เสียงภาษาไทยและประยุกต์ใช้กับเทคโนโลยีเสียงพูดอื่นๆได้ด้วย

ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง “Locating Phone Boundaries from Acoustic Discontinuities using a Two-staged Approach” โดย ไพโรจน์ ลีลาภัทรกิจ อติวงศ์ สุชาโต และ โปรดปราน บุณยพุกกณะ ในงานประชุมวิชาการ “The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)” ณ. เมืองพิทส์เบิร์ก รัฐเพนซิลเวเนีย ประเทศสหรัฐอเมริกา ในระหว่างวันที่ 17-21 กันยายน 2549

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้จะนำเสนอทฤษฎีพื้นฐานและแนวคิดที่เกี่ยวข้องกับการพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์และการพัฒนาระบบรู้จำเสียงพูด โดยเริ่มจากทฤษฎีทางภาษาศาสตร์ ซึ่งเป็นทฤษฎีที่ศึกษาปรากฏการณ์ของเสียงพูดในด้านต่างๆดังนี้

- สรีรสัทศาสตร์ (Articulatory Phonetics)
- สวนสัทศาสตร์ (Acoustic Phonetics)

จากนั้นจะนำเสนอทฤษฎีการวิเคราะห์เสียงพูดในโดเมนความถี่ด้วยการแปลงฟูริเยร์แบบวิยุต (Discrete Fourier Transform) และสเปกโตรแกรม (Spectrogram) ถัดจากนั้นจะนำเสนอการสกัดลักษณะสำคัญ แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model) การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ และ ซัพพอร์ตเวกเตอร์แมชชีน - เอสวีเอ็ม (Support Vector Machine - SVM) ตามลำดับ รวมทั้งเสนองานวิจัยต่างๆที่เกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์

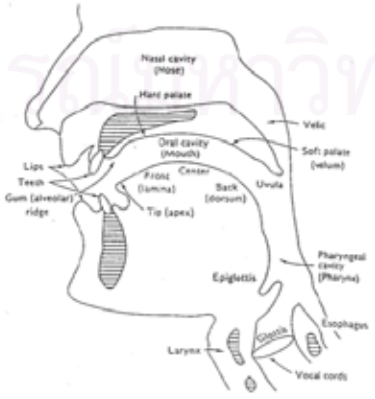
ทฤษฎีที่เกี่ยวข้อง

1. สรีรสัทศาสตร์

สรีรสัทศาสตร์เป็นการศึกษาเสียงพูดจากอวัยวะและการเคลื่อนไหวของอวัยวะที่ทำให้เกิดเสียงพูด โดยในหัวข้อนี้จะนำเสนอเกี่ยวกับอวัยวะที่ทำให้เกิดเสียง และเสียงในภาษาไทยซึ่งได้แก่ เสียงพยัญชนะ เสียงสระ และเสียงวรรณยุกต์ ตามลำดับ

1.1 อวัยวะที่ทำให้เกิดเสียง (The Organs of Speech) [5]

ตำแหน่งของอวัยวะที่ทำให้เกิดเสียงของมนุษย์สามารถแสดงได้ดังรูปที่ 2.1

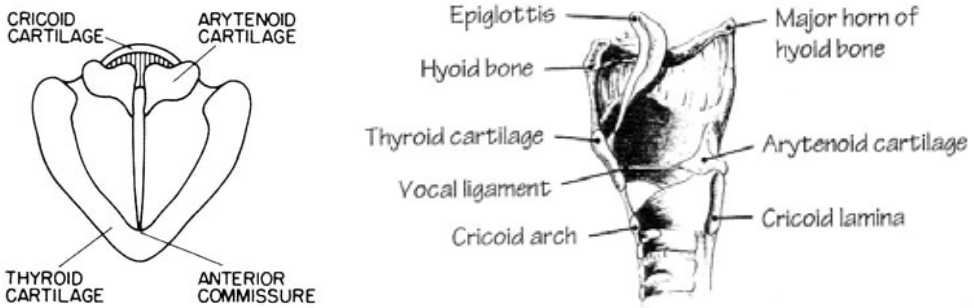


รูปที่ 2.1 อวัยวะภายในของระบบการพูดของมนุษย์ [6]

อวัยวะที่ใช้ในการออกเสียงทั้งหมดมีดังนี้

1. **ริมฝีปาก** (Lips) เป็นอวัยวะส่วนที่สามารถเคลื่อนไหวได้และทำให้เสียงแตกต่างกันได้มาก เราอาจจะบังคับริมฝีปากให้ปิดสนิท ให้เปิดเล็กน้อย ให้เปิดกว้างขึ้น ให้ยื่นออกมา ให้ห่อกลมหรือทำเป็นรูปรีก็ได้ ลักษณะต่างๆ ของริมฝีปากล้วนมีผลต่อการออกเสียงทั้งสิ้น เสียงพยัญชนะที่เกิดจากการกักที่ริมฝีปากเรียกว่าเสียงโอฐชะ (Bilabial Sound)
2. **ฟัน** (Teeth) เป็นอวัยวะที่เป็นฐานหรือตำแหน่งที่เกิดของเสียงหลายชนิด เช่น เมื่อฟันบนกดลงบนริมฝีปากล่าง ลมที่ผ่านออกมาโดยแรงจะลอดช่องที่พอจะผ่านได้ออกมา ทำให้เกิดเป็นเสียงชนิดที่เรียกว่า เสียงเสียดแทรกที่เกิดระหว่างฟันกับริมฝีปาก ถ้าฟันบนกดกับฟันล่าง ลมที่ผ่านออกมาโดยแรงจะทำให้ได้เสียงเสียดแทรกที่เกิดที่ฟัน เป็นต้น นอกจากนี้เนื่องจากปลายลิ้นอยู่ใกล้กับฟัน ปลายลิ้นจึงมักจะทำอาการต่างๆ บริเวณฟันและหลังฟันบ่อยๆ ทำให้เกิดเสียงทันตชะ (Dental Sound)
3. **ปุ่มเหงือก** (Alveolus, Gum Ridge, Tooth Ridge) เป็นส่วนที่นูนออกมาตรงบริเวณหลังฟันด้านบน ถ้าเอาลิ้นแตะดูจะรู้สึกว่ามีลักษณะเป็นคลื่น ลิ้นอาจแตะหรือวางอยู่ใกล้บริเวณปุ่มเหงือก ซึ่งทำให้เกิดเสียงมุทธชะ (Alveolar Sound)
4. **เพดานแข็ง หรือเพดานปาก** (Palate, Hard Palate) หมายถึงส่วนโค้งของเพดานปากส่วนที่เป็นกระดูกแข็ง ซึ่งอยู่ถัดจากปุ่มเหงือกเข้ามา ถ้าลิ้นแตะหรือวางใกล้เพดานแข็งจะทำให้เกิดเสียงตาลุชะ (Palatal Sound)
5. **เพดานอ่อน** (Velum, Soft Palate) คือ ส่วนของเพดานที่อยู่ต่อเพดานแข็งเข้าไปข้างในเป็นกระดูกอ่อนที่ขยับขึ้นลงได้เล็กน้อย เวลาหายใจเพดานอ่อนและลิ้นไก่ซึ่งอยู่ปลายเพดานอ่อนจะลดระดับลงมาเปิดช่องให้ลมออกทางจมูก ฉะนั้นเวลาที่ไม่พูด เพดานอ่อนและลิ้นไก่จะลดระดับลงมา เวลาพูดส่วนใหญ่เพดานอ่อนและลิ้นไก่จะถูกยกขึ้นไปจดกับผนังคอ จะมีแต่เวลาออกเสียงนาสิกเท่านั้นที่เพดานอ่อนจะลดระดับลงมาเพื่อให้ลมออกไปทางจมูกได้ ถ้าลิ้นแตะหรือวางใกล้เพดานอ่อนจะทำให้เกิดเสียงที่เกิดที่เพดานอ่อน (Velar Sound)
6. **ลิ้นไก่** (Uvula) เป็นก้อนเนื้อเล็กๆ อยู่ต่อจากปลายเพดานอ่อนเข้าไปข้างใน และห้อยอยู่ตรงกลางปาก สามารถสั่นรัวได้ เวลาอ้าปากมักจะเห็น ลิ้นไก่ใช้ออกเสียงในบางภาษา เช่น ภาษาเยอรมัน ฝรั่งเศส นอร์เวย์ อาหรับ และอิสราเอล เป็นต้น
7. **ช่องจมูก** (Nasal Cavity) หมายถึง โพรงในช่องจมูกซึ่งอยู่เหนือลิ้นไก่ขึ้นไป เป็นช่องที่ลมซึ่งผ่านเส้นเสียงขึ้นมาจะผ่านออกไปทางจมูกได้เมื่อเวลาหายใจและเวลาออกเสียงนาสิก ในเวลาเปล่งเสียงอื่นๆลิ้นไก่จะถูกยกขึ้นไปปิดช่องจมูกเพื่อให้ลมออกมาทางช่องปาก

8. **ลิ้น** (Tongue) เป็นส่วนที่เคลื่อนไหวได้มากที่สุดในการออกเสียงพูด ส่วนที่เคลื่อนไหวของลิ้นแต่ละส่วนมีผลต่อการออกเสียง เราจึงแบ่งลิ้นออกเป็น 3 ส่วนด้วยกันตามหน้าที่ที่มีในการออกเสียงคือ
    1. ปลายลิ้น (Tip of the Tongue) หรือ ลิ้นส่วนปลายสุด หมายถึงส่วนปลายของลิ้นซึ่งสามารถจะยกขึ้นไปแตะอวัยวะส่วนต่างๆ ในปากตอนบนได้โดยง่าย
    2. หน้าลิ้น (Blade of the Tongue) หรือ ลิ้นส่วนหน้า หมายถึงลิ้นส่วนที่อยู่ตรงข้ามกับเพดานแข็ง ในขณะที่วางลิ้นราบกับปากตอนไม่ได้พูด
    3. หลังลิ้น (Back of the Tongue) หรือ ลิ้นส่วนหลัง หมายถึงส่วนของลิ้นที่อยู่ตรงข้ามกับเพดานอ่อน ในขณะที่วางลิ้นราบกับปากตอนไม่ได้พูด
9. **แผ่นเนื้อปากหลอดลม** (Epiglottis) หรือ ลิ้นปิดกล่องเสียง เป็นก้อนเนื้อเล็กๆ คล้ายลิ้นไก่อยู่ต่อโคนลิ้นลงไปในคอ มีหน้าที่ปิดเปิดช่องหลอดลม เพื่อป้องกันมิให้อาหารตกลงไปในหลอดลม ในเวลาที่กลืนอาหาร แผ่นเนื้อปากหลอดลมปิดลงให้อาหารผ่านไปลงหลอดอาหาร แต่ในเวลาที่พูด แผ่นเนื้อนี้จะเปิดออกเพื่อให้ลมจากหลอดลมออกมา
10. **โพรงคอ** (Pharynx) เป็นโพรงซึ่งอยู่ถัดปากลงไปจากช่องปากจนถึงเส้นเสียงหรือสายเสียง
11. **เส้นเสียง หรือสายเสียง** (Vocal Cords) เป็นอวัยวะสำคัญที่ทำให้เกิดเสียง เส้นเสียงประกอบด้วยเส้นเอ็นและกล้ามเนื้อเป็นแผ่น 2 แผ่น มีความยาวประมาณ 1.2-1.7 เซนติเมตร กว้างประมาณ 0.2-0.3 เซนติเมตร ปิดขวางอยู่ตรงปากของช่องหลอดลม โดยจะวางตัวจากด้านหลังมายังด้านหน้าอยู่ตรงกลางของกล่องเสียง เส้นเสียงทั้งสองสามารถที่จะดึงออกให้ห่างจากกันหรือดึงเข้ามาให้ชิดกันก็ได้ เส้นเสียงเป็นส่วนสำคัญที่ทำให้เกิดเสียงพูด โดยจะเปิดให้ลมผ่านในเวลาหายใจตามปกติ แต่จะอยู่ชิดกันเมื่อมีการเปล่งเสียง
12. **กล่องเสียง** (Larynx) ตั้งอยู่ตอนบนของหลอดลมตรงตำแหน่งที่เรียกว่าลูกกระเดือก (Adam's Apple) กล่องเสียงประกอบด้วยกระดูกอ่อนหลายส่วนด้วยกัน ส่วนที่อยู่ด้านหน้า คือ กระดูกอ่อนไทรอยด์ (Thyroid Cartilage) ปลายด้านหนึ่งของเส้นเสียงทั้งสองจะเชื่อมอยู่กับกระดูกอ่อนไทรอยด์นี้และอยู่ชิดกัน ส่วนปลายอีกด้านหนึ่งของเส้นเสียงทั้งสอง จะเชื่อมอยู่กับกระดูกอ่อนอาริตินอยด์ (Arytenoids Cartilages) ซึ่งเป็นกระดูกอ่อนอีกสองชิ้น กระดูกอ่อนอาริตินอยด์และกล้ามเนื้อในกล่องเสียงจะทำให้เส้นเสียงทั้งสองอยู่ชิดติดกันหรือห่างจากกันได้ เมื่อเส้นเสียงอยู่ห่างจากกันจะเกิดเป็นช่องสามเหลี่ยม ซึ่งเป็นทางให้ลมผ่านเข้าไปถึงปอด หรือผ่านออกมาจากปอดได้ ดังรูปที่ 2.2



รูปที่ 2.2 กล่องเสียง (ซ้าย : กล่องเสียงด้านหน้า, ขวา : กล่องเสียงด้านหลัง) [7]

13. **ช่องระหว่างเส้นเสียง** (Glottis) จะเปิดอยู่ระหว่างที่หายใจเข้าออกตามปกติ แต่จะปิดลงเมื่อมีการเปล่งเสียง ก่อให้เกิดการสั่น และเป็นเสียงดังขึ้น
14. **ช่องปาก** (Oral Cavity) ทำหน้าที่เป็นช่องกำทอน (Resonant Chamber) ซึ่งสามารถเปลี่ยนให้มีรูปร่างต่างๆ กัน ตามท่าทางของอวัยวะภายในช่องปาก โดยอวัยวะภายในช่องปากอาจสามารถแบ่งได้ดังนี้
    1. อวัยวะส่วนกระทำอาการ (Articulator) หมายถึงอวัยวะส่วนที่เคลื่อนไหวเพื่อผลักหรือกักลมในที่ต่างๆ อวัยวะส่วนกระทำอาการที่สำคัญคือลิ้น ซึ่งเคลื่อนไหวได้มากที่สุด อวัยวะส่วนกระทำอาการอาจเรียกว่า “กรณ์”
    2. อวัยวะส่วนเกิดอาการ (Point of Articulation) หมายถึง ตำแหน่งที่อวัยวะส่วนกระทำอาการเคลื่อนไหวไป เพื่อผลักหรือกักลมไว้ อาจเรียกอวัยวะส่วนนี้ว่า “ฐาน” ที่เกิดของหน่วยเสียงต่างๆ ฐานภายในช่องปากที่สำคัญได้แก่ ริมฝีปาก ฟัน ปุ่มเหงือก เพดานแข็ง และเพดานอ่อน
15. **หลอดลม** (Trachea) เป็นทางเดินอากาศจากปอดถึงกล่องเสียง

1.2 เสียงพยัญชนะในภาษาไทย (Thai Consonants) [5]

1.2.1 เสียงพยัญชนะ

เสียงพยัญชนะ (Consonant) หมายถึงเสียงของลมที่ผ่านปอดขึ้นมายังกล่องเสียงแล้วปะทะกับอวัยวะต่างๆ ในช่องปาก ทำให้ลมเพียงส่วนหนึ่งหรือทั้งหมดพบกับอุปสรรคที่อยู่เหนือช่องของเส้นเสียง โดยอุปสรรคเหล่านี้เกิดจากการทำงานประสานกันของอวัยวะในช่องปาก เสียงพยัญชนะที่เกิดขึ้นมาจึงมีหลายแบบแตกต่างกัน ซึ่งเสียงที่แตกต่างกันมักจะทำให้ความหมายของเสียงในภาษาแตกต่างกันไปด้วย คุณสมบัติที่ทำให้เสียงพยัญชนะแตกต่างกันมีดังนี้

1. **คุณสมบัติความก้องของเสียง** ความก้องของเสียงเป็นคุณสมบัติที่ใช้ในการแบ่งแยกเสียงพยัญชนะออกได้เป็นสองชนิด คือ

1. เสียงพยัญชนะโฆษะ (Voiced) หรือเสียงก้อง เป็นเสียงพยัญชนะที่เส้นเสียงสั่นสะเทือนขณะที่เปล่งเสียง
2. เสียงพยัญชนะอโฆษะ (Voiceless) หรือเสียงไม่ก้อง เป็นเสียงพยัญชนะที่เส้นเสียงไม่สั่นสะเทือนขณะที่เปล่งเสียง

2. **ลักษณะของลมที่ผ่านเส้นเสียง** เสียงพยัญชนะสามารถแบ่งตามลักษณะลมที่ผ่านเส้นเสียงออกมาได้ดังนี้
   1. เสียงพยัญชนะหยุด (Stop) อาจแบ่งออกเป็น 2 ลักษณะย่อยๆ ได้แก่ เสียงพยัญชนะระเบิด (Plosive Stop) และเสียงพยัญชนะกัก (Unreleased Stop) เสียงพยัญชนะระเบิดเกิดจากการที่ลมซึ่งเปล่งออกมาถูกกักเอาไว้ ณ ที่ใดที่หนึ่งในช่องปาก แล้วช่องที่กักนั้นเปิดให้ลมพุ่งออกมา เสียงพยัญชนะระเบิดแบ่งออกได้อีกเป็นเสียงพยัญชนะระเบิดมีลม (Aspirated Plosive) หรือธนิต ซึ่งจะมีลมหายใจพ่นออกมาหลังเปล่งเสียงและเสียงพยัญชนะระเบิดไม่มีลม (Unaspirated Plosive) หรือสิถิล ซึ่งไม่มีลมหายใจพ่นออกมา ส่วนเสียงพยัญชนะกักเกิดจากลมซึ่งเปล่งออกมาถูกกักไว้ ณ ที่ใดที่หนึ่งในช่องปาก โดยเสียงพยัญชนะกักนี้มักจะเป็นเสียงตัวสะกดท้ายพยางค์
   2. เสียงพยัญชนะเสียดแทรก (Fricative) เป็นเสียงพยัญชนะที่เมื่อออกเสียงแล้วลมที่ผ่านขึ้นมาถูกบังคับให้ต้องบีบตัวผ่านช่องแคบๆ ที่ใดที่หนึ่งในช่องปาก ซึ่งเสียงเสียดแทรกนี้เราจะทำค้างไว้นานเท่าใดก็ได้ ตราบเท่าที่ลมหายใจจะอำนวย
   3. เสียงพยัญชนะนาสิก (Nasal) เป็นเสียงพยัญชนะที่มีลมผ่านออกมาทางจมูก ซึ่งเกิดจากการที่ลมมากักอยู่ในช่องปาก แล้วเพดานอ่อนและลิ้นไก่ลดระดับลง
   4. เสียงพยัญชนะข้างลิ้น (Lateral) เป็นเสียงที่เกิดจากการนำลิ้นปิดบริเวณปุ่มเหงือกและเพดานแข็งส่วนกลางไว้ แล้วปล่อยให้ลมผ่านออกมาทางข้างลิ้น
   5. เสียงพยัญชนะรัว (Trill) เกิดจากการที่อวัยวะส่วนใดส่วนหนึ่งในช่องปากกระทบกับอวัยวะอีกส่วนหนึ่งในขณะที่ลมถูกพ่นผ่านอวัยวะนั้นออกมาอย่างรุนแรง
   6. เสียงพยัญชนะกึ่งสระ (Semi-vowel, Approximant) หรืออรรธสระ เป็นเสียงเลื่อน (Glide) ที่เกิดขึ้นระหว่างเสียงสระสองเสียง ในการเปล่งเสียงพยัญชนะกึ่งสระอวัยวะที่ใช้ในการออกเสียงจะอยู่ในตำแหน่งของการออกเสียงสระใดสระหนึ่งก่อน แล้วจึงเปล่งเสียงออกมาขณะที่เปลี่ยนตำแหน่งอวัยวะไปสู่การออกเสียงของอีกสระหนึ่ง

3. **ฐานที่เกิดของเสียง** ไม่ว่าลมที่ใช้ในการออกเสียงพยัญชนะนั้นจะมาถูกกัก ถูกดัดแปลงจนเกิดการกัก หรือการเสียดแทรก จำเป็นต้องมีตำแหน่งที่เกิดอยู่ด้วยเสมอในช่องปาก

1.2.2 เสียงพยัญชนะภาษาไทย

พยัญชนะในภาษาไทยมีทั้งหมด 44 รูป 21 หน่วยเสียง แบ่งเป็น 2 กลุ่มใหญ่ๆ คือ พยัญชนะกัก (Stop Consonants) 11 หน่วยเสียง และพยัญชนะไม่กัก (Non-stop Consonants) 10 หน่วยเสียง ดังแสดงในตารางที่ 2.1 ทั้งนี้หน่วยเสียงพยัญชนะทั้ง 21 หน่วยเสียง สามารถที่จะอยู่ในต้นพยางค์ได้ทุกหน่วยเสียง แต่จะมีหน่วยเสียงพยัญชนะที่ปรากฏท้ายพยางค์ได้เพียง 9 หน่วยเสียงเท่านั้น คือ เสียงพยัญชนะกัก 4 หน่วยเสียง (/p/, /t/, /k/, /ʔ/) เสียงพยัญชนะนาสิก 3 หน่วยเสียง (/m/, /n/, /ŋ/) และเสียงพยัญชนะกึ่งสระ 2 หน่วยเสียง (/w/, /j/) ส่วนพยัญชนะต้นควบกล้ำในภาษาไทยแท้เป็นได้ 11 หน่วยเสียง คือ /pr/, /pʰr/, /pl/, /pʰl/, /tr/, /kr/, /kʰr/, /kl/, /kʰl/, /kw/, /kʰw/ ส่วนพยัญชนะต้นควบกล้ำในภาษาไทยทับศัพท์อังกฤษมีได้ 6 หน่วยเสียง คือ /br/, /bl/, /dr/, /fr/, /fl/, /tr/ ส่วนคำไทยที่ยืมมาจากภาษาสันสกฤตก็ควบ /tr/ ได้เช่นกัน

ตารางที่ 2.1 เสียงพยัญชนะภาษาไทย

1 (*) ปรากฏท้ายพยางค์ได้

2 (/…/) ปรากฏเฉพาะในคำไทยทับศัพท์ภาษาอังกฤษ

3 [/…/] ปรากฏในคำไทยทับศัพท์ภาษาอังกฤษ หรือคำไทยที่ยืมมาจากภาษาสันสกฤต

หน่วยเสียง[1]	หน่วยเสียงควบกล้ำ	ลักษณะของลม	การพ่นลม	ความก้อง	ฐานที่เกิด	รูปพยัญชนะ
/p/ (*)	/pr/, /pl/	กัก	ไม่พ่นลม	ไม่ก้อง	ริมฝีปาก	ป
/pʰ/	/pʰr/,/pʰl/	กัก	พ่นลม	ไม่ก้อง	ริมฝีปาก	ผ พ ภ
/b/	/br/,/bl/	กัก	ไม่พ่นลม	ก้อง	ริมฝีปาก	บ
/t/ (*)	/tr/	กัก	ไม่พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฏ ต
/tʰ/	/tʰr/	กัก	พ่นลม	ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ฐ ฑ ฒ ถ ท ธ
/d/	/dr/	กัก	ไม่พ่นลม	ก้อง	ฟัน หรือ ปุ่มเหงือก	ฎ ด
/c/		กัก	ไม่พ่นลม	ไม่ก้อง	เพดานแข็ง	จ
/cʰ/		กัก	พ่นลม	ไม่ก้อง	เพดานแข็ง	ฉ ช ฌ
/k/ (*)	/kr/, /kl/, /kw/	กัก	ไม่พ่นลม	ไม่ก้อง	เพดานอ่อน	ก
/kʰ/	/kʰr/, /kʰl/, /kʰw/	กัก	พ่นลม	ไม่ก้อง	เพดานอ่อน	ข ฃ ค ฅ ฆ
/ʔ/ (*)		กัก	ไม่พ่นลม	ไม่ก้อง	เส้นเสียง	อ
/m/ (*)		นาสิก		ก้อง	ริมฝีปาก	ม
/n/ (*)		นาสิก		ก้อง	ฟัน หรือ ปุ่มเหงือก	ณ น
/ŋ/ (*)		นาสิก		ก้อง	เพดานอ่อน	ง
/f/	/fr/, /fl/	เสียดแทรก		ไม่ก้อง	ริมฝีปาก	ฝ ฟ
/s/		เสียดแทรก		ไม่ก้อง	ฟัน หรือ ปุ่มเหงือก	ซ ศ ษ ส
/h/		เสียดแทรก		ไม่ก้อง	เส้นเสียง	ห ฮ
/r/		รัว		ก้อง	ฟัน หรือ ปุ่มเหงือก	ร
/l/		ข้างลิ้น		ก้อง	ฟัน หรือ ปุ่มเหงือก	ล ฬ
/w/ (*)		กึ่งสระ		ก้อง	ริมฝีปาก–เพดานอ่อน	ว
/j/ (*)		กึ่งสระ		ก้อง	เพดานแข็ง	ญ ย

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

1.3 เสียงสระในภาษาไทย (Thai Vowels) [5]

1.3.1 เสียงสระ

เสียงสระ (Vowel) เป็นเสียงซึ่งถูกเปล่งออกมาทางช่องปากหรือช่องจมูกโดยไม่มีอวัยวะส่วนใดในปากมาเป็นอุปสรรคปิดกั้นทางลมไว้เลย เสียงสระเกิดจากการที่ลมผ่านเส้นเสียงในตำแหน่งที่เส้นเสียงทั้งสองอยู่ชิดกันมากจนเกือบปิดสนิท ทำให้ลมต้องดันตัวออกมาอย่างรุนแรงจนเส้นเสียงเกิดการสั่นสะเทือน และส่งผลทำให้เกิดเสียงดังที่เป็นเสียงก้อง โดยคุณสมบัติที่ทำให้เสียงสระมีความแตกต่างกันมีดังนี้

1. **ส่วนของลิ้นที่ใช้ในการเปล่งเสียง** (Place of Articulation) ในขณะที่ออกเสียงสระต่างๆ ลิ้นหลายส่วนที่ใช้ในการออกเสียงสระ ไม่ว่าจะเป็นลิ้นส่วนหน้า ลิ้นส่วนกลาง หรือลิ้นส่วนหลัง โดยลิ้นส่วนนั้นๆ จะยกขึ้นใกล้เพดานปากในขณะที่ออกเสียงสระหนึ่งๆ ก่อให้เกิดเสียงสระที่แตกต่างกัน โดยถ้าลิ้นส่วนหน้ายกขึ้นให้จุดสูงสุดอยู่ใกล้เพดานแข็ง เราก็จะเรียกเสียงสระนั้นว่าเสียงสระส่วนเพดานแข็ง หรือสระหน้า (Front Vowel) เช่น สระอี สระเอ สระแอ เป็นต้น แต่ถ้าการออกสียงสระใดใช้ลิ้นส่วนหลัง โดยทำการยกลิ้นส่วนหลังขึ้นให้จุดสูงสุดอยู่ใกล้เพดานอ่อนเราก็จะเรียกเสียงสระนั้นว่าเป็นเสียงสระส่วนเพดานอ่อน หรือสระหลัง (back vowel) เช่นสระอู สระโอ สระออ เป็นต้น ส่วนถ้าในการออกเสียงสระใดลิ้นส่วนกลางถูกยกขึ้นไปยังส่วนกลางของเพดานปาก เราก็จะเรียกเสียงสระนั้นว่า สระกลาง (Central Vowel) เช่นสระอือ สระเออ สระอา เป็นต้น
2. **ระยะห่างระหว่างลิ้นและเพดานปากหรือความสูงของลิ้น** (Degree of Stricture) ระยะห่างระหว่างลิ้นและเพดานปากเป็นลักษณะที่สำคัญอย่างหนึ่งในการแบ่งชนิดของเสียงสระ โดยระยะห่างนี้จะเป็นตัวกำหนดว่าเสียงสระที่เปล่งออกมาเป็นสระเปิดหรือสระปิด ถ้าหากลิ้นอยู่ห่างจากเพดานปากมาก หรือลิ้นอยู่ในระดับต่ำ ทำให้ช่องโพรงปากกว้างลมก็จะผ่านออกมาได้มาก เสียงสระที่ได้จะเป็นสระเปิด (Open Vowel) เช่น สระอา ในทางตรงกันข้าม ถ้าตำแหน่งของลิ้นอยู่ใกล้กับเพดานปากมาก หรือลิ้นอยู่ในระดับสูง ช่องโพรงในปากก็จะแคบ ทำให้ลมผ่านออกมาได้น้อย เสียงสระที่ได้จะเป็นสระปิด (Close Vowel) เช่น สระอี สระอู เป็นต้น แต่ถ้าระยะห่างระหว่างลิ้นกับเพดานปากอยู่ในระหว่างสระเปิดและสระปิด เช่นเสียงสระที่เปิดกว้างกว่าสระปิดเล็กน้อย เราก็จะเรียกว่าเป็นสระกลางปิดหรือสระกึ่งปิด (Close-mid, Half-close Vowel) เช่น สระเอ สระโอ เป็นต้น แต่ถ้าเปิดกว้างขึ้นอีก จะเรียกว่าเป็นสระกลางเปิดหรือสระกึ่งเปิด (Open-mid, Half-open Vowel) เช่นสระ แอ สระออ เป็นต้น

3. **การห่อริมฝีปาก** (Labialization) หมายถึงการที่ริมฝีปากทั้งสองเคลื่อนไหวโดยยื่นตัวไปข้างหน้า แล้วห่อกลมมากน้อยเพียงใด ถ้าริมฝีปากยื่นออกไปข้างหน้าแล้วห่อกลมมากเสียงสระที่ได้จะเรียกว่าสระกลม (Rounded Vowel) เช่น สระอู สระโอ สระออ เป็นต้น แต่ถ้าริมปากทั้งสองฉีกออกหรือไม่ห่อกลมขณะเปล่งเสียง สระที่ได้ก็จะเป็นสระไม่กลม (Unrounded Vowel) เช่น สระอี สระเอ สระแอ สระอา เป็นต้น
4. **ลักษณะนาสิก** (Nasalization) เป็นลักษณะในการออกเสียงสระที่ทำให้เกิดเสียงสระขึ้นจมูกหรือสระนาสิก (Nasal Vowel) ขึ้น ซึ่งจะทำให้เสียงแตกต่างจากสระโอฐชะ (Oral Vowel) กล่าวคือ ในการเปล่งเสียงสระโอฐชะนั้น เพดานอ่อนจะยกขึ้นปิดโพรงจมูก อากาศจึงไม่สามารถผ่านออกไปทางช่องจมูกได้ แต่ออกมาทางปากทั้งหมด สำหรับสระนาสิกนั้นเพดานอ่อนจะลดต่ำลง และปล่อยให้อากาศผ่านออกทางช่องจมูกด้วยในเวลาเดียวกัน โดยในการเขียนสัทอักษรสากลจะใช้เครื่องหมาย ~ กำกับอยู่เหนือสระที่ออกเสียงแบบสระนาสิก เช่นในภาษาฝรั่งเศสจะมีหน่วยเสียงนาสิกอยู่ 4 หน่วยเสียงด้วยกัน แต่สำหรับภาษาไทยปกติแล้วไม่มีการออกเสียงสระนาสิก แต่ในบางครั้งก็อาจได้รับอิทธิพลจากการเปล่งเสียงพยัญชนะนาสิกที่อยู่ใกล้เคียง เช่น คำว่า นั้น เป็นต้น
5. **ความยาวในการออกเสียง** (Duration) ความสั้นยาวของการออกเสียงนั้นมีความสำคัญมากในภาษาไทย เพราะหน่วยเสียงที่ใช้ความยาวในการออกเสียงต่างกันจะทำให้ความหมายของพยางค์แตกต่างกันได้ เช่นคำว่า ขุด และคำว่า ขูด โดยสระจะถูกแบ่งออกเป็นสองประเภทตามความสั้นยาวของสระ คือ สระเสียงสั้น (รัสสระ) และสระเสียงยาว (ฑีฆสระ) โดยในการเขียนสัทอักษรสากลจะใช้สัญลักษณ์ ː (Length Mark) เพื่อแสดงว่าเสียงสระนั้นถูกยืดออกไป

1.3.2 เสียงสระภาษาไทย

สระในภาษาไทยตามไวยากรณ์ดั้งเดิม [8] มีทั้งหมด 21 รูป 32 หน่วยเสียง แบ่งเป็น 3 กลุ่ม คือ

1. สระเดี่ยว (Monophthongs) เป็นสระเสียงแท้ ซึ่งการออกเสียงสระตั้งแต่เริ่มต้นจนสิ้นสุดไม่มีการเปลี่ยนรูปร่างของลิ้นและช่องปาก สระเดี่ยวในภาษาไทยมีทั้งสิ้น 18 หน่วยเสียงเป็นสระเสียงสั้น 9 หน่วยเสียง และสระเสียงยาว 9 หน่วยเสียง
2. สระประสม (Diphthongs) เป็นสระที่เกิดจากการออกเสียงผสมกันของสระแท้ โดยลิ้นและช่องปากจะเปลี่ยนจากรูปร่างการออกเสียงของสระหนึ่งไปยังอีกสระหนึ่งอย่างค่อนข้างกลมกลืนและรวดเร็ว สระประสมในภาษาไทยมีทั้งสิ้น 6 หน่วยเสียง เป็นสระเสียงสั้น 3 หน่วยเสียง และสระเสียงยาว 3 หน่วยเสียง

3. สระเกิน (Vowel Letter) ในภาษาไทยมีรูปสระที่เกิดจากการรวมของเสียงสระกับตัวสะกดหรือคำควบเข้าไว้ด้วยกัน ซึ่งมีทั้งหมด 8 หน่วยเสียง เสียงสระในภาษาไทยสามารถสรุปได้ดังตารางที่ 2.2

ตารางที่ 2.2 ตารางเสียงสระภาษาไทย

หน่วยเสียง[1]	ส่วนของลิ้นที่ใช้เปล่งเสียง	ความสูงของลิ้น	การห่อริมฝีปาก	ความยาวเสียง	รูปสระ
สระเดี่ยว					
/i/	หน้า	ปิด	ไม่ห่อ	สั้น	อิ
/iː/	หน้า	ปิด	ไม่ห่อ	ยาว	อี
/e/	หน้า	กึ่งปิด	ไม่ห่อ	สั้น	เอะ
/eː/	หน้า	กึ่งปิด	ไม่ห่อ	ยาว	เอ
/æ/	หน้า	กึ่งเปิด	ไม่ห่อ	สั้น	แอะ
/æː/	หน้า	กึ่งเปิด	ไม่ห่อ	ยาว	แอ
/ɨ/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	สั้น	อึ
/ɨː/	หลัง ค่อนมาทางกลาง	ปิด	ไม่ห่อ	ยาว	อือ
/ɜ/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	สั้น	เออะ
/ɜː/	หลัง ค่อนมาทางกลาง	กึ่งปิด	ไม่ห่อ	ยาว	เออ
/a/	กลาง	เปิด	ไม่ห่อ	สั้น	อะ
/aː/	กลาง	เปิด	ไม่ห่อ	ยาว	อา
/u/	หลัง	ปิด	ห่อ	สั้น	อุ
/uː/	หลัง	ปิด	ห่อ	ยาว	อู
/o/	หลัง	กึ่งปิด	ห่อ	สั้น	โอะ
/oː/	หลัง	กึ่งปิด	ห่อ	ยาว	โอ
/ɔ/	หลัง	กึ่งเปิด	ห่อ	สั้น	เอาะ
/ɔː/	หลัง	กึ่งเปิด	ห่อ	ยาว	ออ
สระประสม	ส่วนประกอบ				
/ia/	/i/ + /a/			สั้น	เอียะ
/iːa/	/iː/ + /a/			ยาว	เอีย
/ɨa/	/ɨ/ + /a/			สั้น	เอือะ
/ɨːa/	/ɨː/ + /a/			ยาว	เอือ
/ua/	/u/ + /a/			สั้น	อัวะ
/uːa/	/uː/ + /a/			ยาว	อัว
สระเกิน	ส่วนประกอบ				
/am/	/a/ + /m/			สั้น	อำ
/aj/	/a/ + /j/			สั้น	ไอ ใอ
/aw/	/a/ + /w/			สั้น	เอา
/ri/, /rɨ/	/r/ + /i/, /r/ + /ɨ/			สั้น	ฤ
/riː/, /rɨː/	/r/ + /iː/, /r/ + /ɨː/			ยาว	ฤา
/li/, /lɨ/	/l/ + /i/, /l/ + /ɨ/			สั้น	ฦ
/liː/, /lɨː/	/l/ + /iː/, /l/ + /ɨː/			ยาว	ฦา

[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

1.4 เสียงวรรณยุกต์ในภาษาไทย (Thai Tones) [5]

เสียงวรรณยุกต์นั้นคือเสียงสูงต่ำในภาษา ซึ่งเกิดจากการสั่นสะเทือนของเส้นเสียงในอัตราความถี่ที่ต่างกันไป ดังนั้นเสียงวรรณยุกต์จะปรากฏอยู่ในส่วนของเสียงสระ เพราะเสียงสระเป็นเสียงที่เกิดจากการสั่นของเส้นสียง นอกจากนี้ยังอาจมีเสียงวรรณยุกต์ปรากฏอยู่บ้างในบางส่วนของเสียงพยัญชนะ แต่จะต้องเป็นส่วนของเสียงพยัญชนะที่เป็นเสียงก้องหรือพยัญชนะนาสิกเท่านั้น เพราะเสียงพยัญชนะไม่ก้องนั้นไม่ได้เกิดจากการสั่นของเส้นเสียง จึงไม่สามารถมีเสียงวรรณยุกต์อยู่ด้วยได้

สำหรับในภาษาไทย วรรณยุกต์นั้นถือได้ว่าเป็นหน่วยเสียงที่สำคัญ เพราะสามารถใช้แยกแยะความแตกต่างทางความหมายของคำในภาษาไทยได้ ตรงกันข้ามกับบางภาษา เช่น ภาษาอังกฤษ ซึ่งไม่จัดว่าเสียงวรรณยุกต์เป็นหน่วยเสียงในภาษา เพราะไม่ว่าเราจะพูดภาษาอังกฤษด้วยเสียงสูงต่ำอย่างไร ก็ไม่ทำให้ความหมายของคำเปลี่ยนไป แต่ก็อาจจะต้องมีการใช้เสียงวรรณยุกต์ประกอบบ้าง ทั้งนี้เพื่อใช้เน้นให้ความหมายโดยรวมของประโยคเปลี่ยนไป ภาษาไทยจึงจัดได้ว่าเป็นภาษามีวรรณยุกต์ (Tonal Language) เสียงวรรณยุกต์ภาษาไทยสามารถแบ่งออกเป็น 2 ชนิดใหญ่ๆ คือ

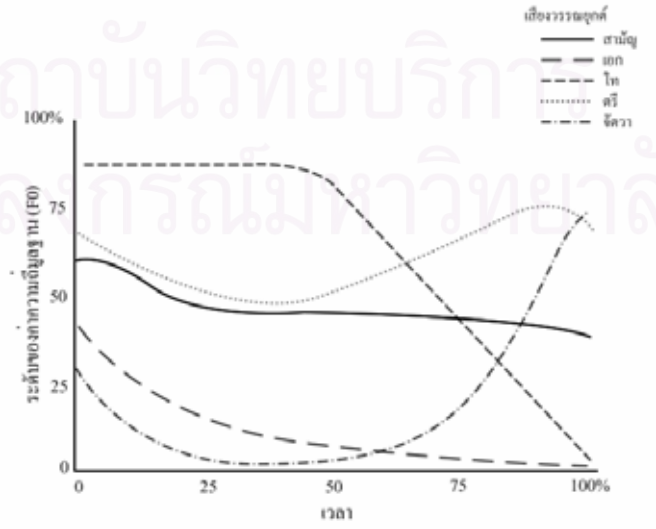
1. เสียงวรรณยุกต์ระดับ (Level Tone) เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ค่อนข้างคงที่ตลอดพยางค์ ถึงแม้ว่าในการออกเสียงพูดโดยปกตินั้น เสียงต้นพยางค์มักจะไม่ได้มีความถี่และความดังเท่ากันกับเสียงท้ายพยางค์ โดยเสียงต้นพยางค์มักมีมีระดับความถี่สูงกว่าและดังกว่าเสียงท้ายพยางค์ แต่ในทางสัทศาสตร์แล้ว ความถี่ที่ต่างกันหรือการเปลี่ยนแปลงของระดับเสียงนี้ถือว่าเล็กน้อยมาก เมื่อเทียบกับการเปลี่ยนระดับความถี่ของเสียงในพยางค์อีกจำพวกหนึ่งซึ่งจะได้กล่าวต่อไป สำหรับเสียงวรรณยุกต์ระดับในภาษาไทยนั้น มีอยู่ด้วยกัน 3 เสียงดังนี้คือ
    1. เสียงวรรณยุกต์สามัญ (Mid Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ปานกลางประมาณ 120 เฮิรตซ์ และคงที่อยู่ที่ระดับนั้นจนกระทั่งปลายพยางค์ จึงจะลดต่ำลงมาจนเกือบถึงประมาณ 110 เฮิรตซ์ เสียงวรรณยุกต์สามัญนี้จะไม่ปรากฏในพยางค์ที่มีพยัญชนะกักเป็นพยัญชนะท้าย หรือที่เรียกกันว่าคำตาย
    2. เสียงวรรณยุกต์เอก (Low Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ต้นเสียงปานกลางประมาณ 120 เฮิรตซ์ แล้วลดต่ำลงมาเหลือประมาณ 100 เฮิรตซ์อย่างรวดเร็ว และคงที่อยู่ในระดับนี้ สำหรับเสียงวรรณยุกต์เอกจะปรากฏกับพยางค์ได้ทุกรูปแบบทั้งคำเป็นและคำตาย
    3. เสียงวรรณยุกต์ตรี (High Tone) เสียงวรรณยุกต์นี้มีระดับความถี่ค่อนข้างสูง โดยจะค่อยๆ สูงขึ้นทีละน้อยจากต้นพยางค์ซึ่งมีความถี่ประมาณ 125 เฮิรตซ์ ไปจนถึง

ประมาณ 135 – 140 เฮิรตซ์เมื่อสิ้นพยางค์ หรืออาจจะลดต่ำลงตอนปลายพยางค์มาอยู่ที่ประมาณ 130 เฮิรตซ์ก็ได้ ขึ้นอยู่กับว่าพยางค์นั้นๆ จบลงด้วยเสียงประเภทใด ถ้าพยางค์นั้นคำเป็น ระดับของเสียงตอนปลายของพยางค์จะไม่ลดต่ำลงมา แต่ถ้าพยางค์นั้นเป็นคำตาย ระดับเสียงตอนปลายจะลดต่ำลงอย่างรวดเร็ว

2. เสียงวรรณยุกต์เปลี่ยนระดับ (Contour Tone) เป็นเสียงวรรณยุกต์ที่มีระดับความถี่ของการออกเสียงเปลี่ยนแปลงมากในช่วงพยางค์หนึ่งๆ เช่น ต้นพยางค์ออกเสียงให้มีระดับสูงแล้วลดระดับเสียงลงอย่างรวดเร็วไปสู่ระดับต่ำที่ท้ายพยางค์ หรือต้นพยางค์ออกเสียงให้มีระดับต่ำ แล้วเพิ่มระดับเสียงอย่างรวดเร็วไปเป็นระดับสูงที่ท้ายพยางค์ นอกจากนี้ ยังอาจจะเกิดจากการเปลี่ยนระดับเสียงจากสูงแล้วไปต่ำแล้วไปสูงอีก หรือเปลี่ยนจากต่ำแล้วไปสูงแล้วไปต่ำอีกก็ได้ สำหรับในภาษาไทยนั้นมีเสียวรรณยุกต์เปลี่ยนระดับอยู่ 2 เสียงดังนี้

   1. เสียงวรรณยุกต์โท (Falling Tone) ระดับเสียงจะเริ่มต้นที่ระดับความถี่ประมาณ 140 เฮิรตซ์ แต่เมื่อถึงประมาณ 1 ใน 4 ของความยาวช่วงพยางค์ ระดับความถี่จะเริ่มลดลงเรื่อยๆ จนต่ำกว่า 100 เฮิรตซ์ที่ปลายพยางค์ หรืออาจจะมีการเปลี่ยนระดับความถี่สูงขึ้นจากต้นพยางค์เล็กน้อยก่อนที่จะลดระดับเสียงลงอย่างรวดเร็วก็ได้ เสียงวรรณยุกต์โทนี้จะไม่ปรากฏในคำตาย ยกเว้นในคำเลียนเสียงธรรมชาติหรือคำลงท้ายประโยคบางคำ เช่น "พลัก" หรือ "ละ" เป็นต้น
   2. เสียงวรรณยุกต์จัตวา (Rising Tone) ระดับเสียงจะเริ่มที่ระดับความถี่ประมาณ 110 เฮิรตซ์ แล้วมักจะลดลงเล็กน้อยก่อนจะเพิ่มความถี่ขึ้นอย่างรวดเร็วจนสูงถึงประมาณ 140 เฮิรตซ์ที่ท้ายพยางค์ เสียงวรรณยุกต์จัตวานี้จะไม่ปรากฏที่คำตาย

การเปลี่ยนแปลงความถี่ของเสียงในวรรณยุกต์ภาษาไทยสามารถแสดงได้ดังรูปที่ 2.3



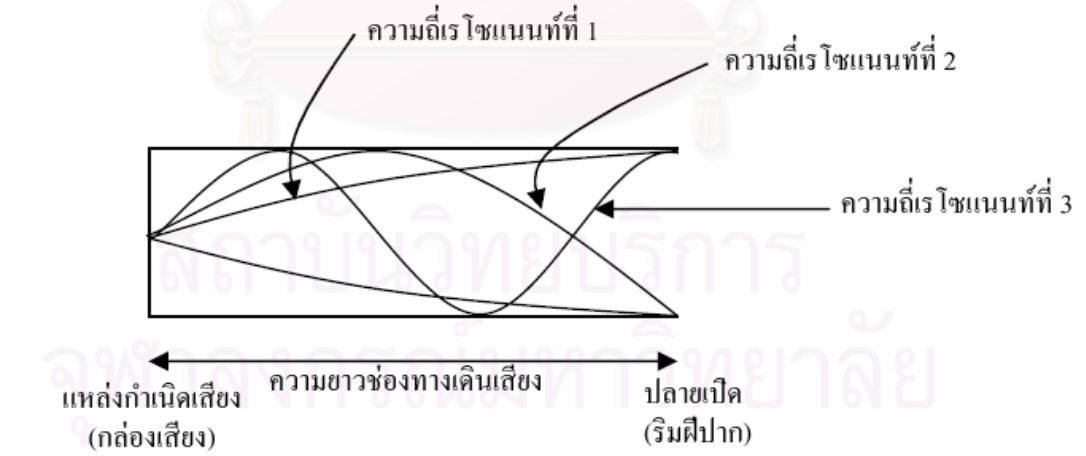
รูปที่ 2.3 การเปลี่ยนแปลงความถี่ของเสียงในววรณยุกต์ภาษาไทย

2. สวนสัทศาสตร์

2.1 กระบวนการสร้างเสียงพูด (Speech Production)

ระบบเสียงพูดสามารถพิจารณาได้ว่าประกอบไปด้วยลำดับของท่อ และช่องที่ต่อออกมาจากปอดไปยังปากและจมูก ท่อและช่องนี้จะมีความยาวโดยประมาณ 7 นิ้ว เส้นเสียงจะอยู่ในตำแหน่งปลายที่ตรงข้ามกับขั้วปอด ทำหน้าที่ควบคุมการไหลของลมให้ผ่านปอดเข้าสู่ช่องทางเดินเสียง (Vocal tract) ภายใต้การควบคุมของกล้ามเนื้อ ส่วนประกอบของช่องทางเดินเสียงที่มีลักษณะเป็นท่อจะสามารถเปลี่ยนรูปร่างได้ในอัตราถึง 10 ครั้งต่อวินาที ส่วนเส้นเสียงนั้นจะสามารถเปิด-ปิดด้วยอัตราเร็วประมาณ 100 - 300 ครั้งต่อวินาที การเปลี่ยนแปลงรูปร่างของช่องทางเดินเสียงและรูปร่างและตำแหน่งของสื่อกลางที่ทำให้เกิดเสียงดังกล่าวนี้รวมเรียกว่ากระบวนการทำให้เกิดเสียง

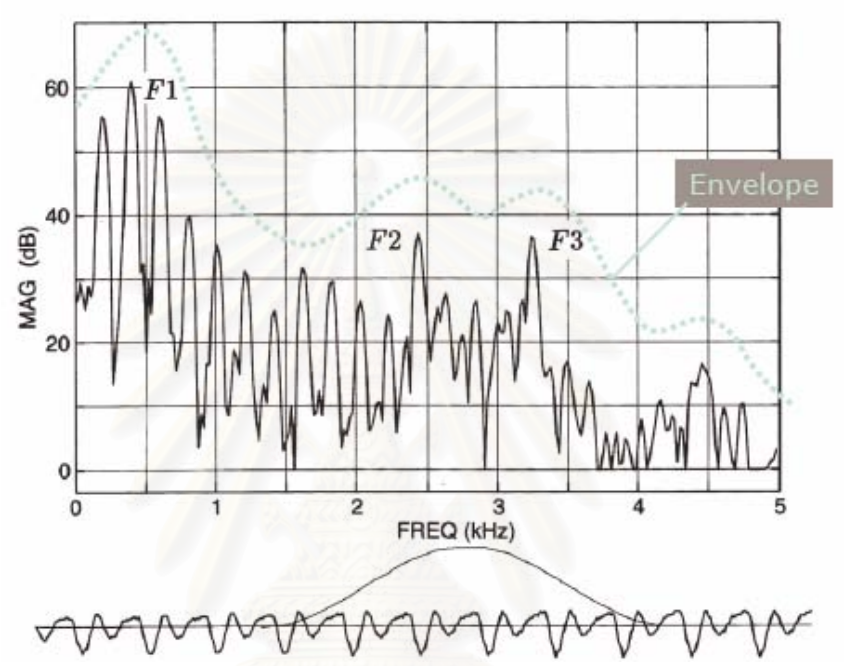
รูปแบบจำลองอย่างง่ายของช่องทางเดินเสียง ก็อาจมองได้เป็นลักษณะของท่อทรงกระบอกที่มีต้นกำเนิดเสียงอยู่ที่ปลายข้างหนึ่ง (ส่วนของกล่องเสียง) และปลายอีกข้างหนึ่งจะเปิด (ส่วนของปาก) ดังรูปที่ 2.4 ดังนั้นมันจะเกิดเรโซแนนท์ภายในท่อได้ที่ความยาวเท่ากับ 4L, 4L/3, 4L/5, ... เฮิร์ต โดยที่ L คือความยาวท่อ ถ้าคิดเป็นความถี่ที่เกิดเรโซแนนท์จะได้ความถี่ที่ c/4L, 3c/4L, 5c/4L, ... เฮิร์ต โดยที่ c คือ ค่าความเร็วของเสียงในอากาศ และถ้าจะคำนวณหาความถี่ในการเรโซแนนท์ของช่องทางเดินเสียงของคน ซึ่งปกติช่องทางเดินเสียงของคนเราจะมีความยาวประมาณ 7 นิ้ว หรือ 17 เซนติเมตร และ c มีค่าเท่ากับ 340 เมตรต่อวินาที ดังนั้นจึงมีเรโซแนนท์ที่ความถี่ประมาณ 500 เฮิรต์, 1500 เฮิรต์, 2500 เฮิรต์, ... เป็นต้น



รูปที่ 2.4 การเกิดเรโซแนนท์ภายในแบบจำลองของช่องทางเดินเสียง

ความถี่เรโซแนนท์ต่างๆ จะเปลี่ยนไปเมื่อพื้นที่หน้าตัดของท่อในบริเวณต่างๆ เปลี่ยนแปลงไป การเปลี่ยนแปลงพื้นที่หน้าตัดของท่อตามเวลานี้ เป็นการจำลองรูปร่างที่เปลี่ยนไปของช่องทางเสียง ซึ่งมีสาเหตุจากการเคลื่อนไหวของอวัยวะส่วนการกระทำอาการ

เสียงพูดของมนุษย์เกิดจากต้นกำเนิดเสียง ที่อาจเป็นสัญญาณกึ่งคาบ (Quasi-periodic Signal) เช่น ต้นกำเนิดที่เกิดจากการสั่นของเส้นเสียง หรือสัญญาณที่มีลักษณะเหมือนสัญญาณรบกวน (Noise) เมื่อต้นกำเนิดเป็นสัญญาณกึ่งคาบ ช่องทางเสียงที่ถูกกระตุ้นด้วยต้นกำเนิดเสียง จะสร้างคลื่นเสียงที่มีองค์ประกอบทางความถี่สูงเด่นขึ้นมาในบริเวณของความถี่เรโซแนนท์ ในขณะนั้นความถี่เรโซแนนท์ในกรณีนี้ถูกเรียกว่า ความถี่ฟอร์แมนต์ (Formant Frequency)



รูปที่ 2.5 สเปกตรัมของพลังงานเสียง [7]

ฟอร์แมนท์ที่มีความถี่ต่ำที่สุดจะเรียกว่าฟอร์แมนท์ที่หนึ่ง ซึ่งจะมีค่าประมาณ 200 – 1200 เฮิรต์ ทั้งนี้ขึ้นอยู่กับขนาดของช่องทางเดินเสียงด้วย ส่วนฟอร์แมนท์ที่สองที่อยู่ถัดไปก็จะมีค่าประมาณ 500 – 2500 เฮิรต์ และฟอร์แมนท์ที่สามจะมีค่าประมาณ 1500 – 3500 เฮิรต์ เป็นต้น ในทำนองเดียวกัน เมื่อต้นกำเนิดเสียงเป็นลักษณะสัญญาณรบกวน เสียงที่ถูกสร้างจากช่องทางเสียงจะมีลักษณะเป็นสัญญาณรบกวนเช่นเดียวกับต้นกำเนิด แต่ขนาดของสเปกตรัมที่ความถี่เรโซแนนท์จะถูกขยายขึ้นอย่างมาก รูปร่างของสเปกตรัมของเสียงนี้สามารถนำมาวิเคราะห์เพื่อประมาณเหตุการณ์ที่เกิดขึ้นในช่องทางเสียงของผู้พูดได้

2.2 สัทลักษณ์ (Phonetic Features)

สัทลักษณ์โดยทั่วไปที่ใช้ในการวิเคราะห์และอธิบายคุณสมบัติของแต่ละประเภทหน่วยเสียงนั้นแบ่งออกเป็น 3 ประเภทคือคุณสมบัติความก้องของเสียงตามลักษณะของแหล่งกำเนิดเสียง (Source Characteristics), ลักษณะการออกเสียง (Manner of Articulation) และ ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง (Place of Articulation) Chomsky และ Halle [9] ได้ศึกษาและให้นิยามของ สัทลักษณ์โดยกำหนดให้แต่ละสัทลักษณ์มีค่าเป็น + หรือ – เท่านั้น

2.2.1 คุณสมบัติความก้องของเสียง

เสียงพูดของมนุษย์เมื่ออากาศถูกดันออกมาจากปอดด้วยแรงดันที่มากจะทำให้เส้นเสียงสั่นทำให้สัญญาณที่เกิดจากแหล่งกำเนิดเสียงเป็นลักษณะเป็นคาบจะแสดงได้ด้วยค่าคุณสมบัติความเป็นเสียงก้อง (Voiced) แทนด้วยสัญลักษณ์ ‘+’ หรือ [+voiced] ส่วนสัญญาณเสียงที่เกิดจากโดยไม่มีการสั่นของเส้นเสียงและมีลักษณะไม่เป็นคาบจะแสดงได้ด้วยค่าคุณสมบัติความเป็นเสียงไม่ก้อง (Unvoiced) แทนด้วยสัญลักษณ์ ‘–‘หรือ [–voiced] ทั้งสัญญาณเสียงที่ก้องและไม่ก้องอาจปรากฏอยู่ในเสียงที่มนุษย์ได้ยินเนื่องจากมีการเปลี่ยนแปลงของสัญญาณเสียงในช่องทางเดินเสียง

2.2.2 ลักษณะการออกเสียง

ลักษณะการออกเสียงพิจารณาจากลักษณะของช่องทางเดินเสียงว่ามีการเปิด/ปิดอย่างไร มีการกักเสียงไว้มากน้อยแค่ไหน และการออกเสียงนั้นอากาศไหลผ่านบริเวณช่องปากหรือผ่านไปในช่องโพรงจมูกบ้างหรือไม่ เสียงที่ผ่านช่องปากไปโดยไม่ได้ถูกกักเอาไว้เพียงพอต่อการสร้างเสียงรบกวนหรือกักการไหลของอากาศจะเรียกว่าเสียงที่มีเสียงสั่น (Sonorant) ซึ่งประกอบไปด้วยเสียงสระ (Vowels) เสียงกึ่งสระ (Semi-vowels) และเสียงนาสิก (Nasals) คุณสมบัติความเป็นเสียงที่มีเสียงสั่นแสดงได้ด้วยสัญลักษณ์ [+sonorant] ส่วนเสียงเสียงที่ไม่มีคุณสมบัติความเป็นความเป็นเสียงที่มีเสียงสั่น ได้แก่ เสียงพยัญชนะกัก (Stop Consonants) และเสียงเสียดแทรก (Fricatives) จะแสดงได้ด้วยสัญลักษณ์ [–sonorant] เสียงที่มีและไม่มีคุณสมบัติเป็นความเป็นเสียงที่มีเสียงสั่นสามารถแยกต่อไปได้อีก ตารางที่ 2.3 แสดงความสัมพันธ์ระหว่างการแบ่งประเภทของหน่วยเสียงกลุ่มใหญ่ๆ (เสียงสระ, เสียงพยัญชนะกัก, เสียงเสียดแทรก และ เสียงนาสิกและเสียงกึ่งสระ) กับคุณสมบัติของลักษณะการออกเสียงในแต่ละกลุ่ม คุณสมบัติความเป็นเสียงที่เป็นเสียงพยางค์ (Syllabic) คือเสียงที่เกิดจากการช่องทางเดินเสียงเปิดอยู่แทนด้วยสัญลักษณ์ [+syllabic] ซึ่งประกอบด้วยพวกเสียงสระ ส่วนเสียงที่เกิดขึ้นโดยที่ช่องทางเดินเสียงปิดอยู่จะแทนด้วยสัญลักษณ์

[–syllabic] ซึ่งประกอบไปด้วยเสียงกึ่งสระและเสียงนาสิก ส่วนเสียงที่มีคุณสมบัติความเป็นเสียงที่มีการกักเสียงเอาไว้โดยที่ช่องทางเดินเสียงอยู่ในสภาพปิดอยู่แต่ยังปิดไม่สมบูรณ์ (Continuant) ตัวอย่างเช่น เสียง /ส/ ที่ใช้ฟันในการกักไม่ให้อากาศไหลผ่านช่องปากได้อย่างเต็มที่ทำให้เกิดเป็นเสียงเสียดแทรก แสดงได้ด้วยสัญลักษณ์ [+continuant] ส่วนเสียงพยัญชนะกักแสดงได้ด้วยสัญลักษณ์ [–continuant]

ตารางที่ 2.3 ตารางแสดงความสัมพันธ์ระหว่างประเภทของหน่วยเสียง
เปรียบเทียบกับสัทลักษณ์ของลักษณะการออกเสียง

สัทลักษณ์	[-continuant]	[+continuant]
[-sonorant]	เสียงพยัญชนะกัก	เสียงเสียดแทรก
[+sonorant, -syllabic]	เสียงนาสิกและเสียงกึ่งสระ	
[+sonorant, +syllabic]	เสียงสระ	

2.2.3 ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียง

ตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงในกรณีของเสียงพยัญชนะกักและเสียงพยัญชนะเสียดแทรกจะพิจารณาจากตำแหน่งที่มีใช้อวัยวะเช่นฟัน ลิ้น หรือริมฝีปากในการกั้นไม่ให้อากาศไหลผ่านช่องปากออกไปได้โดยตรง ในกรณีของเสียงสระจะพิจารณาจากตำแหน่งของลิ้นในขณะที่อากาศเดินทางผ่านช่องปากออกไป ดังจำแนกด้วยคุณสมบัติดังตารางที่ 2.4

ตารางที่ 2.4 ตารางแสดงความสัมพันธ์ระหว่างสัทลักษณ์ของตำแหน่งของอวัยวะที่เป็นฐานในการออกเสียงเปรียบเทียบกับเสียงพยัญชนะกักในภาษาไทย

สัทลักษณ์	อวัยวะที่มีความสัมพันธ์กับการออกเสียง	/บ/ /พ/	/ด/ /ท/	/ค/ /ก/
Velar	เกิดการกักเสียงระหว่างตัวลิ้นกับเพดานอ่อน	–	–	+
Alveolar	เกิดการกักเสียงปลายลิ้นกับเพดานส่วนหน้า	–	+	–
Labial	เกิดการกัดเสียงบริเวณริมฝีปาก	+	–	–

3. การรู้จำเสียงพูด

การรู้จำเสียงพูด (Speech Recognition) เป็นกระบวนการสกัดลำดับของคำ (Sequence of Words) ที่อยู่ในสัญญาณเสียงออกมา โดยในเชิงทฤษฎีสารสนเทศ (Information Theory) เราพิจารณาเสียงพูดที่ได้ยินว่าเป็นสัญญาณที่ถูกรบกวนผ่านทางช่องสัญญาณรบกวน (Noisy Channel) และการรู้จำเสียงพูดคือการถอดรหัส (Decode) สัญญาณนั้น

3.1 ปัญหาการรู้จำเสียงพูด

ปัญหาการรู้จำเสียงพูดหรือการถอดรหัสของสัญญาณเสียงพูด สามารถมองได้ว่าเป็นการหาลำดับของคำที่ดีที่สุดสำหรับลำดับของข้อมูลทางเสียงที่สังเกตได้ (Observation Sequence) โดยกำหนดให้ สัญลักษณ์ของลำดับข้อมูลที่สังเกตได้เป็น $O = o_1 o_2 o_3 ... o_t$ สัญลักษณ์ของลำดับของคำเป็น $W = w_1 w_2 w_3 ... w_n$ และ L แทนภาษาที่พิจารณา

ให้ W^* เป็นลำดับของคำที่ดีที่สุด เพราะฉะนั้นเราจะได้ว่า

$$W^* = \arg\max_{W \in L} P(W \mid O)$$

ซึ่งในทางปฏิบัติแล้ว การหาค่า $P(W \mid O)$ โดยตรงทำได้ยาก จึงเขียนสูตรข้างต้นใหม่โดยใช้กฎของเบย์ได้ดังนี้

$$\begin{aligned} W^* &= \arg\max_{W \in L} \frac{P(O \mid W)P(W)}{P(O)} \\ &= \arg\max_{W \in L} P(O \mid W)P(W) \end{aligned}$$

ในที่นี้ $P(W)$ คือความน่าจะเป็นที่ลำดับของคำ W จะเกิดขึ้นในภาษา จึงเรียก $P(W \mid O)$ ว่าน่าจะเป็นก่อนหน้า (Prior) หรือแบบจำลองทางภาษา (Language Model) ส่วน $P(O \mid W)$ เป็นความน่าจะเป็นที่ลำดับของข้อมูลที่สังเกตได้เป็น O เมื่อลำดับของคำที่เป็นที่มาของ O คือ W และเรียก $P(O \mid W)$ ว่าความเป็นไปได้ (Likelihood) หรือแบบจำลองทางเสียง (Acoustic Model)

3.2 ประเภทของระบบรู้จำเสียงพูด

เราสามารถแบ่งระบบรู้จำเสียงพูดตามเกณฑ์ต่างๆ ซึ่งเป็นส่งผลต่อความยากง่ายในการพัฒนา และประสิทธิภาพของระบบรู้จำเสียงพูด ได้ดังต่อไปนี้

1. **จำนวนคำศัพท์** (Vocabulary Size)

ระบบที่รองรับจำนวนคำศัพท์น้อย เช่น รู้จำเสียงพูดของตัวเลข ศูนย์ ถึง เก้า สามารถพัฒนาได้ง่ายกว่าและให้ความผิดพลาดน้อยกว่าระบบที่ต้องรองรับจำนวนคำศัพท์มาก เช่น รู้จำเสียงพูดของทุกคำในภาษา ซึ่งอาจมีมากถึง 20,000 คำ

2. **ความขึ้นต่อผู้พูด** (Speaker Dependency)
    1. การรู้จำเสียงพูดแบบคำโดดเดี่ยว (Isolated Word Recognition) จะพิจารณาหนึ่งคำต่อการเปล่งเสียงหนึ่งครั้ง โดยระบบจะพยายามตัดสินใจให้รู้จำคำที่ใกล้เคียงที่สุด
    2. การรู้จำเสียงพูดแบบเสียงพูดต่อเนื่อง (Continuous Speech Recognition) จะพิจารณาเป็นวลี หรือประโยค ซึ่งการรู้จำแบบนี้ระบบจะแบ่งการทำงานออกเป็น 2 ขั้นตอน คือขั้นตอนของการแบ่งเสียงที่เข้ามาออกเป็นคำย่อยๆ และขั้นตอนของการรู้จำประโยคโดยอาศัยบริบท เพื่อเข้าใจความหมายของประโยค นอกจากนี้ต้องพิจารณาการเชื่อมเสียงระหว่างคำต่างๆ ในประโยค โดยใช้ข้อมูลทางไวยกรณ์ของภาษาที่ซับซ้อน
3. **ลักษณะการพูด** (Speaking Style)
    1. ขึ้นกับผู้พูด (Speaker-Dependent) จะทำให้แบบจำลองที่เราใช้อ้างอิงต้องเปลี่ยนแปลงตามผู้พูด
    2. ไม่ขึ้นกับผู้พูด (Speaker-Independent) ระบบจะสามารถรู้จำคำได้ไม่ว่าอินพุทของเสียงที่เข้ามาในระบบจะเป็นของผู้พูดคนไหน แต่ระบบแบบนี้ระบบของเราจะต้องมีความซับซ้อนอย่างมาก เพราะระบบของเราจะต้องสามารถรองรับสถานการณ์ทุกรูปแบบของเสียงที่จะเป็นอินพุทให้ได้
4. **ระดับของหน่วยเสียง** (Sound Unit)
    1. หน่วยเสียงระดับคำ (Word-Based) เป็นการรู้จำคำพูดโดยใช้หน่วยของ “คำ” เป็นหน่วยย่อยที่สุดในการรู้จำคำพูด โดยจะใช้โมเดลแต่ละโมเดลเพื่อรู้จำคำแต่ละคำ และไม่มีการใช้โมเดลร่วมกันระหว่างคำ ซึ่งจะทำให้มีการใช้พื้นที่หน่วยความจำสูงมาก และจะประสบปัญหาในการพิจารณาการรู้จำคำในประโยคที่เราต้องพิจารณาการเชื่อมเสียงระหว่างคำเป็นบริบท
    2. หน่วยเสียงระดับเล็กกว่าคำ (Sub-Word Based) เป็นการรู้จำคำพูดโดยใช้หน่วยเสียงที่เป็นหน่วยที่เล็กกว่า “คำ” ในการรู้จำคำพูด โดยจะใช้โมเดลแต่ละโมเดลเพื่อรู้จำแต่ละหน่วยที่เล็กกว่าคำ ซึ่งเราสามารถใช้โมเดลร่วมกันหลายๆ โมเดลได้เพื่อรู้จำคำ โดยแต่ละหน่วยที่เล็กกว่าคำ จะต้องพึ่งบริบทในการพิจารณาการรู้จำคำศัพท์ เพื่อพิจารณาการเชื่อมเสียง นอกจากนี้วิธีนี้ยังใช้พื้นที่ในหน่วยความจำไม่สูงนัก

4. การแปลงฟูริเยร์แบบวิยุต

ในกระบวนการการวิเคราะห์เสียงพูดในโดเมนความถี่นั้น เราจะใช้การแปลงฟูริเยร์แบบวิยุตกับเสียงพูดที่ได้รับเข้ามาให้อยู่ในรูปของเวกเตอร์คุณสมบัติของเสียงเพื่อนำไปวิเคราะห์เปรียบเทียบกับเสียงตัวอย่าง โดยเสียงตัวอย่างที่ถูกส่งเข้ามาในระบบนี้จะเป็นสัญญาณที่อยู่ในโดเมนเวลาซึ่งเป็นฟังก์ชันคาบ และเมื่อได้รับการแปลงฟูริเยร์แบบวิยุตแล้วจะทำให้สัญญาณถูกเปลี่ยนไปอยู่ในโดเมนความถี่ แต่เนื่องจากระบบคอมพิวเตอร์จะจัดเก็บสัญญาณดังกล่าวขึ้นอยู่กับอัตราการสุ่ม (Sampling Rate) ดังนั้นเราจึงสามารถทำการวิเคราะห์สัญญาณในโดเมนความถี่ดังกล่าว ให้อยู่ในรูปแบบฟังก์ชันคาบอีกครั้งหนึ่ง

การแปลงฟูริเยร์แบบต่อเนื่อง (Continuous Fourier Transform) ได้นิยามไว้ ดังนี้

$$f(v) = \mathbf{F}[f(t)](v) = \int_{-\infty}^{+\infty} f(t)e^{-2\pi ivt}dt$$

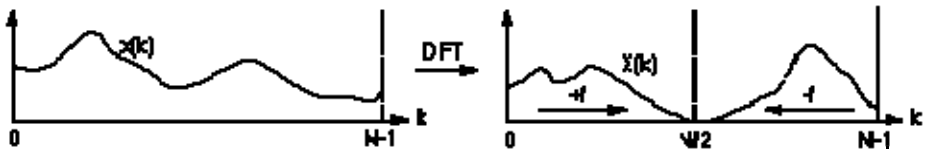
จากนิยามของการแปลงฟูริเยร์แบบต่อเนื่อง เราจะทำการพิจารณาในกรณีที่เป็นฟังก์ชันไม่ต่อเนื่อง $f(t) \to f(t_k)$ โดยให้ $f_k \equiv f(t_k)$ และ $t_k \equiv k\triangle$ เมื่อให้ $k = 0,1,2,...,N-1$ จะได้การแปลงฟูริเยร์แบบวิยุตดังนี้ $F_n = \mathbf{F}_k[\{f_k\}_{k=0}^{N-1}](n)$

$$F_n = \sum_{k=0}^{N-1} f_k e^{-2\pi ink/N}$$

การแปลงฟูริเยร์แบบวิยุต จะทำให้เห็นถึงช่วงคาบและความสัมพันธ์ของแต่ละช่วงสัญญาณที่เข้ามาอย่างชัดเจน ซึ่งการแปลงฟูริเยร์แบบวิยุตของลำดับจำนวนจริงจะเท่ากับการแปลงฟูริเยร์แบบวิยุตของลำดับจำนวนเชิงซ้อนที่มีจำนวนพจน์เท่ากัน โดยสรุปแล้ว ถ้า f_k เป็นจำนวนจริงแล้ว F_{N-n} และ F_n จะมีความสัมพันธ์ดังนี้

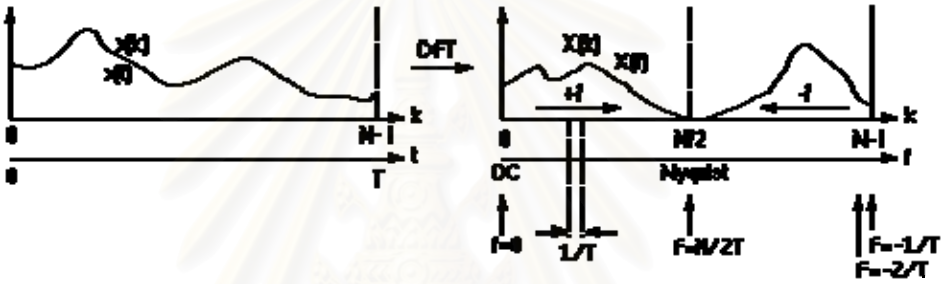
$$F_{N-n} = \bar{F}_n$$

สำหรับ $n = 0,1,2,...,N-1$ และ $\bar{z}$ แทนจำนวนเชิงซ้อนผัน (Complex Conjugate) ซึ่งหมายความว่าส่วน F_0 จะเป็นจำนวนจริงเสมอ และจากความสัมพันธ์ดังกล่าว ฟังก์ชันคาบ (Periodic Function) จะประกอบด้วยจุดสูงสุดของสัญญาณ 2 จุดด้วยกัน ซึ่งหมายความว่าช่วงคาบของสัญญาณที่เข้ามาเป็นสองส่วน คือช่วงความถี่บวก (Positive Frequency) และช่วงความถี่ลบ (Negative Frequency) ซึ่งเป็นช่วงความถี่ที่เป็นจำนวนเชิงซ้อน



รูปที่ 2.6 การแปลงฟูริเยร์แบบวิยุต

กราฟในรูปที่ 2.6 แสดงความสัมพันธ์ระหว่างหน่วยของอนุกรมและหน่วยของความถี่ (กราฟดังกล่าวเป็นเพียงการจำลองภาพขึ้นเท่านั้น) ซึ่งในปกติแล้วการแปลงฟูริเยร์แบบวิยุตจะมีข้อมูลเป็นอนุกรมของจำนวนเชิงซ้อน ดังตัวอย่าง ถ้าอนุกรมแทนลำดับของเวลาที่มีความยาว T แล้ว จะสามารถแสดงค่าความถี่ในรูปที่ 2.7 ได้ดังนี้



รูปที่ 2.7 การแปลงฟูริเยร์แบบวิยุต (เมื่ออนุกรมแทนลำดับของเวลามีความยาว T)

การแปลงฟูริเยร์แบบเร็ว (Fast Fourier Transform - FFT) เป็นอัลกอริทึมของการแปลงฟูริเยร์แบบวิยุต ซึ่งสามารถลดเวลาในการคำนวณได้สำหรับกลุ่มตัวอย่างสัญญาณจำนวน N จุดจาก $2N^2$ ให้เหลือเพียง $2N\log N$ เท่านั้น อัลกอริทึมของการแปลงฟูริเยร์แบบเร็วสามารถนั้นแยกออกเป็นสองคลาส คือ การลดจำนวนของหน่วยเวลา และการลดจำนวนของหน่วยความถี่

ในการลดจำนวนของหน่วยเวลา โดยการใช้การแปลงฟูริเยร์แบบเร็วด้วยอัลกอริทึมของคูเลย์-เตอกีย์ (Cooley-Turkey) จะทำการจัดเรียงข้อมูลที่เข้ามาให้อยู่ในรูปของบิตที่เรียงลำดับถอยหลัง (Bit-reversed Order) แล้วทำการคำนวณผลลัพธ์ออกมาด้วยวิธีนี้จะเป็นการแบ่งการแปลงฟูริเยร์ขนาดความยาว N ให้ออกเป็นการแปลงฟูริเยร์สองส่วนที่มีความยาวเป็น $N/2$

$$\begin{aligned}\sum_{n=0}^{N-1} a_n e^{-2\pi ink/N} &= \sum_{n=0}^{N/2-1} a_{2n} e^{-2\pi i(2n)k/N} + \sum_{n=0}^{N/2-1} a_{2n+1} e^{-2\pi i(2n+1)k/N} \\ &= \sum_{n=0}^{N/2-1} a_n^{even} e^{-2\pi ink/(N/2)} + \sum_{n=0}^{N/2-1} a_n^{odd} e^{-2\pi ink/(N/2)}\end{aligned}$$

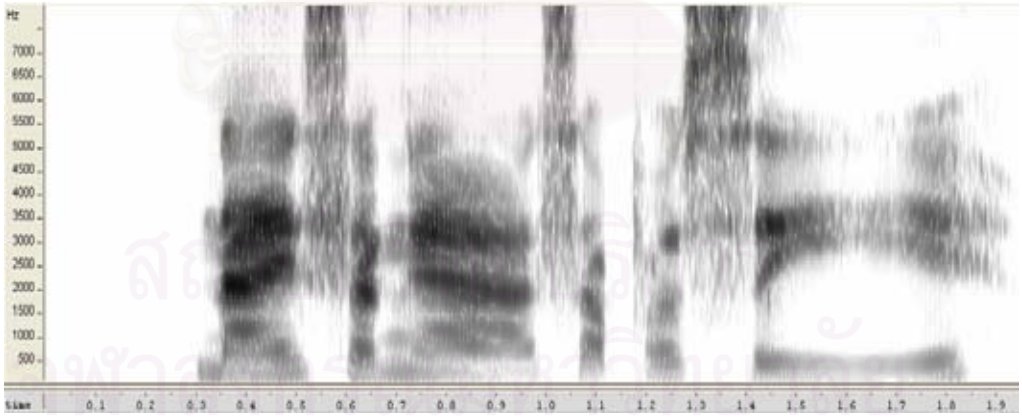
5. สเปกโตรแกรมของเสียงพูด

สัญญาณเสียงมีลักษณะรูปร่างเป็นคลื่นที่สั่นแกว่งไปมา เราไม่สามารถจะอ่านหน่วยเสียงในรูปแบบของคลื่นในโดเมนเวลา แต่ถ้าเราวิเคราะห์รูปแบบคลื่นในโดเมนความถี่ เราจะได้สเปกโตรแกรมซึ่งสามารถจะนำมาถอดรหัสได้

ช่วงความถี่ที่มนุษย์สามารถได้ยินจะอยู่ระหว่าง 20 - 20000 เฮิร์ต (20 กิโลเฮิร์ต) มนุษย์ไม่สามารถได้ยินความสั่นสะเทือนซึ่งเกิดขึ้นที่ความถี่ต่ำกว่า 20 ครั้งต่อวินาที และไม่สามารถรับรู้ความถี่ที่สูงกว่า 20 กิโลเฮิร์ต เสียงคำพูดจะประกอบด้วยพลังงานที่ระดับความถี่ต่างๆในช่วงที่มนุษย์สามารถได้ยินได้ โดยที่เสียงทั้งหมดจะอยู่ที่ระดับต่ำกว่า 8,000 เฮิร์ต

เรานำเทคนิคทางคณิตศาสตร์ที่เรียกว่าการวิเคราะห์ฟูริเยร์มาใช้กับรูปแบบคลื่นคำพูด เพื่อที่จะหาว่า ความถี่ใดที่เกิดขึ้นในเวลาต่างๆ ในสัญญาณคำพูด ผลจากการทำการวิเคราะห์ฟูริเยร์ก็คือ สเปกตรัม (Spectrum) หลังจากเราคำนวณสเปกตรัมสำหรับช่วงเวลาสั้นๆ (5-20 มิลลิวินาที) ของคำพูด เราจะคำนวณสเปกตรัมของช่วงต่อไปเรื่อยๆจนสิ้นสุดรูปแบบของคลื่นโดยทั่วๆไป สเปกตรัมที่อยู่ติดกันจะเปลี่ยนแปลงอย่างช้าๆและราบรื่น สะท้อนถึง (ชี้ให้เห็นถึง) การเคลื่อนไหวอย่างช้าๆของเสียงเทียบกับช่วงเวลาที่เราวิเคราะห์

การวิเคราะห์ดังกล่าวเรียกว่า การวิเคราะห์ฟูริเยร์ในช่วงเวลาสั้น (Short-time Fourier analysis) ซึ่งชุดของสเปกตรัมที่ได้มาจากช่วงเวลาต่างๆ นั้น สามารถนำไปแสดงผลได้ในรูปแบบของสเปกโตรแกรม

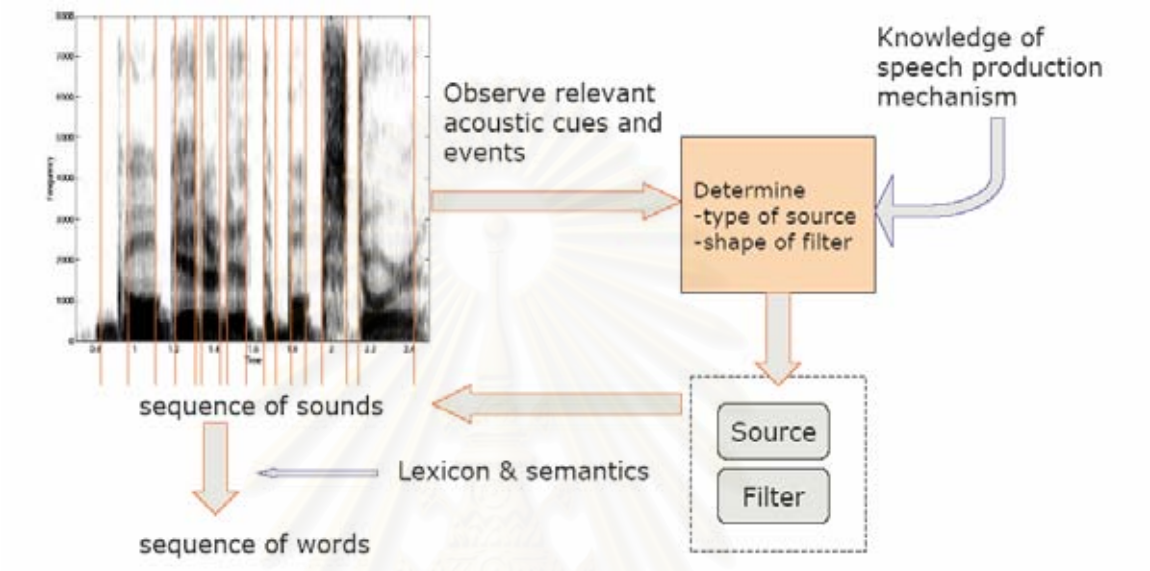


รูปที่ 2.8 สเปกโตรแกรมของเสียงพูดคำว่า “นายสง่าสรรพศรี”

สเปกโตรแกรมตัวอย่างในรูปที่ 2.8 เกิดขึ้นโดยการแสดงสเปกตรัมทั้งหมดที่คำนวณจากรูปแบบคลื่นของคำพูด แกนตั้งของสเปกตรัมแสดงความถี่โดยที่ระดับฐานจะเท่ากับ ศูนย์เฮิร์ต เส้นที่มองเห็นได้ในสเปกโตรแกรมแต่ละเส้นแสดงระดับ 1000 เฮิร์ต ตามแกนความถี่ ดังนั้น สเปกโตรแกรมจะประกอบด้วยความถี่ทั้งหมด 8000 เฮิร์ต สเปกตรัมที่คำนวณจากการแปลงฟูริเยร์ทั้งหมดจะถูกแสดงขนานกับแกนตั้ง แกนนอนแสดงถึงเวลา การเลื่อนไปทางขวาตามแกน x แสดงถึง

สเปกตรัมตามเวลาที่เพิ่มขึ้น สเปกโตรแกรมจะถูกคำนวณค่าพลังงานของเสียงและเก็บไว้ในอาเรย์ขนาดสองมิติ สำหรับสเปกโตรแกรม S ใดๆ ความแรงของสัญญาณความถี่ f ที่เวลา t ในสัญญาณเสียงพูดจะแสดงโดยความเข้มหรือสีในกราฟที่จุด $S(t, f)$

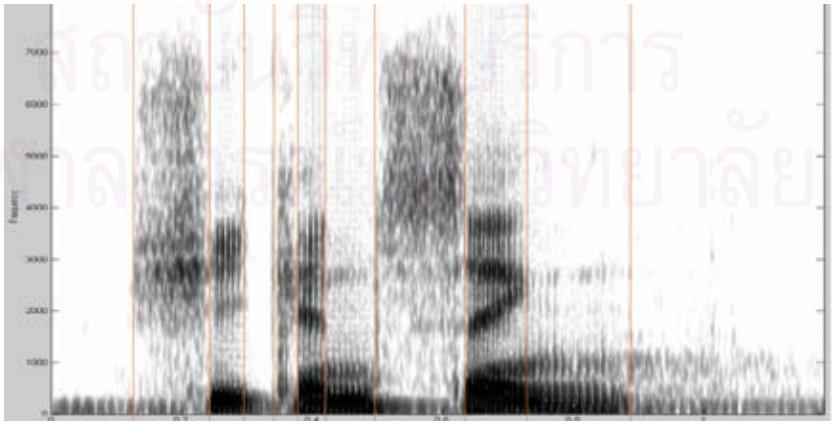
การอ่านสเปกโตรแกรมจะใช้พื้นฐานความรู้ทางด้านการออกเสียงคำพูดเพื่อแบ่งแยกลำดับของคำพูดจากสัญญาณเสียงที่ส่งเข้ามา โดยวิเคราะห์จากสเปกโตรแกรม



รูปที่ 2.9 การแบ่งแยกเสียงจากสัญญาณเสียงที่ได้รับเข้ามา

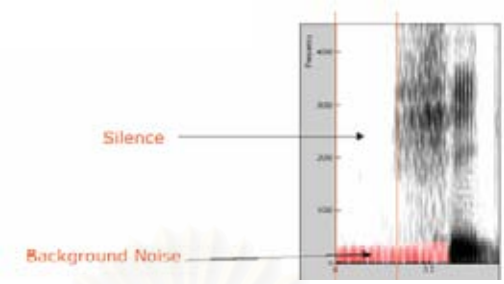
จากรูปที่ 2.9 จะเห็นได้ว่าเมื่อเราได้รับชุดลำดับของเสียงเข้ามา เราจะทำการแบ่งแยกชุดเสียงโดยใช้หลักเกณฑ์ต่างๆ เช่น ประเภทของแหล่งกำเนิด, รูปทรงของตัวกรองเสียง และจะใช้ความยาวกับไวยกรณ์เพื่อแปลความหมายจากชุดลำดับของเสียง ให้เกิดเป็นชุดลำดับของคำ โดยขั้นตอนการแปลความหมายของสเปกโตรแกรมมีดังนี้

1. ทำการกำกับขอบเขตหน่วยเสียงซึ่งสามารถพิจารณาได้จากสเปกโตรแกรม



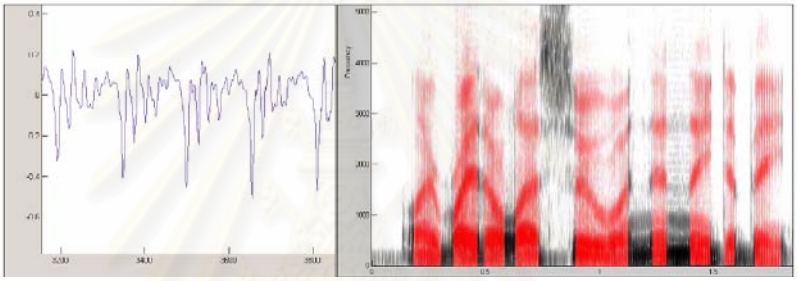
รูปที่ 2.10 การกำกับขอบเขตหน่วยเสียง

2. ทำการแบ่งประเภทกลุ่มของเสียง (Classes of Sound) ซึ่งประกอบด้วย
    1. เสียงเงียบ เป็นเสียงที่ไม่มีพลังงานใดๆ



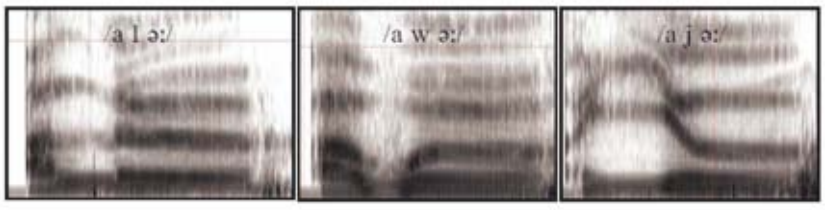
รูปที่ 2.11 สเปกโตรแกรมแสดงความแตกต่างระหว่างเสียงเงียบกับเสียงประเภทอื่นๆ

    2. เสียงสระ (Vowels) จะมีพลังงานสูง และลมมาก ซึ่งมีรูปแบบที่ชัดเจน



รูปที่ 2.12 สเปกโตรแกรมแสดงสัญญาณเสียงสระ

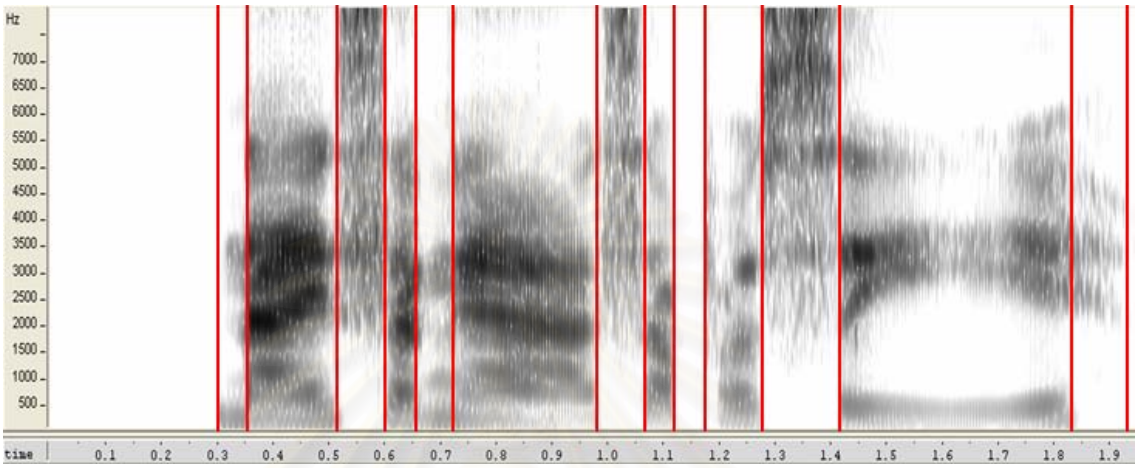
3. เสียงพยัญชนะ (Consonants)
    1. เสียงพยัญชนะกัก (Stop Consonant) ไม่มีพลังงานในช่วงปิด และมีเสียงระเบิดในช่วงเปิดปาก
    2. เสียงเสียดแทรก (Fricatives) มีรูปของเสียงรบกวน (noise) ในช่วงต้น
    3. เสียงกึ่งเสียดแทรก (Affricates) เป็นเสียงกักและต่อด้วยเสียงเสียดแทรก
    4. เสียงนาสิก (Nasal Consonants) พลังงานช่วงความถี่กลางและสูงมีน้อย ความถี่ F1 เพิ่มขึ้น
    5. เสียงกึ่งสระ (Semi-Vowel) มีการเคลื่อนตำแหน่งของความถี่ฟอร์แมนท์เร็วกว่าเสียงสระแต่ไม่มากเท่าเสียงพยัญชนะ



รูปที่ 2.13 สเปกโตรแกรมแสดงสัญญาณเสียงกึ่งสระ

6. ขอบเขตของหน่วยเสียง

ขอบเขตของหน่วยเสียงคือตำแหน่งบอกเวลาเริ่มต้นและสิ้นสุดของหน่วยเสียง ดังแสดงได้ด้วยสเปกโตรแกรมในรูปที่ 2.14 โดยเส้นตรงแต่ละเส้นแทนขอบเขตของหน่วยเสียง



รูปที่ 2.14 สเปกโตรแกรมแสดงการกำกับขอบเขตของหน่วยเสียง

การหาขอบเขตของหน่วยเสียงสามารถทำได้โดยอาศัยการอ่านสเปกโตรแกรม วิเคราะห์รูปแบบสเปกโตรแกรมของหน่วยเสียงแต่ละประเภท เช่นขอบเขตของเสียงสระจะเป็นส่วนที่มีแถบเข้มมากเนื่องจากเป็นเสียงที่มีพลังงานมาก เป็นต้น

ข้อมูลเสียงที่มีการกำกับขอบเขตของหน่วยเสียงไว้แล้วนี้ สามารถนำไปใช้ประโยชน์ด้านการสร้างระบบรู้จำเสียงพูด หรือการสร้างคลังข้อมูลเสียงสำหรับระบบสังเคราะห์เสียงได้

7. การสกัดลักษณะสำคัญ

การสกัดค่าลักษณะสำคัญ คือการวิเคราะห์หาค่าที่จะใช้แทนสัญญาณเสียง เพื่อนำไปใช้ในขั้นตอนการรู้จำ แบ่งได้เป็น 3 กลุ่มหลัก กลุ่มแรกเป็นค่าลักษณะสำคัญระดับสูง (High level feature) ได้แก่ สำเนียงการพูด รูปแบบในการพูด และความเร็วในการพูด เป็นต้น ในกลุ่มที่สอง จะใช้ค่าลักษณะสำคัญทางฉันทลักษณ์ (Prosodic feature) เช่น ค่าความถี่มูลฐาน (Fundamental frequency) ความถี่ฟอร์แมนท์ (Formant frequency) และระดับพลังงาน (Energy profile) เป็นต้น ถึงแม้ว่าค่าลักษณะสำคัญแบบนี้จะมีประสิทธิภาพสูงในการรู้จำ แต่ยากในการสกัดจากสัญญาณ กลุ่มสุดท้ายเรียกว่าค่าลักษณะสำคัญแบบเอนเวโลปเชิงสเปกตรัม (Spectral envelop feature) [10] เป็นกลุ่มที่นิยมใช้กันมาก เนื่องจากค่าลักษณะสำคัญส่วนใหญ่สำหรับการรู้จำเสียงจะรวมอยู่ในข้อมูลเชิงสเปกตรัมนี้ อีกทั้งยังง่ายและสะดวกในการคำนวณหาค่าด้วย ตัวอย่างค่าลักษณะสำคัญแบบนี้ได้แก่ สัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear prediction coefficients: LPC), สัมประสิทธิ์เซปสตรัม (Cepstral coefficient) และพัฒนาการอีกมากมายจากเซปสตรัมปกติ [11] อาทิเช่น สัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel frequency cepstral coefficients: MFCC) เซปสตรัมแบบหักลบค่าเฉลี่ย (Cepstral mean substraction: CMS) และเซปสตรัมแบบผ่านตัวกรองภายหลัง (Post filtered cepsturm: PFL) เป็นต้น นอกจากนั้น ยังมีการคำนวณค่าการเปลี่ยนแปลง (Derivative หรือ Delta) ของสัมประสิทธิ์เหล่านี้มาใช้เป็นค่าลักษณะสำคัญเพิ่มเติมได้ด้วย

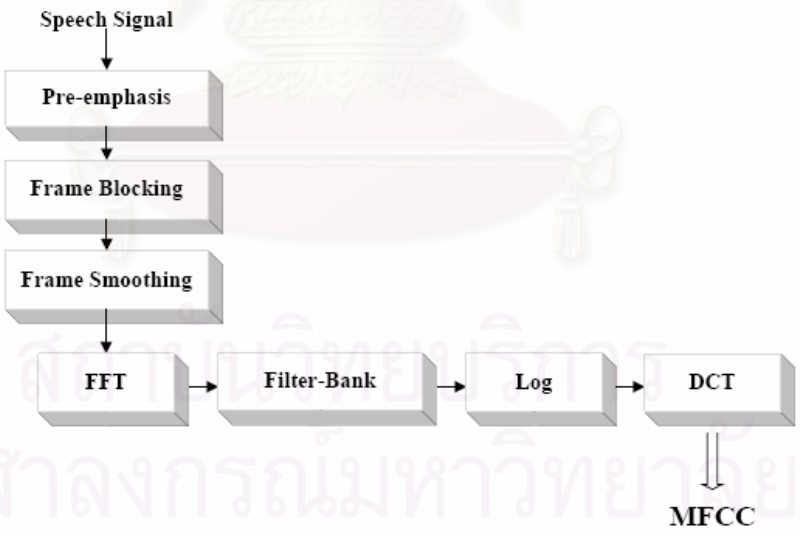
สำหรับการคำนวณค่าลักษณะสำคัญแบบเอนเวโลปของสเปกตรัมจะมีขั้นตอนดังนี้ [12]

1. การเน้นสัญญาณขั้นต้น (Preemphasis) เป็นขั้นตอนในการบีบอัดสัญญาณเสียงโดยนำสัญญาณเสียงผ่านตัวกรองลำดับหนึ่ง (First-order filter) ซึ่งจะเพิ่มอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to noise ratio)
2. การแบ่งเป็นส่วนย่อย (Frame) เป็นขั้นตอนในการแบ่งสัญญาณเสียงเป็นส่วนย่อยขนาดความยาวประมาณ 10 – 40 มิลลิวินาที ซึ่งทำให้สัญญาณเสียงมีคุณสมบัติเปลี่ยนแปลงตามเวลาน้อยมาก หรือไม่มีเลย เพื่อให้สามารถสร้างแบบจำลองการกระจายของหน่วยสัญญาณเสียงย่อยทางสถิติได้
3. การลดขอบด้วยฟังก์ชันหน้าต่างสำหรับปรับสัญญาณให้ราบเรียบ (Smoothing window)
4. การสกัดค่าลักษณะสำคัญ (Feature extraction) ในส่วนนี้ จะทำการคำนวณค่าลักษณะสำคัญของสัญญาณเสียงในแต่ละส่วนย่อย ผลลัพธ์อยู่ในรูปแบบของเวกเตอร์ของค่าลักษณะสำคัญ (Feature vector) สำหรับแต่ละส่วนย่อย

7.1 สัมประสิทธิ์เซปสตรัมบนสเกลเมล (Mel frequency cepstral coefficients - MFCC)

ค่าสัมประสิทธิ์เซปสตรัมเป็นค่าลักษณะสำคัญที่นิยมมากทั้งในระบบรู้จำผู้พูดและเสียงพูด โดยพื้นฐานแล้วเซปสตรัมสามารถคำนวณได้จาก การแปลงโคซายน์แบบไม่ต่อเนื่อง (Discrete cosine transformation) ของค่าลอการิธึม (Logarithm) ของสเปกตรัม ของสัญญาณเสียงแต่ละส่วนย่อย สเปกตรัมของสัญญาณเสียงสามารถหาได้โดยการแปลงฟูริเยร์แบบวิยุต หรือการแปลงฟูริเยร์แบบเร็ว ขั้นตอนดังกล่าวตั้งอยู่บนพื้นฐานแนวคิดที่ว่า สเปกตรัมของสัญญาณเสียงกำเนิดจากส่วนประกอบ 2 ส่วนคือ เอนเวโลปของสเปกตรัม (Spectral envelop) และโครงสร้างรายละเอียดของสเปกตรัม (Spectral fine structure) ทั้ง 2 ส่วนสามารถแยกกันได้ด้วยการใส่ลอการิธึม สัมประสิทธิ์เซปสตรัมเป็นการแทนสัญญาณในส่วนเอนเวโลปของสเปกตรัมเท่านั้น

พัฒนาการหนึ่งของเซปสตรัม คือการผ่านสเปกตรัมของสัญญาณเสียงเข้าไปในกลุ่มของตัวกรอง (Filter bank) ซึ่งกระจายอยู่บนสเกลความถี่แบบไม่สม่ำเสมอ เช่น การกระจายตามสเกลเมล (Mel scale) [10] ซึ่งออกแบบมาให้เหมาะสมกับการรับฟังของหู เป็นต้น ค่าพลังงานของสเปกตรัมของเสียงที่ได้จากตัวกรองแต่ละตัวจะถูกนำมาใช้คำนวณค่าสัมประสิทธิ์เซปสตรัมแทนค่าสเปกตรัมปกติ ค่าสัมประสิทธิ์เซปสตัมที่ได้จากการกระทำเช่นนี้จึงได้ชื่อว่า MFCC



รูปที่ 2.15 ขั้นตอนการคำนวณค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล

8. แบบจำลองฮิดเดนมาร์คอฟ

ระบบแบบจำลองฮิดเดนมาร์คอฟ เป็นขั้นตอนวิธีการจำแนกรูปแบบที่ดีที่สุดวิธีการหนึ่งที่มีอยู่ในขณะนี้ [13] ซึ่งอาศัยวิธีการทางสถิติ เหตุผลที่ระบบแบบจำลองฮิดเดนมาร์คอฟ เป็นที่นิยมมีด้วยกันสองประการคือ

ประการแรก แบบจำลองนี้อาศัยโครงสร้างทางคณิตศาสตร์ และสามารถเปลี่ยนแปลงทฤษฎีพื้นฐานเพื่อประยุกต์ใช้งานได้อย่างกว้างขวาง

ประการที่สอง แบบจำลองนี้สามารถทำงานได้เป็นอย่างดีเมื่อประยุกต์ใช้อย่างเหมาะสม โดยเราจะพิจารณาเนื้อหาของแบบจำลองฮิดเดนมาร์คอฟ โดยแบ่งออกเป็นส่วนย่อยๆ ดังนี้

8.1 องค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ

แบบจำลองฮิดเดนมาร์คอฟ ประกอบไปด้วยพารามิเตอร์ต่างๆดังนี้

1. จำนวนสถานะ (State) ที่อยู่ภายในแบบจำลอง (Model) ใช้สัญลักษณ์แทนด้วย “ N ” แต่ละสถานะแสดงได้ด้วย $S=\{S_1,S_2,...,S_N\}$ โดยมีสถานะที่เวลา t แสดงได้ด้วย q_t มีค่าแปรเปลี่ยนได้ ตามที่กำหนดจนกว่าจะได้ผลลัพธ์การรู้จำที่น่าพอใจ
2. จำนวนสัญลักษณ์ของค่าสังเกตต่อสถานะ ใช้สัญลักษณ์แทนด้วย “ M ” แต่ละสัญลักษณ์แสดงได้ด้วย $V=\{v_1,v_2,...,v_M\}$
3. การแจกแจงของความน่าจะเป็นในการเปลี่ยนสถานะ (State Transition Probability Distribution) ใช้สัญลักษณ์แทนด้วยเมตริก (Matrix) “ A ” โดย $A=\{a_{ij}\}$ เมื่อ $a_{ij}=P[q_{t+1}=S_j \mid q_t=S_i],\ 1\le i,\ j<N$ ในกรณีเฉพาะที่สถานะใดๆ สามารถเข้าถึงสถานะอื่นๆได้ภายในขั้นตอนเดียว จะกำหนดให้ ส่วนในกรณีอื่นนอกเหนือจากนี้จะกำหนดให้ สำหรับ $(i,\ j)$ เพียงคู่เดียวหรือมากกว่า
4. การแจกแจงของความน่าจะเป็นของสัญลักษณ์ของค่าสังเกต (Observation Symbol Probability Distribution) ใช้สัญลักษณ์แทนด้วยเมตริก (Matrix) “ B ” โดย $B=\{b_j(k)\}$ ในสถานะที่ j เมื่อ

$$b_j(k)=P[v_k\ at\ t \mid q_t=S_j],\ 1\le j\le N, 1\le k\le M$$

5. การแจกแจงสถานะเริ่มต้น (Initial State Distribution) ซึ่งก็คือความน่าจะเป็นของแบบจำลองที่จะเริ่มต้นแบบจำลองด้วยสถานะ i ใดๆ ใช้สัญลักษณ์แทนด้วย “ π ” โดย $\pi=\pi_i$ เมื่อ

$$\pi_i=P[q_1=S_i],\ 1\le i\le N$$

โดยการกำหนดค่าที่เหมาะสมให้กับองค์ประกอบ N, M, A, B, π ของแบบจำลองฮิดเดนมาร์คอฟซึ่งใช้ในการกำหนดลำดับค่าสังเกต เมื่อแต่ละค่าสังเกต O_t เป็นสัญลักษณ์ที่ได้จาก V และ T เป็นจำนวนค่าสังเกตทั้งหมดที่มีในลำดับซึ่งมีขั้นตอนวิธีการดังนี้

ตารางที่ 2.5 ขั้นตอนการกำหนดค่าองค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ

ขั้นตอนที่	รายละเอียดการดำเนินการ
1	เลือกสถานะเริ่มต้น $q_1 = S_i$ ที่สัมพันธ์กับการกระจายของสภาวะเริ่มต้น π
2	กำหนดให้ $t = 1$
3	เคลื่อนย้ายไปยังสถานะใหม่ $q_{t+1} = S_j$ ที่สัมพันธ์กับการกระจายความน่าจะเป็นในการเปลี่ยนแปลงสถานะสำหรับสถานะ S_i เช่น a_{ij}
4	กำหนดให้ $t = t + 1$ แล้วกลับไปทำซ้ำขั้นตอนที่ 3 ใหม่ถ้า $t < T$ นอกเหนือจากนี้ให้ยุติกระบวนการ

ขั้นตอนดังกล่าวนี้เป็นได้ทั้งการกำเนิดค่าสังเกต และเป็นแบบจำลองเพื่อบอกถึงความเหมาะสมในการกำเนิดลำดับค่าสังเกตด้วยแบบจำลองฮิดเดนมาร์คอฟ ดังนั้นการกำหนดคุณสมบัติเฉพาะของแบบจำลองฮิดเดนมาร์คอฟ จึงต้องการคุณสมบัติเฉพาะของพารามิเตอร์ของแบบจำลอง (นั่นคือ N และ M) คุณสมบัติเฉพาะของสัญลักษณ์ของค่าสังเกต และ คุณสมบัติเฉพาะของการวัดค่าความน่าจะเป็นอันได้แก่ A, B, π โดยทั้งหมดนี้สามารถเขียนให้อยู่ในรูปแบบ แบบย่อเพื่อบ่งบอกถึงชุดพารามิเตอร์ที่สมบูรณ์ของแบบจำลองได้ดังนี้

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

แบบจำลองฮิดเดนมาร์คอฟก็เปรียบเสมือนเครื่องจักรสถานะ (State Machine) แบบหนึ่งซึ่งนำไปใช้อธิบายรูปแบบการเกิดของเวกเตอร์คุณสมบัติของเสียง (Feature Vector) โดยการที่จะรู้ว่าเวกเตอร์คุณสมบัติของเสียงนั้นเกิดได้อย่างไรสามารถทำได้ด้วยการย้อนรอยสถานะเพื่อหาเส้นทางการรู้จำของเสียงว่าเวกเตอร์คุณสมบัติของเสียงนั้นได้มาจากหน่วยเสียงใดบ้างตามสถานะที่เสียงนั้นๆผ่าน

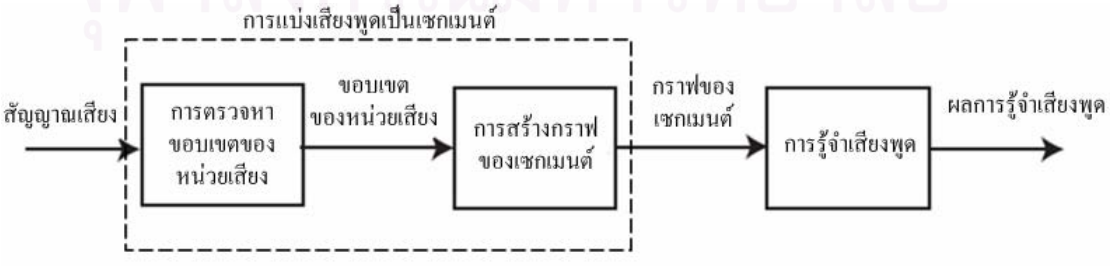
9. การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

กว่าสองทศวรรษที่แบบจำลองฮิดเดนมาร์คอฟได้รับความนิยมมากในการนำมาใช้กับการพัฒนาเครื่องรู้จำเสียงพูด สาเหตุหนึ่งมาจากการที่แบบจำลองฮิดเดนมาร์คอฟมีรากฐานทางคณิตศาสตร์ที่ออกแบบมาเป็นอย่างดี และสามารถนำไปใช้แก้ปัญหาการรู้จำเสียงพูดได้เป็นอย่างมีประสิทธิภาพ แต่แบบจำลองเสียงพูดในเครื่องรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟมีคุณสมบัติสำคัญที่ทำให้เกิดข้อจำกัดสองประการ [14] คือ

คุณสมบัติประการแรก คือการพิจารณาลำดับของเวกเตอร์ลักษณะสำคัญที่คำนวณมาจากสัญญาณเสียงในแต่ละกรอบเวลาสั้นๆที่มีขนาดตายตัว (โดยทั่วไปมีขนาด 10 มิลลิวินาทีต่อหนึ่งกรอบเวลา) แต่ในกระบวนการสร้างเสียงพูด อวัยวะส่วนกระทำอาการจะเคลื่อนไหวช้า ทำให้เวกเตอร์ลักษณะสำคัญที่คำนวณจากสัญญาณเสียงในกรอบเวลาที่อยู่ติดกันมีความคล้ายคลึงกันและความเกี่ยวพันกันสูงโดยเฉพาะอย่างยิ่งเวกเตอร์ลักษณะสำคัญที่คำนวณมาจากสัญญาณเสียงพูดที่มีหน่วยเสียงเหมือนกัน ซึ่งขัดกับคุณสมบัติหนึ่งของแบบจำลองฮิดเดนมาร์คอฟที่สมมุติให้เวกเตอร์ลักษณะสำคัญแต่ละเวกเตอร์เป็นอิสระต่อกันแบบมีเงื่อนไข

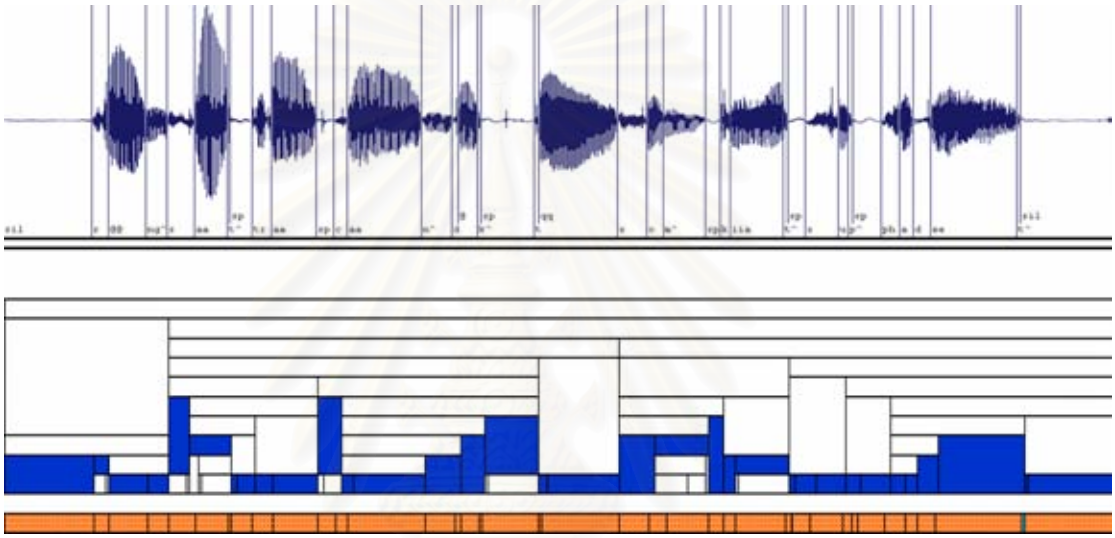
คุณสมบัติประการที่สอง คือการคำนวณหาเวกเตอร์ลักษณะสำคัญจากสัญญาณเสียงในแต่ละกรอบเวลาจะต้องเลือกใช้ลักษณะสำคัญที่เหมือนกันทั้งหมด (เช่น ใช้สัมประสิทธิ์เซปสตรัมบนสเกลเมลเป็นลักษณะสำคัญเพียงแบบเดียว) ทำให้ยากต่อการนำสารสนเทศทางสวนสัทศาสตร์ หรือลักษณะสำคัญอื่นๆ เช่น เสียงวรรณยุกต์ มารวมกันเพื่อเพิ่มประสิทธิภาพการรู้จำเสียงพูดให้ดีขึ้น

เพื่อผ่อนปรนข้อจำกัดเหล่านี้ จึงมีการเสนอวิธีการรู้จำเสียงพูดแบบอาศัยเซกเมนต์ [1] ซึ่งเป็นวิธีการรู้จำเสียงพูดทางสถิติ ซึ่งแตกต่างจากการรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟตรงที่การรู้จำเสียงพูดจะพิจารณาข้อมูลเสียงเป็นเซกเมนต์ โดยที่แต่ละเซกเมนต์ไม่จำเป็นต้องมีขนาดเท่าๆกัน แล้วจำแนกว่าแต่ละเซกเมนต์นั้นมีหน่วยเสียงเป็นอะไร ด้วยเหตุนี้สิ่งที่จำเป็นและส่งผลต่อประสิทธิภาพของเครื่องรู้จำเสียงแบบอาศัยเซกเมนต์โดยตรง คือการมีข้อมูลเสียงที่มีการกำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว ซึ่งเป็นผลมาจากกระบวนการแบ่งเสียงพูดเป็นเซกเมนต์ที่มีความถูกต้องแม่นยำและทำงานได้อย่างรวดเร็ว



รูปที่ 2.16 แผนภาพส่วนประกอบของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์

ภาพรวมส่วนประกอบของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์แสดงได้ด้วยรูปที่ 2.16 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์จะพิจารณาหน่วยเสียงให้เป็นเซกเมนต์ โดยขั้นตอนการทำงานจะประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนที่ทำงานต่อเนื่องกัน คือ ขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ และขั้นตอนการรู้จำเสียงพูด จุดประสงค์ของขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์ ก็เพื่อสันนิษฐานหาขอบเขตของเซกเมนต์แล้วนำมาประกอบกันเป็นกราฟของเซกเมนต์ โดยในขั้นตอนการรู้จำเสียงพูด กราฟนี้จะถูกใช้เป็นข้อมูลอินพุตเพื่อค้นหาลำดับของหน่วยเสียงที่ดีที่สุดออกมา ดังรูปที่ 2.17

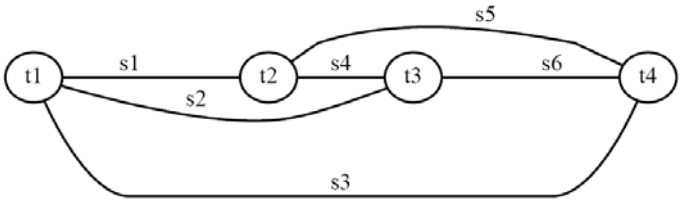


รูปที่ 2.17 สัญญาณเสียงที่กำกับขอบเขตของหน่วยเสียงไว้แล้ว (บน) กราฟของเซกเมนต์ (ล่าง) (สี่เหลี่ยมสีเข้มแทนเซกเมนต์ที่ให้ระบบรู้จำเสียงพูดเลือกเป็นคำตอบ สี่เหลี่ยมสีขาวแทนเซกเมนต์ของหน่วยเสียงอื่นๆที่เป็นไปได้ และสี่เหลี่ยมสีเทาแทนเซกเมนต์ของเสียงพูดที่ถูกต้อง)

9.1 การแบ่งเสียงพูดเป็นเซกเมนต์

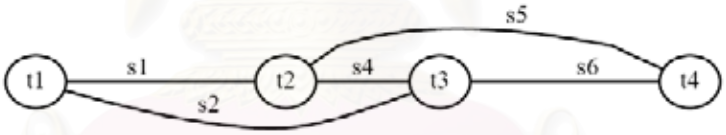
การแบ่งเสียงพูดเป็นเซกเมนต์ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ คือการหาขอบเขตของเซกเมนต์ในที่นี้คือหน่วยเสียง แล้วนำมาประกอบกันสร้างเป็นกราฟของเซกเมนต์ โดยการทำงานจะประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนคือ ขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง และขั้นตอนการสร้างกราฟของเซกเมนต์ โดยในแต่ละขั้นตอนสามารถใช้วิธีการได้หลายวิธี ตัวอย่างเช่น อาจใช้วิธีรู้จำเสียงพูดในระดับหน่วยเสียงออกมาก่อนแล้วจึงพิจารณาเวลาเริ่มต้นหรือสิ้นสุดลงของหน่วยเสียงที่รู้จำออกมาได้ให้เป็นเขตของหน่วยเสียง หรืออาจใช้วิธีดูการเปลี่ยนแปลงสเปกตรัมของสัญญาณเสียงแล้วพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงมากเกินกว่าระดับที่กำหนด

ไว้ให้เป็นขอบเขตของหน่วยเสียง หรืออาจใช้วิธีการจำแนกเสียงพูดออกเป็นประเภทกว้างๆ แล้วพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงประเภทหน่วยเสียงนั้นๆให้เป็นขอบเขตของหน่วยเสียงก็ได้



รูปที่ 2.18 ตัวอย่างกราฟของเซกเมนต์แบบเชื่อมต่อกันหมด

ลำดับของขอบเขตของหน่วยเสียงที่ตรวจหามาได้ จะนำมาประกอบกันสร้างเป็นกราฟของเซกเมนต์ กราฟที่ดีจะต้องมีขนาดเล็กและครอบคลุมหน่วยเสียงที่แท้จริงไว้ได้เป็นจำนวนมาก ในทางทฤษฎีแล้วการรู้จำเสียงพูดแบบอาศัยเซกเมนต์สามารถใช้กราฟของเซกเมนต์ที่ทุกปมของกราฟเชื่อมต่อกันหมดมารู้จำเสียงพูดก็ได้ แต่เนื่องจากจำนวนเซกเมนต์ในกราฟนั้นมีขนาดใหญ่มาก ทำให้ขั้นตอนการรู้จำใช้เวลาในการทำงานมากตามไปด้วย ด้วยเหตุนี้วิธีการแบบอาศัยเซกเมนต์ส่วนใหญ่จึงต้องมีการลดขนาดของกราฟลงก่อนด้วยการตัดเส้นเชื่อมบางเส้นออกไป ตัวอย่างเช่นรูปที่ 2.18 เป็นกราฟของเซกเมนต์ที่ทุกปมเชื่อมต่อกันหมด และรูปที่ 2.19 แสดงกราฟของเซกเมนต์แบบที่ตัดเซกเมนต์ s_3 ออกไป



รูปที่ 2.19 กราฟของเซกเมนต์แบบที่ยอมให้มีการเชื่อมต่อกันบางส่วนเท่านั้น

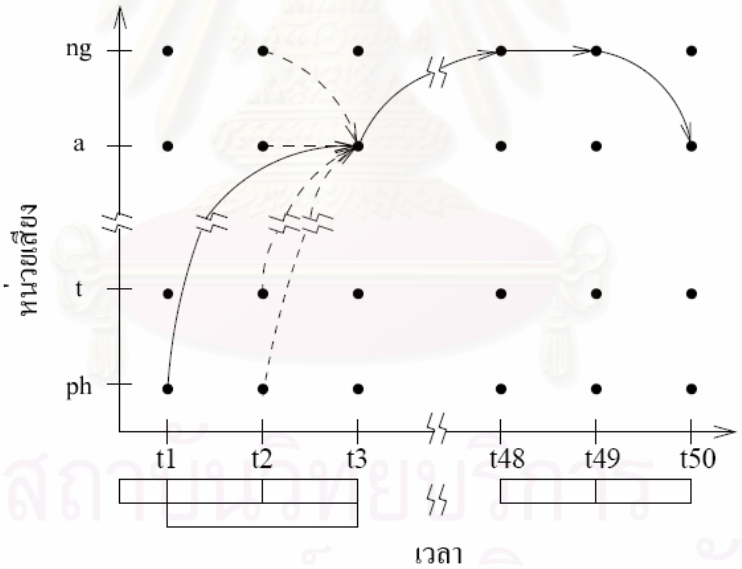
คุณภาพของกราฟที่ได้จากขั้นตอนการแบ่งเสียงพูดเป็นเซกเมนต์จะเป็นการแลกเปลี่ยนกันในแง่ของขนาดของกราฟและความครอบคลุม ซึ่งจะส่งผลถึงเวลาที่ใช้ในการทำงานและความถูกต้องในการรู้จำเสียงพูดในขั้นตอนการค้นหาและรู้จำเสียงพูดต่อไป

9.2 การรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การรู้จำเสียงพูดแบบอาศัยเซกเมนต์ จะเป็นการใช้อัลกอริทึมแบบกำหนดการพลวัต (Dynamic Programming) ค้นหาลำดับของเซกเมนต์ที่ดีที่สุดออกมาจากกราฟของเซกเมนต์ โดยอาศัยค่าลักษณะสำคัญทั้งแบบอาศัยกรอบเวลาและแบบอาศัยเซกเมนต์ ซึ่งค่าลักษณะสำคัญแบบอาศัยกรอบเวลาจะคำนวณมาจากสัญญาณเสียงในแต่ละกรอบเวลาสั้นๆที่มีขนาดตายตัว ตัวอย่างเช่น ค่าสัมประสิทธิ์เซปสตรัมบนสเกลเมล หรือ ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น

เป็นต้น ส่วนค่าลักษณะสำคัญแบบอาศัยเซกเมนต์ จะคำนวณมาจากสัญญาณเสียงของแต่ละเซกเมนต์ ตัวอย่างเช่น ค่าความยาวของเซกเมนต์ เป็นต้น โดยค่าลักษณะสำคัญเหล่านี้จะนำมาใช้คำนวณเป็นค่าคะแนนสะท้อนความน่าจะเป็นว่าสัญญาณเสียงของแต่ละเซกเมนต์นั้นน่าจะเป็นหน่วยเสียงใด

การค้นหาลำดับของเซกเมนต์ที่ดีที่สุดออกมาจากกราฟของเซกเมนต์จะพิจารณากราฟของเซกเมนต์ควบคู่ไปกับหน่วยเสียงประเภทต่างๆ ดังรูปที่ 2.20 โดยในแนวแกนนอนคือเวลา แสดงลำดับของเซกเมนต์ แกนตั้งแสดงเซตของหน่วยเสียงทั้งหมด และเส้นโค้งเชื่อมต่อระหว่างปมในกราฟ จะเป็นเส้นทางเดินของเหตุการณ์ที่เซกเมนต์นั้นๆมีหน่วยเสียงเป็นอะไร ตัวอย่างเช่นเส้นโค้งที่เชื่อมต่อจาก t_1 ไป t_3 ซึ่งชี้ตรงไปยังปมของหน่วยเสียง a ก็คือเส้นทางการเดินของเหตุการณ์ที่เซกเมนต์ซึ่งมีเวลาเริ่มต้นอยู่ที่ t_1 และสิ้นสุดที่เวลา t_3 เป็นหน่วยเสียง a โดยที่แต่ละเส้นโค้งนั้นจะมีการคำนวณค่าคะแนนกำกับไว้ และการค้นหาลำดับของเซกเมนต์ที่ดีที่สุดจะต้องพิจารณาเส้นทางเดินในกราฟของเซกเมนต์ให้ครบทุกเส้นทางเพื่อเลือกเส้นทางที่มีคะแนนรวมสูงที่สุดเป็นผลการรู้จำเสียงพูด



รูปที่ 2.20 การค้นหาลำดับของเซกเมนต์จากกราฟของเซกเมนต์

10. ซัพพอร์ตเวกเตอร์แมชชีน – เอสวีเอ็ม [15]

ซัพพอร์ตเวกเตอร์แมชชีน [16] เป็นเทคนิคการเรียนรู้เชิงสถิติที่ Vapnik เริ่มคิดค้นตั้งแต่ช่วงปี ค.ศ. 1960 แต่ยังไม่เป็นที่นิยมจนถึงช่วงทศวรรษที่ผ่านมาเริ่มได้รับความสนใจมาก สาเหตุหนึ่งมาจากการที่มีการนำไปใช้แก้ปัญหาต่างๆและพบว่าได้ผลดีมาก เนื้อหาในหัวข้อนี้จะกล่าวถึงแนวคิดของเอสวีเอ็มและเอสวีเอ็มแบบหลายประเภท

10.1 แนวคิดพื้นฐานของซัพพอร์ตเวกเตอร์แมชชีน

เอสวีเอ็มเกิดจากแนวคิดพื้นฐานดังนี้

1. การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด (Structural risk minimization) เป็นแนวคิดที่แสดงขอบเขตของความเสี่ยงของเครื่องเรียนรู้ ซึ่งขึ้นกับความเสี่ยงเชิงทดลองที่มาจากความผิดพลาดในการสอน กับช่วงความเชื่อมั่นที่เป็นฟังก์ชันของมิติวีซี (VC dimension) ที่แสดงถึงว่าฟังก์ชันที่ใช้จำแนกมีลักษณะทั่วไปเพียงใด ในทางปฏิบัติเราไม่สามารถลดความเสี่ยงที่แท้จริงได้ จึงพยายามลดความเสี่ยงจากความผิดพลาดให้ต่ำสุดแทน
2. ระนาบหลายมิติแบ่งแยกดีสุด (Optimal separating hyperplane) ระนาบหลายมิตินี้จะต่างจากของข่ายงานประสาทเทียมตรงที่เป็นระนาบที่แบ่งคอนเวกซ์ฮัล (Convex hull) ของข้อมูลสองกลุ่มออกจากกันด้วยระยะห่างที่กว้างที่สุด
3. ฟังก์ชันเคอร์เนล (Kernel Function) เป็นเทคนิคที่ช่วยขยายความสามารถของเอสวีเอ็มให้จัดการกับปัญหาที่ไม่สามารถแบ่งแยกแบบเชิงเส้นได้ โดยการแมปข้อมูลในปริภูมินำเข้า (Input Space) ไปสู่ปริภูมิคุณลักษณะ (Feature Space) ที่มีอันดับสูงขึ้นไปซึ่ง ณ ปริภูมิอันดับสูงนี้ จะสามารถใช้ระนาบหลายมิติแบบเชิงเส้นในการแยกข้อมูลสองกลุ่มออกจากกันได้

10.2 การลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด

ในปัญหาการรู้จำแบบเราต้องการหาฟังก์ชันที่มาประมาณตัวจำแนกประเภท (Classifier) ที่แท้จริง ซึ่งค่าความผิดพลาดคือผลต่างระหว่างค่าที่ได้จากฟังก์ชันประมาณกับค่าที่ได้จากฟังก์ชันที่แท้จริง ซึ่งโดยปกติเราจะต้องการลดความเสี่ยงจากการจำแนกประเภทผิดให้ต่ำที่สุด แต่ในทางปฏิบัติเราไม่รู้ฟังก์ชันที่แท้จริง จึงไม่สามารถคำนวณหาค่าผิดพลาดที่แท้จริงได้

ทางหนึ่งที่ทำได้คือ ในการสอนเราจะพิจารณาค่าผิดพลาดจากกลุ่มตัวอย่างแทน และลดค่าของความเสี่ยงเชิงโครงสร้างซึ่งประกอบด้วย ความเสี่ยงเชิงทดลอง (Empirical risk) กับช่วงความ

เชื่อมั่น (confidence interval) ให้ต่ำสุดแทน ความเสี่ยงเชิงโครงสร้างนี้มีพื้นฐานมาจากข้อเสนอของ Vapnik เรื่องขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่องกับมิติวีซี

10.2.1 ขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง

Vapnik ได้เสนอขอบเขตของความผิดพลาดในการรู้จำแบบของเครื่อง โดยสมมติให้เรามีข้อมูลอยู่ l ตัวซึ่งแต่ละข้อมูลประกอบด้วยคู่ของ เวกเตอร์ $\mathbf{x}$ กับฉลากแสดงประเภท y โดยที่

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_l) \in \mathrm{R}^{\mathrm{N}} \times \{\pm 1\}$$

ตัวอย่างเช่น $\mathbf{x}$ อาจหมายถึงค่าสีของแต่ละจุดภาพของรูปตัวอักษรที่ต้องการรู้จำ ส่วน y แสดงถึงประเภทของรูปภาพว่าเป็นตัวอักษรที่สนใจหรือไม่ ถ้า y เป็น 1 หมายถึงว่าเป็นตัวอักษรที่สนใจ ส่วนถ้า y เป็น -1 ก็แสดงว่าไม่ใช่ตัวอักษรที่สนใจ และหากลองพิจารณาในแง่ของการรู้จำแบบ ภายใต้การแจกแจงความน่าจะเป็น $P(\mathbf{x}, y)$ ใดๆ สมมติว่าเราต้องการเครื่องที่จะเรียนรู้การจำแบบ $\mathbf{x}_i \Rightarrow y_i$ ซึ่งเรานิยามให้เป็นเซตของการรู้จำแบบที่เป็นไปได้ $\mathbf{x} \Rightarrow f(\mathbf{x}, \alpha)$ ซึ่ง α นี้เป็นตัวแปรที่สามารถเปลี่ยนแปลงค่าได้ และเครื่องที่ได้รับการสอนแล้วก็จะมีค่า α ที่แน่นอนของตัวเอง อย่างเช่นในข่ายงานประสาทเทียม ค่า α หมายถึงน้ำหนักและค่าขีดแบ่ง

เรานิยามค่าผิดพลาดสำหรับเครื่องที่ได้รับการสอนนี้เป็น

$$R(\alpha) = \int \frac{1}{2} \left| y - f(\mathbf{x}, \alpha) \right| dP(\mathbf{x}, y)$$

โดยที่ค่า $R(\alpha)$ เป็นความผิดพลาดที่แท้จริง และค่า $1/2\left|y - f(\mathbf{x}, \alpha)\right|$ จะมีค่าเป็น 0 หรือ 1 ส่วน $R_{emp}(\alpha)$ เป็นค่าความผิดพลาดโดยเฉลี่ยในข้อมูลสอนเท่านั้นและสามารถหาค่าได้ดังนี้

$$R_{emp}(\alpha) = \frac{1}{2l} \sum^{l} \left| y - f(\mathbf{x}, \alpha) \right|$$

ค่า $R_{emp}(\alpha)$ จะคงที่สำหรับ α และข้อมูลสอน $\left\{(\mathbf{x}_i, y_i)\right\}$ ชุดหนึ่งๆ เราเรียก $R_{emp}(\alpha)$ ว่าความเสี่ยงเชิงทดลอง

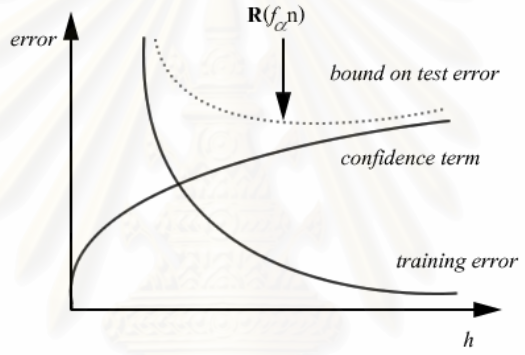
เมื่อเลือก η ตัวหนึ่งโดยที่ $0 \le \eta \le 1$ จะได้ว่า ที่ระดับความเชื่อมั่น $1-\eta$ ความผิดพลาดที่แท้จริงจะมีขอบเขตดังนี้

$$R(\alpha) \le R_{emp}(\alpha) + \sqrt{\frac{h\left(\log\left(\frac{2l}{h}\right)+1\right) - \log\left(\frac{\eta}{4}\right)}{l}}$$

เราเรียก h ซึ่งเป็นจำนวนเต็มบวกหรือศูนย์ เรียกว่ามิติวีซี (Vapnik Chervonenkis dimension – VC dimension) และเรียกพจน์ขวามือว่าขอบเขตความเสี่ยง (Risk bound) ซึ่งพบว่าขอบเขตนี้ไม่ขึ้นกับการแจกแจงความน่าจะเป็น $P(\mathbf{x}_i, y)$

โดยทั่วไป เราไม่สามารถคำนวณค่าของพจน์ทางซ้ายมือได้แต่ถ้ารู้ h จะสามารถคำนวณพจน์ทางขวามือได้ ดังนั้นโดยหลักการแล้ว ในการเรียนรู้ เราจึงกำหนดค่า η เป็นค่าคงที่ต่ำๆแล้วเลือกเครื่องที่ลดค่าทางขวามือให้ต่ำที่สุด เราก็จะได้เครื่องที่ให้ขอบเขตความเสี่ยงต่ำที่สุด ซึ่งแนวคิดนี้เป็นแนวคิดที่สำคัญของการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

เราเรียกพจน์ที่สองทางขวามือของสมการขอบเขตความผิดพลาดแท้จริงข้างต้นนี้ว่าช่วงความเชื่อมั่น พบว่าช่วงความเชื่อมั่นนี้เป็นฟังก์ชันของมิติ VC ที่แทนด้วย h ซึ่งการลดค่า h ให้ต่ำที่สุดจะเป็นการลดความเสี่ยงเชิงโครงสร้างด้วย แต่โดยปกติเมื่อค่า h ลดลงความเสี่ยงเชิงทดลองจะสูงขึ้นตามดังนั้นในการลดความเสี่ยงเชิงโครงสร้างจึงต้องหาจุดที่ผลจากทั้งความเสี่ยงเชิงทดลองและช่วงความเชื่อมั่นต่ำที่สุด โดยเลือกฟังก์ชันในการเรียนรู้เครื่อง $f(\mathbf{x},\alpha)$ ที่มีขอบเขตความเสี่ยงต่ำสุดดังในรูปที่ 2.21

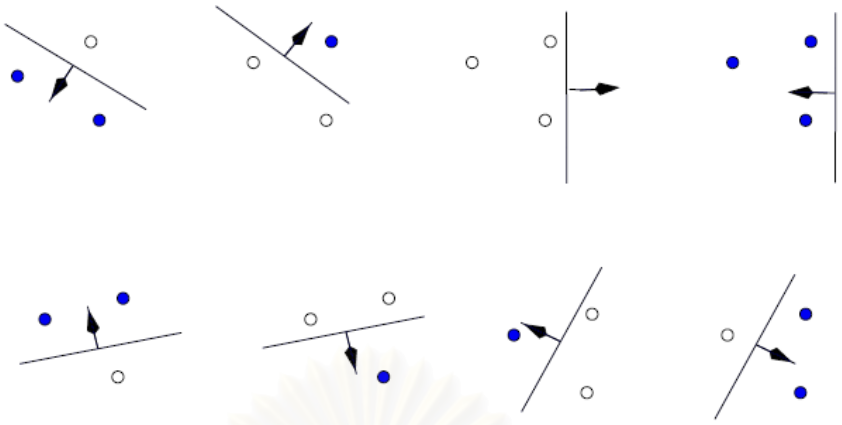


รูปที่ 2.21 ความสัมพันธ์ระหว่าง VC(h) กับค่าผิดพลาด

10.2.2 มิติวีซี

มิติวีซีเป็นคุณลักษณะของเซตของฟังก์ชัน $\{f(\alpha)\}$ ที่แสดงถึงความสามารถของฟังก์ชันในการแบ่งแยกกลุ่มของจุดข้อมูลออกจากกัน โดยทั่วไปฟังก์ชันที่มีมิติวีซีสูงจะมีความสามารถในการแบ่งแยกสูง ในที่นี้จะขออธิบายเฉพาะฟังก์ชันสำหรับกรณีการรู้จำแบบที่มี 2 ประเภทเท่านั้นคือ $f(\mathbf{x},\alpha) \in \{-1,1\}\ \forall \mathbf{x},\alpha$

กำหนดให้มีขอบเขตของจุด l จุด ซึ่งสามารถกำหนดว่าจุดแต่ละจุดอยู่ประเภทใดได้ทั้งหมด 2^l แบบและถ้าในแบบแต่ละแบบมีฟังก์ชันอย่างน้อยหนึ่งฟังก์ชันใน $\{f(\alpha)\}$ ที่สามารถกำหนดประเภทให้กับจุดแต่ละจุดได้อย่างถูกต้อง จะเรียกได้ว่าเซตของจุดเหล่านี้ถูกทำให้แตก (Shattered) โดยเซตของฟังก์ชันนี้ ดังแสดงในรูปที่ 2.22



รูปที่ 2.22 เซตของจุด 3 จุดใน R^2 ถูกทำให้แตกโดยเส้นที่มีทิศทาง

เรานิยามมิติวีซีของฟังก์ชัน $\{f(\alpha)\}$ ให้เป็นจำนวนสูงสุดของจุดในข้อมูลสอนที่สามารถทำให้แตกด้วย $\{f(\alpha)\}$ ได้ โดยมีข้อสังเกตคือ ถ้าให้มิติวีซีมีค่าเป็น h แล้ว จะมีอย่างน้อยหนึ่งเซตของจุดจำนวน h จุดที่ถูกทำให้แตกออกได้ แต่โดยทั่วไปไม่จำเป็นว่าทุกเซตของจุดจำนวน h จุดจะถูกทำให้แตกออกเป็นส่วนได้เสมอไป

ทฤษฎีบทที่ 2.1 พิจารณาเซตที่ประกอบด้วยจุด m จุดใน R^n ถ้าเลือกจุดใดจุดหนึ่งเป็นจุดกำเนิด จะได้ว่าจุด m จุดเหล่านี้จะถูกจำแนกประเภทโดยระนาบหลายมิติแบบมีทิศทาง (Oriented hyperplane) ได้ก็ต่อเมื่อเวกเตอร์บอกตำแหน่งของจุดที่เหลือเป็นอิสระเชิงเส้นต่อกัน

ผลที่ตามมาคือมิติวีซีของเซตของฟังก์ชันระนาบหลายมิติแบบมีทิศทางใน R^n มีค่าเป็น n+1 ซึ่งพิสูจน์ได้ดังนี้ เนื่องจากเราสามารถเลือกจุดจำนวน n+1 จุดแล้วให้จุดหนึ่งจุดเป็นจุดกำเนิด ดังนั้นจุดที่เหลือ n จุดย่อมต้องเป็นอิสระเชิงเส้นต่อกันแน่นอน (เช่น ให้จุดแต่ละจุดอยู่บนแกนแต่ละแกนใน R^n)

มีข้อสังเกตว่าฟังก์ชันที่มีพารามิเตอร์มากไม่จำเป็นต้องมีมิติวีซีมาก และในทางกลับกันฟังก์ชันที่มีพารามิเตอร์เพียงหนึ่งเดียวอาจมีมิติวีซีเป็นอนันต์ได้ แต่แม้ฟังก์ชันจะมีมิติวีซีเป็นอนันต์ก็อาจจะไม่สามารถแยะจุดเพียงไม่กี่จุดได้

ในทางปฏิบัติมิติวีซีที่ต่ำๆย่อมทำให้ขอบเขตของความผิดพลาดแท้จริงต่ำด้วย แต่ก็ไม่ได้หมายความว่าฟังก์ชันที่มีมิติวีซีสูงๆจะใช้งานได้ไม่ดี เช่น เอสวีเอ็มแบบอาร์บีเอฟ (RBF-Radial Basis Function) ซึ่งมีมิติวีซีเป็นอนันต์ก็เป็นฟังก์ชันที่ใช้งานได้ดีฟังก์ชันหนึ่ง อย่างไรก็ตามการเลือกฟังก์ชันเคอร์เนลฟังก์ชันสำหรับเครื่องเรียนรู้ ไม่ได้มีรูปแบบที่แน่นอน หากแต่ต้องการอ้างอิงจากผลการทดลอง เนื่องจากยังไม่มีข้อสรุปทางทฤษฎีว่าเคอร์เนลแบบใดเหมาะสมกับปัญหาแบบใด

10.2.3 วิธีการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด

ขอบเขตบนของความผิดพลาดอันอาจเกิดจากการทำให้มีลักษณะทั่วไปสามารถเขียนให้อยู่ในรูปสมการดังต่อไปนี้

$$R(\alpha) \le R_{emp}(\alpha) + \Phi\left(\frac{l}{h}\right)$$

ซึ่งพจน์แรกคือ ความเสี่ยงเชิงทดลอง ส่วนพจน์หลังคือ ช่วงความเชื่อมั่นซึ่งเป็นฟังก์ชันของมิติวีซี (h) ในการลดความเสี่ยงเชิงโครงสร้างให้ต่ำที่สุด เราจะต้องคำนึงถึงพจน์ทั้งสองของสมการข้างต้น โดยมีอยู่สองแนวทางดังนี้

แนวทางแรก ให้คงค่าช่วงความเชื่อมั่นให้คงที่แล้วลดความเสี่ยงเชิงทดลองให้ต่ำสุด แนวทางนี้จะลดพจน์แรกของสมการให้ต่ำสุดโดยการออกแบบเครื่องเรียนรู้ให้ซับซ้อน แต่ถ้าออกแบบเครื่องที่ซับซ้อนเกินไป จะทำให้ช่วงความเชื่อมั่นกว้างขึ้น ซึ่งแม้จะสามารถลดความเสี่ยงเชิงทดลองลงจนหมดสิ้นไปได้ แต่ความผิดพลาดที่เกิดขึ้นในการใช้งานจริงยังอาจสูงอยู่ ปรากฏการณ์ที่เกิดขึ้นนี้เรียกว่าการปรับเหมาะเกินไป อย่างไรก็ตามถ้าเราเลือกเครื่องที่มีความซับซ้อนต่ำ เพื่อให้ช่วงความเชื่อมั่นแคบ เราก็จะประสบปัญหาในการหาฟังก์ชันที่จะนำมาใช้ประมาณปัญหาได้ยาก อันจะทำให้เกิดความผิดพลาดเชิงทดลองสูง เพื่อที่จะลดปัญหาการประมาณที่ไม่ดีและการปรับเหมาะเกินไปเราจะต้องเลือกสถาปัตยกรรมของเครื่องที่เหมาะสม โดยอาศัยความรู้เกี่ยวกับลักษณะของปัญหา แล้วจึงหาฟังก์ชันในเครื่องนี้ที่สามารถลดค่าผิดพลาดในข้อมูลสอนให้ต่ำที่สุดได้

แนวทางที่สอง ให้คงค่าของความเสี่ยงเชิงทดลองให้คงที่ (อาจเป็นศูนย์) แล้วลดช่วงความเชื่อมั่นให้ต่ำสุด แนวทางนี้ทำโดยการกำหนดขอบเขตความเสี่ยงเชิงทดลองสูงสุดที่ยอมรับได้ แล้วเปลี่ยนรูปแบบของฟังก์ชันเพื่อลดช่วงความเชื่อมั่นให้ต่ำสุด แนวทางนี้พบในเอสวีเอ็ม โดยเอสวีเอ็มจะแบ่งกลุ่มของฟังก์ชันออกเป็นเซตย่อย แล้วหาเซตของเคอร์เนลที่ให้ความเสี่ยงเชิงทดลองต่ำสุด จากนั้นจึงหาฟังก์ชันในเซตนั้นที่มีมิติวีซีต่ำที่สุด แล้วบันทึกค่าขอบเขตความเสี่ยงที่ได้ ทำการพิจารณาเช่นเดิมกับเคอร์เนลกลุ่มถัดมาที่ให้ความเสี่ยงเชิงทดลองต่ำสุด ทำวนซ้ำจนได้ฟังก์ชันที่ให้ขอบเขตความผิดพลาดที่แท้จริงต่ำสุด

10.3 ระนาบหลายมิติแย่งแยกดีสุด

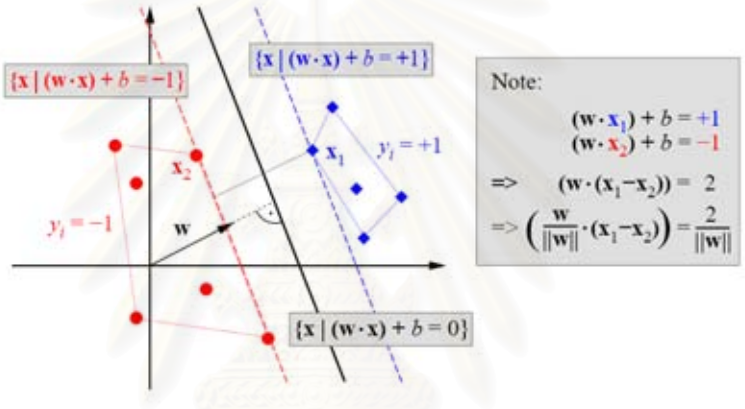
ในการออกแบบขั้นตอนการเรียนรู้ เราจะต้องใช้ฟังก์ชันที่สามารถคำนวณขีดความสามารถได้ ตัวจำแนกเอสวีเอ็มมีพื้นฐานจากฟังก์ชันประเภทระนาบหลายมิติ โดยที่ระนาบหลายมิติทำหน้าที่คล้ายเป็นตัวแยกเขตแดนระหว่างข้อมูลทั้งสองประเภท ดังสมการ

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0, \quad \mathbf{w} \in \mathrm{R}^{\mathrm{N}},\ b \in \mathrm{R}$$

และสอดคล้องกับฟังก์ชันตัดสินใจ

$$f(\mathbf{x}) = sign((\mathbf{w}\cdot\mathbf{x}) + b)$$

ระนาบหลายมิติแย่งแยกดีสุดมีนิยามเป็นระนาบที่มีระยะห่างของการแบ่งแยกระหว่างข้อมูลทั้งสองประเภทมากที่สุด เราพบว่าระยะห่างระหว่างกลุ่มข้อมูลทั้งสองประเภทที่เรียกว่าระยะขอบ (Margin) คือ $2/\|\mathbf{w}\|$ ดังแสดงในรูปที่ 2.23 โดยที่ $y(\mathbf{w}\cdot\mathbf{x}+b)\geq 1$ ต้องเป็นจริงด้วย ในรูปนี้จะมีข้อมูลที่สำคัญจริงๆ ซึ่งส่งผลต่อตำแหน่งของระนาบหลายมิติแบ่งแยกดีสุด ซึ่งคือ x_1 (ของตัวอย่างประเภท +1) และ x_2 กับ x_3 (ของตัวอย่างประเภท -1) เราเรียกข้อมูลทั้งสามตัวนี้ว่าซัพพอร์ตเวกเตอร์ (Support Vector) ซึ่งทำหน้าที่สนับสนุนการสร้างระนาบหลายมิติแบ่งแยกดีสุดนี้ ส่วนข้อมูลอื่นๆ แม้ว่าจะถูกตัดออกไป ก็จะไม่ส่งผลกระทบต่อการสร้างระนาบ



รูปที่ 2.23 ระนาบหลายมิติที่ใช้แยกดีสุดจะมีระยะห่างระหว่างข้อมูลทั้งสองกลุ่มเป็น $2/\|\mathbf{w}\|$

ปัญหาของการหาระนาบหลายมิติแบ่งแยกดีสุดนี้มีลากรองเกียนคือ

$$L(w,b,\alpha) = \tfrac{1}{2}\mathbf{w}\cdot\mathbf{w} - \sum\alpha_i((\mathbf{w}\cdot\mathbf{x}_i + b)y_i - 1)$$

ซึ่ง α คือตัวคูณลากรองจ์โดยต้องหาค่าต่ำสุดเมื่อเทียบกับ $\mathbf{w}, b$ และหาค่าสูงสุดเมื่อเทียบกับ $\alpha_i \geq 0$ ที่จุดที่ดีที่สุด คำตอบ $\mathbf{w}_0, b_0$ และ α_i^0 จะสอดคล้องกับคุณลักษณะของระนาบหลายมิติที่ดีที่สุดดังนี้

$$\sum\alpha_i^0 y_i = 0,\ \ \alpha_i^0 \geq 0,\ \ i = 1,...,l$$
$$\mathbf{w}_0 = \sum\alpha_i^0 y_i \mathbf{x}_i\ \ \ \alpha_i^0 \geq 0,\ \ i = 1,...,l$$

โดยที่จริงแล้วเฉพาะสัมประสิทธิ์ α_i^0 ของซัพเพอร์เวกเตอร์เท่านั้นที่ไม่เป็นศูนย์ ดังนั้นในการหาค่าของเวกเตอร์ $\mathbf{w}_0$ เราจึงหาจากผลรวมเชิงเส้นของข้อมูลตัวที่เป็นซัพพอร์ตเวกเตอร์เท่านั้น

$$b_0 = \frac{1}{2}\left[(\mathbf{w}_0\cdot\mathbf{x}^*(1)) + (\mathbf{w}_0\cdot\mathbf{x}^*(-1))\right]$$

โดยที่ $\mathbf{x}^*(1)$ คือซัพพอร์ตเวกเตอร์ใดๆ ที่อยู่ในประเภท +1 และ $\mathbf{x}^*(-1)$ คือซัพพอร์ตเวกเตอร์ใดๆที่อยู่ในประเภท -1 และจะได้ฟังก์ชันการตัดสินใจในรูปแบบ

$$f(\mathbf{x}) = sign(\sum_{SupportVector} \nu_i(\mathbf{w}_i \cdot \mathbf{x}_i) + b_0)$$

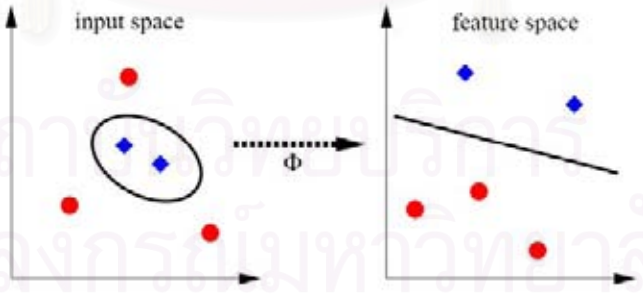
$$\nu_i = y_i \alpha_i^0$$

คำตอบนี้ใช้ได้เฉพาะกรณีที่สามารถแบ่งแยกแบบเชิงเส้นได้เท่านั้น แต่สำหรับกรณีที่แบ่งแยกแบบเชิงเส้นไม่ได้ ต้องมีการปรับข้อจำกัดเล็กน้อย คือ α_i จะมีขอบเขต $0 \le \alpha_i \le C$ โดยเราสามารถแปรค่า C ได้ โดยเป็นการแลกเปลี่ยนระหว่างความถูกต้องกับความเร็วในการสอน

สังเกตได้ว่าทั้งการแก้ปัญหาการหาฟังก์ชันของระนาบหลายมิติและตัวฟังก์ชันการตัดสินใจต่างก็ขึ้นอยู่กับผลคูณเชิงสเกลาร์ระหว่างเวกเตอร์ สิ่งนี้เองที่ทำให้เราสามารถขยายอัลกอริทึมสำหรับกรณีที่ไม่เป็นเชิงเส้นได้ในปริภูมิอันดับสูง

10.4 ปริภูมิระดับสูงและเคอร์เนล

การแมปข้อมูลเข้าไปสู่ปริภูมิคุณลักษณะที่มีระดับสูงขึ้น จะช่วยให้สามารถแยกข้อมูลสองประเภทออกจากกันได้โดยใช้ฟังก์ชันเชิงเส้น โดยมีสมมติฐานว่าข้อมูลที่มีความสัมพันธ์ไม่เป็นเชิงเส้นในปริภูมิอันดับต่ำ เมื่อแมปไปสู่ปริภูมิอันดับสูงจะมีความสัมพันธ์แบบเชิงเส้นได้ อย่างไรก็ตามการแมปไปสู่ปริภูมิอันดับสูงอาจต้องการการคำนวณที่สูงเกินไป ฟังก์ชันเคอร์เนลช่วยหลีกเลี่ยงปัญหาการคำนวณหาฟังก์ชันในการแมปได้ โดยยอมให้คำนวณผลคูณสเกลาร์ของตัวแปรสองตัวในปริภูมิอันดับสูงได้โดยไม่ต้องคำนวณหาฟังก์ชันที่ใช้แมปการแสดงด้วยสมการจะช่วยให้เข้าใจถึงแนวคิดนี้ได้ง่ายขึ้น



รูปที่ 2.24 แนวคิดการแมปแบบไม่เชิงเส้น

แนวคิดเบื้องต้นของเอสวีเอ็ม ดังแสดงในรูปที่ 2.24 เป็นการแมปข้อมูลไปสู่ปริภูมิผลคูณเชิงสเกลาร์อันดับสูงขึ้นไป (เขียนแทนด้วย F) ผ่านการแมปแบบไม่เชิงเส้น

$$\Phi : \mathrm{R}^{\mathrm{N}} \rightarrow \mathrm{F}$$

แล้วจึงใช้ขั้นตอนวิธีเชิงเส้นนี้ใน F ซึ่งสิ่งที่ต้องทำก็เพียงการหาค่าของผลคูณสเกลาร์

$$k(\mathbf{x},\mathbf{y}) := (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$$

ถ้า F มีอันดับสูง ย่อมเท่ากับว่าพจน์ด้านขวามือของสมการข้างต้นจะคำนวณได้ยากมาก อย่างไรก็ตามในบางกรณีจะมีเคอร์เนล k ที่ง่ายต่อการคำนวณตัวอย่างเช่นเคอร์เนลแบบพหุนาม

$$k(\mathbf{x},\mathbf{y}) := (\mathbf{x} \cdot \mathbf{y})^d$$

ซึ่งสามารถแสดงให้เห็นว่าสอดคล้องกับการแมป Φ ไปสู่ปริภูมิที่แสปนโดยผลคูณทั้งหมดของอันดับ d ใน R^{N} ตัวอย่างเช่นในกรณีที่ $d = 2$ และ $x, y \in \mathrm{R}^2$ ตัวอย่างเช่น เรามี

$$\begin{aligned}(\mathbf{x} \cdot \mathbf{y})^2 &= ((x_1, x_2) \cdot (y_1, y_2))^2 \\ &= ((x_1^{\,2}, \sqrt{2}x_1x_2, x_2^{\,2}) \cdot (y_1^{\,2}, \sqrt{2}y_1y_2, y_2^{\,2})) \\ &= (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))\end{aligned}$$

ที่นิยามให้ $\Phi(\mathbf{x}) = (x_1^{\,2}, \sqrt{2}x_1x_2, x_2^{\,2})$

นอกเหนือจากการใช้เคอร์เนลแบบพหุนามแล้วเอสวีเอ็มยังสามารถนำมาใช้กับเคอร์เนลแบบอาร์บีเอฟ (RBF) ซึ่งนิยามดังนี้คือ

$$k(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$$

และเคอร์เนลแบบซิกมอยด์ (ที่มีเกน κ และออฟเซต Θ)

$$k(\mathbf{x},\mathbf{y}) = \tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \Theta)$$

10.5 การเรียนรู้ซัพพอร์ตเวกเตอร์แมชชีน

เอสวีเอ็มเป็นเทคนิคการเรียนรู้ที่มีขีดความสามารถสูงสำหรับการจำแนกข้อมูลแบบสองประเภท โดยสามารถลดความเสี่ยงเชิงโครงสร้างให้ต่ำสุด เอสวีเอ็มใช้ประโยชน์จากการแมปและฟังก์ชันเคอร์เนล ซึ่งระนาบหลายมิติที่ใช้แยกประเภทข้อมูลจะอยู่ในปริภูมิอันดับสูง ถึงจุดนี้จะแทนที่ข้อมูลแต่ละตัวในชุดสอน $\mathbf{x}_i$ ด้วย $\Phi(\mathbf{x}_i)$ และหาระนาบหลายมิติแบ่งแยกดีสุดใน F เนื่องจากเราใช้เคอร์เนล ดังนั้นจึงได้ผลเป็นฟังก์ชันการตัดสินใจในรูปแบบ

$$f(\mathbf{x}) = sign(\sum_{SupportVector} \nu_i \bullet k(\mathbf{x}, \mathbf{x}_i) + b_0)$$

$$\nu_i = y_i \alpha_i^0$$

โดยที่พารามิเตอร์ ν_i สามารถคำนวณได้โดยเป็นคำตอบของปัญหาการโปรแกรมกำลังสอง (Quadratic Programming) โดยลดค่าของฟังก์ชันวัตถุประสงค์ให้ต่ำสุด

$$\frac{1}{2}\sum \alpha_i Q_j \alpha_j - \sum \alpha_i$$

โดยขึ้นกับข้อกำหนดต่อไปนี้

$$0 \le \alpha_i \le C$$
$$\sum y_i \alpha_i = 0$$

Q อยู่ในรูปของ $y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ เป็นเมทริกซ์มิติ $\mathrm{N} \times \mathrm{N}$ ซึ่งขึ้นอยู่กับขนาดของข้อมูลสอน $\mathbf{x}_i$ แบะฉลากของประเภท y_i กับรูปแบบของฟังก์ชันที่จะใช้ ส่วน C เป็นค่าคงที่ที่สามารถกำหนดได้ เมื่อเราลดค่าฟังก์ชันวัตถุประสงค์โดยการแก้ปัญหาการโปรแกรมกำลังสองแล้ว เราจะได้พารามิเตอร์ $\alpha_i, C, \mathbf{w}_0, b_0$ ที่ให้ฟังก์ชันต่ำสุด ซึ่งก็คือการเรียนรู้ของเอสวีเอ็ม

ในปริภูมิอันดับสูงจะได้ระนาบหลายมิติเป็นแบบเชิงเส้น ส่วนในปริภูมิอันดับต่ำระนาบหลายมิติจะสอดคล้องกับฟังก์ชันการตัดสินใจแบบไม่เชิงเส้นซึ่งรูปแบบจะถูกกำหนดโดยเคอร์เนล และโดยการเปลี่ยนเคอร์เนล เราจะได้สถาปัตยกรรมที่ต่างออกไป ดังเช่นตัวจำแนกแบบพหุนาม ตัวจำแนกแบบอาร์บีเอฟ และข่ายงานประสาทเทียมแบบสามระดับ เป็นต้น ในปัญหาต่างกัน เราต้องการเคอร์เนลที่อาจไม่เหมือนกัน แต่โดยปกติแล้วไม่ว่าจะใช้เคอร์เนลชนิดใด ก็มักได้ซัพพอร์ตเวกเตอร์ชุดที่ใกล้เคียงกัน สิ่งนี้สนับสนุนแนวคิดที่ว่าซัพพอร์ตเวกเตอร์เป็นคุณลักษณะเฉพาะสำหรับปัญหาหนึ่งๆ

งานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยส่วนใหญ่เกี่ยวกับการพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ตั้งแต่อดีตจนถึงปัจจุบันนั้น จะมุ่งเน้นการหาลำดับของหน่วยเสียงที่เรียงต่อเนื่องกันเป็นเส้นตรงเพียงลำดับเดียว (Linear Sequence of Segments) [17 – 21] อย่างไรก็ตามการจะนำลำดับของเซกเมนต์เพียงลำดับเดียวมาใช้กับระบบรู้จำเสียงแบบอาศัยเซกเมนต์นั้น ลำดับของเซกเมนต์จะต้องมีความถูกต้องแม่นยำ โดยแม้จะมีความผิดพลาดเพียงเล็กน้อย ก็อาจทำให้ผลของการรู้จำเสียงพูดผิดพลาดไปมาก ในปัจจุบันยังไม่มีวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่สามารถทำงานได้ถูกต้องสมบูรณ์แบบ งานวิจัยบางส่วนจึงมุ่งเน้นไปที่การแบ่งเสียงพูดเป็นเซกเมนต์ให้ผลลัพธ์ออกมาอยู่ในรูปกราฟของเซกเมนต์แทน [22] ในหัวข้อนี้จะนำเสนองานวิจัยเกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ ซึ่งสามารถแบ่งได้เป็นสามกลุ่มดังนี้

1. การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยการจำแนกเสียงพูดเป็นประเภทกว้าง

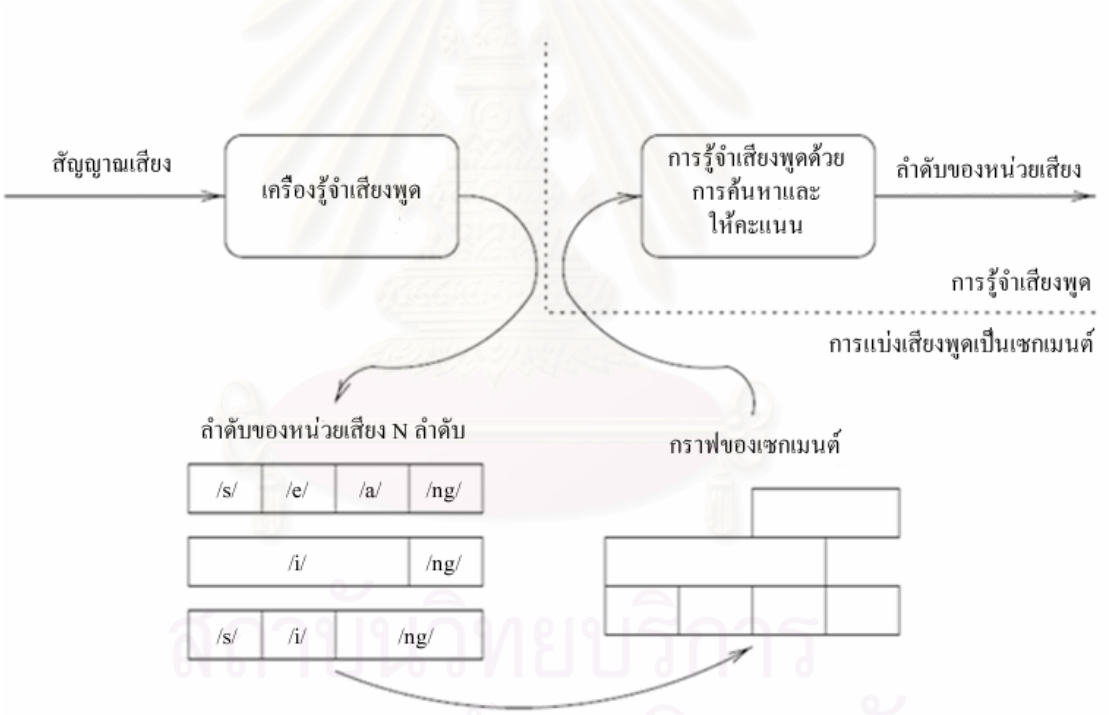
Cole และ Fanty [23] เสนอการแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้การจำแนกหน่วยเสียงในแต่กรอบเวลาสั้นๆ แบ่งหน่วยเสียงออกเป็นประเภทกว้างๆ (Segmentation using Broad-class Classification) ในการค้นหาขอบเขตของหน่วยเสียง รวมทั้งนำโครงข่ายประสาทเทียมมาใช้ในการจำแนกประเภทของเสียงในแต่ละกรอบเวลาออกเป็น 22 ประเภทเพื่อใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับนำไปใช้ในระบบรู้จำเสียงพูดตัวอักษรภาษาอังกฤษที่มีลักษณะของเสียงพูดแบบไม่ต่อเนื่อง โดยสามารถรู้จำอักษรในภาษาอังกฤษได้เปอร์เซ็นต์ความถูกต้องสูงถึง 95%

2. การแบ่งเสียงพูดเป็นเซกเมนต์จากการเปลี่ยนแปลงทางสัญญาณเสียง

การแบ่งเสียงพูดเป็นเซกเมนต์จากการเปลี่ยนแปลงทางสัญญาณเสียง (Acoustic Segmentation) จะใช้วิธีวัดจากปริมาณที่แสดงถึงการเปลี่ยนแปลงหรือความไม่ต่อเนื่องกันของสัญญาณเสียง (Acoustic Discontinuity) โดยอาศัยหลักการที่ว่าสัญญาณเสียงที่บริเวณขอบเขตของหน่วยเสียงจะมีความไม่ต่อเนื่องสูงกว่าบริเวณที่เป็นหน่วยเสียง Wang, Lu และ Zhang [24] เสนอวิธีในการแบ่งเสียงพูดเป็นเซกเมนต์โดยอาศัยการวัดปริมาณสำคัญๆอย่าง เวลาของแต่ละหน่วยเสียง (Duration), อัตราการออกเสียง (Rate of Speech: ROS) และลำดับของการออกเสียง (Phonetic Sequence) มาคำนวณเป็นปริมาณที่ใช้แสดงถึงความไม่ต่อเนื่องของเสียงเพื่อใช้ระบุหาขอบเขตของหน่วยเสียง ผลลัพธ์ที่ได้คือกราฟของเซกเมนต์ ที่ได้จากการเชื่อมกันของขอบเขตทั้งหมด โดยแม้ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบนี้จำทำงานได้อย่างรวดเร็ว แต่ขนาดของกราฟของเซกเมนต์ก็มีใหญ่มากเกินความจำเป็น

3. การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด [4] จะหาขอบเขตของหน่วยเสียงโดยอาศัยเครื่องรู้จำเสียงพูดในระดับหน่วยเสียงแบบอาศัยกรอบเวลา (Frame-based Phonetic Recognizer) มารู้จำเสียงพูดออกมาเป็นลำดับของหน่วยเสียง N ลำดับที่ดีที่สุด แล้วจึงนำเซกเมนต์ของหน่วยเสียงที่รู้จำได้มาประกอบรวมกันเป็นกราฟของเซกเมนต์ดังแสดงด้วยรูปที่ 2.25 โดยที่ N ซึ่งเป็นตัวแปรที่กำหนดจำนวนลำดับที่ดีที่สุดที่ต้องการรู้จำ จะเป็นตัวกำหนดขนาดของกราฟของเซกเมนต์อีกที เครื่องรู้จำเสียงพูดในที่นี้จะสร้างโดยอาศัยแบบจำลองฮิดเดนมาร์คอฟ หรืออาจใช้การค้นหาด้วยวิธีของไวเทอร์บิแบบไปข้างหน้า (Forward Viterbi) และวิธีของ A^* แบบย้อนกลับ (Backward A^*) ก็ได้ แม้การแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีการดังกล่าวจะสามารถทำงานได้ผลดี แต่ก็มีข้อเสียอยู่ที่ต้องใช้การคำนวณนาน ไม่เหมาะกับระบบรู้จำเสียงแบบที่ต้องการความรวดเร็ว



รูปที่ 2.25 การแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

บทที่ 3
การแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้สารสนเทศสวนสัทศาสตร์

บทนี้จะนำเสนอเกี่ยวกับการแบ่งเสียงพูดเป็นเซกเมนต์ โดยจะเริ่มจากภาพรวมของการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ การตรวจหาขอบเขตของหน่วยเสียง การสร้างกราฟของเซกเมนต์ และการให้คะแนนขอบเขตของหน่วยเสียง ตามลำดับ

ภาพรวมของการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์

การแบ่งเสียงพูดเป็นเซกเมนต์ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้นประกอบไปด้วยขั้นตอนย่อยสองขั้นตอนคือ การตรวจหาขอบเขตของหน่วยเสียง และการสร้างกราฟของเซกเมนต์ ดังแสดงด้วยแผนภาพในรูปที่ 3.1 โดยการทำงานจะเริ่มต้นจากการป้อนสัญญาณเสียงพูดเข้าไปเป็นข้อมูลอินพุตของการตรวจหาขอบเขตของหน่วยเสียง ได้เอาท์พุตออกมาเป็นเซตหรือลำดับของขอบเขตของหน่วยเสียงที่เป็นขอบเขตของหน่วยเสียง แล้วผ่านเอาท์พุตนี้ไปเป็นอินพุตของการสร้างกราฟของเซกเมนต์ต่อไป สรุปสุดท้ายได้ผลลัพธ์ออกมาเป็นกราฟของเซกเมนต์สำหรับนำไปใช้ต่อในขั้นตอนการรู้จำเสียงพูดต่อไป

รูปที่ 3.1 แผนภาพแสดงส่วนประกอบของกระบวนการแบ่งเสียงพูดเป็นเซกเมนต์

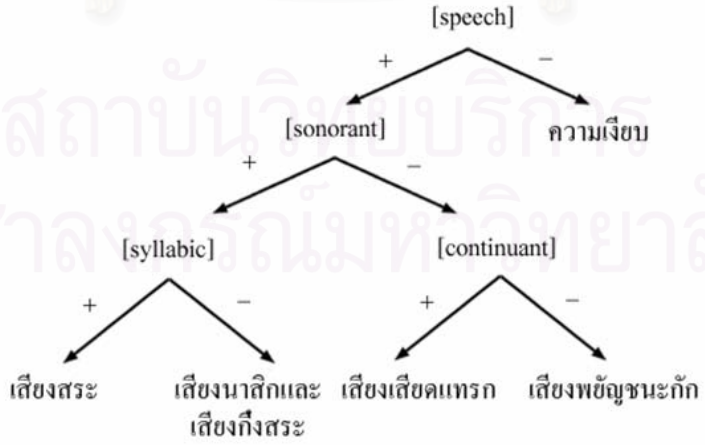
ทั้งการตรวจหาขอบเขตของหน่วยเสียง และการสร้างกราฟของเซกเมนต์นั้นสามารถทำได้หลายวิธี ซึ่งคุณภาพของขอบเขตของหน่วยเสียงและกราฟของเซกเมนต์ที่ได้ จะส่งผลโดยตรงต่อประสิทธิภาพในการรู้จำเสียงพูด งานวิจัยนี้จึงพัฒนาวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ โดยนำสารสนเทศสวนสัทศาสตร์มาช่วยเพิ่มประสิทธิภาพในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง รวมถึงเสนอวิธีการสร้างกราฟของเซกเมนต์แบบต่างๆ เพื่อให้ได้ขอบเขตของหน่วยเสียงและเซกเมนต์กราฟที่มีคุณภาพมากขึ้น อีกทั้งยังสามารถทำงานได้รวดเร็วแบบทันกาลอยู่ในระดับที่นำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ได้

การตรวจหาขอบเขตของหน่วยเสียง

จากความรู้ทางสัทศาสตร์ที่ว่า หน่วยเสียงประเภทเสียงพยัญชนะกัก เสียงพยัญชนะเสียดแทรก เสียงพยัญชนะนาสิก และเสียงสระ จะมีลักษณะการออกเสียงที่แตกต่างกัน ซึ่งสามารถอธิบายได้ด้วยสัทลักษณ์แสดงความมีคุณสมบัติเกี่ยวกับลักษณะการออกเสียงนั้นๆ ดังนั้นตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียง ก็จะเป็นขอบเขตของหน่วยเสียงเสมอ ดังนั้นในวิทยานิพนธ์นี้จะเสนอวิธีการตรวจหาขอบเขตของหน่วยเสียง โดยอาศัยผลของการจำแนกลักษณะการออกเสียง โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบซัพพอร์ตเวกเตอร์แมชชีน และการสกัดลักษณะสำคัญโดยใช้สารสนเทศสวนสัทศาสตร์เพื่อเพิ่มประสิทธิภาพในการจำแนกลักษณะการออกเสียง

การวิเคราะห์และอธิบายคุณสมบัติของเสียงพูดได้จะอาศัยสัทลักษณ์ ดังที่กล่าวไว้ในบทที่ 2 จากความสัมพันธ์ระหว่างสัทลักษณ์ลักษณะการออกเสียง กับประเภทของหน่วยเสียงในตารางที่ 2.3 เราสามารถแสดงให้อยู่ในรูปของโครงสร้างของสัทลักษณ์ลักษณะการออกเสียงเป็นลำดับชั้นได้ดังรูปที่ 3.2

ด้วยโครงสร้างแบบนี้ เราสามารถจำแนกเสียงพูดในแต่ละกรอบเวลาตามลักษณะการออกเสียงได้โดยการแบ่งแยกเสียงพูดที่มีและไม่มีสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบออกจากกันเป็นลำดับชั้น เพื่อระบุว่าสัญญาณเสียงนั้นมีลักษณะการออกเสียงเป็นแบบใด ต่อจากนั้นจึงค่อยพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียงให้เป็นขอบเขตของหน่วยเสียงต่อไป ซึ่งวิธีการนี้มีข้อได้เปรียบตรงที่สามารถเลือกใช้ค่าลักษณะสำคัญที่เหมาะสมต่อการจำแนกสัทลักษณ์แต่ละแบบแตกต่างกันได้



รูปที่ 3.2 โครงสร้างลำดับชั้นของสัทลักษณ์ลักษณะการออกเสียง

ปัญหาการจำแนกลักษณะการออกเสียงนี้ สามารถแตกย่อยเป็นปัญหาการแบ่งแยกข้อมูลสองกลุ่มออกจากกันคือข้อมูลเสียงที่มีและไม่มีสัทลักษณ์ลักษณะการออกเสียง ซึ่งเราสามารถนำเอสวีเอ็มมาใช้ช่วยแก้ปัญหานี้ได้ และเมื่อพิจารณาจากโครงสร้างลำดับชั้นจะเห็นว่าจะต้องใช้ตัวแบ่งแยกเอสวีเอ็มทั้งสิ้น 4 ตัวได้แก่ ตัวแบ่งแยกสัทลักษณ์ [sonorant] ตัวแบ่งแยกสัทลักษณ์ [syllabic] ตัวแบ่งแยกสัทลักษณ์ [continuant] และตัวแบ่งแยกเสียงพูดกับความเงียบซึ่งแทนด้วยสัญลักษณ์ [speech]

ในหัวข้อนี้จึงจะนำเสนอเกี่ยวกับการสกัดลักษณะสำคัญของเสียง กระบวนการเรียนรู้การจำแนกลักษณะการออกเสียง และกระบวนการตรวจหาขอบเขตของหน่วยเสียงจากผลการจำแนกลักษณะการออกเสียง โดยมีรายละเอียดดังต่อไปนี้

1. การสกัดลักษณะสำคัญของเสียง

ลักษณะสำคัญของเสียงเพื่อการเรียนรู้ในที่นี้แบ่งออกเป็นสองส่วนคือ 1) สัมประสิทธิ์เซปสตรัมบนสเกลเมล 2) ลักษณะสำคัญอื่นๆที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ ซึ่งมีรายละเอียดดังต่อไปนี้

1.1 สัมประสิทธิ์เซปสตรัมบนสเกลเมล

ลักษณะสำคัญของเสียงพูดที่ใช้ในที่นี้ได้แก่ สัมประสิทธิ์เซปสตรัมบนสเกลเมลซึ่งมีค่าพลังงานรวมอยู่ด้วย อัตราการเปลี่ยนแปลง (Delta) และความเร่ง (Accelerations) จากสัญญาณเสียงทุกๆกรอบเวลายาว 25 มิลลิวินาที โดยแต่ละกรอบเวลาจะมีระยะเวลาห่างกัน 10 มิลลิวินาที ได้ออกมาเป็นเวกเตอร์ลักษณะสำคัญที่มีขนาด 39 มิติ โดยต่อจากนี้จะขออ้างอิงชุดลักษณะสำคัญที่ด้วยสัญลักษณ์ MFCC_E_D_A

1.2 ลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์

สารสนเทศสวนสัทศาสตร์คือข้อมูลที่ซึ่งเกี่ยวกับความสัมพันธ์ระหว่างลักษณะของสัญญาณกับสัทลักษณ์ ของเสียงพูด โดยเมื่อเราเข้าลักษณะความสัมพันธ์นี้ ก็จะช่วยให้เราเลือกลักษณะสำคัญที่สามารถแบ่งเสียงพูดที่มีสัทลักษณ์แตกต่างกันได้ โดยสารสนเทศสวนสัทศาสตร์ที่เกี่ยวข้องกับเสียงพูดที่มีสัทลักษณ์ลักษณะการออกเสียง [sonorant] [syllabic] และ [continuant] มีรายละเอียดดังนี้

1. เสียงที่มีสัทลักษณ์ [sonorant] เป็นเสียงที่เกิดจากอากาศไหลผ่านช่องปากไปโดยไม่ได้ถูกกักเอาไว้เพียงพอต่อการสร้างเสียงรบกวนหรือกักการไหลของอากาศ ทำให้ได้ยินเป็นเสียงดัง มีพลังงานมากในช่วงความถี่ต่ำ ส่วนใหญ่จะเป็นเสียงสระ เสียงพยัญชนะนาสิก หรือเสียงกึ่งสระ โดยจะมีระดับพลังงานสูงในช่วงความถี่ต่ำ
2. เสียงที่มีสัทลักษณ์ความเป็น [syllabic] เป็นเสียงที่ไม่มีการกักลมไว้ที่บริเวณช่องปากและโพรงจมูกทำให้ไม่มีการสูญเสียพลังงานในระหว่างการเปล่งเสียงซึ่งก็คือลักษณะของเสียงสระ โดยมีลักษณะเป็นเสียงดังที่เป็นเสียงก้องเพราะมีการสั่นสะเทือนเส้นเสียงมาก ทำให้มีสัญญาณเสียงลักษณะเป็นคาบ และมีระดับพลังงานมากในช่วงความถี่ต่ำถึงปานกลาง
3. เสียงที่มีสัทลักษณ์ความเป็น [continuant] เป็นเสียงที่มีการกักเสียงเอาไว้โดยที่ช่องทางเดินเสียงอยู่ในสภาพปิดอยู่แต่ยังปิดไม่สมบูรณ์เป็นเหมือนช่องแคบ ตัวอย่างเช่น เสียง /ส/ ที่ใช้ฟันในการกักไม่ให้อากาศไหลผ่านช่องปากได้อย่างเต็มที่ทำให้เกิดเป็นเสียงเสียดแทรก มีลักษณะของเสียงที่ไม่มีก้องเพราะไม่มีการสั่นสะเทือนของเส้นเสียง ลักษณะของสัญญาณคล้ายกับสัญญาณรบกวนซึ่งเกิดจากกระแสลมถูกขับผ่านช่องแคบ พลังงานของสัญญาณนี้จะหนาแน่นที่ช่วงความถี่ใดนั้นขึ้นอยู่กับตำแหน่งของช่องแคบที่ใช้ในการสร้างเสียงเสียดแทรกนั้นๆ แต่โดยมากแล้วจะมีระดับพลังงานจะหนาแน่นมากที่ช่วงความถี่ตั้งแต่ความถี่ปานกลางไปจนถึงสูง ส่วนเสียงพยัญชนะกักซึ่งไม่มีสัทลักษณ์ความเป็น [continuant] จะเป็นสัญญาณเสียงที่พลังงานตลอดทั้งช่วงความถี่หายไป ซึ่งการที่พลังงานหายไปนี้เกิดจากการสร้างช่องปิดสนิท แต่อาจจะมีพลังงานที่ความถี่ต่ำอันเกิดจากการสั่นของเส้นเสียงในขณะที่เกิดช่องปิด โดยพลังงานที่ถูกส่งผ่านออกมาในอากาศนี้ ผ่านออกมาจากการแผ่รังสีจากกระพุ้งแก้ม มิได้เกิดจากลมจากช่องปากโดยตรง โดยในขณะที่ช่องปิดถูกปล่อยออกอย่างรวดเร็ว อาจเกิดสัญญาณรบกวนสั้นๆ ซึ่งสังเกตได้จากพลังงานที่มีรูปร่าง ยาวออกไปทางแนวตั้ง ในสเปกโตรแกรมหลังจากช่วงที่เป็นช่องปิด

จากสารสนเทศสวนสัทศาสตร์ที่กล่าวมานี้ ผู้วิจัยจึงเลือกใช้ค่าระดับพลังงานที่ช่วงความถี่ต่างๆ ระดับความไม่เป็นคาบของสัญญาณเสียง (Aperiodicity degree) และระดับการสั่นสะเทือนเส้นเสียง (Voicing degree) หรือความก้องของเสียงซึ่งคำนวณมาจากค่าการกระจายของพลังงาน (Energy distribution) อัตราการตัดศูนย์ (Zero-crossing rate) และอัตสหสัมพันธ์ (Autocorrelation) มาเป็นลักษณะสำคัญในการจำแนกสัทลักษณ์ลักษณะการออกเสียง โดยออกแบบเป็นเซตของลักษณะสำคัญสำหรับการจำแนกสัทลักษณ์แต่ละแบบได้ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 สารสนเทศสวนสัทศาสตร์และค่าลักษณะสำคัญที่ใช้จำแนกสัทลักษณ์แต่ละประเภท

1 E[a, b] พลังงานในช่วงความถี่ตั้งแต่ a จึงถึง b (หน่วยเฮิร์ต)

สัทลักษณ์	สารสนเทศสวนสัทศาสตร์	ลักษณะสำคัญ
[sonorant]	มีพลังงานมากในช่วงความถี่ต่ำ	E[100,400] E[0,2000] / E[2000,8000]
[syllabic]	มีพลังงานมากในช่วงความถี่ปานกลาง สัญญาณมีลักษณะเป็นคาบ	E[640,2800] E[2000,3000] Voicing degree
[continuant]	มีสัญญาณรบกวนและมีความปั่นป่วนของสัญญาณในช่วงความถี่ตั้งแต่ปานกลางไปจนถึงความถี่สูง	E[2000,8000] Aperiodicity degree

การสกัดค่าลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ในที่นี้ จะต้องอาศัยค่าพลังงาน อัตสหสัมพันธ์ และอัตราการตัดศูนย์มาคำนวณเป็นระดับความก้องของเสียง และระดับความไม่เป็นคาบของสัญญาณเสียง ในหัวข้อต่อไปนี้จะนำเสนอเกี่ยวกับการหาค่าลักษณะสำคัญดังกล่าวโดยมีรายละเอียดดังต่อไปนี้

1.2.1 ค่าพลังงาน

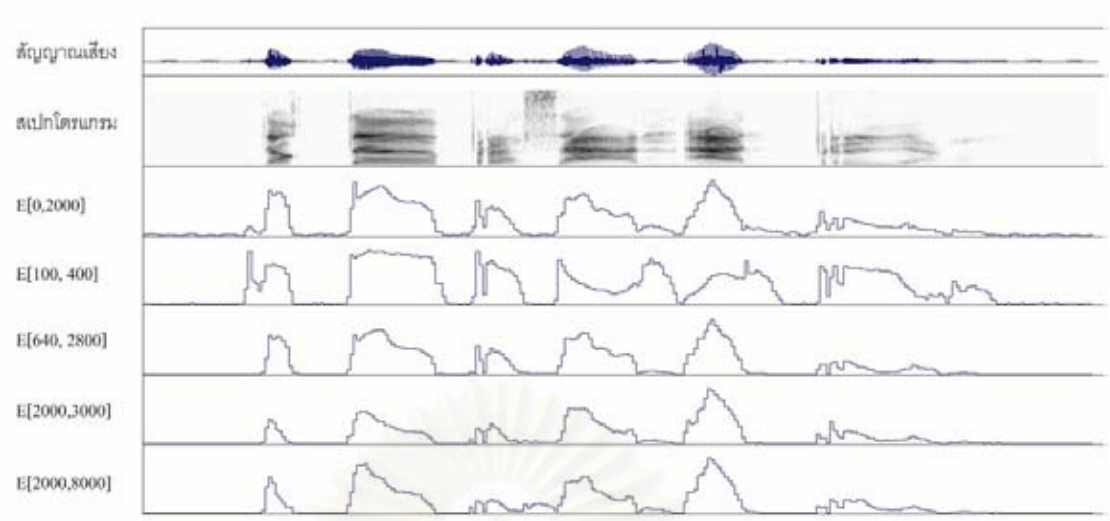
พลังงานของสัญญาณเป็นค่าลักษณะสำคัญที่นิยมนำมาใช้วิเคราะห์สัญญาณเสียง โดยค่าพลังงานของสัญญาณ $s(n)$ ใดๆที่แปรตามเวลาสามารถนิยามได้ว่า

$$E = \sum_{n=-\infty}^{+\infty} s^2(n)$$

โดยในการสกัดค่าพลังงานของสัญญาณเสียงจะต้องแบ่งสัญญาณออกมาพิจารณาเป็นกรอบเวลาด้วยฟังก์ชันหน้าต่าง $w(m)$ ที่มีขนาด N ดังนั้นค่าพลังงานของสัญญาณเสียงที่กรอบเวลาที่ m เขียนแทนด้วยสัญลักษณ์ $E(m)$ จะสามารถคำนวณได้ดังสมการต่อไปนี้

$$E(m) = \sum_{n=0}^{N-1} [w(m)s(m-n)]^2$$

และในการหาค่าพลังงานในช่วงความถี่ต่างๆ จะทำการกรองความถี่ของสัญญาณเสียงโดยใช้ตัวกรองความถี่ชนิดแถบความถี่ผ่าน (Band-pass filter) ซึ่งจะยอมให้สัญญาณที่มีความถี่ในขอบเขตที่กำหนดผ่านไปได้และจะกรองสัญญาณที่ความถี่ส่วนอื่นๆทิ้ง แล้วจึงนำสัญญาณที่กรองได้นี้ไปคำนวณค่าพลังงานของสัญญาณเสียงในช่วงความถี่ที่ต้องการดังแสดงได้ด้วยรูปที่ 3.3



รูปที่ 3.3 ค่าพลังงานของสัญญาณเสียงบนช่วงความถี่ต่างๆ

1.2.2 อัตราการตัดศูนย์

ค่าอัตราการตัดศูนย์ของสัญญาณเสียง คือจำนวนครั้งของการเปลี่ยนแปลงเครื่องหมายจากบวกเป็นลบหรือจากลบเป็นบวกของแอมพลิจูดของสัญญาณเสียงต่อหนึ่งหน่วยเวลา ค่าลักษณะสำคัญนี้นิยมใช้กันมากในระบบรู้จำเสียงพูด โดยสามารถนำมาใช้วัดระดับเสียงรบกวนหรือนำมาไปประยุกต์วัดระดับความก้องของสัญญาณเสียงได้ อัตราการตัดศูนย์สามารถคำนวณได้จากสมการดังต่อไปนี้

$$zcr = \frac{1}{T}\sum_{t=0}^{T-1}\prod\{s_t s_{t-1} < 0\}$$

เมื่อกำหนดให้ $zcr(m)$ คืออัตราการตัดศูนย์ของสัญญาณเสียง s ที่มีความยาว T และ $\prod\{A\}$ จะเป็นฟังก์ชันที่ให้ค่า 1 เมื่อประพจน์ A มีค่าความจริงเป็นจริง ในที่นี้คือประพจน์ที่บอกเงื่อนไขการตัดศูนย์ของสัญญาณเสียงที่เวลา t

1.2.3 อัตสหสัมพันธ์ (Autocorrelation coefficients)

ค่าอัตสหสัมพันธ์เป็นค่าลักษณะสำคัญที่สามารถนำมาใช้วัดระดับความเป็นคาบของสัญญาณเสียง โดยพื้นฐานแล้วค่าอัตสหสัมพันธ์สามารถคำนวณได้จากการเปรียบเทียบสัญญาณเสียงที่เวลาหนึ่งกับสัญญาณเสียงเดียวกันที่ตำแหน่งซึ่งเยื้องหรือหน่วงเวลาออกไปดังแสดงด้วยรูปที่ 3.4 จากกราฟด้านบนของรูปจะเป็นสัญญาณเสียงสระซึ่งมีลักษณะเป็นคาบ ซึ่งการหา

ค่าอัตสหสัมพันธ์จะพิจารณาสัญญาณเสียงที่ตำแหน่งที่ต้องการหาค่าอัตสหสัมพันธ์ว่ามีความคล้ายกันกับสัญญาณเสียงที่ตำแหน่งที่มีการหน่วงเวลาไปมากน้อยแค่ไหน

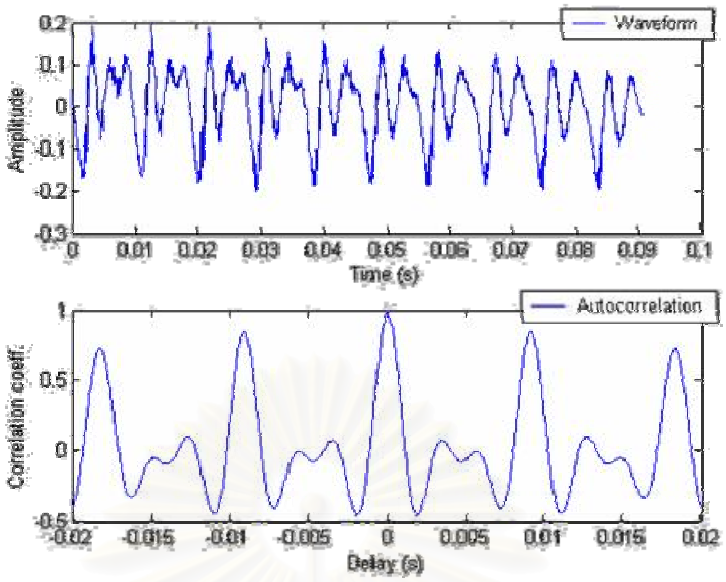
ค่าอัตสหสัมพันธ์ $autocorr(m,k)$ ของสัญญาณเสียงที่กรอบเวลาที่ m เมื่อหน่วงเวลาไปเป็นระยะเวลา k สามารถหาได้จากสมการต่อไปนี้

$$autocorr(m,k) = \frac{1}{N\sigma^2}\sum_{n=1}^{N}\left[s(n)w(m-n)-\mu\right]\left[s(n+k)w(m-n+k)-\mu\right]$$

โดยที่ $s(n)$ คือสัญญาณเสียงตำแหน่งที่ n และ $w(m)$ คือฟังก์ชันหน้าต่างที่มีขนาดความกว้าง N ส่วน μ และ σ คือค่าเฉลี่ยและค่าความแปรปรวนของ $s(n)$ ตามลำดับ

รูปที่ 3.4 สัญญาณเสียงสระที่ระยะเวลาหน่วงต่างๆ

สัญญาณเสียงที่มีลักษณะเป็นคาบจะมีค่าอัตสหสัมพันธ์สูงเมื่อหน่วงเวลาไปเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น ดังรูปที่ 3.5 จากกราฟในรูปจะสังเกตเห็นว่าค่าอัตสหสัมพันธ์จะมีค่าสูงสุดเมื่อไม่มีการหน่วงเวลา และค่าอัตสหสัมพันธ์สูงสุดอันดับต่อๆไปจะอยู่ที่ระยะหน่วงเวลาซึ่งเป็นจำนวนเท่าของคาบของสัญญาณเสียงนั้น



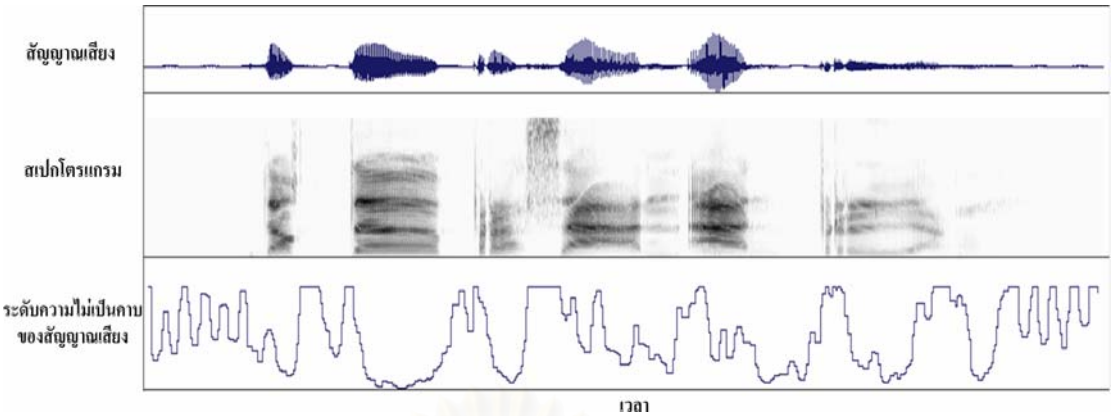
รูปที่ 3.5 สัญญาณเสียงที่มีลักษณะเป็นคาบ และค่าอัตสหสัมพันธ์ที่ระยะเวลาหน่วงต่างๆ

1.2.4 ค่าระดับความไม่เป็นคาบของสัญญาณเสียง

การประมาณระดับความไม่เป็นคาบของสัญญาณเสียงที่เวลาใดๆจะคำนวณโดยอาศัยค่าอัตสหสัมพันธ์ดังที่กล่าวไว้ในหัวข้อที่แล้ว โดยในที่นี้ค่าระดับความไม่เป็นคาบของสัญญาณเสียงในกรอบเวลาที่ m เขียนแทนด้วยสัญลักษณ์ $aperiodicity(m)$ จะคำนวณได้จากอัตราส่วนระหว่างค่าอัตสหสัมพันธ์ต่ำสุดและค่าอัตสหสัมพันธ์เมื่อไม่มีการหน่วงเวลาซึ่งเป็นค่ามากสุด ได้ดังสมการต่อไปนี้

$$aperiodicity(m) = \frac{autocorr(m, k_{min})}{autocorr(m, 0)}$$

เมื่อกำหนดให้ $autocorr(m, k_{min})$ คือค่าอัตสหสัมพันธ์ที่มีค่าน้อยที่สุดที่มีระยะหน่วงเวลาเป็น k_{min} ซึ่งในที่นี้จะพิจารณาหาค่าอัตสหสัมพันธ์ที่มีค่าน้อยสุดในช่วงระยะหน่วงเวลาที่ความถี่ตั้งแต่ 50 ถึง 400 เฮิร์ตซึ่งเป็นช่วงความถี่ต่ำ ดังรูปที่ 3.6 จากรูปส่วนของสัญญาณเสียงเสียดแทรกซึ่งมีลักษณะของสัญญาณคล้ายเสียงรบกวนไม่เป็นคาบ ค่าอัตสหสัมพันธ์ที่คำนวณออกมาได้จะมีค่ามาก



รูปที่ 3.6 ระดับความไม่เป็นคาบของสัญญาณเสียง

1.2.5 ค่าระดับความก้องของเสียง

สัญญาณเสียงที่มีระดับความก้องสูงจะมีลักษณะสัญญาณเป็นคาบ มีค่าพลังงานสูง และมีอัตราการตัดศูนย์ต่ำ ในที่นี้การประมาณระดับความก้องของสัญญาณเสียงที่กรอบเวลา m เขียนแทนด้วยสัญลักษณ์ $vdegree(m)$ จะคำนวณโดยอาศัยค่าลักษณะสำคัญสามอย่างได้แก่ ค่าอัตสหสัมพันธ์ ค่าพลังงาน และอัตราการตัดศูนย์ของสัญญาณเสียง โดยคำนวณค่าระดับความก้องของสัญญาณเสียงจากอัตราส่วนระหว่าง ผลคูณของคะแนนที่ได้จากค่าอัตสหสัมพันธ์ คะแนนที่ได้จากค่าพลังงาน และคะแนนที่ได้จากอัตราการตัดศูนย์ เทียบกับคะแนนที่มีค่าต่ำสุด ดังสมการต่อไปนี้

$$vdegree(m) = \frac{ap(m) \times ep(m) \times zp(m)}{\min\{ap(m), zp(m), ep(m)\}}$$

โดยที่ $ap(m)$ คือคะแนนที่ได้จากค่าอัตสหสัมพันธ์ $ep(m)$ คือคะแนนที่ได้จากค่าพลังงาน และ $zp(m)$ คือคะแนนที่ได้จากอัตราการตัดศูนย์คะแนนทั้งสามจะคำนวณจากสัญญาณเสียงที่กรอบเวลา m และจะมีค่าอยู่ในช่วง 0 ถึง 1 โดยคะแนนที่ได้จากค่าค่าอัตสหสัมพันธ์จะคำนวณจากสมการต่อไปนี้

$$ap(m) = \frac{1}{1 + e^{-\left(\frac{autocorr(m) - 0.75}{0.1}\right)}}$$

เมื่อกำหนดให้ $autocorr(m)$ คือค่าอัตสหสัมพันธ์ที่มีค่ามากที่สุดที่มีระยะหน่วงเวลาที่ความถี่ตั้งแต่ 50 ถึง 400 เฮิร์ต โดยถ้าคะแนนของอัตสัมพันธ์ที่คำนวณออกมามีค่ามากก็จะสะท้อนถึงลักษณะความเป็นคาบ ซึ่งเป็นหนึ่งในลักษณะของสัญญาณเสียงที่มีความก้อง

ค่าคะแนนที่ได้จากค่าพลังงานของสัญญาณเสียงจะคำนวณจากสมการดังนี้

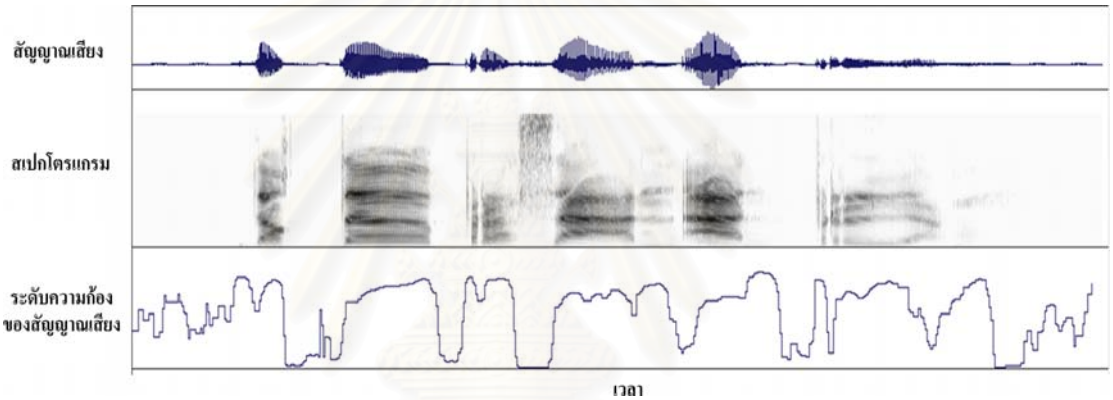
$$ep(m) = \frac{1}{1 + e^{-\left(\frac{E(m) - maxenergy}{5}\right)}}$$

เมื่อกำหนดให้ $E(m)$ คือค่าพลังงานที่กรอบเวลา m และ *maxenergy* คือค่าพลังงานมากสุดของสัญญาณเสียงนั้น โดยถ้าคะแนนของค่าพลังงานที่คำนวณออกมามีค่ามากก็จะหมายความว่าสัญญาณเสียงนั้นมีพลังงานมาก ซึ่งเป็นหนึ่งในลักษณะของสัญญาณเสียงที่มีความก้อง

ค่าคะแนนที่ได้จากอัตราการตัดศูนย์จะคำนวณจากสมการต่อไปนี้

$$zp(m) = \frac{1}{1+e^{\left(\frac{zcr(m)-1000}{200}\right)}}$$

เมื่อกำหนดให้ $zcr(m)$ คืออัตราการตัดศูนย์ของสัญญาณเสียงที่กรอบเวลา m เมื่อสัญญาณเสียงมีอัตราการตัดศูนย์มากๆคือสัญญาณเสียงที่มีลักษณะคล้ายเสียงรบกวนหรือเสียงเสียดแทรกซึ่งไม่เป็นลักษณะของเสียงก้อง ค่าคะแนนนี้ก็จะมีค่าเข้าใกล้ 0



รูปที่ 3.7 ระดับความก้องของสัญญาณเสียง

2. การเรียนรู้การจำแนกลักษณะการออกเสียง

การเรียนรู้การจำแนกลักษณะการออกเสียงในที่นี้ ใช้วิธีการเรียนรู้ของเครื่องแบบซัพพอร์ตเวกเตอร์แมชชีนดังที่กล่าวไว้แล้วในบทที่ 2 โดยอาศัยลักษณะสำคัญที่ได้จากขั้นตอนที่แล้ว เป็นตัวอย่างข้อมูลในการเรียนรู้ โดยนำมาใช้ในการเรียนรู้แบบจำลองการแบ่งแยกสัทลักษณ์สี่แบบได้แก่ [speech] [sonorant] [syllabic] และ [continuant] หน่วยเสียงที่มีและไม่มีสัทลักษณ์ลักษณะการออกเสียง ตามประเภทต่างๆ จะถูกแบ่งกลุ่มไว้ดังตารางที่ 3.2

การเรียนรู้ของสัทลักษณ์ [sonorant] จะให้เวกเตอร์ลักษณะสำคัญที่สกัดจากเสียงสระ เสียงกึ่งสระ และเสียงนาสิกมีฉลากเป็น +1 และให้เวกเตอร์ลักษณะสำคัญที่สกัดจากเสียงพยัญชนะกักและเสียงเสียดแทรกมีฉลากเป็น -1 การเรียนรู้ของสัทลักษณ์ [syllabic] จะให้เวกเตอร์ลักษณะสำคัญที่สกัดจากเสียงสระ มีฉลากเป็น +1 และให้เวกเตอร์ลักษณะสำคัญที่สกัดจากเสียงกึ่งสระ และเสียงนาสิกมีฉลากเป็น -1 การเรียนรู้ของสัทลักษณ์ [continuant] จะให้เวกเตอร์ลักษณะสำคัญที่สกัดจาก

เสียงพยัญชนะกักมีฉลากเป็น +1 และให้เวกเตอร์ลักษณะสำคัญที่สกัดจากเสียงเสียดแทรกมีฉลากเป็น -1 โดยจะสุ่มเลือกข้อมูลมาสร้างเวกเตอร์ลักษณะสำคัญที่มีฉลาก +1 และ -1 อย่างละ 10,000 ตัวอย่าง โดยในที่นี้จะเรียนรู้เครื่องจำแนกลักษณะการออกเสียงในที่นี้ จะเรียนรู้ซัพพอร์ตเวกเตอร์แมชชีนโดยใช้ฟังก์ชันเคอร์เนลสองแบบ คือซัพพอร์ตเวกเตอร์แมชชีนแบบที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้น และซัพพอร์ตเวกเตอร์แมชชีนแบบที่ใช้ฟังก์ชันเคอร์เนลพหุนาม และต่อจากนี้ไปจะใช้สัญลักษณ์ SVM_{linear} และ $SVM_{polynomial}$ แทนเครื่องจำแนกลักษณะการออกเสียงที่ใช้เอสวีเอ็มที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้น และที่เครื่องจำแนกลักษณะการออกเสียงที่ใช้เอสวีเอ็มที่ใช้ฟังก์ชันเคอร์เนลแบบพหุนาม ตามลำดับ

ตารางที่ 3.2 หน่วยเสียงที่มีและไม่มีคุณสมบัติตามสัทลักษณ์แบ่งตามลักษณะการออกเสียง

สัทลักษณ์	หน่วยเสียง[1]	
	มีคุณสมบัติตามสัทลักษณ์ (+)	ไม่มีคุณสมบัติตามสัทลักษณ์ (−)
[sonorant]	/i/ /iː/ /e/ /eː/ /æ/ /æː/ /ɨ/ /ɨː/ /ɜ/ /ɜː/ /a/ /aː/ /u/ /uː/ /o/ /oː/ /ɔ/ /ɔː/ /ia/ /iːa/ /ɨa/ /ɨːa/ /ua/ /uːa/ /m/ /n/ /ŋ/ /r/ /l/ /w/ /j/	/p/ /pr/ /pl/ /ph/ /phr/ /phl/ /b/ /br/ /bl/ /t/ /tr/ /th/ /thr/ /d/ /dr/ /k/ /kr/ /kl/ /kw/ /kh/ /khr/ /khl/ /khw/ /ʔ/ /c/ /ch/ /s/ /h/ /f/ /fr/ /fl/
[syllabic]	/i/ /iː/ /e/ /eː/ /æ/ /æː/ /ɨ/ /ɨː/ /ɜ/ /ɜː/ /a/ /aː/ /u/ /uː/ /o/ /oː/ /ɔ/ /ɔː/ /ia/ /iːa/ /ɨa/ /ɨːa/ /ua/ /uːa/	/m/ /n/ /ŋ/ /r/ /l/ /w/
[continuant]	/c/ /ch/ /f/ /fr/ /fl/ /s/ /h/	/p/ /pr/ /pl/ /ph/ /phr/ /phl/ /b/ /br/ /bl/ /t/ /tr/ /th/ /thr/ /d/ /dr/ /k/ /kr/ /kl/ /kw/ /kh/ /khr/ /khl/ /khw/
[speech]	/i/ /iː/ /e/ /eː/ /æ/ /æː/ /ɨ/ /ɨː/ /ɜ/ /ɜː/ /a/ /aː/ /u/ /uː/ /o/ /oː/ /ɔ/ /ɔː/ /ia/ /iːa/ /ɨa/ /ɨːa/ /ua/ /uːa/ /p/ /pr/ /pl/ /ph/ /phr/ /phl/ /b/ /br/ /bl/ /t/ /tr/ /th/ /thr/ /d/ /dr/ /c/ /ch/ /k/ /kr/ /kl/ /kw/ /kh/ /khr/ /khl/ /khw/ /ʔ/ /m/ /n/ /ŋ/ /f/ /fr/ /fl/ /s/ /h/ /r/ /l/ /w/ /j/	/sil/ /sp/

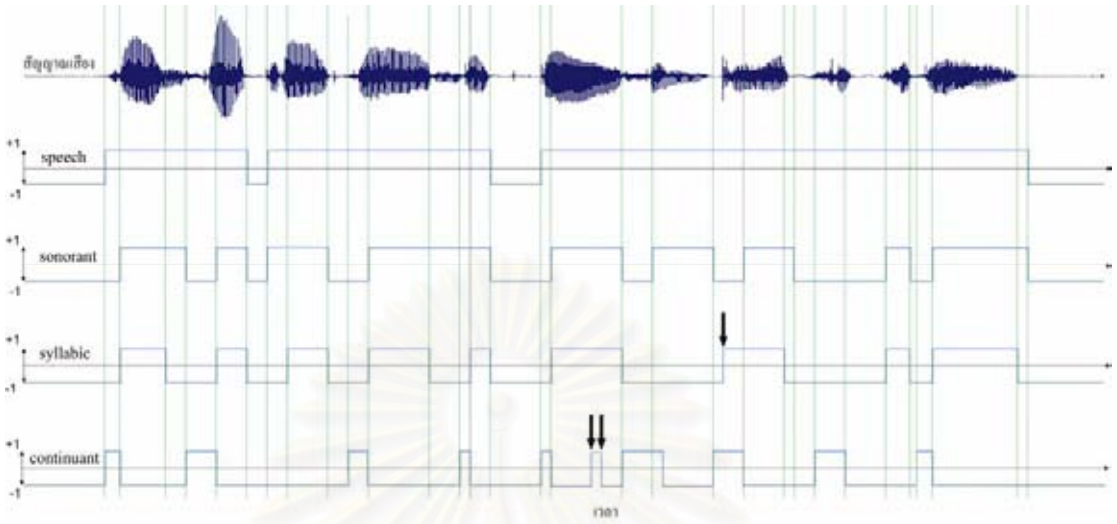
[1]ใช้หน่วยเสียงตามสัทอักษรสากล (International Phonetic Alphabet – IPA)

3. การตรวจหาขอบเขตของหน่วยเสียงจากผลการจำแนกลักษณะการออกเสียง

การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนนี้ จะอาศัยผลที่ได้จากการจำแนกลักษณะการออกเสียง โดยพิจารณาตำแหน่งที่มีการเปลี่ยนแปลงลักษณะการออกเสียงออกมาเป็นขอบเขตของหน่วยเสียง ดังแสดงด้วยรูปที่ 3.8 ซึ่งขอบเขตของหน่วยเสียงนี้จะมีความละเอียดอยู่ในระดับเดียวกันกับระยะห่างของแต่ละกรอบเวลาที่กำหนดเป็นพารามิเตอร์ไว้ในขั้นตอนการสกัดลักษณะสำคัญเพื่อการเรียนรู้และการจำแนก โดยในที่นี้ความละเอียดจะอยู่ที่ 10 มิลลิวินาที

ผลลัพธ์ที่ได้จากการแบ่งแยกสัทลักษณ์ [speech] [sonorant] [syllabic] และ [continuant] ซึ่งมีเป็นค่า +1 หรือ –1 จะนำมาใช้จำแนกประเภทของเสียงพูดตามลักษณะการออกเสียงตามโครงสร้างลำดับชั้นในรูปที่ 3.2 จากบนลงล่าง โดยเริ่มต้นจากการพิจารณาผลการแบ่งแยกสัท-ลักษณ์ [speech] ก่อน ถ้าหากผลที่ได้มีค่าเป็น –1 ก็จะจำแนกให้สัญญาณเสียงส่วนนั้นเป็นความเงียบ แต่ถ้ามีค่าเป็น +1 ก็จะพิจารณาผลการแบ่งแยกสัทลักษณ์ [sonorant] ต่อไป โดยถ้ามีค่าเป็น +1 จะแยกไปพิจารณาผลการแบ่งแยกสัทลักษณ์ [syllabic] แต่หากมีค่าเป็น –1 จะแยกไปพิจารณาผลการแบ่งแยกสัทลักษณ์ [continuant] แทน ในกรณีที่ผลการแบ่งแยกสัทลักษณ์ [syllabic] มีค่าเป็น +1 ก็จะจำแนกให้เป็นเสียงสระ แต่ถ้ามีค่าเป็น –1 ก็จะจำแนกให้เป็นเสียงประเภทกึ่งสระและเสียงนาสิก ในกรณีที่ผลการแบ่งแยกสัทลักษณ์ [sonorant] มีค่าเป็น –1 ก็จะพิจารณาผลการแบ่งแยกสัทลักษณ์ [continuant] ซึ่งหากมีค่าเป็น +1 ก็จะจำแนกให้เป็นเสียงเสียดแทรก และถ้ามีค่าเป็น –1 ก็จะจำแนกให้เป็นเสียงพยัญชนะกัก

เนื่องจากการจำแนกประเภทของเสียงพูดตามลักษณะการออกเสียงด้วยโครงสร้างลำดับชั้นนี้จะพิจารณาจากบนลงล่าง ดังนั้นผลของการแบ่งแยกสัทลักษณ์ที่ไม่สอดคล้องกับโครงสร้างลำดับชั้นจะไม่ได้นำมาใช้ประกอบการพิจารณาจำแนกลักษณะการออกเสียง ตัวอย่างเช่น สัญญาณเสียงในช่วงเวลาหนึ่ง นำไปเข้ากระบวนการแบ่งแยกสัทลักษณ์ [speech] [sonorant] [syllabic] และ [continuant] ได้ผลลัพธ์เป็น +1 +1 +1 และ –1 ตามลำดับ เมื่อพิจารณาตามโครงสร้างลำดับชั้นจากบนลงล่างพบว่าสัญญาณเสียงนี้เป็นเสียงสระ โดยในการพิจารณาเราจะไม่นำผลของการแบ่งแยกสัทลักษณ์ [continuant] มาประกอบการพิจารณา เนื่องจากผลของการแบ่งแยกสัทลักษณ์ [sonorant] มีค่าเป็น +1 ด้วยเหตุนี้การเปลี่ยนแปลงค่าใดๆของสัทลักษณ์ [continuant] จาก +1 ไปเป็น –1 หรือจาก –1 เป็น +1 ในส่วนของสัญญาณเสียงนี้จะถูกกรองออก โดยจะไม่นำมาใช้เป็นขอบเขตของหน่วยเสียง ดังแสดงด้วยรูปที่ 3.8 จากรูปจะสังเกตเห็นว่าบริเวณที่ลูกศรชี้คือตำแหน่งที่มีการเปลี่ยนแปลงค่าสัทลักษณ์ [continuant] และ [syllabic] ที่ไม่สอดคล้องกับโครงสร้างลำดับชั้นในที่นี้จึงกรองออก ไม่ได้นำมาพิจารณาเป็นขอบเขตของหน่วยเสียง



รูปที่ 3.8 ผลการแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงด้วยซัพพอร์ตเวกเตอร์แมชชีน ตำแหน่งที่ลูกศรชี้คือขอบเขตของหน่วยเสียงที่ถูกกรองออก

การสร้างกราฟของเซกเมนต์

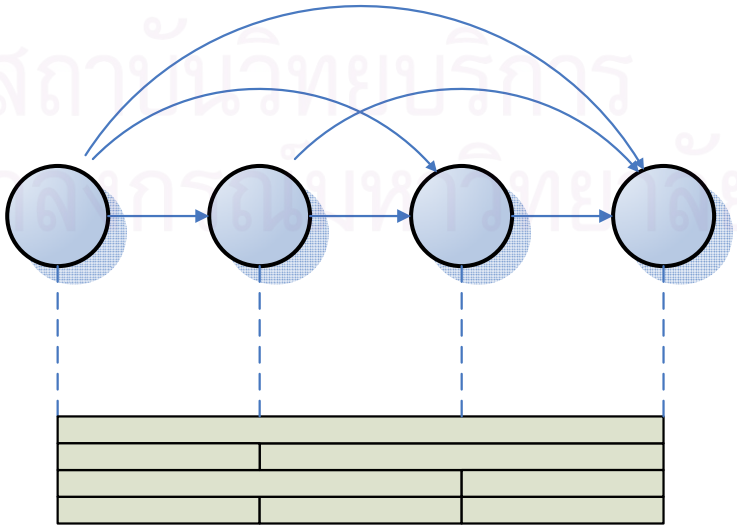
การสร้างกราฟของเซกเมนต์ เป็นขั้นตอนที่อาศัยลำดับของขอบเขตของหน่วยเสียงที่ได้จากกระบวนการตรวจหาขอบเขตของหน่วยเสียงมาสร้างกราฟของเซกเมนต์ เพื่อนำมาใช้ในการรู้จำเสียงพูดในขอบข่ายงานของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยในหัวข้อนี้จะเริ่มต้นด้วยการนำเสนอเกี่ยวกับวิธีในการสร้างกราฟของเซกเมนต์สามวิธีได้แก่

- การสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง (Full Segmentation)
- การสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม (Hard/Soft Segmentation)
- การสร้างกราฟของเซกเมนต์แบบหลายระดับ (Multi-Level Segmentation - MLS) [25][26]

ต่อจากนั้นจะนำเสนอเกี่ยวกับ การทดลองเปรียบเทียบประสิทธิภาพในการสร้างกราฟของเซกเมนต์ของทั้งสามวิธี และวิเคราะห์ผลการทดลอง

1. การสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

การสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง เป็นวิธีการสร้างกราฟของเซกเมนต์แบบที่ง่ายที่สุด โดยจะอาศัยลำดับของขอบเขตของหน่วยเสียงที่ได้จากขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงมาเป็นอินพุต แล้วพิจารณาสร้างเซกเมนต์ขึ้นมาด้วยการจับคู่ขอบเขตของหน่วยเสียงทุกๆคู่ขึ้นมาเป็นขอบเขตของเซกเมนต์ ให้ครบทุกกรณี ดังแสดงด้วยภาพตัวอย่างในรูปที่ 3.9 วงกลมแต่ละวงแทนขอบเขตของหน่วยเสียงแต่ละตำแหน่ง เมื่อเชื่อมต่อครบทุกขอบเขตของหน่วยเสียงก็จะได้เซกเมนต์ซึ่งแสดงด้วยสี่เหลี่ยมผืนผ้า

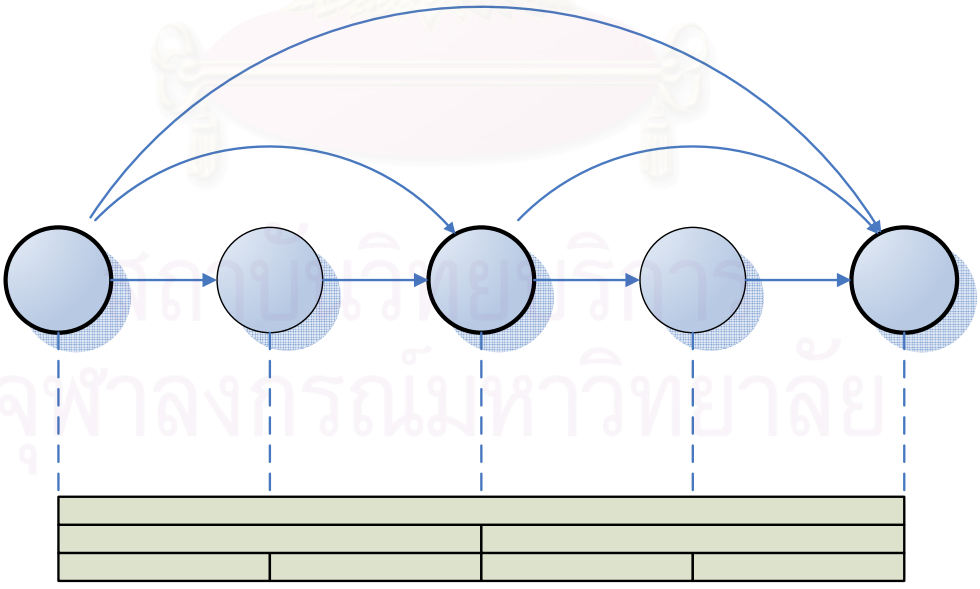


รูปที่ 3.9 กราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

2. การสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

การสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม เป็นวิธีการสร้างกราฟของเซกเมนต์แบบหนึ่งที่อาศัยการเปลี่ยนแปลงสเปกตรัมมาตัดบางเซกเมนต์ออกไป เพื่อให้ได้กราฟที่มีขนาดเล็กลงกว่าวิธีการแบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงเข้ากันหมด โดยใช้แนวคิดที่ว่าตำแหน่งของสัญญาณเสียงที่มีการเปลี่ยนแปลงสเปกตรัมสูงมักจะเป็นขอบเขตของหน่วยเสียง โดยจะวัดค่าการเปลี่ยนแปลงสเปกตรัมที่ทุกๆขอบเขตของหน่วยเสียงเอาไว้ และพิจารณาขอบเขตของหน่วยเสียงออกเป็นสองพวกคือ ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง และขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำ โดยจะใช้ค่าเฉลี่ยของการเปลี่ยนแปลงสเปกตรัมจากทุกๆขอบเขตของหน่วยเสียงในเสียงพูดนั้นมากำหนดระดับของการแบ่งแยก

การเชื่อมต่อขอบเขตของหน่วยเสียงด้วยวิธีการนี้ ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำทั้งหมดที่อยู่ระหว่างขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูงจะเชื่อมต่อกันหมด และขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูงทั้งหมดจะเชื่อมต่อกันเอง โดยจะไม่ยินยอมให้ขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำเชื่อมต่อข้ามขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง ดังแสดงด้วยภาพตัวอย่างในรูปที่ 3.10 วงกลมที่มีเส้นรอบวงหนาจะใช้แทนขอบเขตของหน่วยเสียงที่มีการเปลี่ยนสเปกตรัมสูง วงกลมที่มีเส้นรอบวงบางจะแทนขอบเขตของหน่วยเสียงที่มีการเปลี่ยนแปลงสเปกตรัมต่ำ



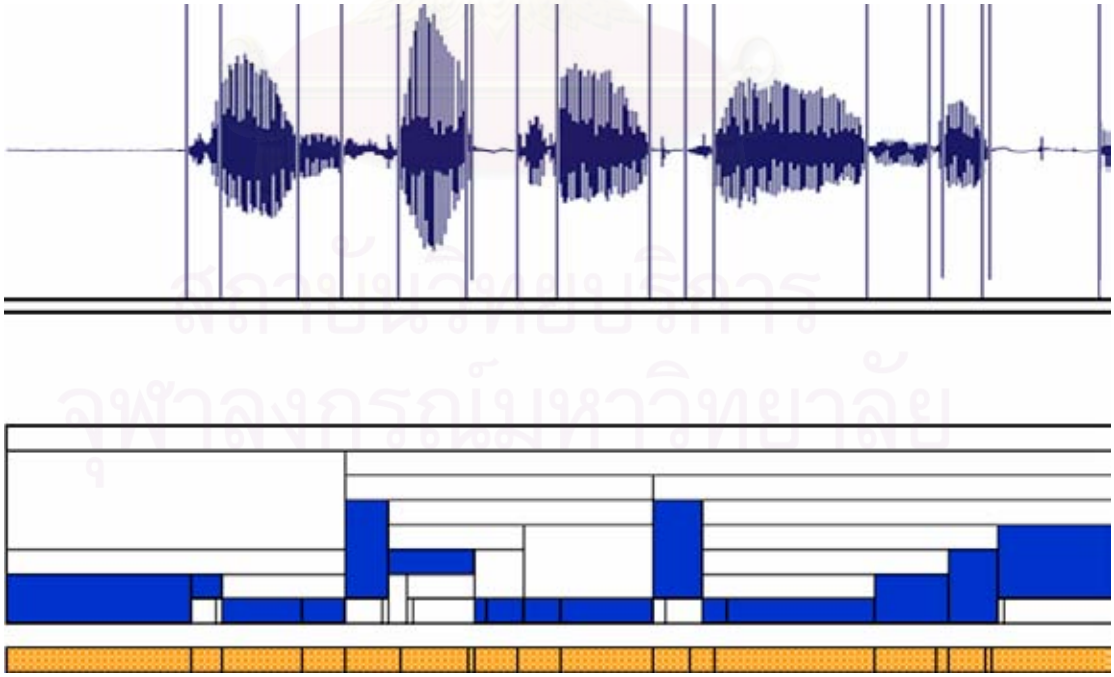
รูปที่ 3.10 กราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

3. การสร้างกราฟของเซกเมนต์แบบหลายระดับ

วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับ เป็นการสร้างกราฟของเซกเมนต์ให้มีลักษณะเหมือนเดนโดแกรมดังรูปที่ 3.11

โดยอาศัยหลักการที่พิจารณาให้เสียงพูดเป็นลำดับของของเซกเมนต์ที่เป็นหน่วยเสียง และสัญญาณเสียงที่อยู่ในเซกเมนต์เดียวกันจะมีความคล้ายคลึงกันมากกว่าสัญญาณเสียงที่อยู่ในเซกเมนต์อื่นซึ่งอยู่ติดกัน จากหลักการนี้ลักษณะของปัญหาการสร้างกราฟของเซกเมนต์สามารถเปลี่ยนไปเป็นปัญหาการรวมกลุ่ม (Clustering) โดยอาศัยการวัดค่าความคล้ายคลึงกันของสัญญาณเสียง แล้วค่อยๆรวมเซกเมนต์ที่มีความคล้ายกันเข้าด้วยกัน อัลกอริทึมการสร้างกราฟของเซกเมนต์แบบหลายระดับแสดงด้วยอัลกอริทึมดังตารางที่ 3.3

ในที่นี้เราใช้เวกเตอร์ลักษณะสำคัญของค่าสัมประสิทธิ์สเปกตรัมบนสเกลเมลที่ได้จากขั้นตอนการสกัดลักษณะสำคัญมาเป็นตัวเปรียบเทียบความคล้ายคลึงกันระหว่างสัญญาณเสียง การพิจารณาหาความคล้ายกันระหว่างสัญญาณเสียง จะวัดและแสดงอยู่ในรูปผลต่างของการกระจัดแบบยูคลิเดียน (Euclidean Distance) ระหว่างสัญญาณ โดยหากผลต่างของการกระจัดจากการเปรียบเทียบกันระหว่างหน่วยเสียงมีค่าน้อยเข้าใกล้ศูนย์ ก็จะสรุปได้ว่าสัญญาณเสียงคู่นั้นมีความคล้ายคลึงกันมาก



รูปที่ 3.11 กราฟของเซกเมนต์แบบหลายระดับ [26]

ตารางที่ 3.3 อัลกอริทึมการแบ่งเสียงพูดเป็นเซกเมนต์แบบหลายระดับ [26]

Find boundaries	
	$\{b_i, 0 < i < N\},\ t_i < t_j,\ \forall i < j$
Create initial region set	
	$R = \{r_0(i), 0 < i < N\},\ r_0(i) \equiv r(i, i+1)$
Create initial distance set	
	$D_0 = \{d_0(i), 0 < i < N\},\ d_0(i) \equiv d(r_0(i), r_0(i+1))$
Until	$R = \{r_N(0)\} \equiv r(0, N)$
For any k such that :	
	$d_j(k-1) > d_j(k) < d_j(k+1)$
(a)	$r_{j+1}(i) = r_j(i),\ 0 \le i < k$
(b)	$r_{j+1}(k) = merge(r_j(k),\ r_j(k+1))$
(c)	$r_{j+1}(i) = r_j(i+1),\ k < i < N - j - 1$
(d)	$R_{j+1} = \{r_{j+1}(i),\ 0 \le i < N - j - 1$
(e)	$d_{j+1}(i) = d_j(i),\ 0 \le i < k - 1$
(f)	$d_{j+1}(k-1) = \max(d_j(k-1), d(r_j(k-1), r_j(k)))$
(g)	$d_{j+1}(k) = \max(d_j(k+1), d(r_{j+1}(k), r_j(k+1)))$
(h)	$d_{j+1}(i) = d_j(i+1),\ k \le i < N - j - 1$
(i)	$D_{j+1} = \{d_{j+1}(i),\ 0 \le i < N - j - 1\}$

เมื่อกำหนดให้

- b_i เป็นขอบเขตของหน่วยเสียงที่เวลา t_i
- $r(i, j)$ คือเซกเมนต์ที่มีขอบเขตระยะเวลาตั้งแต่ t_i ถึง t_j
- $r_j(i)$ คือเซกเมนต์ที่ i ในรอบการทำงานที่ j
- $d(i, j)$ คือการกระจัดระหว่างเซกเมนต์ที่ i และเซกเมนต์ที่ j
- $d_j(i)$ คือการกระจัดที่ i ในรอบการทำงานที่ j
- $d_j(-1) = d_j(N - j) = \infty$
- $merge(r(i, j),\ r(j, k))$ คือการรวมเซกเมนต์สองอันที่อยู่ติดกันเป็นเซกเมนต์ใหม่ที่มีขอบเขตระยะเวลาตั้งแต่ t_i ถึง t_k

การให้คะแนนขอบเขตของหน่วยเสียง

ในการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับนำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์นั้น กราฟของเซกเมนต์ที่สร้างออกมาจะนำไปใช้ต่อในขั้นตอนการค้นหาและให้คะแนนเพื่อการรู้จำเสียงพูด งานวิจัยนี้ได้เสนอให้มีการคิดคะแนนขอบเขตของหน่วยเสียงให้กับกราฟของเซกเมนต์ที่สร้างออกมา โดยเป็นคะแนนแสดงความมั่นใจว่าขอบเขตของหน่วยเสียงนั้นเป็นขอบเขตของหน่วยเสียงจริง ซึ่งจะเป็นประโยชน์ในขั้นตอนการรู้จำเสียงพูดต่อไป

โดยอาศัยแนวคิดที่ว่าตำแหน่งของสัญญาณเสียงที่เป็นขอบเขตของหน่วยเสียงโดยส่วนใหญ่แล้วจะมีการเปลี่ยนแปลงสเปกตรัมสูง [27] ดังนั้นค่าคะแนนของขอบเขตของหน่วยเสียงในที่นี้จะอาศัยค่าการเปลี่ยนแปลงสเปกตรัม มาคำนวณออกมาเป็นความน่าจะเป็นที่ตำแหน่งทางเวลานั้นจะเป็นขอบเขตของหน่วยเสียง ซึ่งการเปลี่ยนแปลงสเปกตรัมนิยามให้เป็นไปตามสมการต่อไปนี้

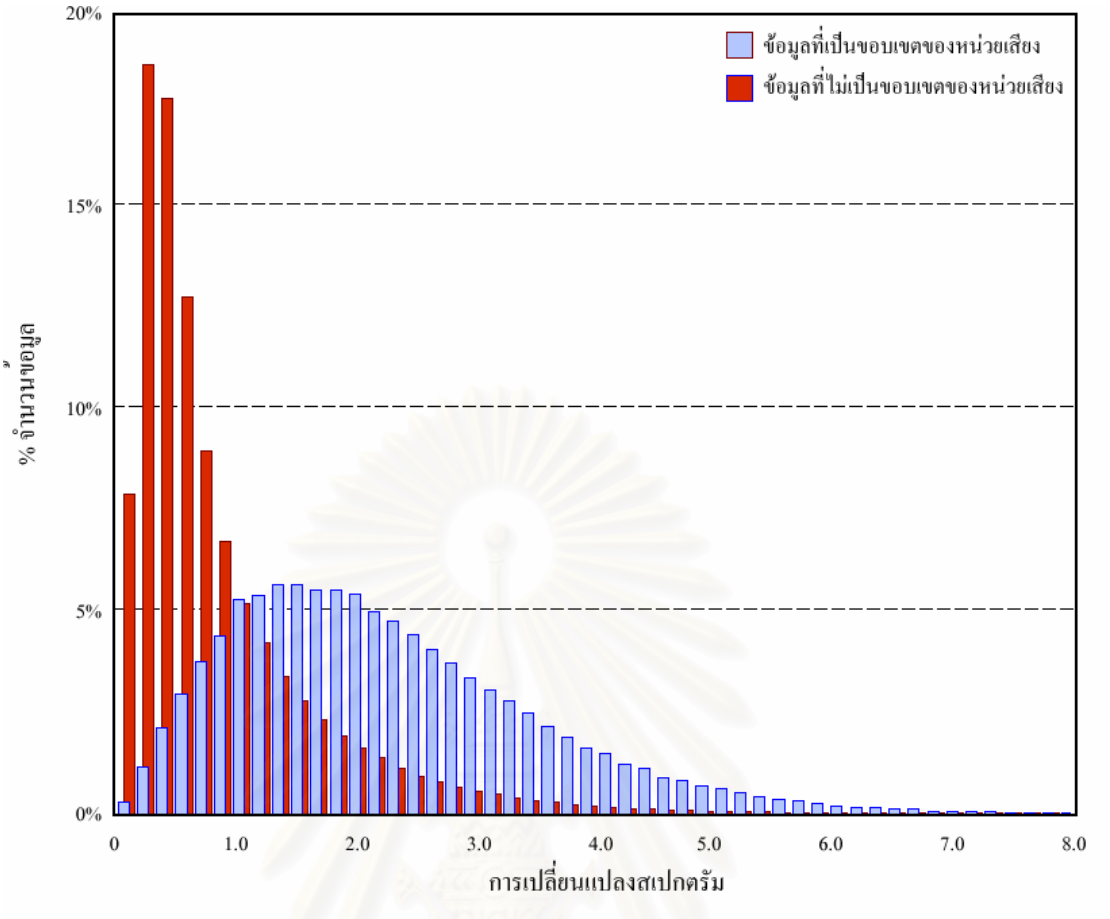
$$STM(m) = \frac{\sum_{i=1}^{D} a_i^2(m)}{D}$$

เมื่อกำหนดให้ D คือจำนวนมิติของเวกเตอร์ลักษณะสำคัญ (ในนี้ใช้สัมประสิทธิ์สเปกตรัมบนเมลสเกล $MFCC\_E\_D\_A$ ซึ่งมีขนาด 39 มิติ ดังที่ได้กล่าวไว้ในขั้นตอนการเตรียมข้อมูลเพื่อการตรวจหาขอบเขตของหน่วยเสียง) $a_i(m)$ คืออัตราการเปลี่ยนแปลงสเปกตรัมของสัมประสิทธิ์สเปกตรัมบนเมลสเกล $MFCC_i$ ซึ่งนิยามได้ดังนี้

$$a_i(m) = \frac{\sum_{n=-I}^{I} MFCC_i(n+m) * n}{\sum_{n=-I}^{I} n^2}$$

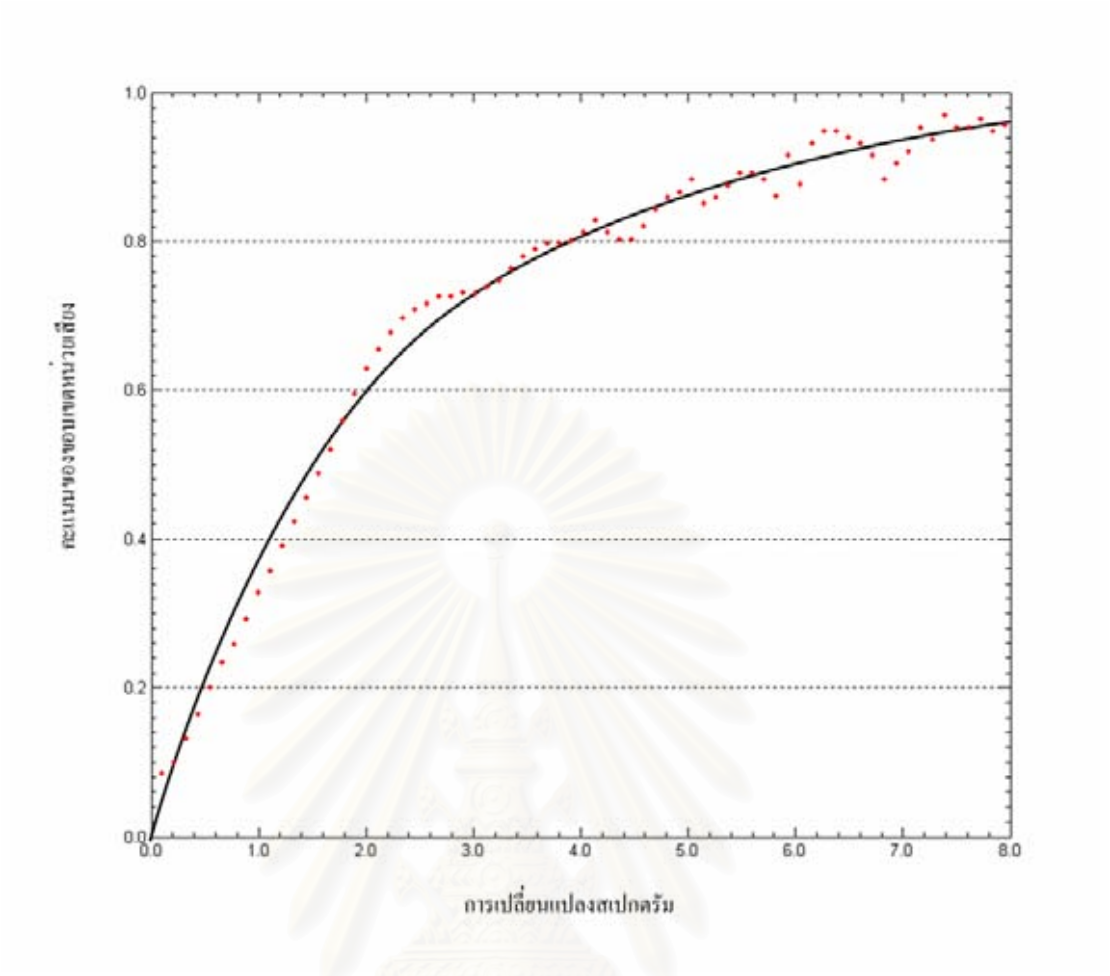
โดยที่ n แทนดัชนีของกรอบเวลา และ I แทนจำนวนของกรอบเวลาสำหรับใช้คำนวณอัตราการเปลี่ยนแปลงสเปกตรัม ในที่นี้กำหนดให้ $I = 2$ ดังนั้นจะเป็นการนำกรอบเวลาที่อยู่ติดกันสองกรอบเวลาก่อนหน้าและสองกรอบเวลาตามหลังมาคำนวณหาการเปลี่ยนแปลงสเปกตรัมที่ตำแหน่งกึ่งกลางตรงกลาง

เพื่อตรวจสอบความถูกต้องและความเป็นไปได้ของการนำแนวคิดนี้มาใช้ งานวิจัยนี้จึงทดลองวัดการกระจายค่าการเปลี่ยนแปลงสเปกตรัมของข้อมูลที่เป็นและไม่เป็นขอบเขตของหน่วยเสียงให้อยู่ในรูปของฮิสโตแกรม เพื่อนำมาอธิบายการแจกแจงความน่าจะเป็นในการเป็นขอบเขตของหน่วยเสียงที่การเปลี่ยนแปลงสเปกตรัมต่างๆ ดังรูปที่ 3.12



รูปที่ 3.12 ฮิสโตแกรมของข้อมูลที่เป็นและไม่เป็นขอบเขตของหน่วยเสียง

จากฮิสโตแกรม พบว่าตำแหน่งของสัญญาณเสียงพูดที่เป็นขอบเขตของหน่วยเสียงส่วนใหญ่จะมีระดับการเปลี่ยนแปลงสเปกตรัมสูงกว่าตำแหน่งของสัญญาณเสียงพูดที่ไม่ได้เป็นขอบเขตของหน่วยเสียง โดยค่าความน่าจะเป็นของการที่ตำแหน่งของสัญญาณเสียงนั้นเป็นขอบเขตของเสียงพูดที่ระดับการเปลี่ยนแปลงสเปกตรัมใดๆจะคำนวณมาจากอัตราส่วนระหว่างจำนวนข้อมูลที่เป็นขอบเขตของหน่วยเสียง ต่อจำนวนข้อมูลทั้งหมดที่มีระดับการเปลี่ยนแปลงสเปกตรัมนั้นๆ แทนด้วยจุดต่างๆ และสามารถประมาณเส้นโค้งของกราฟความสัมพันธ์ระหว่างการเปลี่ยนแปลงสเปกตรัมและคะแนนขอบเขตของหน่วยเสียงได้ดังรูปที่ 3.13



รูปที่ 3.13 ความสัมพันธ์ระหว่างการเปลี่ยนแปลงสเปกตรัมและคะแนนของขอบเขตหน่วยเสียง

บทที่ 4
การทดลองแบ่งเสียงพูดเป็นเซกเมนต์

บทนี้จะนำเสนอเกี่ยวกับการทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ โดยจะเริ่มจากนำเสนอฐานข้อมูลเสียงที่ใช้ในการทดลอง องค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูดที่นำมาใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด การทดลองและผลการทดลองของการแบ่งเสียงพูดเป็นเซกเมนต์ และสรุปผลการทดลอง ตามลำดับ

ฐานข้อมูลเสียงเพื่อการแบ่งเสียงพูดเป็นเซกเมนต์

ข้อมูลเสียงที่นำมาใช้ในการทดลองการแบ่งเสียงพูดเป็นเซกเมนต์คือข้อมูลเสียงจากฐานข้อมูลเสียงโลตัส (LOTUS) [28] ซึ่งเป็นฐานข้อมูลเสียงพูดภาษาไทยขนาดใหญ่ที่มีคำศัพท์จำนวนมากแบบเสียงพูดต่อเนื่อง (Large Vocabulary Continuous Speech Recognition: LVCSR) โดยฐานข้อมูลนี้จะมีข้อมูลเสียง ชุดหน่วยเสียงสมดุล (Phonetically Distributed Set) ใช้สำหรับการเรียนรู้แบบจำลองเสียง การเรียนรู้การกำกับหน่วยเสียงอัตโนมัติ และเป็นชุดเสียงสำหรับการทดลองระบบที่มีการปรับผู้พูด (Speaker Adaptation) ฐานข้อมูลยังประกอบด้วยชุดเสียงอีก 3 ชุด สำหรับฝึกฝนแบบจำลองเสียงและแบบจำลองภาษา ชุดสำหรับทดสอบเพื่อการพัฒนา และชุดสำหรับทดสอบเพื่อประเมินผล ฐานข้อมูลเสียงทั้ง 3 ชุดจะครอบคลุมคำศัพท์ภาษาไทยจำนวนไม่ต่ำกว่า 5,000 คำ จากฐานข้อมูลบทความข่าวหรือบทความทั่วไป

ฐานข้อมูลเสียงโลตัสบันทึกเสียงพูดผ่านไมโครโฟน 2 ประเภท ประเภทแรกเป็นไมโครโฟนสำหรับพูดระยะใกล้ (Close-talk) คุณภาพสูง แบบทิศทางเดียว (Unidirectional) ระดับคุณภาพปานกลาง และทำการบันทึกเสียงใน 2 สภาพแวดล้อม คือ สภาพแวดล้อมแบบห้องเงียบ และ สภาพแวดล้อมแบบสำนักงาน โดยเก็บข้อมูลเสียงผ่านแถบบันทึกเสียงดิจิตัล (Digital Audio Tape) ก่อนแปลงเป็นไฟล์อิเล็กทรอนิกส์

ตารางที่ 4.1 สถิติเกี่ยวกับจำนวนขอบเขตหน่วยเสียงต่อการเปล่งเสียงหนึ่งครั้ง สำหรับข้อมูลเสียงในชุดหน่วยเสียงสมดุล (PD Set)

	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ต่ำสุด	สูงสุด
จำนวนขอบเขตหน่วยเสียงต่อการเปล่งเสียงหนึ่งครั้ง	52.7	30.4	12	190

ในงานวิจัยนี้เลือกใช้ข้อมูลเสียงจากชุดหน่วยเสียงสมดุล (Phonetically Distributed Set, PD Set) จากฐานข้อมูลเสียงโลตัสซึ่งประกอบไปด้วยเสียงผู้พูด 48 คนแบ่งเป็นเพศชายและหญิงในจำนวนเท่ากัน และมีการกำกับขอบเขตหน่วยเสียงไว้แล้ว โดยข้อมูลทางสถิติของข้อมูลเสียงเกี่ยวกับจำนวนขอบเขตของหน่วยเสียงต่อการเปล่งเสียงหนึ่งครั้งแสดงด้วยตารางดังต่อไปนี้

โดยจะแบ่งข้อมูลเสียงของชุดหน่วยเสียงสมดุลออกเป็นสองส่วนเท่าๆกัน ส่วนแรกเป็นข้อมูลเสียงเพื่อการเรียนรู้ อีกส่วนหนึ่งเป็นข้อมูลเสียงเพื่อการทดสอบ โดยแต่ละชุดจะมีข้อมูลเสียงจากผู้พูดชายและผู้พูดหญิงเท่ากันรวมทั้งสิ้น 48 คน คิดเป็นข้อมูลเสียงจำนวน 1680 ไฟล์ ความยาวประมาณ 3 ชั่วโมง

องค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูด

ในวิทยานิพนธ์นี้สร้างเครื่องรู้จำเสียงพูดเพื่อนำมาใช้ในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงของวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด โดยอาศัยชุดเครื่องมือฮิดเดนมาร์คอฟ – เอชทีเค (Hidden Markov Toolkit – HTK) [29] ซึ่งเป็นชุดเครื่องมือสำหรับสร้างเครื่องรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ โดยรายละเอียดเกี่ยวกับขั้นตอนการเรียนรู้และการรู้จำสามารถดูได้ที่ภาคผนวก เครื่องรู้จำเสียงพูดที่ใช้ในการทดลองนี้มีองค์ประกอบและประสิทธิภาพ สรุปได้ดังตารางที่ 4.2

จากประสิทธิภาพในการรู้จำเสียงพูดในตารางที่ 4.2 จะพบว่าเปอร์เซ็นต์ความแม่นยำในการรู้จำเสียงพูดเท่ากับ 47.8% ซึ่งอยู่ในเกณฑ์ที่ยอมรับได้เนื่องจากเป็นการรู้จำเสียงพูดในระดับหน่วยเสียง โดยทั่วไปแล้วเปอร์เซ็นต์ความแม่นยำของเครื่องรู้จำหน่วยเสียงจะอยู่ระหว่าง 40 ถึง 70 เปอร์เซ็นต์ ขึ้นอยู่กับปัจจัยหลายๆอย่างทั้งลักษณะการเรียนรู้และปริมาณข้อมูลเพื่อการเรียนรู้ โดยในการทดลองนี้จำเป็นจะใช้ข้อมูลเสียงแบบที่มีการกำกับหน่วยเสียงอ้างอิงไว้ด้วย ทำให้มีตัวเลือกที่จำกัดกว่าการเรียนรู้เครื่องรู้จำเสียงโดยปกติ

ตารางที่ 4.2 ตารางสรุปองค์ประกอบและประสิทธิภาพของเครื่องรู้จำเสียงพูดที่นำมาใช้เปรียบเทียบการตรวจหาขอบเขตของหน่วยเสียง

	รายละเอียด	
ค่าลักษณะสำคัญ	สัมประสิทธิ์เซปสตรัมบนสเกลเมลที่มีค่าพลังงานรวมอยู่ด้วย อัตราการเปลี่ยนแปลง (Delta) และความเร่ง (Accelerations) โดยคำนวณจากสัญญาณเสียงทุกๆกรอบเวลายาว 25 มิลลิวินาที แต่ละกรอบเวลาจะมีระยะเวลาห่างกัน 10 มิลลิวินาที ได้ออกมาเป็นเวกเตอร์ลักษณะสำคัญที่มีขนาด 39 มิติ	
จำนวนหน่วยเสียง	หน่วยเสียงภาษาไทยจำนวน 74 หน่วยเสียงตามฐานข้อมูลเสียงโลตัส	
แบบจำลองเสียงพูด	เป็นแบบจำลองเสียงพูดแบบไม่ขึ้นกับบริบทรอบข้าง (Context-independent Phone Model) โดยจะใช้การกระจายของค่าความน่าจะเป็นแบบเกาส์เซียนที่มีเมทริกซ์ความแปรปรวนร่วมเกี่ยวตามแนวทแยง (Diagonal Covariance Gaussian Distribution)	
แบบจำลองเสียงภาษา	แบบจำลองภาษาแบบอาศัยค่าความน่าจะเป็นของการที่หน่วยเสียงหนึ่งจะปรากฏอยู่ติดกับอีกหน่วยเสียงหนึ่ง (Bigram Language Model)	
ข้อมูลเสียง	ข้อมูลเสียงชุดหน่วยเสียงสมดุล จากฐานข้อมูลเสียงโลตัส โดยใช้เสียงครึ่งหนึ่งสำหรับฝึกฝน และอีกครึ่งหนึ่งสำหรับทดสอบประสิทธิภาพ	
ประสิทธิภาพ	ความถูกต้องในการรู้จำหน่วยเสียง (Accuracy)	47.80%
	ความผิดพลาดตัดออก (Deletion error)	4.55%
	ความผิดพลาดแบบแทนที่ (Substituition error)	30.7%
	ความผิดพลาดแบบแทรก (Insertion error)	16.91%

การทดลองและผลการทดลอง

การทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ในที่นี้จะทดลองเพื่อวัดประสิทธิภาพของวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่เสนอไว้ในบทที่ 3 เปรียบเทียบกับวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด โดยจะประกอบไปด้วยการทดลองทั้งหมด 3 การทดลอง คือ การทดลองเพื่อวัดประสิทธิภาพการจำแนกลักษณะการออกเสียง การทดลองเพื่อเปรียบเทียบประสิทธิภาพการตรวจหาขอบเขตของหน่วยเสียง และการทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์ ซึ่งการทดลองแต่ละส่วนมีรายละเอียดดังต่อไปนี้

1. การทดลองเพื่อวัดประสิทธิภาพการจำแนกลักษณะการออกเสียง

การทดลองจำแนกลักษณะการออกเสียงในที่นี้ จะทำโดยป้อนข้อมูลที่ได้จากกระบวนการสกัดลักษณะสำคัญเข้าเป็นอินพุตของเอสวีเอ็มที่ได้จากการเรียนรู้ ผลที่ได้คือค่าของฟังก์ชันตัดสินใจ $f(\mathbf{x}) = sign((\mathbf{w} \cdot \mathbf{x}) + b)$ ของแต่ละกรอบเวลาของสัญญาณเสียงซึ่งได้จากซัพพอร์ตเวกเตอร์แมชชีนดังที่กล่าวไว้แล้วในบทที่ 2 โดยหากค่าที่ได้คือ +1 ก็หมายความว่าสัญญาณเสียงกรอบเวลานั้นมีสมบัติของสัทลักษณ์นั้น แต่หากค่าที่ได้เป็น -1 ก็หมายความว่าไม่มีสมบัติของสัทลักษณ์ดังกล่าว แล้วจึงพิจารณาจำแนกประเภทลักษณะการออกเสียงด้วยโครงสร้างลำดับชั้นในรูปที่ 3.2

การจำแนกลักษณะการออกเสียงในแต่ละกรอบเวลาของสัญญาณเสียงจะใช้ เครื่องจำแนกลักษณะการออกเสียงที่ใช้เอสวีเอ็มที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้น เขียนแทนด้วยสัญลักษณ์ SVM_{linear} และที่เครื่องจำแนกลักษณะการออกเสียงที่ใช้เอสวีเอ็มที่ใช้ฟังก์ชันเคอร์เนลแบบพหุนาม เขียนแทนด้วยสัญลักษณ์ $SVM_{polynomial}$ ผลลัพธ์ที่ได้จากการจำแนกลักษณะการออกเสียง จะนำมาวัดประสิทธิภาพโดยมีรายละเอียดการวัดประสิทธิภาพและผลการทดลองดังต่อไปนี้

1.1 การวัดประสิทธิภาพ

การวัดประสิทธิภาพการจำแนกลักษณะการออกเสียงแบบอาศัยเอสวีเอ็ม จะพิจารณาแยกตามเครื่องแบ่งแยกสัทลักษณ์เอสวีเอ็มแต่ละตัว โดยดูจากเปอร์เซ็นต์ความแม่นยำของการแบ่งแยกสัทลักษณ์ เขียนแทนด้วยสัญลักษณ์ $\%Accuracy$

แต่เนื่องจากจำนวนตัวอย่างที่มีฉลากเป็น +1 และ -1 ที่นำมาใช้ทดสอบกับตัวแบ่งแยกนั้นมีจำนวนไม่เท่ากัน ดังนั้นเพื่อให้มีโอกาสในการพบตัวอย่างของข้อมูลทั้งสองแบบอย่างเท่าเทียมกัน คือมีโอกาสเป็น 50% เปอร์เซ็นต์ความแม่นยำจึงต้องถูกปรับให้เป็นบรรทัดฐานโดยคำนวณได้จากสูตรดังต่อไปนี้

$$\%Accuracy = 50 \times \frac{N_{1,1}}{N_{1,1} + N_{1,-1}} + 50 \times \frac{N_{-1,-1}}{N_{-1,1} + N_{-1,-1}}$$

เมื่อ $N_{i,j}$ คือจำนวนเวกเตอร์ลักษณะสำคัญที่มีฉลากเป็น i และถูกจำแนกให้มีฉลากเป็น j โดยที่ $i, j \in \{-1,1\}$

1.2 ผลการทดลองการจำแนกลักษณะการออกเสียง

ผลการจำแนกลักษณะการออกเสียงของเสียงพูดจากตัวแบ่งแยกสัทลักษณ์เอสวีเอ็มทั้งสี่ตัว แสดงด้วยตารางที่ 4.3

ตารางที่ 4.3 ประสิทธิภาพการแบ่งแยกสัทลักษณ์ของลักษณะการออกเสียง
(เปอร์เซ็นต์ความแม่นยำถูกปรับเพื่อให้โอกาสพบข้อมูลที่มีและไม่มีสัทลักษณ์เป็น 50% เท่ากัน)

เครื่องจำแนกลักษณะการออกเสียง	เปอร์เซ็นต์ความแม่นยำในการแบ่งแยกสัทลักษณ์			
	[speech]	[sonorant]	[syllabic]	[continuant]
SVM_{linear}	94.03	77.57	82.71	81.90
$SVM_{polynomial}$	86.75	84.17	86.22	87.67

1.3 วิเคราะห์ผลการทดลอง

จากการทดลองวัดประสิทธิภาพการจำแนกลักษณะการออกเสียง โดยพิจารณาจากผลการแบ่งแยกสัทลักษณ์ในตารางที่ 4.3 จะพบว่าเอสวีเอ็มทั้งแบบเชิงเส้นและแบบพหุนาม สามารถแบ่งแยกสัทลักษณ์ได้เปอร์เซ็นต์ความแม่นยำโดยรวมอยู่ที่ 80 – 90% โดยตัวแบ่งแยกเสียงพูดและความเงียบเอสวีเอ็มแบบเชิงเส้นมีเปอร์เซ็นต์ความแม่นยำสูงถึง 94% ส่วนเอสวีเอ็มแบบพหุนามสามารถแบ่งแยกสัทลักษณ์ [sonorant] [syllabic] [continuant] ได้เปอร์เซ็นต์ความแม่นยำสูงกว่าเอสวีเอ็มแบบเชิงเส้น อย่างไรก็ตามการตัดสินว่าเครื่องจำแนกลักษณะการออกเสียงแบบใดมีประสิทธิภาพในการตรวจหาขอบเขตของหน่วยเสียงได้ดีกว่ากัน ไม่อาจตัดสินโดยพิจารณาจากเปอร์เซ็นต์ความแม่นยำในการแบ่งแยกสัทลักษณ์ของเอสวีเอ็มแต่ละตัวเพียงอย่างเดียว เนื่องจากจุดประสงค์สำคัญที่แท้จริงของการนำเอสวีเอ็มมาใช้ก็เพื่อที่จะนำไปใช้ในการตรวจหาขอบเขตของหน่วยเสียง

2. การทดลองเพื่อเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง

ผลลัพธ์ที่ได้จากการจำแนกลักษณะการออกเสียงด้วยซัพพอร์ตเวกเตอร์แมชชีน และจากการรู้จำหน่วยเสียงด้วยเครื่องรู้จำเสียงพูด จะนำมาใช้วัดประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง โดยมีรายละเอียดเกี่ยวกับการวัดประสิทธิภาพ และผลการทดลองดังต่อไปนี้

2.1 การวัดประสิทธิภาพ

การทดลองตรวจหาขอบเขตของหน่วยเสียงจะวัดประสิทธิภาพเปรียบเทียบกับการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด โดยพิจารณาจากความแม่นยำ (Precision) ความครอบคลุม (Recall) และเวลาที่ใช้ในการทำงานโดยวัดจากเรียลไทม์แฟคเตอร์ (Real-time factor -RTF) โดยมีรายละเอียดดังต่อไปนี้

2.1.1 ความแม่นยำและความครอบคลุม

การพิจารณาค่าความแม่นยำและค่าความครอบคลุม จะต้องใช้ผลของการตัดสินใจว่าขอบเขตของหน่วยเสียงที่หามาได้นั้นมีความถูกต้องตรงกับขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงหรือไม่ ขอบเขตของหน่วยเสียงที่ดีนั้นไม่จำเป็นต้องอยู่ที่ตำแหน่งเดียวกันกับขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงก็ได้ โดยอาจยินยอมให้คลาดเคลื่อนกันได้เป็นระยะเวลาสั้นๆในระดับมิลลิวินาที ขึ้นอยู่กับการนำไปใช้งาน ดังนั้นในการทดลองนี้จึงมีการกำหนดระดับความคลาดเคลื่อนยินยอม (Tolerance)

เปอร์เซ็นต์ความแม่นยำและเปอร์เซ็นต์ความครอบคลุมเขียนแทนด้วยสัญลักษณ์ $\%Pr$ และสัญลักษณ์ $\%Re$ ตามลำดับ ซึ่งสามารถคำนวณได้จากสมการต่อไปนี้

$$\%Pr = 100 \times \frac{C}{D}, \quad \%Re = 100 \times \frac{C}{T}$$

เมื่อกำหนดให้ D คือจำนวนขอบเขตของหน่วยเสียงทั้งหมดที่ตรวจหามาได้ T คือจำนวนขอบเขตหน่วยเสียงที่เป็นตัวอ้างอิง และ C คือจำนวนขอบเขตของหน่วยเสียงทั้งหมดที่ตรงกับขอบเขตของหน่วยเสียงอ้างอิงโดยยินยอมให้คลาดเคลื่อนไปจากขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงได้ไม่เกินระดับความคลาดเคลื่อนยินยอมที่กำหนด

เกณฑ์การตัดสินคุณภาพของขอบเขตของหน่วยเสียงที่หามาได้ว่ายอมรับได้หรือไม่ ส่วนใหญ่แล้วจะพิจารณาโดยดูจากความแม่นยำของขอบเขตที่หามาได้ที่ระดับความคลาดเคลื่อนยินยอมที่แตกต่างกัน ขึ้นอยู่กับความต้องการและประเภทของงานที่นำไปใช้ ในงานวิจัยเกี่ยวกับด้านการ

รู้จำเสียงพูดและการสังเคราะห์เสียงพูดซึ่งต้องใช้ข้อมูลเสียงที่กำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว โดยส่วนใหญ่จะพิจารณาความคลาดเคลื่อนของขอบเขตของหน่วยเสียงที่ระดับไม่เกิน 20 ถึง 30 มิลลิวินาที [30][31][32] เนื่องจากขอบเขตของหน่วยเสียงที่ได้ มีความใกล้เคียงกันกับขอบเขตของหน่วยเสียงที่กำกับไว้ด้วยคนมาก เพียงพอต่อการนำไปใช้งานได้อย่างมีประสิทธิภาพ โดยเสียงพูดที่สังเคราะห์ได้นั้นมีความเป็นธรรมชาติอยู่ในเกณฑ์ที่ยอมรับได้

ดังนั้นเกณฑ์ในการยอมรับได้ของขอบเขตของหน่วยเสียงสำหรับนำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่แนะนำ คือที่ระดับความคลาดเคลื่อนไม่เกิน 30 มิลลิวินาที เช่นเดียวกันกับที่ใช้ในงานวิจัยอื่นๆ แต่อย่างไรก็ตามในงานวิจัยนี้ผู้วิจัยสนใจจะวัดผลลัพธ์ที่ระดับความคลาดเคลื่อนยินยอมที่ต่างๆกัน เพื่อให้เห็นแนวโน้มชัดเจนกว่าการวัดคุณภาพของขอบเขตของหน่วยเสียงที่ระดับความคลาดเคลื่อนยินยอมเพียงค่าเดียว โดยจะใช้ระดับความคลาดเคลื่อนยินยอมที่ 10, 20, 30 และ 40 มิลลิวินาทีตามลำดับ

2.1.2 เรียลไทม์แฟคเตอร์

การเปรียบเทียบเวลาที่ใช้ในการตรวจหาขอบเขตของหน่วยเสียงจะใช้เรียลไทม์แฟคเตอร์เขียนแทนด้วยสัญลักษณ์ *RTF* ซึ่งเป็นตัววัดความเร็วที่ใช้กันโดยทั่วไปในระบบรู้จำเสียงพูด โดยมีนิยามดังนี้

$$RTF = \frac{P}{I}$$

เมื่อกำหนดให้ P คือเวลาที่ใช้ในการทำงาน และ I คือความยาวของเสียงพูดที่เข้ามา ตัวอย่างเช่น หากระบบรับเสียงพูดความยาว 5 วินาทีเข้ามาและใช้เวลาในการทำงานทั้งสิ้น 10 วินาที *RTF* จะมีค่าเป็น 2 ดังนั้นหาก *RTF* เป็น 1 กระบวนการทำงานนั้นจะสามารถทำได้แบบทันกาล

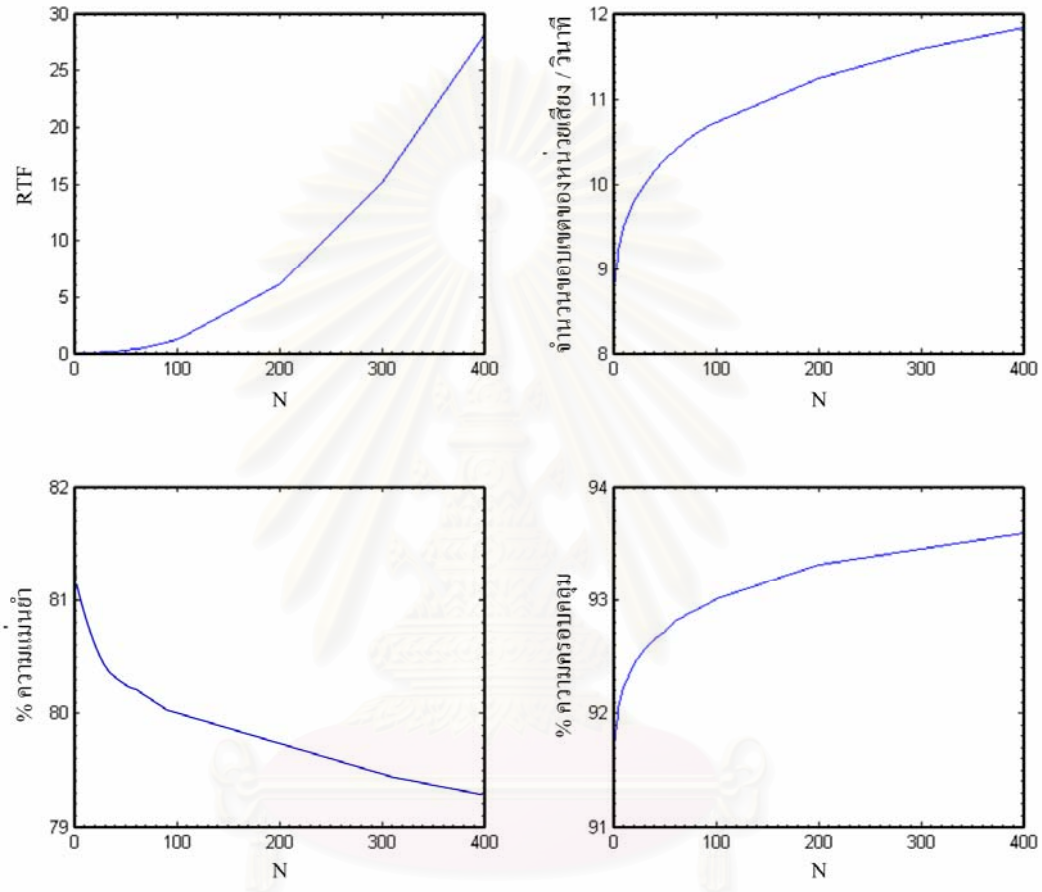
ค่า *RTF* นั้นขึ้นอยู่กับความเร็วของอุปกรณ์ฮาร์ดแวร์ที่ใช้ โดยในอุปกรณ์ฮาร์ดแวร์ที่ใช้ในการทดลองทั้งหมดในวิทยานิพนธ์นี้คือ เครื่องคอมพิวเตอร์พีซีที่มีสถาปัตยกรรมแบบ 32 บิต ความเร็วซีพียู 2,793 เมกะเฮิร์ต ใช้หน่วยความจำหลักขนาด 1 กิกะไบต์ ทำงานบนระบบปฏิบัติการลินุกซ์

2.1.3 จำนวนขอบเขตของหน่วยเสียงที่ตรวจหามาได้

จำนวนขอบเขตของหน่วยเสียงที่ตรวจหามาได้ จะวัดเป็นจำนวนขอบเขตของหน่วยเสียงต่อหนึ่งหน่วยวินาที โดยค่านี้จะสะท้อนให้เห็นขนาดของอินพุตที่จะผ่านเข้าไปยังขั้นตอนการสร้างกราฟของเซกเมนต์ต่อไป โดยหากมีปริมาณมากเกินไปก็จะทำให้กราฟของเซกเมนต์มีขนาดใหญ่ตามไปด้วย แต่ถ้าหากมีน้อยเกินไปก็จะไม่ครอบคลุมขอบเขตของหน่วยเสียงที่ต้องการ

2.2 ผลการทดลองของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด

ประสิทธิภาพของกระบวนการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดจะขึ้นอยู่กับตัวแปร N หรือจำนวนของผลการรู้จำ N อันดับที่ดีที่สุด ดังแสดงด้วยกราฟในรูปที่ 4.1



รูปที่ 4.1 กราฟแสดงประสิทธิภาพการตรวจหาขอบเขตหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด (ที่ระดับความคลาดเคลื่อนยินยอม 30 มิลลิวินาที)

2.3 ผลการทดลองของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

ประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ ในด้านความเร็วในการทำงานและจำนวนของขอบเขตของหน่วยเสียงที่ตรวจหามาได้แสดงไว้ในตารางที่ 4.4 ส่วนประสิทธิภาพของการตรวจหาขอบเขตของหน่วย

เสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนในแง่ของความแม่นยำและความครอบคลุมแสดงไว้ในตารางที่ 4.5 และ 4.6 ตามลำดับ

ตารางที่ 4.4 ความเร็วในการทำงานและจำนวนขอบเขตของหน่วยเสียงที่หามาได้ของกระบวนการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน แบบเชิงเส้นเทียบกับแบบพหุนาม

	RTF	จำนวนขอบเขตของหน่วยเสียงต่อวินาที
SVM_{linear}	0.57	11.1
$SVM_{polynomial}$	4.83	10.9

ตารางที่ 4.5 ความแม่นยำของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

	เปอร์เซ็นต์ความแม่นยำที่ระดับความคลาดเคลื่อน			
	10 มิลลิวินาที	20 มิลลิวินาที	30 มิลลิวินาที	40 มิลลิวินาที
SVM_{linear}	58.1	76.3	85.0	89.9
$SVM_{polynomial}$	57.4	76.4	85.3	90.6

ตารางที่ 4.6 ความครอบคลุมของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน

	เปอร์เซ็นต์ความครอบคลุมที่ระดับความคลาดเคลื่อน			
	10 มิลลิวินาที	20 มิลลิวินาที	30 มิลลิวินาที	40 มิลลิวินาที
SVM_{linear}	72.6	87.2	94.4	97.0
$SVM_{polynomial}$	70.2	85.4	92.4	95.1

2.4 ผลการทดลองของการเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียง

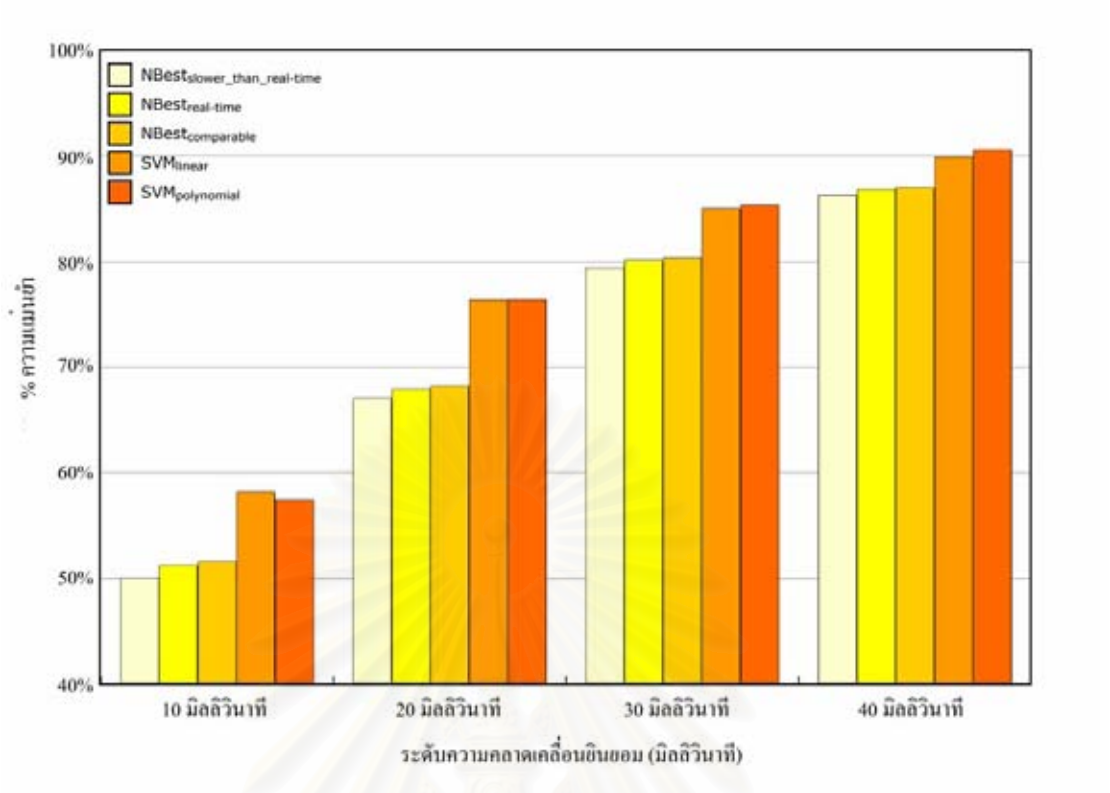
เนื่องจากประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด สามารถปรับเปลี่ยนได้ตามตัวแปร N โดยในการทดลองเพื่อเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงระหว่างวิธีที่อาศัยเครื่องรู้จำเสียงพูด กับวิธีที่อาศัยซัพ

พอร์ตเวกเตอร์แมชชีนนี้ จะเปรียบเทียบประสิทธิภาพกับเครื่องรู้จำเสียงพูดทั้งสิ้นสามเครื่องได้แก่ เครื่องรู้จำเสียงพูดที่ทำงานได้รวดเร็วเท่าเทียมกันกับการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน เขียนแทนด้วยสัญลักษณ์ $NBest_{comparable}$ เครื่องรู้จำเสียงพูดที่ทำงานได้แบบทันกาล เขียนแทนด้วยสัญลักษณ์ $NBest_{real-time}$ และเครื่องรู้จำเสียงพูดที่ทำงานได้ช้ากว่าทันกาล เขียนแทนด้วยสัญลักษณ์ $NBest_{slower\ than\ real-time}$

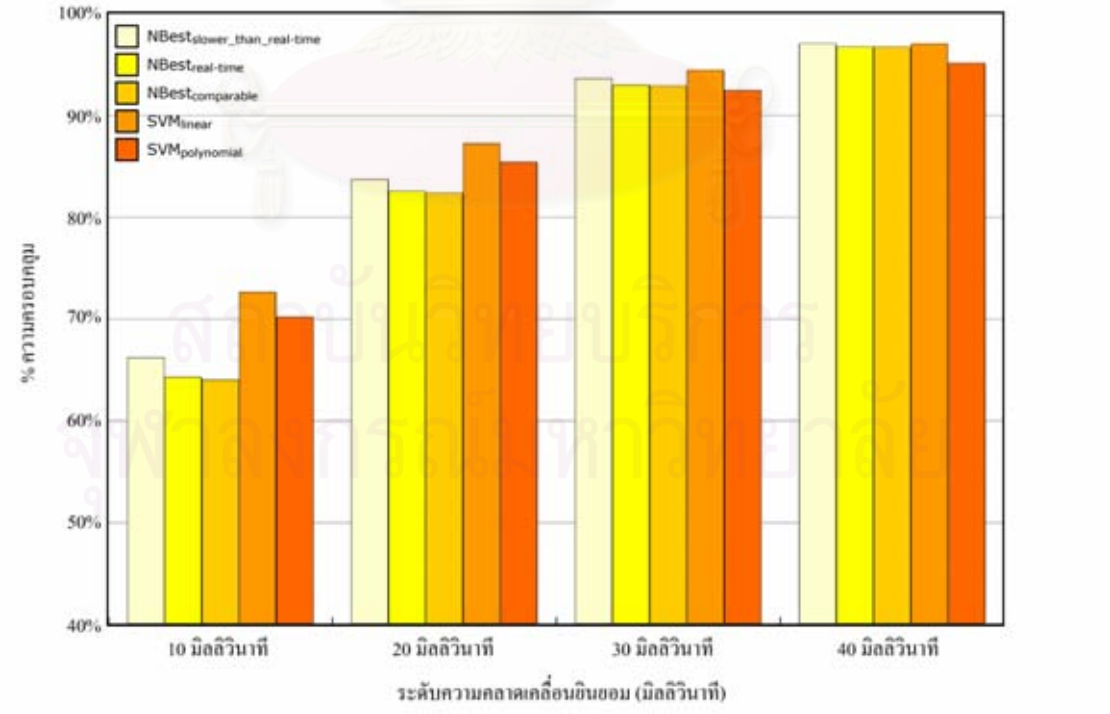
การเปรียบเทียบประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงของเครื่องตรวจหาขอบเขตของหน่วยเสียงทั้งห้าเครื่องในตารางที่ 4.7 ในด้านความแม่นยำและความครอบคลุมของขอบเขตของหน่วยเสียงที่หามาได้แสดงด้วยกราฟในรูปที่ 4.2 และ 4.3

ตารางที่ 4.7 เครื่องตรวจหาขอบเขตของหน่วยเสียงที่นำมาใช้เปรียบเทียบประสิทธิภาพ

เครื่องตรวจหาขอบเขตของหน่วยเสียง	N	RTF
$NBest_{comparable}$	70	0.57
$NBest_{real-time}$	90	0.98
$NBest_{slower\_than\_real-time}$	400	28.2
SVM_{linear}	-	0.57
$SVM_{polynomial}$	-	4.83



รูปที่ 4.2 กราฟแสดงเปอร์เซ็นต์ความแม่นยำในการตรวจหาขอบเขตของหน่วยเสียง



รูปที่ 4.3 กราฟแสดงเปอร์เซ็นต์ความครอบคลุมในการตรวจหาขอบเขตของหน่วยเสียง

2.5 วิเคราะห์ผลการทดลอง

จากการทดลองเปรียบเทียบประสิทธิภาพในการตรวจหาขอบเขตของหน่วยเสียง โดยเริ่มต้นจากการพิจารณาประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดจากกราฟในรูปที่ 4.1 จะพบว่าเมื่อเราปรับให้ N มีค่ามากขึ้น ก็จะทำให้ได้จำนวนขอบเขตของหน่วยเสียงเพิ่มมากขึ้น ส่งผลให้ขอบเขตของหน่วยเสียงที่หามาได้ครอบคลุมขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิงมากยิ่งขึ้นด้วย แต่ทั้งนี้ก็ต้องแลกเปลี่ยนกับเปอร์เซ็นต์ความแม่นยำที่ลดลงและเวลาในการทำงานที่เพิ่มมากขึ้นตามไปด้วย

จากการทดลองเปรียบเทียบประสิทธิภาพเครื่องตรวจหาขอบเขตของหน่วยเสียง 5 เครื่อง ในตารางที่ 4.7 จะพบว่าวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ สามารถตรวจหาขอบเขตของหน่วยเสียงได้แม่นยำและครอบคลุมขอบเขตของเสียงพูดที่เป็นตัวอ้างอิงมากกว่าขอบเขตของหน่วยเสียงที่ได้จากเครื่องรู้จำเสียงพูด อีกทั้งยังทำงานได้รวดเร็วกว่าด้วย โดยที่ระดับความคลาดเคลื่อนยินยอม 20 มิลลิวินาที การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ มีเปอร์เซ็นต์ความแม่นยำและเปอร์เซ็นต์ความครอบคลุมสูงกว่าการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดที่ทำงานได้เร็วพอๆกันอยู่ถึง 8.3% และ 5.1% ตามลำดับ และเมื่อเปรียบเทียบกับเครื่องรู้จำเสียงพูดที่ยอมเสียเวลาทำงานนานๆเพื่อให้ได้คำตอบที่ครอบคลุมมากขึ้นซึ่งมี $RTF = 28.2$ พบว่าการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนมีเปอร์เซ็นต์ความแม่นยำและเปอร์เซ็นต์ความครอบคลุมสูงกว่า 9.3% และ 3.52% ตามลำดับ ด้วยเหตุนี้จึงสามารถสรุปได้ว่า การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์มีประสิทธิภาพดีกว่าการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด ทั้งในด้านคุณภาพของขอบเขตของหน่วยเสียงที่หามาได้ และความเร็วในการทำงาน

และเมื่อพิจารณาประสิทธิภาพของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนด้วยกันเอง จากตารางที่ 4.4 4.5 และ 4.6 จะพบว่าความแม่นยำและความครอบคลุมของขอบเขตของหน่วยเสียงที่ได้จากเอสวีเอ็มแบบเชิงเส้นและแบบพหุนามอยู่ในระดับที่ใกล้เคียงกันมาก โดยที่ระดับความคลาดเคลื่อนยินยอม 20 มิลลิวินาที กว่า 76% ของขอบเขตของหน่วยเสียงที่ตรวจหามาได้นั้น ตรงกันกับขอบเขตของหน่วยเสียงที่เป็นตัวอ้างอิง และครอบคลุมกว่า 87% ของขอบเขตของหน่วยเสียงอ้างอิงทั้งหมด ทั้งนี้อาจเป็นเพราะข้อมูลที่ใช้ในการทดสอบอาจจะสามารถแบ่งแยกออกจากกันด้วยเส้นตรงได้ ทำให้การแมปไปยังปริภูมิที่สูงกว่าด้วยฟังก์ชันเคอร์เนลแบบพหุนามไม่มีผลมากนัก แต่เนื่องจากเอสวีเอ็มแบบพหุนามนั้นใช้เวลาในการทำงานมากกว่ามาก โดยที่ RTF ของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้

ฟังก์ชันเคอร์เนลแบบพหุนามมีค่าสูงถึง 4.83 ในระหว่างที่ *RTF* ของการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นมีค่าเพียง 0.57 ด้วยเหตุนี้จึงสามารถสรุปได้ว่า วิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นนั้นมีประสิทธิภาพดีกว่าวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบพหุนาม

เนื่องจากเราต้องการวิธีการตรวจหาขอบเขตของหน่วยเสียงที่สามารถทำงานได้รวดเร็ว ดังนั้นตัวเลือกที่เหมาะสมที่สุดในที่นี้ก็คือ การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้นเนื่องจากสามารถทำงานได้รวดเร็ว อีกทั้งยังสามารถตรวจหาขอบเขตของหน่วยเสียงที่มีความแม่นยำและความครอบคลุมใกล้เคียงกับการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลพหุนาม ด้วยเหตุนี้ในการทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์ในหัวข้อถัดไปจะนำขอบเขตของหน่วยเสียง ที่ได้จากวิธีตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้นไปใช้เท่านั้น

3. การทดลองเพื่อเปรียบเทียบประสิทธิภาพการสร้างกราฟของเซกเมนต์

การทดลองในที่นี้จะเป็นการทดลองวัดประสิทธิภาพในการสร้างกราฟของเซกเมนต์ โดยจะเปรียบเทียบวิธีการสร้างกราฟทั้งหมดสามวิธีตามที่เสนอไว้ในบทที่ 2 คือ วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง วิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และวิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับ โดยอาศัยขอบเขตของหน่วยเสียงที่ได้จากการตรวจหาขอบเขตของหน่วยเสียงทั้งแบบอาศัยเครื่องรู้จำเสียงพูดและแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีน แล้วนำกราฟของเซกเมนต์ที่ได้มาวัดประสิทธิภาพเปรียบเทียบกัน ซึ่งมีรายละเอียดการวัดประสิทธิภาพและผลการทดลองดังต่อไปนี้

3.1 การวัดประสิทธิภาพ

การแบ่งเสียงพูดเป็นเซกเมนต์ที่ดีนั้นจะต้องสามารถสร้างกราฟของเซกเมนต์ที่มีคุณภาพรวมถึงทำงานได้อย่างรวดเร็ว ความหมายคุณภาพในที่นี้คือกราฟของเซกเมนต์ที่มีขนาดเล็ก และครอบคลุมเซกเมนต์ของหน่วยเสียงอ้างอิง ในส่วนของความเร็วในการแบ่งเสียงพูดเป็นเซกเมนต์เนื่องจากเวลาที่ในการทำงานเฉพาะขั้นตอนในการสร้างกราฟของเซกเมนต์นั้นอยู่ในระดับที่ยอมกันได้ โดยใช้เวลาในการทำงานสั้นมากเมื่อเทียบกับขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง ดังนั้นการเปรียบเทียบความเร็วในการแบ่งเสียงพูดเป็นเซกเมนต์จึงจะพิจารณาจากความเร็วในการตรวจหาขอบเขตของหน่วยเสียงเป็นหลัก

3.1.1 ขนาดของกราฟของเซกเมนต์

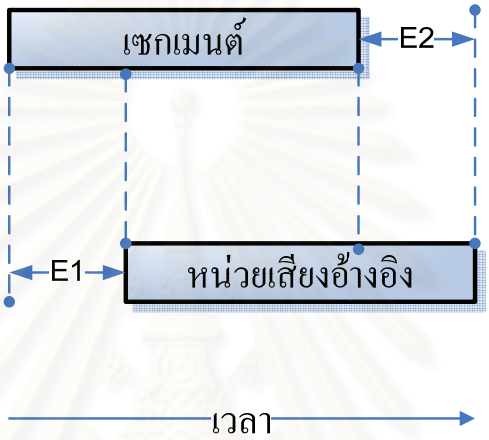
การวัดขนาดของกราฟของเซกเมนต์จะวัดจากจำนวนเซกเมนต์ที่พบในกราฟต่อหนึ่งหน่วยวินาที โดยค่านี้จะสะท้อนให้เห็นขนาดของกราฟของเซกเมนต์ที่จะนำไปใช้ต่อในกระบวนการค้นหาและให้คะแนนเพื่อการรู้จำเสียงพูดในขอบข่ายงานของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ต่อไป โดยหากมีปริมาณมากเกินไปก็จะทำให้กราฟของเซกเมนต์มีขนาดใหญ่ แต่ถ้าหากมีน้อยเกินไปก็อาจจะได้จำนวนเซกเมนต์ที่ไม่ครอบคลุมหน่วยเสียงในระดับที่ต้องการ

3.1.2 ความครอบคลุม

การพิจารณาความครอบคลุมของกราฟของเซกเมนต์ จะต้องใช้ผลของการตัดสินใจว่าเซกเมนต์ที่หามาได้นั้นมีความถูกต้องตรงกับหน่วยเสียงที่เป็นตัวอ้างอิงหรือไม่ โดยเซกเมนต์ที่ดี

นั้นไม่จำเป็นต้องอยู่ที่ตำแหน่งเดียวกันกับหน่วยเสียงที่เป็นตัวอ้างอิงก็ได้ โดยอาจยินยอมให้คลาดเคลื่อนกันได้เป็นระยะเวลาสั้นๆในระดับมิลลิวินาที ค่าความคลาดเคลื่อนของเซกเมนต์จะคิดจากผลรวมระหว่างค่าความคลาดเคลื่อนของขอบเขตทางด้านซ้ายของเซกเมนต์เทียบกับขอบเขตทางด้านซ้ายของหน่วยเสียงอ้างอิง และค่าความคลาดเคลื่อนของขอบเขตทางด้านขวาของเซกเมนต์เทียบกับขอบเขตทางด้านขวาของหน่วยเสียงอ้างอิง โดยแสดงได้ด้วยรูปที่ 4.4

$$\text{ค่าความคลาดเคลื่อน} = E1 + E2$$



รูปที่ 4.4 การวัดความคลาดเคลื่อนของเซกเมนต์

เปอร์เซ็นต์ความครอบคลุมเขียนแทนด้วยสัญลักษณ์ $\%Re$ สามารถคำนวณได้จากสมการต่อไปนี้

$$\%Re = 100 \times \frac{C}{T}$$

เมื่อกำหนดให้ T คือจำนวนหน่วยเสียงที่เป็นตัวอ้างอิง และ C คือจำนวนเซกเมนต์ทั้งหมดที่ถูกต้องอยู่ตรงกับหน่วยเสียงที่เป็นตัวอ้างอิงโดยยินยอมให้มีค่าคลาดเคลื่อนไปจากหน่วยเสียงที่เป็นตัวอ้างอิงได้ไม่เกินระดับความคลาดเคลื่อนยินยอมที่ 10, 20, 30 และ 40 มิลลิวินาที ตามลำดับ

3.2 ผลการทดลอง

ประสิทธิภาพของการสร้างกราฟของเซกเมนต์ ในด้านความครอบคลุมและขนาดของกราฟแสดงไว้ในตารางที่ 4.8 4.9 และ 4.10

ตารางที่ 4.8 คุณภาพกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

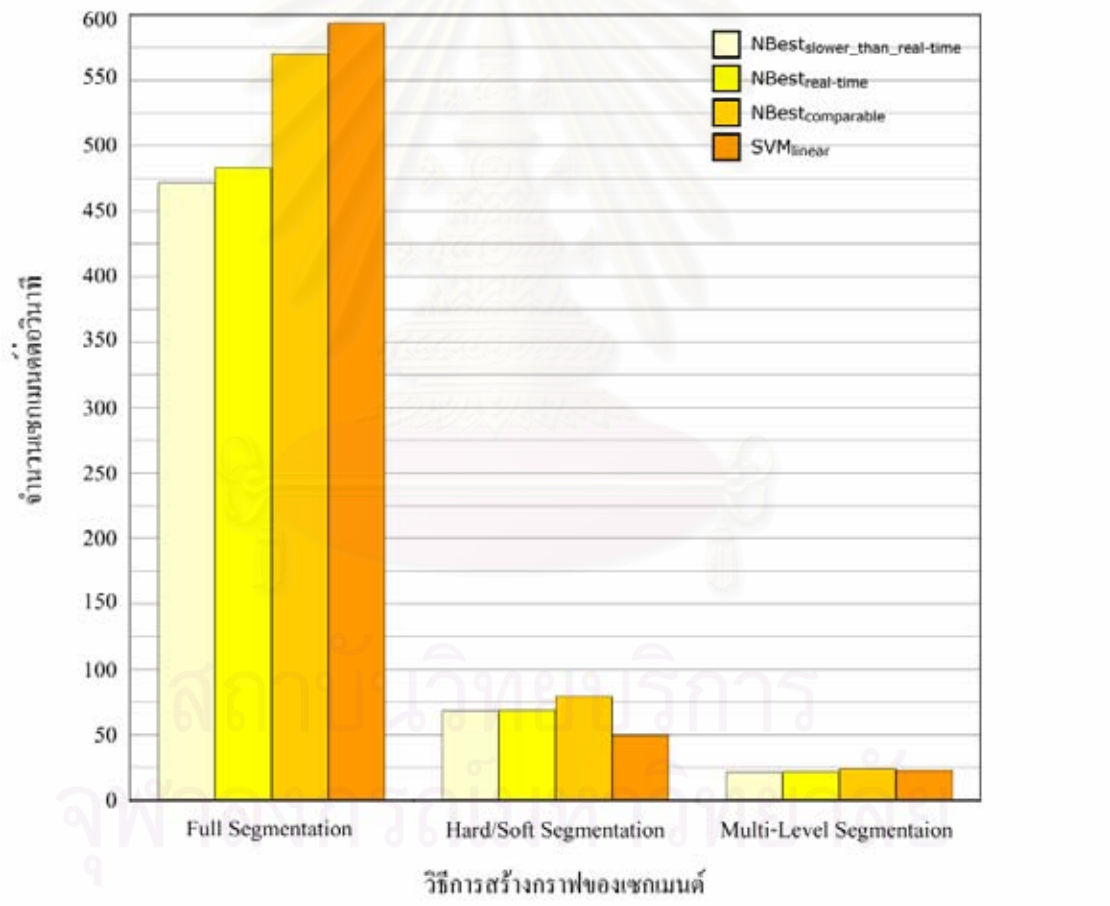
	เปอร์เซ็นต์ความครอบคลุมของเซกเมนต์ที่ระดับความคลาดเคลื่อน (มิลลิวินาที)				จำนวนเซกเมนต์ต่อวินาที
	10	20	30	40	
$NBest_{comparable}$	40.01	63.34	82.83	91.28	471.4
$NBest_{real-time}$	40.40	64.71	83.10	91.41	483.1
$NBest_{slower\ than\ real-time}$	43.10	67.09	84.72	92.39	570.0
SVM_{linear}	50.95	74.80	88.98	98.63	593.7

ตารางที่ 4.9 คุณภาพกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

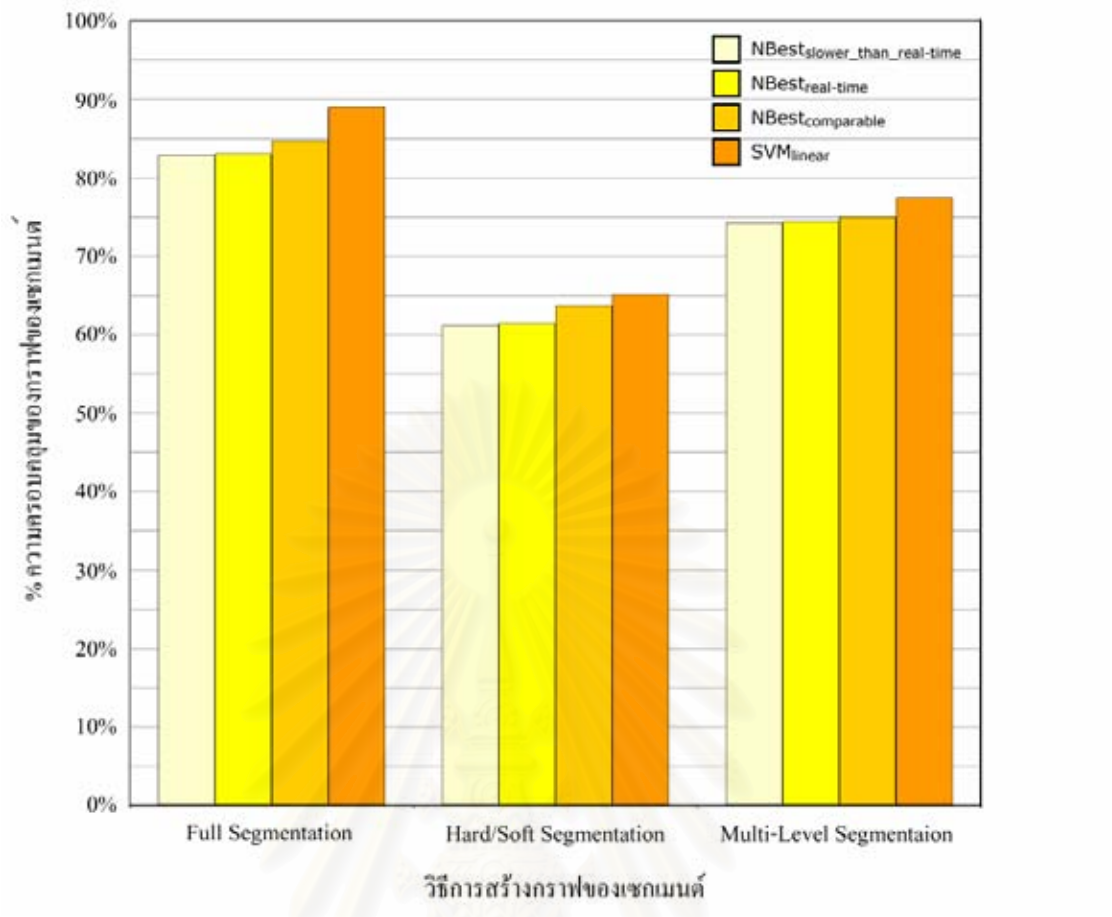
	เปอร์เซ็นต์ความครอบคลุมของเซกเมนต์ที่ระดับความคลาดเคลื่อน (มิลลิวินาที)				จำนวนเซกเมนต์ต่อวินาที
	10	20	30	40	
$NBest_{comparable}$	27.02	45.72	61.13	68.76	68.0
$NBest_{real-time}$	27.30	46.02	61.40	69.00	68.7
$NBest_{slower\ than\ real-time}$	29.32	48.43	63.65	71.00	79.1
SVM_{linear}	32.90	51.85	65.02	74.82	49.3

ตารางที่ 4.10 คุณภาพกราฟของเซกเมนต์แบบหลายระดับ

	เปอร์เซ็นต์ความครอบคลุมของเซกเมนต์ที่ระดับความคลาดเคลื่อน (มิลลิวินาที)				จำนวนเซกเมนต์ต่อวินาที
	10	20	30	40	
$NBest_{comparable}$	32.00	53.90	74.21	85.61	21.1
$NBest_{real-time}$	32.10	54.13	74.40	85.70	21.5
$NBest_{slower\ than\ real-time}$	32.44	54.79	74.98	86.02	23.8
SVM_{linear}	39.58	65.20	77.40	89.16	22.4



รูปที่ 4.5 กราฟเปรียบเทียบขนาดของกราฟของเซกเมนต์ที่สร้างด้วยวิธีต่างๆ



รูปที่ 4.6 กราฟเปรียบเทียบเปอร์เซ็นต์ครอบคลุมของกราฟของเซกเมนต์
ที่ระดับความคลาดเคลื่อน 30 มิลลิวินาที

3.3 วิเคราะห์ผลการทดลอง

จากการทดลองเปรียบเทียบประสิทธิภาพในการสร้างกราฟของเซกเมนต์ทั้งสามวิธี โดยดูจากกราฟในรูปที่ 4.5 จะพบว่ากราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงมีขนาดใหญ่มากเมื่อเทียบกับกราฟที่ได้จากวิธีอื่นๆ โดยกราฟของเซกเมนต์ที่มีจำนวนเซกเมนต์สูงถึงประมาณ 500 ถึง 600 เซกเมนต์ต่อวินาทีในระหว่างที่กราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมมีจำนวนเซกเมนต์อยู่ในระดับ 50 – 80 เซกเมนต์ต่อวินาที ซึ่งมีขนาดเล็กกว่าประมาณ 7 – 8 เท่า และกราฟของเซกเมนต์แบบหลายระดับซึ่งมีขนาดเล็กที่สุด โดยมีจำนวนเซกเมนต์อยู่ประมาณ 20 – 24 เซกเมนต์ต่อวินาทีซึ่งมีขนาดเล็กกว่ากราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงประมาณ 14 – 15 เท่า และจากการทดลองเปรียบเทียบประสิทธิภาพในการสร้างกราฟของเซกเมนต์ทั้งสามวิธี โดยดูจากกราฟในรูปที่ 4.6 จะพบว่ากราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงจะมีเปอร์เซ็นต์ความครอบคลุมสูงที่สุด โดยที่ระดับความคลาดเคลื่อน 30 มิลลิวินาที กราฟที่ได้จะมีเปอร์เซ็นต์ความครอบคลุมโดยประมาณอยู่ในระดับ 82 –

89% รองลงมาจะเป็นกราฟของเซกเมนต์แบบหลายระดับซึ่งมีเปอร์เซ็นต์ความครอบคลุมอยู่ในระดับ 74 – 77% ในระหว่างที่กราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมมีมีเปอร์เซ็นต์ความครอบคลุมต่ำที่สุดอยู่ในระดับ 60 – 65% แม้ว่ากราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงจะมีเปอร์เซ็นต์ความครอบคลุมสูงที่สุดแต่ก็มีขนาดใหญ่ที่สุดด้วย ซึ่งอาจไม่เหมาะสมต่อการนำไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ต้องการความรวดเร็วในการทำงาน และเมื่อใช้วิธีการสร้างกราฟแบบอาศัยการเปลี่ยนแปลงสเปกตรัมมาเพื่อลดขนาดของกราฟลงก็พบว่าเปอร์เซ็นต์ความครอบคลุมลดลงไปกว่า 20% เมื่อเปรียบเทียบกับวิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง ซึ่งจะไปทำให้ความถูกต้องในการรู้จำเสียงพูดลดลงไปด้วย แต่เมื่อใช้วิธีการสร้างกราฟแบบหลายชั้นมาใช้จะพบว่าสามารถลดขนาดของกราฟได้มากกว่า อีกทั้งยังได้เปอร์เซ็นต์ความครอบคลุมสูงกว่าวิธีการสร้างกราฟแบบอาศัยการเปลี่ยนแปลงสเปกตรัมประมาณ 10 – 15%

ด้วยเหตุนี้จึงสรุปได้ว่า วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับนั้นสามารถนำมาใช้สร้างกราฟของเซกเมนต์ที่มีคุณภาพดีกว่าวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และเมื่อพิจารณาความต้องการที่จะนำการแบ่งเสียงพูดเป็นเซกเมนต์นี้ไปใช้ในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่จะต้องทำงานได้รวดเร็วแบบทันกาล วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับจึงมีความเหมาะสมกว่าวิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียง

สรุปผลการทดลอง

การทดลองการแบ่งเสียงพูดเป็นเซกเมนต์ก็เพื่อเป็นการเปรียบเทียบประสิทธิภาพของขั้นตอนย่อยของการแบ่งเสียงพูดเป็นเซกเมนต์ซึ่งประกอบไปด้วยสองขั้นตอน คือ ขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงและขั้นตอนการสร้างกราฟของเซกเมนต์ โดยสามารถสรุปการเปรียบเทียบวิธีการที่ใช้ในขั้นตอนต่างๆได้ดังแผนภาพในรูปที่ 4.7

รูปที่ 4.7 แผนภาพเปรียบเทียบวิธีการที่ใช้ในแต่ละขั้นตอนของการแบ่งเสียงพูดเป็นเซกเมนต์

ในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง งานวิจัยนี้เสนอวิธีการตรวจหาขอบเขตของเสียงพูดแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ มาเปรียบเทียบกับวิธีตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด จากการวิเคราะห์ผลการทดลองเปรียบเทียบประสิทธิภาพของทั้งสองวิธีพบว่าวิธีการตรวจหาขอบเขตของเสียงพูดแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ มีประสิทธิภาพดีกว่าทั้งในด้านคุณภาพของขอบเขตของหน่วยเสียงที่ได้ และความเร็วในการทำงาน

ในขั้นตอนการสร้างกราฟของเซกเมนต์ งานวิจัยนี้เสนอวิธีการสร้างกราฟของเซกเมนต์สามวิธีได้แก่ วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงซึ่งเป็นวิธีดั้งเดิมที่ใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด วิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และวิธีการสร้างกราฟแบบหลายระดับ จากการวิเคราะห์ผลการทดลองพบว่า วิธีการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียงสามารถสร้างกราฟของเซกเมนต์ที่ได้ความครอบคลุมมากที่สุดแต่กราฟนั้นจะมีขนาดใหญ่มากไม่

เหมาะกับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่ต้องการความรวดเร็วในการทำงาน และจากการวิเคราะห์ผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีการสร้างกราฟอีกสองวิธีที่เหลือ พบว่าสามารถลดขนาดของกราฟลงได้มาก แต่วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับนั้นมีประสิทธิภาพดีกว่าวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมเนื่องจากสามารถลดขนาดของกราฟลงได้มากกว่าอีกทั้งยังสามารถรักษาระดับความครอบคลุมไว้ได้มากกว่าด้วยเหตุนี้เมื่อพิจารณาความจำเป็นของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่จะต้องทำงานได้แบบทันกาล วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับจะมีความเหมาะสมที่สุด

ด้วยเหตุนี้เมื่อพิจารณาผลการทดลองเปรียบเทียบประสิทธิภาพของวิธีการต่างๆที่นำมาใช้ในแต่ละขั้นตอน จะสามารถสรุปได้ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่ใช้วิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์ และใช้วิธีการสร้างกราฟแบบหลายระดับ มีประสิทธิภาพดีที่สุด ซึ่งสามารถสร้างกราฟของเซกเมนต์ที่มีคุณภาพดีกว่า อีกทั้งยังทำงานได้รวดเร็วกว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด

ด้วยข้อจำกัดของวิธีการตรวจหาขอบเขตของหน่วยเสียง ในบางครั้งขอบเขตของหน่วยเสียงที่ตรวจหามาได้จะอยู่ติดกันมาก โดยมีระยะห่างระหว่างขอบเขตของหน่วยเสียงที่สั้นประมาณ 10 มิลลิวินาที ทำให้เซกเมนต์ที่สร้างจากขอบเขตของหน่วยเสียงนี้มีขนาดเล็กเกินกว่าจะเป็นหน่วยเสียงได้ ซึ่งหากนำเอาขอบเขตของหน่วยเสียงนี้ไปใช้ต่อก็จะส่งผลต่อคุณภาพกราฟของเซกเมนต์โดยตรง

การทดลองจำแนกลักษณะการออกเสียงในงานวิจัยนี้ ทำการทดลองกับเฉพาะซัพพอร์ตเวกเตอร์แมชชีนที่มีฟังก์ชันเคอร์เนลแบบเชิงเส้นและแบบพหุนามเท่านั้น ไม่ได้ทดลองกับซัพพอร์ตเวกเตอร์แมชชีนที่มีฟังก์ชันเคอร์เนลแบบอื่นๆอย่างครบถ้วน และด้วยข้อจำกัดทางด้านทรัพยากรเกี่ยวกับข้อมูลเสียงที่นำมาใช้ในการทดลอง ซึ่งจะต้องมีการกำกับขอบเขตของหน่วยเสียงไว้ก่อนแล้ว ทำให้ไม่สามารถนำข้อมูลเสียงพูดในฐานข้อมูลเสียงโลตัสมาใช้ได้ทั้งหมด ทำให้เลือกใช้ได้เฉพาะข้อมูลเสียงชุดหน่วยเสียงสมดุล ซึ่งเป็นชุดข้อมูลเสียงเพียงชุดเดียวในฐานข้อมูลเสียงโลตัสที่มีการกำกับขอบเขตของหน่วยเสียงไว้

บทที่ 5
สรุปผลการวิจัย

ผลการวิจัย

การแบ่งเสียงพูดเป็นเซกเมนต์ประกอบไปด้วยขั้นตอนย่อยสองขั้นตอน คือ ขั้นตอนการตรวจหาขอบเขตของหน่วยเสียงและขั้นตอนการสร้างกราฟของเซกเมนต์ วิทยานิพนธ์นี้นำเสนอวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบซัพพอร์ตเวกเตอร์แมชชีนโดยใช้สารสนเทศสวนสัทศาสตร์เพื่อนำไปใช้กับระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ โดยเปรียบเทียบประสิทธิภาพกับวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูดที่ใช้อยู่เดิม ด้วยความหวังที่ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่เสนอนี้จะสามารถทำงานได้รวดเร็วกว่าและสร้างกราฟของเซกเมนต์ที่มีคุณภาพสูงกว่าได้ ในหัวข้อนี้จะนำเสนอผลการวิจัยซึ่งแบ่งออกเป็นสองส่วนคือ การตรวจหาขอบเขตของหน่วยเสียง และการสร้างกราฟของเซกเมนต์ซึ่งมีรายละเอียดดังนี้

1. การตรวจหาขอบเขตของหน่วยเสียง

กระบวนการตรวจหาขอบเขตของหน่วยเสียงตามวิธีที่เสนอนี้ เริ่มต้นจากการจำแนกหน่วยเสียงออกเป็นประเภทตามลักษณะการออกเสียง โดยอาศัยซัพพอร์ตเวกเตอร์แมชชีนมาช่วยแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบ ซึ่งการแบ่งแยกนี้จะใช้ลักษณะสำคัญของเสียงพูดที่ได้จากการใช้ความรู้และสารสนเทศทางสวนสัทศาสตร์ หลังจากนั้นจึงนำผลการจำแนกลักษณะการออกเสียงมาแยกเอาข้อมูลที่เป็นขอบเขตของหน่วยเสียงออกมา แล้วคำนวณค่าคะแนนความมั่นใจให้กับขอบเขตของหน่วยเสียงที่หามาได้ โดยวิธีการนี้มีจุดเด่นสามประการคือ

1. ในการตรวจหาขอบเขตของหน่วยเสียงด้วยวิธีการนี้ การแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบจะสามารถใช้ลักษณะสำคัญที่แตกต่างกันได้ ทำให้สามารถเลือกใช้ลักษณะสำคัญที่เหมาะสมกับการจำแนกสัทลักษณ์ลักษณะสำคัญแต่ละประเภทได้อย่างอิสระ
2. การตรวจหาขอบเขตของหน่วยเสียงด้วยวิธีการนี้ สามารถเพิ่มเติมลักษณะสำคัญที่น่าจะเป็นประโยชน์ต่อการแบ่งแยกสัทลักษณ์แต่ละแบบได้โดยง่าย ทำให้สามารถเพิ่มประสิทธิภาพการแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบให้มีความแม่นยำมากขึ้นได้ หากอาศัยสารสนเทศสวนสัทศาสตร์มาวิเคราะห์เลือกใช้ลักษณะสำคัญเพิ่ม
3. การตรวจหาขอบเขตของหน่วยเสียงด้วยวิธีการนี้มีการคิดคะแนนให้กับขอบเขตของหน่วยเสียงที่หามาได้ ค่าคะแนนนี้จะแทนระดับความมั่นใจของการที่ขอบเขตของ

หน่วยเสียงจะตรงกันกับขอบเขตของหน่วยเสียงอ้างอิง ซึ่งสามารถนำไปใช้ประโยชน์ต่อในขั้นตอนการรู้จำเสียงพูดเพื่อเพิ่มประสิทธิภาพการรู้จำเสียงพูดต่อไปได้

และจากการทดลองตรวจหาขอบเขตของหน่วยเสียง โดยเมื่อเปรียบเทียบวิธีการตรวจหาขอบเขตของหน่วยเสียงที่เสนอกับวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด พบว่าวิธีการตรวจหาขอบเขตของหน่วยเสียงที่เสนอนี้มีประสิทธิภาพดีกว่าทั้งในทางด้านคุณภาพของขอบเขตของหน่วยเสียงที่ตรวจหามาได้และความเร็วในการทำงาน โดยที่ระดับความคลาดเคลื่อนยินยอม 20 มิลลิวินาที ขอบเขตของหน่วยเสียงที่ได้จากวิธีการที่เสนอนี้มีเปอร์เซนต์ความแม่นยำและความครอบคลุมสูงกว่าขอบเขตของหน่วยเสียงที่ได้จากวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดที่ทำงานได้เร็วพอๆกัน อยู่ถึง 8.3% และ 5.1% ตามลำดับ และเมื่อเปรียบเทียบกับวิธีการตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูดที่ยอมเสียเวลาทำงานนานๆเพื่อให้ได้คำตอบที่ครอบคลุมมากขึ้น พบว่าวิธีการตรวจหาขอบเขตของหน่วยเสียงที่เสนอนี้ยังสามารถตรวจหาขอบเขตของหน่วยเสียงได้ความแม่นยำและความครอบคลุมสูงกว่าถึง 9.3% และ 3.52% ตามลำดับ

2. การสร้างกราฟของเซกเมนต์

วิทยานิพนธ์นี้ได้เสนอวิธีการสร้างกราฟของเซกเมนต์สามวิธีได้แก่ วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงซึ่งเป็นวิธีดั้งเดิมที่ใช้ในการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูดที่นำมาใช้เปรียบเทียบกับวิธีการสร้างกราฟอีกสองแบบที่เหลือ วิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม และวิธีการสร้างกราฟแบบหลายระดับ

วิธีการสร้างกราฟของเซกเมนต์แบบเชื่อมต่อทุกขอบเขตของหน่วยเสียงมีข้อดีอยู่ที่ การสร้างกราฟแบบนี้จะเชื่อมต่อขอบเขตของหน่วยเสียงครบทุก ทำให้กราฟที่ได้มีความครอบคลุมสูงกว่าวิธีอื่นๆทั้งหมด แต่มีข้อเสียสำคัญคือกราฟนั้นจะมีขนาดใหญ่มาก ทำให้ขั้นตอนการรู้จำเสียงพูดซึ่งนำกราฟนี้ไปใช้ต่อจะต้องเสียเวลาในการทำงานมากตามไปด้วย ส่วนวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมมีข้อดีตรงที่สามารถลดขนาดของกราฟได้เป็นจำนวนมากอีกทั้งยังใช้เวลาในการทำงานน้อยมากเมื่อเทียบกับเวลาที่ใช้ในขั้นตอนการตรวจหาขอบเขตของหน่วยเสียง แต่มีข้อเสียคือกราฟที่ได้จะมีความครอบคลุมลดลงมาก ในระหว่างที่วิธีการสร้างกราฟแบบหลายระดับมีข้อได้เปรียบตรงที่สามารถลดขนาดของกราฟได้มากกว่า อีกทั้งยังรักษาระดับความครอบคลุมไว้ได้มากกว่ากราฟที่ได้จากวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม

โดยในการทดลองเปรียบเทียบประสิทธิภาพวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมกับวิธีการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียง พบว่า กราฟที่ได้จากวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัมมีขนาดที่เล็กกว่า 7 – 8 เท่าโดยประมาณ แต่มีความครอบคลุมน้อยกว่ากราฟที่ได้จากการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียงถึง 23.96% (จาก 88.98% เป็น 65.02%) และในการทดลองเปรียบเทียบประสิทธิภาพวิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับกับวิธีการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียงพบว่า กราฟที่ได้จากวิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับมีขนาดที่เล็กกว่าประมาณ 14 เท่าซึ่งเล็กกว่ากราฟที่ได้จากวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม ในขณะที่มีความครอบคลุมน้อยกว่ากราฟที่ได้จากการสร้างกราฟของเซกเมนต์แบบจับคู่ทุกขอบเขตของหน่วยเสียงประมาณ 11.58% (จาก 88.98% เป็น 77.40%) ซึ่งลดลงไปน้อยกว่ากราฟที่ได้จากวิธีการสร้างกราฟของเซกเมนต์แบบอาศัยการเปลี่ยนแปลงสเปกตรัม ดังนั้นเมื่อพิจารณาความจำเป็นของระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ที่จะต้องทำงานได้อย่างรวดเร็ว วิธีการสร้างกราฟของเซกเมนต์แบบหลายระดับจึงมีความเหมาะสมที่สุด

โดยสรุปแล้ววิทยานิพนธ์นี้เสนอวิธีการแบ่งเสียงพูดเป็นเซกเมนต์สำหรับการรู้จำเสียงพูดแบบอาศัยเซกเมนต์แบบใหม่ ที่นำเอาสารสนเทศสวนสัทศาสตร์และซัพพอร์ตเวกเตอร์แมชชีนมาประยุกต์ใช้เพื่อให้ประสิทธิภาพทั้งในด้านคุณภาพของผลลัพธ์ และความเร็วในการทำงานดีกว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด ดังนั้นวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ที่เสนอนี้ จึงสามารถนำไปใช้งานในระบบรู้จำเสียงพูดแบบอาศัยเซกเมนต์ได้เป็นอย่างดีและส่งผลให้การรู้จำเสียงพูดมีความถูกต้องมากยิ่งขึ้น นอกจากนี้วิธีการแบ่งเสียงพูดที่เสนอยังอาจนำไปประยุกต์ใช้ในการเตรียมข้อมูลเสียงสำหรับระบบสังเคราะห์เสียงพูด หรือนำไปเตรียมข้อมูลฝึกฝนสำหรับเครื่องรู้จำเสียงพูดอื่นๆได้

ข้อเสนอแนะ

แม้ว่าวิธีการแบ่งเสียงพูดเป็นเซกเมนต์โดยใช้สารสนเทศสวนสัทศาสตร์ที่เสนอไว้ในวิทยานิพนธ์นี้ จะมีประสิทธิภาพดีกว่าทั้งในด้านคุณภาพของผลลัพธ์และความเร็วในการทำงานเมื่อเปรียบเทียบกับวิธีการแบ่งเสียงพูดเป็นเซกเมนต์แบบอาศัยเครื่องรู้จำเสียงพูด แต่ประสิทธิภาพที่ได้นั้น ยังคงห่างจากประสิทธิภาพของวิธีการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยคนอยู่ ดังนั้นงานวิจัยนี้จึงยังสามารถพัฒนาได้อีกในหลายทาง

การเพิ่มขั้นตอนการกรองเซกเมนต์ที่มีขนาดเล็กเกินกว่าจะเป็นหน่วยเสียงได้ออกไป ก็เป็นหนทางหนึ่งในการเพิ่มประสิทธิภาพของการแบ่งเสียงพูดเป็นเซกเมนต์ โดยอาจเลือกกรองเฉพาะเซกเมนต์ที่มีระยะห่างระหว่างขอบเขตด้านซ้ายและด้านขวาซึ่งสั้นกว่า 10 มิลลิวินาที เพื่อทำให้กราฟของเซกเมนต์ที่ได้มีขนาดเล็กลง ทำให้กราฟมีคุณภาพมากยิ่งขึ้น

นอกจากนี้อาจทำการทดลองใช้ซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบอื่นๆ นอกเหนือไปจากซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นและแบบพหุนาม เช่น ซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ฟังก์ชันเคอร์เนลแบบอาร์บีเอฟ มาใช้จำแนกเสียงพูดตามลักษณะการออกเสียง ซึ่งอาจจะสามารถแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงได้แม่นยำมากยิ่งขึ้น ส่งผลให้ได้ขอบเขตของหน่วยเสียงที่มีคุณภาพเพิ่มมากขึ้นตามไปด้วย

และส่วนสำคัญอีกประการหนึ่งของการปรับปรุงประสิทธิภาพการแบ่งเสียงพูดเป็นเซกเมนต์ด้วยวิธีการนี้คือ การใช้ลักษณะสำคัญเพิ่มเติมโดยใช้สารสนเทศสวนสัทศาสตร์มาประกอบการพิจารณาสำหรับนำไปใช้ในการแบ่งแยกสัทลักษณ์ลักษณะการออกเสียงแต่ละแบบให้เหมาะสม เพื่อช่วยให้สามารถตรวจหาขอบเขตของหน่วยเสียงได้แม่นยำมากยิ่งขึ้น ทำให้ได้กราฟของเซกเมนต์ที่นำขอบเขตของหน่วยเสียงนี้ไปใช้มีคุณภาพเพิ่มมากขึ้นด้วย

รายการอ้างอิง

[1] Glass, J.R. A probabilistic frameworkfor segment-based speech recognition. Computer Speech and Language 17 (2003): 137–152.

[2] Kvale, K. On the Connection Between Manual Segmentation Conventions and Errors Made by Automatic Segmentation. Proceedings of ICSLP'94, pp. 1667–1670, 1994.

[3] Zue, V., Glass, J.R., Phillips, M., and Sene, S. The MIT SUMMIT speech recognition system: a progress report. Proceedings of the Speech and Natural Language Workshop, pp. 179–189, 1989.

[4] Lee, S.C. Probabilistic Segmentation for Segment-Based Speech Recognition. Master's Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, 1998.

[5] กาญจนา นาคสกุล, ระบบเสียงภาษาไทย, พิมพ์ครั้งที่ 4, โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2541.

[6] Ladefoged, P. A Course in Phonetics. Harcourt Brace Jovanovich Inc., 1975.

[7] Stevens, K.N. Acoustic Phonetics. Cambridge, MA: MIT Press, 1999.

[8] อุปกิตศิลปสาร, พระยา, หลักภาษาไทย. ไทยวัฒนาพานิช, 2533.

[9] Chomsky, N., Halle, N. The Sound Pattern of English. Cambridge, MA: MIT Press, 1968.

[10] Rabiner, L.R., Juang, B.H. Fundamentals of Speech Recognition. A. Oppenheim, Series Editor, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[11] Mammone, R.J., Zhang, X., and Ramachandran, R.P. Robust Speaker Recognition, A Feature-based Approach, IEEE Signal Processing Magazine, pp. 58-71, 1996.

[12] Furui, S. Digital Speech Processing, Synthesis, and Recognition. New York and Basel: Marcel Dekker Inc., 1989.

[13] Roe, D.B., Wilpon, J.G. Whither Speech Recognition: the next 25 years, IEEE Comm. Magazine, Vol. 31, 1993.

[14] Ostendorf, M., Digalakis, V., and Kimball, O. From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. IEEE Transactions on Speech and Audio Proceesing, pp. 360-378, 1996.

[15] บุญเสริม กิจศิริกุล, การเรียนรู้ของเครื่อง. โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2549

[16] Vapnik, V. Statistical Learning Theory. New York: Wiley.

[17] Kipp, A., Wesenick, M. B., and Schiel, F. Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. Proc. of ICSLP'96, pp. 106-109, 1996.

[18] Brugnara, F., Falavigna, D., and Omologo, M. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. Speech Communication (1993): 357-370.

[19] Hosom, J.P. Automatic Time Alignment of Phonemes Using Acousitc-Phonetic Information. Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, 2000.

[20] Kim, Y.J., and Conkie, A. Automatic Segmentation Combining an HMM-based Approach and Spectral Boundary Correction. In Proc. ICSLP'2002, pp. 145-148, 2002.

[21] Leelaphattarakij, P., Punyabukkana, P., and Suchato, A. Locating Phone Boundaries from Acoustic Discontinuities using a Two-staged Approach. The Ninth International Conference on Spoken Language Processing: Interspeech2006, Pittsburgh, PA, 2006.

[22] Chang, J., Glass, J.R. Segmentation and modeling in segment-based recognition, Proc. Eurospeech'97, pp. 1199-1202, 1997.

[23] Cole, R., Fanty, M. Spoken letter recognition. In Proc. Third DARPA Speech and Natural Language Workshop, pp. 385-390, 1990.

[24] Wang, D., Lu, L., and Zhang H.-J. Speech Segmentation without Speech Recognition. Proc. of IEEE ICASSP'03, pp. 468-471, 2003.

[25] Glass, J.R., Zue, V.W. Multi-Level Acoustic Segmentation of Continuous Speech. Proc. of IEEE ICASSP'88, pp. 429-432, 1988.

[26] Glass, J.R. Finding Acoustic Regularities in Speech: Application to Phonetic Recognition. Ph.D Thesis, MIT press, 1988.

[27] Sorin, D., Rabiner, L.R. On the Relation between Maximum Spectral Transition Positions and Phone Boundaries. The Ninth International Conference on Spoken Language Processing: Interspeech2006, Pittsburgh, PA, 2006.

[28] Kasuriyam, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., and Thatphithakkul, N. Thai Speech Corpus for Thai Speech Recognition. Proc. COCOSDA'03, pp. 54-61, 2003.

[29] Young, S., Jansen, J., Odell, J. Ollason, D., and Woodland, P. The HTK Book (for HTK Version 3.3). Cambridge, 2005.

[30] Makashay, M. J., Wightman, C. W., Syrdal, A. K., and Conkie, A. Perceptual evaluation of automatic segmentation in text-to-speech synthesis. In Proc. ICLSP2000, pp. 431-434, 2000.

[31] Kominek, J. K., Bennett, C., and Black, A. W. Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In Proc. EUROSPEECH'2003, pp. 313-316, 2003.

[32] Kawai, H., and Toda, T. An evaluation of automatic phone segmentation for concatenative speech synthesis. In Proc. ICASSP'2004, pp.677–680, Quebec, Canada, May 2004.

[33] Huckvale, M. Speech Filing System Vs3.0 – Computer Tools for Speech Research, London, U.K.: University College, 1996.

[34] Scholkopf, B. Smola, A. J., and Muller, K. R. Kernel principal component analysis. Advances in Kernel Methods - Support Vector Learning. MIT Press, 1999.

ภาคผนวก

ใหหัวข้อนี้จำนำเสนอเกี่ยวกับขั้นตอนการเรียนรู้และการรู้จำของเครื่องรู้จำเสียงพูดที่นำมาใช้ในการตรวจหาขอบเขตของหน่วยเสียง การสกัดลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ รวมถึงนำเสนอเกี่ยวกับการเรียนรู้และการจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน

ขั้นตอนการเรียนรู้และการรู้จำของเครื่องรู้จำเสียงพูด

ในวิทยานิพนธ์นี้ทำการสร้างเครื่องรู้จำเสียงพูดโดยใช้โปรแกรม Hidden Markov Toolkit - HTK [29] ซึ่งเป็นชุดเครื่องมือสำหรับสร้างเครื่องรู้จำเสียงพูดแบบอาศัยแบบจำลองฮิดเดนมาร์คอฟ ในหัวข้อนี้จะนำเสนอเกี่ยวกับขั้นตอนการการเรียนรู้เสียงพูด การรู้จำเสียงพูด และการตรวจหาขอบเขตของหน่วยเสียงจากผลการรู้จำเสียงพูด

1. กระบวนการเรียนรู้

กระบวนการเรียนรู้มีขั้นตอนต่างๆดังนี้

1.1 การหาลักษณะสำคัญของเสียงเพื่อการเรียนรู้

ลักษณะสำคัญของเสียงเพื่อการเรียนรู้ในที่นี้คือ สัมประสิทธิ์เซปสตรัมบนสเกลเมลดังที่กล่าวไว้ในหัวข้อการสกัดลักษณะสำคัญ โดยใช้โปรแกรม HCopy ซึ่งเป็นเครื่องมือที่อยู่ในชุดเครื่องมือ HTK มาใช้ในการสกัดลักษณะสำคัญจากสัญญาณเสียง โดยเลือกใช้พารามิเตอร์ต่างๆ ซึ่งจะระบุไว้ในไฟล์ code.config ดังนี้

```
# HTK Configuration Parameters for Generating MFCC_E _D_A

SOURCEFORMAT=WAV        # source forwat is WAV
SOURCEKIND = WAVEFORM # simple waveform
SOURCERATE = 625              # source sampling frequency is [16kHz]

# mel-frequency cepstral coefs, energy, their deltas, and accelerations

TARGETKIND = MFCC_E _D_A
```

```
TARGETRATE=100000.0     # frame interval is 10msec
WINDOWSIZE=250000.0     # window length is 25msec
USEHAMMING=T            # use HAMMING window
PREEMCOEF=0.97          # apply highpass filtering
NUMCHANS=24             # number of filterbank for MFCC is 24
NUMCEPS=12              # number of parameters for MFCC presentation

# Rather local Parameters

ENORMALISE=T            # normalize energy
ESCALE=1.0              # energy scale is 1.0
```

เมื่อใช้พารามิเตอร์ต่างๆที่กำหนด โปรแกรม HCopy จะคำนวณหาสัมประสิทธิ์เซปสตรัมบนสเกลเมลโดยมีค่าพลังงานรวมอยู่ด้วย อัตราการเปลี่ยนแปลง (delta) และความเร่ง (accelerations) จากสัญญาณเสียงทุกๆกรอบเวลายาว 25 มิลลิวินาที โดยแต่ละกรอบเวลาจะมีระยะเวลาห่างกัน 10 มิลลิวินาที ได้ออกมาเป็นเวกเตอร์ลักษณะสำคัญที่มีขนาด 39 มิติ โดยต่อจากนี้จะขออ้างอิงชุดลักษณะสำคัญที่ด้วยสัญลักษณ์ MFCC_E_D_A

การใช้งานโปรแกรม HCopy จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HCopy -C code.config -S CCHCopy.scp
```

โดยที่

- code.config เป็นไฟล์ที่กำหนดค่าพารามิเตอร์ในการหาลักษณะสำคัญของเสียง
- –C code.config เป็นการกำหนดให้ใช้ไฟล์ code.config เป็นไฟล์กำหนดค่าพารามิเตอร์ในการหาลักษณะสำคัญของเสียง
- CCHCopy.scp เป็นไฟล์ข้อความซึ่งแต่ละบรรทัดเป็นชื่อของไฟล์สัญญาณเสียง .wav ที่เป็นอินพุต และไฟล์ลักษณะสำคัญที่เป็นเอาท์พุตโดยมีเครื่องหมายเว้นวรรคคั่นกลาง ตัวอย่างเช่น CCF001_Pa001_001.wav CCF001_Pa001_001.mfc เป็นการกำหนดให้ใช้ไฟล์ CCF001_Pa001_001.wav เป็นไฟล์สัญญาณเสียงอินพุต และให้เขียนไฟล์เอาท์พุตชื่อ CCF001_Pa001_001.mfc ออกมา

- -S CCHCopy.scp เป็นการกำหนดให้ใช้ไฟล์ CCHCopy.scp เป็นไฟล์กำหนดไฟล์สัญญาณเสียง .wav ที่เป็นอินพุต และไฟล์ลักษณะสำคัญที่เป็นเอาท์พุต ในการหาลักษณะสำคัญของเสียง

1.2 การเตรียมข้อมูลเพื่อนำไปใช้ในการเรียนรู้

ข้อมูลที่ต้องเตรียมเพื่อนำไปใช้ในการเรียนรู้การรู้จำเสียงพูดแบ่งออกเป็นสองส่วนคือ ส่วนที่เป็นไฟล์กำหนดพารามิเตอร์และลักษณะการเรียนรู้ซึ่งได้แก่

- ไฟล์พจนานุกรม (Dictionary File) เป็นไฟล์ข้อความแสดงรายการของคำศัพท์ของเครื่องรู้จำเสียงพูด โดยในงานวิจัยนี้ต้องการสร้างเครื่องรู้จำที่รู้จำเสียงพูดในระดับหน่วยเสียง ดังนั้นพจนานุกรมที่ใช้จึงอยู่ในรูปหน่วยเสียงทั้งหมดแทนที่จะอยู่ในรูปของคำ โดยในที่นี้จะใช้ชื่อไฟล์เป็น CCphn.dict
- ไฟล์รายการหน่วยเสียง (Phone List File) เป็นไฟล์ข้อความแสดงรายการหน่วยเสียงทั้งหมด โดยแต่ละบรรทัดจะแทนแต่ละหน่วยเสียง โดยในที่นี้จะใช้ชื่อไฟล์เป็น CCphn.list
- ไฟล์ต้นแบบทอพอโลยีของแบบจำลองเสียง (HMM Prototype File) เป็นไฟล์ที่เก็บพารามิเตอร์ตั้งต้นของแบบจำลองฮิดเดนมาร์คอฟเอาไว้ โดยจะนำมาใช้เป็นไฟล์ต้นแบบสำหรับสร้างแบบจำลองเสียง เราสามารถกำหนดทอพอโลยี ปรับเปลี่ยนพารามิเตอร์ตั้งต้น และจำนวนสถานะของแบบจำลองฮิดเดนมาร์คอฟได้จากไฟล์ต้นแบบนี้ ในที่นี้เลือกใช้แบบจำลองฮิดเดนมาร์คอฟที่มีจำนวนสถานะ 5 สถานะโดยมีทอพอโลยีเป็นแบบจากซ้ายไปขวา โดยในที่นี้จะใช้ชื่อไฟล์เป็น proto5s

ข้อมูลอีกส่วนหนึ่งจะเป็นไฟล์ข้อมูลอินพุตเพื่อการเรียนรู้ซึ่งได้แก่

- ไฟล์ลักษณะสำคัญ ซึ่งเป็นไฟล์ที่ได้จากขั้นตอนในหัวข้อการหาค่าลักษณะเพื่อการเรียนรู้ โดยใช้ไฟล์เสียงจากชุดข้อมูลเสียงเพื่อการเรียนรู้ดังที่ได้กล่าวไว้ในหัวข้อฐานข้อมูลเสียงเพื่อการแบ่งเสียงพูดเป็นเซกเมนต์เป็นอินพุต รวมแล้วจะได้ไฟล์ลักษณะสำคัญเพื่อการเรียนรู้ 840 ไฟล์
- ไฟล์กำกับหน่วยเสียง (Label File) ซึ่งเป็นไฟล์ข้อความที่แสดงลำดับของหน่วยเสียงและระยะเวลาของการเกิดหน่วยเสียงต่างๆของแต่ละไฟล์เสียง โดยไฟล์เหล่านี้จะมีมาให้อยู่แล้วในชุดข้อมูลเสียงเพื่อการเรียนรู้ รวมแล้วจะได้ไฟล์กำกับหน่วยเสียง 840

ไฟล์เท่ากันกับจำนวนของไฟล์ลักษณะสำคัญ และ ไฟล์เสียงของชุดข้อมูลเสียงเพื่อการเรียนรู้

- ไฟล์กำกับหน่วยเสียงสำคัญ (Master Label File) ซึ่งเป็นไฟล์ข้อความที่รวบรวมไฟล์กำกับหน่วยเสียงย่อยทั้งหมดมาไว้ในไฟล์เดียว เพื่อสะดวกต่อการนำไปใช้ในการเรียนรู้ โดยในที่นี้จะใช้ชื่อไฟล์เป็น CCphn.mlf

1.3 การเรียนรู้

การเรียนรู้ในที่นี้ใช้แบบจำลองฮิดเดนมาร์คอฟดังที่กล่าวไว้ในบทที่ 2 โดยใช้ชุดเครื่องมือ HTK เป็นเครื่องมือสำหรับสร้างเครื่องรู้จำเสียงพูด โดยนำมาใช้ในการเรียนรู้แบบจำลองภาษา และแบบจำลองเสียงพูด โดยแบบจำลองเสียงพูดที่เรียนรู้จะมีจำนวนเท่ากับจำนวนหน่วยเสียงทั้งหมดที่ต้องการรู้จำ ขั้นตอนการเรียนรู้ในที่นี้แบ่งออกเป็นสองขั้นตอนคือ การเรียนรู้แบบจำลองภาษา และการเรียนรู้แบบจำลองเสียง โดยมีรายละเอียดดังนี้

1.3.1 การเรียนรู้แบบจำลองภาษา

การเรียนรู้แบบจำลองภาษาในที่นี้ ใช้แบบจำลองภาษาแบบอาศัยค่าความน่าจะเป็นของการที่หน่วยเสียงหนึ่งจะปรากฏอยู่ติดกับอีกหน่วยเสียงหนึ่ง (Bigram Language Model) โดยใช้โปรแกรม HLStats มาวิเคราะห์เก็บรวบรวมสถิติค่าความน่าเป็นออกมาจากไฟล์กำกับหน่วยเสียงสำคัญ การใช้งานโปรแกรม HLStats จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HLStats –C code.config -b CCphn.big CCphn.lst CCphn.mlf
```

โดยที่

- -C code.config เป็นการกำหนดให้ใช้ไฟล์ code.config เป็นไฟล์กำหนดค่าพารามิเตอร์ในการหาลักษณะสำคัญของเสียง
- -b CCphn.big เป็นการกำหนดให้เก็บสถิติค่าความเป็นของการที่หน่วยเสียงหนึ่งจะปรากฏอยู่ติดกับอีกหน่วยเสียงหนึ่งเก็บอยู่ในไฟล์ชื่อ CCphn.big
- CCphn.list คือไฟล์รายการหน่วยเสียงที่ใช้เป็นอินพุต
- CCphn.mlf คือไฟล์กำกับหน่วยเสียงสำคัญที่ใช้เป็นอินพุต

แบบจำลองภาษาที่ใช้จะอยู่ในรูปของโครงข่ายไวยากรณ์ของการเกิดหน่วยเสียงต่างๆ โดยใช้โปรแกรม HBuild มารับไฟล์สถิตค่าความเป็นของการที่หน่วยเสียงหนึ่งจะปรากฏอยู่ติดกับอีกหน่วยเสียงเข้ามาเป็นอินพุตแล้วแปลงให้อยู่ในรูปของโครงข่ายไวยากรณ์

การใช้งานโปรแกรม HBuild จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HBuild -C code.config -n CCphn.big CCphn.lst CCphn.net
```

โดยที่

- –C code.config เป็นการกำหนดให้ใช้ไฟล์ code.config เป็นไฟล์กำหนดค่าพารามิเตอร์ในการหาลักษณะสำคัญของเสียง
- -n CCphn.big เป็นการกำหนดให้ใช้ไฟล์ CCphn.big เป็นไฟล์เก็บสถิตค่าความเป็นของการที่หน่วยเสียงหนึ่งจะปรากฏอยู่ติดกับอีกหน่วยเสียง
- CCphn.list คือไฟล์รายการหน่วยเสียงที่ใช้เป็นอินพุต
- CCphn.net คือไฟล์แบบจำลองภาษาที่เป็นเอาท์พุต

1.3.2 การเรียนรู้แบบจำลองเสียงพูด

การเรียนรู้แบบจำลองเสียงพูดในที่นี้ จะสร้างแบบจำลองเสียงพูดสำหรับหน่วยเสียงในภาษาไทยทั้งหมด ให้เป็นแบบจำลองเสียงพูดแบบไม่ขึ้นกับบริบทรอบข้าง (Context-independent Phone Model) โดยจะใช้การกระจายของค่าความน่าจะเป็นแบบเกาส์เซียนที่มีเมทริกซ์ความแปรปรวนร่วมเกี่ยวตามแนวทแยง (Diagonal Covariance Gaussian Distribution) มาประมาณความน่าจะเป็นของการเกิดลักษณะสำคัญในแบบจำลองฮิดเดนมาร์คอฟ การเรียนรู้แบบจำลองเสียงมีขั้นตอนต่างๆดังนี้

1. การกำหนดลักษณะทอพอโลยี

ในลำดับแรกเราจะต้องกำหนดลักษณะทอพอโลยีตั้งต้นของแบบจำลองเสียงพูด โดยใช้โปรแกรม HCompV และไฟล์ต้นแบบทอพอโลยีของแบบจำลองเสียง ดังที่กล่าวไว้ในหัวข้อการเตรียมข้อมูลเพื่อใช้ในการเรียนรู้ การใช้งานโปรแกรม HCompV จะต้องพิมพ์คำสั่งดังนี้

```
HCompV -C train.config -f 0.01 -m –S CCTrain.scp -M hmm_0 proto5s
```

โดยที่

- -C train.config เป็นการกำหนดให้ใช้ไฟล์ train.config เป็นไฟล์กำหนดค่าพารามิเตอร์ในการเรียนรู้ โดยจะระบุพารามิเตอร์สามอย่างได้แก่ SOURCEKIND TARGETKIND และ RAWENERGY โดยใช้ค่าเดียวกันกับพารามิเตอร์ในไฟล์ code.config ทุกประการ
- -f 0.01 เป็นการกำหนดให้ค่าความแปรปรวนชักลง (Variance floor) เป็น 0.01
- -m เป็นการกำหนดให้มีการคำนวณพารามิเตอร์ ค่าเฉลี่ย เช่นเดียวกันกับคำนวณค่าความแปรปรวน
- -S CCTrain.scp เป็นการกำหนดให้ใช้ไฟล์ลักษณะสำคัญที่แสดงเป็นรายการอยู่ในไฟล์ CCTrain.scp มาใช้ในการเรียนรู้
- -M hmm_0 เป็นการกำหนดให้เก็บเอาท์พุตแบบจำลองเสียงพูดที่ได้ไว้ในไดเร็กทอรีชื่อ hmm_0
- proto5s เป็นการกำหนดให้ใช้ลักษณะทอพอโลยีที่กำหนดไว้แล้วในไฟล์ proto5s

ผลลัพท์ที่ได้คือไฟล์เป็นภาพรวมของแบบจำลองเสียงชื่อ macros และไฟล์ค่าความแปรปรวนของแต่ละลักษณะสำคัญชื่อ vFloor ซึ่งอยู่ในไดเร็กทอรี hmm_0

2. **การกำหนดค่าพารามิเตอร์ตั้งต้น**

ขั้นตอนถัดไปจะเป็นการกำหนดค่าตั้งต้นให้กับพารามิเตอร์ต่างๆให้กับแบบจำลองเสียงพูดของแต่ละหน่วยเสียง โดยใช้โปรแกรม HInit มากำหนดพารามิเตอร์ตั้งต้นจากลักษณะสำคัญที่ได้จากแต่ละหน่วยเสียง

การใช้งานโปรแกรม HInit จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HInit -I CCphn.mlf -S CCTrain.scp -H hmm_0/macros -C train.config -M hmm_1 -l $phn
proto5s
```

โดยที่

- -I CCphn.mlf เป็นการกำหนดให้โปรแกรม HInit ไปอ่านข้อมูลจากไฟล์กำกับหน่วยเสียงสำคัญชื่อ CCphn.mlf ขึ้นมาเพื่อดูว่าหน่วยเสียงแต่ละแบบอยู่ในช่วงเวลาใดบ้าง เพื่อที่จะได้เลือกลักษณะสำคัญในช่วงเวลาดังกล่าวมากำหนดค่าพารามิเตอร์ของแบบจำลองเสียงพูดของหน่วยเสียงนั้นๆ
- -S CCTrain.scp เป็นการกำหนดให้ใช้ไฟล์ลักษณะสำคัญที่แสดงเป็นรายการอยู่ในไฟล์ CCTrain.scp มาใช้ในการเรียนรู้

- -H hmm_0/macros เป็นการกำหนดให้อ่านไฟล์ภาพรวมของแบบจำลองเสียงที่อยู่ในไดเร็กทอรี hmm_0 เข้ามา
- -C train.config เป็นการกำหนดให้ใช้ไฟล์ train.config เป็นไฟล์กำหนดค่าพารามิเตอร์เพื่อการเรียนรู้
- -M hmm_1 เป็นการกำหนดให้เก็บเอาท์พุตแบบจำลองเสียงพูดที่ได้ไว้ในไดเร็กทอรีชื่อ hmm_1
- -l *$phn* เป็นการกำหนดพารามิเตอร์ให้แบบจำลองเสียงพูดของหน่วยเสียง *$phn*
- proto5s เป็นการกำหนดให้ใช้ลักษณะทอพอโลยีที่กำหนดไว้แล้วในไฟล์ proto5s

โดยจะต้องเรียกใช้โปรแกรม HInit เพื่อประมาณพารามิเตอร์ตั้งต้นของแบบจำลองเสียงพูดของหน่วยเสียงทุกๆหน่วยเสียง เพื่อให้ได้แบบจำลองเสียงพูดครบตามที่ต้องการ

ผลที่ได้จะเป็นไฟล์แบบจำลองเสียงพูดของแต่ละหน่วยเสียงโดยมีชื่อไฟล์ตรงกับสัญลักษณ์หน่วยเสียงนั้นๆเก็บไว้ในไดเร็กทอรี hmm_1

3. **การประมาณค่าพารามิเตอร์**

ขั้นตอนถัดไปจะเป็นการประมาณค่าพารามิเตอร์ต่างๆของแบบจำลองเสียงพูด โดยใช้โปรแกรม HRest มาประมาณพารามิเตอร์จากลักษณะสำคัญที่ได้จากแต่ละหน่วยเสียง

การใช้งานโปรแกรม HRest จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HRest –I CCphn.mlf -S CCTrain.scp -H hmm_1\macros -M hmm_2 -l $phn hmm_1\$phn
```

โดยที่

- -I CCphn.mlf เป็นการกำหนดให้โปรแกรม HInit ไปอ่านข้อมูลจากไฟล์กำกับหน่วยเสียงสำคัญชื่อ CCphn.mlf ขึ้นมาเพื่อดูว่าหน่วยเสียงแต่ละแบบอยู่ในช่วงเวลาใดบ้าง เพื่อที่จะได้เลือกลักษณะสำคัญในช่วงเวลาดังกล่าวมากำหนดค่าพารามิเตอร์ของแบบจำลองเสียงพูดของหน่วยเสียงนั้นๆ
- -S CCTrain.scp เป็นการกำหนดให้ใช้ไฟล์ลักษณะสำคัญที่แสดงเป็นรายการอยู่ในไฟล์ CCTrain.scp มาใช้ในการเรียนรู้
- -H hmm_1\macros เป็นการกำหนดให้อ่านไฟล์ภาพรวมของแบบจำลองเสียงที่อยู่ในไดเร็กทอรี hmm_1 ซึ่งได้จากขั้นตอนที่แล้วเข้ามา

- -M hmm_2 เป็นการกำหนดให้เก็บเอาท์พุตแบบจำลองเสียงพูดที่ได้ไว้ในไดเร็กทอรีชื่อ hmm_2
- -l *$phn* เป็นการกำหนดพารามิเตอร์ให้แบบจำลองเสียงพูดของหน่วยเสียง *$phn*
- hmm_1*$phn* เป็นการกำหนดให้นำไฟล์แบบจำลองเสียงพูดของหน่วยเสียง *$phn* ที่ได้จากขั้นตอนที่แล้วมาใช้เป็นอินพุต

ผลที่ได้คือไฟล์แบบจำลองเสียงพูดของแต่ละหน่วยเสียงเก็บไว้ในไดเร็กทอรี hmm_2

ต่อจากนั้นจะเป็นการประมาณพารามิเตอร์ซ้ำๆกันอีกหลายครั้งเพื่อให้ได้ค่าพารามิเตอร์ที่ใกล้เคียงกันกับลักษณะสำคัญของเสียงพูดมากยิ่งขึ้น โดยในครั้งนี้จะใช้โปรแกรม HERest มาประมาณพารามิเตอร์แทนการใช้โปรแกรม HRest เนื่องจากโปรแกรม HERest สามารถประมาณค่าพารามิเตอร์ของทุกๆแบบจำลองเสียงไปพร้อมกันได้แบบคู่ขนาน โดยในที่นี้เราจะประมาณค่าพารามิเตอร์ซ้ำเป็นจำนวน 10 รอบ ได้ผลลัพท์ออกมาเป็นแบบจำลองเสียงพูดของแต่ละหน่วยเสียงเก็บอยู่ในไดเร็กทอรี hmm_12

การใช้งานโปรแกรม HERest จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
HERest -C train.config -I CCphn.mlf -S CCTrain.scp -H hmm_$i\macros -M hmm_$j
CCphn.list
```

โดยที่

- -C train.config เป็นการกำหนดให้ใช้ไฟล์ train.config เป็นไฟล์กำหนดค่าพารามิเตอร์ในการเรียนรู้ โดยจะระบุพารามิเตอร์สามอย่างได้แก่ SOURCEKIND TARGETKIND และ RAWENERGY โดยใช้ค่าเดียวกันกับพารามิเตอร์ในไฟล์ code.config ทุกประการ
- -I CCphn.mlf เป็นการกำหนดให้โปรแกรม HInit ไปอ่านข้อมูลจากไฟล์กำกับหน่วยเสียงสำคัญชื่อ CCphn.mlf ขึ้นมาเพื่อดูว่าหน่วยเสียงแต่ละแบบอยู่ในช่วงเวลาใดบ้าง เพื่อที่จะได้เลือกลักษณะสำคัญในช่วงเวลาดังกล่าวมากำหนดค่าพารามิเตอร์ของแบบจำลองเสียงพูดของหน่วยเสียงนั้นๆ
- -S CCTrain.scp เป็นการกำหนดให้ใช้ไฟล์ลักษณะสำคัญที่แสดงเป็นรายการอยู่ในไฟล์ CCTrain.scp มาใช้ในการเรียนรู้
- -H hmm_*$i*\macros เป็นการกำหนดให้อ่านไฟล์ภาพรวมของแบบจำลองเสียงที่อยู่ในไดเร็กทอรี hmm_*$i* เข้ามา

- -M hmm_*$j* เป็นการกำหนดให้เก็บเอาท์พุตแบบจำลองเสียงพูดที่ได้ไว้ในไดเร็กทอรีชื่อ hmm_*$j* เมื่อให้ตัวแปร $\$j = \$i - 1$
- CCphn.list คือไฟล์รายการหน่วยเสียงที่ใช้เป็นอินพุต

2. กระบวนการรู้จำ

กระบวนการรู้จำเป็นการนำผลที่ได้จากขั้นตอนการเรียนรู้ ซึ่งในที่นี้คือแบบจำลองเสียงพูดและแบบจำลองภาษาที่ผ่านการเรียนรู้แล้ว รวมกันเป็นเครื่องรู้จำเสียงพูดที่รู้จำเสียงพูดได้ในระดับหน่วยเสียง

กระบวนการรู้จำมีขั้นตอนต่างๆดังต่อไปนี้

2.1 การหาค่าลักษณะสำคัญของเสียงเพื่อการรู้จำ

ลักษณะสำคัญของเสียงเพื่อการรู้จำจะใช้สัมประสิทธิ์เซปสตรัมบนสเกลเมลเหมือนกันกับค่าลักษณะสำคัญของเสียงเพื่อการเรียนรู้ โดยค่าพารามิเตอร์ที่ใช้ในการหาลักษณะสำคัญของเสียงในที่นี้จำเป็นต้องเหมือนกับค่าที่ใช้ในการหาลักษณะสำคัญของเสียงเพื่อการเรียนรู้ทุกประการ

2.2 การเตรียมข้อมูลเพื่อนำไปใช้ในการรู้จำ

ข้อมูลที่ต้องเตรียมเพื่อนำไปใช้ในการรู้จำเสียงพูดในที่นี้ได้แก่

- ไฟล์ลักษณะสำคัญ ซึ่งเป็นไฟล์ที่ได้จากขั้นตอนการหาค่าลักษณะสำคัญของเสียงเพื่อการรู้จำโดยใช้ไฟล์เสียงจากชุดข้อมูลเสียงเพื่อการทดสอบดังที่ได้กล่าวไว้ในหัวข้อฐานข้อมูลเสียงเพื่อการแบ่งเสียงพูดเป็นเซกเมนต์ รวมแล้วจะได้ไฟล์ลักษณะสำคัญเพื่อการรู้จำ 840 ไฟล์
- ไฟล์กำกับหน่วยเสียง (Label File) ซึ่งเป็นไฟล์ข้อความที่แสดงลำดับของหน่วยเสียงและระยะเวลาของการเกิดหน่วยเสียงต่างๆของแต่ละไฟล์เสียง โดยไฟล์เหล่านี้จะมีมาให้อยู่แล้วในชุดข้อมูลเสียงเพื่อการทดสอบ รวมแล้วจะได้ไฟล์กำกับหน่วยเสียง 840 ไฟล์

2.3 การรู้จำ

การรู้จำจะทำโดยการป้อนข้อมูลที่ได้จากกระบวนการเตรียมข้อมูลเพื่อนำไปใช้ในการรู้จำเข้าเป็นอินพุตของเครื่องรู้จำเสียงพูดที่ได้จากการเรียนรู้ เครื่องรู้จำเสียงพูดจะคำนวณหาลำดับของหน่วยเสียงที่ให้ค่าความน่าจะเป็นสูงสุด N อันดับ ซึ่งคิดมาจากความน่าจะเป็นของแบบจำลองเสียง และแบบจำลองภาษาร่วมกัน โดยการรู้จำจะใช้โปรแกรม HVite มารู้จำเสียงพูดด้วยการพิมพ์คำสั่งดังนี้

```
HVite -S CCTest.scp -H hmm_12/macros -w CCphn.net -n $i N -l '*' -i recout.mlf
CCphn.dict CCphn.list
```

โดยที่

- -S CCTest.scp เป็นการกำหนดให้ใช้ไฟล์ลักษณะสำคัญที่แสดงเป็นรายการอยู่ในไฟล์ CCTest.scp มาใช้ในการเรียนรู้
- -H hmm_12\macros เป็นการกำหนดให้อ่านไฟล์ภาพรวมของแบบจำลองเสียงที่อยู่ในไดเร็กทอรี hmm_12 เข้ามา
- -w CCphn.net เป็นการกำหนดใช้ใช้ไฟล์แบบจำลองภาษาชื่อ CCphn.net
- -n i N$ เป็นการกำหนดให้เครื่องรู้จำทำงานโดยให้ผลลัพธ์ที่มีค่าความน่าจะเป็นสูงสุด N อันดับแรกออกมา
- -l '*' -i recout.mlf เป็นการกำหนดเก็บผลลัพธ์การรู้จำไว้ที่ไฟล์ recout.mlf โดยไฟล์นี้จะมีรูปแบบเนื้อความเหมือนกับไฟล์กำกับหน่วยเสียงสำคัญ
- CCphn.dict คือไฟล์พจนานุกรมหน่วยเสียง
- CCphn.list คือไฟล์รายการหน่วยเสียง

3. การพิจารณาขอบเขตของหน่วยเสียงจากผลการรู้จำเสียงพูด

การตรวจหาขอบเขตของหน่วยเสียงแบบอาศัยเครื่องรู้จำเสียงพูด จะอาศัยผลที่ได้จากการรู้จำเสียงพูด โดยเลือกเฉพาะข้อมูลส่วนที่เป็นขอบเขตทางเวลาของหน่วยเสียงออกมาเป็นขอบเขตของหน่วยเสียง ซึ่งขอบเขตของหน่วยเสียงนี้จะมีความละเอียดอยู่ในระดับเดียวกันกับระยะห่างของแต่ละกรอบเวลาที่กำหนดเป็นพารามิเตอร์ไว้ในขั้นตอนการสกัดลักษณะสำคัญเพื่อการเรียนรู้และการรู้จำ โดยในที่นี้ความละเอียดจะอยู่ที่ 10 มิลลิวินาที

ตัวอย่างการตรวจหาขอบเขตของหน่วยเสียง เช่น เมื่อเนื้อความในไฟล์ผลการรู้จำมีลักษณะเป็นดังนี้

```
"*/CCF003_Pa003_012.rec"
0 2100000 sil -1024.600843
2100000 3100000 thr -780.772339
3100000 3800000 l -575.756409
3800000 5300000 x -904.084290
5300000 6100000 b -618.631958
6100000 7000000 a -599.543152
7000000 8100000 ng^ -784.267822
8100000 8800000 sp -459.062744
8800000 9300000 p -439.476685
9300000 13600000 ii -2227.795166
13600000 15700000 sp -1454.947998
15700000 17600000 f -1381.702393
17600000 18300000 o -565.726929
18300000 19900000 n^ -1211.035156
19900000 20700000 k -639.093384
20700000 21300000 a -452.169922
21300000 22200000 m -695.446106
22200000 23900000 xx -1044.781738
23900000 24700000 l -573.536621
24700000 28800000 aa -1923.821899
28800000 29600000 f^ -599.704895
29600000 32200000 sil -1300.840332
///
0 2100000 sil -1024.600843
2100000 3100000 thr -780.772339
3100000 3800000 l -575.756409
3800000 5300000 x -904.084290
5300000 6100000 b -618.631958
6100000 7000000 a -599.543152
7000000 8100000 ng^ -784.267822
8100000 8800000 sp -459.062744
8800000 9300000 p -439.476685
9300000 13600000 ii -2227.795166
13600000 15700000 sp -1454.947998
15700000 17600000 f -1381.702393
17600000 18300000 o -565.726929
18300000 19900000 n^ -1211.035156
19900000 20700000 k -639.093384
20700000 21300000 a -452.169922
21300000 22100000 m -617.664970
22100000 23900000 aa -1122.411133
23900000 24700000 l -575.246582
24700000 28800000 aa -1923.821899
28800000 29600000 f^ -599.704895
29600000 32200000 sil -1300.840332
.
```

จะได้ลำดับของขอบเขตของหน่วยเสียง $L = \{0, 21, 31, 38, 53, 61, 70, 81, 88, 93, 136, 157, 176, 183, 199, 207, 213, 221, 222, 239, 247, 288, 296, 322\}$ โดยที่ขอบเขตของหน่วยเสียงแต่ละอันจะเป็นตัวเลขบอกตำแหน่งทางเวลาซึ่งอยู่ในหน่วย 10 มิลลิวินาที ตัวอย่างเช่นขอบเขตของหน่วยเสียงที่สามของ L แทนด้วยเลข 31 จะหมายความว่าขอบเขตของหน่วยเสียงนั้นอยู่ที่เวลา 310 มิลลิวินาทีของสัญญาณเสียง

การสกัดลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์

การคำนวณหาลักษณะสำคัญที่ได้จากการใช้สารสนเทศสวนสัทศาสตร์ในตารางที่ 3.1 จะใช้โปรแกรม Speech Filling System – SFS [33] ซึ่งเป็นเครื่องมือสำหรับวิเคราะห์และสกัดลักษณะสำคัญจากสัญญาณเสียง การสกัดค่าพลังงานในช่วงความถี่ต่างๆจะใช้คำสั่ง mktrack โดยเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
mktrack -l lowfreq -h highfreq -n 4 -r 100 file
```

โดยที่

- -l lowfreq –h highfreq เป็นการกำหนดให้หาค่าพลังงานในช่วงความถี่ตั้งแต่ lowfreq ไปจนถึง highfreq (หน่วยเฮิร์ต)
- -n 4 เป็นการกำหนดให้ใช้จำนวนอันดับตัวกรองสัญญาณทั้งสิ้น 4 อันดับ
- -r 100 เป็นการกำหนดอัตราสุ่มของค่าพลังงานเป็น 100 เฮิร์ต
- file เป็นการกำหนดไฟล์อินพุตที่ต้องการหาค่าพลังงาน

การสกัดค่าระดับการสั่นสะเทือนเส้นเสียงหรือความก้องของเสียงจะอาศัยค่าการกระจายของพลังงาน (Energy distribution) อัตราการตัดศูนย์ (Zero-crossing rate) และอัตสหสัมพันธ์ (Autocorrelation) มาประมาณระดับความก้องของสัญญาณเสียงโดยเรียกใช้คำสั่ง vdegree ด้วยการพิมพ์คำสั่งดังนี้

```
vdegree file
```

โดยที่

- file เป็นการกำหนดไฟล์อินพุตที่ต้องการหาค่าพลังงาน

การสกัดค่าระดับของภาวะความไม่เป็นคาบของจะใช้คำสั่ง noisanal มาประมาณระดับของภาวะความไม่เป็นคาบในสัญญาณเสียง โดยเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
noisanal -w  0.025file
```

โดยที่

- -w windowsize เป็นการกำหนดขนาดของกรอบเวลาที่ทำการวิเคราะห์ให้เป็น 25 มิลลิวินาที

การเรียนรู้และการจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน

ในงานวิทยานิพนธ์นี้เลือกใช้เอสวีเอ็มไลท์ SVM^{light} [34] มาเป็นเครื่องมือสำหรับสร้างตัวแบ่งแยกเอสวีเอ็ม โดยรายละเอียดเกี่ยวกับใช้งานเครื่องมือเพื่อการเรียนรู้และการจำแนกข้อมูลมีดังนี้

1. การเรียนรู้ของซัพพอร์เวกเตอร์แมชชีน

ข้อมูลที่ต้องเตรียมเพื่อนำไปใช้ในการเรียนรู้ในที่นี้ จะอาศัยไฟล์ลักษณะสำคัญที่ได้จากขั้นตอนการสกัดลักษณะสำคัญของเสียงในบทที่ 4 มาสร้างไฟล์เวกเตอร์ลักษณะสำคัญเพื่อการเรียนรู้ของเอสวีเอ็ม โดยไฟล์เวกเตอร์ลักษณะสำคัญจะเป็นไฟล์ข้อความที่แต่ละบรรทัดแทนลักษณะสำคัญของแต่ละกรอบเวลา ซึ่งแต่ละบรรทัดแสดงด้วยรูปแบบดังนี้

```
<target> <feature>:<value> <feature>:<value> ... <feature>:<value>
```

โดยที่

- <target> คือฉลากกำกับเวกเตอร์ลักษณะสำคัญโดยในที่นี้คือ +1 หรือ -1
- <feature> คือตัวเลขจำนวนเต็มสำหรับบอกลำดับของลักษณะสำคัญ
- <value> คือค่าตัวเลขจำนวนจริงของลักษณะสำคัญ

ตัวอย่างเช่น

```
-1 1:0.432 :0.12  3:0.2
```

แทนเวกเตอร์ที่มีฉลากกำกับเป็น -1 โดยเป็นเวกเตอร์ 3 มิติที่มีลักษณะสำคัญลำดับที่หนึ่งเป็น 0.43 ลักษณะสำคัญลำดับที่สองเป็น 0.12 และลำดับสุดท้ายเป็น 0.2

โดยในที่นี้เลือกเรียนรู้เอสวีเอ็มสองแบบคือแบบที่ใช้ฟังก์ชันเคอร์เนลเชิงเส้น และแบบที่ใช้ฟังก์ชันเคอร์เนลพหุนาม โดยใช้โปรแกรม svm_learn ที่มีมากับชุดเครื่องมือ SVM^{light}

การใช้งานโปรแกรม svm_learn จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
svm_learn -t 0 train_sonorant.tok linear_sonorant.svm
```

โดยที่

- -t 0 เป็นการกำหนดฟังก์ชันเคอร์เนลที่ต้องการใช้งานเป็นแบบเชิงเส้น ถ้าต้องการใช้ฟังก์ชันเคอร์เนลแบบพหุนามให้กำหนดเป็น 1
- train_sonorant.tok เป็นไฟล์เวกเตอร์ลักษณะสำคัญที่นำมาใช้เป็นตัวอย่างข้อมูลเพื่อการเรียนรู้เอสวีเอ็ม
- linear_sonorant.svm คือไฟล์แบบจำลองที่เป็นเอาท์พุต

2. การจำแนกข้อมูลด้วยซัพพอร์เวกเตอร์แมชชีน

การจำแนกข้อมูลด้วยซัพพอร์เวกเตอร์แมชชีน จะใช้โปรแกรม svm_classify ซึ่งมีอยู่ในชุดเครื่องมือ SVM^{light} โดยการใช้งานโปรแกรม svm_classify จะต้องเรียกใช้ด้วยการพิมพ์คำสั่งดังนี้

```
svm_classify example_file model_file output_file
```

โดยที่

- example_file คือไฟล์เวกเตอร์ลักษณะสำคัญที่ต้องการทดสอบ
- model_file คือไฟล์แบบจำลองเอสวีเอ็มที่ต้องการนำมาใช้แบ่งแยกข้อมูล
- output_file คือไฟล์ผลลัพธ์ค่าของฟังก์ชันตัดสินใจที่เป็นเอาท์พุต

ประวัติผู้เขียนวิทยานิพนธ์

นายไพโรจน์ ลีลาภัทรกิจ เกิดเมื่อวันที่ 27 มกราคม พ.ศ. 2527 ที่จังหวัดกรุงเทพฯ สำเร็จการศึกษาระดับมัธยมศึกษาตอนต้นและตอนปลายจากโรงเรียนเทพศิรินทร์ สำเร็จการศึกษาระดับปริญญาบัณฑิต ในสาขาวิชาวิศวกรรมคอมพิวเตอร์ จากคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยในปีการศึกษา 2548